

December 2019

Cloud Based IoT Architecture

NATHAN ROEHL

University of Wisconsin-Milwaukee

Follow this and additional works at: <https://dc.uwm.edu/etd>



Part of the [Computer Sciences Commons](#)

Recommended Citation

ROEHL, NATHAN, "Cloud Based IoT Architecture" (2019). *Theses and Dissertations*. 2333.
<https://dc.uwm.edu/etd/2333>

This Thesis is brought to you for free and open access by UWM Digital Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of UWM Digital Commons. For more information, please contact open-access@uwm.edu.

CLOUD BASED IOT ARCHITECTURE

by

Nathan Roehl

A Thesis Submitted in
Partial Fulfillment of the
Requirements for the Degree of

Master of Science
in Computer Science

at

The University of Wisconsin-Milwaukee

December 2019

ABSTRACT

CLOUD BASED IOT ARCHITECTURE

by

Nathan Roehl

The University of Wisconsin-Milwaukee, 2019
Under the Supervision of Professor Tian Zhao

The Internet of Things (IoT) and cloud computing have grown in popularity over the past decade as the internet becomes faster and more ubiquitous. Cloud platforms are well suited to handle IoT systems as they are accessible and resilient, and they provide a scalable solution to store and analyze large amounts of IoT data. IoT applications are complex software systems and software developers need to have a thorough understanding of the capabilities, limitations, architecture, and design patterns of cloud platforms and cloud-based IoT tools to build an efficient, maintainable, and customizable IoT application. As the IoT landscape is constantly changing, research into cloud-based IoT platforms is either lacking or out of date. The goal of this thesis is to describe the basic components and requirements for a cloud-based IoT platform, to provide useful insights and experiences in implementing a cloud-based IoT solution using Microsoft Azure, and to discuss some of the shortcomings when combining IoT with a cloud platform.

© Copyright by Nathan Roehl, 2020
All Rights Reserved

TABLE OF CONTENTS

Abstract	ii
List of Figures	vi
1 Introduction	1
1.1 IoT platforms	1
1.2 Requirements of IoT Applications	2
1.3 Cloud-based IoT platform	4
2 Related Works	9
3 Azure IoT Platform	13
3.1 IoT Devices and Gateways	14
3.2 Storage	19
3.3 Data Analysis	21
3.4 User interface	27
4 Experiences with Azure	29
4.1 Data Set	30
4.2 Azure Set Up	30
4.3 Resource Groups	31

4.4	IoT Hub and Storage	31
4.5	Stream Analytics	36
4.6	Azure Functions	39
4.7	Azure Machine Learning	46
4.8	Power BI	59
4.9	Experiment Observations	61
5	High Velocity and High Volume Data	64
5.1	Cloud IoT Pitfalls	64
5.2	Possible Solutions	69
5.2.1	Distributed Model	69
5.2.2	Cost Efficient Model	72
6	AWS vs Azure vs GCP	77
6.1	IoT Related Services	77
6.2	IoT Storage Services	78
6.3	Streaming, Analytics, and Visualization Services	80
7	Conclusion	82
	Bibliography	85

LIST OF FIGURES

3.1	Cloud IoT platform architecture, where solid line denotes direction of data flow and dashed line denotes direction of control.	13
3.2	Communication protocols between IoT devices and cloud gateway, where solid line denotes the direction of device-to-cloud messages and the dash line denotes the direction of cloud-to-device messages	14
3.3	Communication patterns between IoT devices and Azure <i>IoT/Event Hubs</i> , where solid line denotes the direction of device-to-cloud messages and the dash line denotes the direction of cloud-to-device messages	17
3.4	Relation between IoT devices and device identity store, device provisioning service, and device and application model.	18
3.5	Dataflow between IoT gateway, data transformation, stream processing, cold storage, warm storage, and data visualization modules.	19
3.6	Options of warm and cold storage with different query methods.	20
3.7	Possible dataflow paths for IoT analytics.	22
3.8	Stateful data analysis using durable <i>Azure Functions</i>	26

4.1	View of the <i>Azure portal</i> . The left most menu contains favorite services. When clicking on any of these tabs, a window opens to the right. These windows are called blades. The <i>All services</i> tab was selected, showing only a small portion of all available services.	29
4.2	Creating an <i>IoT Hub</i>	32
4.3	Different tiers for <i>IoT Hubs</i> . IoT Edge enabled means this tier can send/receive messages from IoT Edge devices. Device Management means this tier can use device twin, query, direct method, and jobs [64].	33
4.4	Real-time graphs showing incoming messages to the <i>IoT Hub</i>	34
4.5	Creating a storage account.	35
4.6	Overview of inputs and outputs for <i>Stream Analytics</i> . [101]	37
4.7	<i>Stream Analytics</i> query with multiple parts. First selection takes all data sent to an <i>IoT Hub</i> aliased as <code>IoTHubData</code> and outputs to <i>Table Storage</i> aliased as <code>OutTableStorage</code> . The second selection reads data from the same <i>IoT Hub</i> , except when the parameters of the <code>where</code> statement are true it executes an <i>Azure Function</i>	38
4.8	Results of testing query in Figure 4.7.	39
4.9	<i>Azure Functions</i> triggers, inputs, and output options [102]. Other services are available but were not included as they do not relate to IoT.	40
4.10	A functions trigger and bindings json file.	42
4.11	Classes used to help <i>Azure Function</i> read and write to table storage. See Figure 4.12 for main method of <i>Azure Function</i>	44
4.12	<i>Azure Function</i> used to read and write to table storage. See Figure 4.11 for classes required to complete function.	45

4.13	Plotting relations between each feature and Cycle in <i>Power BI Desktop</i> . Battery Current, Battery Voltage, Load Voltage, and Temperature are all instances of the data set when the Load Voltage dropped below 2.7. Capacity and Time are computed from the start of each cycle to the point where Load Voltage drops below 2.7.	48
4.14	Some visualization features comparing real to predicted values.	52
4.15	Underfitting, good fitting, and overfitting examples for machine learning [12].	53
4.16	Experiment using two different methods to train a model.	54
4.17	Result of converting Figure 4.16 experiment to a web service prediction model.	55
4.18	Automated Machine Learning results.	57
4.19	Output created using python script module to show prediction.	58
4.20	Power BI report showing prediction compared to actual values. Temperature (in Celsius) is the instance of the data set when the Battery Voltage dropped below 2.7. Capacity (in Coulomb) and Time (in seconds) are computed from the start of each cycle to the point where Battery Voltage drops below 2.7. . .	62
6.1	Comparison of cloud-based IoT services between AWS, Azure, and GCP. . .	78
6.2	Comparison of cloud-based storage services between AWS, Azure, and GCP.	79
6.3	Comparison of cloud-based streaming, analytics, and visualization services between AWS, Azure, and GCP.	80

1 Introduction

IoT applications are complex software systems that include software components for devices, gateways, message aggregation, data analysis, storage, and visualization. As an emerging technology, IoT has been used in areas such as predictive maintenance [43, 127], smart metering [1, 39], asset tracking [129, 130], connected vehicles [9, 35, 117], and fleet management [10]. These applications collect vast volumes of telemetry messages from distributed IoT devices and processed them in real-time to gain business insights and to take actions based on the insights. The sensitive data collected by IoT devices must be protected during data transmission and processing. Stable storage is needed to archive historical data. Scalable data analysis tools are needed to learn knowledge in real-time from streaming data or as offline analysis from historical data. Data visualization and graphic user interface are needed to provide visual feedback and user control of data analysis and IoT devices [36, 120]. Each one of these components is a software program that requires significant engineering effort. Composing them so that they work together properly is even more challenging. In this context, developers of IoT applications may find it more productive to use existing IoT frameworks and other commercial off-the-shelf software components to build their final products.

1.1 IoT platforms

In recent years, many IoT platforms have emerged to answer the demand for rapid prototyping of robust and scalable IoT applications. These IoT platforms include cloud-based system such as AWS, Google, Azure, IBM IoT [36, 118, 120] and standalone systems such as NodeRed [15], Thingworx [36, 120], and Thingsboard [38]. While differing in interfaces,

implementation, performance, and costs, these IoT systems all offer device clients, messaging hubs, stable storage, data analysis modules, data routing mechanisms, and data visualization tools. From the users' perspective, experience gained from working with one platform can be translated to others since these IoT platforms have similar design principles.

IoT applications are large software systems that are difficult to develop from the ground up. IoT platforms are designed to alleviate this burden. Despite the recent advances of IoT platforms, the complexity of IoT applications should not be underestimated. IoT applications have to be compatible with diverse and distributed sensors, able to process large volume of data in real-time, be tolerant to missing or erroneous data, runtime exceptions, and unreliable networks, scalable to big data analysis and storage, and extensible to complex workflow and business logic [36, 120]. As common to any complex software, utilization of appropriate software architecture and design patterns are essential to the success of IoT development projects. When making choices of the architecture and design patterns, it is important to understand the requirements of the IoT applications and balance them against the performance, limitations, and design tradeoffs of the software components of an IoT platform [36].

1.2 Requirements of IoT Applications

Different types of IoT applications have different functional and non-functional requirements. Based on these requirements, developers need to select the suitable architecture, design patterns, and software components to be used on an IoT platform. Functional requirements refer to the specific functionalities of a software system while non-functional requirements refer to the properties of the software application such as performance, security, fault tolerance, and cost.

For example, predictive maintenance and smart metering applications require device monitoring, abnormal event detection, and real-time and offline data analysis [1, 39, 43]. For applications such as asset tracking and fleet management, spatial locations of the sensor devices needs to be maintained [10, 129, 130]. For connected vehicles, field gateways are often required to collect data from vehicles of each road segment and generate traffic alerts and send aggregate information to the cloud servers [9, 35, 117]. For connected vehicles, there are hard real-time constraints, which require low latencies [9, 35, 117]. For smart metering and asset tracking, high throughput is needed to process large volume of telemetry data [1, 39, 129, 130]. For asset tracking and fleet management, offline analysis such as mixed integer linear programming is needed to find optimal scheduling [10, 129]. For predictive maintenance, offline analysis such as machine learning is needed to generate models that can be used to suggest proactive measures in real-time [43]. All these applications require graphic user interface (GUI) for data visualization and control, user management, device management, device provisioning, and data security. While functional requirements often are related to individual IoT components, the non-functional requirements such as data security, latency, scalability, and fault tolerance are cross-cutting concerns that affect the entire IoT system. Below is a summary of the functional and non-functional requirements of IoT applications.

1. Functional requirements

- (a) Real-time and/or offline data analysis.
- (b) Sensor data archive. Some types of data may be compressed.
- (c) Requirement and availability of gateways for IoT devices.
- (d) Real-time control of IoT devices.
- (e) GUI for data visualization and control.

- (f) Data processing workflow and/or business logic.
- (g) Distributed machine learning.
- (h) Device management and provisioning.
- (i) User management.

2. Non-functional requirements

- (a) Scalability to large number of IoT devices or telemetry messages.
- (b) Strong data security.
- (c) Hard real-time constraints with low latency or high throughput with soft real-time constraints.
- (d) Tolerance of faults such as erroneous data and unreliable network.
- (e) Cost requirements.

1.3 Cloud-based IoT platform

IoT applications can be implemented using on-premise servers or cloud-based platforms. The choice mostly depends on the non-functional requirements of the application such as latency and security. Cloud-based software systems have gained tremendous growth in recent years with many legacy software systems migrated to cloud platforms to offer software as services rather than installation-based products. The benefits of cloud-based software systems include transparent updates, distributed and load-balanced processing, scalable computing and storage, and flexible offering of diverse products and services [111]. The drawback is that users of the cloud-based system need to maintain network connectivity for some product features to function correctly. Also, unstable network or latencies of a slow network can cause unpredictable issues for applications with hard real-time constraints.

Fundamentally, a cloud platform is a collection of network-connected distributed computing and storage servers that provide virtual environments for different operation systems, application containers (such as Google's *app engine*), and computing services [111]. A software system can run on a cloud platform by deploying its instances on virtual machines of the cloud platform, though the advantage of this is limited to eliminating the need to acquire and maintain physical hardware. Users still have to set up the required computation environments such as dependency libraries, networking, and storage within the cloud servers. Application containers reduce some of these burdens in that the containers provide interfaces to computation environments that users do not need to separately set up for each deployment of the software system on a cloud platform.

Utilization of cloud-based platforms can be divided into four types: IaaS, PaaS, FaaS, and SaaS with increasing level of dependency on the cloud platform [28, 30, 111].

1. IaaS (Infrastructure as a Service) is the ability of a cloud platform to act as virtual machines for software applications that normally run in local servers. *Amazon EC2* (Elastic Compute Cloud) is such an example, where users can deploy software applications by launching instances of virtual servers, uploading the applications, and executing them. The virtual machines can be distributed and can use stable storage such as *Amazon S3* (Simple Storage Service) buckets.
2. PaaS (Platform as a Service) is a managed computation environment for specialized applications such as Web servlets. Google's *app engine* is such an example, where users can develop software programs using *app engine* development tool and deploy the programs to the *app engine* to execute. In PaaS, users still need to develop complete software programs as in IaaS except that users rely on the API of the platform for runtime support, network, and storage.

3. FaaS (Function as a Service) such as *AWS Lambda* and *Azure Functions* allow users to implement lightweight applications as stateless functions, which can be used for high throughput processing such as data transformation, filtering, and event detection. FaaS functions connect to storage and network through APIs. FaaS functions are usually considered serverless since there is no dedicated servers allocated for running the functions and the cost is based on calls to the functions.
4. SaaS (Software as a Service) refers to software applications hosted by cloud servers such as Dropbox, Microsoft Office 365, and Google Apps that deliver functionalities over the network.

Most of the software components of a cloud-based IoT platform fit the mode of PaaS (e.g. *Azure Event Hubs*, *Cosmos DB*, *Redis Cache*, and *Stream Analytics*), FaaS (e.g. *Azure Functions*), or SaaS (e.g. *Azure IoT Central*). *Azure IoT Central* is a software service for developing IoT applications through a Web interface, which is supposed to be simple to use with the least amount of control and customizability. PaaS and FaaS are used for customizable IoT applications that can be reasonably handled by the services such as *IoT Hub* and *Stream Analytics*. Solutions based on IaaS are usually reserved for customized and complex IoT applications that require a high-level of control. Developers can utilize open-source software stack such as Spark, Mesos, Akka, Cassandra, and Kafka to build customized IoT applications and deploy them in IaaS virtual machines. Since IaaS-based IoT applications need to reserve virtual machines, it is more suitable for computation-intensive applications with high rate of utilization of the virtual machines. SaaS such as *Azure IoT Central* is intended for simple IoT applications with limited functionalities while IaaS-based IoT applications are essentially locally developed applications deployed on IaaS virtual machines. Therefore, this research will focus on cloud-based IoT components provided in forms of PaaS and FaaS.

The contributions of this thesis are as follows.

1. The IoT landscape is constantly evolving, where past survey papers are either out of date or their scope of research focused on a few key aspects and not the entire system. This thesis presents a current in-depth analysis and description on many of the capabilities, limitations, architecture, design patterns, and services of modern cloud-based IoT platforms and applications, providing a better understanding of each part and how it relates to the whole.
2. Cloud platforms can offer an overwhelming amount of services to choose from, creating a steep learning curve for new users. Chapter 5 gives guidance for how to navigate a large cloud platform such as Microsoft Azure, and use it to implement an IoT application and avoid many of the pitfalls that are not immediately noticeable.
3. The combination of IoT with cloud platforms create an accessible, resilient, and scalable solution that can handle many devices and can store and process large volumes of data. The combination is not without its shortcomings, as there are limits to throughput and query efficiency. Two proposals are described, one distributed and one unique to the scenario focusing on reducing costs.
4. Lastly, corporations such as Amazon, Microsoft, and Google have become leading providers for cloud-based IoT platforms as demand for IoT solutions from industries such as healthcare, automotive, industrial, security, and many more increase. Many of the services from each platform are compared, showing current similarities and differences between the platforms.

The remainder of this thesis is organized in the following order. Chapter 2 reviews prior surveys on IoT systems. Chapter 3 describes the main aspects of IoT platforms using

Microsoft's Azure as a model. Chapter 4 shares our experiences using Microsoft Azure to create a small IoT application. Chapter 5 discusses the short comings of cloud-based IoT platforms and possible solutions. Chapter 6 compares the services offered by Amazon, Microsoft, and Google with respect to their cloud-based IoT platforms. Chapter 7 concludes. Portions of this thesis were published in a chapter for a book, where I was a contributing author [11].

2 Related Works

This Chapter reviews survey papers relating to IoT and cloud computing, which described general concepts and architecture, compared current cloud-based IoT platforms, explained analytics and its uses in IoT, and detailed the relation of IoT to big data.

Past research [40, 36, 33] has provided high level IoT reference architectures. Goyal *et al.* [40] discusses a five layer architecture that consists of perception, transport, processing, application, and business layers. The perception layer is the device layer which collects information. Transport layer is the connection created between device layer and processing layer. Processing layer analyzes the data that is collected. The application layer is software that interacts with perception and processing layer. The business layer manages the work of the entire system, such as privacy and security.

Hejazi *et al.* [36] proposed an IoT architecture consisting of four components, sensing, communication and identification, computation and cloud, and services and applications components. The sensing component consists of devices and data collection. Communication and identification connect devices to the computation and cloud components. Computation and cloud, also called IoT Integration Middleware, provide processing capabilities and acts as a combination layer for devices and applications. The services and applications components are the many tools provided by cloud platforms that collect, analyze, and visualize the data. They also discuss several considerations when selecting an IoT platform, such as stability, scalability, and cost as well as the benefits of using open source over proprietary software. Lastly, they provide a table comparing 20 different IoT platforms.

Guth *et al.* [33] noticed a lack of standards among the increasing number of IoT platforms, creating confusion and lack of comprehension when selecting or building an IoT platform.

They proposed an abstract IoT reference architecture that consists of five main layers. They are sensors/actuators, devices, gateways, IoT Integration Middleware, and application. Sensors and actuators are part of the devices, which send their data to a gateway. The gateway is responsible for bi-directional communication and authentication with devices. IoT Integration Middleware processes the data, and the application layer is software that uses the middleware layer to gain insights. Because the design is very high level, it allows for easier comparisons with existing IoT architectures such as OpenMTC, FIWARE, SiteWhere, and AWS IoT. It is clear that all three share similar design ideas, with minor variations on layers and naming.

Pflanzer *et al.* [118] listed and compared many cloud platforms with an emphasis on PaaS IoT related services, providing an extensive list of options for application developers. Platforms discussed range from large corporations like Amazon's AWS, Microsoft Azure, and Google Cloud Platform, to alternate options such as Heroku, CloudFoundry, and DreamFactory. Parse, bought and run by Facebook trying to compete with AWS, Azure, and GCP, is mentioned as an option but was discontinued.

Verma *et al.* [131] stressed the importance of real time analytics in IoT, with a focus on analytical network methodologies. IoT analytics can be separated into historical and proactive. Historical analytics focuses on learning insights from historical data. This category is separated into descriptive, such as energy use patterns, and diagnostic, generating disaster alerts. Proactive analytics can be separated into stream IoT analytics, where streams or batches of IoT data are monitored with low or no real-time constraints, and real-time IoT analytics. Real-time IoT analytics have harder real-time constraints and are becoming more popular. Real-time IoT analytics are separated into predictive, such as predicting power usage, and prescriptive, determining if or when repairs can be made to a machine. They argue that the massive amounts of IoT data create many challenges for real-time analytics,

and that current networking technologies, similar to ones used in cloud infrastructures, are not adequate to support real-time IoT analytics in their current state.

Siow *et al.* [126] provides an extensive survey on the many aspects of analytics in IoT. Several definitions of IoT are discussed, with four common principles to each. They exist at a global scale, are made of uniquely identifiable Things that monitor the physical world, the Things are connected so data can be shared, and they have the ability to impact our society through advanced services. Advanced services are use case scenarios in different industries, with a vast number of examples given spanning industries such as health, transport, living, environment, and industry. Five categories are discussed listing different analytical capabilities: prescriptive, prediction, discovery, diagnosis, and description. The category descriptions are similar to Verma *et al.*, except no hierarchy is established among the analytical types, and stream processing is replaced by discovery. An extensive list of infrastructure components are given, focusing on data generation, collection, aggregation, storage, and compute. Lastly, challenges are discussed with relation to analytics and IoT. They argue there is a trade-off between distribution and interoperability, performance, and analytical value.

Chen *et al.* [14] provides an in depth analysis of big data, with several sections discussing IoT as a big data producer. Currently IoT data is not the majority of big data, but by 2030 HP predicts it will be. They argue that even though the technologies are inter-dependent they should be developed jointly, as research and advancement in one area helps the other. Data generated from IoT devices have four common characteristics. First characteristic is large-scale data, where current and historical data are collected to generate large volumes of data. Second is heterogeneity, as having a variety of devices means a variety of data is collect. Third is a strong time and space correlation. Devices are placed in specific locations where each message is time stamped, creating a time and space correlation when performing

analysis. Lastly, much of the data collected will not be useful in many cases. There example is of a traffic camera, where the only useful footage is when accidents or violations occur. The rest of the data is normal traffic with little value.

Past research surveys of cloud computing and IoT either lack specifics [40, 36, 33], are out of date [118], provide in depth analysis of specific parts of an IoT platform [131, 126], or focus on a different subject that overlaps with IoT [14]. Our research aims to fix these shortcomings by providing an up to date, comprehensive, and focused survey of the requirements and parts of a cloud-based IoT platform. Not only does our research improve on past research in these specific areas, it also provides insights and experiences using Microsoft Azure to implement a cloud-based IoT solution.

3 Azure IoT Platform

The architecture of cloud-based IoT platforms are mostly driven by the functional requirements of IoT applications. The main requirements are to collect and analyze information collected from IoT sensors, to store the collected information, to report the results of analysis to users, and to integrate the analysis results with business systems [40]. The non-functional requirements such as security and latency are cross-cutting concerns that impact multiple components of the IoT platform. In this chapter, we discuss the main components of an IoT application, design choices, and related crosscutting concerns using Azure IoT platform for examples and as a guide.

Azure IoT platform refers to a suite of software tools offered on Azure cloud platform that can be used in combination to build IoT applications. While some software tools such as *Azure IoT Hub* [62], *Stream Analytics* [77], and *Power BI* [80] are designed IoT applications, others such as *Blob storage* [61], *Cosmos DB* [78], and *Azure Functions* [47] are for general-purpose applications but frequently used in IoT applications as part of the solution. The overall architecture of a cloud-based IoT system is shown in Figure 3.1.

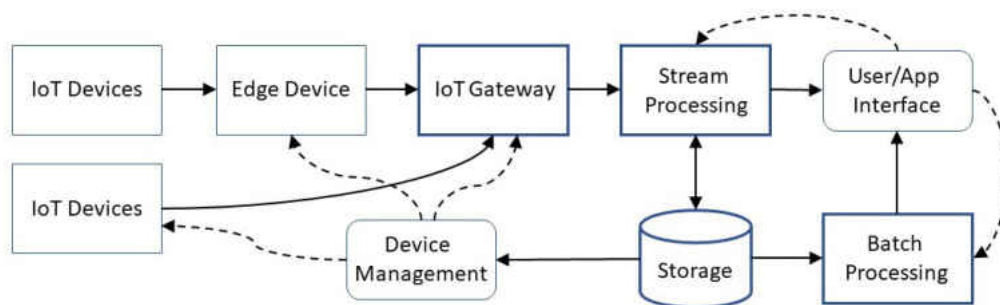


Figure 3.1: Cloud IoT platform architecture, where solid line denotes direction of data flow and dashed line denotes direction of control.

3.1 IoT Devices and Gateways

The primary functionality of an IoT application is to enable IoT devices to send telemetry messages to a cloud gateway which enables further processing of the messages. To implement this functionality, a messaging protocol must be chosen for communication between IoT devices and cloud gateways. IoT client must be implemented for the IoT devices so that they can establish secure communication with a specific cloud gateway. In order to manage the IoT devices, an IoT application must also maintain storage for device identities, device entities, and device provisioning.

Messaging protocol IoT devices communicate with IoT gateways through messaging protocols such as MQTT, AMQP, and HTTP as shown in Figure 3.2.

1. MQTT (Message Queue Telemetry Transport) is a lightweight client-server messaging protocol. It has a small memory footprint on the device and sends compact messages that use less network bandwidth [119]. Thus, MQTT is a popular choice for devices with limited computing resources and network connectivity. However, the compact messaging format also presents challenges to IoT applications. Since MQTT does not support message metadata in its header, IoT clients and gateways must agree on the

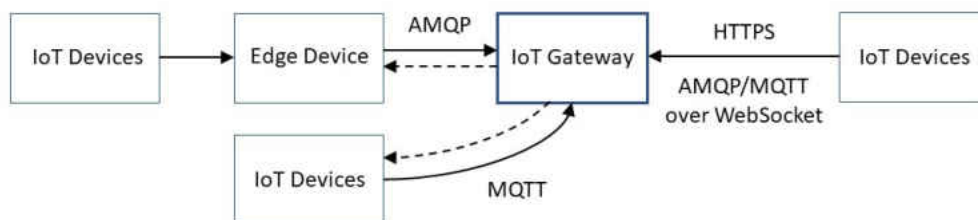


Figure 3.2: Communication protocols between IoT devices and cloud gateway, where solid line denotes the direction of device-to-cloud messages and the dash line denotes the direction of cloud-to-device messages

format of the message content. MQTT has three levels of *quality of service* related to the message delivery: *at most once*, *at least once*, and *exactly once* delivery [114]. Azure *IoT Hub* supports the first two but not the last one since it increases the latency while reducing the availability of distributed IoT infrastructure.

2. AMQP (Advanced Message Queuing Protocol) is a connection-oriented, multiplexing messaging protocol with compact message format, which is suitable for devices that require long-running connections and high-throughput communication with cloud gateways [112]. Both AMQP and MQTT are implemented over TCP layer. While MQTT is more lightweight and suitable for devices that connect to cloud gateway directly with per-device credentials, AMQP allows more complex device topology so that multiple devices can connect with cloud gateway using the same secure connection.
3. IoT devices can also directly communicate with Azure *IoT Hub* via HTTP though the implementation is more complex and the message overhead is higher than the binary format of MQTT and AMQP messages [112]. Moreover, cloud-to-device messaging is less efficient over HTTP since the device has to periodically poll messages while AMQP and MQTT can push cloud messages to a device with far less latency [112]. AMQP and MQTT can run over WebSocket for devices with restriction on open ports.

Device Client There are a variety of IoT devices with different processors, operating systems, networking capabilities, and runtime support of programming languages. For example, Intel NUC runs Ubuntu Linux while Raspberry Pi can run Windows 10. For devices that run an operation system, IoT clients may be implemented using Azure IoT SDK [55] using various programming languages such as Java, C, .NET, Python, and JavaScript (Node.js). For the devices supported by Azure IoT SDK, the implementation of IoT clients is a relatively straightforward process that first establishes connection to cloud gateway with device

credentials and then starts the loop of encoding messages and sending them asynchronously to the gateway. For FPGA-based devices such as CompactRIO, IoT clients can be developed using vendor-specific languages such as LabView, though such client may not be compatible with specific cloud gateways such as *Azure IoT Hub* due to its security requirement.

Cloud Gateway IoT devices may be connected to cloud gateway directly or connect through field gateways such as edge devices, which then connect to the cloud gateway. In the former case, IoT client program runs in IoT devices while in the latter case, the IoT client runs in the field gateways. Field gateways may be used in cases where the IoT devices are low-power sensors that lack the computing power and network connectivity to reliably transmit IoT messages. Field gateways may be used to aggregate, filter, and preprocess raw telemetries from IoT sensors to reduce network congestion and workload of cloud gateway. Field gateway may also be used to connect to devices such as CompactRIO that lack suitable implementation of secure communication layers or to devices that can only be connected with industrial protocols such as OPC UA. In some cases, a custom cloud gateway may be developed using *Azure IoT protocol gateway* [48] to perform the functions of a field gateway where the connection from IoT devices to the custom gateway may be secured using network tunneling such as VPN (Virtual Private Network).

Azure platform supports two types of IoT gateways: *Azure IoT Hub* and *Azure Event Hubs* with available connection patterns shown in Figure 3.3. *Azure IoT Hub* is more suitable for applications that require bidirectional messaging between devices and cloud gateways [62]. *Azure Event Hubs* is more suitable for applications that only need to send large number of messages to cloud gateway [58] and it does not support MQTT protocol. MQTT is based on publisher-subscriber model where message producers register as publishers in a MQTT broker and message consumers also register at the broker to listen on messages of

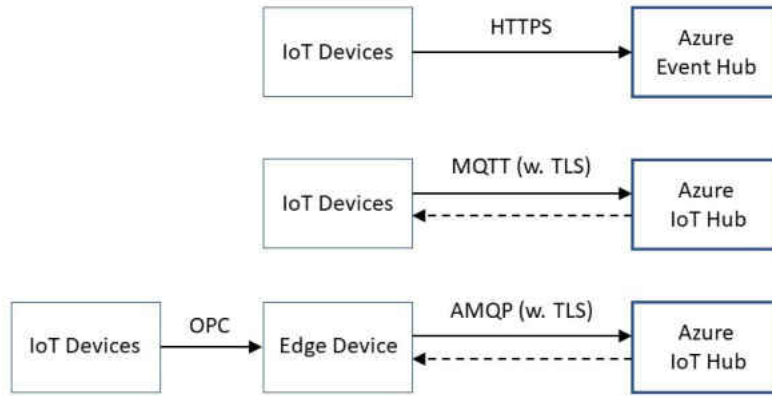


Figure 3.3: Communication patterns between IoT devices and Azure *IoT/Event Hubs*, where solid line denotes the direction of device-to-cloud messages and the dash line denotes the direction of cloud-to-device messages

selected topics. *Azure IoT Hub* is such a broker except that MQTT protocol does not specify security requirement. To protect sensitive IoT data transmitted over untrusted networks, *Azure IoT Hub* requires connections be established over TLS (Transportation Layer Security) protocol [60]. In order to connect to *Event Hubs*, the clients need to have the access keys associated with the *IoT Hub* account that they will connect to. Applications can retrieve the messages through *IoT Hub* using the same access keys.

After installing IoT clients on devices and setting up cloud gateway such as *Azure Event Hubs*, developers need to make sure that only authorized devices can be connected, the messages from the devices can be understood by the rest of the application, and the devices can be managed and controlled remotely through cloud gateway and other components. To this end, an IoT application needs to have device identity store, device provisioning service, and device and application models as shown in Figure 3.4.

While developers can configure small number of devices manually, for commercially produced IoT devices such as smart meters and smart thermostats, bulk processing of the device

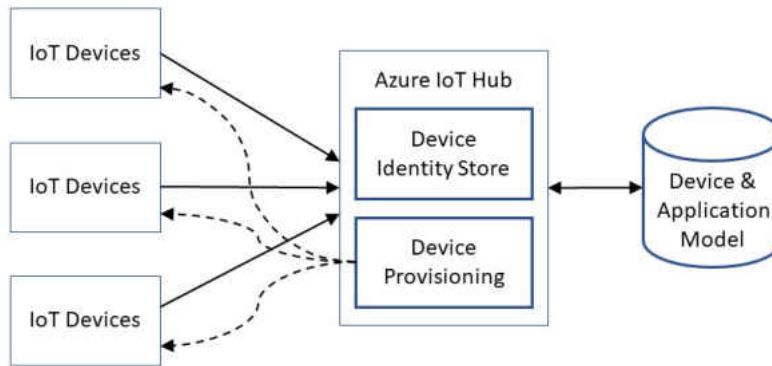


Figure 3.4: Relation between IoT devices and device identity store, device provisioning service, and device and application model.

configuration and setup may be necessary. *Azure IoT Hub* provides *Device Provisioning Service* (DPS), which is a global service to support device registration and configuration [76]. DPS provides API for device configuration and registration using *Azure IoT Hub's device identity store* to provide per-device security credentials for authentication and authorization. DPS can also automate the registration of the devices to store the device model and application model.

The device model defines the schemas of device metadata, output, control parameters, and control actions. The application model contains the semantic relation between devices such as device topology and the relation between device and other application components. The device and application models are core design abstractions of each IoT application. For example, devices such as smart thermostats are modeled in relation to the room and building where they are located. A building management system may define a static hierarchical structure to model the devices based on their physical locations [130]. For applications such as fleet management, the device topology is more dynamic since the IoT devices may change their grouping as the vehicles carrying the devices are assigned to different tasks. Regardless of the specific designs, these models should be made persistent in storage such

as SQL database or Azure *Cosmos DB*, which includes a graph API for convenient query of device topology.

3.2 Storage

IoT applications generate increasingly large amounts of data that must be processed in real-time or by offline batch applications. Storing IoT data in an on-premise database is not suitable for industrial IoT applications due to the lack of scalability of local storage. Cloud-based storage offers elastic capability that can scale up without significantly degrading performance [111]. Cloud-based storage can be divided into warm and cold storage with some tradeoff between latency, query capabilities, and cost. Warm storage such as *Cosmos DB* and *Azure SQL Database* offer lower latency and flexible query interface but at a higher cost. Cold storage such as *Blob storage* and *Data Lake Storage* have higher latency and simpler interface but lower cost. Figure 3.5 shows the possible dataflow path between cloud gateway, data processing computation, and storage.

The distribution of data between warm and cold storage should be based on the requirement of IoT applications. To reduce cost, data that is not frequently used should be stored

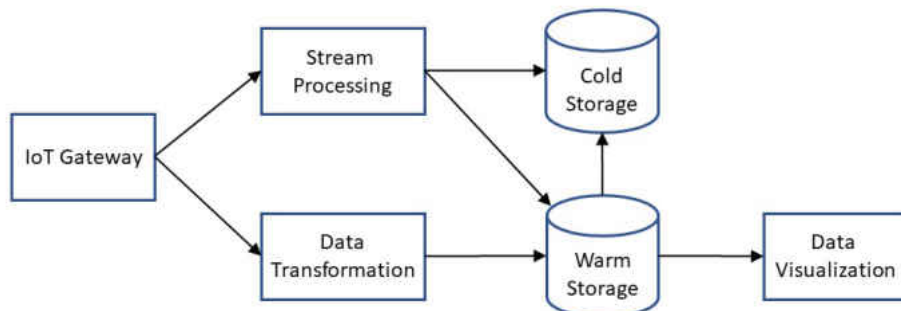


Figure 3.5: Dataflow between IoT gateway, data transformation, stream processing, cold storage, warm storage, and data visualization modules.

in cold storage. Data used in real-time processing components such as *Stream Analytics* and data visualization should be kept in warm storage [71]. A straightforward way of managing data for different storage is to store recently-transmitted data in warm storage first and then gradually migrate old data into cold storage after a predetermined amount of time has passed.

There are a few options for warm and cold storage on Azure IoT platform as shown in Figure 3.6. For warm storage, low latency is required. Azure provides *Cosmos DB*, which is a NoSQL database that supports five query methods including traditional SQL, MongoDB, graph, table, and Cassandra Query Language (CQL). Data in *Cosmos DB* can be set with an expiration date, after which the data is automatically deleted. Another option is *Azure SQL Database* [79], which provides relational database model with transactional support, which may be necessary for applications with strong integrity constraints. Both databases have flexible query interfaces but they also have higher cost than that of cold storage such as *Blob storage* and *Azure Data Lake Storage* [74], which stores data as files or objects and uses

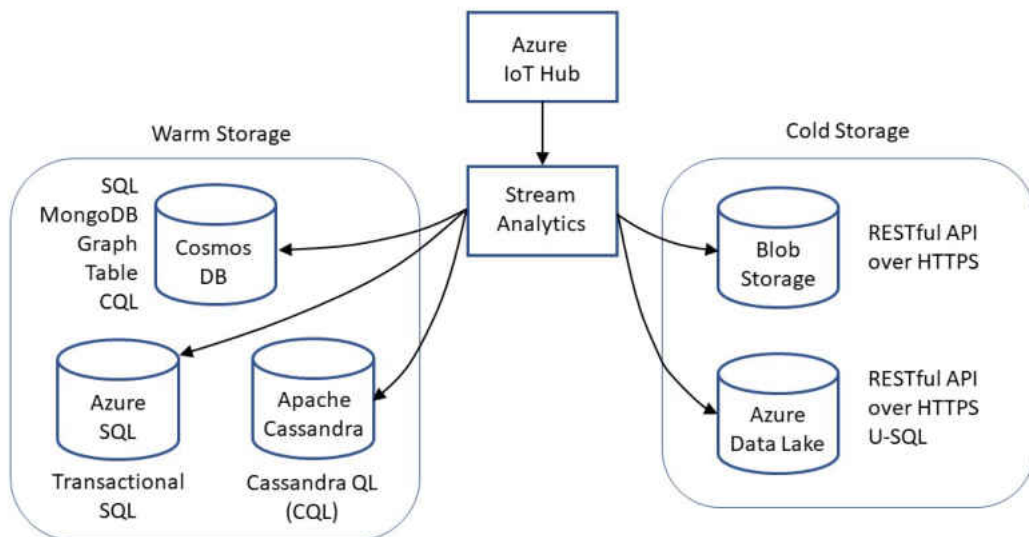


Figure 3.6: Options of warm and cold storage with different query methods.

RESTful API for query access. Query interface to cold storage is much more restrictive than that of warm storage. Thus, for computation such as visualizing power usage of a freezer over the past 30 days, warm storage should be used since the computation needs to issue queries that can return precise results with low latency. For computation such as analyzing the average power consumption of all freezers over the past year, cold storage can be used since the analysis can be computed offline with aggregate data set as input.

3.3 Data Analysis

The main purpose of IoT applications is to learn insight from data emitted from IoT devices and use the insight in making logical decisions towards business objectives [126] such as reducing operating cost of fleet vehicles or preventative maintenance of industrial equipment. Data analysis is the crucial step in realizing this goal. Depending on applications, IoT data analysis can be online processing of data streams or off-line processing of data batches. For online processing, the computation can also be divided as stateless or stateful. Figure 3.7 shows the possible dataflow paths for the different types of IoT data processing on Azure platform.

For online processing of IoT data, latency is an important consideration [125]. Latency in this context refers to the amount of time it takes to analyze the IoT data collected over a time window (such as 5 seconds) to generate the input for a backend application such as a data visualization tool. For example, latency in the scale of seconds is not suitable for applications such as connected vehicles, which may have real-time constraints in milliseconds. Throughput is another consideration for online processing. Throughput refers to the amount of data that can be processed within a time period. Applications with high data ingestion rates require sufficient throughput to handle incoming telemetry data to avoid data

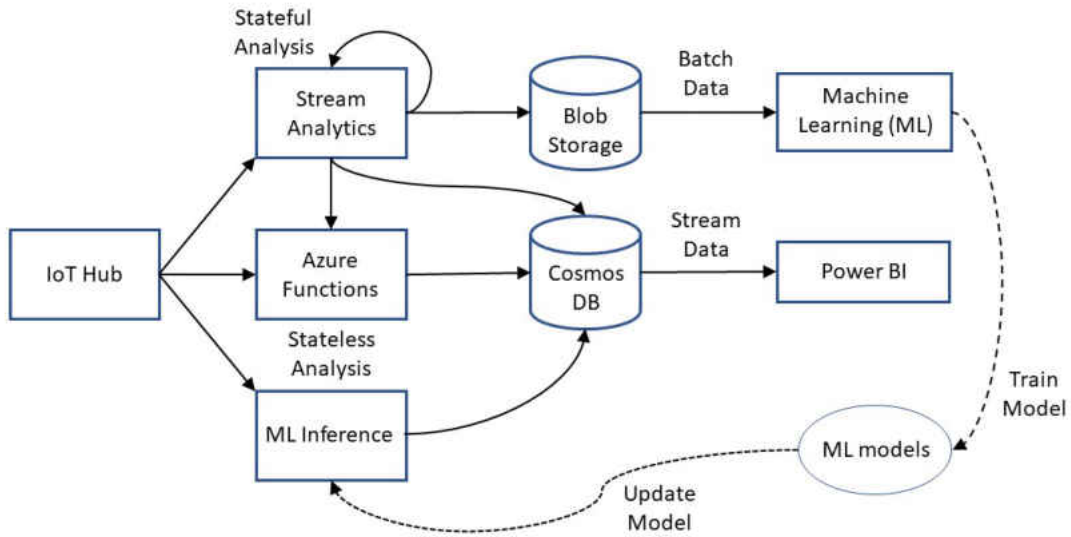


Figure 3.7: Possible dataflow paths for IoT analytics.

loss or excessive latency. Insufficient throughput can increase latency since unprocessed data has to be buffered and if the buffer overflows, unhandled telemetry data is dropped.

An application can reduce the requirement of throughput by slowing down data ingestion rates using methods such as sampling or aggregation. Cloud-based IoT platforms such as Azure can usually satisfy high throughput requirements of IoT applications since users can allocate more resources to process data in parallel. However, latencies of these platforms are more difficult to reduce. For applications that demand near real-time response time, on-premise data processing is more appropriate. For example, field gateways can be used to perform preliminary data integration, data analysis, and aggregation before sending the lower frequency data to the cloud gateway [29] such as Azure *IoT Hub*. This also can reduce the computation cost associated with cloud platform usage and the network bandwidth.

Online data processing includes data transformation, data filtering, event detection, alert generation, data aggregation, and other streaming data analysis such as Key Performance Indicator (KPI) calculation, classification, and anomaly detection. Online data processing may be stateless or stateful computation. Stateless computation such as data transformation only needs its input data at time t to obtain its output at time t . Stateful computation needs to use both its input at time t and its states at previous time stamps such as $t - 1$ to compute its output and states at time t .

Azure includes two types of data processing systems: *Azure Functions* and *Azure Stream Analytics*. Both systems can take streaming inputs from *IoT/Event Hubs* and output results to Azure storage and reporting tools such as *Power BI*. *Stream Analytics* implements query computation on data streams while *Azure Functions* provides general-purpose computation triggered by events. A common design pattern is to use *Stream Analytics* to preprocess data streams from *IoT Hubs* and then forward the results to *Azure Functions* for advanced processing.

Azure Stream Analytics provides a domain-specific way for processing IoT data streams using a SQL-like (stream analytics) query language [46] that treats data streams as tables, stream events as records, and stream event fields as record fields. For example, `select V, I from Power where V > 10` returns a stream of values that are the fields (V, I) from the stream `Power` where value of V is greater than 10. Nested queries can be made using *with* clause to define result sets of inner queries to be used in the outer query. In this query language, streams can be joined with the addition of an *on* operator that specifies the time bound of the joined stream events. The time bound is calculated with the *datediff* function using the timestamps of the stream events. Selected stream events can be *grouped by* field names with the addition of a *window* function or system timestamp. A

window function specifies the grouping of stream events based on the event timestamps so that aggregate function can be applied to the events in each time window to output a single event. Azure *window* functions include *tumbling window* (fixed-sized and non-overlapping), *hopping window* (with fixed overlap), *sliding window* (all distinct windows with fixed size), and *session window* (events grouped by similar timestamps) [75]. This query language also provides a list of builtin functions to support computation on scalars, aggregates, records, spatial values, and temporal values (e.g. lagged events, last event, and first event in an interval).

Azure *Stream Analytics* can invoke user-defined function (UDF) [56] written in JavaScript in its queries. UDF allows users to perform operations such as math, array, regexp, and remote data access, which is not supported by builtin functions of the query language. *Stream Analytics* can also utilize user-defined aggregate (UDA) [49] that specifies how values are (de-)accumulated over event sequences grouped by window functions in a query. UDA is defined as a function that constructs a JavaScript object with fixed-named methods. *Stream Analytics* can also use *Azure Functions* to perform additional processing by sending query results via HTTP request to an *Azure Function* in batches [70]. The batch size is limited to 256KB while each batch is limited to 100 events by default. Note that Azure *Stream Analytics* can perform both stateless and stateful data processing. Through *window* and *analytic* functions, *Stream Analytics* can perform stateful computation on event streams subject to time-based constraints.

Azure *Stream Analytics* jobs can be launched through Azure's Web interface by filling in the query statements in a textbox and specifying the input and output sinks of the analytics jobs. Launching a *Stream Analytics* job takes a few minutes while error messages are few and uninformative, which makes it difficult to perform testing since mistakes in query statement such as typo of field names may simply yield empty results. Azure *Stream Analytics* does,

however, allow faster tests by executing the queries on sampled datasets though this kind of testing does not reveal all possible faults with the queries. Also, the UDF and UDA associated with each *Stream Analytics* query are in separate scopes, which prevents sharing across multiple queries.

Azure Functions provide *serverless* computation triggered by events. *Azure Functions* can be implemented in languages such as Java, C#, and JavaScript to provide stateless computation in response to events from sources such as HTTP, *Blob storage*, *Event Hubs*, *Cosmos DB*, *Storage Queue*, and *Service Bus* [47]. *Azure Functions* allows the use of general-purpose programming languages to define complex logic to process event data and redirect the result to a destination such as *Blob storage*.

For stateful data processing, users can use **durable** *Azure Functions* [57], which is an extension of *Azure Functions* that implicitly uses Azure storage to manage stateful computation. As shown in Figure 3.8, durable *Azure Functions* includes *orchestration* function that can compose multiple *activity* functions to run in sequence or in parallel. Durable *Azure Functions* also includes *orchestration* client that can start a long-running process that can be monitored for progress via HTTP endpoints.

As an evaluation of the throughput and latency of high-volume event processing with *Azure Functions*, Microsoft conducted a performance test [50] to demonstrate some of the optimization strategies. This test partitioned a stream of weather and seismic event data into two streams to be processed independently. The stream processing used Azure *Event Hubs* to act as buffers, *Azure Functions* to separate the two types of event data and push them into their respective *Event Hubs* for further processing. In this test, a stream of 100,000 events per second was processed for 9 days. During the test, half of the events were processed with about 1 second of latency while 90% of the events were processed with less than 3

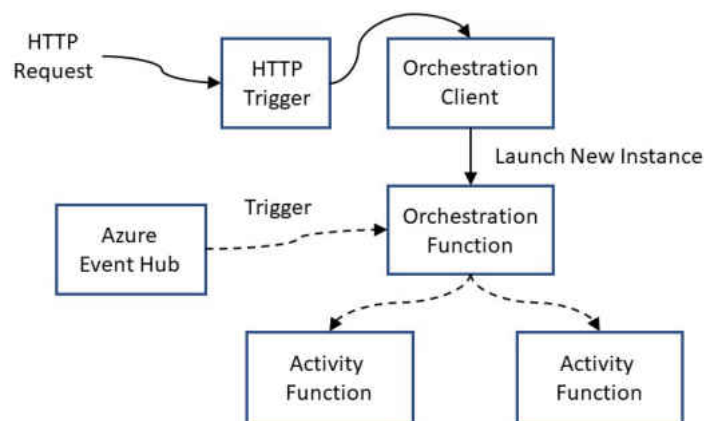


Figure 3.8: Stateful data analysis using durable *Azure Functions*.

seconds of latency. However, max latency was about 4 minutes. The test achieved this level of throughput by configuring each *Event Hub* with 100 partitions, which translates to 100 virtual machines that execute the *Azure Functions* in parallel. This test also processed the streaming events in batches instead of individual events in *Azure Functions*, which increases throughput by reducing the overhead of launching *Azure Functions*.

Offline data processing includes KPI calculation using historical data, clustering, regression, and classification using unsupervised or supervised machine learning. Many of these offline data analysis tasks can be completed in *Azure HDInsight* [72], which is a PaaS cluster platform for launching distributed computing services. Apache Spark, which is a distributed in-memory programming library, is available in *HDInsight* as a computation environment for big data analysis [69]. With Spark, users can develop parallel applications that can perform data analysis using data from storage facilities such as *Azure Data Lake Storage*, *Blob storage* and gateways such as *Event Hubs* (through Spark streaming). Spark utilizes a map-reduce programming model to support distributed computation and its memory-based resilient distributed dataset (RDD) abstraction provides better performance on big data than disk-based

distributed file system such as Hadoop does [122]. Spark supports offline analysis such as machine learning. For example, using Spark, an IoT application can train a regression model offline using historical data and use the model to make online predictions for streaming data as shown in Figure 3.7.

Note that Spark uses a cluster of machines to complete its tasks. When running Spark in *HDInsight*, a cluster of VMs of a minimum configuration such as a master and a slave must be created and the cluster will exist until the task completes. Keeping a cluster running is far more expensive than services such as *Azure Functions*. Thus, it is not cost effective to keep a long-running data analysis using Spark in *HDInsight*, which is more suitable for offline data analysis that is both computation intensive and has a finite duration.

3.4 User interface

Some IoT applications such as smart thermostats can use the results of data analysis to generate automatic control signals such as setting the thermostat temperature. However, for most applications, graphic user interface is needed to render data visualization and generate reports for business decision making. The interface can also be used to monitor the status of IoT devices and domain abstractions such as buildings that contain smart meters, to schedule an offline data analysis such as machine learning on historical data of certain time period, and to change an online data analysis such as the formula of a KPI. Azure IoT platform provides several PaaS and SaaS components for this purpose.

Visualization *Azure Power BI* is a versatile dashboard tool for data visualization and report generation of static or low frequency data. Using the graphic interface of *Power BI*, non-programmers can create dashboard with various data charts from a range of data sources. *Power BI* includes two versions: *Power BI Desktop* and *Power BI Service* (online).

Power BI Desktop has more functionalities in terms of visualization and report generation. However, only *Power BI Service* supports real-time streaming where its data ingest rate is limited to 5 requests per second for its streaming dataset and 1 request per second for push dataset [59]. Higher frequency data may be streamed or pushed to *Power BI* in batches with some added latency. For visualization and analysis on IoT data as time series, Azure *Time Series Insights* [51] may be used. *Time Series Insights* is also a type of warm storage that stores data in memory and solid state drives, though its data is only kept for a limited amount of time and its query latency is still too high for on-demand applications.

Web and mobile apps Azure *App Service* provides environment for building Web app, mobile backend, and RESTful APIs. Users can create a Web interface using Azure *App Service* to connect to IoT components such as Azure *IoT Hub* through WebSocket. For mobile devices with unreliable network and limited energy budget, push notification may be used to send IoT data such as alerts and event notification. To this end, Azure provides *Notification Hubs* [73], which allows applications to push messages to mobile devices of platforms such as iOS and Android. Azure *Notification Hubs* can also be integrated with Azure *App Service Mobile Apps*, which works by retrieving device PNS (platform notification system) handles, registering the devices with notification hubs, and sending notifications from app backend through notification hubs.

4 Experiences with Azure

The Azure Cloud offers a wide range of services, where each service belongs to a broader category to help with organization. As of this writing, there are 22 categories, where the number of services each category has ranges from as low as 1 to as high as 50. Many services are unique to each category, but some services do overlap and are placed in other categories as well. Information and examples of each service can be found at Microsoft's main Azure website. See Figure 4.1 for a view of the *Azure portal* and the many services offered. The wide range of services, categories, and amount of information online creates a steep learning curve to those unfamiliar with the Azure platform. To study the platform and ease of use, a set of tasks were selected to be performed on a data set. To illustrate the implementation of an IoT application using Azure platform, an experiment was conducted using Azure to

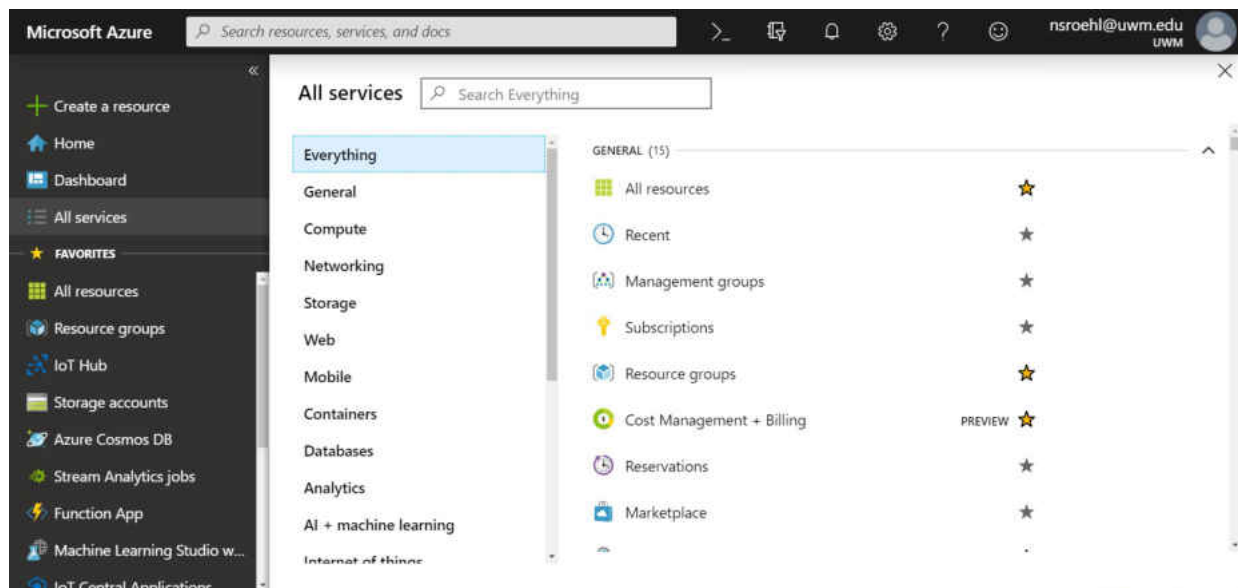


Figure 4.1: View of the *Azure portal*. The left most menu contains favorite services. When clicking on any of these tabs, a window opens to the right. These windows are called blades. The *All services* tab was selected, showing only a small portion of all available services.

analyze a set of time-series data of Li-ion batteries. The remainder of this chapter details the data set, the process, and our experiences using Azure. The experiences are based on the current iteration of Azure, which is always changing. Some features or services may not be available at the time this is read.

4.1 Data Set

The data set used details the operational use of a number of Li-ion batteries collected over an extended period of time by NASA [121]. This experiment used the data of battery number 5. The data was collected over many cycles, each of which includes a full charging phase followed by a full discharging phase while metrics such as current, voltage, temperature, sampling time, and the impedance were collected. As the number of cycles increased, the time to fully discharge the battery decreased. The data was originally in Matlab data files and was converted to CSV (Comma Separated Value) files before it is used for the simulation of IoT telemetries in the experiments.

4.2 Azure Set Up

The main goal of this experiment is to create a IoT system using Azure. The experiment simulated telemetry data from sensors connected to the battery and sent it to an Azure *IoT Hub*. The data is stored, processed, then analyzed during each cycle. The analysis will then reveal when the battery should be replaced as its capacity degrades over time. The list of different services used during the set up are *IoT Hub*, *Storage Accounts*, *Stream Analytics Jobs*, *Function App*, *Machine Learning Studio*, *Resource Groups*, and *Cost Management + Billing*.

To use any Azure service, a user must have an account along with a subscription. The

subscription keeps track of what services were used and how much to charge. A nice feature is that many of the services in Azure only charge a user when they are running. If a user is looking to cut costs, turning off a service during low-peak hours is a viable option. The user still must be careful as some services may incur charges just for creating a resource.

4.3 Resource Groups

Before going in to detail about many of the services and what was done, the importance of resources and Resource Groups must be stressed. When using any of these services, a resource from that service must first be created. For example, when using the *IoT Hub*, a user must create an *IoT Hub* resource first. Since multiple resources can be created from the same service, a unique name is given to each one to distinguish them. Once a resource is created, it must be assigned to a resource group. Resource group is a convenient way to store multiple resources together in one place. When creating some resources Azure may create other resources unaware to the user. The resources created automatically will appear in the same resource group, keeping related resources together. This makes it easier to locate, delete, and manage multiple resources. Creating a resource can be done via command line inputs using the *cloud shell* or through the *Azure portal*. Both work through a browser and allow a user to create and monitor resources, but the *cloud shell* requires the user to know all the commands where the *Azure portal* is an interactive graphic interface. *Azure portal* is used for the remainder of the chapter.

4.4 IoT Hub and Storage

After creating a resource group, an *IoT Hub* resource is needed as a location to receive messages [62]. An *IoT Hub* can also send messages, but this functionality is not required

for this experiment. When creating an *IoT Hub* resource, a tier must be selected to specify how many messages this hub is expected to receive. Tier selection helps Microsoft allocate physical hardware resources on their end to meet the users needs. The free tier was used in this experiment, as it provides sufficient throughput and is cost effective. See Figure 4.2 for the blade used to create an *IoT Hub* and Figure 4.3 for more details on *IoT Hub* tiers. IoT devices must be added to the *IoT Hub* after an *IoT Hub* resource is created. Each device requires its own device ID in the *IoT Hub* to distinguish one device from another. To add a device, select the *IoT Devices* tab under the newly created *IoT Hub* resource. Here a user can add devices where Azure will auto generate connection strings for each device. The connection string is critical in maintaining a secure connection between the device and the *IoT Hub*.

To send messages from a device, Microsoft provides many SDKs and example code to connect various types of IoT devices to the *IoT Hub* [100]. Various programming languages and

IoT hub
Microsoft

Create an IoT Hub to help you connect, monitor, and manage billions of your IoT assets. [Learn More](#)

PROJECT DETAILS

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

* Subscription ⓘ Pay-As-You-Go ()

* Resource Group ⓘ Battery5Testing
[Create new](#)

* Region ⓘ Central US

* IoT Hub Name ⓘ battery5IoTHub ✓

[Review + create](#) [Next: Size and scale »](#) [Automation options](#)

Figure 4.2: Creating an *IoT Hub*.

Tier Name	Messages per minute per unit	Message size per minute per unit	Number of units allowed	Cloud-to-device Messaging	IoT Edge Enabled	Device Management	Cost (USD) per month per unit
B1: Basic Tier	278	1.111 MB	200	N	N	N	\$10.00
B2: Basic Tier	4,167	16 MB	200	N	N	N	\$50.00
B3: Basic Tier	208,333	814 MB	10	N	N	N	\$500.00
S1: Standard Tier	278	1.111 MB	200	Y	Y	Y	\$25.00
S2: Standard Tier	4,167	16 MB	200	Y	Y	Y	\$250.00
S3: Standard Tier	208,333	814 MB	10	Y	Y	Y	\$2,500.00
F1: Free Tier	8,000 (per day)	Unknown	1	Y	Y	Y	\$0

Figure 4.3: Different tiers for *IoT Hubs*. IoT Edge enabled means this tier can send/receive messages from IoT Edge devices. Device Management means this tier can use device twin, query, direct method, and jobs [64].

platforms such as Java, C, .NET, Node.js, Python, and iOS frameworks are all supported. This experiment used Java, which needed Microsofts SDKs and Maven. It is recommended to use Maven as a build tool for Java-based IoT client, as Microsoft provides a dependency snippet that can be copied and pasted into the Maven build file (commonly named as pom.xml) which will then import all the appropriate library dependencies (namely jar files) to connect to the *IoT Hub*. Modifying one of the example codes provided by Microsoft was an easy way to start a project and send messages. The default connection string given in the example must be updated to match the devices ID in the *IoT Hub*, if it is not the messages will not send. To simulate telemetry messages, the battery data was first read from the CSV file, converted to strings in JSON format, and sent as MQTT messages to the *IoT Hub* using the libraries provided. A quick way to view whether the messages are being sent or not is to examine the real-time graphs that are shown on the overview page of the *IoT Hub* (Figure 4.4). There will be some latency, but if the messages are sent correctly the graphs will eventually update.

The type of storage account determines what services can be used to store data. StorageV2 allows storage in blobs, files, queues, and tables. StorageV1 is similar to StorageV2 except is used for legacy accounts. BlockBlobStorage accounts are blob only storage tailored

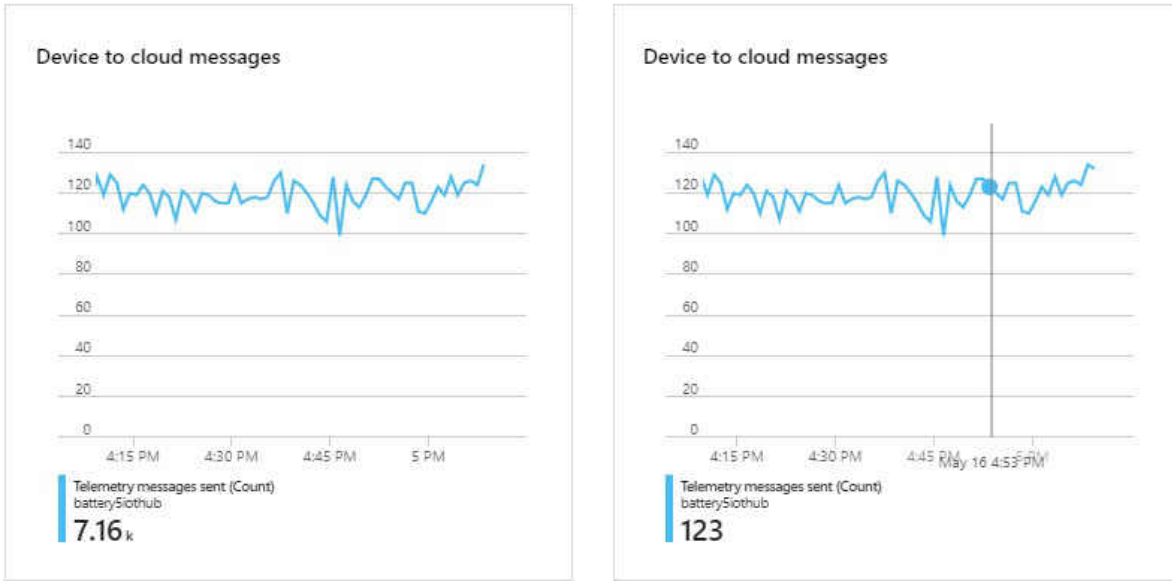


Figure 4.4: Real-time graphs showing incoming messages to the *IoT Hub*.

to blob performance. FileStorage accounts are file only storage tailored to file performance. BlobStorage accounts are blob only storage but for legacy accounts [98]. The replication option determines where duplicated data is stored geographically. Azure creates multiple copies of all data, the further away from the source the more expensive storage gets, and replicated data is updated synchronously.

First storage option is locally redundant storage (LRS). LRS maintains three replicas in the same storage scale unit, which is a collection of storage racks. LRS is the most cost-effective way but it provides the least durability in case of hardware failure or natural disasters. Zone-redundant storage (ZRS) maintains three replicas in the same data center but in different zones. If only the storage scale unit that contains source data malfunctions, replicated data will still be accessible with ZRS. This is the not case with LRS as the source and replicas are in the same storage scale unit. The last two options are geo-redundant storage (GRS) and read-access geo-redundant storage (RA-GRS). Both store three replicas in the same data center and three replicas at a different data center in a region hundreds

of miles away. RA-GRS is the only option that allows read only access to replicated data. Latency could be a problem with RA-GRS, as replicas at different data centers are updated asynchronously unlike updates done in the same data center. Lastly, cold or hot storage can be selected. Hot storage should be used for frequent reads and updates. Use cold storage otherwise.

After creating a storage account with the appropriate settings, see Figure 4.5 for example of this in *Azure portal*, messages sent to the *IoT Hub* can now be stored. If no other services are needed and storage is the only concern, the *IoT Hub's* Message Routing options can be updated to output data into four locations. They are *Event Hubs*, *Service Bus queue*, *Service Bus topic*, and *Blob storage* [99]. Having a message reroute to an *Event Hub* seems redundant, since the *IoT Hub* is essentially an *Event Hub* with bidirectional communication, but if a user wanted to combine information from an *IoT Hub* and an existing *Event hub* that collects

The screenshot displays the 'Create storage account' configuration page in the Azure portal. At the top, the 'Subscription' is set to 'Pay-As-You-Go' and the 'Resource group' is 'Battery5 Testing'. Below this, the 'INSTANCE DETAILS' section explains the deployment model. The main configuration fields are: 'Storage account name' (battery5storageaccount), 'Location' ((US) Central US), 'Performance' (Standard selected), 'Account kind' (StorageV2), 'Replication' (Locally-redundant storage), and 'Access tier' (Hot selected). At the bottom, there are three buttons: 'Review + create' (highlighted in blue), 'Previous', and 'Next : Advanced >'.

Figure 4.5: Creating a storage account.

data from multiple non IoT sources for data analysis this could be a useful option. A service bus is a message broker, where the *IoT Hub* would be the publisher and any number of Azure services could be a subscriber [65]. A *Service Bus queue* stores messages in a single queue where a subscriber service reads from the queue when ready. Typically, *Service Bus queues* are meant for point to point communication. A *Service Bus topic* is similar to a *Service Bus queue* except that multiple queues are used. This allows multiple subscriber services to read from one source and is more similar to the typical publisher/subscriber model. If the data isn't needed immediately, *Blob storage* can be used for long term storage or for analysis at a later time.

When using the *Azure portal*, certain storage options don't allow a user to view the contents of stored messages. It is recommended to download and install Microsoft Azure Storage Explorer. This program is free and links up with an existing storage account after providing the correct credentials. After installation a user can now access all stored messages in their storage account from their desktop with out having to go through the *Azure portal*.

4.5 Stream Analytics

If more destinations for storage or processing are needed, *Stream Analytics* is another service for routing messages sent to the *IoT Hub* [101]. The basic set up of a *Stream Analytics* job are the inputs, query, and outputs (Figure 4.6). Inputs are differentiated between data streams and reference data [104]. Data streams are unbounded streams of data, like *Event Hubs*, *Blob storage*, or an *IoT Hub*. Reference data is typically static data that is used for correlation and look ups. Currently only *SQL Databases* and *Blob Storage* are offered as reference data inputs. The query language used by *Stream Analytics* is similar to SQL, and considered a subset of Transact-SQL (also known as T-SQL) [46]. Queries can determine

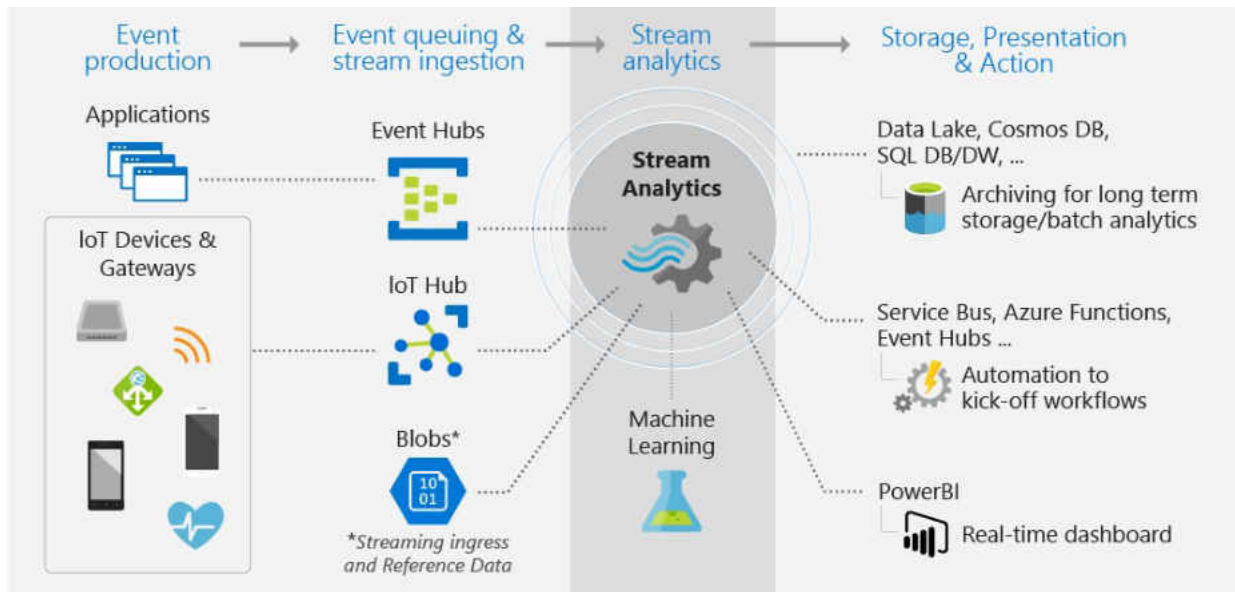


Figure 4.6: Overview of inputs and outputs for *Stream Analytics*. [101]

or limit what gets stored, and can be combined with windowing to perform analytics over designated periods of time. Output is where the data gets stored after being processed by the user defined query [105]. Options for output are *Event Hub*, *SQL Database*, *Blob Storage*, *Table Storage*, *Service Bus topic*, *Service Bus queue*, *Cosmos DB*, *Power BI*, and *Data Lake Storage Gen1*. *Azure Functions* is also an option for output, but these details will be discussed later. Unlike the *IoT Hub*, *Stream Analytics* jobs must be started to be used. A start button is visible after creating a *Stream Analytics* resource, click this to start a job after all inputs, queries, and outputs are added. It may take a few moments for the job to start, but it will inform the user when it has started. A user is only charged when a job is running, so remember to stop the job when it is not in use.

When adding inputs or outputs, an alias is used in a *Stream Analytics* query to refer to an actual resource. Multiple inputs and outputs can be used for a single *Stream Analytics* job, reducing the amount of services required. For the experiment, a *Stream Analytics* job was created that used a *IoT Hub* resource as input and a *Table Storage* resource as output.

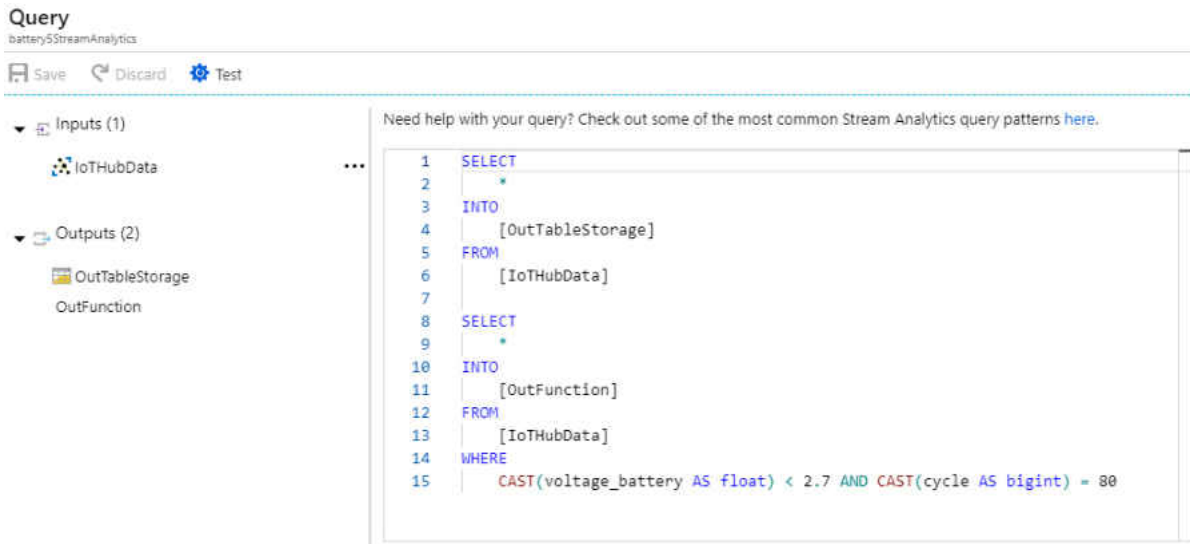


Figure 4.7: *Stream Analytics* query with multiple parts. First selection takes all data sent to an *IoT Hub* aliased as `IoTHubData` and outputs to *Table Storage* aliased as `OutTableStorage`. The second selection reads data from the same *IoT Hub*, except when the parameters of the `where` statement are true it executes an *Azure Function*.

See Figure 4.7 for the example. Several other storage options could have been selected, but *Table Storage* provides all the functionality required. Tables are a NoSQL data store that can store lots of data, can scale with the data, can be queried easily and quickly, and they are viable input and output sources for *Azure Functions* [102, 103]. When adding a table as output for a *Stream Analytics* job, it will ask for a Partition and Row key as field values which form the primary key for each row entry in the table. When combined they must be unique, if not it will over write the existing data with the matching primary key. Both keys must reference a named datatype in the streaming data. If not *Stream Analytics* will display a yellow warning symbol next to the output saying it is degraded. During testing no messages were updated when a *Stream Analytics* job became degraded.

Azure does offer query testing and some syntax error detection which helps save time and potential head aches as starting up a *Stream Analytics* job can be cumbersome [68]. Files to upload must be in JSON, CSV, or Avro format. Running the test after uploading, if done

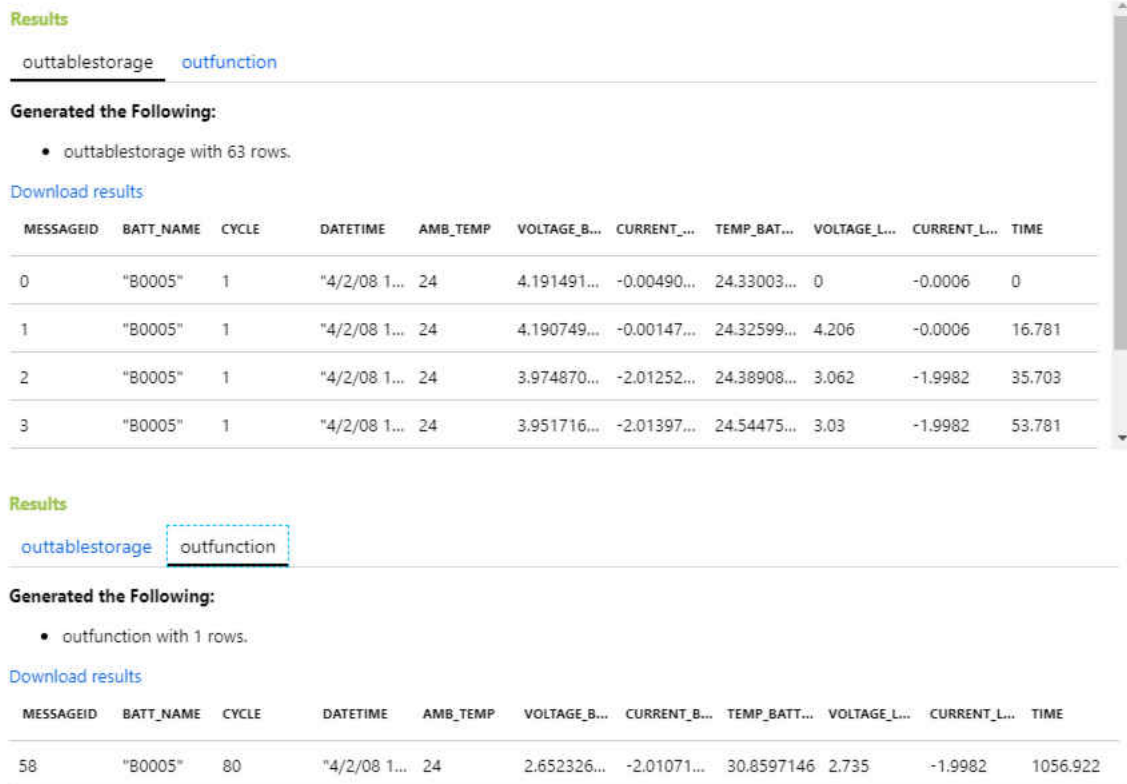


Figure 4.8: Results of testing query in Figure 4.7.

correctly, will give results to what is expected to happen during execution (Figure 4.8). The testing that is offered is helpful, but it is not fool proof. Queries can still be syntactically correct but produce no output. Azure does not provide any debugging tools to understand why something is not working. It is also unclear whether testing works with windowing, as windowing reads messages based on the time they were received and whether they reside inside a designated time slot.

4.6 Azure Functions

Once the data is stored, an *Azure Function* can be triggered to read and perform some computations on the data. The main parts of an *Azure Function* are the trigger, bindings,

and code implementation. The trigger causes a function to execute, and a function can only have one trigger [102]. Previously it was mentioned that *Azure Functions* can be used as output for a *Stream Analytics* job. The output is not a data stream, but an `HTTPTrigger` that is sent when a certain condition is met to execute the function. Bindings can be distinguished by input vs output. They are not required, but they provide an easier way to connect to other data sources for input or output in the function. See Figure 4.9 for a list of some of the triggers and bindings offered.

Azure Functions can be written in C#, JavaScript, Java, and Python [67]. All languages listed can be written in an integrated development environment (IDE) on a desktop and then uploaded to Azure. An IDE is provided in *Azure portal* through the browser, but it only supports C# scripts and JavaScript. Each option has its advantages and disadvantages. Using a desktop IDE gives the developer the choice to use a setup and programming language they are more familiar with. A desktop IDE also offers debugging and syntax error detection. Developing in the portal offers only two programming languages, and no debugging or syntax error detection. Mistakes are only found when the code is executed. A downside to using a desktop IDE is having to set up the programming environment to be able to connect to Azure. This means having to go through Microsoft’s online documentation which can be

Type	Trigger	Input	Output
Blob Storage	X	X	X
CosmosDB	X	X	X
Table Storage		X	X
Queue Storage	X		X
HTTP & Webhooks	X		X
Service Bus	X		X
Event Hubs	X		X
Timer	X		

Figure 4.9: *Azure Functions* triggers, inputs, and output options [102]. Other services are available but were not included as they do not relate to IoT.

difficult to follow. Desktop IDE's also require a user to upload their functions to Azure, where anytime a function is updated it must be re-uploaded. The portal avoids both of these issues.

To create a function in the portal, use the *Function App* service to create a *Function App* resource. The resource creation stage is where the programming language is selected. All functions that reside in this resource will use that language. For the experiment, the .NET framework was chosen to create a C# script because it did not require much code, many of the better examples online were in C#, and ease of environment setup. The rest of the setup described is based on the user selecting the .NET framework in the portal to develop a function. The other programming language options were not tested and will not be discussed in great detail. Once the resource is created and opened in a new blade, a function can be made by clicking the plus symbol. After selecting the trigger to be used and providing a name, Azure will generate a C# script template, and a `function.json` file. For script files, a `function.json` file is required to list the trigger, and all input and output bindings [53, 102]. See Figure 4.10 for an example of the `function.json` file.

The process of updating the json file and connecting to sources outside the function are the most difficult aspects when creating an *Azure Function*. The idea was very simple for the experiment: have a function read data from a table, perform some calculations on the data, and then output the new information to another table. One would think having the storage and processing all done in Azure would be a simple process, but sadly it is not. Microsoft provides a lot of documentation and examples online for how to add input and output bindings, but the amount of information and lack of explanations for each example given make it more difficult than it needs to be.

To update the bindings, select the *integrate* option underneath the selected function. A new blade will open up which allows a user to update the trigger and add any bindings.

function.json

```
1 {
2   "bindings": [
3     {
4       "authLevel": "function",
5       "name": "req",
6       "type": "httpTrigger",
7       "direction": "in",
8       "methods": [
9         "get",
10        "post"
11      ]
12    },
13    {
14      "type": "table",
15      "name": "inputTable",
16      "tableName": "battery5TableStorage",
17      "take": 50,
18      "connection": "AzureWebJobsStorage",
19      "direction": "in"
20    },
21    {
22      "type": "table",
23      "name": "outputTable",
24      "tableName": "battery5CoulombTable",
25      "connection": "AzureWebJobsStorage",
26      "direction": "out"
27    }
28  ],
29  "disabled": true
30 }
```

Figure 4.10: A functions trigger and bindings json file.

When using table storage, it needs the connection string to the appropriate storage account, the name of the table where the data is stored, and a parameter name. There are other fields to fill out but they are optional and not required for the experiment. The connection string can be updated in the storage account connection field option. Clicking the *new* option next to the field reveals all storage accounts associated with the current subscription. Selecting the correct storage account will automatically import the connection string. When inputting the table, only the table name is required. Unlike *Blob storage*, table storage does not allow sub containers so all tables are stored in the same location under one storage account [103].

The parameter name is the name used in the function when referencing the actual table. For example, the experiment's table name was `battery5TableStorage` and the parameter name was `inputTable`. Any time a read operation was performed on the `battery5TableStorage` table in the function, the `inputTable` parameter name would be used. The documentation online on how to perform this basic setup is lacking. While intuitive in hindsight, it was initially very confusing. After updating the bindings through the portal, the existing json file should update automatically to reflect the changes made.

Once the bindings are updated, modifications can be made to the function code. The parameter names used in the json file must be used as actual parameter names for the function. The datatype for these parameters depends on which function runtime is being used. The two versions are 1.x, for legacy runtimes, and 2.x, for newer runtimes [66]. If using 1.x, the datatype used is called `IQueryable`. If using 2.x, the datatype used is called `CloudTable`. Developing functions in the portal automatically uses version 2.x, so each parameter representing a table should be a `CloudTable` datatype. Continuing the example from above, the full function parameter name would be `CloudTable inputTable`.

After the function parameters are declared, it would be nice if the process was over and a user could simply use the `CloudTable` parameters directly. But there is still more setup required. To get information from the table, it can be queried using `TableQuery` and `TableQuerySegment` datatypes [54]. A query is provided to a `TableQuery`, which in turn is passed to a `TableQuerySegment` to get information from the table. After the results are returned from the query, one final step remains. A class must be declared whose instance variables match the names of the columns in the table to be read. The user may select which columns to use as instance variables for the class, but the partition key and row key are included by default. Take note of the datatypes of each column, the datatype of each instance variable must match the datatype of the corresponding column. If unsure of the

data types, use the Microsoft Azure Storage Explorer to view specific details about each column. Once the class is created, data can finally be processed and analyzed by iterating over each row in the results sequentially. Each rows data is stored in an object of the classes instance variables allowing for easy access and modification of the data.

To store data to a table from a function, the output binding must be added to the json file similar to how input bindings were added. The chosen parameter name must be added to the function parameter list as well. For example, if an output table was declared in the bindings called `outputTable`, the full function parameter name would be `CloudTable outputTable`. No querying is necessary to write to the output table, so the `TableQuery` and `TableQuerySegment` datatypes are not needed. Instead, a `TableOperation` datatype is used [54]. An object must be created which stores any relevant data for a row to be written. This object is passed to the `TableOperation` which then writes the data to storage. Online documentation shows other ways to read and write from table storage, but the examples are limited and don't provide much detail if an error occurs. See Figures 4.11 and 4.12 for function used.

```
75 //Class to store input data from table storage.
76 public class ReadEntity : TableEntity
77 {
78     public string Batt_name { get; set;}
79     public string dateTime { get; set;}
80     public long MessageID { get; set;}
81     public long cycle { get; set;}
82     public double time { get; set;}
83     public double voltage_battery { get; set;}
84     public double voltage_load { get; set;}
85     public double current_battery { get; set;}
86     public double current_load{ get; set;}
87     public double temp_battery { get; set;}
88     //Variable not part of input table.
89     //Added as new compute variable to help prediction.
90     public double capacity { get; set; }
91 }
92
93 //Class to sort all inputs by messageID.
94 class TableSorter : IComparer<ReadEntity>
95 {
96     public int Compare(ReadEntity x, ReadEntity y)
97     {
98         return (int)(x.MessageID - y.MessageID);
99     }
100 }
```

Figure 4.11: Classes used to help *Azure Function* read and write to table storage. See Figure 4.12 for main method of *Azure Function*.

```

1 #r "Newtonsoft.Json"
2 #r "Microsoft.WindowsAzure.Storage"
3
4 using Microsoft.WindowsAzure.Storage.Table;
5 using Microsoft.WindowsAzure.Storage;
6 using System.Threading.Tasks;
7 using Microsoft.Extensions.Logging;
8 using Microsoft.AspNetCore.Mvc;
9 using Microsoft.Extensions.Primitives;
10 using Newtonsoft.Json;
11 using System.Net;
12
13 public static async void Run(HttpRequest req, CloudTable inputTable, CloudTable outputTable, ILogger log)
14 {
15     //TableQuery used to store query.
16     TableQuery<ReadEntity> query = new TableQuery<ReadEntity>().
17         Where(TableQuery.GenerateFilterCondition("Batt_name", QueryComparisons.Equal, "80005"));
18     TableContinuationToken token = null;
19     List<ReadEntity> allEntities = new List<ReadEntity>();
20
21     do
22     {
23         //TableQuerySegment uses TableQuery to search inputTable.
24         TableQuerySegment<ReadEntity> resultSegment =
25             await inputTable.ExecuteQuerySegmentedAsync(query, token);
26         //Token used to continue searching large tables.
27         //If not used, TableQuerySegment will only return 1000 rows.
28         token = resultSegment.ContinuationToken;
29
30         //InputTable may not be sorted.
31         //Must store in list to sort before performing computations.
32         foreach (ReadEntity entity in resultSegment.Results)
33         {
34             allEntities.Add(entity);
35         }
36     } while (token != null);
37
38     allEntities.Sort(new TableSorter());
39
40     double coulomb = 0;
41     bool hasWritten = false;
42     ReadEntity prev, cur;
43
44     for(int i = 1; i < allEntities.Count; i++)
45     {
46         prev = allEntities[i-1];
47         cur = allEntities[i];
48         if(prev.cycle == cur.cycle)
49         {
50             if(prev.voltage_battery < 2.7)
51             {
52                 //Add coulomb data to current object and write to table.
53                 prev.coulomb = coulomb/3600;
54                 TableOperation insertOperation = TableOperation.Insert(prev);
55                 await outputTable.ExecuteAsync(insertOperation);
56                 hasWritten = true;
57             }
58             else if(!hasWritten)
59             {
60                 //Compute coulomb.
61                 coulomb += (Convert.ToDouble(cur.time) - Convert.ToDouble(prev.time)) *
62                     Convert.ToDouble(cur.current_battery);
63             }
64         }
65         else
66         {
67             //New cycle has started. Reset Values.
68             coulomb = 0;
69             hasWritten = false;
70         }
71     }
72     log.LogInformation("Ending process.");
73 }

```

Figure 4.12: *Azure Function* used to read and write to table storage. See Figure 4.11 for classes required to complete function.

4.7 Azure Machine Learning

When the telemetry data has been processed and stored, a machine learning model can analyze the processed data to perform predictive analytics. The process of creating a predictive model is straight forward. First data must be collected and processed, where processing means transforming the data into a more readable format for a machine learning algorithm to analyze. Then an algorithm must be selected to act as the predictive backbone for the model. Data is then passed to the selected algorithm to train and create a model. Once the model is trained and fully tested, it can be deployed and used for real time prediction.

When selecting an algorithm there are three main categories, supervised, unsupervised, and reinforcement learning [8, 19]. Supervised learning uses clearly labeled historical data to train a model and provide predictions. The equation $f(x) = y$ is a simple way to explain how it works. The function f , the model, accepts the variable x , the historical input data often referred to as the features, as a parameter which computes y , the variable to predict. Typically, the input data comes in a table format, where each row represents a single entity and each column is an attribute for each entity. For example, a table could contain information about a list of houses. Each row is a different house, and the columns describe the number of the bathrooms, bedrooms, square footage, and so on for each house.

Supervised learning can be separated into classification and regression, with each category offering a different type of prediction. Classification algorithms focus on creating groups with similar traits, where prediction is used to classify new data as belonging to one of these groups. Classification is well suited for spam detection, where a model is trained using millions of different emails where each email is labeled as spam or not spam. When the model is trained, new emails that are not labeled can be passed to the model to predict the group it belongs to. Regression algorithms are used to predict future values, typically numerical. A common example for regression is predicting housing prices. Feature selection can include

many of the attributes listed previously and any other relevant information. The model is trained using these features along with the houses current value. Once trained, data from new houses can be input to the model to predict its future price. For the battery experiment, some form of regression would be the best choice for predicting what cycle the battery will need to be replaced during.

While unsupervised and reinforcement learning were not used, they will be discussed briefly here to complete the machine learning picture. Unsupervised learning is used for learning without knowing the desired outcome. It is more of an exploratory process where there is no specific output value to predict unlike supervised learning. Some uses are clustering, association, and anomaly detection. Clustering and association find hidden relations amongst the data. Clustering is like supervised classification where similar or dissimilar data is grouped together, but no predictions are made. It is just grouping the data together based on relations. Advertising firms can use clustering to group populations to better understand and target an area for a product. Association can be used by retailers to find correlation between buying certain products to help increase sales. Anomaly detection is used for finding unexpected behavior. Credit card companies can use this for fraud detection, where unexplained spending can be detected and the customer can be alerted. Reinforcement learning involves a model being able to learn in an environment where rewards and punishments are given based on its actions. The goal is to maximize the rewards while minimizing punishments. The company DeepMind Technologies is famous for creating AlphaZero, a program that uses reinforcement learning to play the games of chess, shogi, and go at superhuman levels [124].

When selecting features to train the model, it is common practice to visualize the data before training a model. Visualizing the battery experiment data, where the x-coordinates

are one feature and the y-coordinates are another, can reveal several relations between different features helping with feature selection. See Figure 4.13 for an example, which shows the battery data imported and visualized with different plots in *Power BI Desktop*. The cycle of the battery will be a feature, so comparing other features to it may reveal significant trends. **Capacity** and **Time**, where **Time** is the length of each cycle, have the strongest relation to a cycle, so they should be included in the feature selection. **Temperature** also has some relation so it may also prove useful. The rest are more erratic with no real trends, which means they may not be the best choices to include.

The visuals can also be used to help with algorithm selection. Many of the less erratic visuals have a sloping trend, so linear or polynomial regression could be good choices. Linear regression tries to find a straight line that closely represents the general trend of the data, and polynomial regression creates a similar line except the line can curve to follow the data, like a polynomial function. Trial and error will always be a part of the process, where many

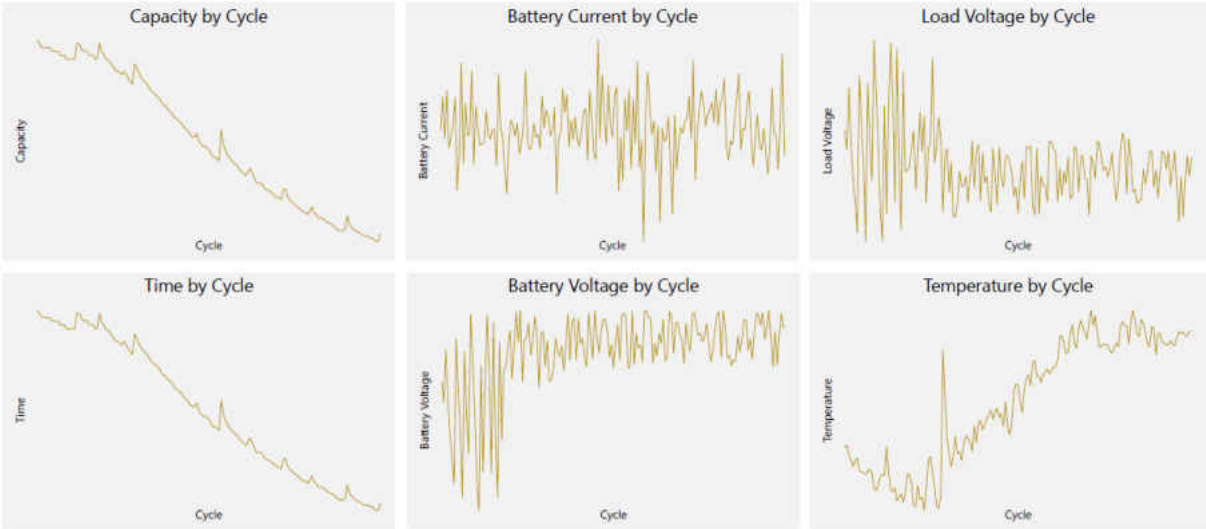


Figure 4.13: Plotting relations between each feature and **Cycle** in *Power BI Desktop*. **Battery Current**, **Battery Voltage**, **Load Voltage**, and **Temperature** are all instances of the data set when the **Load Voltage** dropped below 2.7. **Capacity** and **Time** are computed from the start of each cycle to the point where **Load Voltage** drops below 2.7.

features should be tested against many different algorithms. Using visuals can help narrow down the best features and algorithms to choose from.

Understanding the process to create a model is straight forward, knowing how to collect and process data, feature selection, selecting an algorithm, and being able to write programs to combine all these parts successfully is not trivial. Azure provides services for all skill levels, where even users new to machine learning can create models in minutes. Selecting one of these services can be confusing though, as there are cloud-based, on-premise, and development platforms and tool options. Also, some of the names used for different services are very similar, which can add to the confusion. The cloud-based options are *Azure Machine Learning Studio*, and *Azure Machine Learning Services*. On-premises options are *SQL Server Machine Learning Services*, and *Microsoft Machine Learning Server*. The development platforms and tools are *Azure Data Science Virtual Machine*, *Azure Databricks*, *ML.NET*, and *Windows ML* [88]. The cloud-based options will be the focus for this section, with an emphasis on *Azure Machine Learning Studio* as it was used for the battery experiment. *Azure Machine Learning Studio* is great for creating, training, and deploying models quickly with ease. It is also better for smaller projects, where scaling the model to accept larger amounts of data are not a issue.

Azure Machine Learning Studio is a drag and drop interface which focuses on placing and connecting modules to build a model [91]. It does not run in the *Azure portal*, but can be accessed through the *Azure portal* using a *Machine Learning Studio Workspaces* resource or via normal web site addressing. A similar version is available in the portal called *Visual Interface*, but it is accessible using a *Machine Learning service workspaces* resource which will be discussed later. The layout of the interface is very intuitive and easy to understand, with several tabs to make organization and creating models easier. The projects tab is used for creating projects, which act as containers for organizing experiments and files. The

experiments tab is where the design and creation of a model begins. Take note that models are called training or predictive experiments in studio. The web services tab contains all web services. A web service is a fully trained predictive experiment that is accessible by other users and programs.

A free and standard version of *Azure Machine Learning Studio* are available where there are similarities and differences between both. Both versions can create predictive models and deploy classic web services. The standard version does have a monthly subscription fee, but it offers more storage, allows for larger models to be built, can train models faster, has a service level agreement, and can deploy new web services which offer more features compared to classic web services. If using the standard version, a *Machine Learning Studio Workspaces* resource must be created in the portal so compute and storage accounts can be allocated. The free version does not require a workspaces resource or need any connection to the portal.

To create a new experiment, select the experiments tab and click the new button to bring up a list of options. The user can create a blank experiment or use one of the many examples Microsoft provides as a starting point or guide to creating their own [86]. Selecting a blank experiment opens a new work space where modules can now be selected, dragged, and dropped in to. After dropping a few modules in the experiment, they can be connected with lines denoting the control flow of the experiment.

Every module is self-contained meaning no code is required to use it [84]. Each module can be customized to some extent, but most of the implementation is hidden from the user. Each module performs some task which sends its results to the next module, and the names of the modules typically describe their task. For example, a module titled **Import Data** is used for importing a data set. This module can connect to azure storage, but the users subscription level may limit this feature [83]. If this is the case, the data can be uploaded directly to

the work space. There are a wide array of modules to select from whose operations range from data transformation, statistical analysis, performing SQL commands, machine learning algorithms, and many more. Of the machine learning algorithms, regression, classification, anomaly detection, and clustering algorithms are all available with each category offering multiple choices. If a module does not exist for a certain task, the **Execute Python Script** and **Execute R Script** modules are available to implement custom code. This allows users to perform custom data transformations, create and use different metrics for analysis, and even run different algorithms not available as an existing module.

Once the data is imported and features are selected, one way to create a model is to use the **Split Data**, **Train Model**, **Score Model**, and **Evaluate Model** modules. The **Split Data** module takes data as input, and splits the data into training and testing sets. The training set is used to train the model, where the testing set is used to see how accurate the model is once trained. By default, 70% of the data becomes training, where 30% becomes testing but this number can be changed by the user. Randomization can also be applied when splitting the data, and it is highly recommended to do so. The training set is used as input for the **Train Model** module, along with the output of the module for the algorithm chosen. For the battery experiment, the *Linear Regression* module was chosen as the algorithm.

The **Train Model** module trains and creates a model for prediction. Each algorithm module has different hyperparameters that can be adjusted to train the model differently. Feature selection and model selection are extremely important, as changing hyperparameters and selecting different features can increase or decrease prediction accuracy. Visualizing the data before selection and training is one way to help select both as described earlier. The output model is combined with the testing set in the **Score Model** module to compute predictions for the testing set. This module does not give an overall score for the model, but each individual prediction can be compared to its actual value. See Figure 4.14 for an example comparing

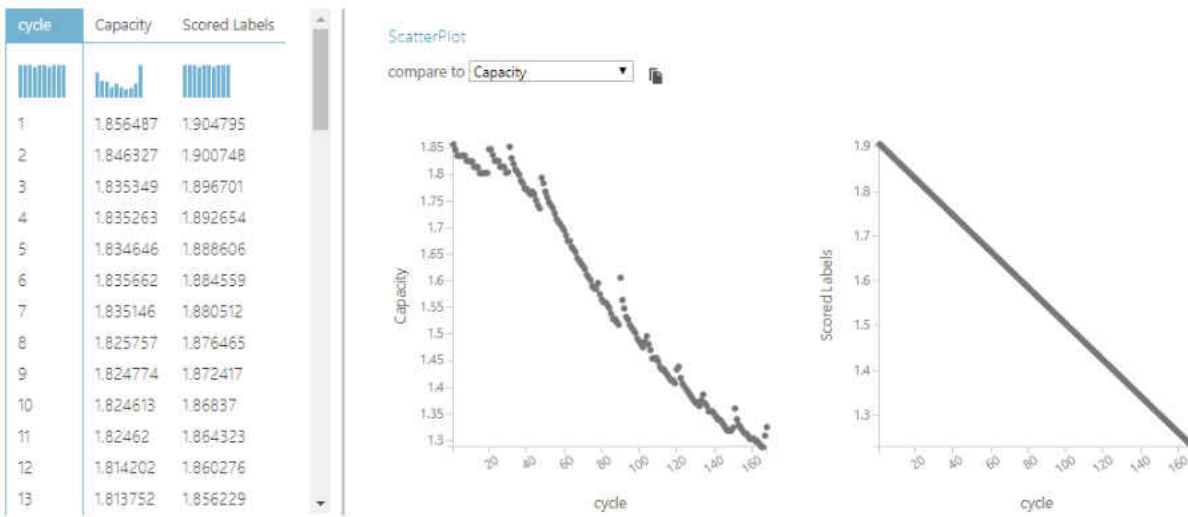


Figure 4.14: Some visualization features comparing real to predicted values.

the two values along with some other visualization features offered in studio.

To get the overall accuracy of the model, the output from the score module can be sent to the Evaluate Model module. Good accuracy scores can be anything above 75%, but there is no set number for a good score as it may vary based on the complexity of the model and the predictions being made. If a number is too high, say 99% or 100%, the model is probably overfitted. This means too much noise is being captured when training which causes the model to fit the training data too well. If the number is very low, the model is probably underfitted. This occurs when no general trends are found in the data and the model is oversimplified [12]. If either occurs, adding or removing features from the data, modifying the hyperparameters of the current model, or selecting a new algorithm are possible fixes. Both scenarios can lead to very poor predictions of new data if not addressed. See Figure 4.15 for examples of underfitting, a good fit, and overfitting.

Another way to create a model is to use the *Cross Validate Model* module. It splits, trains, scores, and evaluates the model all in one module. It uses a technique called cross validation to split the data randomly into equal folds, or partitions [82]. The default number

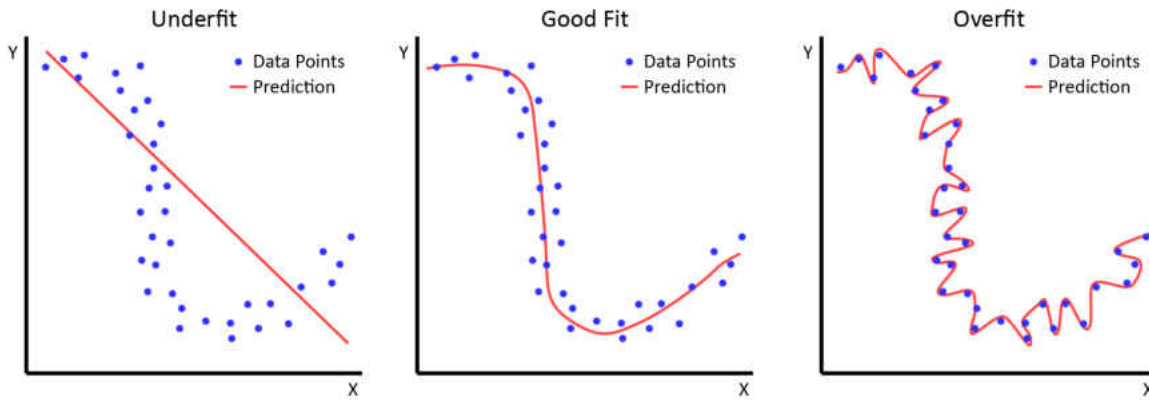


Figure 4.15: Underfitting, good fitting, and overfitting examples for machine learning [12].

of folds is 10 but this can be changed using the **Partition and Sample** module if desired. The model is trained using all but one of the folds. The fold not used, called the holdout fold, is used for testing the trained model. This process is repeated so each fold is used as testing data. For example, if there are 10 folds the model would be trained 10 times. During each training iteration, a different holdout fold is selected where the remaining 9 folds are used for training. Evaluation is done during each iteration and stored, if each iterations score is vastly different than there may be problems with the model or data. The advantages of using cross validation over splitting the data are it uses more data for training and testing making it better for smaller data sets, it can help prevent overfitting, and it evaluates the quality of the data being used as well. A disadvantage is it is more computationally expensive, so larger data sets may take more time. See Figure 4.16 for an example of a battery training experiment using both methods described to create a model.

When a user thinks the training experiments prediction capabilities are sufficient, it must be deployed as a web service so it may be used against new data. The process is very easy, click the **Set Up Web Service** option then select **Predictive Web Service**. This will automatically start shuffling, adding, and removing modules to convert the training

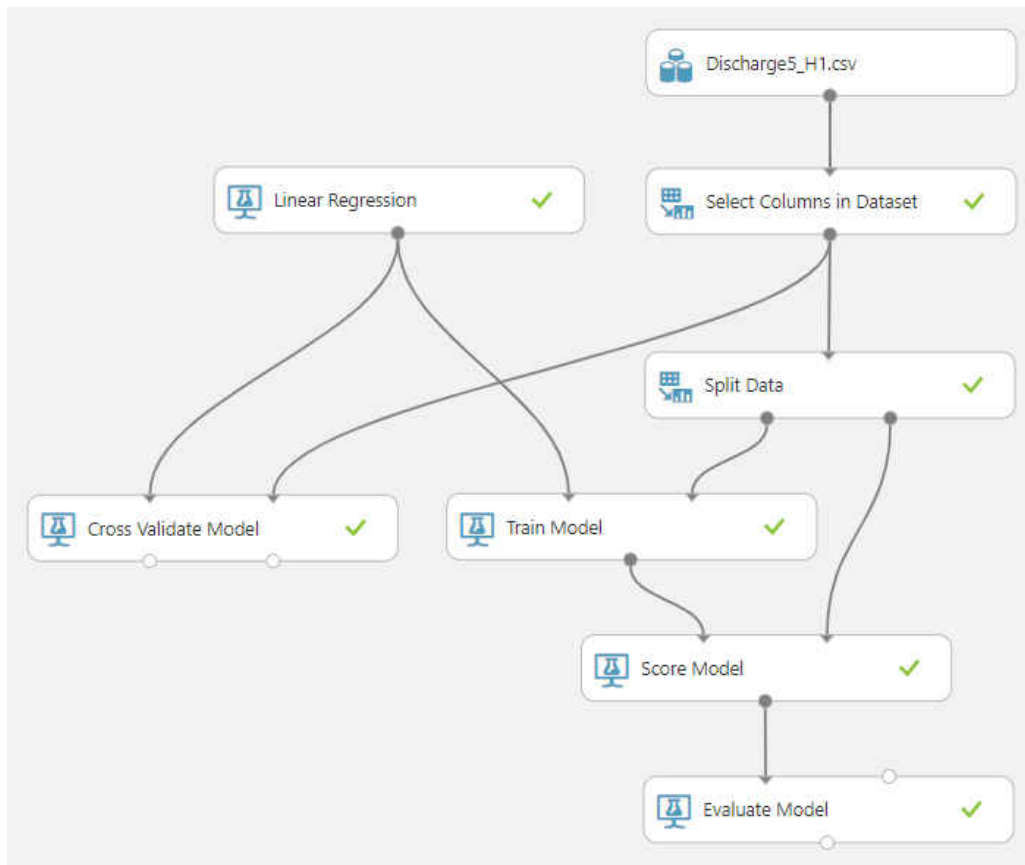


Figure 4.16: Experiment using two different methods to train a model.

experiment to a predictive experiment. If multiple algorithm modules are being tested in one experiment, it is recommended to remove all unwanted modules unrelated to the selected algorithm so only the relevant modules remain [87]. The conversion process will not remove the module used to import data, as it is used as a template for what features will be passed into the predictive experiment. When building the training experiment the variable to predict was allowed to pass through the **Select Columns in Dataset** module as it was needed for training the model, with new data this value is unknown so it can be removed from feature selection in the predictive experiment. See Figure 4.17 for the result of converting Figure 4.16 to a predictive experiment.

When complete, the **Deploy as Web Service** option can be selected which will add the

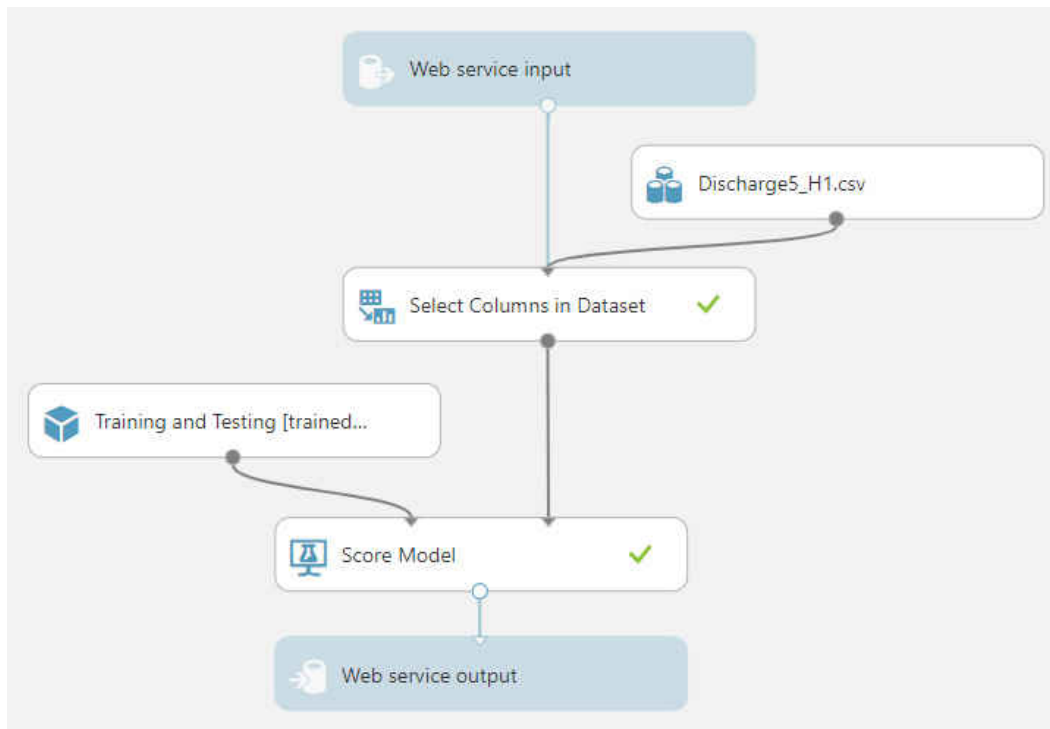


Figure 4.17: Result of converting Figure 4.16 experiment to a web service prediction model.

model to the Web services tab as a classic or new web service. As mentioned previously, the free version of studio can only create classic web services, where the standard version can create both. From the web services tab in *Azure Machine Learning Studio*, the newly created web service should appear and can now be used for prediction. Single values or batch testing can be done through the browser using *Azure Machine Learning Web Services* or excel. *Azure Machine Learning Web Services* is another option for managing, testing, and deleting web services created in studio [52]. Selecting the *New Web Services Experience* link from the web service tab in studio will open up *Azure Machine Learning Web Services* in a new web page. The *Test* tab contains all options for running predictions against the model. If using excel, the add-in for Azure Machine Learning must be installed and the URL and API key must be entered. The URL and API keys are found in the *Consume* tab from the *Azure Machine Learning Web Services* page. The user may also click the excel link from the

studio or web services page, this will download a blank excel file with the Azure Machine Learning add-in, URL, and API keys installed. Data in excel can be selected and sent to the web service for prediction, with the results being returned and populated in excel.

Azure Machine Learning Services is another cloud-based machine learning platform in Azure. It offers multiple ways to build, train, and manage predictive models from multiple sources in the cloud. It is *Azure portal* based, meaning a user must create a *Machine Learning Service Workspaces* resource to use it. Predictive models can be created using *Visual Interface*, *Automated Machine Learning*, or the *Azure Machine Learning Python SDK* can be used with multiple programming interfaces [90]. Currently several of the tools are in preview stage, which means the service should primarily be used for testing or proof of concept. Services labeled as preview can be removed at any time, have different service level agreements, and are used as is without expectations of further updates [81].

Visual Interface is a partial clone of *Azure Machine Learning Studio* with some minor differences. *Visual Interface* is a drag and drop workspace to create and train models using prebuilt modules, but everything is done in the *Azure portal* so it is fully integrated with *Azure Machine Learning Services* unlike studio [93]. It also has fewer modules to choose from, and it relies on a user's own compute resources to train and test models [91]. Compute resources are created by the user, used by the workspace to run and train models, and can be located on a local machine or in the cloud. Like other Azure services, compute resources running in the cloud are only billed when they are active. When training and testing, it is recommended to use a local machine with smaller data sets. When the model is ready, compute resources can be pushed to the cloud where they can scale up or down based on demand.

Automated Machine Learning, or sometimes referred to as *Automated ML*, is a way for Azure to train, test, and tune multiple models at a time [89]. The user only needs to

provide a data set, feature selection, prediction variable, and whether they want to perform regression, classification, or forecasting. When automation is started, Azure will train many models using multiple algorithms, modifying the hyperparameters during each iteration in an attempt to create the best model for prediction. When complete, all the models trained are presented to the user with prediction scores and other evaluation metrics. *Automated ML* can be run in the portal or by using the *Azure Machine Learning Python SDK* in Python. Figure 4.18 shows a sample run using the *Azure portal* on the battery data set. Each blue dot is not a different model, but a models score during one of its training iterations. The sample run took about 15 min. and created 26 different models. Many of the scores are extremely high though, so there is a good chance many of the models are overfitted.

If a user prefers more control and writing their own code, the *Azure Machine Learning Python SDK* can be used with *Jupyter Notebook*, any Python IDE, or Visual Studio Code to connect to *Azure Machine Learning Services* [92]. *Jupyter Notebook* is an open source

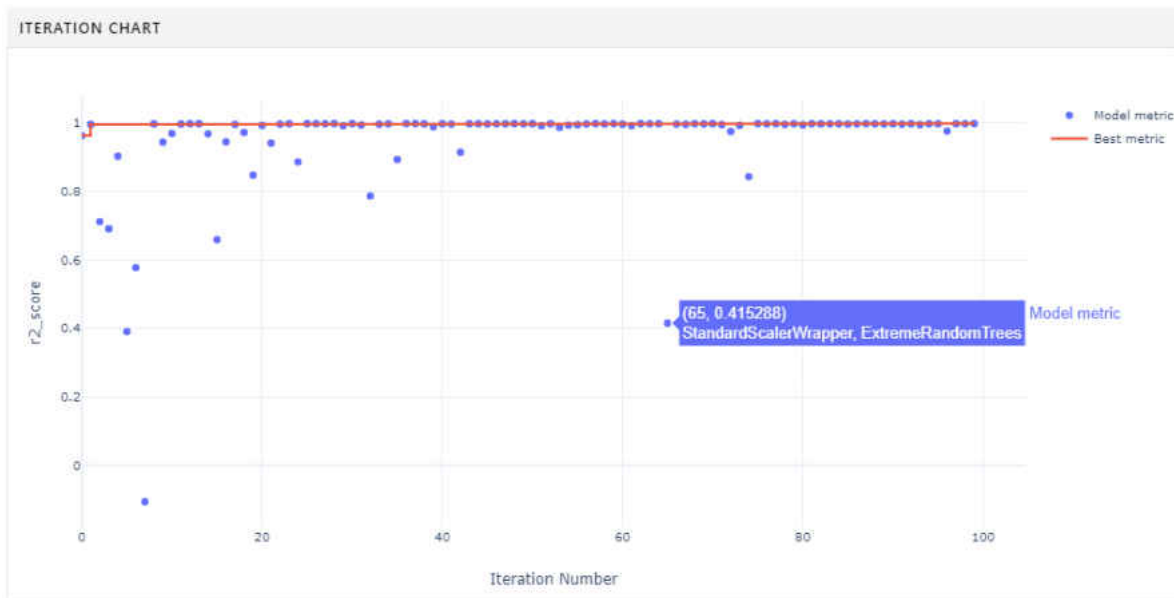


Figure 4.18: Automated Machine Learning results.

web application that allows users to write and share code through a browser [45]. Microsoft also offers a service called *Azure Notebooks*, which essentially is Jupyter notebooks but it runs in the cloud and requires no installation [85]. By importing the SDK, any Python environment can manage datasets or cloud resources, train models, deploy web services, and use *Automated Machine Learning*. Using the SDK requires more knowledge about programming, where *Automated ML* and *Visual Interface* are code free.

Many of the steps used to to create a predictive model for the battery experiment have been detailed in the previous chapter focusing on *Azure Machine Learning Studio*, even though many of the tools in *Azure Machine Learning Services* could have been used as well. Figure 4.19 shows a plot using the *Execute Python Script* module comparing the training data to the prediction line. For this data set the predicted replacement cycle is only off by three compared to when the battery should be replaced.

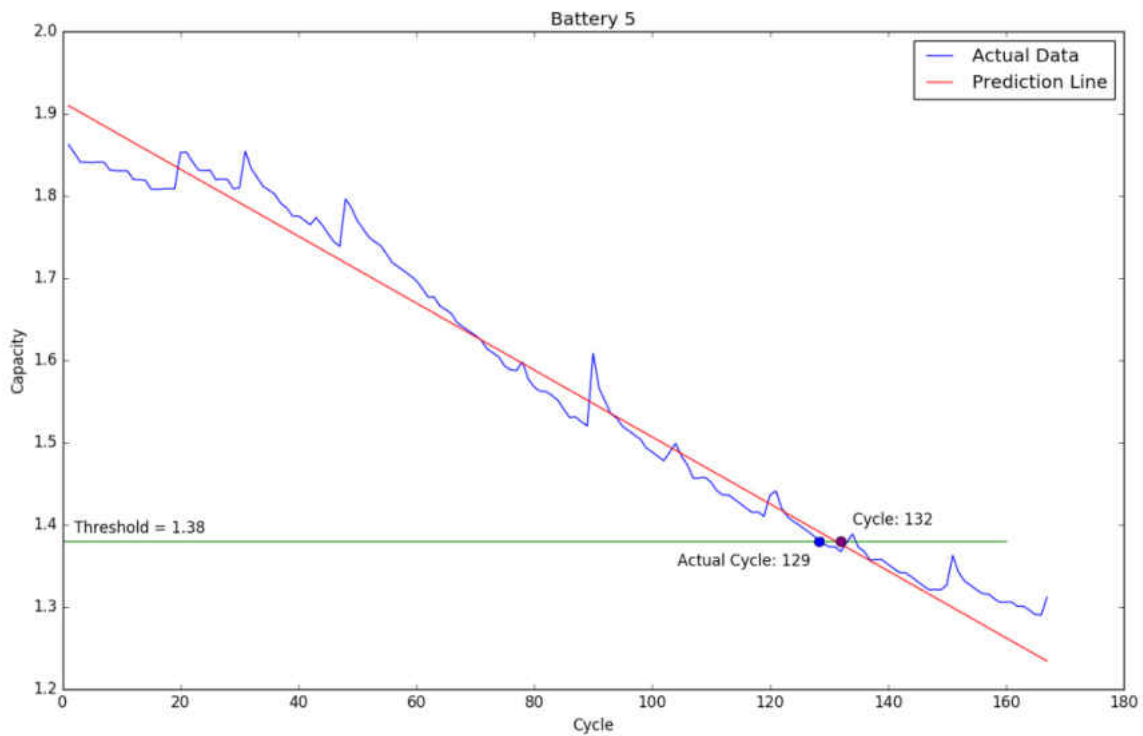


Figure 4.19: Output created using python script module to show prediction.

4.8 Power BI

Many of Azure services provide some basic visualization options for data analysis, but they are limited in terms of customization and creating presentations. Microsoft offers *Power BI*, a collaborative drag and drop visualization tool that allows users to transform their data into interactive presentations. There are several services offered in *Power Bi*, but three useful tools for creating presentations and delivering actionable business insights are *Power BI Desktop*, *Power BI service*, and *Power BI mobile apps* [97].

Power BI Desktop is a free, downloadable software that allows users to connect, import, modify, and visualize their data on their own computer [80]. When importing or connecting to the data source, the *Query Editor* is used to remove columns, change data types, or perform other modifications to the data before creating visuals. Modifications made are saved in the query, so each time the file is opened again the data remains in the same format. This also keeps the source data intact and allows queries to be reversed [96]. Special attention should be placed on the column data type, as data types determine how the data is visualized. When the data is in the correct format, the *Query Editor* can be closed and the data can be used to create visualizations. Visualizations, sometimes called visuals, are pre-configured templates for bar charts, line charts, scatter charts, maps, and many other options. Each visualization has options that can be further edited to fit a specific visual style, such as modifying colors, adding text, and selecting a range for the data. Visuals can be placed on report pages, which are canvases to drag and drop visuals on to. Typically *Power BI Desktop* is used for creating reports, and not sharing them. To share reports, they can be published and uploaded to *Power BI service*.

Power BI service is a cloud-based, Software as a Service (SaaS) tool primarily used for collaboration on pre-built models from *Power BI Desktop*. Both can create and edit reports, but *Power BI desktop* is better suited for it [63]. The key components of a *Power*

BI Service are workspaces, dashboards, reports, datasets, and workbooks [94]. Workspaces are containers for the other key components of a *Power BI service*. Workspaces can either be a *my workspace* or *app workspaces*. A *my workspace* is a personal workspace for a single user, where only they can access and modify its content. *App workspaces* are for collaboration and sharing workspace content with other users. They also are used for app creation where end users can interact with the data but not modify it. **Dashboards** are a single, shareable canvas that can display visualizations from multiple reports, datasets, and other tools such as excel. They are useful for grouping many related visualizations together to aid in business decisions. **Reports** consist of one or more pages of visualizations in a workspace. A report can only get data from one dataset, but reports can belong to multiple dashboards. **Datasets** are data a user wishes to visualize. They can be imported or connected to, and be used in multiple workspaces. **Workbooks** are similar to datasets, except the data is from an excel workbook.

Embedded analytics is the process of inserting meaningful visuals, reports, and dashboards into an application through application program interfaces (APIs) to give a customer or an organization business insight for better decision making. Embedding can be done using the APIs in *Power BI service*, or through the *Azure portal* with *Power BI Embedded service* [95]. Embedding for a customer and a organization are two different tasks. Embedding for a customer allows a client to access visuals, reports, and dashboards without having a *Power BI* account. The customer does not even need to know how to use *Power BI*, as much of the implementation is hidden from them. Embedding for an organization requires all users of the organization to have a *Power BI* account. Each user is given access to information only required for their work, narrowing the information flow and giving the user more focus to relevant information for their task. Another way to view information is through *Power BI mobile apps*. The apps are available for apple, android, and windows 10 devices. Users

simply download the app and sign in. The app will list all dashboards and reports available so they can be monitored from any device. Dashboards and reports cannot be created using *Power BI mobile apps*, which must be done in *Power BI service* or *Power BI Desktop*.

For the battery experiment, an excel file with actual and predicted data was loaded into *Power BI Desktop*. After the file is loaded, the *Query Editor* will open allowing for the modification of the data. When complete, clicking ‘close and apply’ will cause the query editor to exit after applying all changes to the data set. With the data loaded, visualizations can be added. Clicking any of the options in the Visualizations tab will cause a generic box to appear in the report page. With the generic box selected, columns can be dragged and dropped into the settings of the box, which are below the Visualizations tab. After a few columns are added, the box will populate itself with the corresponding values. Visualizations can be changed to any other visualization at any time. If a visual is a stacked bar chart, selecting stacked area chart will update the visual automatically. Some visualizations may not have a one to one correlation, such as a line chart and a pie chart, so some of the settings will need to be modified. The whole process is very intuitive, and as it is a design process, trial and error will be involved. See Figure 4.20 for example output from the data set.

4.9 Experiment Observations

Throughout this chapter, many of the steps used to create the experiment have been described and shown in great detail. After all the setup is complete and the predictive model is released as a web service, new batteries can be connected to sensors which will send their messages to the *IoT Hub*. *Stream Analytics* can monitor these messages and store the data in a table while scanning the current battery cycle. When the battery voltage dips below 2.7 for the specified cycle, it will trigger a *Azure Function* to read all relevant data for the

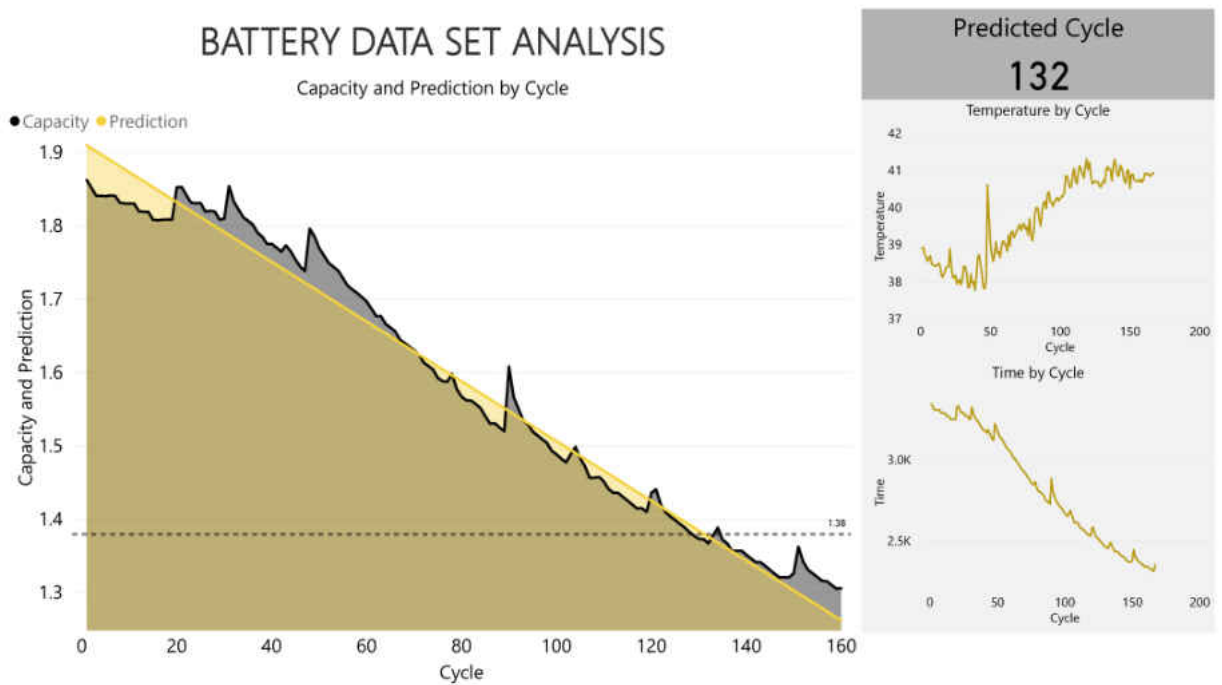


Figure 4.20: Power BI report showing prediction compared to actual values. Temperature (in Celsius) is the instance of the data set when the Battery Voltage dropped below 2.7. Capacity (in Coulomb) and Time (in seconds) are computed from the start of each cycle to the point where Battery Voltage drops below 2.7.

new battery and compute the necessary values. The computed values can be sent to the predictive web service to get a better idea when the battery should be changed out.

If a user wants to cut costs or to ensure that the amount of data sent to the *IoT Hub* is manageable for human observers, the user can create a custom endpoint for the *IoT Hub* resource, which stores all telemetry data to *Blob storage*. The user can then monitor the cycles on their own and wait till the desired cycle is reached to launch a slightly modified *Azure Function* to read from *Blob storage* and to compute the new values. This eliminates *Stream Analytics* from the process, which can greatly reduce the overall costs. For the battery experiment, it is possible to do this as the time intervals between the MQTT messages are large enough for a human observer to manage. There will be some loss in efficiency, as *Stream*

Analytics triggers the *Azure Function* the moment it finds the value where a human will have to continuously check. A hybrid approach is also possible, where telemetry data sent to the *IoT Hub* is stored in *Blob storage* and *Stream Analytics* is started much later in the batteries life cycle. This reduces the overall costs and maintains efficiency. If multiple battery's are being monitored or the time intervals decreased to a speed humans cannot perceive, *Stream Analytics* should be used and the costs are unavoidable.

5 High Velocity and High Volume Data

5.1 Cloud IoT Pitfalls

The amount of data generated by IoT devices is based on the number of devices times how frequently they send telemetry. IoT data can be considered a subset of Big Data, which is often described using a different set of words that start with the letter V. Typically 3 V's are used, but sometimes 5, 6, 7, or even as many as 51 can be applied [41]. For IoT big data, Mohammadi et al. described 6 V's. They are volume, velocity, variety, veracity, variability, and value [110]. These six traits can be applied to any IoT system, but not all of the 6 V's will apply to every IoT system. Many IoT systems will be very specific in their scope, limiting the type of data that is collected and controlling the frequency the data is sent at. For example, a study in the United Kingdom collected data from 2,000 smart plugs installed in 40 homes and generated 4 billion messages a month [21]. For this study, volume, velocity, and value would be the best choices to describe the data because the type and frequency of the data is known.

IoT systems, like the UK housing study, can run in to problems with public cloud platforms when they need to increase the total number of sensors or messages increasing the volume and velocity of the telemetry data. High volume, high velocity data doesn't just come from many devices sending a few messages each second, but can also come from few devices sending tens of thousands of messages per second. The problem created is many cloud gateways used to connect and receive data have a throughput limit, meaning the gateway

can only process so many messages in a given time frame.

For example, looking at Figure 4.3 shows the seven tiers of an *Azure IoT Hub*. The tiers with the highest ingestion rates are B3 and S3, which if either increase their units to 10 units can process 3 billion messages per day, or up to 93 billion messages per month at a cost of \$25,000 per month. The UK housing study produced 4 billion messages a month from 40 homes, a single *Azure IoT Hub* B3 or S3 tier would be able to handle 930 houses. If the study were to expand to all 26.4 million households in the United Kingdom [126], there would need to be 28,388 *Azure IoT Hub* B3 or S3 tiers at a cost of about \$142 million per month if using B3 or \$710 million dollars per month if using S3. What if instead of having millions of devices, only a few devices send tens of thousands of messages per second? The *Azure IoT Hub* B3 or S3 tier with 10 units can handle 34,722 messages per second. If a single device can send 10,000 messages per second, the highest *Azure IoT Hub* tiers can handle at most 3 devices. This severely limits the amount of devices that can be used while still being very expensive. It is possible to increase throughput by performing batch updates and reduce costs of the gateway [42]. Each IoT Device collects multiple readings, compresses the data, and uploads a batch of messages. If devices with very low processing power are used, an edge device can be added between the device and gateway to receive, compress, and perform batch uploads.

Once the data passes through the cloud gateway it must be stored to allow for easy access and fast queries. Because the data is high velocity it is also high volume, meaning lots of memory will be required to store the data. Many cloud platforms offer numerous ways to store large amounts of data, but being able to query this data efficiently is very difficult. Some common options provided for storing large amounts of data are Relational databases, NoSQL databases, Data Lakes, and Blob storage. Each option has its own strengths and weaknesses, where selection of the storage format will be project dependent.

Relational databases are one of the most common ways to store, modify, and query data. Relational databases use SQL, Structured Query Language, as the backbone for inserting, removing, and querying elements in the database. They have been around since the 1970's and follow the ACID properties, which are Atomicity, Consistency, Isolation, and Durability [133]. ACID properties ensure that integrity of the database is maintained before and after each transaction is performed in the database. The schema for a relational database is very strict, which allows only structured data to be entered.

NoSQL, Not Only SQL, databases started to gain popularity in the early 2000's with the increase in popularity of the internet. More internet traffic and larger data sets created a higher demand for faster web sites. Developers attempted to optimize their relational databases by simplifying their schema to improve availability and decrease latency. This did not work as it does not address the horizontal scalability of the database [133].

NoSQL databases were invented to fix the short comings of relational databases with Big Data, not replace them entirely. A characteristic common to NoSQL databases are a schemaless design allowing for unstructured, semi-structured, and structured data to be stored. A relaxation of the ACID properties allows for higher availability, lower latency with Web applications, and data is replicated and distributed among many servers to allow for horizontal scalability [17]. The data can be stored in different formats such as key-value, wide-column, document, and graph. Each format provides different functionality suited for different applications, providing efficient queries over large amounts of data. Because it is a distributed database, it follows the CAP theorem which stands for Consistency, Availability, and Partition tolerance. The theorem states that any distributed system can only satisfy up to two of the three properties at the same time, never all three [17].

Research done by Samanta et al. [123] compared the performance of a relational database, MySQL, to two NoSQL databases, MongoDB and Cassandra. The data set used in the

testing was a 100 MB CSV file with 31 columns and 1,000,000 rows. The tests included 11 queries that were performed against all three databases. Each query involved a different set of tasks, with execution time being the average of repeating the query at least 10 times. After inserting all the data, the storage requirements for MySQL was 143.43 MB, MongoDB was 961 MB, and Cassandra was 212.8 MB. The results showed for each query MySQL had the lowest execution time, which was typically followed by Cassandra and then MongoDB.

Yassien et al. performed a similar analysis with a data set of 11 GB and different queries, but instead of Cassandra they used HBase. Their results showed MySQL had the lowest read latency, but MongoDB and HBase had lower write latency [133]. According to Khan et al. [41], big data volume is on the scale of Petabytes, Exabytes, and Zettabytes, which are orders of magnitude larger than Megabytes or Gigabytes. If data of this scale was used in either test MySQL's performance would have been much worse.

Data Lakes were invented as another solution to handle the rise of Big Data. Data Lakes can store massive amounts of structured, semi-structured, and unstructured data of any kind in a central place to be queried and analyzed at a later time. Data Lakes are typically divided in to three parts. First is a storage system often built on Hadoop Distributed File System (HDFS). Second is a metadata system. Each stored piece of data has it's own metadata tag which allows users to understand and access the data. The third part is a tool to access and analyze the Data Lake. Analysis of the Data Lake is usually done with a schema-on-read, which applies a schema to the data as it queries it. Two common tools for analysis are MapReduce and Spark [113]. For projects that involve high volume and velocity data, Data Lakes offer a viable and affordable storage solution. Fast writing is done because no schema is every applied to the data, it simply gets added to the Data Lake. Because no schema is applied though, querying the data can take longer which makes it a poor choice if time sensitive information needs to be found quickly.

Blob, or object, storage is a cost effective and simple way of storing data. Blob stands for Binary Large Object and any type of data can be stored in a blob. Blobs can store massive amounts of data, with limits varying by type of blob and cloud provider. For example, Microsoft's Azure can store Block Blobs up to 4.7 TB and Page Blobs up to 8 TB [61]. A user can create as many blobs as needed, allowing for near unlimited and affordable storage. Access policies allow for easier sharing and protection. Each blob is assigned an ID and is associated with the access key of its container. Sharing these keys allows for specific blobs to be shared while hiding sensitive information in private blobs. If millions of small files need to be stored, storing all the files together in one blob alleviates the maintenance nightmare by only needing to keep track of a few blobs instead of millions of files [20]. Blobs can also have extremely high throughput depending on the cloud provider. Azure announced possible throughput of up to 12.5 GB per second when using the High Throughput Block Blob [109]. There are many advantages to blobs, but query performance is often very poor. Blobs are not stored in a hierarchical format and they can contain any type of information. Similar to Data Lakes, the metadata of each blob needs to be stored and made searchable to improve query performance.

While IoT based cloud platforms are used in many IoT scenarios, IoT projects that involve high velocity and high volume data can place a large burden on a cloud platform. As shown, cloud IoT gateways cannot handle the high velocity data without adding more gateways at great cost or performing batch updates using edge devices. Similarly with the storage options, each has its own benefits and trade offs which need to be accounted for when making a selection. The goal, if possible, is to provide a way to store, index, and query IoT high volume and velocity data efficiently and in a cost-effective manner using cloud based technology.

5.2 Possible Solutions

5.2.1 Distributed Model

The first solution uses open source software to create a distributed IoT system than can run on cloud virtual machines. It is composed of several parts of the Apache Software Foundation, which provides many open source software solutions that can be used and combined in different ways. The proposed combination uses Kafka [24] for ingestion, Druid [23, 132] for processing and querying, Zookeeper [26] for cluster management, and Hadoop Distributed File System (HDFS) [25] for data storage.

Apache Kafka is a distributed streaming platform which follows the publish/subscribe model, where IoT devices can publish their telemetry data and subscribers can read from it. Kafka is well suited for communications inside a shared network, but if Kafka is running in a cloud it is not immediately usable for ingesting IoT data without some modifications [115]. A major concern is devices need to be able to directly connect to Kafka. Because the IoT devices exist outside the cloud and Kafka is in the cloud, Kafka would need to be exposed to the public internet to be visible to the devices. To keep it encapsulated in the cloud, a server running an MQTT broker is required to act as a secure bridge for telemetry data between devices and Kafka. The MQTT Broker can be custom built for the project, or it can be open sourced or purchased such as the different options offered from HiveMQ. If Kafka struggles with the high velocity data, the server running the MQTT Broker can perform batch uploads of multiple messages to increase throughput.

The subscriber to Kafka would be Druid, which is an online analytical processing (OLAP) distributed data store. It is designed for real-time analytics, is column orientated, and is optimized for low query latency on large data sets. IoT data can vary greatly between sensors, as each sensor is tailored to a unique scenario or environment. A common trait with

IoT data is it is time series based, meaning all IoT data will have a timestamp related to when the information was collected. Time series databases (TSDB) such as InfluxDB [37], TimescaleDB [128], and OpenTSDB [116] were designed to store and query large volumes of time series data. As stated previously parts of this design can be swapped out, such as replacing Druid and HDFS with OpenTSDB and HBase. Druid was selected because it pairs well with Kafka, and it performs well not only on time series data but multi field data. Currently Airbnb, Alibaba, Cisco, eBay, Hulu, Lyft, Metamarkets, Netflix, Paypal, Twitter, Walmart and many more use Druid to monitor large ingestion streams and perform analytics. The highest ingestion rates come from Metamarkets, which claims to ingest 200 billion events per day which is about 2.4 million events/s [22].

Druid's architecture is based around having a cluster of nodes where each node performs a unique operation based on its type. The four types of nodes used are real-time, historical, broker, and coordinator. Because it is a distributed system, all nodes share metadata information with Zookeeper, a coordination system for distributed file systems. Real-time nodes are responsible for ingesting and querying event streams. Kafka can be used as a single point of entry for telemetry data and act as a message bus which real-time nodes can read from. They are also responsible for writing the data to deep storage, where deep storage can be any compatible storage service with Druid. For this solution HDFS was chosen as it is part of the Apache Framework and is widely used and tested.

Historical nodes load and serve the data created by real-time nodes. All data that is stored is immutable. Immutable data offers efficient replication, read consistency, and parallelization across all historical nodes. Broker nodes read the metadata stored in Zookeeper to determine which data segments are queryable and route the queries to the correct real-time and historical nodes to be processed. Coordinator nodes monitor the data in historical nodes. They tell historical nodes when to load, remove, and replicate data based on a set

of rules loaded in to each coordinator node. They also monitor load balancing to distribute data for better query performance.

Being distributed makes the system horizontally scalable, making it well suited for ingestion, storing, and querying of high velocity and high volume data. Using open source software has many benefits as well. It reduces costs as open source is typically free, it can have a robust support network of active users, and it gives more flexibility as the software can be tailored to unique scenarios. Running this solution in the cloud alleviates having to manage all hardware, which includes networking, operating system updates, and fixing, adding, or replacing machines.

A few downsides are costs and expertise. The costs of running multiple virtual machines can add up quickly. Knowing how much data will be processed can help determine what type of virtual machines will be needed, but sometimes this data isn't readily available or it may change as the project continues making estimation difficult. To further complicate things, understanding Quality of Service (QoS), Service Level Agreements (SLA), and the pricing structures for different virtual machines can be very confusing as they vary between cloud providers [18]. For example, virtual machines charge differently for compute and storage services per virtual machine. A Kafka cluster ingesting from 1 MB/s to 32 MB/s can cost between a few hundred dollars to over \$1500 per month just for storage for that cluster using *AWS EC2* [2]. Using many of the tools selected will require a solid knowledge base, as getting everything to work together as a cohesive unit will require skill and time. For some developers this may not be a feasible solution as the knowledge required and costs may be too high.

5.2.2 Cost Efficient Model

An alternative solution is to combine many of the existing services from a cloud provider with custom software to create a design that is cost efficient, scalable, and queryable at the expense of added latency. The design consists of edge devices, cloud provided SDK's, blob storage, storage queues, serverless functions, and a single virtual machine for processing queries. Edge devices, a stand alone server, will be responsible for collecting IoT telemetry data from devices, compressing batches of data based on a set time frame, and performing batch uploads directly to blob storage using cloud provided SDK's. Since it is time series data, each entry in the batch should be sorted by time making it easier to keep track of which messages belong to which batch. The batch is stored as a blob in compressed format, which helps reduce bandwidth and storage needed in the cloud. For example, a CSV file with 200,000 different entries with 13 fields was created spanning 10 seconds to mimic device telemetry. With out any compression the file size is about 36MB, when using basic windows compression it is only 3.5MB. The batch is only decompressed when being scanned by the serverless function or when returning a query, both which are discussed later. When the batch of data has been successfully uploaded as a blob, the blob URL can be obtained and uploaded to a storage queue in the cloud.

The storage queues sole responsibility is to alert a serverless function a new blob has been created. Combining serverless functions and queue storage take advantage of the pay-as-you-go billing that many cloud providers offer, as serverless functions are only charged when they are running and they will only run when the queue triggers them. The queue creates an alert which triggers the function to execute, which reads the blob URL at the head of the queue. The function uses the blob URL to download the actual blob, decompress it, and the read the telemetry data creating a minimum and maximum range for each field. It is also possible to calculate ranges for other fields that are not part of the initial data set, such as

key performance indicators (KPI) like power levels or total harmonic distortion (THD). If certain fields are used for error detection, a minimum or maximum threshold can be applied when scanning. If the either threshold is met, the blob URL and a warning message can be added to a warnings queue. The developer is then alerted to when a message is added and they can inspect further.

After reading the entire blob, a metadata tag is created which contains the device id, current year, current month, current day, start time of blob, end time of blob, all minimum and maximum values for each field and KPI, and the blob URL. To reduce cloud computing costs the edge device can perform many of the calculations of each batch of data. A downside is this will put a greater strain on the edge device, requiring a more powerful and expensive server to handle the increased load. It is up to the developer which avenue to choose, but having each part of the system perform one job allows for easier implementation and debugging.

When the metadata for a blob is complete, it is sent to a single virtual machine which is responsible for the storing metadata tags, and answering queries. The storage structure used takes ideas from Colmenares *et al.* [16] MultiDimensional DataStore (MDDS), which focused on creating a single node that can handle ingestion, indexing, storage, and querying of massive amounts of data. To store and index data, a two-level indexing structure of R*-trees with kd-trees was used to increase query performance. A R*-tree is a variation of the R-tree, which creates minimum bounding rectangles that are useful in storing multidimensional data. The difference between the two is the R*-tree will “self balance” itself, which involves a complicated process of removal and reinsertion of data to reduce overlap and increase query performance. A kd-tree is a data structure similar to a binary search tree, only at each node depth a different dimension is used to sort the data. Kd-trees are useful for storing multidimensional data, as well as performing range queries, point queries, and finding a

nearest neighbor. R*-trees and kd-trees can suffer from the curse of dimensionality, which is when the number of dimensions exceeds a certain threshold causing poor query performance. High dimensions usually means 10 to 16, but can be as low as 5 or 6 in some cases [13].

The virtual machine will have a varying level indexing structure, where the number of levels is determined by how much metadata is going to be stored. Because each device will be creating different telemetry, the first level will always be sorted by device id to keep the device data separate. Since IoT data is time series based, the next levels of the structure will focus on year, month, day, and hour. Year, month, and day are the varying part of the structure, as the amount of metadata to be stored will determine which values to use. If the number of metadata tags to store is low, year, month, and day can all be used. If the number is very high, then the developer needs to determine which levels to include as the memory of the virtual machine is finite. If one of these is not included in the indexing structure it will still be used when storing the indexing structure to blob storage. For example, if month and day were used and year was not, every year would have its own indexing structure stored as a blob. When a query is looking for data spanning years, each individual indexing structure relating to a year in the range of the query is loaded from blob storage in to the virtual machine for scanning.

The second lowest level will always be hour, where each hour level is broken up in to 24 separate buckets for each hour of the day. Each bucket points to an R*-tree where all metadata tags are stored for that time period. Breaking the indexing structure in to a tree where each level focuses on a different unit of time helps reduce the amount of data being stored in each R*-tree. It can also be advantageous to know what fields will have very little variance, as these fields do not need to be indexed to help reduce the curse of dimensionality. If the number of metadata tags created per hour is low, a list can be used instead of an R*-tree to store the metadata tags.

When a query is entered, the virtual machine will work its way through the levels and store all metadata tags returned from the R*-trees. When all tags are retrieved, the blobs are downloaded to the virtual machine to display the results to the user. If more processing is required, say only specific values are needed and not entire chunks of data, the virtual machine can parse out the relevant information from each blob. If the number of devices and time frame used are constrained limiting the search scope, the amount of blobs returned will be manageable with most latency coming from downloading and decompressing the blobs. If a query searches many devices with larger time frames this can drastically increase the amount of blobs returned. After all metadata tags are retrieved, the virtual machine can perform a quick estimation based on the number of blobs and the size of each blob for how long it will take to download and further process. The user can then decide whether to proceed or not based on the result, or they can further trim the results by modifying their original query. This design does involve re-computing derived fields such as KPIs. If the query limits the scope, the amount of blobs returned should be manageable for most computers to be able to download and process with acceptable latency.

This proposal offers a more affordable solution to using a distributed system. Costs are saved by using blob storage, bypassing cloud gateways with cloud provided SDKs, taking advantage of serverless functions with pay-as-you-go billing, and running a single virtual machine to perform queries. The solution is not perfect, as every design will have trade offs. Latency is increased, making it difficult to use in time critical situations. Using existing cloud services can be very restraining and at times frustrating as documentation can be lacking and confusing at the same time. Fine tuning parameters such as time frame size, error detection thresholds, and which fields to index in the R*-tree for optimal query performance are trial and error based. If many devices are required, each device requires its own indexing structure limiting the number of devices the system can handle. The goal was to provide a way to

store, index, and query IoT high volume and velocity data efficiently and in a cost-effective manner in the cloud. If costs are no issue, using a distributed system will provide good performance. Otherwise the alternative is to combine many of the existing technologies into a custom solution.

6 AWS vs Azure vs GCP

According to Harvey et al. [34], the top three cloud providers ordered from first to third are Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP). AWS's maturity, continued ingenuity, broad distribution of data centers, and extensive set of tools are reasons why developers and companies continue to make it the number one cloud platform. It is not surprising then that these are the top three cloud-based IoT platforms as well. According to an IoT developer Survey from the Eclipse Foundation [27], 34% of respondents use AWS, 23% use Azure, and 20% use GCP. As each platform competes for more market share, the services provided become more similar. The similarities and differences are made clear as the remainder of this chapter will compare AWS, Azure, and GCP based on IoT related services. Pricing is not directly addressed even though it is a major factor when selecting a cloud provider. Each provider uses different metrics for compute resources, services can have different tiers with varying pricing, and prices are constantly changing as new services are added or updated making it difficult to capture the total costs.

6.1 IoT Related Services

IoT related services are specific to IoT use cases. These services include IoT cloud gateways, acceptable message formats, support for edge computing, and provided software development kits (SDKs). Looking at Figure 6.1, the similarities are evident as each category has a similar service for each provider. The main differences are gateway throughput limits, possible message formats, and provided device SDKs. Each provider offers different levels of throughput, where limitations and scaling are unique to each gateway. Azure accepts MQTT, HTTP, and AMQP, where AWS and GCP only accept MQTT, and HTTP. In terms

IoT Related Services					
Cloud Platform	Gateway (Bi-Directional)	Gateway (Ingestion Only)	Message Format	Edge Computing	Device SDKs
Amazon AWS	AWS IoT Core	Kinesis Data Firehose, Kinesis Data Streams	MQTT, HTTP	AWS Greengrass	Embedded C, C++, Java, JavaScript, Python, Android, Arduino Yun, iOS
Microsoft Azure	IoT Hub	Event Hubs	MQTT, HTTP, AMQP	IoT Edge	C, Java, Python, Node.js, .NET, IOS
Google Cloud Platform	IoT Core	Cloud Pub/Sub	MQTT, HTTP	Edge TPU	Embedded C

Figure 6.1: Comparison of cloud-based IoT services between AWS, Azure, and GCP.

of device SDKs, C or embedded C SDKs are offered by all platforms. AWS and Azure offer Java, Python, and iOS SDKs. Azure is the only provider for Node.js and .NET SDKs. AWS is the only platform to offer Arduino Yun, a popular microcontroller, and Android device SDKs.

While not listed, AWS and Azure offer multiple other services tailored to IoT scenarios. *AWS freeRTOS* [3] is an open source operating system specifically for microcontrollers, allowing easier connections to *AWS IoT Core* and *AWS IoT Greengrass*. *AWS IoT 1-Click* [4] allows for simple devices to trigger AWS Lambda functions. This makes it easier to alert staff, and monitor asset tracking. *Azure Time Series Insights* [108] is a managed service that is meant to store, visualize, and query large amounts of time series data, combining several services into one. *Azure IoT Central* [107] provides a fully managed interface for managing simple IoT projects.

6.2 IoT Storage Services

Each provider offers a multitude of storage options which can serve to store IoT data. Figure 6.2 lists the different types of Object storage, SQL databases, NoSQL databases, Data Warehouses, and Data Lake options for each provider. There is not much that differentiates

Cloud Storage Services					
Cloud Platform	Object	NoSQL Database	SQL Database	Data Warehouse	Data Lake
Amazon AWS	Amazon Simple Storage Service (S3)	Amazon DynamoDB, Amazon SimpleDB, Amazon EMR (Apache Hbase)	Amazon RDS (Relational Database Service)	Amazon Redshift	AWS Lake Formation, AWS Data Lake Solutions
Microsoft Azure	Blob Storage	CosmosDB, Table, HDInsight (Apache Hbase)	SQL Database, PostgreSQL	SQL Data Warehouse	Data Lake Storage
Google Cloud Platform	Cloud Storage	Cloud BigTable, Cloud Firestore	Cloud SQL, Cloud Spanner	BigQuery	Cloud Storage, BigQuery

Figure 6.2: Comparison of cloud-based storage services between AWS, Azure, and GCP.

the three platforms as many of the services are replicated to provide similar functionality. Each version of object storage offers warm, cold, and archival storage. SQL databases are available for structured data. Object, NoSQL databases, Data Warehouses, and Data Lakes can be used for big data storage.

One category that requires further explanation is Data Lake Storage, as the services offered vary more compared to other storage categories. *Azure Data Lake Storage* has two versions, Gen1 [106] and Gen2 [74]. *ADLS Gen1* is an Apache Hadoop file system, running on an *HDInsight* cluster, for big data storage. *ADLS Gen2* combines several features from Gen1 but uses *Azure Blob storage* instead of an *HDInsight* cluster, reducing overall costs while still offering good performance.

AWS Lake Formation [5] is a service that can help set up a data lake using *Amazon S3* for storage. AWS also offers many guides and solutions for how to create a data lake, where ingestion, storage, and analytics are all integrated into one unit and the design can be customized to the problem [7]. One example uses a provided template file which is read by *AWS CloudFormation* service to create all services used by the data lake [6]. The process creates a solution using *AWS Lambda*, *Amazon Elasticsearch*, *Amazon Cognito*, *AWS Glue*, *Amazon Athena*, *Amazon S3* and *Amazon DynamoDB*. An odd thing about AWS naming is if a service starts with “Amazon” it means it is a stand alone service, where if it starts with “AWS” it is a utility service that typically uses other services.

GCP provides a design architecture for implementing data lake storage using *Cloud Storage* service [31]. Several techniques are discussed for ingestion and analysis, but from our research no services are provided to automatically create a data lake such as AWS and Azure. Another possibility is to use *BigQuery*, which will act as both *Azure Data Lake Storage* and *Data Lake Analytics*, reducing the amount of services to monitor [32].

6.3 Streaming, Analytics, and Visualization Services

The remaining services do not fit in the IoT specific or storage related categories from the previous sections, which is why they are compared in this final section. Real time stream processing, analytics, serverless computing, machine learning, and visualization services are offered by all three platforms as seen in Figure 6.3. Similar to gateways, real time streaming will vary in terms of throughput limits per platform. All three provide their own unique solutions for analytics, such as *AWS IoT Analytics*, *Azure Data Lake Analytics*, and *BigQuery*. *Amazon EMR*, *Azure HDInsight*, and *GCP Cloud Dataproc* repackages the Apache Framework, such as Spark, Hive, HBase, and Kafka, to more easily integrate with many of their other services such as gateways, storage, and visualization for more analytical options.

Streaming / Analytics / Visualization					
Cloud Platform	Real Time Stream Processing	Analytics	Serverless Compute	Machine Learning Services	Visualization
Amazon AWS	Amazon Kinesis Data Analytics	AWS IoT Analytics, Amazon Athena, Amazon EMR(Apache Framework Tools)	AWS Lambda	SageMaker	Amazon Quicksight
Microsoft Azure	Stream Analytics	Data Lake Analytics, HDInsight (Apache Framework Tools)	Functions	Machine Learning Services	Power BI Desktop, Power BI Service
Google Cloud Platform	Cloud Dataflow	BigQuery, Cloud Dataproc (Apache Framework Tools)	Cloud Functions	Cloud Machine Learning Engine	Cloud Datalab, Data Studio

Figure 6.3: Comparison of cloud-based streaming, analytics, and visualization services between AWS, Azure, and GCP.

McGrath et al [44] performed two tests on all three serverless functions, *AWS Lambda*, *Azure Functions*, and *GCP Cloud Functions*. The first test focused on concurrency and scaling, with *AWS Lambda* almost scaling linearly and having the highest throughput. *GCP Cloud Functions* performed the worst, with performance appearing to decrease after a certain threshold. *Azure Functions* had good performance but does not exhibit obvious scaling with concurrency. The second test focused on cold start times and expiration behaviors for each function. *AWS Lambda* and *GCP Cloud Functions* had similar performance with minimal latency, and *Azure Functions* performed the worst.

Machine learning has become another hot topic in computer science, and each platform offers different yet similar services. *AWS SageMaker*, *Azure Machine Learning*, and *GCP Cloud Machine Learning Engine* offer multiple options and algorithms for creating a trained model to integrate into their IoT designs. Similar to machine learning and many of the categories before it, visualization services are offered for each platform. Microsoft's *Power BI* may have the upper hand with respect to the other services, as *Power BI* integrates with many of the programs from the Microsoft's Office Suite that many people are familiar with already.

7 Conclusion

Developing an IoT application on a cloud platform requires thoroughly understanding the programming API, configuration, performance, limitation, and design patterns of the tools of the platform. In this thesis, we reviewed the different aspects required to create a cloud-based IoT system, such as device clients, gateways, IoT communication protocols, storage, analysis, and visualization using many of the tools from Azure as examples. Following the requirements and design patterns for a cloud-based IoT system, we provided our experiences using Microsoft Azure to create a custom IoT solution. Our insights and experiences can be used by many as an introduction to the various services Azure provides, or as a starting point to implement their IoT solution.

New services for IoT platforms are emerging every year, with performance and API of existing tools rapidly improving as well. While cloud-based IoT platforms are already scalable, extensible, and comprehensive development environments, it still has many limitations. One limitation discussed focused on ingestion and storage of extremely high velocity and high volume data, where current gateways cannot handle the high traffic without adding great costs to the user. Two solutions were proposed, one implementing a distributed IoT system running on cloud virtual machines, and another custom solution combining existing cloud services, on-premise edge devices, and a unique indexing structure used for querying. The distributed model reduces latency, but costs more. The custom solution is less expensive, but has higher latency. There are many possible solutions that can be implemented, but there will always be a trade-off between latency and costs which need to be factored into design choices.

This thesis also provides a comparison of the current IoT related services from the top

three cloud platforms, AWS, Azure, and GCP. All provide similar services for many of the requirements of implementing an IoT system. For example, each platform offers a bi-directional gateway with similar functionality that varies in terms of throughput and costs. This is a trend in most categories, where services provided are similar with minor nuances. Some platforms offer unique solutions to different problems, such as *Azure Time Series Insights* and *AWS IoT 1-Click*. Each platform also provides unique services tailored to their platform, such as *AWS IoT Analytics*, *Data Lake Analytics*, and *BigQuery*.

There are several areas in this thesis that can benefit from further work and research. In Chapter 4, the IoT system created used only Microsoft Azure. It would be beneficial to replicate the same IoT system in AWS and GCP, explaining the different steps and comparing the overall process done to implement the system. Even though many of the services are similar, implementing and maintaining each service may vary greatly between platforms. If each platform is running the same IoT system using platform specific services, the performance of each IoT system can also be compared to create more distinctions between the providers. Another issue is the IoT system created in Azure was not designed to put a lot of stress on Azure. A more robust system could be designed where performance can be measured across all services used, comparing the results to the performance claims of Azure. This can also be done on AWS and GCP, not only comparing the results against each cloud providers claims, but against each other as well.

Two solutions were proposed in Chapter 5 for handling extremely high velocity and high volume data, but neither were implemented. Detailing the experiences of creating each solution can offer more insights into the many services offered by a platform. Analysis and performance metrics can be compared to show the feasibility, strengths, and weaknesses of each design. Costs can also be compared revealing potential savings and trade-offs with performance.

Lastly, further analysis can be gained when comparing the many different services between the cloud providers in Chapter 6. Costs were not discussed at all, even though this could be the single determining factor between each provider. One possible idea is to create many different IoT scenarios, each having varying ingestion, storage, analytical, and visualization requirements. Using the pricing calculators for each platform, overall costs can be estimated and compared. More information can also be provided between the many services offered, but navigating and reading through each platform's online documentation can be very cumbersome and confusing. Key details are often not discussed or buried in other pages, links to other pages can lead you in circles, and examples are limited or outdated. Beyond these shortcomings, this thesis provides a wide depth of knowledge and information with respect to cloud-based IoT systems that all skill levels can use to gain insights and better understanding of the evolving world of IoT.

Bibliography

- [1] D. Alahakoon and X. Yu. Smart electricity meter data intelligence for future energy systems: A survey. *IEEE Transactions on Industrial Informatics*, 12(1):425–436, Feb 2016.
- [2] Amazon. *Amazon MSK pricing*, 2019 (accessed November 1, 2019). <https://aws.amazon.com/msk/pricing/>.
- [3] Amazon. *Amazon FreeRTOS*, 2019 (accessed November 12, 2019). <https://aws.amazon.com/freertos/>.
- [4] Amazon. *AWS IoT 1-Click*, 2019 (accessed November 12, 2019). <https://aws.amazon.com/iot-1-click/>.
- [5] Amazon. *AWS Lake Formation*, 2019 (accessed November 12, 2019). <https://aws.amazon.com/lake-formation/>.
- [6] Amazon. *Data Lake on AWS*, 2019 (accessed November 12, 2019). <https://aws.amazon.com/solutions/data-lake-solution/>.
- [7] Amazon. *Data Lakes on AWS*, 2019 (accessed November 12, 2019). <https://aws.amazon.com/big-data/datalakes-and-analytics/featured-partners/>.

- [8] S. Athmaja, M. Hanumanthappa, and V. Kavitha. A survey of machine learning algorithms for big data analytics. In *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pages 1–4, March 2017.
- [9] M. Bacco, A. Berton, E. Ferro, C. Gennaro, A. Gotta, S. Matteoli, F. Paonessa, M. Ruggeri, G. Virone, and A. Zanella. Smart farming: Opportunities, challenges and technology enablers. In *2018 IoT Vertical and Topical Summit on Agriculture - Tuscany (IOT Tuscany)*, pages 1–6, May 2018.
- [10] J. Backman, J. Vre, K. Frmling, M. Madhikermi, and O. Nyknen. Iot-based interoperability framework for asset and fleet management. In *2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA)*, pages 1–4, Sep. 2016.
- [11] Dr. Farhad Balali, Dr. Jessie Nouri, Dr. Adel Nasiri, and Dr. Tian Zhao. *Data Intensive Industrial Asset Management*. Springer, In Press.
- [12] Anup Bhande. *What is underfitting and overfitting in machine learning and how to deal with it.*, 2018 (accessed June 15, 2019). <https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76>.
- [13] Christian Böhm, Stefan Berchtold, and Daniel A. Keim. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. *ACM Comput. Surv.*, 33(3):322–373, September 2001.
- [14] Min Chen, Shiwen Mao, and Yunhao Liu. Big data: A survey. *Mob. Netw. Appl.*, 19(2):171–209, April 2014.

- [15] Diego Clerissi, Maurizio Leotta, Gianna Reggio, and Filippo Ricca. Towards an approach for developing and testing node-red iot systems. In *Proceedings of the 1st ACM SIGSOFT International Workshop on Ensemble-Based Software Engineering, EnSEmble 2018*, pages 1–8, New York, NY, USA, 2018. ACM.
- [16] J. A. Colmenares, R. Dorriv, and D. G. Waddington. A single-node datastore for high-velocity multidimensional sensor data. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 445–452, Dec 2017.
- [17] Ali Davoudian, Liu Chen, and Mengchi Liu. A survey on nosql stores. *ACM Comput. Surv.*, 51(2):40:1–40:43, April 2018.
- [18] L. L. Dhirani, T. Newe, and S. Nizamani. Can iot escape cloud qos and cost pitfalls. In *2018 12th International Conference on Sensing Technology (ICST)*, pages 65–70, Dec 2018.
- [19] H. U. Dike, Y. Zhou, K. K. Deverasetty, and Q. Wu. Unsupervised learning based on artificial neural network: A review. In *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, pages 322–327, Oct 2018.
- [20] A. Elomari, A. Maizate, and L. Hassouni. Data storage in big data context: A survey. In *2016 Third International Conference on Systems of Collaboration (SysCo)*, pages 1–4, Nov 2016.
- [21] Raul Castro Fernandez, Matthias Weidlich, Peter Pietzuch, and Avigdor Gal. Scalable stateful stream processing for smart grids. In *Proceedings of the 8th ACM International Conference on Distributed Event-Based Systems, DEBS '14*, pages 276–281, New York, NY, USA, 2014. ACM.

- [22] Apache Software Foundation. *Powered by Apache Druid (incubating)*, 2019 (accessed October 18, 2019). <https://druid.apache.org/druid-powered.html>.
- [23] Apache Software Foundation. *Apache Druid (incubating) is a high performance real-time analytics database.*, 2019 (accessed October 20, 2019). <https://druid.apache.org/>.
- [24] Apache Software Foundation. *Apache Kafka: A Distributed Streaming Platform*, 2019 (accessed October 20, 2019). <https://kafka.apache.org/>.
- [25] Apache Software Foundation. *HDFS Architecture*, 2019 (accessed October 20, 2019). <https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html>.
- [26] Apache Software Foundation. *Welcome to Apache Zookeeper*, 2019 (accessed October 20, 2019). <https://zookeeper.apache.org/index.html>.
- [27] Eclipse Foundation. *IoT Developer Survey 2019 Results*, 2019 (accessed November 10, 2019). <https://iot.eclipse.org/resources/iot-developer-survey/iot-developer-survey-2019.pdf>.
- [28] P. Garca Lpez, M. Snchez-Artigas, G. Pars, D. Barcelona Pons, . Ruiz Ollobarren, and D. Arroyo Pinto. Comparison of faas orchestration systems. In *2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion)*, pages 148–153, Dec 2018.
- [29] N. K. Giang, R. Lea, M. Blackstock, and V. C. M. Leung. Fog at the edge: Experiences building an edge computing platform. In *2018 IEEE International Conference on Edge Computing (EDGE)*, pages 9–16, July 2018.

- [30] J. Gibson, R. Rondeau, D. Eveleigh, and Q. Tan. Benefits and challenges of three cloud computing service models. In *2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN)*, pages 198–205, Nov 2012.
- [31] Google. *Cloud Storage as a data lake*, 2019 (accessed November 12, 2019). <https://cloud.google.com/solutions/build-a-data-lake-on-gcp>.
- [32] Google. *Google Cloud Platform for Azure professionals*, 2019 (accessed November 12, 2019). <https://cloud.google.com/docs/compare/azure/>.
- [33] Jasmin Guth, Uwe Breitenbücher, Michael Falkenthal, Frank Leymann, and Lukas Reinfurt. Comparison of iot platform architectures: A field study based on a reference architecture. In *2016 Cloudification of the Internet of Things (CIoT)*, pages 1–6. IEEE, Nov 2016.
- [34] C. Harvey and A. Patrizio. *AWS vs. Azure vs. Google: Cloud Comparison [2019 Update]*, 2019 (accessed November 12, 2019). <https://www.datamation.com/cloud-computing/aws-vs-azure-vs-google-cloud-comparison.html>.
- [35] W. He, G. Yan, and L. D. Xu. Developing vehicular data cloud services in the iot environment. *IEEE Transactions on Industrial Informatics*, 10(2):1587–1595, May 2014.
- [36] H. Hejazi, H. Rajab, T. Cinkler, and L. Lengyel. Survey of platforms for massive iot. In *2018 IEEE International Conference on Future IoT Technologies (Future IoT)*, pages 1–8, Jan 2018.
- [37] InfluxData. *InfluxDB*, 2019 (accessed October 29, 2019). <https://www.influxdata.com/products/influxdb-overview/>.

- [38] A. A. Ismail, H. S. Hamza, and A. M. Kotb. Performance evaluation of open source iot platforms. In *2018 IEEE Global Conference on Internet of Things (GCIoT)*, pages 1–5, Dec 2018.
- [39] S. Jain, , A. Paventhan, V. Kumar Chinnaiyan, and V. Arnachalam and. Survey on smart grid technologies- smart metering, iot and ems. In *2014 IEEE Students' Conference on Electrical, Electronics and Computer Science*, pages 1–6, March 2014.
- [40] A Rastogi K. Goyal, A. Garg and S Singhal. A literature survey on internet of things (iot). *Int. J Advanced Networking and Applications*, 9:3663–3668, 2018.
- [41] Nawsher Khan, Arshi Naim, Mohammad Rashid Hussain, Quadri Noorulhasan Naveed, Naim Ahmad, and Shamimul Qamar. The 51 v's of big data: Survey, technologies, characteristics, opportunities, issues and challenges. In *Proceedings of the International Conference on Omni-Layer Intelligent Systems, COINS '19*, pages 19–24, New York, NY, USA, 2019. ACM.
- [42] Jay Kreps. *Benchmarking Apache Kafka: 2 Million Writes Per Second (On Three Cheap Machines)*, 2014 (accessed October 20, 2019). <https://engineering.linkedin.com/kafka/benchmarking-apache-kafka-2-million-writes-second-three-cheap-machines>.
- [43] Jay Lee, Hung-An Kao, and Shanhu Yang. Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia CIRP*, 16:3 – 8, 2014. Product Services Systems and Value Creation. Proceedings of the 6th CIRP Conference on Industrial Product-Service Systems.

- [44] G. McGrath and P. R. Brenner. Serverless computing: Design, implementation, and performance. In *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pages 405–410, June 2017.
- [45] Microsoft. *The Jupyter Notebook Introduction*, 2015 (accessed June 15, 2019). <https://jupyter-notebook.readthedocs.io/en/stable/notebook.html>.
- [46] Microsoft. *Stream Analytics Query Language Reference*, 2016 (accessed April 10, 2019). <https://docs.microsoft.com/en-us/stream-analytics-query/stream-analytics-query-language-reference>.
- [47] Microsoft. *Azure Functions*, 2017 (accessed April 10, 2019). <https://docs.microsoft.com/en-us/azure/azure-functions>.
- [48] Microsoft. *Azure IoT protocol gateway*, 2017 (accessed April 10, 2019). <https://docs.microsoft.com/en-us/azure/iot-hub/iot-hub-protocol-gateway>.
- [49] Microsoft. *Azure Stream analytics – User Defined Aggregates*, 2017 (accessed April 10, 2019). <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-javascript-user-defined-aggregates>.
- [50] Microsoft. *Processing 100,000 events per second on Azure Functions*, 2017 (accessed April 10, 2019). <https://azure.microsoft.com/en-us/blog/processing-100-000-events-per-second-on-azure-functions>.
- [51] Microsoft. *Azure Time Series Insights Documentation*, 2017 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/time-series-insights>.
- [52] Microsoft. *Azure Machine Learning Studio Web Services: Deployment and consumption*, 2017 (accessed June 15, 2019). <https://docs.microsoft.com/en-us/azure/machine-learning/studio/deploy-consume-web-service-guide>.

- [53] Microsoft. *Azure Functions C# script (.csx) developer reference*, 2017 (accessed May 12, 2019). <https://docs.microsoft.com/en-us/azure/azure-functions/functions-reference-csharp>.
- [54] Microsoft. *How to get started with Azure Table storage and Visual Studio connected services*, 2017 (accessed May 12, 2019). <https://docs.microsoft.com/en-us/azure/visual-studio/vs-storage-aspnet5-getting-started-tables>.
- [55] Microsoft. *Azure IoT SDK*, 2018 (accessed April 10, 2019). <https://docs.microsoft.com/en-us/azure/iot-hub/iot-hub-devguide-sdks>.
- [56] Microsoft. *Azure Stream analytics – User Defined Functions*, 2018 (accessed April 10, 2019). <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-javascript-user-defined-functions>.
- [57] Microsoft. *Durable Azure Functions*, 2018 (accessed April 10, 2019). <https://docs.microsoft.com/en-us/dotnet/standard/serverless-architecture/durable-azure-functions>.
- [58] Microsoft. *Azure Event Hubs A big data streaming platform and event ingestion service?*, 2018 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-about>.
- [59] Microsoft. *Real-time streaming in Power BI*, 2018 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/power-bi/service-real-time-streaming>.
- [60] Microsoft. *Security for Internet of Things (IoT) from the ground up*, 2018 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/iot-fundamentals/iot-security-ground-up?context=azure/iot-hub/rc/rc>.

- [61] Microsoft. *What is Azure Blob storage?*, 2018 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blobs-overview>.
- [62] Microsoft. *What is Azure IoT Hub?*, 2018 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/iot-hub/about-iot-hub>.
- [63] Microsoft. *Comparing Power BI Desktop and the Power BI service*, 2018 (accessed June 16, 2019). <https://docs.microsoft.com/en-us/power-bi/service-service-vs-desktop>.
- [64] Microsoft. *Choose the right IoT Hub tier for your solution*, 2018 (accessed May 10, 2019). <https://docs.microsoft.com/en-us/azure/iot-hub/iot-hub-scaling#basic-and-standard-tiers>.
- [65] Microsoft. *What is Azure Service Bus?*, 2018 (accessed May 10, 2019). <https://docs.microsoft.com/en-us/azure/service-bus-messaging/service-bus-messaging-overview>.
- [66] Microsoft. *Azure Table storage bindings for Azure Functions*, 2018 (accessed May 12, 2019). <https://docs.microsoft.com/en-us/azure/azure-functions/functions-bindings-storage-table>.
- [67] Microsoft. *Supported languages in Azure Functions*, 2018 (accessed May 12, 2019). <https://docs.microsoft.com/en-us/azure/azure-functions/supported-languages>.
- [68] Microsoft. *Test a Stream Analytics query with sample data*, 2018 (accessed May 12, 2019). <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-test-query>.
- [69] Microsoft. *Apache Spark in Azure HDInsight*, 2019 (accessed April 10, 2019). <https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-overview>.

- [70] Microsoft. *Run Azure Functions from Azure Stream Analytics jobs*, 2019 (accessed April 10, 2019). <https://github.com/MicrosoftDocs/azure-docs/blob/master/articles/stream-analytics/stream-analytics-with-azure-functions.md>.
- [71] Microsoft. *Azure Blob storage: hot, cool, and archive access tiers*, 2019 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-storage-tiers>.
- [72] Microsoft. *Azure HDInsight Documentation*, 2019 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/hdinsight/>.
- [73] Microsoft. *Azure Notification Hubs*, 2019 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/notification-hubs>.
- [74] Microsoft. *Introduction to Azure Data Lake Storage Gen2*, 2019 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction>.
- [75] Microsoft. *Introduction to Stream Analytics windowing functions*, 2019 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>.
- [76] Microsoft. *Provisioning devices with Azure IoT Hub Device Provisioning Service*, 2019 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/iot-dps/about-iot-dps#when-to-use-device-provisioning-service>.
- [77] Microsoft. *Stream Analytics Documentation*, 2019 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/stream-analytics/>.

- [78] Microsoft. *Welcome to Azure Cosmos DB*, 2019 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/cosmos-db/introduction>.
- [79] Microsoft. *What is Azure SQL Database service*, 2019 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/azure/sql-database/sql-database-technical-overview>.
- [80] Microsoft. *What is Power BI Desktop?*, 2019 (accessed April 11, 2019). <https://docs.microsoft.com/en-us/power-bi/desktop-what-is-desktop>.
- [81] Microsoft. *Supplemental Terms of Use for Microsoft Azure Previews*, 2019 (accessed June 14, 2019). <https://azure.microsoft.com/en-us/support/legal/preview-supplemental-terms/>.
- [82] Microsoft. *Cross-Validate Model*, 2019 (accessed June 15, 2019). <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/cross-validate-model>.
- [83] Microsoft. *Import Data*, 2019 (accessed June 15, 2019). <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/import-data>.
- [84] Microsoft. *Machine Learning module descriptions*, 2019 (accessed June 15, 2019). <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-module-descriptions>.
- [85] Microsoft. *Overview of Azure Notebooks*, 2019 (accessed June 15, 2019). <https://docs.microsoft.com/en-us/azure/notebooks/azure-notebooks-overview>.
- [86] Microsoft. *Quickstart: Create your first data science experiment in Azure Machine Learning Studio*, 2019 (accessed June 15, 2019). <https://docs.microsoft.com/en-us/azure/machine-learning/studio/create-experiment>.

- [87] Microsoft. *Tutorial 3: Deploy credit risk model - Azure Machine Learning Studio*, 2019 (accessed June 15, 2019). <https://docs.microsoft.com/en-us/azure/machine-learning/studio/tutorial-part3-credit-risk-deploy>.
- [88] Microsoft. *What are the machine learning products at Microsoft?*, 2019 (accessed June 15, 2019). <https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/data-science-and-machine-learning?context=azure/machine-learning/studio/context/ml-context>.
- [89] Microsoft. *What is automated machine learning?*, 2019 (accessed June 15, 2019). <https://docs.microsoft.com/en-us/azure/machine-learning/service/concept-automated-ml>.
- [90] Microsoft. *What is Azure Machine Learning service?*, 2019 (accessed June 15, 2019). <https://docs.microsoft.com/en-us/azure/machine-learning/service/overview-what-is-azure-ml>.
- [91] Microsoft. *What is Azure Machine Learning Studio?*, 2019 (accessed June 15, 2019). <https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-ml-studio>.
- [92] Microsoft. *What is the Azure Machine Learning SDK for Python?*, 2019 (accessed June 15, 2019). <https://docs.microsoft.com/en-us/python/api/overview/azure/ml/intro?view=azure-ml-py>.
- [93] Microsoft. *What is the visual interface for Azure Machine Learning service?*, 2019 (accessed June 15, 2019). <https://docs.microsoft.com/en-us/azure/machine-learning/service/ui-concept-visual-interface>.
- [94] Microsoft. *Basic concepts for designers in the Power BI service*, 2019 (accessed June 16, 2019). <https://docs.microsoft.com/en-us/power-bi/service-basic-concepts>.

- [95] Microsoft. *Embedded analytics with Power BI*, 2019 (accessed June 16, 2019).
<https://docs.microsoft.com/en-us/power-bi/developer/embedding>.
- [96] Microsoft. *Query overview in Power BI Desktop*, 2019 (accessed June 16, 2019).
<https://docs.microsoft.com/en-us/power-bi/desktop-query-overview>.
- [97] Microsoft. *What is Power BI?*, 2019 (accessed June 16, 2019).
<https://docs.microsoft.com/en-us/power-bi/power-bi-overview>.
- [98] Microsoft. *Azure storage account overview*, 2019 (accessed May 10, 2019).
<https://docs.microsoft.com/en-us/azure/storage/common/storage-account-overview>.
- [99] Microsoft. *Reference - IoT Hub endpoints*, 2019 (accessed May 10, 2019).
<https://docs.microsoft.com/en-us/azure/iot-hub/iot-hub-devguide-endpoints>.
- [100] Microsoft. *Understand and use Azure IoT Hub SDKs*, 2019 (accessed May 10, 2019).
<https://docs.microsoft.com/en-us/azure/iot-hub/iot-hub-devguide-sdks>.
- [101] Microsoft. *What is Azure Stream Analytics*, 2019 (accessed May 10, 2019).
<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction>.
- [102] Microsoft. *Azure Functions triggers and bindings concepts*, 2019 (accessed May 12, 2019).
<https://docs.microsoft.com/en-us/azure/azure-functions/functions-triggers-bindings>.
- [103] Microsoft. *Azure Table storage overview*, 2019 (accessed May 12, 2019).
<https://docs.microsoft.com/en-us/azure/cosmos-db/table-storage-overview>.

- [104] Microsoft. *Understand inputs for Azure Stream Analytics*, 2019 (accessed May 12, 2019). <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-add-inputs>.
- [105] Microsoft. *Understand outputs from Azure Stream Analytics*, 2019 (accessed May 12, 2019). <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-define-outputs>.
- [106] Microsoft. *What is Azure Data Lake Storage Gen1?*, 2019 (accessed Nov 12, 2019). <https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-overview>.
- [107] Microsoft. *What is Azure IoT Central?*, 2019 (accessed November 12, 2019). <https://docs.microsoft.com/en-us/azure/iot-central/core/overview-iot-central>.
- [108] Microsoft. *What is Azure Time Series Insights?*, 2019 (accessed November 12, 2019). <https://docs.microsoft.com/en-us/azure/time-series-insights/time-series-insights-overview>.
- [109] Microsoft. *High-Throughput with Azure Blob Storage*, 2019 (accessed October 6, 2019). <https://azure.microsoft.com/en-au/blog/high-throughput-with-azure-blob-storage/>.
- [110] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani. Deep learning for iot big data and streaming analytics: A survey. *IEEE Communications Surveys Tutorials*, 20(4):2923–2960, Fourthquarter 2018.
- [111] M. Moravcik, P. Segec, and M. Kontsek. Overview of cloud computing standards. In *2018 16th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pages 395–402, Nov 2018.

- [112] N. Naik. Choice of effective messaging protocols for iot systems: Mqtt, coap, amqp and http. In *2017 IEEE International Systems Engineering Symposium (ISSE)*, pages 1–7, Oct 2017.
- [113] Iuri D. Nogueira, Maram Romdhane, and Jérôme Darmont. Modeling data lake metadata with a data vault. In *Proceedings of the 22Nd International Database Engineering & Applications Symposium, IDEAS 2018*, pages 253–261, New York, NY, USA, 2018. ACM.
- [114] Oasis. *MQTT Version 3.1.1*, 2014 (accessed April 11, 2019). <http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.html>.
- [115] D. Obermaier and I. Skerrett. *HiveMQ and Apache Kafka - Streaming IoT Data and MQTT Messages*, 2019 (accessed October 29, 2019). <https://www.hivemq.com/blog/streaming-iot-data-and-mqtt-messages-to-apache-kafka/>.
- [116] OpenTSDB. *The Scalable Time Series Database*, 2019 (accessed October 29, 2019). <http://opentsdb.net/>.
- [117] B. Padmaja, P. V. Narasimha Rao, M. Madhu Bala, and E. K. Rao Patro. A novel design of autonomous cars using iot and visual features. In *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on*, pages 18–21, Aug 2018.
- [118] T. Pflanzner and A. Kertesz. A survey of iot cloud providers. In *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 730–735, May 2016.

- [119] A. Polianytsia, O. Starkova, and K. Herasymenko. Survey of the iot data transmission protocols. In *2017 4th International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S T)*, pages 369–371, Oct 2017.
- [120] P. Ray. A survey of iot cloud platforms. *Future Computing and Informatics Journal*, 1(1-2):35–46, Dec. 2016.
- [121] B. Saha and K. Goebel. “Battery data set”, *NASA AMES prognostics data repository*., 2007. <https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/>.
- [122] Y. Samadi, M. Zbakh, and C. Tadonki. Comparative study between hadoop and spark based on hibench benchmarks. In *2016 2nd International Conference on Cloud Computing Technologies and Applications (CloudTech)*, pages 267–275, May 2016.
- [123] A. K. Samanta, B. B. Sarkar, and N. Chaki. Query performance analysis of nosql and big data. In *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pages 237–241, Nov 2018.
- [124] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *CoRR*, abs/1712.01815, 2017.
- [125] S. Singh. Optimize cloud computations using edge computing. In *2017 International Conference on Big Data, IoT and Data Science (BID)*, pages 49–53, Dec 2017.
- [126] E. Siow, T. Tiropanis, and W. Hall. Analytics for the internet of things: A survey. *ACM Comput. Surv.*, 51(4):74:1–74:36, July 2018.

- [127] Daniel Sonntag, Sonja Zillner, Patrick van der Smagt, and András Lörincz. *Overview of the CPS for Smart Factories Project: Deep Learning, Knowledge Acquisition, Anomaly Detection and Intelligent User Interfaces*, pages 487–504. Springer International Publishing, Cham, 2017.
- [128] Timescale. *TimescaleDB*, 2019 (accessed October 29, 2019). <https://www.timescale.com/products>.
- [129] F. J. Valente and A. C. Neto. Intelligent steel inventory tracking with iot / rfid. In *2017 IEEE International Conference on RFID Technology Application (RFID-TA)*, pages 158–163, Sep. 2017.
- [130] R. Vargheese and H. Dahir. An iot/ioe enabled architecture framework for precision on shelf availability: Enhancing proactive shopper experience. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 21–26, Oct 2014.
- [131] S. Verma, Y. Kawamoto, Z. M. Fadlullah, H. Nishiyama, and N. Kato. A survey on network methodologies for real-time analytics of massive iot data and open research issues. *IEEE Communications Surveys Tutorials*, 19(3):1457–1477, thirdquarter 2017.
- [132] Fangjin Yang, Eric Tschetter, Xavier Léauté, Nelson Ray, Gian Merlino, and Deep Ganguli. Druid: A real-time analytical data store. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD ’14, pages 157–168, New York, NY, USA, 2014. ACM.
- [133] Amal W. Yassien and Amr F. Desouky. Rdbms, nosql, hadoop: A performance-based empirical analysis. In *Proceedings of the 2Nd Africa and Middle East Conference on Software Engineering*, AMECSE ’16, pages 52–59, New York, NY, USA, 2016. ACM.