IoT Solutions in Microsoft's Azure IoT Suite

Data Acquisition and Analysis in the Real World

—

Building for the Internet of Things

—

Scott Klein

Apress®

IoT Solutions in Microsoft's Azure IoT Suite

Data Acquisition and Analysis in the Real World

Scott Klein

Apress®

IoT Solutions in Microsoft's Azure IoT Suite: Data Acquisition and Analysis in the Real World

Scott Klein
Redmond, Washington, USA

ISBN-13 (pbk): 978-1-4842-2142-6 ISBN-13 (electronic): 978-1-4842-2143-3
DOI 10.1007/978-1-4842-2143-3

Library of Congress Control Number: 2017939347

Copyright © 2017 by Scott Klein

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director: Welmoed Spahr
Editorial Director: Todd Green
Acquisitions Editor: Jonathan Gennick
Development Editor: Laura Berendson
Technical Reviewer: Richard Conway
Coordinating Editor: Jill Balzano
Copy Editor: Mary Behr
Compositor: SPi Global
Indexer: SPi Global
Artist: SPi Global
Cover image designed by Freepik

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit http://www.apress.com/rights-permissions.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/9781484221426. For more detailed information, please visit http://www.apress.com/source-code.

Printed on acid-free paper

To the people who mean the most to me:
my wife, my children, my parents, and my family.

Contents at a Glance

Contents

About the Author

Scott Klein is a Microsoft Senior Program Manager with a passion for Microsoft's data services and technologies. He spent the four previous years traveling the globe evanglizing SQL Server and big data, and he spent the last year working with and sharing his excitement for Microsoft's IoT, Intelligence, and Analytics services, so much so that he can frequently be found burried underneath a pile of Raspberry Pis and other devices. Not forgetting his roots, he still works with SQL Server to the point that he spends most of his day in buildings 16 and 17.

You can find Scott hosting a few Channel 9 shows, including Data Exposed (https://channel9.msdn.com/shows/data-exposed), The Internet of Things Show (https://channel9.msdn.com/Shows/Internet-of-Things-Show), and SQL Unplugged (https://channel9.msdn.com/Shows/sql-unplugged). Scott was one of the four original SQL Azure MVPs, and even though they don't exist any more, he still claims it. Scott thinks the word "grok" is an awesome word, and he is still trying to figure out how to brew the perfect batch of root beer. To see what Scott is up to, follow him on Twitter at @SQLScott or on his blog at http://aka.ms/SQLScott.

About the Technical Reviewer

Richard Conway has been programming since his ZX81 days through a morass of jobs in a number of verticals and some spectacularly failing startups. He is a Microsoft Regional Director and Azure Most Valuable Professional with a penchant for all things cloud and data. He is a founder and director of Elastacloud, a cloud data science consultancy based in London and Derby, and he is a founder and organizer of the UK Azure Group and IoT and Data Science Innovators. His latest project is AzureCraft, a twice-yearly conference for children that teaches about the cloud, data, and AI using Minecraft. Follow him on Twitter at @azurecoder.

Acknowledgments

Apress gave me a page to list all of my acknowledgements. Honestly, I don't think a page will be long enough. The list of people who helped me through this book, provided feedback, added insight and direction, and probably let me bug them way too much is very long. But I'll try.

First and foremost are the awesome, and extremely patient, people at Apress. Jonathan Gennick and Jill Balzano are beyond phenomenal. And patient. Did I mention patient?

Next comes all of my co-workers, the fantastic PMs at Microsoft who build these wonderful services: Elio Damaggio, Ryan Crawcour, Matthew Hicks, Saveen Reddy, Matthew Roche, Ashish Thapliyal, Anand Subbaraj, Sharon Lo, Saurin Shah, Michael Rys, John Taubensee, Konstantin Zoryn, Arindam Chatterjee, Rajesh Dadhia, and a host of others.

Next comes the more-than-amazing technical reviewer, Richard Conway. Richard is a big data, IoT, and analytics rock star. I've known Richard a few years and I could not have been more excited to have Richard review this book. I am pretty sure I will forever be in his debt for all the questions I asked him and for the help I received from him.

Lastly, but most importantly, comes my family. Enough cannot be said about the love and support I received from them. So, please excuse me while I go reintroduce myself to them. ☺

Introduction

I'll cut right to the chase here and not be long-winded. The intent of this book is to provide some insight into how Microsoft's Internet of Things and Intelligence and Analytics services can be used together to build an end-to-end solution. This book takes one example and walks that example through the book, implementing service after service, to help stitch together the end-to-end picture. Along the way, the book will stop and look at other interesting, real-world scenarios to help clarify and broaden the picture, to give you further ideas.

This book does not go deeply into any specific service. Entire books could probably be written on each service covered in this book. That is not the intent of this book. My goal is to discuss each service enough to help you decide how you can use the particular service in an IoT solution. I want to get you excited about working with IoT, its capabilities and possibilities. This area is still in its infancy, and there is so much more that can be done to make the quality of our lives better—not to the WALL-E point because heaven forbid we ever get to that point, but to where we are using IoT to save lives, make things safer, and do great things for this planet and humankind.

I tried my best to keep the screenshots and related information up to date. However, if you work with Microsoft Azure, and probably in cloud service for that matter, you know how fast things can change. If the screenshots changed between when I wrote the chapter and it got printed and into your hands, that's the way working with cloud services is. It should be close enough for you to figure out the differences and work with them.

Almost every author will tell you that when they write a book, they look back and wish they could have added "x" or talked more about "y." Such is the case in this book as well. Not to make excuses, but I had to make decisions on many areas just to get the book out. I couldn't keep holding the book up because feature "x" was coming or improvement "y" was about to be released. Thus, I relied on feedback and input to create what is in your hands. I hope you find it good enough.

There are many things I would have liked to discuss in more detail. For example, I would have liked to discuss more about Azure Stream Analytics sliding windows and spent more time in U-SQL (which probably deserves its own book, by the way). However, to get the book out, I hopefully provided enough information to paint the end-to-end picture I was striving for.

If you want to learn more about these amazing services, especially feature "x" and improvement "y," I'll continue to add to this book via my blog, `http://aka.ms/SQLScott`. If you have ideas or want more information on a specific topic, feel free to ping me via Twitter at `@SQLScott` or my blog. I am always interested in your questions and feedback because that makes us both better.

PART I

Getting Started

CHAPTER 1

The World of Big Data and IoT

If you're reading this book, I would venture to guess it is for one of two primary reasons. First, you *want* to learn and get into the big data/Internet-of-things space and technologies. Second, you *need* to learn about dealing with big data. The first is more from the perspective of curiosity and career growth and you are to be applauded for taking the initiative and first step into the fantastic world of dealing with vast amounts of data. If it's the second, I pray for you because this is a "Holy data explosion, Batman! I need to make sense of all this data!" situation and you have some work ahead of you.

Now, I don't want to scare you into thinking that dealing with big data is scary and overwhelming, so don't hyperventilate. It can certainly appear overwhelming, especially if you look at all the technologies that have emerged over the last few years to deal the exponential growth of data and help solve the data insights problems.

But before we begin the journey into this wonderful space, let's go back in time and look at where things have come from, sort of a "history of data." It wasn't too long ago that data was typically stored in simple text format in flat files in either a comma- or tab-delimited format. Reading and writing data took a fair amount of code. If you're over 40, you remember these days not too fondly. Luckily, the relational database came to the rescue, which made dealing with data much easier. Gone were the days of parsing files and reading one line at time; instead we could write SQL queries and have the processing done more efficiently at the server. The relational database veiled the complexity of the storage layer while at the same time provided built-in relational capabilities that flat files did not.

The fact that relational databases still have value stands as a tribute to their effectiveness and capabilities even when storing and working with terabytes of data. But we need to be realistic. As good and effective as a relational database is, we need to have a clear understanding that today's needs can't be solved by a relational database easily. As the size and different types of data being generated and collected continues to rise, it becomes increasingly obvious that we are moving beyond the capabilities of even the best-of-class enterprise relational database systems.

Another thing to consider is how big data changes how we keep (i.e. store) data. On-premises storage is expensive, but with cloud storage being so inexpensive, companies can hang on to their data much longer. As cheap as cloud storage is, companies can keep every raw data point, from clickstream, weblogs, and telemetry data. Many companies today are storing tens of petabytes of data in the cloud.

Which leads us to the need to understand big data and ultimately the Internet of Things (IoT). In this chapter, I will begin by looking at the overall topic of big data, its characteristics, and how we should think about it. I will use that foundation to discuss the hot topic of the Internet of Things. I'll then wrap up the chapter by looking at several scenarios where both big data and IoT are a common place to provide a visual and foundation for the rest of this book.

© Scott Klein 2017
S. Klein, *IoT Solutions in Microsoft's Azure IoT Suite*, DOI 10.1007/978-1-4842-2143-3_1

Big Data

There is some argument on when the term "big data" was first used or coined, but if you think that the topic is relatively new, you are grossly mistaken. Back in 1941, 76 years ago, the term "information explosion" was first used in the Oxford English Dictionary. So even way back then we were trying to understand the phenomenon of the increasing volume of data. Regardless of when or who came up with the term, the important point is that we now live in a time where data is growing at an exponential rate and we need to understand what it is and how to gain insights into its depths, which is what this section will attempt to do. What follows will provide the foundation of what big data is and what it is not, and how to think about it in terms of your business.

What Is Big Data?

If you stop to think about how application data has evolved over time, you can see what is happening out in the enterprise ecosystem, which ultimately makes us stop and think about how the process of storing and processing data and information has evolved. Figure 1-1 illustrates the progress of both the volume and variety of data into what we know today as the big data era. The X axis represents the complexity and variety of data, and the Y axis the volume of data.

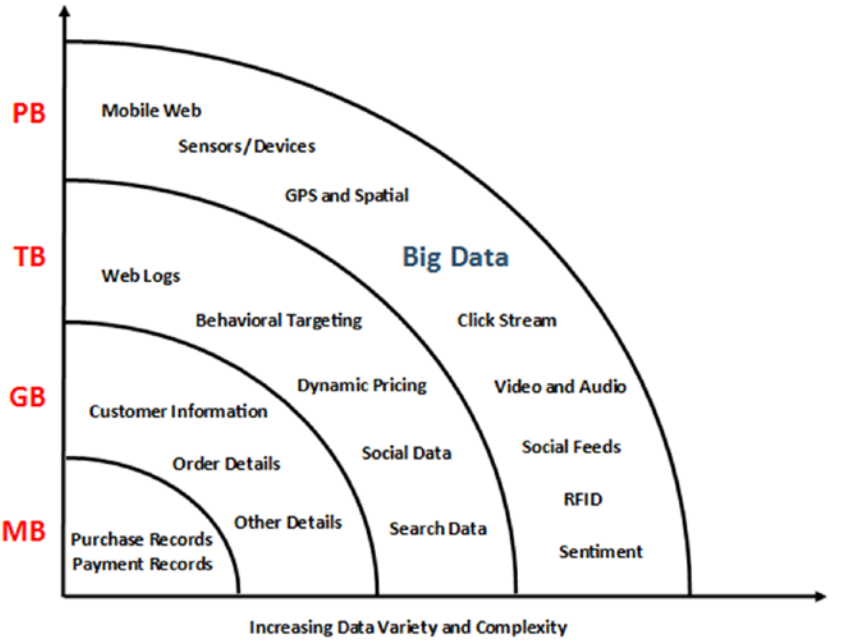

Figure 1-1. *Evolution of application data*

Many of us remember the days when dealing with data meant working in spreadsheets or smaller databases like Microsoft Access or FoxPro. In this era, you dealt with a defined amount and type of information, such as storing types of transactions such as purchases or orders. These scenarios had limited reporting and information insight into the business.

From there, as architectures grew and changed, the need for more advanced data management systems became necessary along with the need to do some level of advanced analytics. For example, a popular scenario was the need to start studying customers and what they were buying. The analytics allowed, for example, businesses to provide purchase recommendations and other opportunities to drive business opportunities.

This scenario, along with more detailed data insight requirements, also brought around more powerful database that added more analytics processing, dealing with cubes, dimensions, and fact tables for doing OLAP analysis. Modern databases provided greater power and additional proficiencies for accessing information through query syntaxes such as SQL. This also abstracted the complexity of the underlying storage mechanism, making it easy to work with relational data and the volume of data it supported.

But there quickly came a point when we needed more than what current enterprise-level relational database systems provided. Exponential data growth was a reality and we needed deeper and distinct analysis on historical, current, and future data. This was critical to business success. The need to look and analyze each individual record was right in front of us, but the infrastructure and current data management systems didn't and couldn't scale. It was difficult to take current systems and infrastructure and apply them to the deeper analytics of analyzing individual transactions.

This isn't to say that the current architecture and systems are going away. There is and will be a need for powerful relational database and business intelligence systems like SQL Server and SQL Server Analysis Services. But as the amount data grew and the need to get deeper insights into that data was needed, it was hard to take the same processing, architecture, and analysis model and scale it to meet the requirements of storing and processing very granular transactions.

The strain on the data architecture along with the different types of data has challenged us to come up with more efficient and scalable forms of storing and processing information. Our data wasn't just relational anymore. It was coming in the forms of social media, sensor and device information, weblogs, and more. About this time we saw the term "No-SQL" bounced around. Many thought that this meant that relational was dead. But in reality the term means "not only SQL," meaning that data comes in all forms and types: relational and non-, or semi-, relational.

Thus, in essence, we can think about big data as representing the vast amounts and different types of data that organizations continue to work with, and try to gain insights into, on a daily basis. One attribute of big data is that it can be generated at a very rapid rate, sometimes called a "fire hose" rate. The data is valuable but previously not practical to store or analyze due to the cost of appropriate methods.

Additionally, when we talk about solutions to big data, it's just not about the data itself; it also includes the systems and processes to uncover and gain insights hidden inside all that wonderful, sweet data. Today's big data solutions are comprised of a set of technologies and processes that allow you to efficiently store and analyze the data to gain the desired data insights.

The Three Vs of Big Data

Companies begin to look at big data solutions when their current traditional database systems reach the limit of performance, scale, and cost. These big data solutions, in addition to helping overcome database system limitations, also provide a way to more effectively provide the much-needed avenue for gaining the greatly required insight into the data to further examine and mine data in a way that isn't possible in database systems. As such, big data is typically described and defined in terms of the three Vs to understand big data and solve big data issues: volume, variety, and velocity.

Volume

The volume of data has reference to the scale of data. Where relational database systems work in the high terabyte range, big data solutions store and process hundreds or thousands of terabytes, petabytes, and even exabytes, with the total volume growing exponentially. A big data solution must have the storage to handle and manage this volume of data and be designed to scale and work efficiently across multiple machines in a distributed environment. Organizations today need to handle mass quantities of data each day.

Variety

Data today comes in many different forms, both structured and semi-structured, relational and non-relational. Gone are the days where the majority of data resides in a relational database. Today, a very large amount of data being generated comes from sensors, devices, social media, and other formats that aren't conducive to a structured or relational format. The key to keep in mind here is that it is just not all about relational data anymore. Today, organizations are dealing with relational (structured) and non-relational (semi-structured or non-structured) data. Typically, the majority of data currently stored in big data solutions is unstructured or semi-structured.

A key factor when dealing with a variety of data is deciding how and where to store the variety of data. Using a traditional relational database system can be challenging and may no longer be a practical solution for storing semi-structured or non-structured due to a lack of a schema application. Big data solutions typically target scenarios where there is a huge volume of unstructured or semi-structured data that must be stored and queried to extract business intelligence.

Velocity

The velocity of data has dual meanings. The obvious meaning applies to how fast data is being generated and gathered. Today, data is being produced and collected at an ever-increasing rate from a wide range of sources, including devices and sensors as well as applications such as social media.

The second meaning of velocity applies to the analysis of the data being streamed in. Businesses and organizations must decide how fast they need to understand the data as it is being streamed in.

To help put these Vs in perspective, here are some examples:

- Microsoft Bing ingests over 7 petabytes of data a month.
- The Twitter community generates over 1 terabyte of tweet data every single day.
- Five years, ago, it was predicted that 7.9 zettabytes of data would make up the digital universe.
- 72 hours of video are uploaded per minute on YouTube. That's 1 terabyte every 4 minutes.
- 500 terabytes of new data are ingested in Facebook databases.
- Sensors from a Boeing jet engine create 20 terabytes of data every hour.
- The proposed Square Kilometer Array telescope will generate "a few exabytes of data per day" per single beam.

I'll give you another example that will help put the data explosion into context. In 1969, the United States wanted to put a man on the moon. Keep in mind this was 48 years ago and it had to work the first time, meaning that we wanted to send a man into space and we also wanted to get him back. So, NASA built Apollo X1 with a weight of almost 30,000 pounds (13,500 kg) and a top speed of almost 2,200 mile per hour (3,500 km/hour). The distance to the moon is about 221,208 miles (356,000 km). NASA got the brightest minds together and when it was all said and done, it took 64Kb of RAM with the code written in Fortran to get Neil Alden Armstrong to the moon and back.

Today, 47 years later, the health messages alone from the game Halo generate gigabytes of data per **second**. PER SECOND! A GAME! Yet it only took 64Kb of RAM to send man to the moon. Mind blowing.

Are There Additional Vs?

While it is universally agreed upon that volume, variety, and velocity make up the three main Vs of big data, some have added additional Vs to the big data problem. There is some disagreement, however, as to what they are. Some list value as the fourth V, while others list veracity as the fourth V. Some still list all way out to seven Vs. Personally, I think that if you're all the way out to seven Vs, you're just looking for words that start with the letter V. Yet, I do find worth in *veracity* and *value* so I will briefly discuss them here.

- **Veracity**: This term has reference to the *uncertainty* of your data. Essentially, how accurate is your data? It's worthless if it's not accurate. You can't make trusted and accurate decisions if you don't trust your data, and your programs are only as good as the data they are working with.
- **Value**: Big data is data that has value, and having access to this data does no good unless you can turn it into something of meaning.

Other Vs have been mentioned, such as variability. However, what's important to you and me is to understand how these Vs help provide insight into the scale and importance of data, including the challenges of dealing with big data.

Why You Should Care About Big Data

Organizations today that deal with the Vs discussed above look to big data solutions to resolve the limitations of traditional database systems. Today's requirements for storing data are outpacing those of a relational data store, and businesses today may not survive into tomorrow without the enabling power and flexibility of big data solutions.

The data warehouses of yesterday are being outgunned by the big data solutions of today and tomorrow simply because the volume of data surpasses the cost and capacity found in relational database systems. That doesn't mean that data warehouses are passé and not needed, but in fact quite the opposite. Organizations start caring about big data solutions as a new way to gain rapid and valuable insights into their growing data when the volume, velocity, and variety of data exceeds the cost and capabilities of a data warehouse storage solution. One does not replace the other; they are complimentary.

The time to start caring about big data is when the Vs of big data are a reality within your organization: when your data is no longer just relational; when you have the need to store and process up to petabytes of data in a cost-effective way; when you have the need to find valuable insights into that ever-increasing volume of data quickly and efficiently.

Big Data Solutions vs. Traditional Databases

As the landscape grows and progresses from traditional database systems to big data solutions, it is helpful to understand how traditional database systems and solutions differ from big data solutions. Table 1-1 summarizes the major differentiators and provides a high-level comparison.

Table 1-1. *Comparing Relational Database Systems to Big Data Solutions*

	RDBMS	Big Data Solutions
Data Size	Gigabytes (terabytes)	Petabytes (hexabytes)
Data Types	Structured	Semi-structured or unstructured
Access	Interactive and batch	Batch
Update pattern	Read/write many times	Write once, read many times
Structure	Static schema	Dynamic schema
Integrity	High (ACID)	Low
Scaling	Nonlinear	Linear

Traditional database systems use a relational design where data is stored and based on predetermined schema, and are designed to handle workloads up to terabytes in size. These systems are meant for OLTP (Online Transaction Processing) transactions, meaning that applications are reading, writing, and updating data within the database at a high frequency, usually using small transactions that affect a few rows at a time.

Scaling a relational database to gain performance needs typically means a nonlinear scalability model, meaning that you are either adding more memory or CPU (or both) to the node where all the transactions take place, and/or applying disk partitioning and filegroups that divide the data into multiple logical chunks, with those chunks residing on the same physical node or storage system, or different physical nodes and storage systems resulting in network latency.

In contrast, big data solutions allow you to store any type and format of data including non-structured and semi-structured data while not requiring nor operating over a predetermined schema. Together, these two features allow data to be stored in its native, raw format and apply a schema only when data is read.

While both traditional databases and big data solutions store and query data, one key difference with big data solutions is how data distribution and query processing takes place, and how data moves across the network. Big data solutions work and function in a cluster in which each node in the cluster contains storage and the initial data processing takes place at each node. With the data already loaded onto each node in the cluster, no data needs to be moved onto each node for processing, thus improving processing performance.

Query processing in big data solutions are primarily batch operations. With massive data volumes, a variety of data formats and types, and a tendency for queries to be a bit more complex, these batch operations are likely to take some time to yield results. These batch queries, however, run as multiple tasks across the cluster, making it much easier to handle the volumes of data, and as such, provide a level of performance that is typically not seen in traditional database or other systems. And since we're talking about querying, it should be pointed out that in big data solutions, batch-type queries are by-and-large not frequently executed. Compared to traditional databases where queries are regularly executed as part of a process or application, big data queries are much less frequently executed, which becomes much less of a shortcoming.

Hopefully the insights into the differences between traditional databases and big data solutions provided above has helped make clear that these two solutions are complimentary. Relational databases will continue to be around for quite some time. In fact, it is quite common to see the results of a big data query to be stored in a relational database or data warehouse to be used for BI (Business Intelligence) reporting or further down-the-stream processes.

Putting all of this into perspective, it is all about how organizations tackle the problem of the three Vs. Organizations look to big data solutions to solve the classic problem of how to discover the hidden gems of useful and meaningful information in their data.

A Quick Data Landscape Comparison

It wasn't too long ago that the data landscape consisted of only a handful of technologies to understand data. Data mining, data warehouses, and BI have been a staple and go-to for data processing and data insights for a long time. Today, it's impossible to have the same conversation without including data analysis (or analytics) and big data in the same breath. You can also throw IoT into the conversation, but that will be discussed in the next section so it will be left out of this comparison.

What I want to do is provide a quick look at how big data and analytics compare to the technologies that many data professionals have been using to those which are more recent, specifically BI, data mining, analytics, and big data. Let's not get hung up on the term "data professional" either. For the sake of argument, a data professional can mean a SQL DBA, SQL developer, data scientist, or BI architect.

As a note, this is not a deep, analytical comparison, but more of a simple discussion toward understanding the boundaries between each area and technology, as these tend to blur when they come up in discussion. In fact, you will probably have different opinions, and that is fine. The goal here is simply to help understand the different concepts and navigate their relationships between each other.

- **Business Intelligence**: BI is data-driven decision making based on the generation, aggregation, analysis, and visualization of data. Discussions about BI go beyond the data into what insights can be gleaned from it. BI is spoken in terms of both technology, such as data transformation (ETL processes), and reporting, as well as general processes that encompass the technologies that support the processes.
- **Data Mining**: Data mining is the process of discovering actionable information from large sets of data. It is sifting through all the evidence looking for previously recognized patterns, and finding answers you didn't know you were looking for.
- **Analytics**: Analytics focuses on all the ways you can break down the data and compare one nugget of data to another, looking for patterns, trends, and relationships. Analytics is tied at the hip with BI. Analytics is about asking questions; BI is about making decisions on those questions. Machine learning is a type of analytics—predictive analytics, asking questions about the future.
- **Big Data**: Solutions and technologies that store and process massive volumes and different types of structured and semi- or unstructured data.

There are certainly similarities between each of these terms; for example, one could argue that analytics and BI are synonymous, or rely on one to the exclusion of the other. What helps blur the line is the great tooling available, such as Power BI. It is a tool that provides both data analysis and data visualization.

Internet of Things (IoT)

A portion of the first section talked about the amount and different types of data being generated and the ways to handle that data, and I also listed some examples of the different types of applications and devices that are generating that data (telescopes, jet engines). In order to understand big data, it helps to understand what is generating the different types and amount of data, and that is what this section is about: the "things" that are generating the data and passing that data around the Internet. Thus, the Internet of Things.

The term "Internet of Things" was coined by the guy who helped create RFID, Kevin Ashton. In fact, it was lipstick that ultimately led to the coining of the IoT term. At the time, Kevin was employed at Procter & Gamble as the Oil of Olay lipstick brand manager, and he noticed that a popular color of their lipstick was continually out of stock. Kevin decided to find out why and in digging into the problem he discovered that there was a data problem between the Procter & Gamble stores and the supply chain. The solution he came up with led him to drive the development and deployment of RFID chips on inventory. Kevin asked himself "What if I took the radio microchip out of the credit card and stuck it in my lipstick? Could I then know what was on the shelf, if I had this shelf talk to the lipstick?"

Kevin's efforts in RFID development led to him being loaned out to MIT to start the technology group Auto-ID Center, which would continue the research of RFID technology. It was during this time that he coined the phrase "Internet of Things" in 1999. As such, Kevin is known as the "Father of the Internet of Things."

What Is the IoT?

Simply put, the Internet of Things refers to devices that collect and transmit data over the Internet. These devices can be anything from your toaster or washing machine to your cell phone or wearable devices such as the Microsoft Band, Apple Watch, or Fitbit, just to name a few. It is said that if a device has an on/off switch, generates data, and can connect to the Internet, chances are that it can be part of the IoT. Today, many cars are connected to the Internet. By 2020, it is estimated that over 250,000 cars will be connected to the Internet. The wearable device market grew 223% in 2015 globally. Many companies are investing in home devices, such as Samsung. It is estimated that by "connecting" kitchens to the Internet, the food and beverage industry could save as much as 15% annually. According to some estimates, the Internet of Things will add $10-$15 trillion to the global GDP in the next 20 years.

Other popular devices that are used today to generate data are the low cost, credit card-sized computers such as the Raspberry Pi, Pi Zero, and Tessel IO boards. These little devices are frequently used to create and generate data in places and situations where normal PCs and laptops can't go or aren't efficient enough. In fact, it is these devices that will be used in this book to generate data.

However, even though the term Internet of Things was coined in 1999, the concept of collecting and transmitting data goes back earlier than that. ATMs did it as far back as 1974. Still, with the recent movement and excitement around big data and IoT, IoT is still a new concept to a lot of people.

There are several key factors that are helping drive the IoT space, and one is the inexpensive costs of the components. For example, both the Raspberry Pi and Tessel IO board are less than $40, and the associated sensors and modules, such as GPS, accelerometer, climate, and ambient modules, are even less. Each of these credit card-sized boards is powerful enough to run advanced software, such as Windows 10 Core. There are other boards as well, including ones from Intel and Arduino. Chapter 2 will cover these boards.

Another factor is the advancement in software that provides rich, dynamic, and high-level data processing and analysis capabilities. The software is becoming more powerful and easy to use, providing the ability to deliver the high-level and enterprise-class analytics big data solutions need.

The explosion of cellular and wireless connectivity has also boosted IoT solutions, allowing big data and IoT solutions to include mobile components and provide connections to the Internet where previously impossible. As connectivity to the Internet continues to improve and become less of a cost barrier, the increase in IoT scenarios has accelerated tremendously.

Lastly, a major key factor in the explosion of IoT scenarios has been the rapid advancement in cloud services and technologies, providing a highly cost-effective solution for data storage, processing, and analysis. Coupled with the fact that the current development tools and technologies used on-premises also work when developing for the cloud makes using cloud-based services and solutions highly advantageous.

The Internet of "Your" Things

It was mentioned earlier that the Internet of Things refers to devices that collect and transmit data over the Internet. This year it is estimated that there will be over 5 billion connected "things," or devices, to the Internet. Gartner estimates that by 2020 there will be over 25 billion connected devices, and some believe that the number will be even higher. Today, businesses are using the data generated by these devices to create tremendous business value through the analysis of this data.

The question you need to ask yourself is, what does IoT mean for you? As much talk as there is about IoT, it is useless unless the discussion includes the things, devices, and data that are important for *your* business. Thus, the discussion is really about the Internet of *Your* Things: the devices and data that have an influential impact on unlocking the value and insights your business needs.

The sole purpose of this book is to clearly lay out Microsoft's view on IoT and help paint a clear picture of Microsoft's IoT ecosystem of services and technologies. Thus, the chapters in this book will walk through the generation of data via devices and sensors, and the consumption and analysis of that data using Microsoft's suite of cloud services and technologies to gain insight and analysis of the generated data.

The first set of chapters will focus on "data on the move," meaning data that is being generated, processed, and analyzed in real time. The second set of chapters will focus on the processing and analysis of data at rest, meaning data that has been consumed from devices and sensors and stored for future analysis. Together, the intent of these chapters is to provide an end-to-end example of how one might use Microsoft's services to gain real-time insight into the data generated by the Internet of "Their" Things.

Before proceeding, however, the next few pages will take a look at several real-world examples used today. You may have heard about them before but they are mentioned here because they provide insight and examples into IoT solutions that are in place today.

Scenarios

There are a plethora of scenarios in which IoT solutions have been created and used to improve and solve countless data problems. This section will discuss a few of the scenarios to help paint a clear picture as to what is possible with IoT solutions as well as provide a solid foundation of what the chapters in this book will provide.

The Connected Car

There is commercial (or advert) in the United States by a popular insurance company that has been on the television for a while. The premise of the commercial simply states that you can save on your auto insurance through safe driving habits. The way it works is this: if you are signed up with their insurance and you sign up for this program, they will send you a little device that you plug into your vehicle's OBDII diagnostic port. When the vehicle is on, this device sends information about your trip to the insurance company. This information includes information such as how fast you accelerate, how hard you brake, speed, time of day you drive, etc. All of this information is sent to the insurance company to determine if your driving habits could turn into a cheaper rate. There may be examples of this in other countries as well with other insurance companies.

Personally, this is too "big brother" for me. Are they going to raise rates if they don't like the way I drive? The next step might be putting a speaker in the device and using cloud-to-device messaging (which I'll discuss in the next chapter or two) to tell me in real time that it noticed I was a bit heavy on the gas and to slow down. Nope, I don't need that (not that I am a speeder ☺).

All kidding aside, this is a very simple but excellent example of a connected car. Data is being generated by a device and sent to a location for real-time or near-real-time analysis. In the bigger picture, think of the number of these devices out on the road and the data being generated by them and sent to the insurance company. Now think about how the data is being used and analyzed.

This data could be very beneficial to not only the insurance company but to auto manufactures and other companies. Taken together, all of this data could be used to make cars safer or prevent accidents from happening by preemptively diagnosing an engine problem. From a predictive analysis perspective, possible accidents could be avoided by looking at people's driving patterns and habits.

Another example is Formula 1. Using the cloud, many race teams use real-time data analysis during race time to finish. Formula 1 cars have hundreds, if not thousands, of sensors on their cars, and today these sensors send data back to the pits in real time. This data is then displayed on a dashboard back in the pits for real-time analysis, making for more efficient pit stops.

More and more car manufacturers are building cars that enable people to perform remote diagnostics on their cars and send the data to the Internet for insight and analysis. Many automakers today have made 4G wireless connectivity available in new cars, paving the way for a plethora of services and information, including better navigation, real-time traffic and parking information, as well as enhanced rider experiences.

Mercedes-Benz has recently introduced models that can link to Nest (https://nest.com/), the IoT-powered smart home system which can remotely activate a home's temperature. Cars in the United States and Canada have an ODB (onboard diagnostic) port, which has been mandatory since 1996. It is through these ODB ports that a rapidly growing business of cloud-connected mobile apps are springing up, providing myriad services including maintenance reminders, diagnostic access, and much more.

Connected Home

The advancement of connected home technology has gained a lot of momentum lately. The company X10 (www.x10.com/) has provided home automation gadgets for almost 40 years now. What started as just gadgets and widgets has morphed into modern versions now offered by X10, for example, as well as fun, simple, yet critical projects such as a simple home security system that uses a Kinect and Microsoft's cloud services, made by good friend and co-worker Brady Gaster (go to https://github.com/bradygaster/Kinectonitor).

Other examples of connected homes are things like the June Intelligent Oven or the Samsung Family Hub refrigerator, both of which let you see what's going on inside from your smartphone.

The advancement of IoT solutions in the area of connected homes and buildings continues to focus on controlling nearly every aspect including remote diagnostics, maintenance, and analytics. The connected home is such a growing industry that it warrants its own conference, the Connected Home and Building conference (www.connectedhomecon.com/east/).

I recently read in a March 2016 Business Insider article that connected home device sales will drive over $61 billion in revenue for 2016, with 52% compound annual growth rate to reach $490 billion by 2019. That same report stated that the connected home device category makes up roughly 25% of shipments in the IoT category, and connected home device shipments will outpace smartphone or tablet shipments. That's significant.

Let's not, however, confuse *smart home* with *connected home*. A smart home provides the ability for me to turn the lights on from my phone. A connected home is about the data the devices in your home are generating and how that data is being used to improve home life, save money, drive new products, and keep people safe.

Connected Cow

The connected cow example is a fun and interesting scenario. Perhaps you have heard it, but if you haven't, have a seat. I have talked about connecting cars and homes, but farm animals? This is a scenario where technology has transformed even the oldest industry.

The connected cow scenario is simply where pedometers were strapped to the legs of cows. Each pedometer was connected to the Internet and simply sent the step count of each cow to a web dashboard back in the farmhouse via Microsoft Azure.

Why was this being done? A farm in Japan was looking to solve two things. First, the farm wanted to detect health issues early and prevent herd loss. Second, the farm wanted to improve cattle production by accurately detecting estrus. Estrus is when the animal, in this case the cow, goes into heat.

The problem was that cattle go into heat for a very small window of time (12-18 hours) every 21 days, and this occurs mostly between 10pm and 8am. The goal was to use technology to more accurately detect the estrus period during which time artificial insemination would take place, so that the pregnancy rate would significantly increase.

Without technology, managing hundreds or thousands of cows and accurately detecting when each cow goes into heat is nearly impossible. With technology, this could be doable, and in actuality, it was.

The farm in Japan contacted Fujitsu for help. Fujitsu created a solution that used pedometers. What they found was that a cow takes many more steps when it goes into heat. So they strapped a pedometer to every cow. These pedometers sent the step count to a Microsoft Azure solution, which analyzed the data and sent alerts.

Using this solution, they were able to get up to 95% accuracy in the detection of estrus as well as the optimum time for artificial insemination. Upon further analysis, Fujitsu and the researchers also found out something else very interesting. They found out that there is a window around the optimal timeframe for artificial insemination. They found that if you perform artificial insemination in the first half of the window, you are more probable to get a female (cow), and if you perform artificial insemination in the second half of the window, you are more probable to get a male (bull), with up to 70% probability. The farmer now has the ability to control production whether he needs more cows or bulls.

Another interesting thing they found is that they can detect from between 8 and 10 different diseases from the step patterns and number of steps returned from the pedometers of the cows.

Also, the farm was able to improve their labor savings simply from not needing to manually monitor the cows constantly. The impact of savings from implementing the solution was quite significant in many aspects.

There are many great examples of how IoT is being used today, and we are just scratching the surface in each of these areas, so there is more to explore and discover. As you explore and learn more about IoT, think about how all of this data can be used to improve our way of life, lower costs, save lives, and much, much more.

Summary

Today it seems like everyone has their own definition of what big data and the Internet of Things means. But there is no question that there is huge opportunity and a lot of potential in this space. This chapter began by defining big data and looking at the characteristics of big data. To help lay the foundation for the rest of the chapter, and the book, the chapter discussed why understanding big data concepts and technologies is important.

Building on the big data theme, the chapter shifted from big data to the Internet of Things, first providing insight into where the term comes from and then taking a good look into what IoT is and means. To provide some context, the chapter finished with several scenarios where the Internet of Things is being used to make the world a better place and set forth a challenge to think about how one might use IoT to change the world for the better.

CHAPTER 2

Generating Data with Devices

The Internet of Things, or IoT, is about devices that generate and transmit data over the Internet. As mentioned in the last chapter, it is estimated that by 2020 there will be over 25 billion "things" (i.e. devices) connected to the Internet; these devices will range from the washer or oven in your house to the watch you wear or the phone in your pocket. The last chapter also covered several scenarios in which the IoT and devices were implemented, including cars, homes, and animals.

This chapter will simply build on that information and show several examples of the types of devices used today, how to use them to generate data, and how to send that data over the Internet and store it for analysis. The devices that this chapter will use will be the Raspberry Pi board from the Raspberry Pi Foundation (www.raspberrypi.org/) and the Tessel board from the Tessel Project (https://tessel.io/). There are a large number of devices available that provide very similar functionality, including the Adafruit Feather HUZZAH, the Edison from Intel, the DragonBoard from Arrow, and the boards from Beaglebone, but it would be unrealistic to discuss and demo them all here. Visit https://azure.microsoft.com/en-us/develop/iot/get-started/ for a full list of IoT boards supported by Microsoft.

Adafruit makes a great Azure IoT starter kit complete with the Feather HUZZAH board, a DHT22 sensor, cables, breadboard, a power cable, and other jumpers and switches. It is available via the Adafruit web site: www.adafruit.com/products/3032. This chapter won't cover how to use it but I have blogged about how to get it set up and running with Azure. Visit my blog for more information.

The reason I have chosen the Raspberry Pi is because of its ability to run Windows 10 IoT Core and the familiar development environment of Visual Studio, which makes it easy to code and deploy solutions. I have chosen to also discuss the Tessel board not only because it is red (my favorite color) and looks cool, but to also illustrate Microsoft's support for open source technologies. In all fairness, I am including the Tessel in this chapter simply to show Microsoft's support for a wide range of operating systems, languages, tools, and frameworks.

Now, a couple of things by way of disclaimer. First, as discussed in the introduction, this book is about Microsoft's vision of big data and IoT via the IoT and Cortana Intelligence suite of Azure services, thus data generated will be routed as such. However, for the sake of the examples of this chapter, the data generated from these devices will simply be routed to the screen for output. Chapter 3 will show in detail how to create, configure, and connect the device to an Azure IoT Hub for data storage and processing, and Chapter 4 will then hook up the devices to an Azure IoT Hub. This chapter is all about using devices to generate data.

So, with that, let's get started.

Raspberry Pi

A popular IoT device, the Raspberry Pi is a small but proficient device that does everything a normal computer does by plugging in a monitor, keyboard, mouse, and Ethernet cable. You can browse the Internet, play games, or even run desktop applications such as Microsoft Word or Excel. However, this tiny but powerful mini-pc is really targeted to those who want to explore the world of devices and maker projects. A "maker" is a hardware hacker, someone who likes to build things that make the world a better place.

© Scott Klein 2017
S. Klein, *IoT Solutions in Microsoft's Azure IoT Suite*, DOI 10.1007/978-1-4842-2143-3_2

In Figure 2-1 you can see two Raspberry Pis. The device on the top is the Raspberry Pi 2 Model B, which has been available since February of 2015. The device on the bottom is the Raspberry Pi 3 Model B, which became available in February of 2016.

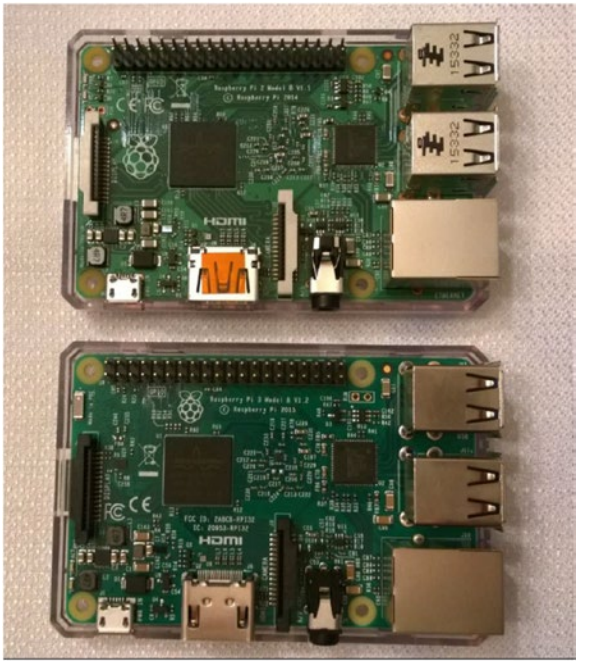

Figure 2-1. *Raspberry Pi 2 (top) and 3 (bottom)*

The Raspberry Pi is about the exact same size as a credit card and both the Pi 2 and Pi 3 vary very little in functionality. Table 2-1 details the main differences between the Pi 2 and the Pi 3.

Table 2-1. *Comparing the Raspberry Pi 2 and 3*

Raspberry Pi 2	Raspberry Pi 3
• 900MHz quad-core ARM Cortex – A7 CPU • 1GB RAM	• 1.2GHz 64-bit quad-core ARMv8 CPU • 1GB RAM • 802.11n Wireless LAN • Bluetooth 4.1 • Bluetooth Low Energy (BLE)