WHITE PAPER

IoT and Big Data

With the incredible growth of the Internet of Things (IoT), Big Data technologies play a vital role in processing the massive amounts of data that will be generated from Internet-connected things. IoT and Big Data are bringing significant business benefits for the early adopters.

Data from IoT poses unique characteristics and challenges for data analytics. The typical architecture of an IoT Big Data ecosystem starts from high-speed ingestion, data storage and covers analytics and reporting from live and historic data. Tools from Azure® and AWS® as well as open source alternatives are available for implementing an IoT Big Data ecosystem.

THINXTREAM®

Introduction

The incredible growth of the Internet of Things (IoT) is changing our world. Two decades ago, the Internet was confined to computers. Since then, it has encompassed mobile devices including phones, smartphones, and tablets and today even the ubiquitous automobiles, TVs, and refrigerators. As this connectivity expands, it provides an opportunity to better understand the relationships between disparate data, derive new insights and make informed decisions. Big Data technologies and techniques play a vital role in processing the massive amounts of data that will be generated from Internet-connected things.

In this white paper, we examine the unique characteristics of IoT data and challenges, the typical architecture of a Big Data system and the tools available to implement this, and finally some real-world applications of IoT and Big Data.

IoT Big Data Characteristics

It is important to understand the key characteristics of IoT and associated challenges for designing an architecture for a big data system.

Number of IoT Devices: With cost reduction, comes increased adoption and drives tremendous scale. We can expect tens of billions of connected devices in the future. IoT products and solutions are going to be tested on a massive scale like never before.

Multiple IoT Devices & Data Types: IoT is a system of systems and often the devices come from multiple manufacturers. Data from an individual device manufacturer or model may be quite dissimilar from that of a nearly identical device. Moreover, they may produce structured, semi-structured, and unstructured data with a wide variety of data types such as multiple text formats like XML, JSON, plain text, audio, video, sensory data, etc.

Veracity of Data: In an IoT setup, there will be instances where IoT devices either send spurious data or completely fail, resulting in data aberrations, or failure to perform a required control function. Veracity refers to the quality, consistency, and trustworthiness of the data, which impacts the accuracy of analytics.

Update Frequency: Devices produce data at varied frequencies. While remote sensors produce data at a low frequency, more sophisticated things like a car produce it at a high frequency. Given this high rate of data reporting, you need advanced tools and technologies to efficiently analyze reported data.

Historical Data: Often Big Data insights are derived from current data along with historical data recorded from IoT devices to enable smart monitoring and control. However, many times you may not have the history yet and building this will require time.

Context Data: Context adds significant value to the collected IoT data. It can be either static/low update frequency such as location or could be more dynamic like weather data. Establishing a relationship between these may pose challenges.

Privacy Issues: Storing IoT data collected from connected devices bring substantial risk relating to data privacy. Sometimes this data is shared with external systems for creating new applications. When this is done, IoT systems should ensure end-users always remain in control of their personal information through security policies.

IoT Big Data Architecture

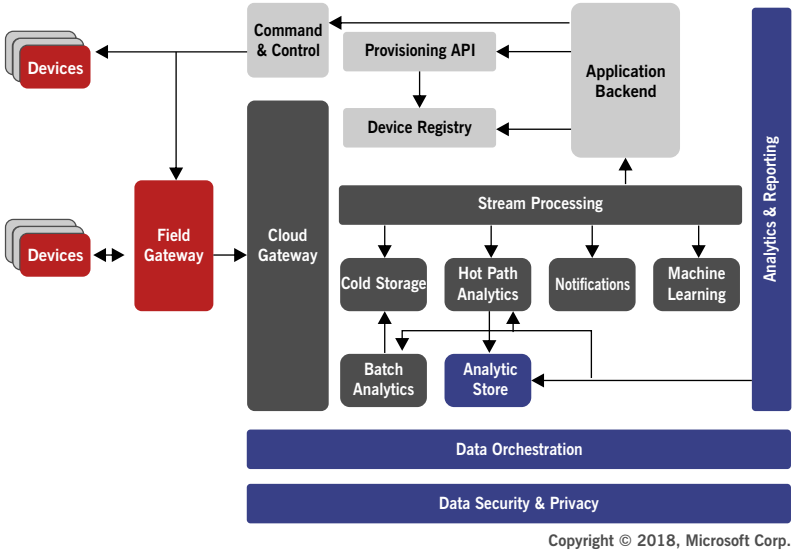

Figure 1. Big Data Architecture for IoT

Solving the challenges concerning Big Data from IoT systems demands well-thought-out design. Many big players have provided reference architectures, which is an outcome of their research and learnings from actual implementation.

Based on a logical Big Data architecture for IoT from Microsoft®, let us review the following logical blocks and the options available from Azure®, AWS® and open source.

Cloud Gateway: Cloud gateway captures and stores messages coming from devices. It acts as a buffer for messages, implements message queuing, reliable delivery and scale-out processing. Cloud gateway options include Azure® Event Hubs, Azure IoT Hub, AWS® Device Gateway, Amazon Kinesis®, and open source alternatives such as Apache Kafka®.

Stream Processing: After capturing real-time stream, the solution must filter, aggregate, and prepare data for the next step and pass it to subscribers. Azure Stream Analytics and Amazon Kinesis provides a managed stream processing service to run SQL queries on a continuous stream of data. Streaming technologies like Apache Storm™ and Apache Spark™ in an HDInsight cluster are some of the open source alternatives.

For near real time analysis of the data stream, a speed layer or hot path examines and detects anomalies, recognizes patterns over time windows, or triggers alerts based on specific stream conditions.

Cold Storage: Data for non-real time processing such as batch is typically stored in an economical distributed file store (referred to as Data Lake) that can hold high volumes of large files in various formats. Options for implementing this storage include Azure Data Lake Store, blob containers in Azure Storage, Amazon S3®, and Amazon Glacier®.

Batch Analytics: For large sets of data, and non-real time use cases, batch jobs are the best fit for filtering, aggregating and preparing data for further analysis. These batch jobs generally involve

steps such as reading source files, processing them, and writing the output to new files. Options include running U-SQL jobs in Azure Data Lake Analytics, using AWS Glue® with Amazon Athena® combinations, or open source Apache Hive™, Apache Pig™, or custom Map/Reduce jobs in an HDInsight Apache Hadoop® cluster or Amazon Elastic MapReduce.

Analytical Data Store: Big Data solutions generally prepare data for analysis and then provide an interface to query this structured data using analytical tools. Azure SQL Data Warehouse and Amazon Redshift® provide a managed service for large-scale, cloud-based data warehousing.

Analysis & Reporting: The key objective of Big Data solutions is to deliver insights into the data through analysis and reporting. For doing this data modeling layer is essential. Azure Analysis Service, for instance, provides a multidimensional Online Analytical Processing (OLAP) cube or tabular data models. Tools such as Microsoft Power BI® and Amazon Quicksight® aid in visualizing the results. For data scientists or data analysts, interactive data exploration may be supported with analytical notebooks such as Jupyter™.

Orchestration: Most big data solutions consist of a fixed workflow of transforming data, routing data between multiple sources and destinations, loading the data into an analytical data store and generate a report/update dashboard. To automate these repeat steps, you can use orchestration technology.

The table overleaf summarizes the various Big Data offering from open source and commercial platforms for each of the blocks (and more) in the architecture diagram.

	Open Source	AWS®	Microsoft Azure ®
Batch Ingest	Apache Sqoop ™ File Transfer Apache Flume ™ StreamSets ™	AWS Data Transfer Services (various options)	Import/Export Service Data Factory
Streaming Ingest	Apache Flume StreamSets	Amazon Kinesis Firehose	Event Hubs IOT Hub
Persistent Storage	HDFS ™ RDBMS	Amazon S3, Amazon Glacier Amazon RDS	Storage Blob HDFS SQL Database
Transient Storage	Apache Kafka	Amazon Kinesis	Event Hubs IOT Hub HDInsight (Kafka)
Batch Processing	Apache Hive Apache Flink®, Apache Spark Hadoop MapReduce ™ PostgreSQL ®	Amazon Elastic MapReduce (EMR) Spark Amazon EMR Hadoop Amazon EMR Presto AWS Batch Amazon Redshift	Azure Batch HDIns ight (Spark/Map Reduce) SQL Data Warehouse Data Lake Analytics Azure Functions
Stream Processing	Apache Flink Apache Spark Apache Beam	Amazon Kinesis Streams Amazon Kinesis Analytics Amazon EMR Spark	Stream Analytics HDInsight (Storm, Spark)

	Open Source	AWS®	Microsoft Azure ®
Machine Learning	Scikit-learn™ TensorFlow™ Apache Spark MLLib TensorFlow , etc. Huge number of libraries	Amazon Lex Amazon Polly Amazon Rekognition Amazon Machine Learning	Azure ML Cognitive Services
Serving Storage Graph	JanusGraph®	Amazon Neptune	CosmosDB
Serving Storage BI/EDW	Apache Impala® + Apache Kudu™	Amazon Redshift Amazon Athena	SQL Data Warehouse Analysis Services (OLAP Cubes)
Serving Storage Search (keywords + facets)	Apache Solr™	Amazon CloudSearch Amazon Elasticsearch	Azure Search
Serving Storage RDBMS	PostgreSQL	Amazon RDS	SQL DB
Serving Storage NoSQL	Apache HBase®	Amazon DynamoDB	HDInsight (HBase) CosmosDB
Sandboxes Notebook	Apache Zeppelin™	Amazon EMR Zeppelin	Azure Notebooks
Clients/Data Apps	Superset (BI)	Amazon Quicksight	PowerBI
Orchestration	Apache Airflow™	AWS Data Pipeline	Data Factory
ETL Tool	N/A	AWS Glue	Data Factory

Figure 2. Big Data Offerings from Open Source, AWS and Microsoft Azure

IoT Big Data Applications

Here are some real-life examples of IoT and Big Data working together in businesses.

- One of the largest shipping companies in the world, UPS, captures 200 data points from each vehicle in a 80,000-strong fleet to monitor speed, miles per gallon, mileage, number of stops, and engine health. These help the company optimize usage of their fleet, decrease fuel consumption and control emissions.
- Barcelona has implemented smart parking meters that give residents real-time updates on their phone regarding availability of parking slots and enables them to pay through the mobile app.
- The John Deere Field Connect system monitors air and soil temperature, wind speed, humidity, solar radiation, rainfall and leaf wetness through environmental sensors. This helps farmers in making timely irrigation decisions.
- Disney uses RFID based wearable bands to collect data on visitor movement throughout the park. Disney uses this information to streamline guest numbers at attractions, adequately staff rides and attractions, and regulate stocks at busy shops and restaurants.
- Jewelry store chain Alex and Ani use Bluetooth® sensors in their stores. This setup can track customer traffic in their stores and push specialized offers and messages to users' phones as they enter the store.

- Food Company King's Hawaiian uses data collected from connected machines in their bread production factories to monitor factory performance, reduce potential downtime of machines and lower maintenance costs.

- BC Hydro allows its users to track their energy use by the hour and see trends in their own usage data. With this, electricity theft has been greatly reduced and outages automatically alert the company when the power is out in a certain area.

Conclusion

Data originating from an IoT solution poses unique challenges to consider when choosing or building an IoT Analytics solution. An IoT Big Data ecosystem, starts from high-speed ingestion and covers analytics and reporting. Popular commercial platforms from Azure and AWS as well as open source alternatives are available for each of the building blocks of an IoT Big Data ecosystem. IoT and Big Data have brought significant business benefits for the early adopters.

As an IoT services provider, Thinxtream has expertise in Big Data for IoT solutions across Azure IoT Hub, AWS IoT and Open Source platforms. By leveraging the IoT expertise built over a decade, Thinxtream ensures cost-effective, quality and timely delivery of IoT solutions.

References

- IoT Big Data Framework Architecture Version 1.0
- Big Data Architecture
- Big Data Architecture Style
- 10 examples of IoT and Big Data Working together

Thinxtream Technologies is a global software company with a portfolio of innovative software platforms, products, components, solutions, patents, competences and services for Internet of Things (IoT) across several industry verticals and applications, successfully enabling leading customers, including Fortune 500 companies, meet their application, product and business goals.

Interested in learning more? For more information contact:

Thinxtream Technologies Pte. Ltd.
220 Orchard Road #05-01
Midpoint Orchard
SINGAPORE 238852
Phone: +65 66358625
Email: info@thinxtream.com

Thinxtream Technologies, Inc.
10260 SW Greenburg Road
Suite 400 Portland, OR 97223,
U.S.A
Phone: +1 503 293-3598
Email: info@thinxtream.com

www.thinxtream.com

LinkedIn/thinxtream

TT-WP-006-1-0918

THINXTREAM®