Data Analysis and AI for IoT Networks

Embedded Interface Design

with **Bruce Montgomery**

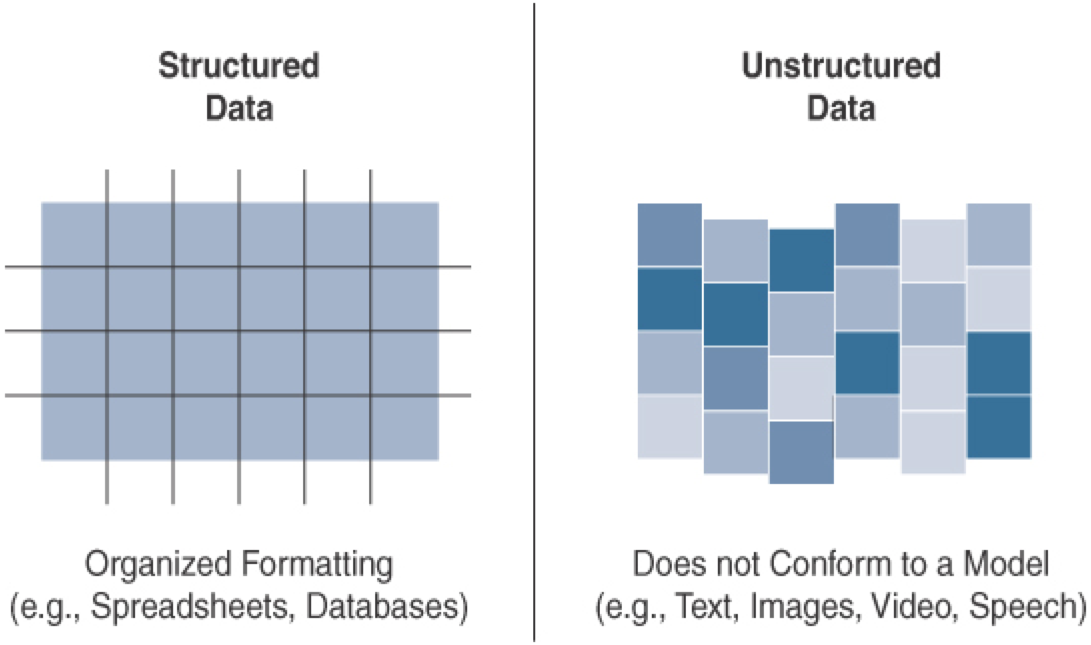
Learning Objectives

Students will be able to…

- Consider data in larger scale IoT networks
- Understand the role of AI and Machine learning in IoT networks
- Consider characteristics and tools for big data and edge analytics
- Consider characteristics and tools for network analytics in FANs (Field Area Networks)

Another view of IoT Data

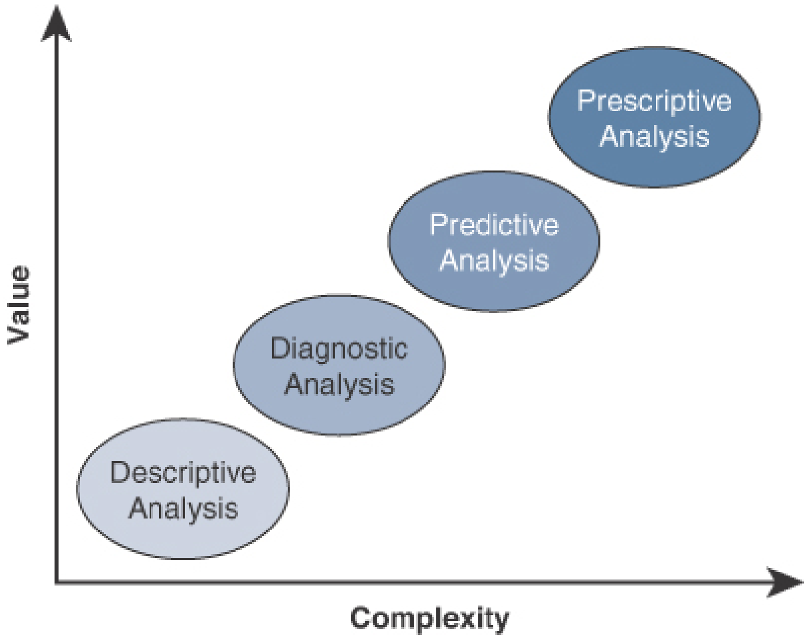
- Structured vs. Unstructured – estimate is 80% of a business' data is unstructured (text, speech, images, video) and requires advanced analysis – machine learning, natural language processing, etc. to extract key information [1]
- Smart objects and IoT devices generate both types.

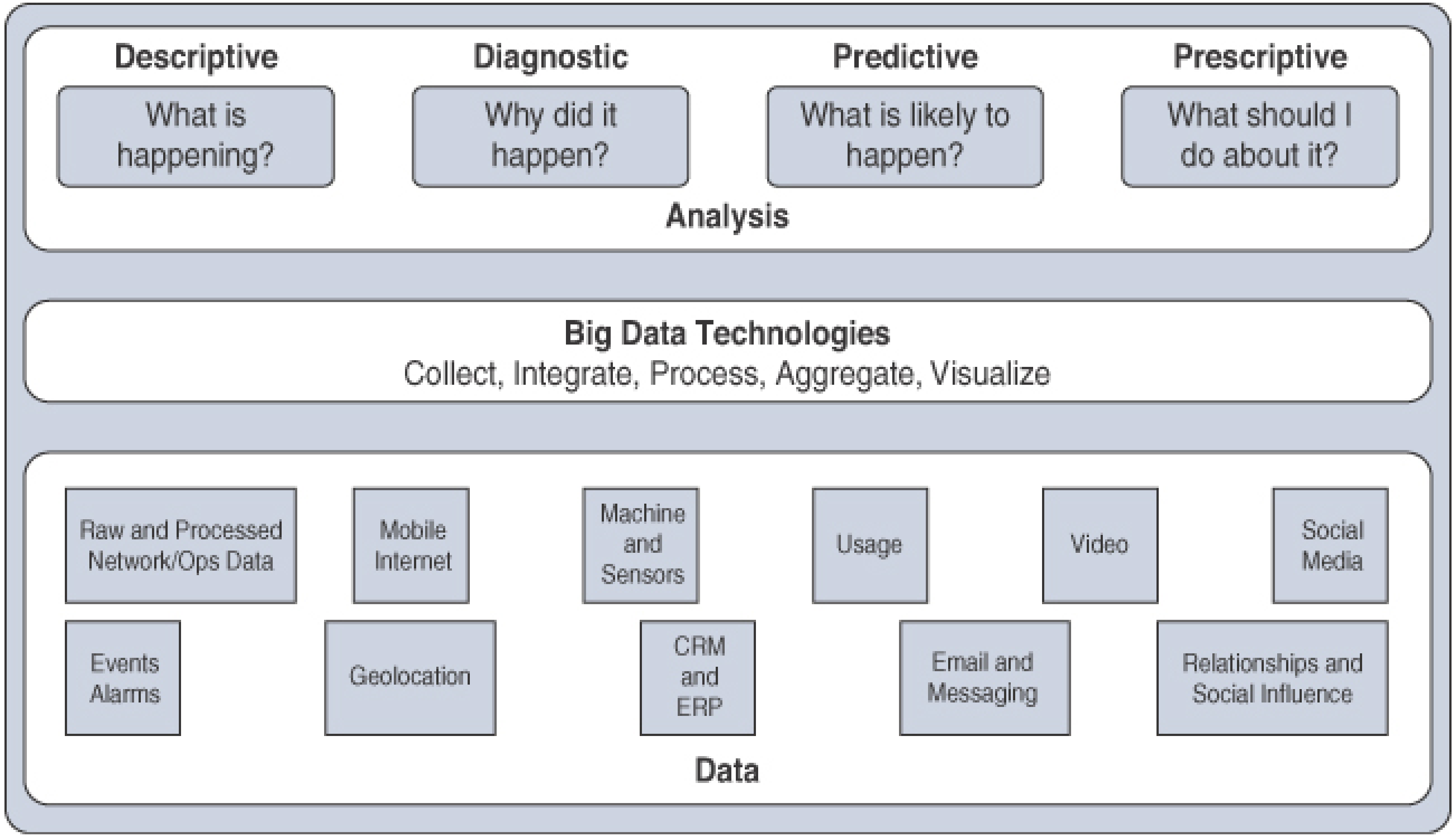

Structured Data
Organized Formatting (e.g., Spreadsheets, Databases)
Unstructured Data
Does not Conform to a Model (e.g., Text, Images, Video, Speech)

Data in Motion, Data at Rest

- Data in IoT networks is in motion or at rest
- Examples of data in motion include traditional client/server exchanges, such as web browsing and file transfers, and email
- Data saved to a hard drive, storage array, or USB drive is data at rest
- Data process at the edge in a fog network is in motion, as it’s eventually pushed out to the cloud
- Tools that work with data in motion are relatively new compared to working with stored data, which has many toolsets
- An example tool for stored data analysis is Hadoop

IoT Data Analytics





Descriptive

What is happening?

Diagnostic

Why did it happen?

Predictive

What is likely to happen?

Prescriptive

What should I do about it?

Analysis

Big Data Technologies

Collect, Integrate, Process, Aggregate, Visualize

Raw and Processed Network/Ops Data

Mobile Internet

Machine and Sensors

Usage

Video

Social Media

Events Alarms

Geolocation

CRM and ERP

Email and Messaging

Relationships and Social Influence

Data

- Scale and volatility of data are key challenges
- Reference [1]

Value of IoT Data Analytics

- Being able to analyze high volumes of streaming data for anomalies or trends, limits or changes, is only valuable if you can work with it in real-time (or nearly so)
- Cloud vendors continue to extend streaming analysis tools and visual dashboards of networked device data behavior
- Network analytics become a key element of effective IoT networks
- Data can be analyzed as big data collections or at the edge near to collection
- Tools like Flexible NetFlow and IPFIX provide examples of this streaming analysis for massive networks as in utilities management (more later)
- Reference [1]

Machine Learning in IoT

- Machine Learning, Deep Learning, Neural Networks, Convolutional Networks are all terms common in discussing analysis of big data from IoT networks…
- At the core, the analysis of the data from an IoT network provides the users the ability to take intelligent actions based on the analysis
- These AI tools are not new, they've been around for some time
- Much like for wearables, the availability of connected networks, increased data sources, open source, and Moore's Law increases in power of hardware have driven them into common use
- Reference [1]

AI, Machine Learning - defined

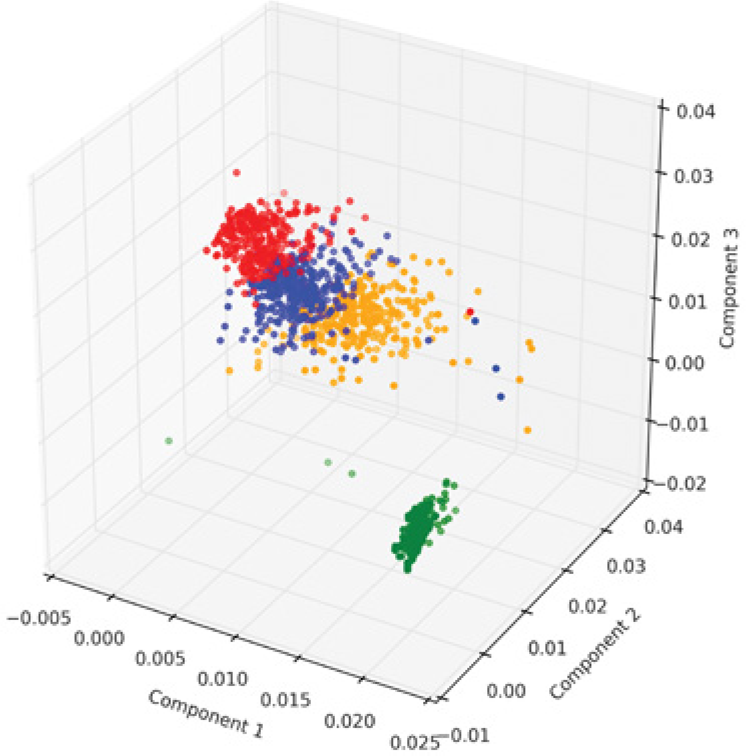
- Artificial Intelligence, the category where these previous analysis tools belong, is simply any technology that allows the mimicking of some aspects of human intelligence – using any technique
- Machine Learning is a process where a set of data is analyzed to perform a task more efficiently, where the parameters for analysis rules can change or are imperfectly understood
- Two types of Machine Learning are common – supervised and unsupervised
- Reference [1]

Supervised machine learning

- Consider an IoT system designed to monitor a tunnel for human presence
- A video sensor can capture shape data
- **Supervised machine learning** would use **training sets** of images to train the parameters of an algorithm what is a human and what's not
- The process of determining from the available data whether an image does represent a human or not is known as classification – does a data set belong to a specific category
- This differs from regression, which is predicting numeric values
- Reference [1]

Unsupervised machine learning

- Consider a set of IoT devices monitoring manufacturing systems in a factory
- You'd like to gather data to find the .1% of bad products being produced
- But the products are made up of 100s of parts, and it is difficult to detect specific defects
- What you can do is measure each production machine's parameters – sound, pressure, temperatures, etc.
- Mathematically these parameters can be grouped in different ways (for example, using K-means clustering) to find mean values
- When you've found a defective product, you can use the multidimensional parameters gathered to find what conditions are more likely to result in defects



Using Machine Learning for IoT Applications

- If you can identify the algorithms and learning models for a given IoT use case, the benefits of understanding your data is obvious
- Machine Learning operations fall into two general categories
  - Local Learning – in the edge node or in the gateway (fog-based)
  - Remote Learning – from collected data sent to the cloud
- There are four major domains for IoT machine learning applications
  - Monitoring
  - Behavior Control
  - Operations Optimization
  - Self-healing or Self-optimization
- Reference [1]

Uses of Machine Learning

- Monitoring – using environmental data to detect failure conditions or unexpected environmental changes – examples: oil flow in a pipe, temperature and pressure at a pump, etc.
- Behavior Control – combined with monitoring, triggering corrective actions – examples: lowering pressure in a pipe, slowing pump rotations
- Operations Optimization – analyzing data to improve operations past simple corrective actions – looking at peak flows or system variations to tune systems to maximum efficiency
- Self-healing, Self-optimizing – having a device change its behavior based on analysis of its condition – reducing sensor counts due to low battery, avoiding movement of a mechanical part to avoid fatigue limits, etc.
- Reference [1]

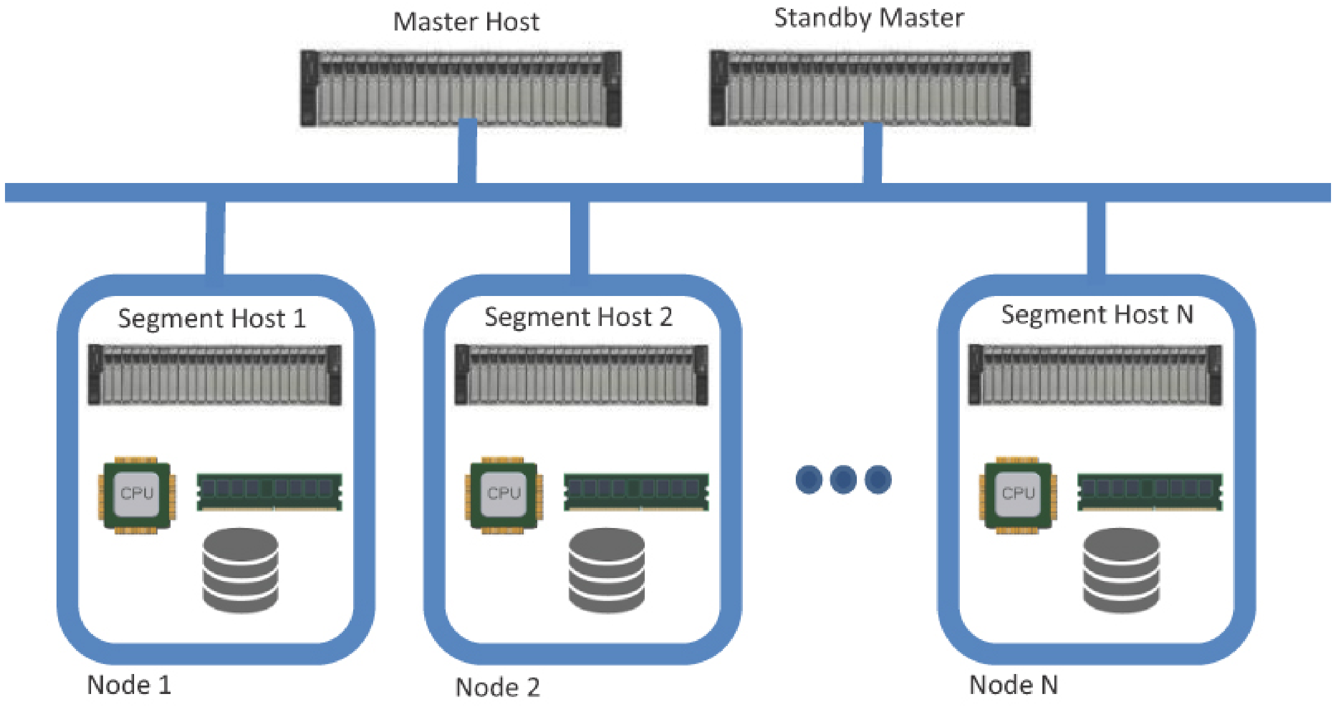
Big Data Analytics

- Three ways to categorize big data
  - Velocity – how quickly is data collected and analyzed
  - Variety – what types of data are being analyzed
  - Volume – what scale of data: gigabytes, petabytes, exabytes?

- Types of big data analysis
  - Machine data – as from IoT devices
  - Transactional data – from systems performing specific functions
  - Social data – high volume data with varying structures
  - Enterprise data – data from a variety of business sources

Big Data Processing Architectures

- We've discussed relational and no-SQL databases
- Beyond traditional Relational databases
- Massively Parallel Processing Databases (MPP Databases)
  - Multiple computers designed in a scale-out architecture with data and processing distributed across multiple systems – operating in a "shared-nothing" fashion



Reference [1]

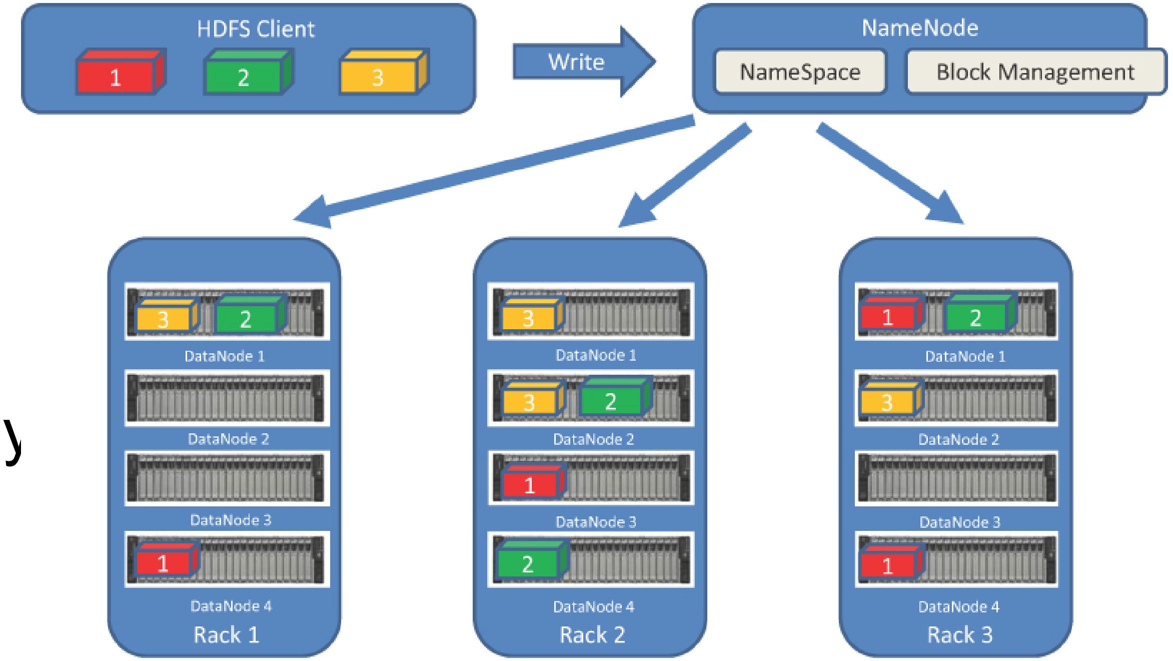
Processing Big Data - Hadoop

- Hadoop is a data repository and processing engine, originally developed from projects at Google and Yahoo!
- Initial focus was to index millions of websites and quickly return search results
- The initial project had two key elements:
  - Hadoop Distributed File System (HDFS) – a system for storing data across multiple nodes
  - MapReduce – a distributed processing engine splitting a large task into smaller ones that can be run in parallel
- Both elements are still part of Hadoop and other big data analysis tools

- Reference [1]

Hadoop Architecture - HDFS

- HDFS uses NameNodes and DataNodes to manage data across a distributed scalable storage and compute system
- NameNodes are critical for data add, move, delete, or read on HDFS – they coordinate and map where data is stored and replicated, typically in 64 MB or 128 MB blocks
- DataNodes are the servers where data is stored, and are often highly replicated for data redundancy



Reference [1]

MapReduce and the Hadoop EcoSystem

- MapReduce uses a similar model to process data across the cluster of nodes – a query is broken down into smaller tasks and distributed across nodes running MapReduce – however, MapReduce is not a real time analysis tool
- Another tool, YARN (Yet Another Resource Negotiator) was added to take over resource negotiation and job/task tracking, leaving MapReduce to focus on data processing
- Over time, Hadoop has grown to cover over 100 software projects (the Hadoop Ecosystem) for data collection, storage, processing, analysis, and visualization – and continues to be applied to processing of IoT network data

Reference [1]

Other Hadoop Elements

- Apache Kafka – A distributed publisher-subscriber message processing system that can flow data produced from masses of IoT devices to real-time consumers for data processing using a topic method similar to MQTT
- Apache Spark – In-memory distributed data analytics for near real-time event processing from streaming sources broken into “microbatches”
- Lambda Architecture – A combined layered system for data management that includes ingesting stream data as well as using Hadoop for batch analysis
- AWS provides Apache Spark, Hadoop, and other technologies via AWS EMR

Reference [1], [3]

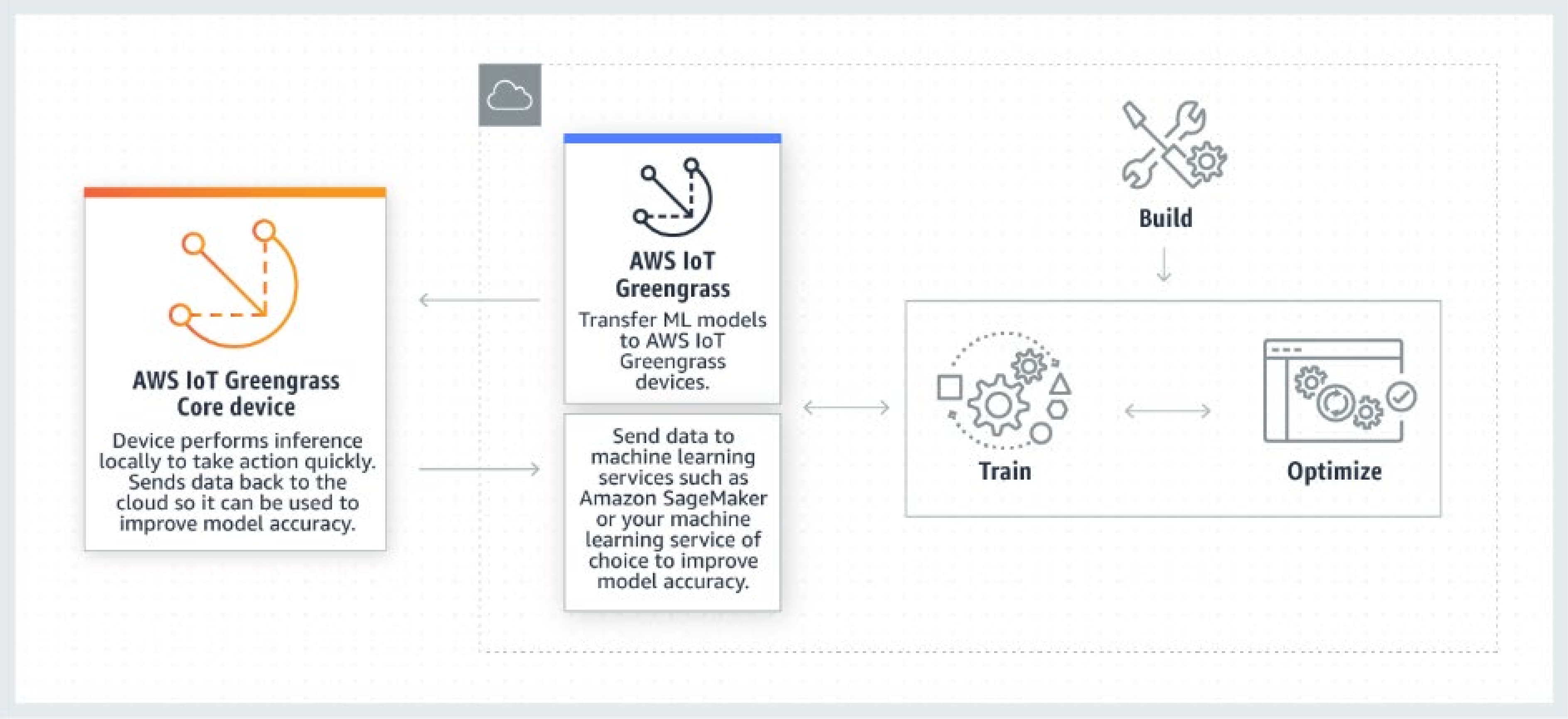
Edge Analytics - Benefits

- The benefits of edge-based streaming analytics for IoT devices include
  - Reducing data at the edge – not passing data to the cloud that uses up expensive bandwidth or network infrastructure
  - Analysis and response at the edge – using the data where there's an immediate need at or near the collecting device
  - Time sensitivity – avoiding multistage data passing and analysis for immediate response to changing conditions

Reference [1]

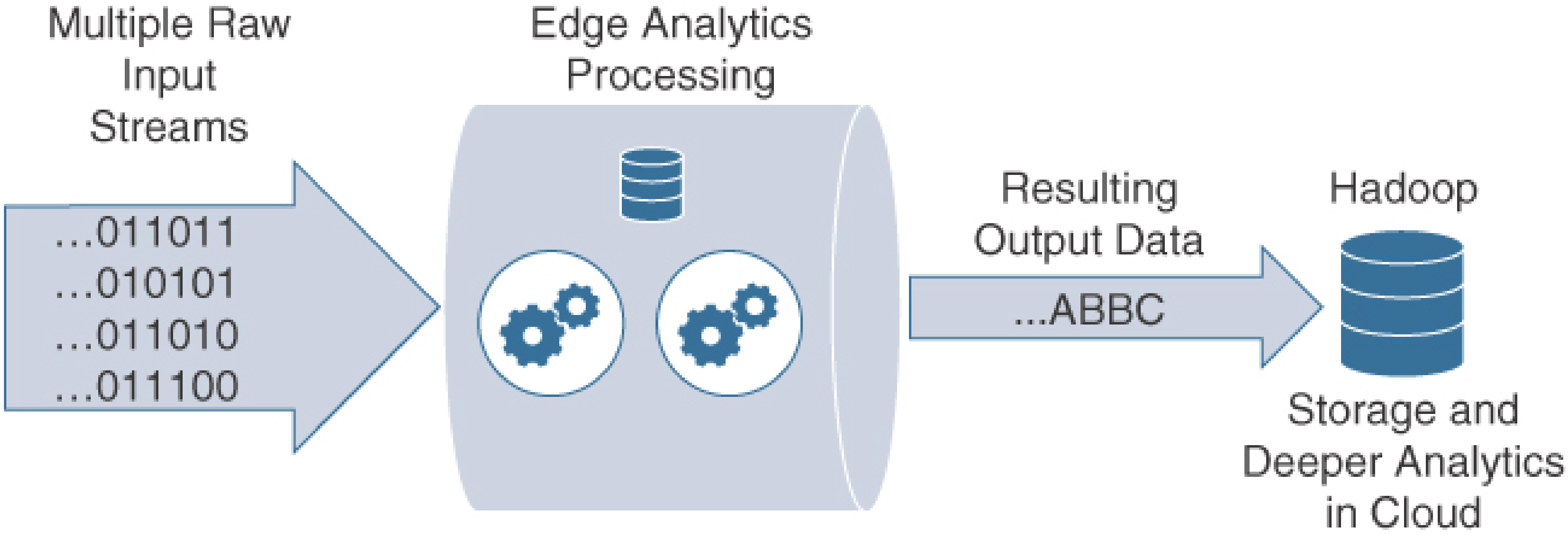
Edge AI – AWS IoT Greengrass ML Inference

- The IoT Greengrass edge support can use models trained in the cloud that are later transferred down to individual devices
- Reference [2]


AWS IoT Greengrass Core device
Device performs inference locally to take action quickly. Sends data back to the cloud so it can be used to improve model accuracy.
AWS IoT Greengrass
Transfer ML models to AWS IoT Greengrass devices.
Send data to machine learning services such as Amazon SageMaker or your machine learning service of choice to improve model accuracy.
Build
Train
Optimize

Edge Analytics – Core Functions

- Raw data input, Analytics processing unit (APU), Output Streams
- The APU may use a local device microservice to act on results


Multiple Raw Input Streams
...011011
...010101
...011010
...011100
Edge Analytics Processing
Resulting Output Data
...ABBC
Hadoop
Storage and Deeper Analytics in Cloud

Reference [1]

Functions of an Edge Analytics Processing Unit (APU)

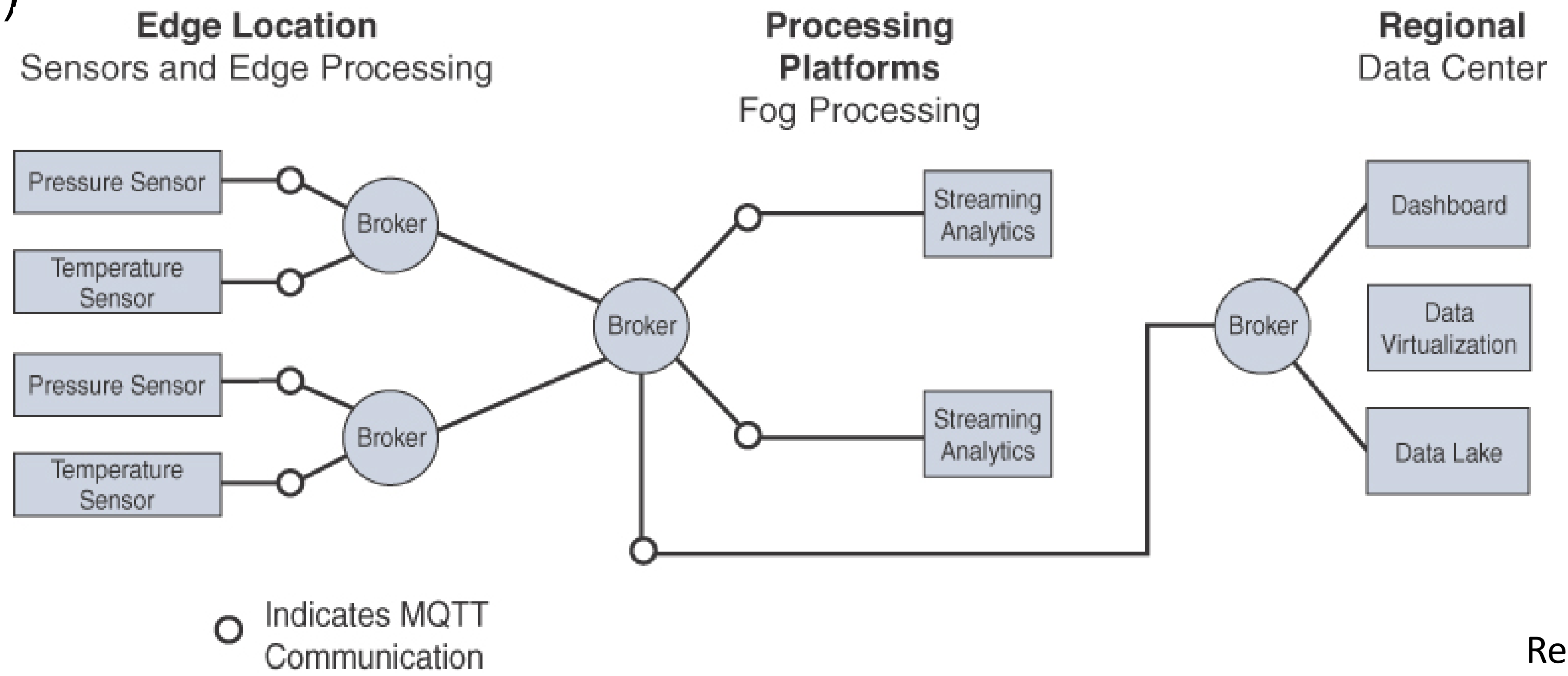
- Filter – trim out supervision data or other data flow not key to local analysis
- Transform – convert raw data into formats for local processing
- Time – provide a timing context for incoming data
  - Example monitoring a group of IoT temperature sensors for vaccine storage producing average estimates every two minutes, storing data for mandatory reporting, and processing alarm or warning conditions for out of bound readings
- Correlate – use data from multiple sources to assess overall state
- Pattern matching – using locally maintained historical data to supplement algorithms looking for anomalous behaviors

- Reference [1]

Reference [1]

Combining Edge and Big Data Analysis

- Example of processing inputs from IoT devices on an oil rig using MQTT as an application protocol – Distributed Analytics for Field Area Networks (FANs)



Reference [1]

Benefits of Large Scale Flow Network Analytics

- Network traffic monitoring and profiling
- Application traffic monitoring and profiling
- Capacity planning – for the network or what's produced
- Security analysis
- Accounting
- Data warehousing and data mining

- Typical tools for network architectures of IoT devices and control systems are Flexible NetFlow (FNF) from Cisco and IETF IPFIX

Reference [1]

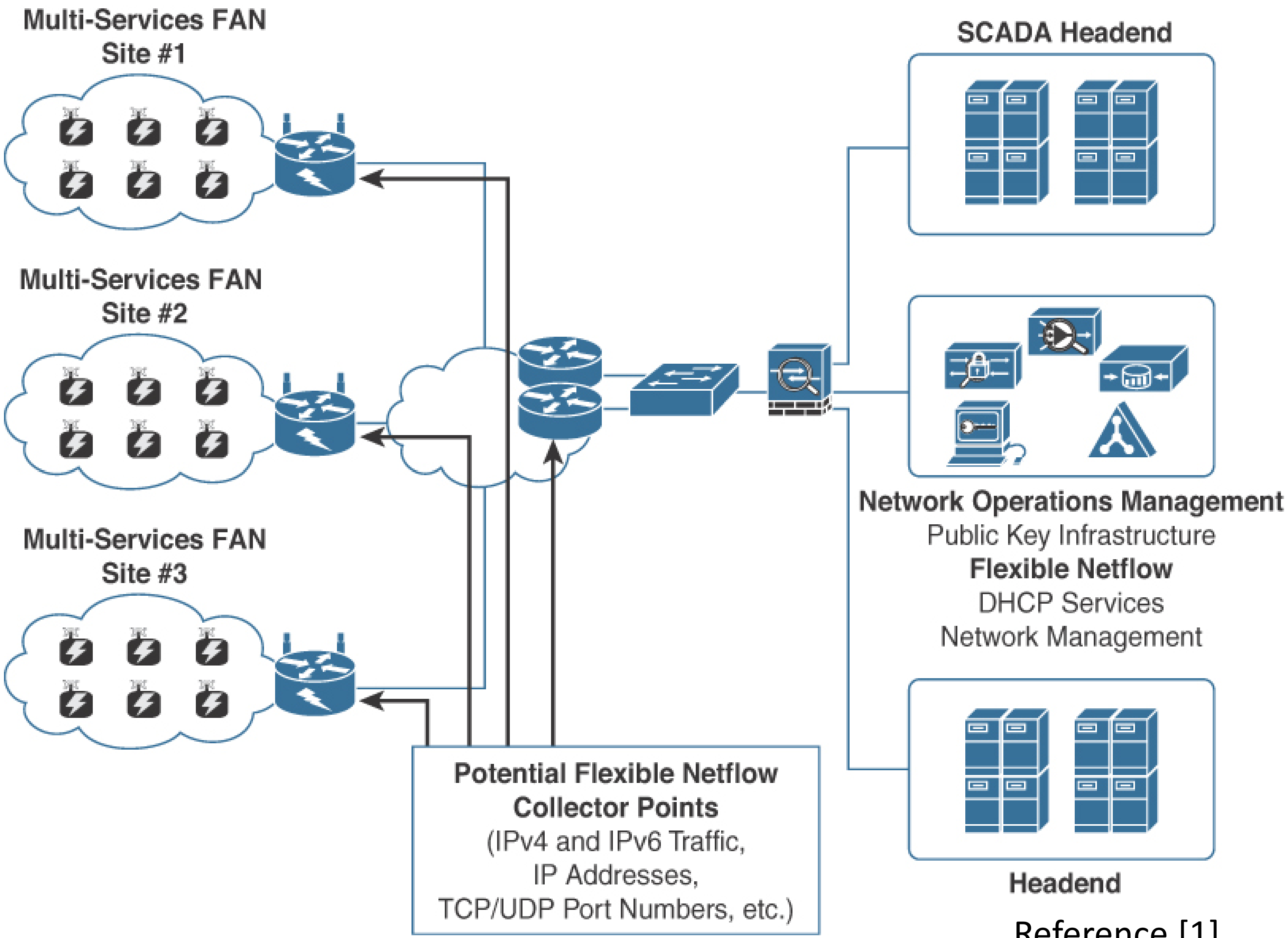
Flexible Netflow (FNF) in an IoT Deployment

- FNF provides
  - Flexibility, scalability, and aggregation of flow data
  - Ability to monitor a wide range of packet information and produce new information about network behavior
  - Enhanced network anomaly and security detection
  - User-configurable flow information for performing customized traffic identification and ability to focus and monitor specific network behavior
  - Convergence of multiple accounting technologies into one accounting mechanism

Reference [1]

Network Analytics example: SCADA

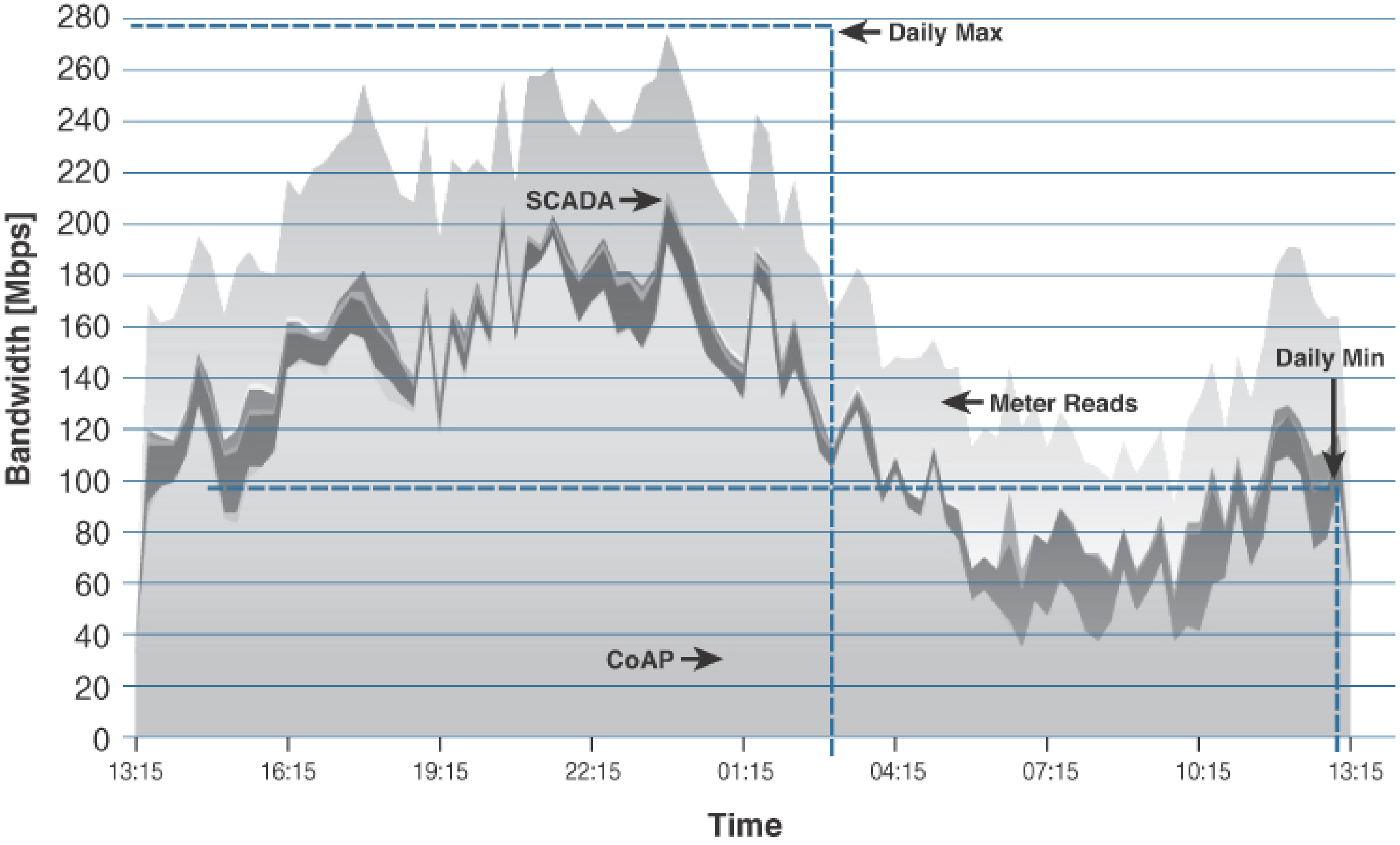
- Example: a Flexible Netflow System as a SCADA system for utility management in a smart grid
- SCADA = Supervisory Control And Data Acquisition
- Reference [1]


Multi-Services FAN Site #1
Multi-Services FAN Site #2
Multi-Services FAN Site #3
SCADA Headend
Network Operations Management
Public Key Infrastructure
Flexible Netflow
DHCP Services
Network Management
Headend
Potential Flexible Netflow Collector Points
(IPv4 and IPv6 Traffic,
IP Addresses,
TCP/UDP Port Numbers, etc.)

Reference [1]

Elements of an FNF System

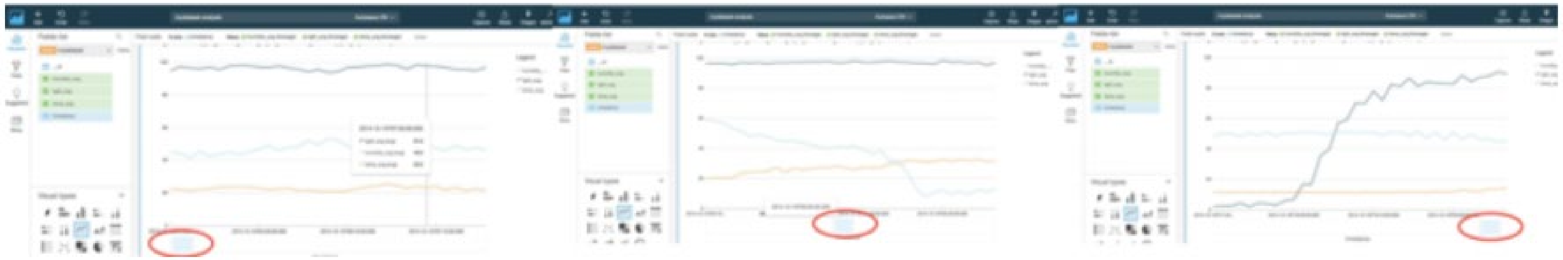
- Flow Monitor
- Flow Records
- Exporter
- Export Timers
- Export Formatting
- Servers for Collecting and Reporting
- Typical Traffic Report from a Smart Grid FAN

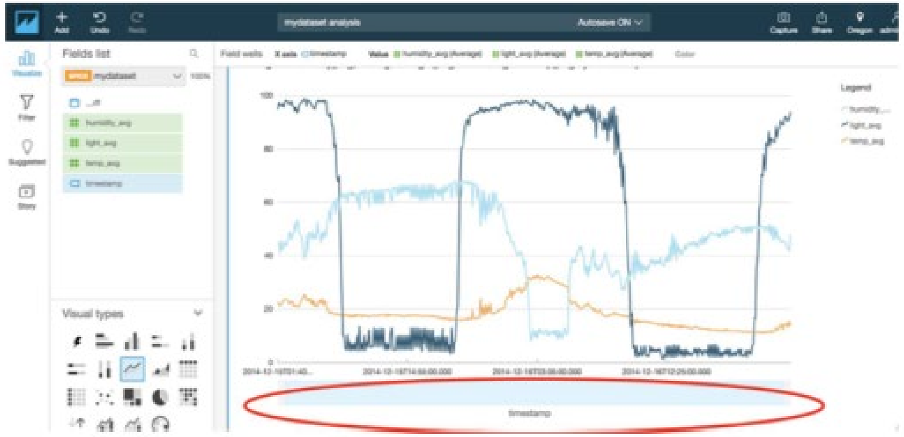


Reference [1]

AWS IoT Device Management and QuickSight

- AWS IoT framework also provides ways to manage and track behavior of networks of IoT devices [4]





Field wells X axis timestamp Value humidity_avg (Average) light_avg (Average) temp_avg (Average) Color

Summary

- IoT networks can produce vast volumes of data
- New technologies have developed for storage and analysis
- AI methods are regularly used for assessing data
- Both big data and edge analysis are appropriate for system elements
- Network analytics also becomes key for managing large systems
- Leading network and cloud vendors support large scale IoT system management with a variety of toolsets

References

[1] IoT Fundamentals: Networking Technologies, Protocols, and Use Cases for the Internet of Things, Salgueiro et al., Cisco Press, 2017

[2] https://aws.amazon.com/greengrass/ml/

[3] https://aws.amazon.com/emr/

[4] https://d1.awsstatic.com/IoT/User%20Guide%20PDFs/04_AWS_Mini-User_Guide_Analytics-and-Visualization_August2018.pdf