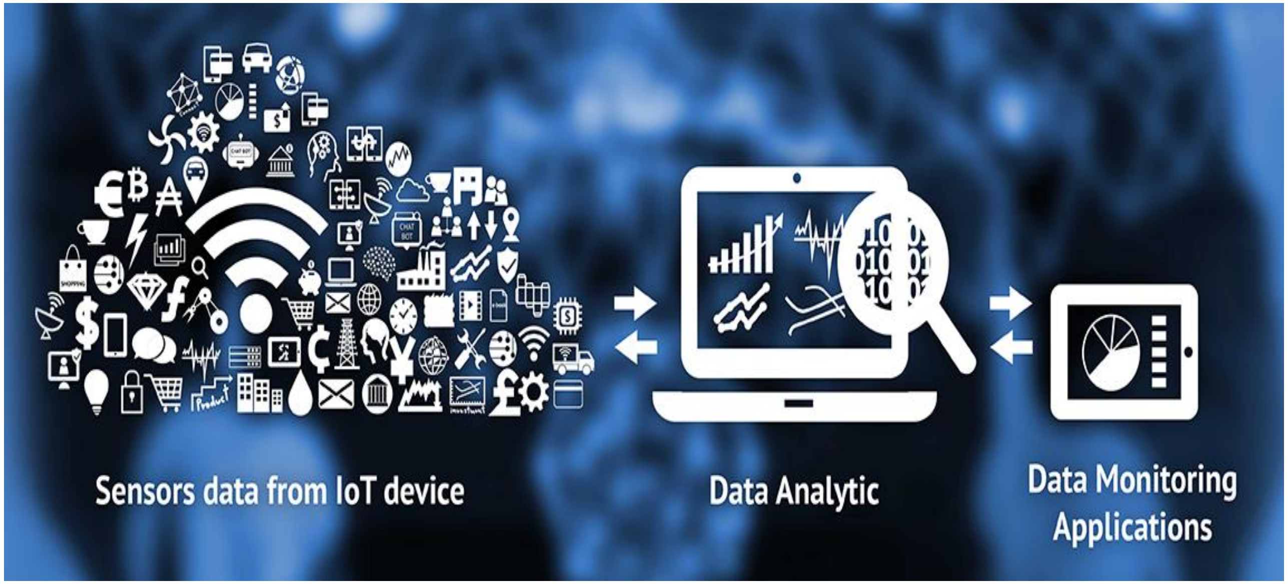
CS578: Internet of Things

Data Analytics in IoT



Dr. Manas Khatua

Assistant Professor

Dept. of CSE, IIT Guwahati

E-mail: manaskhatua@iitg.ac.in

Introduction to Data Analytics for IoT

- One of the biggest challenges in IoT:
  - Management of massive amounts of data generated by sensors.
- Few great examples
  - commercial aviation industry
  - utility industry
- Modern jet engines are fitted with thousands of sensors that generate a whopping 10GB of data per second
- A twin engine commercial aircraft with these engines operating on average 8 hours a day will generate over 500 TB of data daily, and this is just the data from the engines!

Commercial Jet Engine

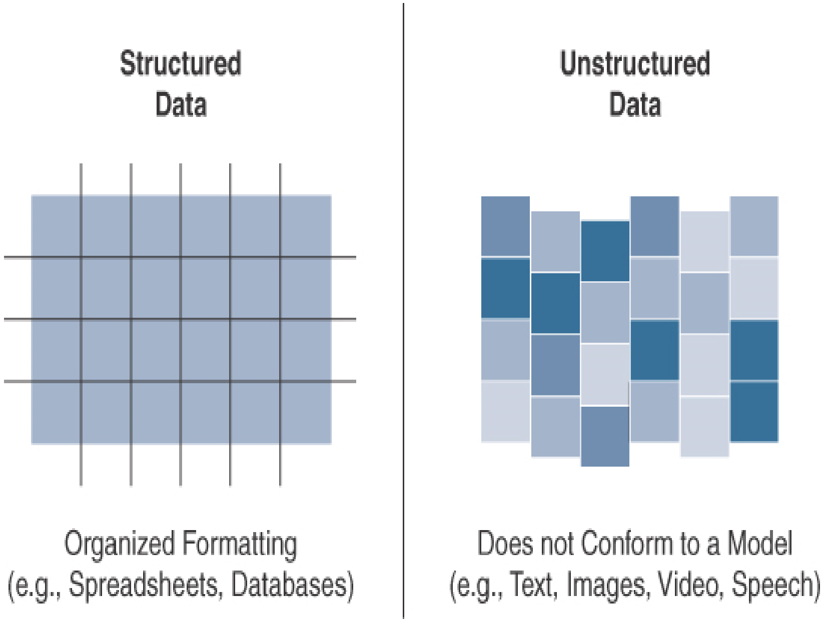
Structured v/s Unstructured Data

- Not all data is the same
- it can be categorized and thus analyzed in different ways.

- Structured data :
  - data follows a model/schema
  - defines data representation
  - e.g. Relational Database
  - easily formatted, stored, queried, and processed
    - has been core type of data used for making business decisions
    - Wide array of data analytics tools are available

- Unstructured data:
  - lacks a logical schema
  - Doesn't fit into predefined data model
  - e.g. text, speech, images, video

- Semi-structured data:
  - hybrid of structured and unstructured data
  - Not relational, but contains a certain schema
  - e.g. Email : fields are well defined, but body and attachments are unstructured

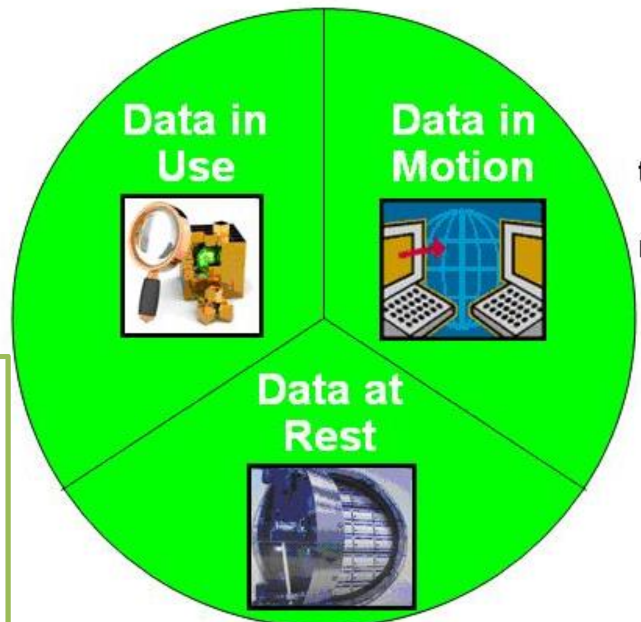


Data in Motion v/s at Rest v/s in Use

- Different states digital data can be in
  - in transit (data in motion)
  - being held/stored (data at rest)

- Data in motion is data that is currently traveling across a network or
- sitting in a computer's RAM ready to be read, updated, or processed.

Data in Use:
Active data under constant change stored physically in databases, data warehouses, spreadsheets etc.



Data in Motion:
Data that is traversing a network or temporarily residing in computer memory to be read or updated.

- Data being processed by one/more applications.
- data in the process of being generated, viewed, updated, appended, or erased.

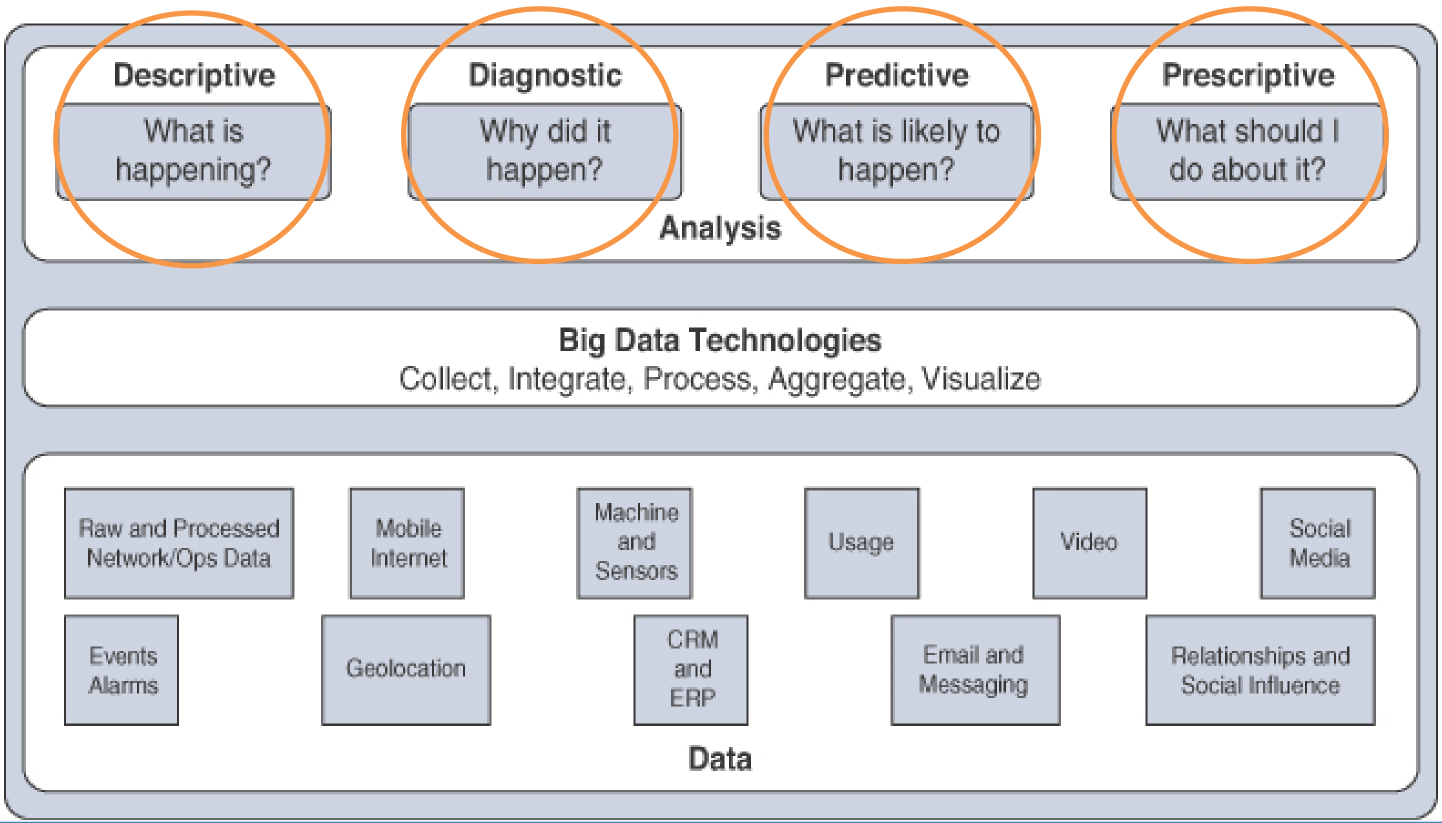
Data at Rest:
Inactive data stored physically in databases, data warehouses, spreadsheets, archives, tapes, off-site backups etc.

- Data at rest is typically in a stable state.
- It is not traveling within the system or network, and
- it is not being acted upon by any application or the CPU.

IoT Data Analytics Overview

- The true importance of IoT data from smart objects is realized only when
  - the analysis of the data leads to actionable business intelligence and insights.

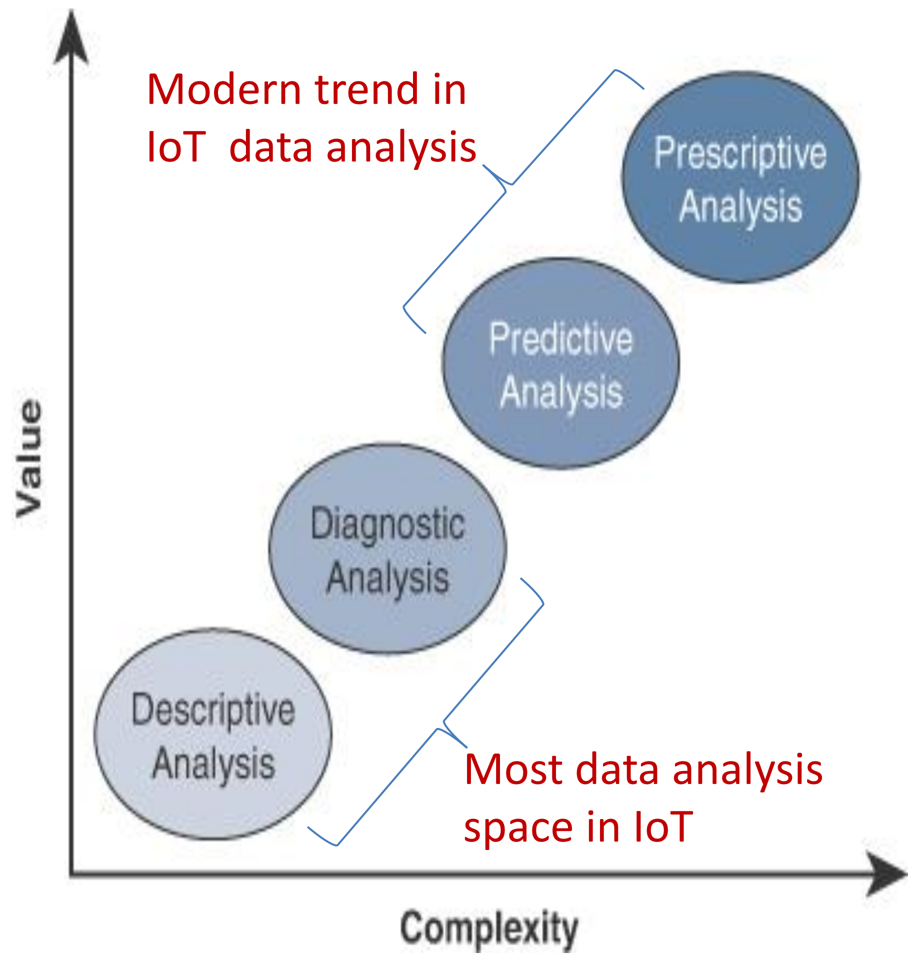
- Data analysis is typically broken down by
  - the types of results that are produced.



Cont…

- Descriptive
  - It tells you what is happening, either now or in the past.
    - e.g., thermometer in a truck engine reports temperature values every second.
- Diagnostic
  - It can provide the answer to “why” it has happened
    - e.g. why the truck engine failed
- Predictive
  - It aims to foretell problems or issues before they occur.
    - e.g., it could provide an estimate on the remaining life of the truck engine.
- Prescriptive
  - It goes a step beyond predictive and recommends solutions for upcoming problems.
    - e.g. it might calculate various alternatives to cost-effectively maintain our truck.

Application of Value and Complexity Factors to the Types of Data Analysis



IoT Data Analytics Challenges

- Traditional data analytics solutions are not always adequate
  - It typically employs a standard RDBMS and corresponding tools

- IoT data places two specific **challenges on relational database data**:
  - Scaling problems:
    - large number of smart objects continually send data,
    - relational databases grow incredibly large very quickly.
    - Results in performance issues which is costly to resolve
  - Volatility of data:
    - In RDBMS, schema is designed from the beginning,
    - changing the scheme later creates problem.
    - IoT data is volatile in the sense that the data model is likely to change and evolve over time.
    - A dynamic schema is often required.

- **Solution**: NoSQL database in used
  - does not use SQL to interact with the database
  - do not enforce a strict schema
  - support a complex, evolving data model
  - databases are inherently much more scalable

Cont…

- IoT brings **challenges to streaming and network analytics**
  - with the live streaming nature of its data, and
  - with managing data at the network level.
    - usually of a very high volume
    - real-time analysis of streaming data
      - Google, Microsoft, IBM, etc., have streaming analytics offerings
  - with the area of network data i.e. network analytics.
    - it can be challenging to ensure that the data flows are effectively managed, monitored, and secure.
      - Network analytics tools: Flexible NetFlow, IPFIX

Machine Learning

- How to make sense of the data that is generated in IoT?
  - by Machine Learning
  - ML is used to find the data relationships that will lead to new business insights

- **ML** is a part of a larger set of technologies commonly grouped under the term *artificial intelligence* (*AI*).

- **AI** includes any technology that allows a computing system to mimic human intelligence
  - e.g., an App that can help you find your parked car.
  - e.g., a GPS reading of your position at regular intervals calculates your speed.

- In more complex cases, static rules cannot be simply inserted into the program
  - because the programs require parameters that can change.
  - e.g., dictation program – it does not know your accent, tone, speed, and so on. You need to record a set of predetermined sentences to help the tool. This process is called machine learning.

Types of ML

- **Supervised**: All data is labeled and the algorithms learn to predict the output from the input data.
- **Unsupervised**: All data is unlabeled and the algorithms learn to inherent structure from the input data.

Supervised	Unsupervised	Semi-Supervised	Reinforcement
• Data has **known labels** or output	• Labels or output unknown • Focus on **finding patterns and gaining insight** from the data	• Labels or output known for **a subset of data** • A blend of supervised and unsupervised learning	• Focus on **making decisions** based on previous experience • Policy-making with feedback
• Insurance underwriting • Fraud detection	• Customer clustering • Association rule mining	• Medical predictions (where tests and expert diagnoses are expensive, and only part of the population receives them)	• Game AI • Complex decision problems • Reward systems

Few ML Algorithms

Machine Learning Algorithms *(sample)*

	Unsupervised	Supervised

Continuous	• Clustering & Dimensionality Reduction
	  ○ SVD
	  ○ PCA
	  ○ K-means
		• Regression
		  ○ Linear
		  ○ Polynomial
		• Decision Trees
		• Random Forests

Categorical	• Association Analysis
	  ○ Apriori
	  ○ FP-Growth
	• Hidden Markov Model
		• Classification
		  ○ KNN
		  ○ Trees
		  ○ Logistic Regression
		  ○ Naive-Bayes
		  ○ SVM

Examples from IoT

Supervised Learning

- Suppose you are training a system to recognize when there is a human in a mine tunnel.
  - sensor equipped with a basic camera can capture shapes
  - return them to a computing system.
  - hundreds or thousands of images are fed into the machine.
  - each image is labelled (human or nonhuman in this case)
  - An algorithm is used to determine common parameters and common differences between the images.
  - This process is called *training*.
  - Each new image is compared with "good images" of human as per training model
  - This process is called *classification*.
  - the machine should be able to recognize human shapes.

  - the learning process is not about classifying in two or more categories but about finding a correct value.
  - regression predicts numeric values, whereas classification predicts categories.

Cont...

Unsupervised Learning

- Suppose that you are processing IoT data from a factory manufacturing small engines. You know that about 0.1% of the produced engines on average need adjustments to prevent later defects.
- Your task is to identify them before they shipped away from the factory.
  - you can test each engine
  - record multiple parameters, such as sound, pressure, temperature of key parts, and so on.
  - Once data is recorded, you can graph these elements in relation to one another.
  - You can then input this data into a computer and use mathematical functions to find groups.
  - A standard function to operate this grouping, *K-means clustering*
  - Grouping the engines this way can quickly reveal several types of engines that all belong to the same category.
  - There will occasionally be an engine in the group that displays unusual characteristics
  - This is the engine that you send for manual evaluation
  - this determination is called *unsupervised learning*.

Applications of ML for IoT

It revolves around four major domains:

I. Monitoring
- ML can be used with monitoring to detect early failure conditions or to better evaluate the environment

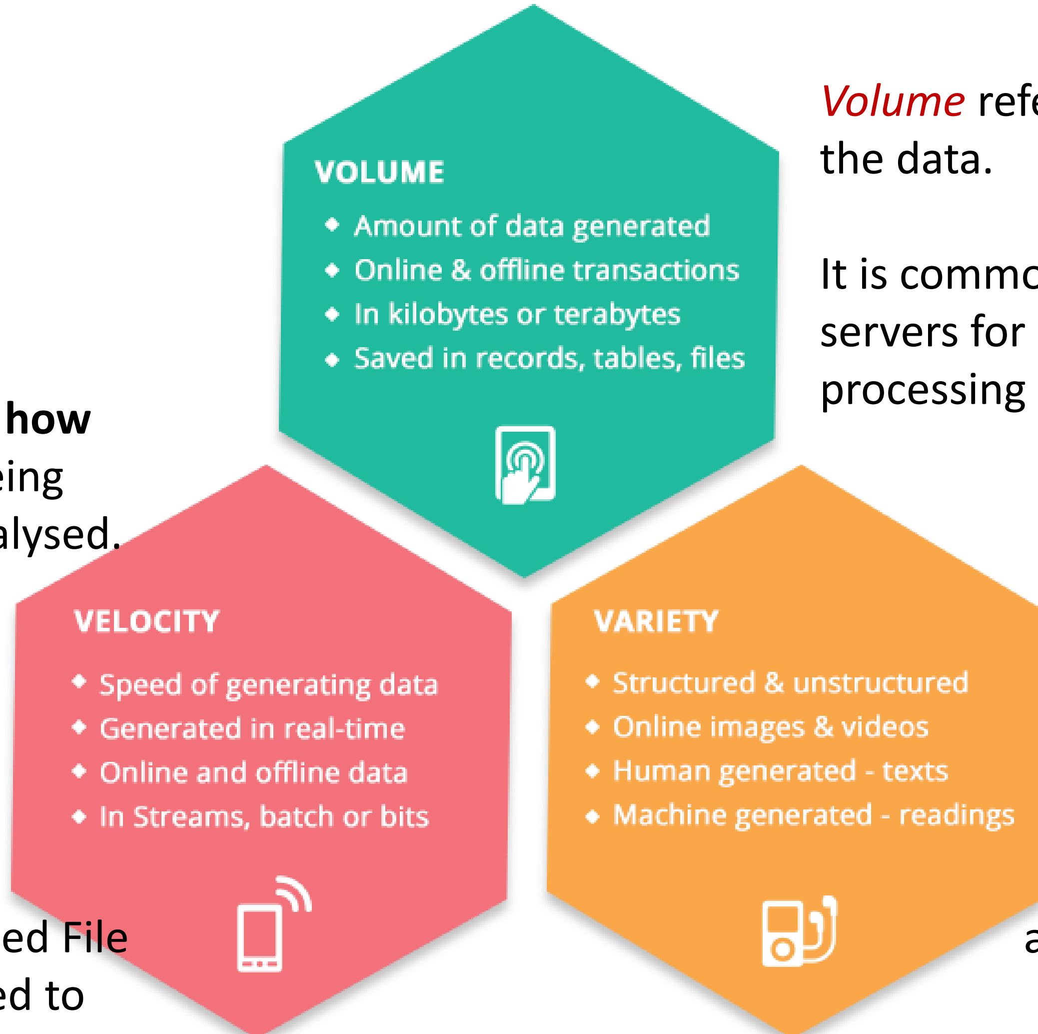
II. Behavior control
- Monitoring commonly works in conjunction with behavior control.
- When a given set of parameters reach a target threshold, monitoring functions generate an alarm OR would trigger a corrective action

III. Operations optimization
- The objective is not merely to pilot the operations but to improve the efficiency and the result of these operations.
- e.g., NN based system for a water purification plant in a smart city estimate the best chemical and stirring mix for a target air temperature

IV. Self-healing, self-optimizing
- The system becomes self-learning and self-optimizing.
- ML engine can be programmed to dynamically monitor and combine new parameters, and automatically deduce and implement new optimizations

What is Big Data?

- Industry looks to **three V**'s to categorize big data



Volume refers to the **scale** of the data.

It is common to see clusters of servers for storing and processing data

Velocity refers to **how quickly** data is being collected and analysed.

Variety refers to **different types** of data.

Hadoop is able to collect and store all three types – structured, unstructured, or semi-structured.

Hadoop Distributed File System is designed to process data very quickly.

Characteristics of Big Data

- can be defined by the sources and types of data
  - Machine data
    - generated by IoT devices and is typically unstructured data.
  - Transactional data
    - from the sources that produce data from transactions on these systems, and, have high volume and structured.
  - Social data
    - which are typically high volume and structured.
  - Enterprise data
    - data that is lower in volume and very structured.

Database Technologies

- Relational databases and historians are mature technologies
  - Relational databases, such as Oracle and Microsoft SQL, are good for transactional, or process, data.
  - Historians are optimized for time-series data from systems and processes

- Database technologies used in an IoT context.
  - NoSQL
    - It is not a specific database technology; rather, it is an umbrella term that encompasses several different types of databases.
    - Can quickly ingest rapidly changing data
    - Can be able to query and analyse data within the database itself
    - built to scale horizontally i.e. database can span to multiple hosts (so distributed)
    - Best fit for IoT data:
      - **Document stores:** stores semi-structured data, such as **XML** or **JSON**.
        - » allowing the database schema to change quickly
      - **Key-value stores:** stores associative arrays where a key is paired with a value.
        - » capable of handling indexing and persistence.
  - Massively Parallel Processing
    - built on the concept of the relational data warehouses
    - designed to allow for fast query processing; often have built-in analytic functions
    - designed in a scale-out architecture such that both data and processing are distributed across multiple systems
  - Hadoop

Hadoop

- Most popular choice as a data repository and processing engine

- Originally developed as a result of projects at Google and Yahoo!
  - original intent was to index millions of websites and quickly return search results for open source search engines.

- Initially, the project had two key elements:
  - **Hadoop Distributed File System (HDFS):** A system for storing data across multiple nodes
  - **MapReduce:** A distributed processing engine that splits a large task into smaller ones that can be run in parallel

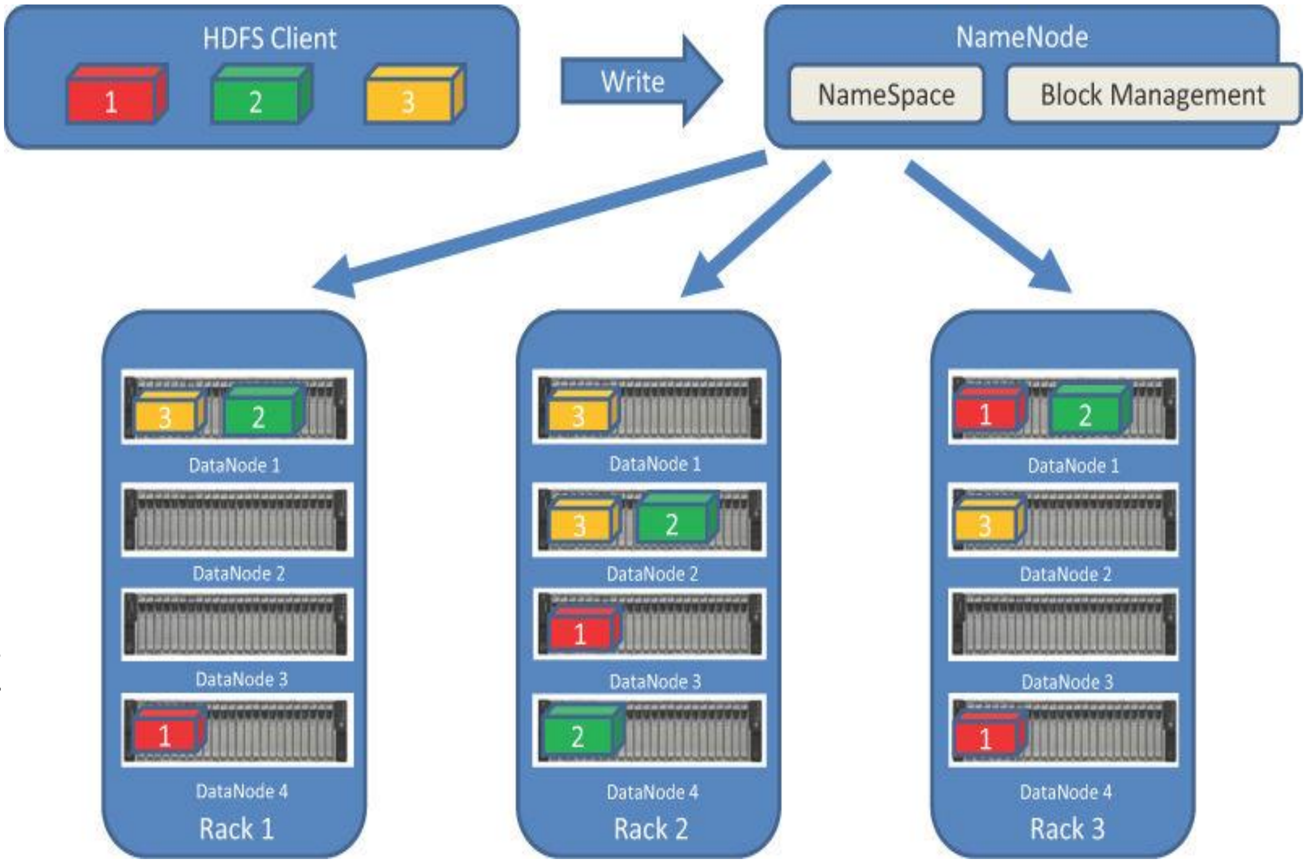
- Hadoop relies on a scale-out architecture
- Both MapReduce and HDFS
  - take advantage of this distributed architecture to store and process massive amounts of data
  - leverages local processing, memory, and storage from all nodes in the cluster

Cont…

- For HDFS, this capability is handled by specialized nodes in the cluster – NameNode and DataNode

- NameNode coordinate where the data is stored, and maintain a map of where each block of data is stored and where it is replicated.

- DataNodes are the servers where the data is stored at the direction of the NameNode.



Hadoop Ecosystem

- Hadoop plays an increasingly big role in the collection, storage, and processing of IoT data

- Hadoop Ecosystem comprises of more than 100 software projects under the Hadoop umbrella
  - Capabale of every element in the data lifecycle, from data collection, to storage, to processing, to analysis and to visualization

- Several of these packages
  - **Apache Kafka**
  - **Apache Spark**
  - **Apache Storm**
  - **Apache Flink**
  - **Lambda Architecture**

Edge Streaming Analytics

- In the world of IoT vast quantities of data are generated on the fly
  - Often they are time sensitive i.e. needs immediate attention,
  - waiting for deep analysis in the cloud simply isn't possible.
  - e.g., automobile racing industry
    - Formula One racing car has 150-200 sensors that generate more than 1000 data points per second
    - enormous insights leading to better race results can be gained by analyzing data on the fly
- Big Data tools like Hadoop and MapReduce are not suited for real-time analysis
  - because of distance from the IoT endpoints and the network bandwidth requirement
- **Streaming analytics** allows you to continually monitor and assess data in real-time so that you can adjust or fine-tune your predictions as the race progresses.
- In IoT, streaming analytics is performed at the edge (either at the sensors themselves or very close to them such as gateway)
- The edge isn't in just one place. The edge is highly distributed.
- Does the streaming analytics replaces big data analytics in the cloud?
  - Not at all.
  - **Big data** analytics is focused on large quantities of **data at rest**, **edge analytics** continually processes streaming flows of **data in motion**.

Key Values of Edge Streaming Analytics

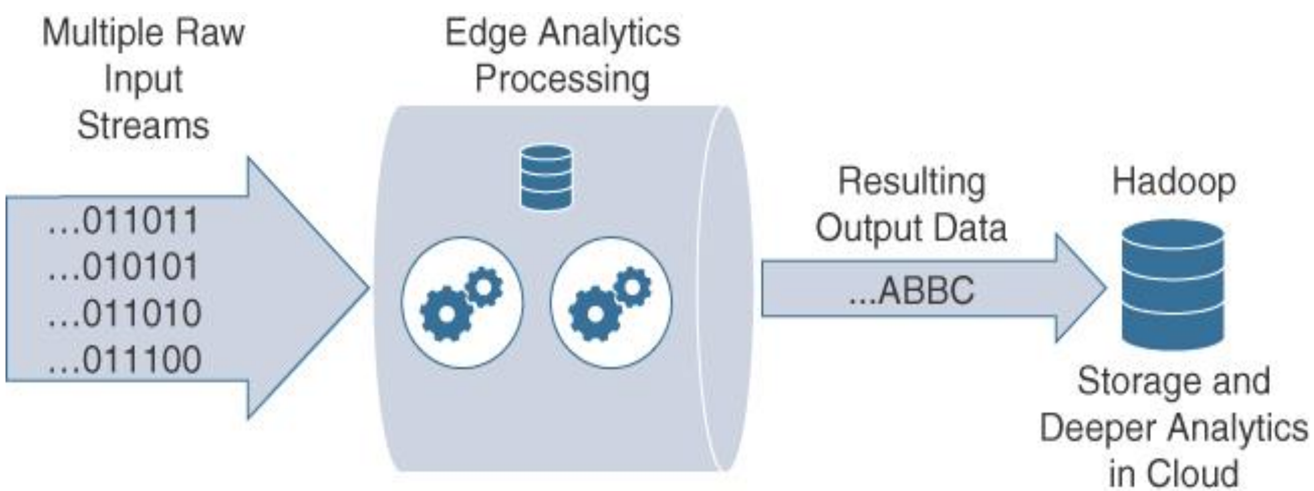
- Reducing data at the edge
  - Passing all data to the cloud is inefficient and is unnecessarily expensive in terms of bandwidth and network infrastructure.

- Analysis and response at the edge
  - Some data is useful only at the edge and for small window of time
  - e.g., Roadway sensors combined with GPS wayfinding apps may tell a driver to avoid a certain highway due to traffic. This data is valuable for only a small window of time.

- Time sensitivity
  - When timely response to data is required, passing data to the cloud for future processing results in unacceptable latency.

Edge Analytics Core Functions

- Raw input data
  - This is the raw data coming from the sensors into the analytics processing unit.
- Analytics processing unit (APU)
  - The APU filters and combines (or separates) data streams, organizes them by time windows, and performs various analytical functions.
- Output streams
  - The data that is output is organized into insightful streams and passed on for storage and further processing in the cloud.



Network Analytics

- This form of analytics extremely important in managing IoT systems
- Data analytics: concerned with finding patterns in the data generated by endpoints
- Network analytics: concerned with discovering patterns in the communication flows
  - It is network-based analytics
  - power to analyze details of communications patterns made by protocols
  - correlate this pattern across the network
  - allows to understand what should be considered normal behavior in a network

Benefits

Benefits of Network Analytics:

- Offer capabilities to cope with capacity planning for scalable IoT deployment
- Security monitoring in order to detect abnormal traffic volume and patterns
  - e.g. an unusual traffic spike for a normally quiet protocol
  - for both centralized or distributed architectures
- Network traffic monitoring and profiling
- Application traffic monitoring and profiling
- Capacity planning
- Security analysis
- Accounting
- Data warehousing and data mining

Challenges

Challenges with deploying flow analytics tools in an IoT network

- Flow analysis at the gateway is not possible with all IoT systems
  - LoRaWAN gateways simply forward MAC-layer sensor traffic to the centralized LoRaWAN network server, which means flow analysis (based on Layer 3) is not possible at this point.
  - A similar problem is encountered when using an MQTT server that sends data through an IoT broker

- Traffic flows are processed in places that might not support flow analytics, and visibility is thus lost.

- IPv4 and IPv6 native interfaces sometimes need to inspect inside VPN tunnels, which may impact the router's performance.

- Additional network management traffic is generated by analytics reporting devices

Thanks!