1. Data and Analytics for IoT: An Introduction to Data Analytics for IoT
2. Structured Versus Unstructured Data
3. Machine Learning
4. Big Data Analytics Tools and Technology
5. Edge Streaming Analytics
6. Network Analytics
7. Data Analytics Challenges
8. Data Acquiring
9. Organizing in IoT/M2M

1. Data Analytics for IoT

The real value of the **Internet of Things (IoT)** is not just in connecting devices (things) but also in the **data** produced by these devices. This data can reveal valuable business insights and enable new services.

IoT creates massive amounts of data from sensors, which can be challenging to manage both in terms of transport and data handling. For example, modern jet engines are equipped with thousands of sensors generating around **10GB of data per second**. Analyzing this vast amount of data efficiently falls under **data analytics**.

An Introduction to Data Analytics for IoT

Data in IoT can be categorized in different ways for analysis. The way data is classified determines the tools and processing methods applied. Two important categorizations are:

1. **Structured vs. Unstructured Data**
2. **Data in Motion vs. Data at Rest**

Structured Data

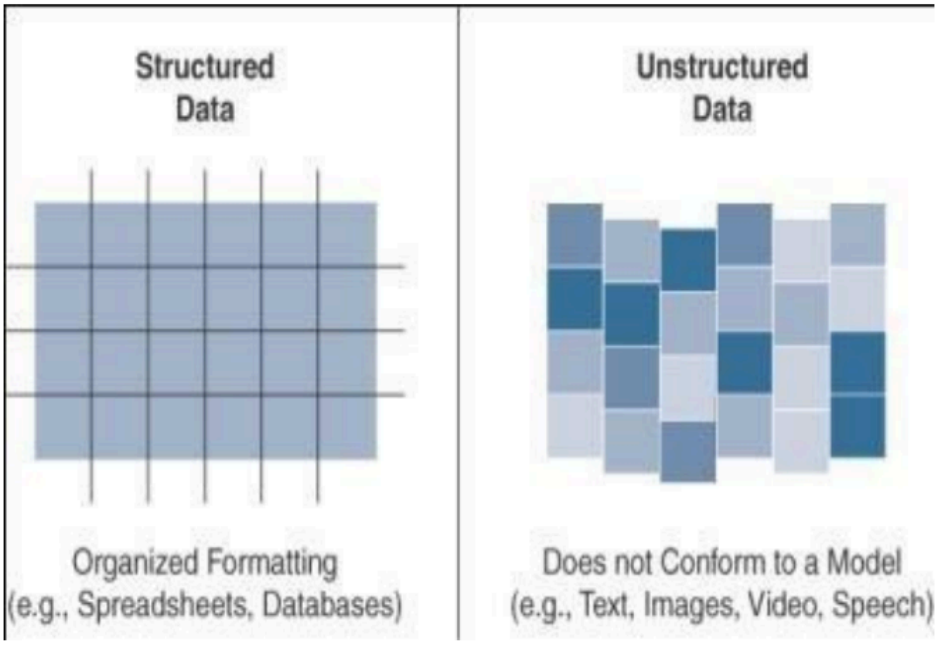
- **Structured data** is organized in a specific model or schema, which makes it easy to work with using traditional database systems (RDBMS).
- It's often in tabular form, like spreadsheets, where data is organized in rows and columns.
- Examples include banking transactions, invoices, and log files. **IoT sensor data** like temperature, pressure, and humidity is typically structured, following a known format.
- **Structured data** is easy to store, process, and query, and can be handled by a wide range of tools like **Excel, Tableau**, or custom scripts.

Unstructured Data

- **Unstructured data** lacks a predefined structure and doesn't fit into a traditional database system.
- Examples include **text, images, speech, and video**.

- A significant portion (around **80%**) of business data is unstructured. Therefore, new analytics methods like **cognitive computing** and **machine learning** are used to process such data.
    - **Machine learning** applications, like **Natural Language Processing (NLP)**, decode speech, and **image recognition** helps extract information from images and videos.

Structured and Unstructured Data in IoT

- IoT networks generate both **structured** and **unstructured data**.
    - **Structured data** is easier to process and manage due to its well-organized nature.
    - **Unstructured data** is more complex and requires specialized analytics tools.
    - 

Data in Motion vs. Data at Rest

- **Data in Motion** refers to data that is actively being transferred over the network, like in client-server exchanges (e.g., web browsing, file transfers).
    - **IoT smart objects** often generate **data in motion**, which may be processed at the **edge** of the network using **fog computing**.
    - **Edge processing** allows data to be filtered or deleted or sent for further processing in a data center or fog node.
- **Data at Rest** is stored data, like on hard drives, storage arrays, or cloud storage.

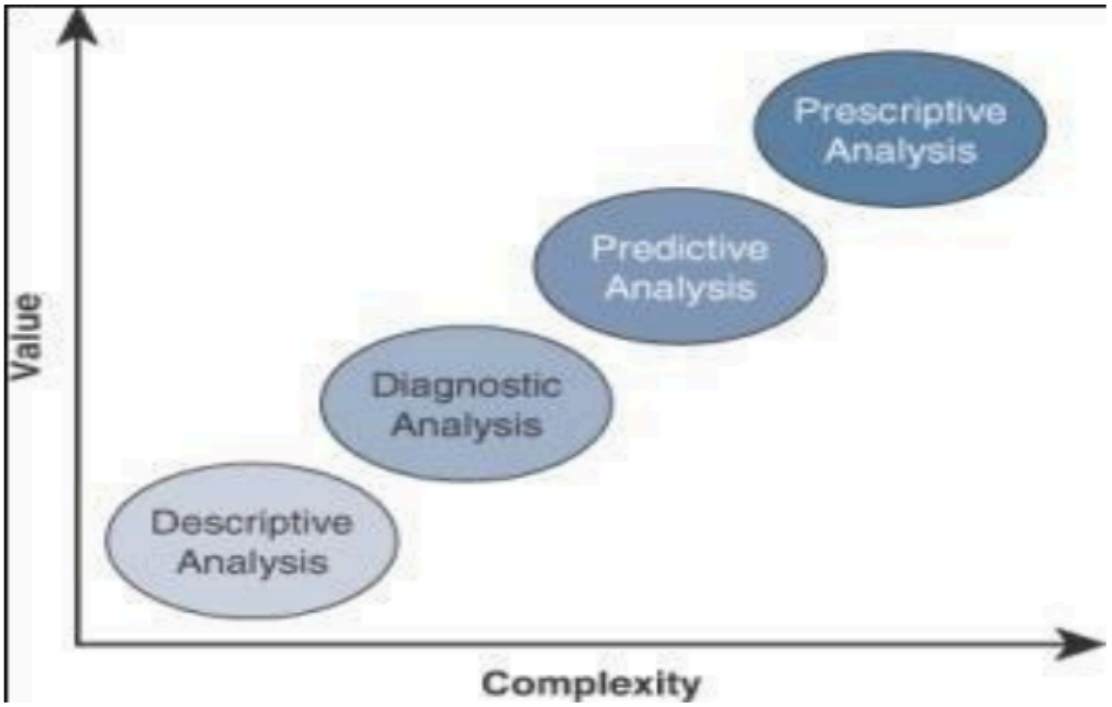
- **IoT data** that is not currently being transferred but is stored for future access, often found in **IoT brokers** or **data centers**.
- **Hadoop** is often used for both storing and processing data.

IoT Data Analytics Overview

The **real value** of IoT data is realized when it is analyzed to produce **business intelligence** and **actionable insights**. The data analysis process is typically categorized into four types based on the kind of results it provides:

Types of Data Analysis Results

1. **Descriptive Analysis**:
   - This type of analysis tells you **what** is happening or has happened.
   - For example, a thermometer in a truck engine reports the **current temperature**. Descriptive analysis can tell you the **current operating condition** of the truck engine, such as identifying if it's running hot or cold.
2. **Diagnostic Analysis**:
   - This analysis answers why something happened.
   - For example, if a truck engine failed, diagnostic analysis might reveal that the engine temperature was too high, causing it to overheat.
3. **Predictive Analysis**:
   - This analysis predicts **what will happen in the future**.
   - Using historical temperature data, predictive analysis might predict the **remaining life** of engine components or predict future failures based on rising temperature trends. It helps prevent problems before they occur.
4. **Prescriptive Analysis**:
   - This type of analysis not only predicts what might happen but also recommends **solutions**.
   - For example, if the truck engine is predicted to overheat, prescriptive analysis could recommend **cost-effective solutions** like more frequent oil changes or upgrading the engine's cooling system.
   - **Prescriptive analysis** looks at multiple factors and provides a variety of potential solutions.



Note: Predictive and prescriptive analyses are more complex and resource-intensive but offer greater value than descriptive and diagnostic analyses.

IoT Data Analytics Challenges

Using traditional **Relational Database Management Systems (RDBMS)** in IoT systems presents several challenges:

1. **Scaling Problems**:
    - IoT generates large volumes of data, making RDBMS difficult to scale. This results in performance issues and high costs for expanding hardware or changing system architecture.
2. **Volatility of Data**:
    - IoT data often changes over time, causing schema changes or data to be inconsistent, which makes it difficult to manage with traditional database systems.

Machine Learning (ML) in IoT

- **Importance of ML in IoT**: In the Internet of Things (IoT), data generated by smart devices needs to be analyzed to make intelligent decisions. Doing this manually is slow and inefficient. Therefore, machines with ML are used to quickly process data and take immediate action when certain conditions are met.

    - **Example**: In self-driving cars, ML helps in recognizing abnormal patterns and making decisions without human intervention.

What is Machine Learning (ML)?

- **Part of AI**: ML is a subset of Artificial Intelligence (AI), which includes any technology that allows a computer system to mimic human intelligence. It uses techniques like advanced logic or simple rules like "if-then-else" decision-making.
    - **Example**: An app that helps you find your parked car uses ML to determine when you're driving and when you've parked, based on your car's speed and location.
- **When Rules Aren't Enough**: In some cases, static rules can't solve the problem, especially when there are changing parameters or complex conditions. This is where ML is needed.
    - **Example**: A voice recognition system needs to adjust to your specific accent, tone, and speed by learning from your spoken data.

Two Main Categories of ML:

1. **Supervised Learning**
    - In supervised learning, the machine is trained using data that already has the correct answer.
        - **Example**: Training a system to recognize humans in images. You provide images labeled as "human" or "non-human" (training set). The system learns to identify human shapes by comparing new images to the known ones.
    - The system uses an algorithm to analyze the images at a pixel level, identifying common patterns and differences.
    - **Classification**: The machine learns to classify images as "human" or "non-human".
    - After training, the machine is tested with new, unlabeled images to check if it can correctly recognize humans. This is called a validation test.
    - **Regression**: Sometimes, instead of categorizing things, the goal is to predict a value. For example, predicting the speed of oil flow based on pipe size and other factors.
        - **Supervised learning** can also predict numerical values using regression.
2. **Unsupervised Learning**
    - In some cases, there's no labeled data available, and the machine must find patterns by itself. This is where unsupervised learning comes in.

    - **Example**: In a factory that makes small engines, you want to detect which engines might need adjustments before they are shipped. There is no direct label of "good" or "bad", but you can measure factors like sound, pressure, and temperature.
    - The machine looks at data like temperature vs. pressure and groups engines that behave similarly.
    - **K-means Clustering**: This method groups similar data points (like engines) together. For instance, engines that make similar sounds or have similar temperatures are grouped into categories.
    - If an engine behaves differently (e.g., outside the usual temperature range), it is flagged for further inspection.
    - **Unsupervised learning** works by finding patterns or clusters in data, without needing pre-defined labels or categories.

Key Differences between Supervised and Unsupervised Learning:

- **Supervised Learning**: The machine is given labeled data to learn from. It uses this data to make predictions or classifications (e.g., recognizing human images or predicting values like oil flow speed).
- **Unsupervised Learning**: The machine is not given labels. It identifies patterns or clusters by analyzing the data, such as grouping similar engines in a factory or detecting unusual behavior.

In Simple Terms:

- **Supervised Learning**: Learn from known examples (labeled data) to classify or predict future data.
- **Unsupervised Learning**: Learn from data without labels to find hidden patterns or groupings.

Neural Networks:

1. **Computing Power**: Processing data with many dimensions (features) takes a lot of computing power. Also, it's hard to decide which parameters (features) should be used for the task or what combinations should raise alarms.
2. **Supervised Learning**: This method is very effective when the dataset is large. Larger datasets generally improve prediction accuracy. But, training machines can be expensive and complicated.
3. **Recognition Challenges**: It's easy for humans to distinguish between a person and a car, but harder to tell apart similar objects, like a human from another mammal, or a pickup truck from a van. This requires more than just shape recognition.
4. **How Neural Networks Work**: Neural networks mimic the human brain. When you see an object, different parts of your brain are activated to recognize

various features like color, shapes, movements, etc., and combine them to conclude what the object is. In neural networks, information is processed by different "units" or layers, where each unit is responsible for analyzing a certain aspect of the data.

5. **Deep Learning**: Information in neural networks is divided into components, each with a weight. These weights are compared to classify information. When the result of one layer is passed to another layer, this process is called **deep learning**. The deeper the network (more layers), the more detailed and accurate the data processing becomes.

Machine Learning (ML):

1. **Types of Learning**:
    - **Local Learning**: Data is processed locally (on a sensor or device itself).
    - **Remote Learning**: Data is collected locally but processed in a centralized unit (like a data center or cloud).
2. **Applications of ML in IoT**:
    - **Monitoring**: Smart devices monitor their environment (e.g., temperature, humidity, gas levels).
    - **Behavior Control**: If certain conditions (like temperature) exceed a set threshold, an alarm is triggered. In more advanced systems, corrective actions are automatically taken, like adjusting air flow or stopping machinery.
    - **Operations Optimization**: Analyzing data helps improve processes, like finding the best chemical to use in a water purification plant.
    - **Self-Healing and Self-Optimizing**: ML systems can learn from data and automatically improve themselves or detect defects early. The system adapts to optimize operations.

Predictive Analytics:

1. **Predictive Analytics with IoT**: When data from multiple systems is combined, predictions can be made. For instance, sensors on trains can measure multiple parameters like weight and speed, which helps predict when maintenance might be needed. This keeps systems running safely and efficiently.

Big Data Analytics Tools and Technology:

1. **The Three Vs of Big Data**:
    - **Velocity**: Refers to how fast data is being collected and processed. For instance, IoT devices generate data rapidly, so systems must handle fast data ingestion.
    - **Variety**: Refers to the different types of data (structured, semi-structured, unstructured). Big data systems like Hadoop can store and process all these types of data.
    - **Volume**: Refers to the large amount of data (from gigabytes to petabytes or even exabytes).
2. **Types of Data in Big Data**:
    - **Machine Data**: Unstructured data from IoT devices.
    - **Transactional Data**: Structured, high-volume data from transactions.
    - **Social Data**: Structured, high-volume data from social media.
    - **Enterprise Data**: Structured, low-volume data from business systems.

Hadoop:

1. **Hadoop**: It's a popular system for storing and processing large amounts of data. It was developed by Google and Yahoo! to index websites. It has two key components:
    - **Hadoop Distributed File System (HDFS)**: A system that stores data across multiple computers.
    - **MapReduce**: A processing engine that splits large tasks into smaller ones and runs them in parallel.

Edge Analytics:

1. **Edge Analytics**: Data is processed closer to the source (at the "edge" of the network) rather than in a centralized system. This helps in real-time processing.
    - **Three Stages of Edge Analytics**:
        1. Raw Input Data
        2. Analytics Processing Unit (APU) - where the data is filtered, transformed, and patterns are matched.
        3. Output Streams - processed data is sent out for decision-making.

Distributed Analytics Systems:

1. **Analytics Locations**: Data can be analyzed at different stages:
    - **Edge**: Close to where the data is generated.
    - **Fog**: At intermediate levels (between the edge and cloud).
    - **Cloud**: Centralized data centers. Fog computing helps look at data from a wider perspective, which can give better insights, like when analyzing pressure and temperature on an oil rig.

IoT Data Analytics Challenges:

1. **Scaling Problems with Traditional Databases**: IoT data grows quickly, and relational databases (like SQL) can't scale fast enough. This leads to performance issues.
2. **Volatility of Data**: IoT data is dynamic, meaning it changes often. Relational databases struggle with this because they need a fixed schema (structure). IoT requires databases that can adapt and change, like **NoSQL** databases.

Network Analytics:

1. **Network Analytics**: With IoT devices constantly communicating, managing the flow of data is challenging. Tools like **Flexible NetFlow** and **IPFIX** help monitor data flow and detect issues in IoT networks.

Summary:

- Neural networks are inspired by how the brain works, and deep learning improves the accuracy of tasks by using more layers.
- Machine learning helps IoT devices to monitor environments, take corrective actions, optimize operations, and self-improve.
- Predictive analytics allows systems to anticipate failures and perform maintenance ahead of time.
- Big data is defined by velocity (speed), variety (types of data), and volume (amount of data), with tools like Hadoop helping to manage and process large datasets.
- Edge and fog analytics help in real-time data processing, allowing smarter decision-making in IoT systems.
- Traditional relational databases face challenges with the vast and changing IoT data, which is better handled by NoSQL databases.

Data Acquiring

1. Data Generation

- Data is created by devices and sent to the Internet through gateways.
- **Types of Data Generation:**
    - **Passive Devices:** Don't have their own power source; rely on external power. Examples: RFID tags, ATM debit cards, barcodes.
    - **Active Devices:** Have their own power source and are more advanced. Examples: Active RFID, streetlight sensors.
    - **Event Data:** Generated only when a specific event occurs. Example: Security cameras detecting an intrusion.
    - **Device Real-Time Data:** Data sent instantly, such as ATM transactions.
    - **Event-Driven Data:** Data generated only once after an event or command. Example: A device sends status updates when asked.

2. Data Acquisition

- Gathering data from IoT (Internet of Things) devices.
- Data is collected through an application which interacts with devices.
- Devices can be programmed to send data at specific intervals or only on demand.
- Data might be filtered, managed, or modified at the gateway before sending to the system.

3. Data Validation

- Ensures that collected data is accurate and consistent.
- Validation software checks if data falls within expected ranges and isn't corrupted.
- Invalid data is filtered out at gateways or even at the devices themselves to save resources.

4. Data Categorization for Storage

- Valid and useful data is stored based on usage:
    - **Raw Data:** Stored if it needs future processing or audits.
    - **Processed Data and Results:** Stored for quick access and reporting.
    - **Streaming Data:** Processed and stored in real-time.
- Huge volumes of data (Big Data) are stored on servers, warehouses, or the cloud.

5. Assembly Software for Events

- Devices create "events" when certain conditions are met (e.g., a temperature sensor triggers an event if it exceeds a limit).
- Events are assigned IDs and timestamps, which are assembled into meaningful information by software.

6. Data Store

- A **data store** is where all the data is kept, such as databases, spreadsheets, or files.
- Some data stores are distributed across multiple systems (e.g., Apache Cassandra).
- The data stored is used for reporting, analytics, or processing.

7. Data Center Management

- Data centers are facilities with powerful computers, servers, and backup systems to store and manage data securely.
- They ensure:
    - Data security.
    - Backup and recovery.
    - Maintenance of an optimal environment (e.g., cooling systems, dust-free areas).

8. Server Management

- Servers must run efficiently 24/7.
- Responsibilities include:
    - Quick issue resolution.
    - Regular updates and maintenance.
    - Ensuring high security and protecting against cyber threats.

9. Spatial Storage

- This is storage used for location-based data, such as tracking goods with RFID tags.
- Spatial databases handle data about geometric objects like maps, parking spaces, or city layouts.
- They allow operations like calculating distances, measuring areas, or checking relationships between locations.