Data Analytics for IoT

An Introduction to Data Analytics for IoT

- The ***real value of IoT*** is not just in connecting things but rather in the *data produced by those things*, the *new services you can enable* via those connected things, and the business insights that the data can reveal.

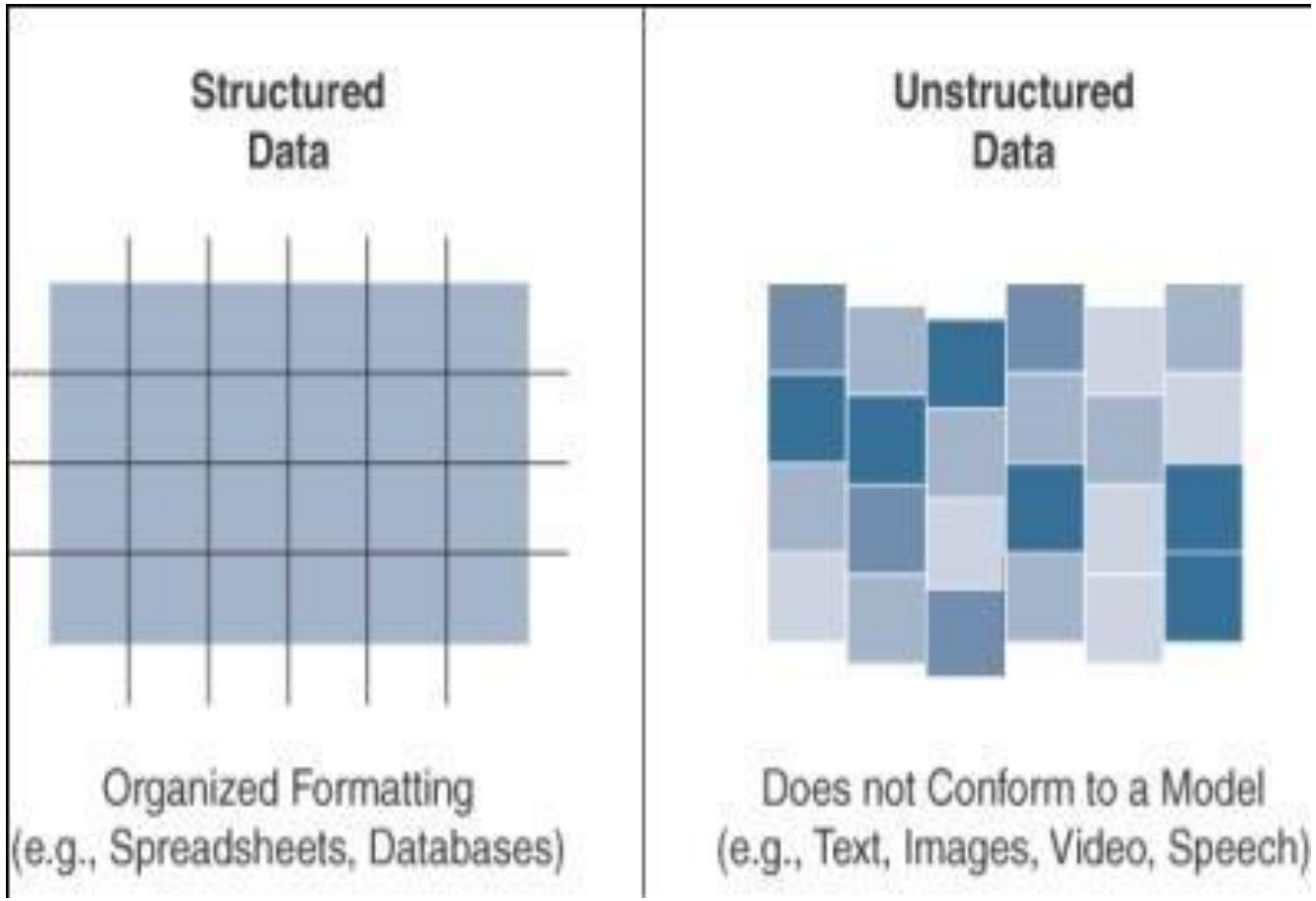
- In the world of IoT, the creation of massive amounts of data from sensors is common and one of the biggest challenges— not only from a ***transport perspective*** **but also from a *data management* standpoint**

- ***Modern jet engines*** are fitted with thousands of sensors that generate a whopping ***10GB of data per second***

- *Analyzing this amount of data in the most efficient manner possible falls under the umbrella of data analytics*

- Not all data is the same; it can be categorized and thus analyzed in different ways.
- Depending on how data is categorized, various data analytics tools and processing methods can be applied.
- **Two important categorizations from an IoT perspective are**

whether the data is structured or unstructured **and** ***whether it is in motion or at rest*****.**


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

Structured Data

- Structured data and unstructured data are important classifications as they typically require *different toolsets* from a data analytics perspective
- **Structured data means** that the data follows a model or schema that defines *how the data is represented or organized, meaning it fits well with a traditional relational database management system (RDBMS).*
- In many cases you will find structured data in a simple ***tabular form***—for example, a ***spreadsheet*** *where data occupies a specific cell and can be explicitly defined and referenced*

Structured Data

- Structured data can be found in most **computing systems** and includes everything from banking transaction and invoices to computer log files and router configurations.
- IoT sensor data often uses **structured values,** such as ***temperature, pressure, humidity, and so on, which are all sent in a known format.***
- Structured data is *easily formatted, stored, queried, and processed*
- Because of the highly organizational format of structured data, a wide array of **data analytics tools** are readily available for processing this type of data.
- From custom scripts to commercial software like **Microsoft Excel** and **Tableau**

Unstructured Data

- **Unstructured data lacks a logical schema** for understanding and decoding the data through traditional programming means.
- **Examples** of this data type include **text, speech, images, and video.**
- As a general rule, **any data that does not fit neatly into a predefined data model is classified as unstructured data**

Unstructured Data

- According to some estimates, ***around 80% of a business's data is unstructured.***
- Because of this fact, **data analytics methods** that can be **applied to unstructured data**, such as cognitive computing and machine learning, are deservedly garnering a lot of attention.
- With machine learning applications, such as natural language processing (NLP), you can decode speech.
- With image/facial recognition applications, you can extract critical information from still images and video

Structured and Unstructured Data

- **Smart objects in IoT networks *generate both structured and unstructured data.***
- Structured data is more easily managed and processed due to its well-defined organization.
- On the other hand, unstructured data can be harder to deal with and typically requires very different analytics tools for processing the data

Data in Motion Versus Data at Rest

- Data in IoT networks is either in transit (“data in motion”) or being held or stored (“data 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.***

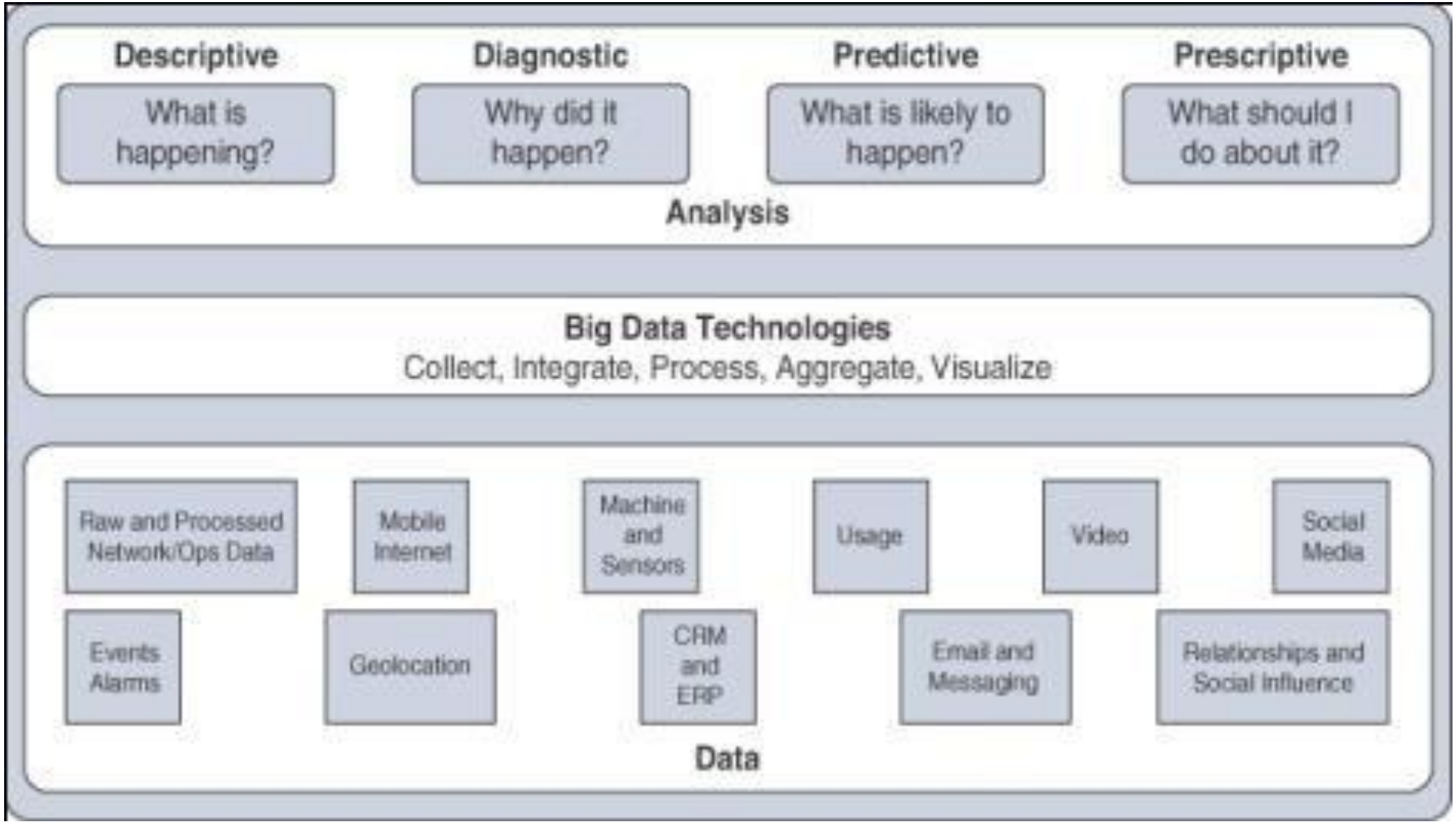
Data in Motion Versus Data at Rest

- **The data from smart objects is considered data in motion** as it passes through the network in route to its final destination.This is often ***processed at the edge, using fog computing.***
- When data is processed at the edge, it may be filtered and deleted or forwarded on for further processing and possible storage at a fog node or in the data center.Data does not come to rest at the edge.
- When data arrives at the data center, it is possible to process it in real-time, just like at the edge, while it is still in motion. **Tools** with this sort of capability, are **Spark, Storm, and Flink**
- **Data at rest in IoT networks** can be typically found in ***IoT brokers*** or in ***some sort of storage array at the data center***
- **Hadoop** not only helps with data processing but also data storage

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***



Types of Data Analysis Results

4 types of data analysis results

1.Descriptive:

- Descriptive data analysis tells you what is happening, either now or in the past.
- For example, a thermometer in a truck engine reports temperature values every second.
- From a descriptive analysis perspective, you can pull this data at any moment to gain insight into the **current operating condition of the truck engine.**
- **If the temperature value is too high**, then **there may be a cooling problem** or the ***engine may be experiencing too much load.***

2. Diagnostic:

- When you are interested in the **"why,"** diagnostic data analysis can provide the answer.
- Continuing with the example of the temperature sensor in the truck engine, ***you might wonder why the truck engine failed.***
- **Diagnostic analysis might show that the temperature of the engine was too high, and the engine overheated**.
- Applying diagnostic analysis across the data generated by a wide range of smart objects can provide a clear picture of why a problem or an event occurred

3. Predictive:

- **Predictive analysis aims to foretell problems or issues before they occur.**
- For example, with historical values of temperatures for the **truck engine, predictive analysis could provide an estimate on the remaining life of certain components in the engine.**
- These components could then be proactively replaced before failure occurs.
- Or perhaps if temperature values of the truck engine start to rise slowly over time, this could indicate the need for an **oil change or some other sort of engine cooling maintenance**

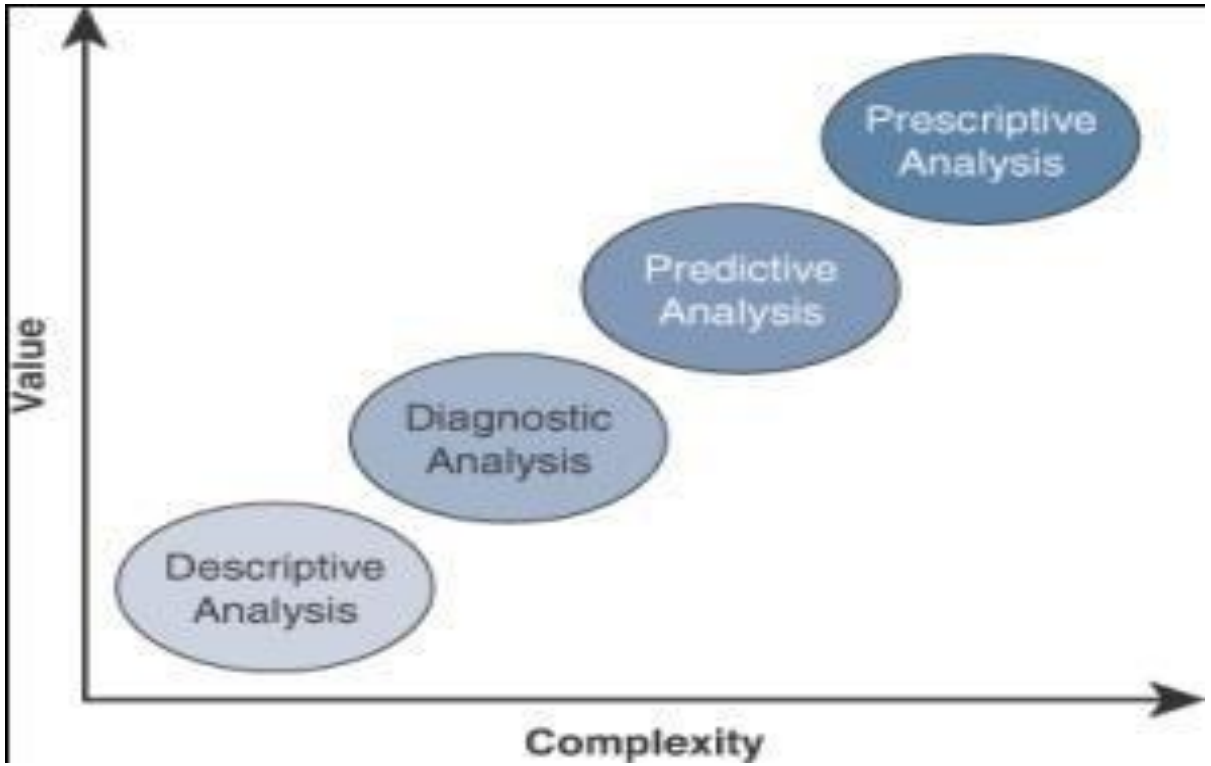
4 types of data analysis results

4. Prescriptive:

- Prescriptive analysis goes a step beyond predictive and recommends solutions for upcoming problems.
- A prescriptive analysis of the temperature data from a truck engine might calculate **various alternatives to cost-effectively maintain our truck**
- *These calculations could range from the cost necessary for more frequent oil changes and cooling maintenance to installing new cooling equipment on the engine or upgrading to a lease on a model with a more powerful engine.*
- Prescriptive analysis looks at a variety of factors and makes the appropriate recommendation

4 types of data analysis results

- Both predictive and prescriptive analyses are more resource intensive and increase complexity, but the value they provide is much greater than the value from descriptive and diagnostic analysis



IoT Data Analytics Challenges

Problems by using RDMS in IoT

1. **Scaling Problems** (performance issues, costly to resolve, req more h/w, architechture changes)

2. **Volatility of Data** (change in schema)

Machine Learning

- ML is central to IoT.
- Data collected by smart objects needs to be **analyzed**, and **intelligent actions** need to be taken based on these analyses.
- Performing this kind of operation manually is almost impossible (or very, very slow and inefficient).

- ***Machines are needed to process information fast and react instantly when thresholds are met***
  - ***Ex: advances in self-driving vehicle--abnormal pattrn recognition in a crowd and automated intelligent and machine-assisted decision system***

Machine Learning Overview

- Machine learning is, in fact, 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** using any technique, from very advanced logic to basic "if-then-else" decision loops.

- Any computer that **uses rules to make decisions** belong to this group

Machine Learning Overview

- **A simple example is an app that can help you find your parked car.**
- A GPS reading of your position at regular intervals calculates your speed.
- A basic threshold system determines whether you are driving (for example, “if speed > 20 mph or 30 kmh, then start calculating speed”).
- When you park and disconnect from the car Bluetooth system, the app simply records the location when the disconnection happens.
- This is where your car is parked.

Machine Learning Overview

- In more **complex cases**, static rules cannot be simply inserted into the program because they require *parameters that can change or that are imperfectly understood*
- A typical example is a **dictation program** that runs on a computer.
- The ***program is configured to recognize the audio pattern of each word in a dictionary,*** but it does not know your voice's specifics—your accent, tone, speed, and so on
- You need to record a **set of predetermined sentences** to help the tool match **well-known words** to **the sounds you make when you say the words.-**This process is called machine learning.
- **ML is concerned with any process where the computer needs to receive a set of data that is processed to help perform a task with more efficiency.**
- ML is a vast field divided in **two main categories:** supervised and

Supervised Learning

- In supervised learning, **the machine is trained with input for which there is a known correct answer**.
- For example, suppose that you are training a system to recognize when there is a human in a **mine tunnel**.
- **A sensor** equipped with a **basic camera** can **capture shapes** and **return them** to a **computing system** that is responsible for **determining whether the shape is a human or something else** (such as a vehicle, a pile of ore, a rock, a piece of wood, and so on.).

Supervised Learning

- With supervised learning techniques, hundreds or thousands of images are fed into the machine, and each image is labelled (human or nonhuman in this case).
- This is called the ***training set.***
- An **algorithm** is used to ***determine common parameters and common differences between the images*.**
- The **comparison** is usually done at the ***scale of the entire image, or pixel by pixel.***
- Images are **resized** to have the same characteristics (resolution, color depth, position of the central figure, and so on), and each point is analyzed.

Supervised Learning

- Each new image is compared to the set of known "good images," and a **deviation is calculated** to ***determine how different, the new image*** is from the ***average human image*** and, therefore, the **probability** that what is shown is a human figure. This process is called ***classification.***

- After training, the machine should be able to recognize human shapes. **Before real field deployments,** the *machine is usually tested with* ***unlabeled pictures****— this is called* ***the validation or the test set***, depending on the ML system used—to verify that the recognition level is at acceptable thresholds. ***If the machine does not reach the level of success expected, more training is needed***

Supervised Learning

- In other cases, the learning process is not about classifying in two or more categories *but about finding a correct value.*
- For example, the **speed of the flow of oil in a pipe** is a function of the *size of the pipe, the viscosity of the oil, pressure, and a few other factors.*
- When you train the machine with measured values, the machine can predict the speed of the flow for a new, and unmeasured, viscosity.
- This process is called ***regression; regression predicts numeric values,*** whereas classification predicts categories

Unsupervised Learning

- In some cases, supervised learning is not the best method for a machine to help with a human decision.
- 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, and your task is to identify them before they get mounted into machines and shipped away from the factory.
- With hundreds of parts, it may be very difficult to detect the potential defects, and it is almost impossible to train a machine to recognize issues that may not be visible

Unsupervised Learning

- However, you can test each engine and 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 (for example, ***temperature as a function of pressure***, ***sound versus rotating speed overtime).***
- You can then input this data into a computer and use mathematical functions to find groups.

Unsupervised Learning

- **For example**, **you may decide to group the engines by the sound they make at a given temperature.**
- A standard function to operate this grouping, ***K-means clustering, finds the mean values for a group of engines (for example, mean value for temperature, mean frequency for sound).***
- Grouping the engines this way can quickly reveal several types of engines that all belong to the same category (for example, *small engine of chainsaw type*, *medium engine of lawnmower type*).
- All engines of the same type produce sounds and temperatures in the same range as the other members of the same group.

Unsupervised Learning

- **There will occasionally be an engine in the group that displays unusual characteristics (slightly out of expected temperature or sound range).**
- **This is the engine that you send for manual evaluation**.
- The computing process associated with this determination is called

  unsupervised learning.

- *This type of learning is* unsupervised ***because there is not a "good" or "bad" answer known in advance.***
- It is the ***variation from a group behavior*** that allows the **computer to learn that something is different**

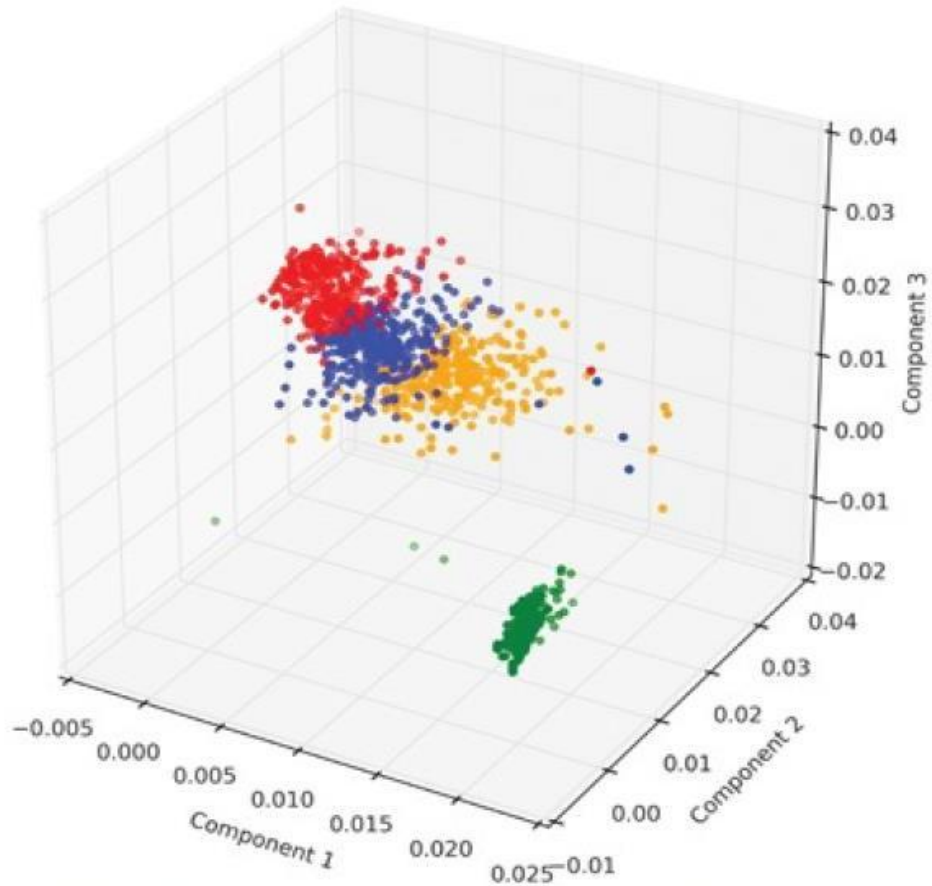


Figure 7-5 *Clustering and Deviation Detection Example*

Neural Networks

- Processing multiple dimensions requires a *lot of computing power*.
- It is also difficult to determine *what parameters to input* and *what combined variations should raise red flags.*
- Similarly, supervised learning is efficient only with a large training set; larger training sets usually lead to higher accuracy in the prediction
- Training the machines was often deemed too expensive and complicated.

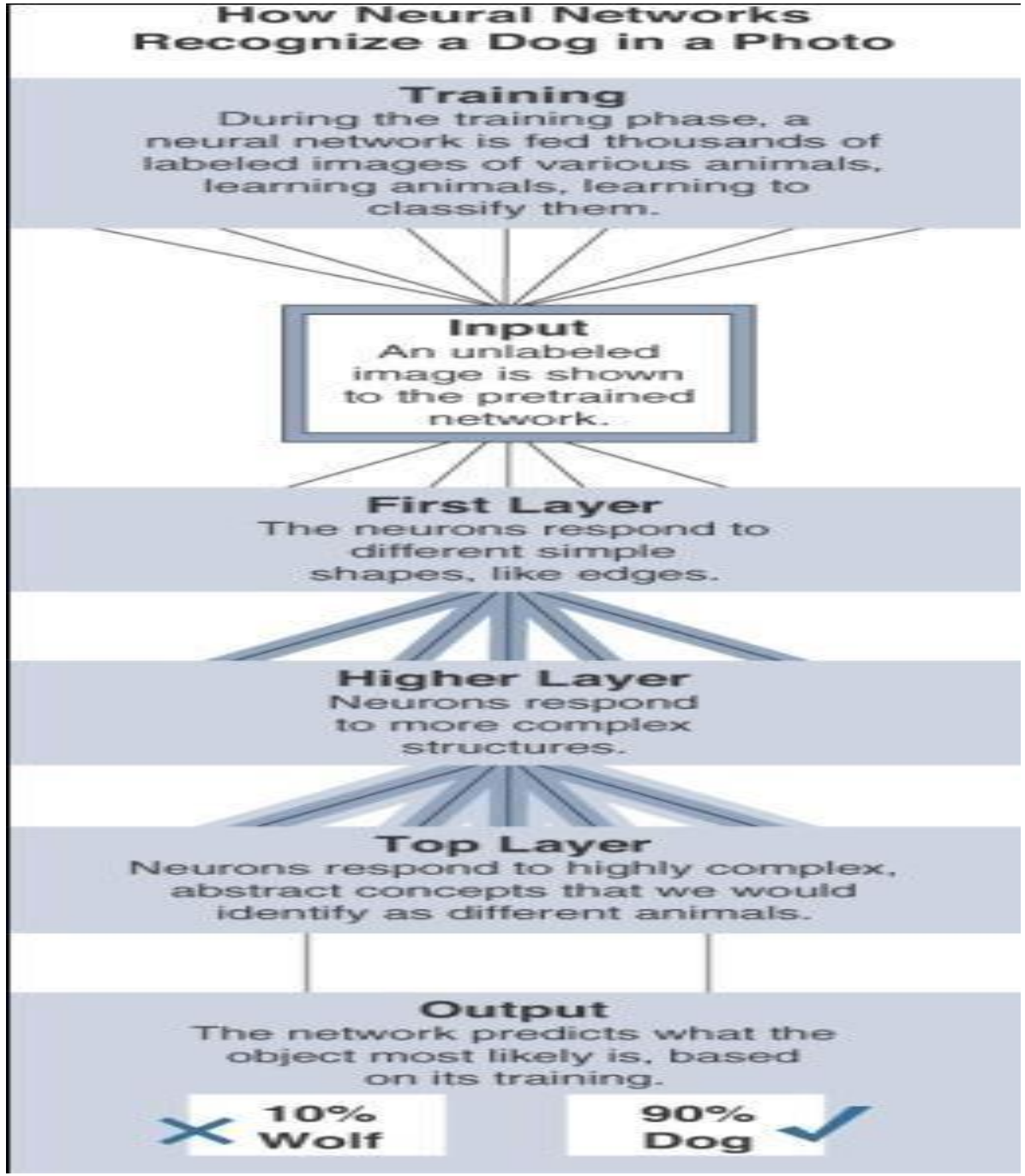
Neural Networks

- **Distinguishing between a human and a car is easy.**
- The computer can recognize that humans have distinct shapes (such as legs or arms) and that vehicles do not.
- But, distinguishing a human from another *mammal* is much more difficult

  The same goes for telling the difference between a pickup truck and a van.
- You can tell when you see one, but training a machine to differentiate them ***requires more than basic shape recognition***

Neural Networks

- Neural networks are ML methods that mimic the way the human brain works.
- When you look at a human figure, multiple zones of your brain are activated to recognize colors, movements, facial expressions, and so on.
- Your brain combines these elements to conclude that the shape you are seeing is human
- Neural networks mimic the same logic.
- The information goes through different algorithms (called *units),* each of which is in charge of processing an aspect of the information


How Neural Networks Recognize a Dog in a Photo
Training
During the training phase, a neural network is fed thousands of labeled images of various animals, learning animals, learning to classify them.
Input
An unlabeled image is shown to the pretrained network.
First Layer
The neurons respond to different simple shapes, like edges.
Higher Layer
Neurons respond to more complex structures.
Top Layer
Neurons respond to highly complex, abstract concepts that we would identify as different animals.
Output
The network predicts what the object most likely is, based on its training.
10% Wolf
90% Dog

Neural Networks

- Neural networks rely on the **idea** that **information is divided into key components, and each component is assigned a weight**.
- The weights compared together decide the classification of this information (no straight lines + face + smile = human).
- When the result of a layer is fed into another layer, the process is called **deep learning** ("deep" because the learning process has more than a single layer).
- One advantage of deep learning is that ***having more layers allows for richer intermediate processing*** and ***representation of the data.***
- At each layer, ***the data can be formatted to be better utilized by the next layer.*** **This process increases the efficiency of the overall result**

Machine Learning and Getting Intelligence from Big Data ML operations can be organized into two broad subgroups:

- **Local learning**
  - Data is collected and processed **locally,** either in the **sensor itself (the edge node)** or in the **gateway (the fog node)**

- **Remote learning**
  - Data is collected and **sent to a central computing unit (typically the data center in a specific location or in the cloud), where it is processed.**

- Regardless of the location where data is processed, ***common applications of ML for IoT revolve around four major domains:***

1.Monitoring

- Smart objects monitor the environment where they operate
- Example such as air temperature, humidity, or presence of carbon dioxide in a mine etc

2. Behavior control

- *Monitoring commonly works in conjunction with behavior control.*
- When a given set of parameters reach a target **threshold — defined in advance (that is, supervised)** or **learned dynamically through deviation from mean values (that is, unsupervised)—*monitoring functions generate an alarm.***
- This alarm can be relayed to a human, ***but a more efficient and more advanced system would trigger a corrective action***, such as *increasing the flow of fresh air in the mine tunnel*, *turning the robot arm*, or *reducing the oil pressure in the pipe.*

3. Operations optimization

- Behavior control typically aims at taking corrective actions based on thresholds.
- However, ***analyzing data can also lead to changes that improve the overall process***
- For example, **a water purification plant in a smart city can implement a system to monitor the efficiency of the purification process** ***based on which chemical (from company A or company B) is used, at what temperature, and associated to what stirring mechanism (stirring speed and depth).***

4. Self-healing, self-optimizing

- The ML engine can be ***programmed to dynamically monitor and combine new parameters (randomly or semi-randomly) and automatically deduce and implement new optimizations when the results demonstrate a possible gain.***
- The system becomes ***self-learning and self optimizing***
- It also detects new K-means deviations that result in pre detection of new potential defects, allowing the system to self-heal

Predictive Analytics

- Machine learning and big data processing for IoT fit very well into th digitization
- The ***advanced stages*** of this model **see the network self- diagnose and self-optimize**.
- In the IoT world, this behavior is what the previous section describes
- ***When data from multiple systems is combined and analyzed together, predictions can be made about the state of the system.*** For example,
- case of sensors deployed on locomotives. Multiple smart objects measure the pull between carriages, the weight on each wheel, and multiple other parameters to offer a form of cruise control optimization for the driver.

- At the same time, cameras observe the state of the tracks ahead, audio sensors analyze the sound of each wheel on the tracks, and multiple engine parameters are measured and analyzed.
- All this data can be returned to a data processing center in the cloud that can re-create a virtual twin of each locomotive.
- Modeling the state of each locomotive and combining this knowledge with anticipated travel and with the states (and detected failures) of all other locomotives of the same type circulating on the tracks of the entire city, province, state, or country allows the analytics platform to make very accurate **predictions on what issue is likely to affect each train and each locomotive**.
- **Such predictive analysis allows preemptive maintenance and increases the safety and efficiency of operations.**
- Similarly, sensors combined with big data can defects or issues in vehicles operating in mines, in manufacturing machines, or any system that can be monitored, along with other similar systems.

- **Big data analytics can *consist of many different software pieces* that *together collect, store, manipulate, and analyze all different data types.***

Generally, the industry looks to the **"three *Vs"*** *to categorize* big data:

Velocity-Refers to ***how quickly data is being collected and analyzed.***

- *Hadoop Distributed File System* is designed to ingest and process data very quickly.
- Smart objects can *generate machine and sensor data at a very fast rate and require database or file systems capable of equally fast ingest functions.*

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

- Often you see *data categorized as structured, semi-structured, or unstructured.*
- *Different database technologies may only be capable of accepting one of these types.*
- ***Hadoop is able to collect and store all three types***

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

- Typically, this is ***measured from gigabytes on the very low end to petabytes or even exabytes of data on the other extreme***

- The **characteristics of big data** can be defined by the **sources** and **types of data**.
- First is **machine data,** which is ***generated by IoT devices*** *and* ***is typically unstructured data.***
- Second is ***transactional data***, which is *from sources that produce data from transactions on these systems*, and, *have* ***high volume and structured.***
- Third is ***social data sources***, which are typically ***high volume and structured.***
- Fourth is ***enterprise data***, which is data that is ***lower in volume*** and ***very structured***
- **Hence big data consists of data from all these separate sources**.

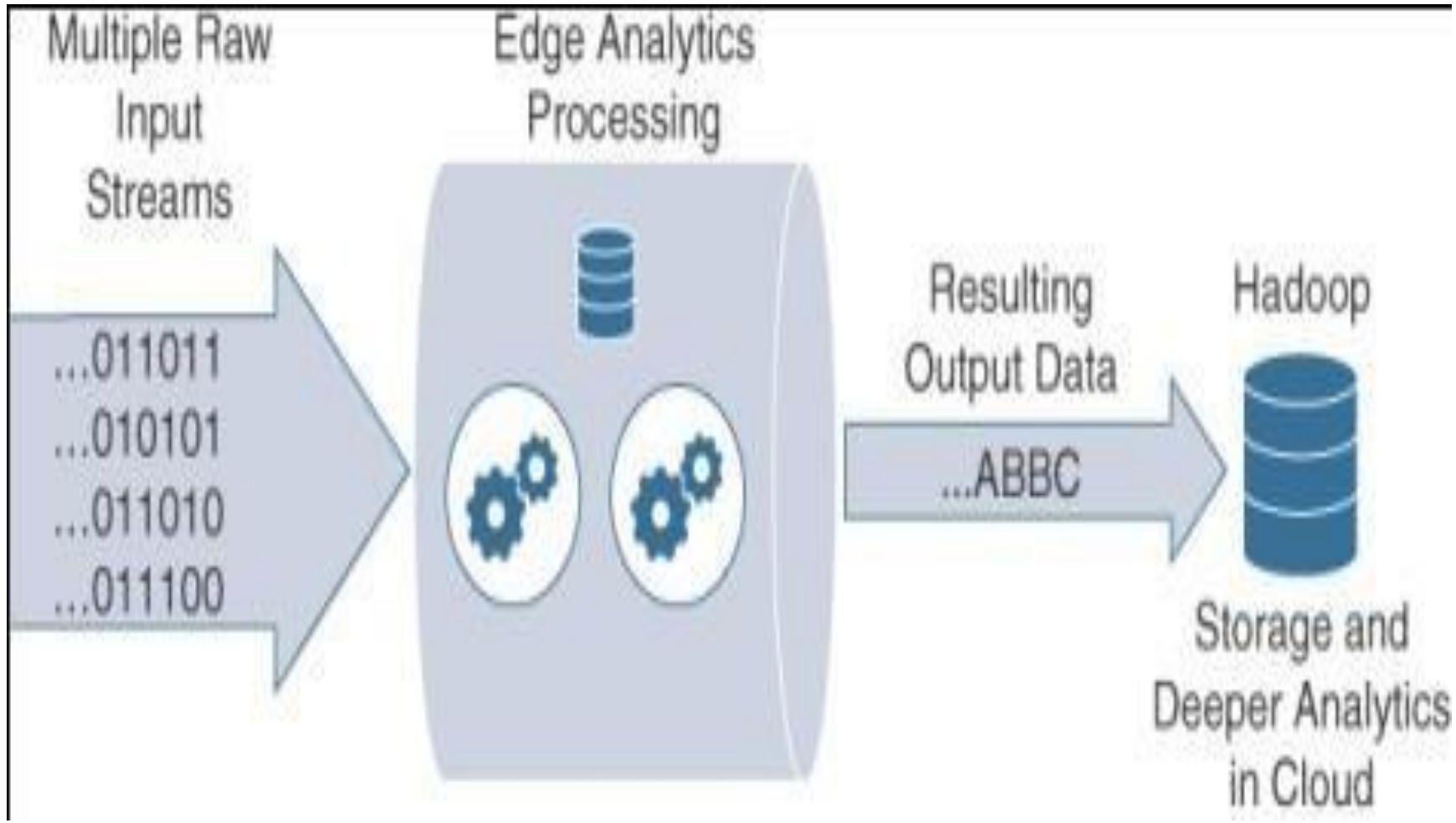
Hadoop

- Hadoop is the most recent entrant into the data management market, but it is arguably the *most popular choice* as a ***data repository and processing engine.***
- Hadoop was originally ***developed as a result of projects at Google andYahoo!***
- The original intent for Hadoop was to ***index millions of websites*** and ***quickly return search results for open source search engines***
- **Hadoop Distributed File System (HDFS):**

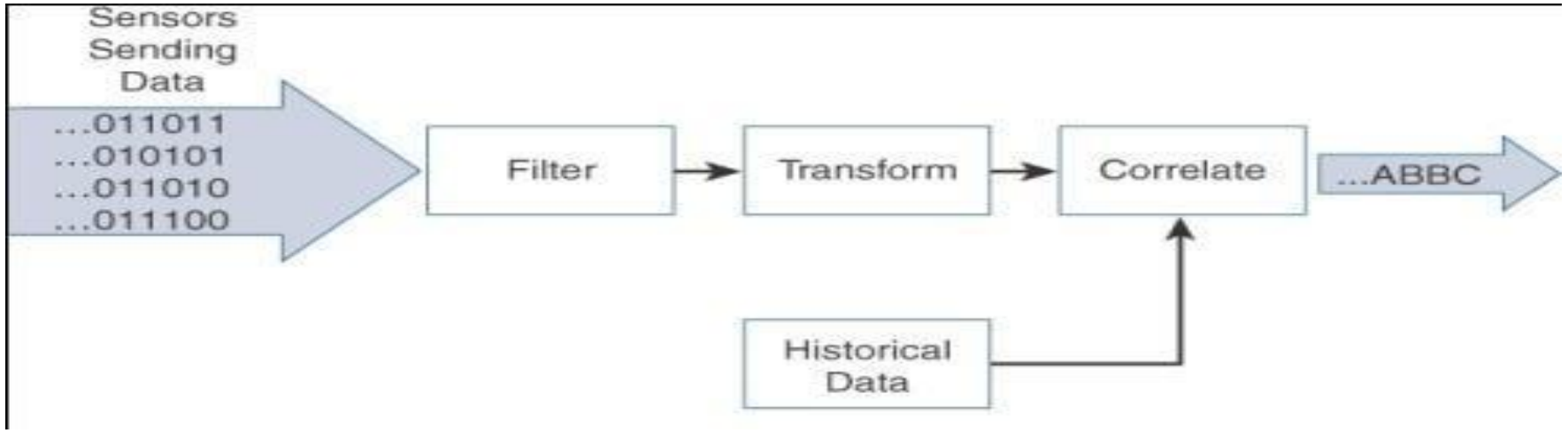
  A system for storing data across multiple nodes
- **MapReduce:**
  - ***A distributed processing engine*** that ***splits a large task into smaller ones that can be run in parallel.***
  - Hadoop relies on a ***scale-out architecture*** that ***leverages local processing, memory, and storage to distribute tasks and provide a scalable storage system for data.***

Edge Analytics Core Functions

- To perform analytics at the edge, data needs to be viewed as real-time flows.
- Whereas big data analytics is focused on large quantities of data at rest, edge analytics continually processes streaming flows of data in motion
- Streaming analytics at the edge can be broken down into three simple stages:
  - **Raw input data**
  - **Analytics processing unit (APU)**
  - **Output streams**


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

- In order to perform analysis in real-time, the APU needs to perform the following functions:
  - **Filter**
  - **Transform**
  - **Time**
  - **Correlate**
  - **Match patterns**
  - **Improve business intelligence**



Distributed Analytics Systems

- Depending on the application and network architecture, analytics can happen at any point throughout the IoT system.
- Streaming analytics may be performed directly at the edge, in the fog, or in the cloud data center.
- There are no hard and- fast rules dictating where analytics should be done, but there are a few guiding principles
- Sometimes better insights can be gained and data responded to more intelligently when we step back from the edge and look at a wider data set.
- This is the value of fog computing.
- Fog analytics allows you to see beyond one device, giving you visibility into an aggregation of edge nodes and allowing you to correlate data from a wider set
- *Example of an oil drilling company that is measuring both pressure and temperature on an oil rig.*

IoT Data Analytics Challenges

As IoT has grown and evolved, traditional data analytics solutions were not always adequate for it
Traditional data analytics typically employs a standard RDBMS and corresponding tools, but the world of IoT is much more demanding. Relational databases often struggle with the nature of IoT data.

IoT data places two specific challenges on a relational database as follows:

Scaling problems:

- Due to the large number of smart objects in most IoT networks that continually send data, relational databases need to grow incredibly large and very quickly.
- This can result in performance issues that can be costly to resolve, often requiring more hardware and architecture changes.

Volatility of data:

- With relational databases, it is critical that the schema be designed correctly from the beginning. Changing it later can slow or stop the database from operating.
- Due to the lack of flexibility, revisions to the schema must be kept at a minimum. IoT data, however, is volatile in the sense that the data model is likely to change and evolve over time.
- A dynamic schema is often required so that data model changes can be made daily or even hourly.
- To deal with challenges like scaling and data volatility, a different type of database, known as NoSQL, is being used. Structured Query Language (SQL) is the computer language used to communicate with an RDBMS.

Major cloud analytics providers, such as Google, Microsoft, and IBM, have streaming analytics offerings, and various other applications can be used in house. (Edge streaming analytics is discussed in depth later in this chapter.)

Another challenge that IoT brings to analytics is in the area of network data, which is referred to as network analytics. With the large numbers of smart objects in IoT networks that are communicating and streaming data, it can be challenging to ensure that these data flows are effectively managed, monitored, and secure.

Network analytics tools such as Flexible NetFlow and IPFIX provide the capability to detect irregular patterns or other problems in the flow of IoT data through a network. Network analytics, including both Flexible NetFlow and IPFIX, is covered in more detail.

Data acquiring

Having learnt about devices, devices-network data, messages and packet communication to the Internet, let us understand the functions required for applications, services and business processes at application-support and application layers. These functions are data acquiring, data storage, data transactions, analytics, results visualisations, IoT applications integration, services, processes, intelligence, knowledge discovery and knowledge management.

Data Analytics Challenges, Data Acquiring, Organizing in IoT/M2M.

Network Analytics

Network Analytics

Network Analytics

Network Analytics

Network Analytics