Module-4 :-

Data and Analytics for IOT

- **Introduction :-**

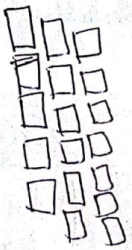
In the world of IOT, the creation of massive amounts of data from sensors is common and one of the biggest challenges - not only from a transport perspective but also from a data management standpoint.

Before diving deeper into data analytics, it is important to define a few key concepts related to data. Depending on how data is categorized, various data analytics tools and processing methods can be applied.

⇒ Structured Versus Unstructured Data

- Structured data and unstructured data are important classifications as they typically require different toolsets from a data analytics perspective.

Fig : comparison Bho Structured & Unstructured

Structured Data	Unstructured data
	
organized formatting (eg, spreadsheets Databases)	Does not conform to a model (eg- text, Images, Video speech).

- 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).
- Structured data can be found in most computing systems and includes everything from banking transaction and invoices to computer log files and router configurations.

- Unstructured data lacks a logical schema for understanding and decoding the data through traditional programming means.
  Eg - data type include text, speech, images and video.

⇒ Data in Motion Versus Data at Rest

- Data in IOT networks is either in transit or being held or stored. Examples of data in motion include traditional client/server exchanges, such as web browsing and file transfers and email.
- 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.
- Data are rest in IOT networks can be typically found in IOT brokers or in some sort of storage array at the data center. The best known of these tools is Hadoop.

⇒ 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 intelligence and insights.

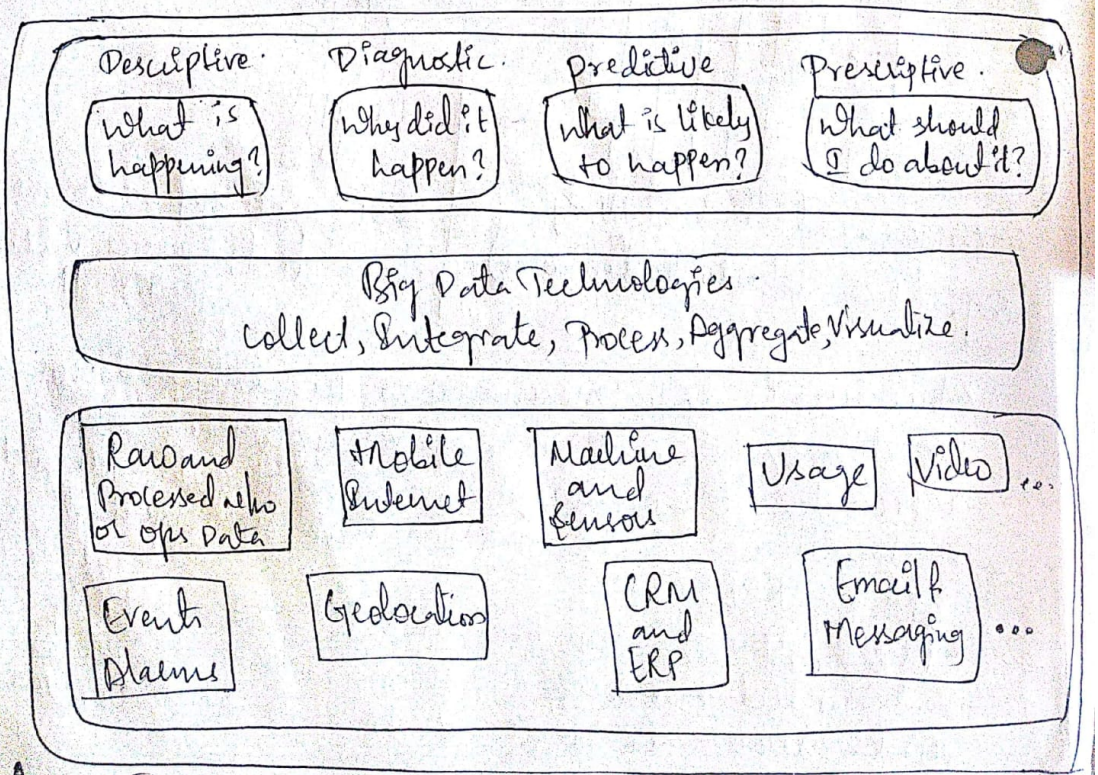


Fig:- Types of Data Analysis Results.

* Descriptive :- Descriptive data analysis tells you what is happening, either now or in the past.
For eg - a thermometer in a truck again engine reports temperature values every second.

* Diagnostic :- when you are interested in the "why", diagnostic data analysis can provide the answer.
For eg - continuing the above example, we might wonder why the truck engine failed. Diagnostic analysis might show that the temperature of the engine was too high, and the engine overheat.

* Predictive :- Predictive analysis aims to foretell problems or issues before they occur.
For eg - with historical values of temperature for the truck engine, predictive analysis could provide an estimate on the remaining life of certain components in the engine.

* 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.

⇒ IOT Data Analytics Challenges.

As IOT has grown and evolved, it has become clear that traditional data analytics solutions were not always adequate.
IOT data places two specific challenges on a relational database:

* Scaling Problems :- Due to the large number of smart objects in most IOT networks that continually send data, relational databases can grow incredibly large very quickly.

* 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.

Machine Learning.

Machine learning, deep learning, neural networks, and convolutional networks are words you have probably heard in relation to big data and IOT. ML is indeed central to IOT.

Machine Learning overview :—

Machine learning is the 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.
- ML is a vast field but can be simply divided in two main categories : Supervised and unsupervised learning.

* Supervised Learning :—

In supervised learning, the machine is trained with input for which there is known correct answer.

- With supervised learning techniques, hundreds or thousands of images are fed into the machine and each image is labeled. This is called the training set.
- 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, the probability that what is shown is a human figure. This process is called classification.

* Unsupervised Learning :—

In some cases, supervised learning is not the best method for a machine to help with a human decision.

There will occasionally be an engine in the group that displays unusual characteristics. This is the engine that you send for manual evaluation. The computing process associated with this determination is called unsupervised learning.

* 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 flages.

Neural networks are ML methods that mimic the way the human brain work. Neural network 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.

The great efficiency of neural network is that each unit processes a simple test, and therefore computation is quite fast.

⇒ Machine Learning and Getting Intelligence from Big Data :-

ML operations enter into two broad subgroups.

* Local learning :- In this group, data is collected and processed locally, either in the sensor itself or in the gateway.
* Remote Learning :- In this group, data is collected and sent to a central computing unit, where it is processed.

- common applications of ML for IOT revolve around four major domains :-

* Monitoring :- Smart objects monitor the environment where they operate. Data is processed to better understand the conditions of operations.
* Behavior control :- Monitoring commonly works in conjunction with behavior control. When a given set of parameters reach a target threshold - defined in advance or learned dynamically through deviations from mean values - monitoring functions generate an alarm.
* Operations optimization :- Behavior control typically aims at taking corrective actions based on thresholds.

The objective is not merely to pilot the operations but to improve the efficiency and the result of these operations.

* Self-healing, self-optimizing :- A fast-developing aspect of deep learning is the closed loop ML-based monitoring triggers changes in machine behavior, and operations optimization.
The system becomes self-learning and self optimizing, it also detects new k-means deviations that results in predetection of new potential defects, allowing the system to self-heal.

Big Data Analytics Tools and Technology.

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 V's" to categorize big data:

- Velocity : Velocity refers to how quickly data is being collected and analyzed. Hadoop Distributed File System is designed to ingest and process data very quickly.
- Variety : Variety refers to different types of data. Data is categorized as structured, semi-structured or unstructured. Hadoop is able to collect and store all three types.
- Volume : 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.
It is common to see clusters of servers that consist of dozens, hundreds, or even thousands of nodes for some large.

- The characteristics of big data can be defined by 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 source that produce data from transactions on these systems and have high volume and structured.

⇒ Massively Parallel Processing Database.

- MPP database were built on the concept of the relational data warehouse but are designed to be much faster, to be efficient, and to support reduced query times.
- MPPs are sometimes referred to as analytics database because they are designed to allow for fast query processing and often have built-in analytics functions.

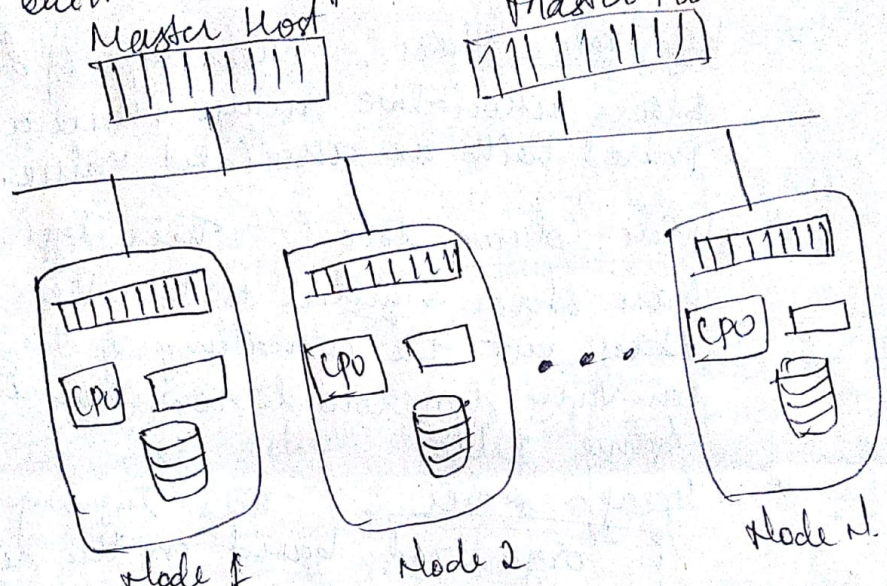


Fig :- MPP Shared-Nothing Architecture.

- In the above figure, we see it typically contains a single master node that is responsible for the coordination of all the data storage and processing across the cluster.
- It operates in a "shared-nothing" fashion, with each node containing local processing common SQL tools and applications.
- The data stored on MPP must still conform and conform to this relational structure, it may not be the only database type used in an EDT implementation.
- The sources and type of data may vary, requiring a database that is more flexible than relational databases allow.

→ NoSQL Databases :-

No SQL ("Not only SQL") is a class of databases that support semi-structured and unstructured data, in addition to the structured data handled by data warehouse and RDBMS.

* Document stores :- This type of database stores semi-structured data, such as XML or JSON.
* Key value stores :- This type of database stores associative arrays where a key is paired with an associated value.
* Wide-column stores :- This type of database stores similar to the key-value stores, but the formatting of the values can vary from row to row, even in the same table.
* Graph stores :- This type of database is organized based on the relationships b/w elements.

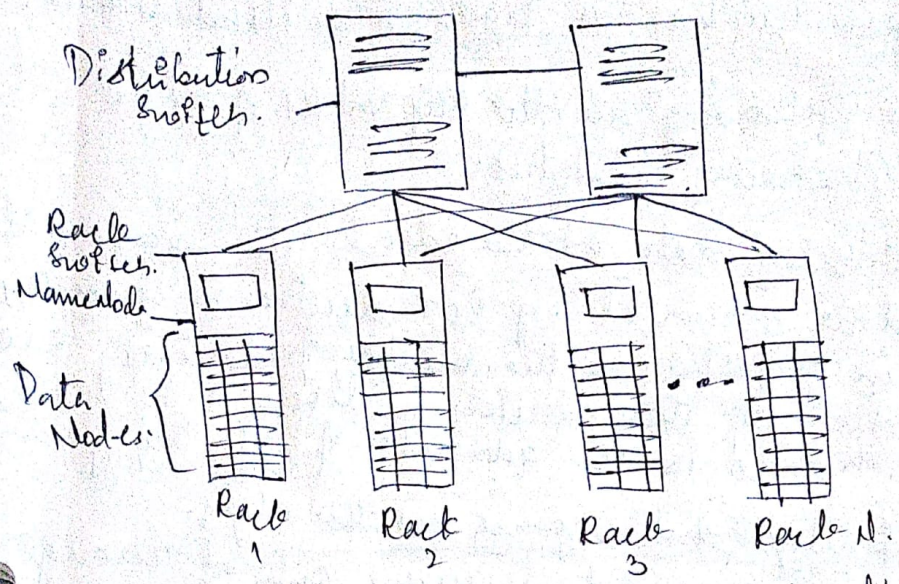
- NoSQL was developed to support the high-velocity, urgent data requirements of modern web applications that typically do not require much repeated use.

⇒ Hadoop :-

Hadoop was originally developed as a result of projects at Google and Yahoo!. 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.

Fig :- Distributed Hadoop Cluster.



- For HDFS, this capability is handled by specialized nodes in the cluster, including NameNodes and DataNodes.

* NameNodes :- These are a critical piece in data adds, moves, deletes and reads on HDFS. They coordinate where the data is stored, and maintain a map of where each block of data is stored and where it is replicated.

The Name Node is also responsible for instructing the DataNodes where replication should occur.

* DataNodes :- These are the servers where the data is stored at the direction of the NameNode. It is common to have many DataNodes in a Hadoop cluster to store the data.

Data blocks are distributed across several nodes and often are replicated three, four or more times across nodes for redundancy.

Once data is written to one of the DataNodes, the DataNodes selects two additional nodes, based on replication policies, to ensure data redundancy techniques such as (RAID) Redundant Array of Independent Disks are generally not used for HDFS because the NameNodes and DataNodes coordinate block level redundancy with this replication techniques.

⇒ Edge Streaming Analytics:-

A major area of evolution for IT in the past few years has been the transition to cloud services

- The following are the key values of edge streaming analytics.

* Reducing data at the edge:- The aggregate data generated by IoT devices is generally in proportion to the number of devices. The scale of these devices is likely to be large, and so is the quantity of data they generate.

* Analysis and response at the edge:- Some data is useful only at the edge. In cases such as this, the data is best analyzed and acted upon where it is generated.

* 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

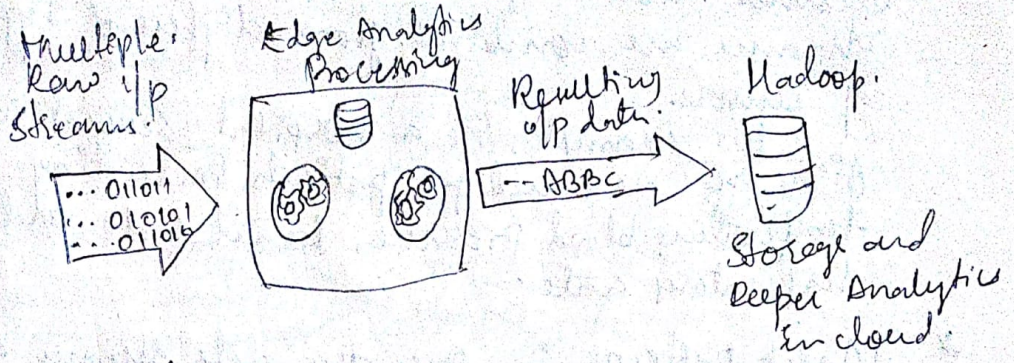
Streaming analytics at the edge can be broken down into three simple stages.

* Raw input data:- This is the raw data coming from the sensor into the analytic processing unit.

* Analytic Processing Unit (APU):- The APU filters and combines data streams, organizes them by time windows, and performs various analytical functions.

* Output Streams:- The data that is organized into meaningful streams and is used to influence the behavior of smart objects, and passed on for storage and further processing in the cloud.

Key :- Illustrates the stages of data processing in an edge App.



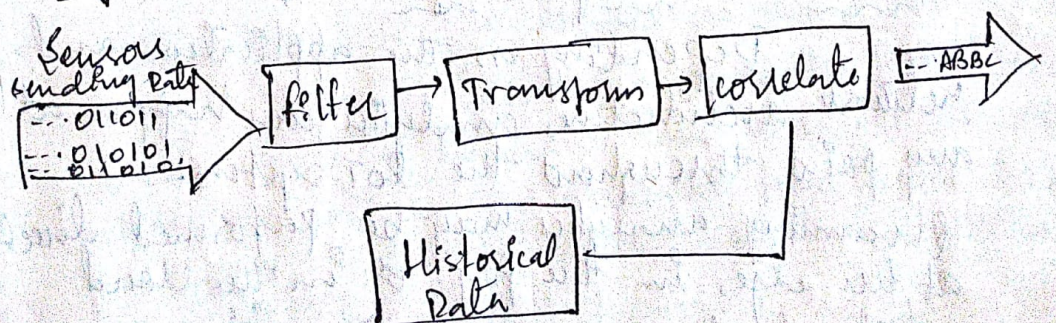
- In order to perform analysis in real-time, the App needs to perform the following functions:

* Filters :- The streaming data generated by IoT endpoints is likely to be very large, and most of it is irrelevant.
  For example - a sensor may simply poll on a regular basis to confirm that it is still reachable.

* Transform :- In the data warehousing world, Extract, Transform, and Load (ETL) operations are used to manipulate the data structure into a form that can be used for other purpose.

* Time :- As the real-time streaming data flows, a timing context needs to be established. This could be to correlated average temperature readings from sensors on a minute-by-minute basis.

Key :- Correlating Data Streams with Historical Data.



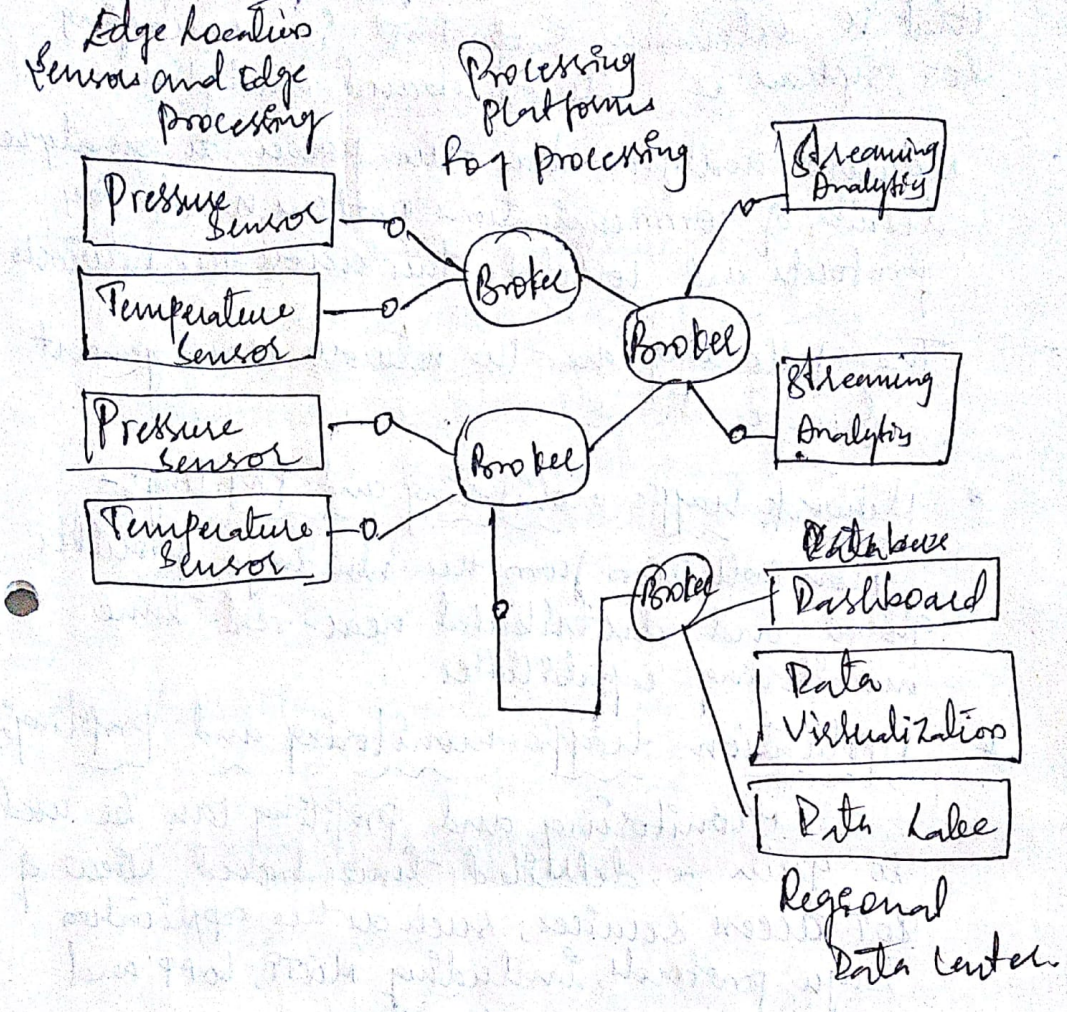
* Correlates :- Streaming data analytics becomes most useful when multiple data streams are combined from different types of sensors.
For example, in a hospital, several vital signs are measured for patients, including body temperature, blood pressure, heart rate and respiratory rate.

* Match patterns :- Once the data streams are properly cleaned, transformed, and related with other live streams as well as historical data sets, pattern matching operations are used to gain deeper insights to the data.
for eg :- If an unexpected event arises, such as a sudden change in heart rate or respiration the pattern matching operator recognizes this as out of the ordinary and can take certain actions.

* Improve business intelligence :-
The value of edge analytics is in the improvements to business intelligence that were not previously available.
For eg - Conducting edge analytics on patients in a hospital allows staff to respond more quickly to the patient's changing needs and also reduces the volume of unstructured data sent to the cloud.

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.

Fig: Distributed Analytics Throughout the IoT systems.



→ The above figure shows an example of distributed analytics of an oil drilling company that is measuring both pressure and temperature on an oil rig.

— While there may be some value in doing analytics directly on the edge, in analytics mode, allowing a broader data set.

— The fog node is located on the same oil rig and performs streaming analytics from several edge node is located on the same oil rig and performs streaming analytics from several edge devices, giving it better insights due to the expanded data sets.

⇒ Network Analytics.

It is another form of analytics that is extremely important in managing IoT systems is network-based analytics.

- Network analytics has the power to analyze details of communications patterns made by protocols and correlate this across the networks.
- The following are the network management services.

Network traffic monitoring and profiling:-

Flow collection from the NMS layer provides global and distributed near-real-time monitoring capabilities.

Application traffic monitoring and profiling:-

Monitoring and profiling can be used to gain a detailed time based view of IoT access services, such as the application layer protocols, including MQTT, CoAP and DNP3 over the network.

Capacity Planning:-

Flow analytics can be used to track and anticipate IoT traffic growth and help in the planning of upgrades when deploying new locations or services by analyzing captured data over a long period of time.

Security analysis:-

Most of IoT devices typically generate a low volume of traffic and always send their data to the same servers, any change in network traffic behavior may indicate a cyber security event, such as denial of service (DoS) attack.

Securing IOT

- Common challenges in IOT security.

The security challenges faced in IOT are by no means new and are not limited to specific industrial environments.

① Erosion of the Architecture :-

Two of the major challenges in securing industrial environments have been initial design and ongoing maintenance.

The initial design challenges arose from the concept that networks were safe due to physical separation from the enterprise with minimal or no connectivity to the outside world.

② Pervasive Legacy Systems :-

Due to the static nature and long lifecycles of equipment in industrial environments, many operational systems may be deemed legacy systems.

Beyond the endpoints, the communication infrastructure and shared centralized compute resources are often not built to comply with modern standards.

③ Insecure Operational Protocols :-

Many industrial control protocols, particularly those that are serial based, were designed without inherent strong security requirements.

- Modbus

Modbus is commonly found in many industries, such as utilities and manufacturing environments, and has multiple variants.

The security challenges that have existed with modbus are not unusual. Authentication of communication endpoints was not a default operation.

- DNP3 (Distributed Nlho Protocol)

DNP3 is found in multiple deployment scenarios and industries. It is common in utilities and is also found in discrete and continuous process systems.

There is an explicit "secure" version of DNP3, but there also remain many insecure implementations of DNP3 as well.

- ICCP (Inter-control center communications Protocol

ICCP is a common control protocol in utilities across North America that is frequently used to communicate between utilities.

ICCP was designed from inception to work across a WAN. Initial versions of ICCP had several significant gaps in the area of security.

- OPC (OLE for process control) :-

OPC is based on the Microsoft Interoperability methodology Object Linking and Embedding (OLE).

This is an example where an IT Standard used within the IT domain and personal computers has been leveraged for use as a control protocol across an industrial network.

XXXXX

Head of the Dept.
Computer Science & Engineering
Basavakalyan Engineering College
BASAVAKALYAN