LECTURE-5
BIG DATA AND ANALYTICS

TOPICS
Introduction
Data Types
Charaterstics of Big Data
Data Sources
Data Acquistion
Data Storage
Big Data Analytics for Industry.4.0

Introduction

Big Data Analytics: Big data analytics involves the collection, storage, and analysis of large volumes of data to uncover patterns, trends, and insights. It utilizes advanced algorithms and machine learning techniques to process and derive actionable information from diverse data sources generated by machines, sensors, and other digital systems.

According to Gartner, the definition of Big Data –
"Big data" is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

This definition clearly answers the "What is Big Data?" question – Big Data refers to complex and large data sets that have to be processed and analyzed to uncover valuable information that can benefit businesses and organizations.

However, there are certain basic tenets of Big Data that will make it even simpler to answer what is Big Data:

- It refers to a massive amount of data that keeps on growing exponentially with time.
- It is so voluminous that it cannot be processed or analyzed using conventional data processing techniques.
- It includes data mining, data storage, data analysis, data sharing, and data visualization.
- The term is an all-comprehensive one including data, data frameworks, along with the tools and techniques used to process and analyze the data.

The History of Big Data

Although the concept of big data itself is relatively new, the origins of large data sets go back to the 1960s and '70s when the world of data was just getting started with the first data centers and the development of the relational database.

Around 2005, people began to realize just how much data users generated through Facebook, YouTube, and other online services. Hadoop (an open-source framework created specifically to store and analyze big data sets) was developed that same year. NoSQL also began to gain popularity during this time.

The development of open-source frameworks, such as Hadoop (and more recently, Spark) was essential for the growth of big data because they make big data easier to work with and cheaper to store. In the years since then, the volume of big data has skyrocketed. Users are still generating huge amounts of data—but it's not just humans who are doing it.

With the advent of the Internet of Things (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance. The emergence of machine learning has produced still more data.

While big data has come far, its usefulness is only just beginning. Cloud computing has expanded big data possibilities even further. The cloud offers truly elastic scalability, where developers can simply spin up ad hoc clusters to test a subset of data.

Data Types

Benefits of Big Data and Data Analytics

- Big data makes it possible for you to gain more complete answers because you have more information.
- More complete answers mean more confidence in the data—which means a completely different approach to tackling problems.

Types of Big Data

Now that we are on track with what is big data, let's have a look at the types of big data:

a) Structured

Structured is one of the types of big data and By structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms. **For instance, the employee table in a company database will be structured as the employee details, their job positions, their salaries, etc.,** will be present in an organized manner.

b) Unstructured

Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data. Email is an example of unstructured data. Structured and unstructured are two important types of big data.

c) Semi-structured

Semi structured is the third type of big data. Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data. To be precise, it refers to the data that although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data. Thus we come to the end of types of data.

Charaterstics of Big Data

Characteristics of Big Data

Back in 2001, Gartner analyst Doug Laney listed the **3 'V's of Big Data – Variety, Velocity, and Volume.** Let's discuss the characteristics of big data. These characteristics, isolated, are enough to know what big data is. Let's look at them in depth:

a) Variety

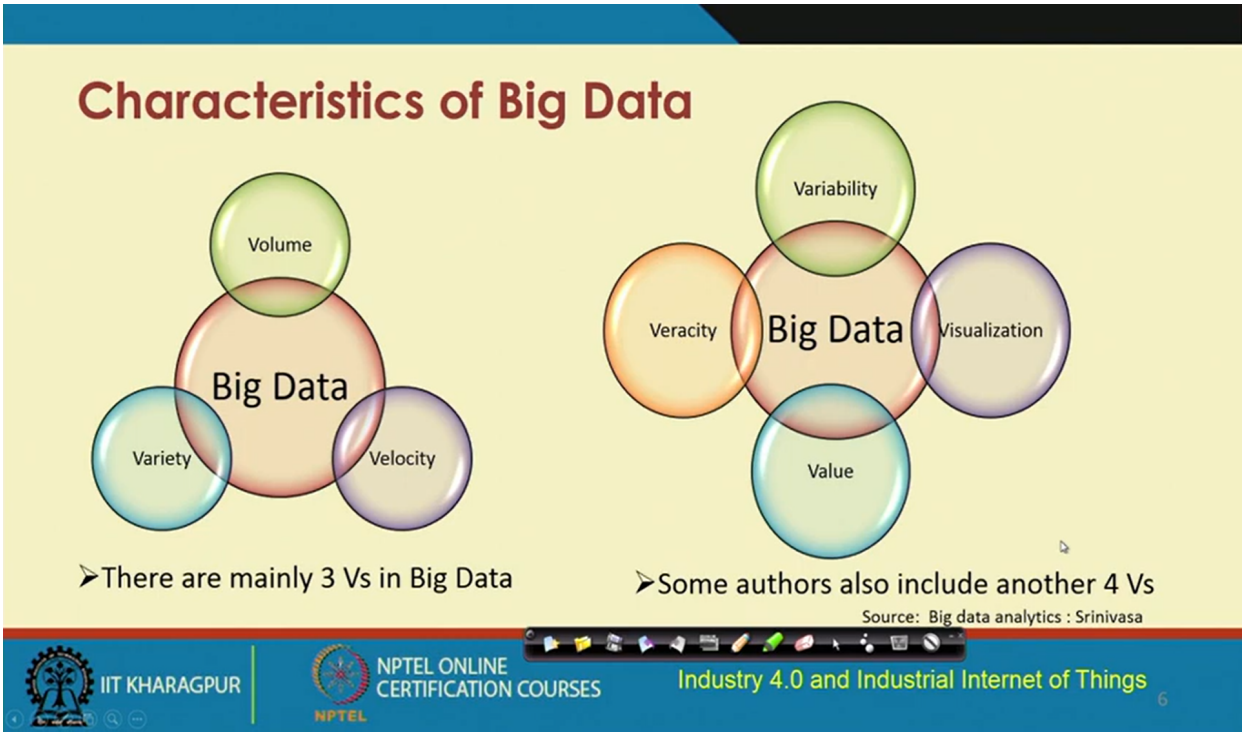
Variety of Big Data refers to structured, unstructured, and semi-structured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more. Variety is one of the important characteristics of big data.

b) Velocity

Velocity essentially refers to the speed at which data is being created in real-time. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.

c) Volume

Volume is one of the characteristics of big data. We already know that Big Data indicates huge 'volumes' of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. Such a large amount of data is stored in data warehouses. Thus comes to the end of characteristics of big data.



Why is Big Data Important?

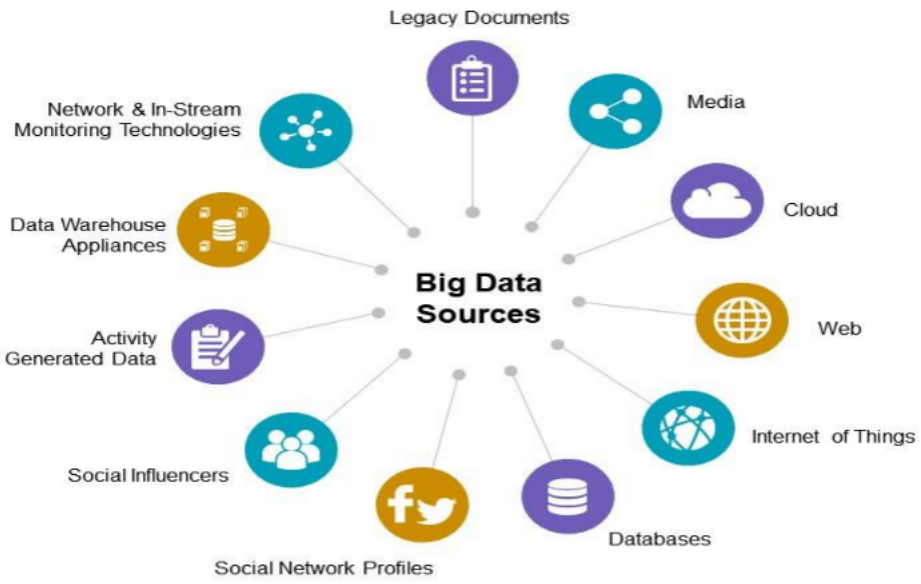
The importance of big data does not revolve around how much data a company has but how a company utilizes the collected data. Every company uses data in its own way; the more efficiently a company uses its data, the more potential it has to grow. The company can take data from any source and analyze it to find answers which will enable:

1. **Cost Savings**: Some tools of Big Data like Hadoop and Cloud-Based Analytics can bring cost advantages to business when large amounts of data are to be stored and these tools also help in identifying more efficient ways of doing business.
2. **Time Reductions: The** high speed of tools like Hadoop and in-memory analytics can easily identify new sources of data which helps businesses analyzing data immediately and make quick decisions based on the learning.
3. **Understand the market conditions**: By analyzing big data you can get a better understanding of current market conditions. For example, by analyzing customers' purchasing behaviors, a company can find out the products that are sold the most and produce products according to this trend. By this, it can get ahead of its competitors.
4. **Control online reputation:** Big data tools can do sentiment analysis. Therefore, you can get feedback about who is saying what about your company. If you want to monitor and improve the online presence of your business, then, big data tools can help in all this.
5. **Using Big Data Analytics to Boost Customer Acquisition and Retention**
   The customer is the most important asset any business depends on. There is no single business that can claim success without first having to establish a solid customer base. However, even with a customer base, a business cannot afford to disregard the high competition it faces. If a business is slow to learn what customers are looking for, then it is very easy to begin offering poor quality products. In the end, loss of clientele will result, and this creates an adverse overall effect on business success. The use of big data allows businesses to observe various customer related patterns and trends. Observing customer behavior is important to trigger loyalty.
6. **Using Big Data Analytics to Solve Advertisers Problem and Offer Marketing Insights**

   Big data analytics can help change all business operations. This includes the ability to match customer expectation, changing company's product line and of course ensuring that the marketing campaigns are powerful.
7. **Big Data Analytics As a Driver of Innovations and Product Development**
   Another huge advantage of big data is the ability to help companies innovate and redevelop their products.

Sources of big data

Main sources of big data can be grouped under the headings of social (human), machine (sensor) and transactional.



Social Networks (human-sourced information): – this source is becoming more and more relevant to organisations. This source includes all social media posts, videos posted etc.(this information is the record of human experiences, previously recorded in books and works of art, and later in photographs, audio and video. Human-sourced information is now almost entirely digitized and stored everywhere from personal computers to social networks. Data are loosely structured and often ungoverned.)

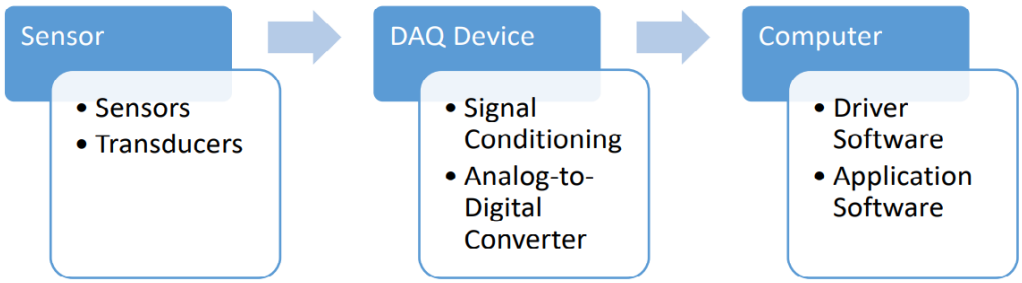
Internet of Things (machine-generated data):– this data comes from what can be measured by the equipment used.: The phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world. The output of these sensors is machine-generated data, and from simple sensor records to complex computer logs, it is well structured. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the information stored and processed by many businesses. Its well-structured nature is suitable for computer processing, but its size and speed is beyond traditional approaches. Data from sensors,Fixed sensors,Home automation , Weather/pollution sensors,Traffic Scientific sensors .Security/surveillance videos/images,Mobile sensors (tracking) ,Mobile phone location ,Cars, Satellite images ,. Data from computer systems, Web logs

Traditional Business systems (process-mediated data):– these processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. The process-mediated data thus collected is highly structured and includes transactions,reference tables and relationships, as well as the metadata that sets its context. Traditional business data is the vast majority of what IT managed and processed, in both operational and BI systems. Usually structured and stored in relational database systems. (Some sources belonging to this class may fall into the category of Administrative data,Data produced by Public Agencies , Medical records,Data produced by businesses ,Commercial transactions ,Banking/stock records ,E-commerce ,Credit cards)

Data Acquistion(COLLECTION,TRANSMISSION,PREPROCESSING,)

Data acquisition has been understood as the process of gathering, filtering, and cleaning data before the data is put in a data warehouse or any other storage solution. The acquisition of big data is most commonly governed by four of the Vs: volume, velocity, variety, and value.There are four methods of acquiring data: collecting new data; converting/transforming legacy data; sharing/exchanging data; and purchasing data. It is a preprocessing before the data is put in a data warehouse or any other storage solution. . Most data acquisition scenarios assume high-volume, high-velocity, high-variety, but low-value data, making it important to have adaptable and time-efficient gathering, filtering, and cleaning algorithms that ensure that only the high-value fragments of the data are actually processed by the data-warehouse analysis.

Big Data architecture has to acquire high speed data from a variety of sources like web, DBMS(OLTP), NoSQL, HDFS and the data is also diverse in nature . It is required to store only data which could be helpful or "raw" data with a lower degree of uncertainty[1]. For that a filter could be established. In some applications, the conditions of generation of data are important, thus it could be interesting for further analysis to capture these metadata and store them with the corresponding data



Protocols that enable the gathering of information from distributed data sources

Frameworks through which the data is collected by using different protocols

Technologies that allow constant storage of data acquired through the frameworks

Frameworks and tech for data gathering

Data gathering frameworks and technologies are very specific when it comes to their functionalities and ideal usage, so it's important to define your overall goals before you lock- in on any of them.When it comes to data gathering, some of the most widely used frameworks and technologies are:

Storm

Storm is an open-source framework for robust distributed real-time computation on streams of data. It supports a wide range of programming languages and storage facilities, and one of its main advantages is that it can be utilized in many data gathering scenarios including stream processing and distributed RPC for solving intensive functions on-the-fly.***"It's used by a number of big systems, with some of the largest ones being Groupon, The Weather Channel and Twitter."***

Simply Scalable Streaming System

Simply Scalable Streaming System or S4 is a distributed, general-purpose platform for developing applications that process streams of data which was launched by Yahoo! Inc. It is designed to

work on commodity hardware, avoiding I/O bottlenecks by relying on an all-in- memory approach.***"provides a simple programming interface for processing data streams in a decentralized, symmetric, and pluggable architecture".***

Kafka

Kafka is a distributed publish-subscribe messaging system designed to support persistent messaging with high-throughput. It aims to unify offline and online processing with its ability to partition real-time consumption over a cluster of machines, and is built in a way that minimizes the network overhead and sequential disk operations.***"It was originally developed at LinkedIn to track the huge volume of activity events generated by the website."***

Flume

Flume is a service whose purpose is to provide a distributed, reliable and available system for efficiently collecting, aggregating and moving large amounts of log data from many different sources to a centralized data store. Its architecture is based on streaming data flows — it is simple and flexible, but also robust and fault tolerant with tuneable reliability mechanisms and many failover and recovery mechanisms.***"Flume was designed with these four key goals in mind — reliability, scalability, manageability and extensibility."***

Hadoop

Hadoop is an open-source project that focuses on developing a framework for reliable, scalable, and distributed computing on big data using clusters of commodity hardware.**"It's used and supported by a large number of big organizations like Facebook, AOL, Baidu, IBM, Imageshack, and Yahoo."**

Big Data Storage

Every month, more companies are adopting it to help them improve their businesses. But with any new technology comes challenges and questions, and big data is no exception.

The first challenge is how much storage you'll need for your extensive data system. If you're going to store large amounts of information about your customers and their behavior, you'll need a lot of space for that data to live.

Companies are constantly gathering new types of information about their customer's habits and preferences, and they're looking at ways they can use this information to improve their products or services.

As a result, big data systems will continue growing exponentially until something stops them. It means it's essential for companies who want to use this technology effectively to plan how they'll deal with it later on down the road when it becomes too much for them alone!

Big Data Storage Key Considerations

Big data storage is a complicated problem. There are many things to consider when building the infrastructure for your big data project, but there are three key considerations you must consider before you move forward.

Data velocity: Your data must be able to move quickly between processing centers and databases for it to be helpful in real-time applications.

Scalability: The system should be able to expand as your business does and accommodate new projects as needed without disrupting existing workflows or causing any downtime.

Cost efficiency: Because big data projects can be so expensive, choosing a system that reduces costs without sacrificing the quality of service or functionality is essential and how long you want your stored data to remain accessible

Key Insights for Big Data Storage

Big data storage is a critical part of any business. The sheer volume of data being created and stored by companies is staggering and growing daily. But without a proper strategy for storing and protecting this data, your business could be vulnerable to hackers—and your bottom line could suffer.So storage methods are one of major key factor

Data Storage Methods

Warehouse and cloud storage are two of the most popular options for storing big data. Warehouse storage is typically done on-site, while cloud storage involves storing your data offsite in a secure location.

Warehouse Storage

Warehouse storage is one of the more common ways to store large amounts of data, but it has drawbacks. For example, if you need immediate access to your data and want to avoid delays or problems accessing it over the internet, there might be better options than this. Also, warehouse storage can be expensive if you're looking for long-term contracts or need extra personnel to manage your warehouse space.

Cloud Storage

Cloud storage is an increasingly popular option since it's easier than ever to use this method, thanks to advancements in technology such as Amazon Web Services (AWS). With AWS, you can store unlimited data without worrying about how much space each file takes up on their servers. They'll automatically compress them before sending them over, so they take up less space overall!

Data Storage Technologies

Apache Hadoop, Apache HBase, and Snowflake are three big data storage technologies often used in the data lake analytics paradigm.

Hadoop

Hadoop has gained considerable attention as it is one of the most common frameworks to support big data analytics. A distributed processing framework based on open-source software, Hadoop enables large data sets to be processed across clusters of computers. Large

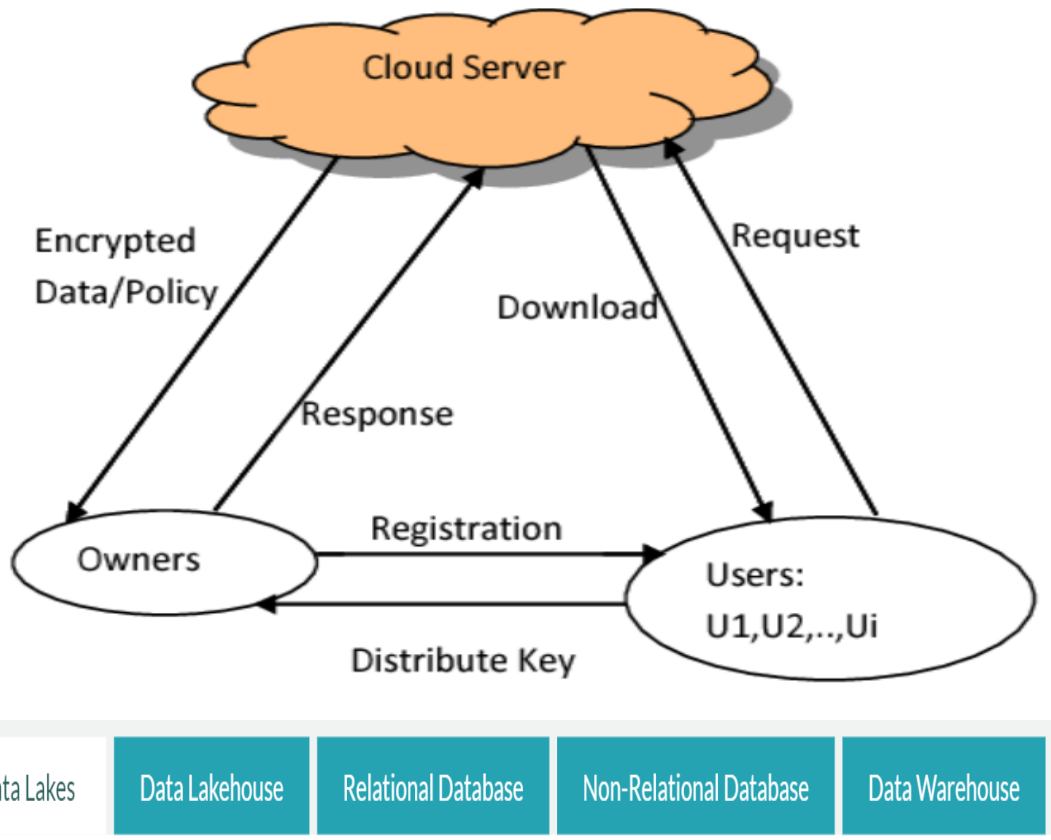
data sets were initially intended to be processed and stored across clusters of commodity hardware.

HBase

With HBase, you can use a NoSQL database or complement Hadoop with a column-oriented store. This database is designed to efficiently manage large tables with billions of rows and millions of columns. The performance can be tuned by adjusting memory usage, the number of servers, block size, and other settings.

Snowflake

Snowflake for Data Lake Analytics is an enterprise-grade cloud platform for advanced analytics applications built on top of Apache Hadoop. It offers real-time access to historical and streaming data from any source and format at any scale without requiring changes to existing applications or workflows. It also enables users to quickly scale up their processing power as needed without having to worry about infrastructure management tasks such as provisioning and



(Big Data Storage solutions)

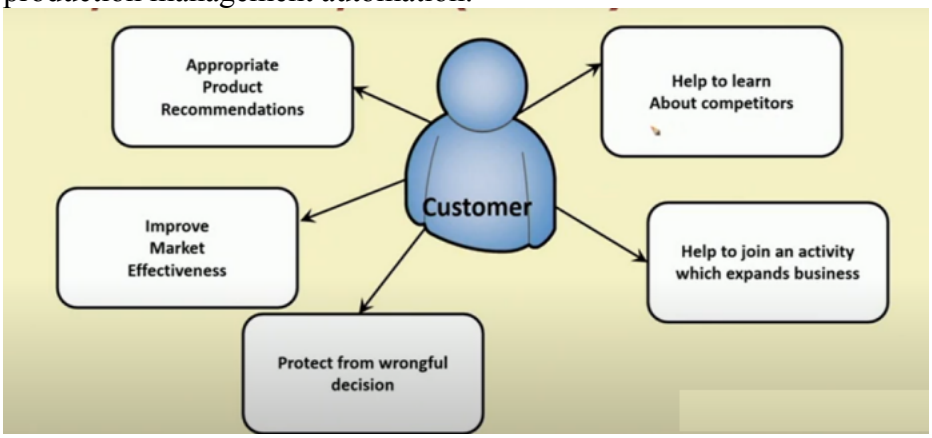
Specifics of each of these Big Data Storage methods reveals:

	Data Lake	Data Lakehouse	Non-Relational Database (NoSQL)	Relational Database (SQL)	Data Warehouse
Data Storage Approach	Store everything	Store everything	Non-relational data in a document model format	Low-level data in a tabular format	Hold business-ready data
Type of Data	Unstructured, semi-structured and structured.	Unstructured, semi-structured and structured.	Documents, key-value, column-oriented or graph types	Tabular form & structure and relational	Tabular form & structure and aggregated / dimensional
Processing	Data is mainly unprocessed	Data is processed however you want	Raw documental data	Raw transactional data	Highly processed data
Adaptability	Fast ingestion of new or changed data	Fast ingestion of new or changed data	Flexible with an emphasis on document processing but can support transactions	Time-consuming to introduce new content	Time-consuming to introduce new content
Storage Cost	Low-cost	Low-cost	Generally cheaper to scale when compared to relational databases	Can become expensive when scaled	Expensive storage

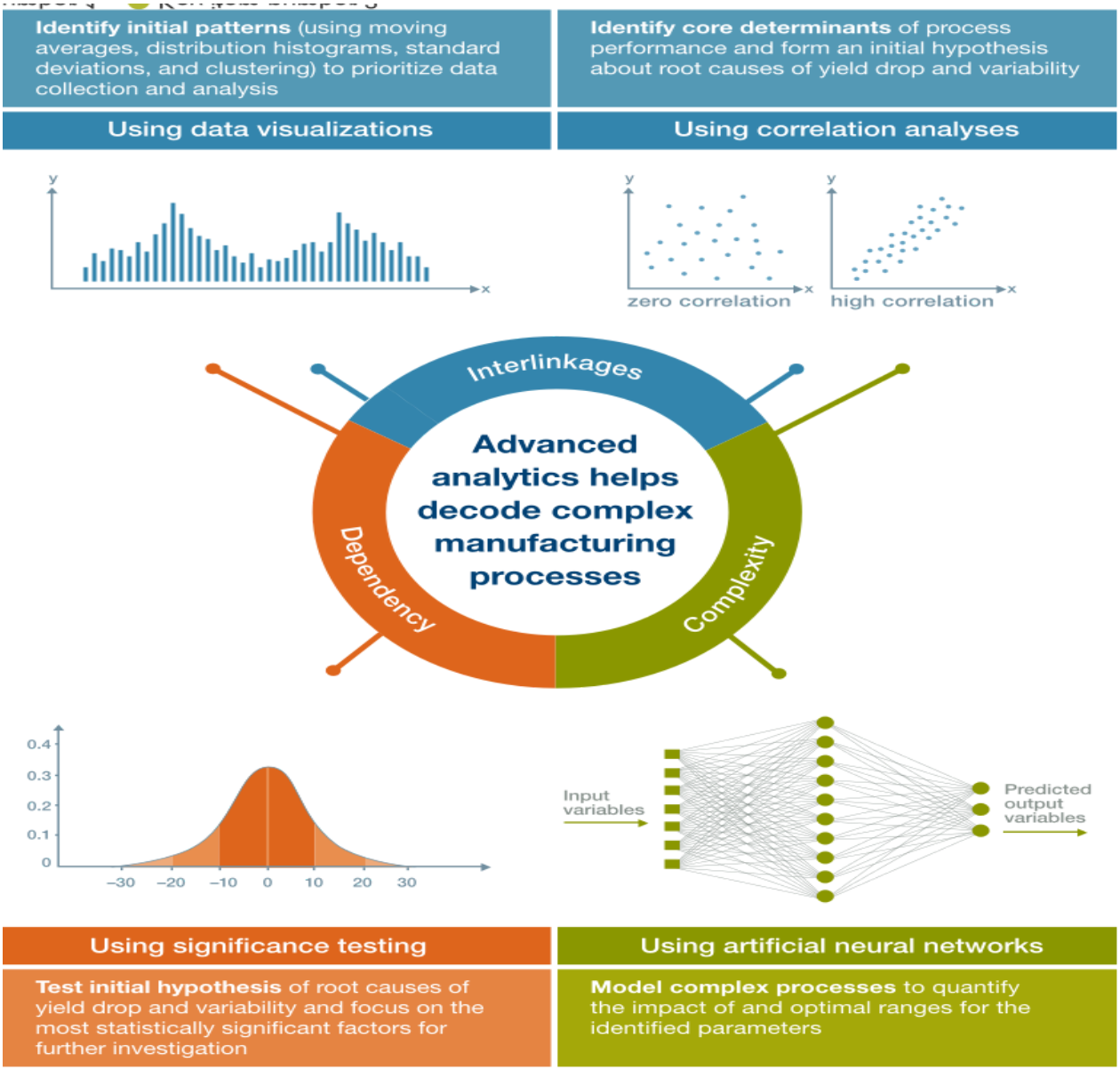
Big Data Analytics for Industry.4.0

Big data analytics

Some data are irrelevant or relevant for system and users.So to increase the accuracy and efficiency utilization data taken from huge amount should be analyze .Big data analytics is the use of advanced computing technologies on huge data sets to discover valuable correlations, patterns, trends, and preferences for companies to make better decisions. In Industry 4.0, big data analytics plays a role in a few areas including in **smart factories**, where sensor data from production machinery is analyzed to predict when maintenance and repair operations will be needed. Through application of it, manufacturers experience production efficiency, understand their real-time data with self-service systems, predictive maintenance optimization, and production management automation.



Analytical techniques include predictive modeling, machine learning, simulation, process optimization and control along with business intelligence tools. However, performing analytics through an all-inclusive technology can be time-consuming and expensive. The infographic below showcases different analytical techniques used to generate manufacturing insights.



Use of big data analytics in business

Businesses use big data analytics to improve business decisions by understand patterns and picking up on trends from huge amounts of customer data.

Big data analytics used in Industry 4.0

Manufacturers use big data analytics in the same way as most other commercial entities except with a narrower focus. They collect huge amounts of data from smart sensors through cloud computing and IIoT platforms that allow them to uncover patterns that help them improve the efficiency of supply chain management.

Big data analytics can help them discover hidden variables causing bottlenecks in production that they didn't even know existed. After identifying the source of the problem, manufacturers

use targeted data analytics to better understand the underlying cause of bottleneck variables. This helps manufacturers improve output while reducing cost and eliminating waste.

Production efficiency and assets are everything the manufacturing industry. The manufacturers' ability to maintain their means of production and keep schedules tight and on track can mean the difference between a good reputation and a bad one. Big data analytics reduces breakdowns and unscheduled downtime by about 25 percent.

Big data analytics is crucial to real-time performance, supply chain optimization, price optimization, fault prediction, product development, and smart factory design.

There are four main types of big data analytics that support and inform different business decisions.

1 Descriptive analytics

Descriptive analytics refers to data that can be easily read and interpreted. This data helps create reports and visualize information that can detail company profits and sales.

Example: During the pandemic, a leading pharmaceuticals company conducted data analysis on its offices and research labs. Descriptive analytics helped them identify unutilized spaces and departments that were consolidated, saving the company millions of dollars.

2. Diagnostics analytics

Diagnostics analytics helps companies understand why a problem occurred. Big data technologies and tools allow users to mine and recover data that helps dissect an issue and prevent it from happening in the future.

Example: A clothing company's sales have decreased even though customers continue to add items to their shopping carts. Diagnostics analytics helped to understand that the payment page was not working properly for a few weeks.

3. Predictive analytics

Predictive analytics looks at past and present data to make predictions. With artificial intelligence (AI), machine learning, and data mining, users can analyze the data to predict market trends.

Example: In the manufacturing sector, companies can use algorithms based on historical data to predict if or when a piece of equipment will malfunction or break down.

4. Prescriptive analytics

Prescriptive analytics provides a solution to a problem, relying on AI and machine learning to gather data and use it for risk management.

Example: Within the energy sector, utility companies, gas producers, and pipeline owners identify factors that affect the price of oil and gas in order to hedge risks.

Otherway methods for BIG DATA ANALYSIS can be represented as

Association rule learning
Classification tree analysis
Genetic algorithms
Machine learning
Regression analysis
Sentiment analysis
Social network analysis

Tools used in big data analytics

Various technology has advanced so that there are many intuitive software systems available for data analysts to use.

Hadoop: *An open-source framework that stores and processes big data sets. Hadoop is able to handle and analyze structured and unstructured data.*

Spark: *An open-source cluster computing framework used for real-time processing and analyzing data.*

Data integration software: *Programs that allow big data to be streamlined across different platforms, such as MongoDB, Apache, Hadoop, and Amazon EMR.*

Stream analytics tools: *Systems that filter, aggregate, and analyze data that might be stored in different platforms and formats, such as Kafka.*

Distributed storage: *Databases that can split data across multiple servers and have the capability to identify lost or corrupt data, such as Cassandra.*

Predictive analytics hardware and software: *Systems that process large amounts of complex data, using machine learning and algorithms to predict future outcomes, such as fraud detection, marketing, and risk assessments.*

Data mining tools: *Programs that allow users to search within structured and unstructured big data.*

NoSQL databases: *Non-relational data management systems ideal for dealing with raw and unstructured data.*

Data warehouses: *Storage for large amounts of data collected from many different sources, typically using predefined schemas.*