WHAT IS BIG DATA?

“Big Data” is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it.Big data is a combination of structured, semistructured and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications.

WHY IS BIG DATA IMPORTANT?

Companies use big data in their systems to improve operations, provide better customer service, create personalized marketing campaigns and take other actions that, ultimately, can increase revenue and profits. Businesses that use it effectively hold a potential competitive advantage over those that don’t because they’re able to make faster and more informed business decisions.

Eg: big data provides valuable insights into customers that companies can use to refine their marketing, advertising and promotions in order to increase customer engagement and conversion rates. Both historical and real-time data can be analyzed to assess the evolving preferences of consumers or corporate buyers, enabling businesses to become more responsive to customer wants and needs.

CHARACTERSTICS OF BIG DATA

1.Data is begin generated fast and need to be processed fast
2.Online Data Analytics
3.Late decisions implies missing opportunities
4.Examples

E-Promotions: Based on your current location, your purchase history, what you like;send promotions right now for store next to you

Healthcare monitoring: sensors monitoring your activities and body;any abnormal measurements require immediate reaction

FOUR V’s OF BIG DATA

1.Volume(data at rest):terabyte to exabyte of existing data to process
2.Velocity(data in motion):streaming data, ms to s to respond
3.Variety(data in various forms):structured, unstructured,text,multimedia
4.Veracity(data in doubt):uncertainity due to data inconsistency and incompleteness, ambiguities, latency, deception, model approximations

DRIVERS FOR BIG DATA

Big Data has quickly risen to become one of the most desired topics in the industry. The main business drivers for such rising demand for Big Data Analytics are:

1. The digitization of society
2. The drop in technology costs
3. Connectivity through cloud computing
4. Increased knowledge about data science
5. Social media applications
6. The rise of Internet-of-Things (IoT)

Example: A number of companies that have Big Data at the core of their strategy like :
Apple, Amazon, Facebook and Netflix have become very successful at the beginning of the 21st century.

WHAT IS BIG DATA ANALYTICS?

Big Data analytics is a process used to extract meaningful insights, such as hidden patterns, unknown correlations, market trends, and customer preferences. Big Data analytics provides various advantages—it can be used for better decision making, preventing fraudulent activities, among other things.

There are many different ways that Big Data analytics can be used in order to improve businesses and organizations. Here are some examples:

•Using analytics to understand customer behavior in order to optimize the customer experience
•Predicting future trends in order to make better business decisions

•Improving marketing campaigns by understanding what works and what doesn't
•Increasing operational efficiency by understanding where bottlenecks are and how to fix them
•Detecting fraud and other forms of misuse sooner

ADVANTAGES OF BIG DATA ANALYTICS

1.Risk management

Use Case: Banco de Oro, a Phillippine banking company, uses Big Data analytics to identify fraudulent activities and discrepancies. The organization leverages it to narrow down a list of suspects or root causes of problems.

2.Product development and innovation

Use Case: Rolls-Royce, one of the largest manufacturers of jet engines for airlines and armed forces across the globe, uses Big Data analytics to analyze how efficient the engine designs are and if there is any need for improvements.

3.Quicker and better decision making

Use Case: Starbucks uses Big Data analytics to make strategic decisions. For example, the company leverages it to decide if a particular location would be suitable for a new outlet or not. They will analyze several different factors, such as population, demographics, accessibility of the location, and more.

4.Improved customer experience

Use Case: Delta Air Lines uses Big Data analysis to improve customer experiences. They monitor tweets to find out their customers' experience regarding their journeys, delays, and so on. The airline identifies negative tweets and does what's necessary to remedy the situation. By publicly addressing these issues and offering solutions, it helps the airline build good customer relations

LIFECYCLE/PHASES OF BIG DATA ANALYTICS

•Stage 1 - Business case evaluation - The Big Data analytics lifecycle begins with a business case, which defines the reason and goal behind the analysis.
•Stage 2 - Identification of data - Here, a broad variety of data sources are identified.
•Stage 3 - Data filtering - All of the identified data from the previous stage is filtered here to remove corrupt data.
•Stage 4 - Data extraction - Data that is not compatible with the tool is extracted and then transformed into a compatible form.
•Stage 5 - Data aggregation - In this stage, data with the same fields across different datasets are integrated.
•Stage 6 - Data analysis - Data is evaluated using analytical and statistical tools to discover useful information.
•Stage 7 - Visualization of data - With tools like Tableau, Power BI, and QlikView, Big Data analysts can produce graphic visualizations of the analysis.
•Stage 8 - Final analysis result - This is the last step of the Big Data analytics lifecycle, where the final results of the analysis are made available to business stakeholders who will take action.

BIG DATA INDUSTRY APPLICATION

1.E-Commerce(predicts customer trends and optimizing prices)
2.Marketing(to drive high roi marketing campaigns which result in improves sales)
3.Education(develop new and improve existing courses based on market requirment)
4.Healthcare(with patient's medical history,bda predicts how likely they are to have health issues)
5.Media and Entertainment(understand demands of shows,movies,songs..to deliver a personalized list to customer)
6.Banking(customer's income and spending pattern helps predict chosing a banking offer like loan,credit card..)
7.Telecommunication(forecast n/w capacity and improve customer experience)
8.Government(lawenforcement and other..)

BID DATA ANALYTICS TOOLS

•Hadoop - helps in storing and analyzing data
•MongoDB - used on datasets that change frequently

•Talend - used for data integration and management
•Cassandra - a distributed database used to handle chunks of data
•Spark - used for real-time processing and analyzing large amounts of data
•STORM - an open-source real-time computational system
•Kafka - a distributed streaming platform that is used for fault-tolerant storage

TYPES OF BIG DATA ANALYTICS
1. Descriptive Analytics
This summarizes past data into a form that people can easily read. This helps in creating reports, like a co mpany's revenue, profit, sales, and so on. Also, it helps in the tabulation of social media metrics.

2. Diagnostic Analytics
This is done to understand what caused a problem in the first place. Techniques like drill-down, data mini ng, and data recovery are all examples. Organizations use diagnostic analytics because they provide an i n-depth insight into a particular problem.

3. Predictive Analytics
This type of analytics looks into the historical and present data to make predictions of the future. Predictiv e analytics uses data mining, AI, and machine learning to analyze current data and make predictions abo ut the future. It works on predicting customer trends, market trends, and so on.

4. Prescriptive Analytics
This type of analytics prescribes the solution to a particular problem. Perspective analytics works with bot h descriptive and predictive analytics. Most of the time, it relies on AI and machine learning.

BIG DATA TECHNOLOGIES
The 2 big data technologies are:-
1.NoSQL(Not Only SQL)
2.HADOOP

NOSQL(NOT ONLY SQL)
It is a light weight ,open source, non relational database that did not expose the standard SQL interface. NoSQL databases are widely used in bigdata and other real time web applications.

FEATURES OF NOSQL
1. NoSQL databases are non-relational
2. Distributed
3. No support for ACID properties but follows CAP(Consistency, Availability, Partition tolerance)
4. No fixed table schema

TYPES OF NOSQL DATABASES
1.Key-Value pair: the data in nosql db are stored in a key-value pair format i.e a big hash table is maintain ed to store these key-value pairs.
2.Document-Type: the nosql db stores the data as a collection in terms of documents.
3.Column-Oriented: the data of nosql db is stored in blocks where each block contains one column data. Eg: cassandra, hbase
4.Graph-based: this type of db is also called the network based model where the data is stored in the nod es of the graph

WHY NOSQL?
1. follows scale-out architecture over monolithic architecture
2. auto-sharding : to balance the work load, nosql automatically partitons the larger chunks into smaller on

es
3. schema-less: nosql db tries to store the data in it's native form
4. replication: tries to maintain multiple copies of the same data
5. support scale-in and scale-out

HADOOP
->Hadoop is an opensource platform for storage and processing of diverse data types that enables data-driven enterprises to rapidly derive the complete value from all their data.
->The scale and variety of data have permanently overwhelmed the ability to cost-effectively extract value using traditional platforms.
->The scalability and elasticity of free, open-source Hadoop running on standard hardware allow organizations to hold onto more data than ever before.
->Hadoop handles a variety of workloads, including search, log processing, recommendation systems, data warehousing, and video/image analysis.
->Apache Hadoop is an open-source project Hadoop is able to store any kind of data in its native format and to perform a wide variety of analyses and transformations on that data.
->Hadoop stores terabytes, and even petabytes, of data inexpensively. It is robust and reliable and handles hardware and system failures automatically, without losing data or interrupting data analyses.
->Hadoop runs on clusters of commodity servers and each of those servers has local CPUs and disk storage that can be leveraged by the system.

COMPONENTS OF HADOOP
1.HDFS(Hadoop Distributed File System)
HDFS is the storage system for a Hadoop cluster. When data lands in the cluster, HDFS breaks it into pieces and distributes those pieces among the different servers participating in the cluster. Each server stores just a small fragment of the complete data set, and each piece of data is replicated on more than one server.
2. Map Reduce
Because Hadoop stores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data. Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into a comprehensive answer. MapReduce is the agent that distributes the work and collects the results.

DATA DISCOVERY
Works the way people's mind works.
Tableau Software and QlikTech International(Qlikview)

CLOUD AND BIG DATA
With a cloud model, you pay on a subscription basis with no upfront capital expense. You don't incur the typical 30 percent maintenance fees, and all the updates on the platform are automatically available.
The ability to build massively scalable platforms—platforms where you have the option to keep adding new products and services for zero additional cost—is giving rise to business models that weren't possible before.

OPEN SOURCE TECHNOLOGY FOR BIG DATA ANALYTICS
•Open-source software is computer software that is available in source code form under an open-source license that permits users to study, change, and improve and at times also to distribute the software.

•Although the source code is released, there are still governing bodies and agreements in place. The most prominent and popular example is the GNU Gen- eral Public License (GPL), which "allows free distribution under the condition that further developments and applications are put under the same license."This ensures that the products keep improving over time for the greater population of users.

•Some other open-source projects are managed and supported by commer- cial companies, such as Clou dera, that provide extra capabilities, training, and professional services that support open-source projects such as Hadoop.

•You can make it into what you want and what you need. If you come up with an idea, you can put it to wo rk immediately. That's the advantage of the open- source stack—flexibility, extensibility, and lower cost."

•"One of the great benefits of open source lies in the flexibility of the adoption model: you download and d eploy it when you need it".

•Pace of software development has accelerated dramatically because of open-source software.

•The old model was top-down, slow, inflexible and expensive. The new software development model is bo ttom-up, fast, flexible, and considerably less costly.

•A traditional proprietary stack is defined and controlled by a single vendor, or by a small group of vendors . It reflects the old com- mand-and-control mentality of the traditional corporate world and the old economi c order.

•An open-source stack is defined by its community of users and con- tributors. No one "controls" an open-source stack, and no one can predict exactly how it will evolve. The open-source stack reflects the new re alities of the networked global economy, which is increasingly dependent on big data.

PREDICTIVE ANALYSIS

It is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of futu re outcomes based on historical data. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future.

WHY PREDICTIVE ANALYTICS?

1.Growing volumes and types of data, and more interest in using data to produce valuable insights.
2.Faster, cheaper computers.
3.Easier-to-use software.
4.Tougher economic conditions and a need for competitive differentiation.
5.With interactive and easy-to-use software becoming more prevalent, predictive analytics is no longer jus t the domain of mathematicians and statisticians. Business analysts and line-of-business experts are usin g these technologies as well.

USE OF PREDICTIVE ANALYTICS/ADVANTAGES/IMPORTANCE

1.Detecting fraud
2.Optimizing marketing campaign
3.Improving operations
4.Reducing risk

HOW DOES PREDICTIVE ANALYTICS WORK?

Predictive models use known results to develop (or train) a model that can be used to predict values for di fferent or new data. Modeling provides results in the form of predictions that represent a probability of the target variable (for example, revenue) based on estimated significance from a set of input variables.