Unit 1

1) Explain Big Data and its importance

Big data is a term defined for data sets that are large or complex that traditional data processing applications are inadequate. Big Data basically consists of analysis zing, capturing the data, data creation, searching, sharing, storage capacity, transfer, visualization, and querying and information privacy.

- **Big Data** is a collection of large datasets that cannot be adequately processed using traditional processing techniques. Big data is not only data it has become a complete subject, which involves various tools, techniques and frameworks.

- Big data term describes the volume amount of data both structured and unstructured manner that adapted in day-to-day business environment. It's important that what organizations utilize with these with the data that matters.
- Big data helps to analyze the in-depth concepts for the better decisions and strategic taken for the development of the organization.

The importance of big data is how you utilize the data which you own. Data can be fetched from any source and analyze it to solve that enable us in terms of

1) Cost reductions

2) Time reductions,

3) New product development and optimized offerings, and

4) Smart decision making.

Combination of big data with high-powered analytics, you can have great impact on your business strategy such as:

- Finding the root cause of failures, issues and defects in real time operations.
- Generating coupons at the point of sale seeing the customer's habit of buying goods.
- Recalculating entire risk portfolios in just minutes.
- Detecting fraudulent behavior before it affects and risks your organization.

2) Explain structured, semi-structured and unstructured data.

Big Data includes huge valume, high velocity, and extensible variaty of data. These are 3 types: Structured data, Semi-structured data, and Unstructured data.

1. **Structured data –**
   Structured data is a data whose elements are addressable for effective analysis. It has been organised into a formatted repository that is typically a database. It concern all data which can be stored in database SQL in table with rows and columns. They have relational key and can easily mapped into pre-designed fields. Today, those data are most processed in development and simplest way to manage information. *Example:* Relational data.
2. **semi-structured data –**
   Semi-structured data is information that does not reside in a rational database but that have some organizational properties that make it easier to analyze. With some process, you can store them in the relation database (it could be very hard for some kind of semi-structured data), but Semi-structured exist to ease space. *Example*: XML data.
3. **Unstructured data –**
   Unstructured data is a data that is which is not organised in a pre-defined manner or does not have a pre-defined data model, thus it is not a good fit for a mainstream relational database. So for Unstructured data, there are alternative platforms for storing and managing, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications. *Example*: Word, PDF, Text, Media logs.

Differences between Structured, Semi-structured and Unstructured data:

PROPERTIES	STRUCTURED DATA	SEMI-STRUCTURED DATA	UNSTRUCTURED DATA
Technology	It is based on Relational database table	It is based on XML/RDF	It is based on character and binary data
Transaction management	Matured transaction and various concurrency technique	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	Versioning over tuples,row,tables	Versioning over tuples or graph is possible	Versioned as whole

PROPERTIES	STRUCTURED DATA	SEMI-STRUCTURED DATA	UNSTRUCTURED DATA
Flexibility	It is sehema dependent and less flexible	It is more flexible than structuded data but less than flexible than unstructured data	it very flexible and there is abbsence of schema
Scalability	It is very difficult to scale DB schema	It's scaling is simpler than sstructured data	It is very scalable
Robustness	Very robust	New technology, not very spread	—
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual query are possible

3) Explain feature of Big Data. (Four Vs)

VOLUME

The main characteristic that makes data "big" is the sheer volume. It makes no sense to focus on minimum storage units because the total amount of information is growing exponentially every year. In 2010, Thomson Reuters estimated in its annual report that it believed the world was "awash with over 800 exabytes of data and growing."

For that same year, EMC, a hardware company that makes data storage devices, thought it was closer to 900 exabytes and would grow by 50 percent every year. No one really knows how much new data is being generated, but the amount of information being collected is huge.

VARIETY

Variety is one the most interesting developments in technology as more and more information is digitized. Traditional data types (structured data) include things on a bank statement like date, amount, and time. These are things that fit neatly in a relational database.

Structured data is augmented by unstructured data, which is where things like Twitter feeds, audio files, MRI images, web pages, web logs are put — anything that can be captured and stored but doesn't have a *meta model* (a set of rules to frame a concept or idea — it defines a class of information and how to express it) that neatly defines it.

Unstructured data is a fundamental concept in big data. The best way to understand unstructured data is by comparing it to structured data. Think of *structured data* as data that is well defined in a set of rules. For example, money will always be numbers and have at least two decimal points; names are expressed as text; and dates follow a specific pattern.

With *unstructured data,* on the other hand, there are no rules. A picture, a voice recording, a tweet — they all can be different but express ideas and thoughts based on human understanding. One of the goals of big data is to use technology to take this unstructured data and make sense of it.

VERACITY

Veracity refers to the trustworthiness of the data. Can the manager rely on the fact that the data is representative? Every good manager knows that there are inherent discrepancies in all the data collected.

VELOCITY

Velocity is the frequency of incoming data that needs to be processed. Think about how many SMS messages, Facebook status updates, or credit card swipes are being sent on a particular telecom carrier every minute of every day, and you'll have a good appreciation of velocity. A streaming application like Amazon Web Services Kinesis is an example of an application that handles the velocity of data.

VALUE

It may seem painfully obvious to some, but a real objective is critical to this mashup of the four V's. Will the insights you gather from analysis create a new product line, a cross-sell opportunity, or a cost-

cutting measure? Or will your data analysis lead to the discovery of a critical causal effect that results in a cure to a disease?

4) What are the drivers for big data?

Business

1. Opportunity to enable innovative new business models
2. Potential for new insights that drive competitive advantage

Technical

1. Data collected and stored continues to grow exponentially
2. Data is increasingly everywhere and in many formats
3. Traditional solutions are failing under new requirements

Financial

1. Cost of data systems, as a percentage of IT spend, continues to grow
2. Cost advantages of commodity hardware & open source software

(https://analyticsweek.com/content/12-drivers-bigdata-analytics/)

5) What do you mean by big data analytics?

Big data analytics is the often complex process of examining large and varied data sets -- or big data -- to uncover information including hidden patterns, unknown correlations, market trends and customer preferences that can help organizations make informed business decisions.

On a broad scale, data analytics technologies and techniques provide a means to analyze data sets and draw conclusions about them to help organizations make informed business decisions. BI queries answer basic questions about business operations and performance.

Big data analytics is a form of advanced analytics, which involves complex applications with elements such as predictive models, statistical algorithms and what-if analysis powered by high-performance analytics systems.

The importance of big data analytics

Driven by specialized analytics systems and software, as well as high-powered computing systems, big data analytics offers various business benefits, including new revenue opportunities, more effective marketing, better customer service, improved operational efficiency and competitive advantages over rivals.

Big data analytics applications enable big data analysts, data scientists, predictive modelers, statisticians and other analytics professionals to analyze growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional business intelligence (BI) and analytics programs. That encompasses a mix of semi-structured and unstructured data -- for example, internet clickstream data, web server logs, social media content, text from customer emails and survey responses, mobile phone records, and machine data captured by sensors connected to the internet of things.

Why is big data analytics important?

Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers. In his report *Big Data in Big Companies*, IIA Director of Research Tom Davenport interviewed more than 50 businesses to understand how they used big data. He found they got value in the following ways:

1. **Cost reduction.** Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business.
2. **Faster, better decision making.** With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they've learned.

3. **New products and services.** With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

6) Explain big data application.

Banking

Large amounts of data streaming in from countless sources, banks have to find out unique and innovative ways to manage big data. It's important to analyze customers needs and provide them service as per their requirements, and minimize risk and fraud while maintaining regulatory compliance. Big data have to deal with financial institutions to do one step from the advanced analytics.

Government

When government agencies are harnessing and applying analytics to their big data, they have improvised a lot in terms of managing utilities, running agencies, dealing with traffic congestion or preventing the affects crime. But apart from its advantages in Big Data, governments also address issues of transparency and privacy.

Education

Educator regarding Big Data provides a significant impact on school systems, students and curriculums. By analyzing big data, they can identify at-risk students, ensuring student's progress, and can implement an improvised system for evaluation and support of teachers and principals in their teachings.

Health Care

When it comes to health care in terms of Patient records. Treatment plans. Prescription information etc., everything needs to be done quickly and accurately and some aspects enough transparency to satisfy stringent industry regulations. Effective management results in good health care to uncover hidden insights that improve patient care.

Manufacturing

Manufacturers can improve their quality and output while minimizing waste where processes are known as the main key factors in today's highly competitive market. Several manufacturers are working on analytics where they can solve problems faster and make more agile business decisions.

Retail

Customer relationship maintains is the biggest challenge in the retail industry and the best way to manage will be to manage big data. Retailers must have unique marketing ideas to sell their products to customers, the most effective way to handle transactions, and applying improvised tactics of using innovative ideas using BigData to improve their business.

7) Explain map-reduce algorithm.

https://www.tutorialspoint.com/map_reduce/map_reduce_algorithm.htm