UNIT-1

INTRODUCTION TO BIG DATA

UNIT-I

INTRODUCTION

Syllabus

What is big data, Meet Hadoop – Data, Characteristics of Big Data, Data Storage and Analysis, Comparison with other systems: Relational Database Management Systems, Grid computing and Volunteer Computing.

1.1 What is Big Data

Definition: Big Data is often described as extremely large data sets that have grown beyond the ability to manage and analyze them with traditional data processing tools.

- The data set has grown so large that it is difficult to manage and even harder to garner value out of it.
- The primary difficulties are the acquisition, storage, searching, sharing, analytics, and visualization of data. Not only the size of the data set but also difficult to process the data.

The data come from everywhere : Sensors used to gather climate information, posts to social media sites, digital pictures and videos posted online, transaction records of online purchases, and cell phone GPS signals etc.,

All of these data have intrinsic value that can be extracted using analytics, algorithms, and other technique.

Why Big Data is Important

- Big Data solutions are ideal for analyzing not only raw structured data, but semi structured and unstructured data from a wide variety of sources.
- Big Data solutions are ideal when all, or most, of the data needs to be analyzed versus a sample of the data; or a sampling of data isn't nearly as effective as a larger set of data from which to derive analysis.
- Big Data solutions are ideal for iterative and exploratory analysis when business measures on data are not predetermined.

- Big Data is well suited for solving information challenges that don't natively fit within a traditional relational database approach for handling the problem at hand.

Big Data has already proved its importance and value in several areas. Organizations such as the National Oceanic and Atmospheric Administration

(NOAA), the National Aeronautics and Space Administration (NASA), several pharmaceutical companies, and numerous energy companies have amassed huge amounts of data and now leverage Big Data technologies on a daily basis to extract value from them.

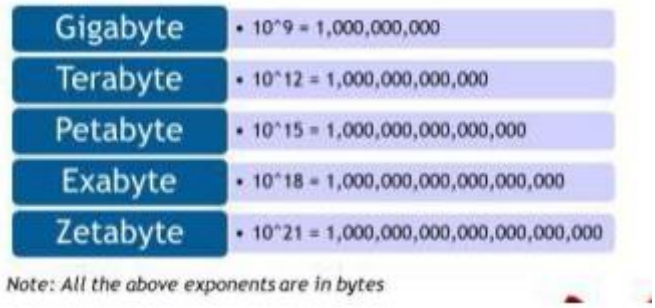
NOAA uses Big Data approaches to aid in climate, ecosystem, weather, and commercial research,

- NASA uses Big Data for aeronautical and other research.
- Pharmaceutical companies and energy companies have leveraged Big Data.
- for more tangible results. such as drug testing and geophysical analysis.
- The New York Times has used Big Data tools for text analysis and Web Mining.
- Walt Disney Company uses them to correlate and understand customer behavior in all of its stores, theme parks.
- Companies such as Facebook, Amazon, and Google rely on Big Data analytics a part of their primary marketing schemes as well as a means of servicing their customers better.
- accomplished by storing each customer's searches and purchases and other piece of information available, and then applying algorithms to that information to compare one customer's information with all other customers information.
- Big Data plays another role in today's businesses: Large organizations
- increasingly face the need to maintain massive amounts of structured and unstructured data—from transaction information in data warehouses to employee tweets, from supplier records to regulatory filings—to comply with government regulations.

1,2 Meet Hadoop- data

Data

Every day zeta bytes or peta bytes of data is generated by People and machines.



If the amount of data is more than hundreds of terabytes then such a data is called as big data.

Data generated by People:

Through individual interactions

-Phone calls

- emails

- documents .

Through social media

–twitter

-facebook

-whatsup etc.

Data generated by Machines:

-RFID readers

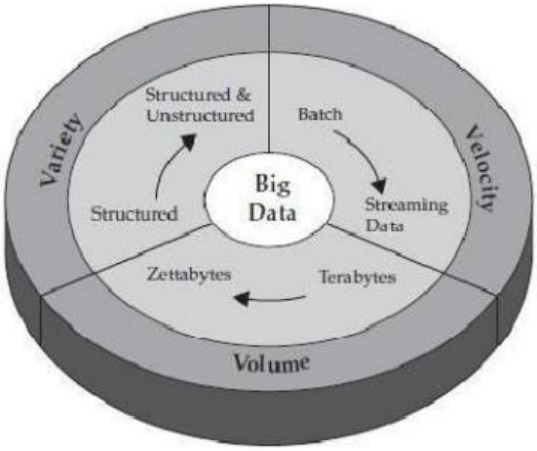
-Sensor networks

-Vehicle GPS traces

-Machine logs

1.3 Characteristics of Big Data

Three characteristics define Big Data: **Volume, Variety**, and **Velocity**



1,3,1 Volume: The amount of data

- The size of available data has been growing at an increasing rate. The Volume of data is growing. Experts predict that the volume of data in the world will grow to 35 Zetta bytes in 2020.
- Twitter alone generates more than 7 tera bytes(TB) of data every day. Facebook 10TB.That same phenomenon affects every business – their data is growing at the same exponential rate too.
- A text file is a few kilo bytes, a sound file is a few mega bytes while a full length movie is a few giga bytes. More sources of data are added on continuous basis.
- For companies, in the old days, all data was generated internally by employees. Currently, the data is generated by employees, partners and customers.
- For a group of companies, the data is also generated by machines. For example, Hundreds of millions of smart phones send a variety of information to the network infrastructure.
- We store everything: Environmental data, financial data, medical data, surveillance data.
- Peta byte data sets are common these days and Exa byte is not far away.

1.3.2 Velocity: How fast it is generated

- Data is increasingly accelerating the velocity at which it is created and at which it is integrated. We have moved from batch to a real-time business.
- Initially, companies analyzed data using a batch process. One takes a chunk of data, submits a job to the server and waits for delivery of the result. That scheme works when the incoming data rate is slower than the batch-processing rate and when the result is useful despite the delay.
- With the new sources of data such as social and mobile applications, the batch process breaks down. The data is now streaming into the server in real time, in a continuous fashion and the result is only useful if the delay is very short.

\1.3.3 Variety: Represents all kinds of data

Data can be classified under several categories : structured data, semi structured data and unstructured data.

Structured data are normally found in traditional databases (SQL or others) where data are organized into tables based on defined business rules. Structured data usually prove to be the easiest type of data to work with, simply because the data are defined and indexed, making access and filtering easier.

Unstructured data, are not organized into tables and cannot be natively used by applications or interpreted by a database. A good example of unstructured data would be a collection of binary image files.

Semi-Structured data fall between unstructured and structured data. Semi-structured data do not have a formal structure like a database with tables and relationships. However, unlike unstructured data, semi-structured data have tags or other markers to separate the elements and provide a hierarchy of records and fields, which define the data.

- Big data extend beyond structured data to include unstructured data off all varieties: text, audio, video, click streams, log files and more.
- The growth in data sources has fuelled the growth in data types. In fact, 80% of the world's data is unstructured and only 20% structured data . Yet most traditional methods apply analytics only to structured information.

1.4 Data Storage and Analysis

- Struggling with storage and analysis of the data.
- Even though the storage capacities of hard drives have increased massively over the years, access speeds(the rate at which data can be read from drives).
- Take long time to read all data on a single drive—and writing is even slower. The obvious way to reduce the time is to read from multiple disks at once
- Ex : if we had 100 drives, each holding one hundredth of the data.

Working in parallel, we could read the data in under two minutes. Even though read and write data in parallel to or from multiple disks , there are some more problems.

First Problem: Hardware failure.

As soon as you start using many pieces of hardware, the chance that one will fail is fairly high.

A common way of avoiding data loss is through replication: redundant copies of the data are kept by the system so that in the event of failure, there is another copy available. This is how RAID(redundant array of inexpensive disks)works.

Second problem : most analysis tasks need to be able to combine the data in some way; i.e data read from one disk may need to be combined with the data from any of the other 99 disks.

Solution for above problems:

Building distributed systems—for data storage, data analysis, and coordination.

Hadoop provides: a reliable shared storage and analysis system. The storage is provided by **HDFS** and analysis by **MapReduce**.

- **HDFS** - Hadoop Disributed File System . It avoids data loss is through replication. Minimum of three replicas for the data.
- **MapReduce-** Programming model. It abstracts the problem from disk reads and writes, transforming it into a computation over sets of keys and values.

1.5 Comparison with other systems

- The approach taken by MapReduce may seem like a **brute-force** approach on the entire dataset—or at least a good portion of it—is processed for each query.
- MapReduce is a **batch** query processor, and the ability to run an ad hoc query against the whole dataset and get the results in a reasonable time is transformative.
- It changes the way you think about data, and unlocks data that was previously archived on tape or disk.
- Why can't we use databases with lots of disks to do large-scale batch analysis? Why is MapReduce needed?
- MapReduce can be seen as a complement to an RDBMS. The differences between the two systems are shown in Table

	Traditional RDBMS	MapReduce
Data Size	Gigabytes	Petabytes
Update	Read and Write Many Times	Write Once, read many times
Structure	Static Schema	Dynamic Schema
Integrity	High	Low
Scaling	Nonlinear	Linear

- MapReduce is a good fit for problems that need to analyze the whole dataset, in a batch fashion, particularly for ad hoc analysis. RDBMS is good for point queries or updates, where the dataset has been indexed to deliver low-latency retrieval and update times of a relatively small amount of data.
- MapReduce suits applications where the data is written once, and read many times.Relational database is good for datasets that are continually updated.
- Another difference is the amount of structure in the datasets that they operate on.

RDBMS operate on Structured data is data that is organized into entities that have a defined format, such as XML documents or database tables that conform to a particular predefined schema. Map Reduce operate on Semistructured and Unstructured data. In Semi-structured data there may be a schema, it is often ignored, so it may be used only as a guide to the structure of the data.

Ex : Spreadsheet, in which the structure is the grid of cells, although the cells themselves may hold any form of data.Unstructured data does not have any particular internal structure.

Ex : plain text or image data. MapReduce works well on unstructured or semistructured data, since it is designed to interpret the data at processing time.

- Relational data is normalized to retain its integrity(assurance of accuracy) and remove redundancy. Normalization poses problems for MapReduce, since it makes reading a record a nonlocal operation, and one of the central assumptions that MapReduce makes is that it is possible to perform (high-speed) streaming reads and writes.

Ex : Web server log is a good example of a set of records that is not normalized .

The client hostnames are specified in full each time, even though the same client may appear many times and this is one reason that logfiles of all kinds are particularly well-suited to analysis with MapReduce.

- MapReduce is a linearly scalable programming model. The programmer writes two functions—a map function and a reduce function—each of which defines a mapping from one set of key-value pairs to another.
- These functions are unmind to the size of the data or the cluster that they are operating on, so they can be used unchanged for a small dataset and for a massive one.
- if you double the size of the input data, a job will run twice as slow. But if you also double the size of the cluster, a job will run as fast as the original one. This is not generally true of SQL queries.

1.6 Grid Computing

1) The HPC and Grid computing doing large scale data processing using APIs as Message Passing Interface(MPI).

The approach of HPC is to distribute the work across a cluster of machines

- Which access shared files system
- Hosted by a Storage Area Network(SAN) - Works well for compute intensive jobs.
- It face problem when nodes need to access larger data volumes i.e hundreds of giga bytes. Reason is the network bandwidth is the bottleneck and computer nodes become idle. (At this point Hadoop starts shines).
- MapReduce tries to collocate the data with the compute node, so data access is fast since it is local. This feature, known as data locality, is at the heart of MapReduce and is the reason for its good performance.
- Network bandwidth is more precious resource in the data center environment(easy to saturate network links by copying data around).
- Hadoop models its network topology by consuming bandwidth as less as possible. It does not prevent high –CPU analysis in hadoop.

2)MPI gives great control to the programmer, but requires that explicitly handle the mechanics of the

- data flow
- exposed via low-level C routines
- constructs, such as sockets
- the higher-level algorithm for the analysis.

MapReduce operates only at the higher level: the programmer thinks in terms of functions of key and value pairs, and the data flow is implicit.

3) Coordinating the processes in a large-scale distributed computation is a challenge.

The hardest aspect is gracefully handling partial failure—you don’t know if a remote process has failed or not.

MapReduce spares the programmer from having to think about failure, since the implementation detects failed map or reduce tasks and reschedules replacements on machines that are healthy.

MapReduce is able to do this since it is a shared-nothing architecture, meaning that tasks have no dependence on one other.

(This is a slight oversimplification, since the output from mappers is fed to the reducers, but this is under the control of the MapReduce system; it needs to take more care rerunning a failed reducer than rerunning a failed map, it has to make sure it can retrieve the necessary map outputs, and if not, regenerate them by running the relevant maps again.)

- the programmer's point of view, the order in which the tasks run doesn't matter.
- By contrast, MPI programs have to explicitly manage their own check pointing and recovery, which gives more control to the programmer, but makes them more difficult to write.
- MapReduce is a restrictive programming model, and in a sense it is: limited to key and value types that are related in specified ways, and mappers and reducers run with very limited coordination between one another.

MapReduce was invented by engineers at Google . It was inspired by older ideas from the functional programming, distributed computing, and database communities.

- Many applications in many industries use MR . It is pleasantly surprising to see the range of algorithms that can be expressed in MapReduce, from image analysis, to graph-based problems, to machine learning algorithms.
- It can't solve every problem, but it is a general data-processing tool.

1.7 Volunteer Computing

- Volunteer computing is one in which volunteers donate CPU time from their idle computers to analyze data.
- Volunteer computing projects work by breaking the problem they are trying to solve in to chunks called work unit.
- Work units are send to computers around the world to be analyzed.

Ex: SETI (the Search for Extra-Terrestrial Intelligence) runs a project SETI@home in which volunteers donate CPU time from their idle computers to analyze radio telescope data for signs of intelligent life outside earth.

- In SETI@home work unit is about 0.35 MB of radio telescope data, and takes hours or days to analyze on a typical home computer.
- When the analysis is completed, the results are sent back to the server, and the client gets another work unit
- As a precaution to combat cheating, each work unit is sent to three different machines and needs at least two results to agree to be accepted.
- SETI@home may be superficially similar to MapReduce (breaking a problem into independent pieces to be worked on in parallel).
- The difference is SETI@home problem is very CPU-intensive, which makes it suitable for running on hundreds of thousands of computers across the world.
- The time to transfer the work unit is very small by the time to run the computation on it. Volunteers are donating CPU cycles, not bandwidth.
- MapReduce is designed to run jobs that last minutes or hours on trusted, dedicated hardware running in a single data center with very high aggregate bandwidth interconnects.
- By contrast, SETI@home perform computation on untrusted machines on the Internet with highly variable connection speeds and no data locality.

ACT-QUESTIONS

A. Objective Questions

What are the main components of big data?

a) HDFS b) MapReduce c) YARN **d) all**

On which of the following platforms does Hadoop run?

a) Debian **b) cross-patform** c)bare metal d) unix

Data in ____ bytes size is called big data

a) Meata b) giga c) tera **d) peta**

Transaction of data of the bank is a type of.

a) Unstructured data **b) Structured data** c) Both a and b d) None of the above

The total forms of big data is ____

a) 1 b) 2 **c) 3** d) 4

Identify the incorrect big data Technologies.

a) Apache Pytorch **b) Apache Kafka** **c)** Apache Hadoop d) Apache Spark

_____ is a collection of data that is used in volume, yet growing exponentially with time

a) Big Database b) Big DBMS c) Big Datafile **d) Big Data**

Choose the primary characteristics of big data among the following

a) Value b) Variety **c) Volume** d) All of the above

Identify the different features of Big Data Analytics.

a) Open-source b) Data recovery c) Scalability **d) All of the above**

Total V's of big data is ____

a) 3 b) 4 **c) 5** d) 6

Amongst which of the following can be considered as the main source of unstructured data.

a) Twitter b) Facebook c) Webpages d) **All**

Big data deals with high-volume, high-velocity and high-variety information assets,

a) True b) False

What is the term used for a collection of large, complex data sets that cannot be processed using traditional data processing tools?

a) BigData b) Small Data c) Medium Data d) d) Mini Data

.Match the Following.

I) Volume	[ ]	a) different data formats
II)Velocity	[ ]	b)rate at which data grows
III)Variety	[ ]	c)uncertainty of available data
IV)Veracity	[ ]	d)amount of data

.Match the Following.

I) Semi structured Data	[ ]	a) images
II) Structured Data	[ ]	b) Bigdatacse@gmail.com
III) Unstructured Data	[ ]	c) Log Files

B. Descriptive Questions

1. Discuss the importance of Big Data?
2. Discuss the problems in data storage and analysis that motivate Big Data Analytics
3. Define Big Data. Explain its characteristics.
4. Compare and contrast the following terms for Big Data

   (i) Structured data (ii) Semi structured data (iii) Unstructured data
5. Illustrate the velocity characteristic of Big Data.
6. Describe about volume of data as a characteristic of Big Data.
7. List the companies who use the Hadoop tool to solve the Real world problems? Differentiate between grid computing and volunteer computing
8. Compare and contrast Hadoop with traditional RDBMS
9. Discuss the comparison of Hadoop MapReduce with other systems like Grid computing, RDBMS, Volunteer computing.