Class Note:

Subject: Big Data Technologies

Faculty: V. Divya

Topic: Big data

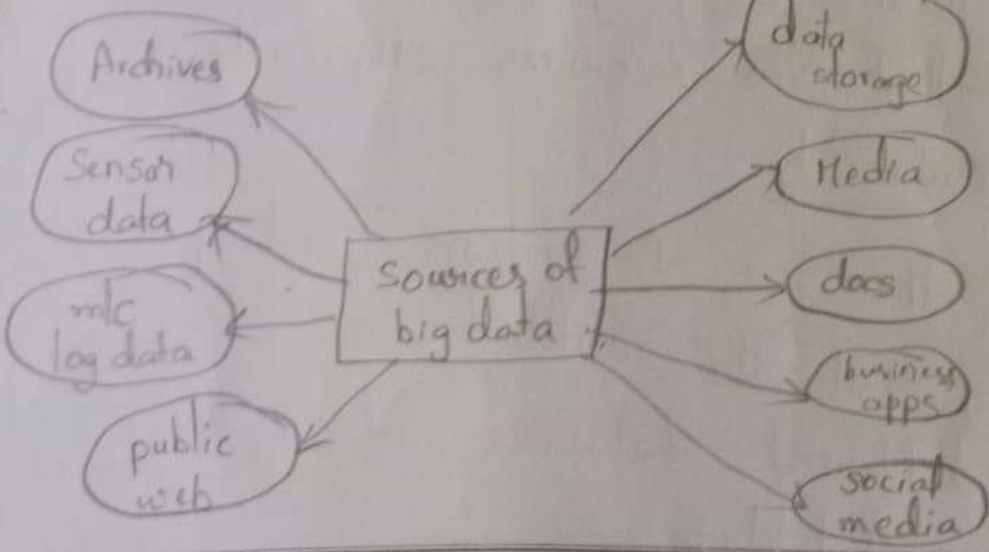
Unit No.: 1
Lecture No.: 1 7
Link to Session
Planner (SP) S.No. ....of SP
Date Conducted:
Page No. 1

Big data:-

- Data is information, usually in the form of facts or statistics that one can analyse or use for further calculations.
- Data is information that can be stored and used by a computer program.
- Data is information presented in numbers, letters & other form.
- Data is information from series of observations, measurement or facts.
- Information from series of behavioural observations, measurements or facts.

Characteristics of big data:-

- Volume
- Velocity
- Variety



- Volume:-

→ **Velocity :-**

Batch processing :- is a method computers use to periodically complete high volume repetitive data jobs.

Periodic processing :- is a function that involves performing tasks at regular intervals such as daily, weekly, monthly or yearly.

Near real time processing :- refers to the time delay b/w an event and use its processed data.

Real time processing :- is a computing method that evaluates input data immediately to produce o/ps in realtime.

→ **Variety :-**

- Structured data.
- Semistructured data.
- Unstructured data.

Scalability & Parallel processing :-

Vertical Scalability :-

- means scaling up the given system resources and increasing the system analytics, reporting and visualization capabilities
- this is an additional way to solve problems of greater complexities.
- Scaling up means designing the algorithm according to the architecture that uses resources efficiently.

Horizontol scalability :-

- Increases the no. of systems working in coherence and scaling out the workload.
- Processing different datasets of a large dataset deploys horizontal scalability.

Class Note:

Subject : Big data technologies

Faculty : V. Divya

Topic: Scalability & parallel processing, Designing data architecture, datasources

Unit No. : 1
Lecture No.: 2
Link to Session Planner (SP) S.No. ....of SP
Date Conducted:
Page No. 2

- scaling out means using more resources and distributing the processes and starts tasks in parallel.
- the easier way to scale up and scale out execution of analytics software is to implement it on a bigger m/c. with more CPUS for greater volume, velocity, variety.

Designing data architecture:-

It consists of five layers.

- Identification of data sources
- acquisition, ingestion, extraction, preprocessing, transformation of data
- Data storage of files, servers, clusters or cloud
- data processing
- data consumption in the number of programs & tools

Layer 5 Data consumption	Export of datasets to cloud, web	Datasets BPS, BLS, KDD	Analytics reporting visualizing
Layer4 Data processing	Processing tech Mapreduce, Hive Pig, Spark	Processing real time scheduled batches	Synchronous or asynchronous processing

Layer 3 Data storage	Consideration of types, formats, compression	Hadoop, DFs, spark S3	NosQL MongodB	
Layer 2 Data ingestion & acquisition	Ingestion using ELT	Data semantic	Pre processing	Ingestion of data
Layer 1 clarification of internal & external sources of data	sources for ingestion of data	Push or pull of data from the sources for ingestion	Datatypes for DB, files, web or series	Data formats semi/un structured

- structured
- Semi structured
- multi structured or unstructured.

- Structured data :-
  - Data source for ingestion, storage and processing can be a file, database or streaming data.
  - structured data sources are SQL server, MySQL, Microsoft access database, oracle DBMS.

→ Unstructured data sources :-
  - Distributed over highspeed netwoorks, data need high velocity processing, sources are from distributed file system
  - .txt, .csv data may be as key value pairs such as hash key value pair.

Datasources: Sensors, signal and GPS.
- The data sources can be sensors, sensor n/w signals from machines, devices controllers and intelligent edge node of different types in the industry M2M communicating and GPS system.

Class Note:

Subject : ......BDT..........

Faculty : .....V. Divya......

Topic: History of Big data, Designing data architecture, data sources

Unit No. : 1
Lecture No.: 3
Link to Session Planner (SP) S.No. ....of SP
Date Conducted:
Page No. 3

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 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 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 emergance of machine language learning has produced still more data.

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.

Big data vs Business intelligence:-

- Although big data and business intelligence are two technologies used to analyze data to help companies in the decision making process, there are differences between both of them.
- They differ in the way they work as much as in the type of data they analyze.
- Traditional BI methodology is based on the principle of grouping all business data into a central server.

Typically, this data is analyzed in offline mode, after storing the information in an environment called Datawarehouse.

The data is structured in a conventional relational database with an additional set of indexes and forms of access to the tables.

Main differences between Big data and business intelligence are:-

- In a Big data environment, information is stored on a

Class Note:

Subject : BDT

Faculty : V. Divya

Topic: Big data vs Business intelligence & Big data vs Datawarehouse.

Unit No. : 1
Lecture No.: 3
Link to Session Planner (SP) S.No. ....of SP
Date Conducted:
Page No. 4

central server. It is a much safer and more flexible space.

- Big data solutions carry the processing functions to the data, rather than the data to the functions. As the analysis is created centered on the information, it's easier to handle larger amounts of information in a more agile way.
- Big data can analyze data in different formats, both structured and unstructured. The volume of unstructured data is growing at levels much higher than the structured data.
- Data processed by Big data solutions can be historical or come from real-time sources. Thus, companies can make decisions that affect their business in an agile and efficient way.
- Big data technology uses parallel mass processing (MPP) concepts, which improves the speed of analysis. With MPP many instructions are executed simultaneously and since the various jobs are divided into several parallel execution parts, at the end the overall results

are reunited and presented. This allows you to analyze large volumes of information quickly.

Big data vs Data warehouse:

- Big data has become the reality of doing business for organizations today. There is a boom in the amount of structured as well as raw data that floods every organization daily. If this data is managed well, it can lead to powerful insights and quality decision making.
- Big data analytics is the process of examining large data sets containing a variety of datatypes to discover some knowledge in databases, to identify interesting patterns and establish relationships to solve problems, markt trends, customer preference. and other useful information. Companies and businesses that implement Big Data Analytics often reap several business benefits. Companies implement Big Data Analytics because they want to make more informed business decisions.
- A data warehouse (DW) is a collection of corporate information and data derived from operational systems and external data sources. A data warehouse is designed to support business decisions by allowing data consolidation, analysis and reporting at different aggregate levels. Data is populated into the Data warehouse through the processes of extraction, transformation and loading (ETL tools). Data analysis tools, such as business intelligence software, access the data within the warehouse.

Class Note:

Subject : ....B.D.T...........

Faculty : ...V.Divya.........

Topic: ......Quality, Preprocessing & storing.

Unit No. : 1
Lecture No.: 4
Link to Session
Planner (SP) S.No. ....of SP
Date Conducted:
Page No. 5

Data quality five R's.

- Relevancy
- Recency
- Range
- Robustness
- Reliability.

Data integrity:

- Data integrity refers to the maintenance of consistency and accuracy in data over its usable life.
- Software, which store, process or retrieve the data, should maintain the integrity of data.

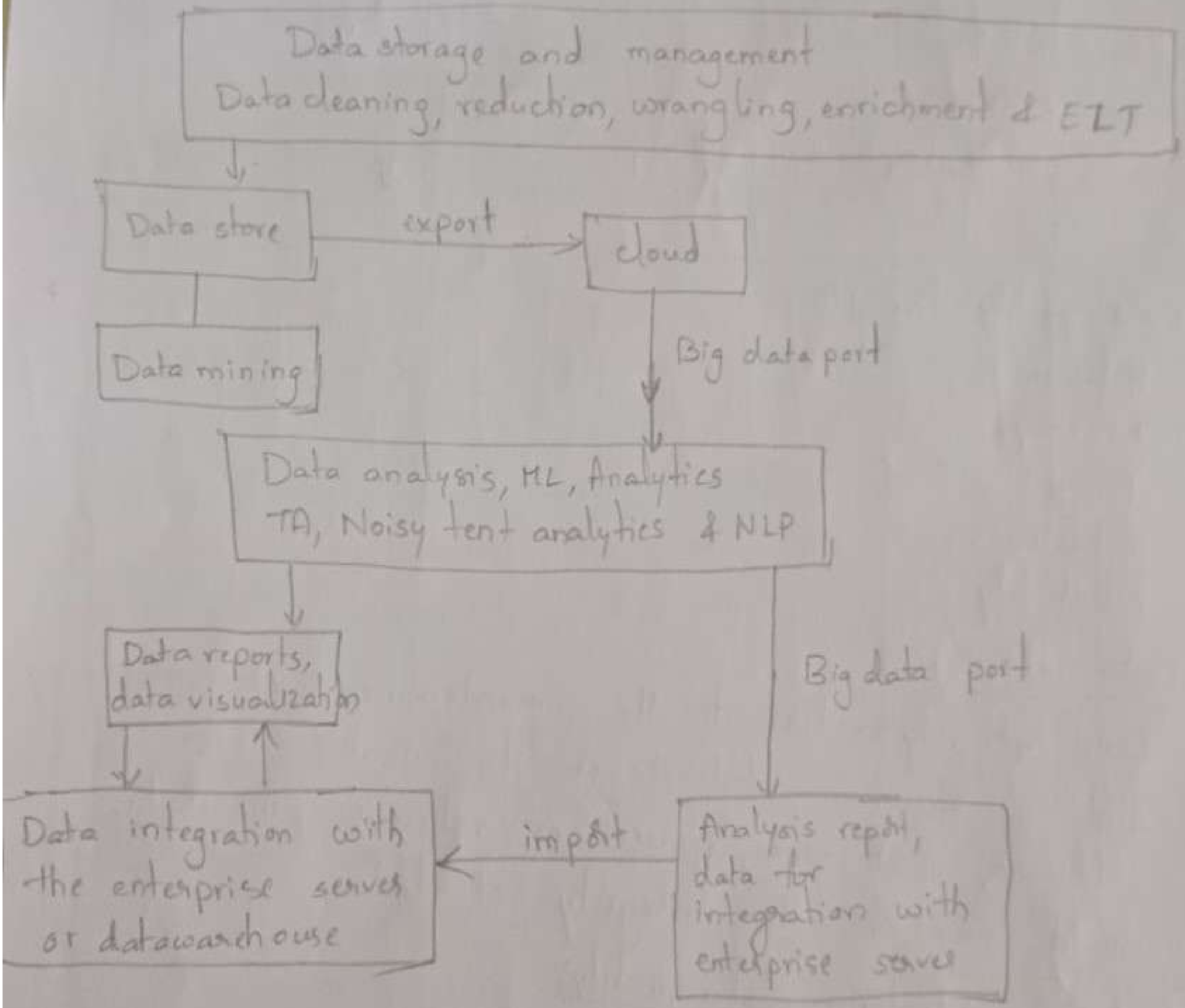
Factors affecting data quality:

- data noise
- Outlier
- Missing value
- Duplicate value

~~nota~~

→ Data cleaning, editing, reduction and wrasing
- Data validation, transformation or transcoding
- ELT processing.

Data export to cloud


Data storage and management
Data cleaning, reduction, wrangling, enrichment & ELT
Data store
export
cloud
Data mining
Big data port
Data analysis, ML, Analytics
TA, Noisy tent analytics & NLP
Data reports,
data visualization
Big data port
Data integration with the enterprise server or datawarehouse
import
Analysis report, data for integration with enterprise server

Class Note:

Subject : BDT

Faculty : V. Divya

Topic: Terminologies used in big data environments.

Unit No. : 1
Lecture No.: 5
Link to Session
Planner (SP) S.No. ....of SP
Date Conducted:
Page No. 6

Terminologies used in big data environments:-

- As-a-service infrastructure:-

Data-as-service, software-as-a-service - all refer to the idea that rather than selling data, licences to use data, or platforms for running Big data technology, it can be provided "as a service", rather than as a product.

This reduces the upfront capital investment necessary for customers to begin putting their data, or platforms to work for them, as the provider bears all of the costs of setting up and hosting the infrastructure. As a customer, as-a-service infrastructure can greatly reduce the initial cost and setup time of getting Big data initiatives up and running.

- Data Science:-

Data science is the professional field that deals with turning data into value such as new insights or predictive models. It brings together expertise from fields including statistics, mathematics, computer science, communication as well as domain expertise such as business knowledge.

- Data mining:-

Data mining is the process of discovering insights from data. In terms of big data, because it is so large, this is generally done by computational methods in an automated way using methods such as decision trees, clustering and analysis and most recently, machine learning. This can be thought of as using the brute mathematical power of computers to spot patterns in data which would not be visible to the human eye due to the complexity of dataset.

- Hadoop:-

- is a framework for Big data computing which has been released into the public domain as open source software and so can freely be used by anyone.
- It consists of a number of modules all tailored for a different vital step of Big data process - from file storage to database to carrying out data operations.
- It has become so popular due to its power and flexibility that it has developed its own industry of retailers, support service providers and consultants.

- Predictive modelling:-

- As its simplest, this is predicting what will happen next based on data about what has happened previously.
- In Big data age, because there is more data around than ever before, predictions are becoming more and more accurate.
- Predictive modeling is a core component of most Big data initiatives, which are formulated to help us choose the course

of action which will lead to the most desirable outcome. The speed of modern computers and the volume of data available means that predictions can be made based on a huge number of variables, allowing an ever increasing number of variables to be assessed for the probability that it will lead to success.

- MapReduce :

- is a computing procedure for working with large datasets, which ~~to~~ was devised due to difficulty of reading and analysing really Big data using conventional computing methodologies. As its name suggest, it consists of 2 procedures - mapping (sorting information into the format needed for analysis - i.e., sorting a list of people according to their age) and reducing (performing an operation, such checking the age of everyone in the dataset to see who is over 21.

- NoSQL :

NoSQL refers to a database format designed to hold more than data which is simply arranged into tables, rows and columns as is the case in a conventional relational database. This database format has proven very popular in Big Data applications because Big data is often messy, unstructured and does not easily fit into traditional database frameworks.

- Python:

- Python is a programming language which has become very popular in the Big data space due to its ability to work very well with large, unstructured datasets.
- It is considered to be easier to learn for a data science beginner than other languages such as R. and more flexible.

- R Programming:

- R is another programming language commonly used in Big data, and can be thought of as more specialized than Python, being geared towards statistics. Its strength lies in its powerful handling of structured data.
- Like Python, it has ~~to~~ an active community of users who are constantly expanding and adding to its capabilities by creating new libraries and extensions.

- Spark:

- Spark is another open source framework like Hadoop but more recently developed and more suited to handling cutting edge Big data tasks involving real time analytics and m/c learning. Unlike Hadoop it does not include its own filesystem though it is designed to work with Hadoop's HDFS or a no. of other options.

Class Note:

Subject : BDT

Faculty : V. Divya

Topic: Data storage & analysis, Big data analytics

Unit No. : 1
Lecture No.: 8, 9
Link to Session Planner (SP) S.No. .... of SP
Date Conducted:
Page No. 8

Data storage and management:-

- traditional systems use structured or semistructured data
- RDBMS, MySQL, DB2, Enterprise server & DW

SQL:-

- Schema creation
- Create catalog
- DDL
- DML
- DCL

Distributed Database Management system:-

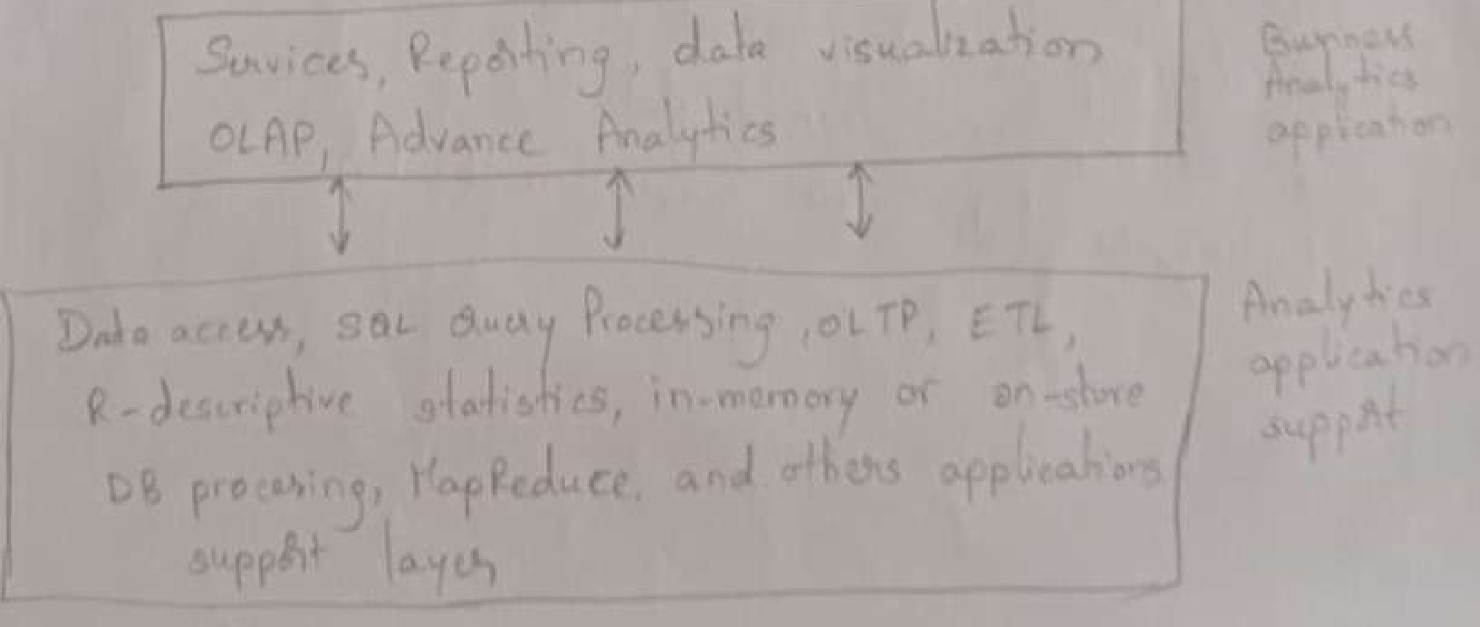
- A distributed DBMS (DDBMS) is a collection of logically interrelated database at multiple system over a computer network.
- in memory column format data
- in memory row format data
- Enterprise data store server and DW

Phases in analytics:-

1. Descriptive analytics enables deriving the additional value from visualization and reports.
2. Predictive analytics is advanced analytics which enables extraction of new facts and knowledge and then predicts/forecast to maximize the profits.
3. Cognitive analytics enables derivation of the additional value and undertake better decision.

Barkley Data Analysis Stack (BDAS)

- Applications, AMP- Genomics and Carat run at the BDAS.
- Data processing s/w component provides in memory processing which processing data effectively across the framework.
- Data processing combines batch, streaming
- Resource management s/w components provide for sharing the infrastructure across various framework.
- Traditional big data analytics reference model



Services, Reporting, data visualization OLAP, Advance Analytics

Business Analytics application

Data access, SQL Query Processing, OLTP, ETL, R-descriptive statistics, in-memory or on-store DB processing, MapReduce, and others applications support layer

Analytics application support

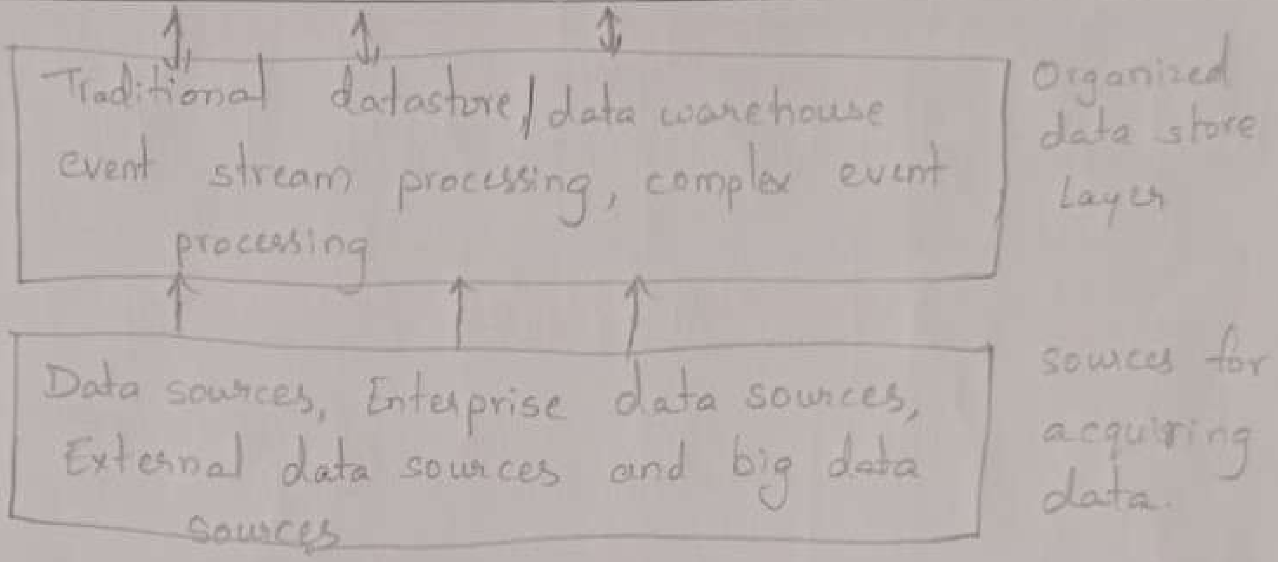
Class Note:

Subject : BDT

Faculty : V.Divya

Topic: Big data Analytics application & case studies.

Unit No. : 1
Lecture No.: 10, 11, 12
Link to Session
Planner (SP) S.No. ....of SP
Date Conducted:
Page No. 9



Traditional datastore / data warehouse event stream processing, complex event processing

Organized data store Layer

Data sources, Enterprise data sources, External data sources and big data sources

sources for acquiring data.

- Big data in marketing & sales
- Big data analytics in detection of marketing frauds
- Big data risks
- Big data in credit card risk management
- Big data in Healthcare
- Big data in medicine
- Big data in Advertising.