UNIT-1

Introduction to Big Data

→ Irrespective of the size of the enterprise (big or small) data continues to be a precious and irreplaceable asset.

→ Data is present internal to the enterprise and also exists outside the four walls and firewalls of the enterprise.

→ Data is present in "homogeneous source" as well as in "heterogeneous sources."

→ The need of the hour is to understand, Manage, process and take the data for analysis to draw valuable insights.

Data → Information

Information → Insights.

Classification of digital data:-

→ Digital data can be broadly classified into three categories

1) Structured data
2) Semi - Structured data
3) Unstructured data

Un-Structured Data:-

→ The table which does not conform to a data model or is not in a form which can be used easily by a computer program.

→ 80% - 90% data of an organization is in this format; for example; memos, powerpoint presentation, images, videos, letters, researches, white papers etc,

Twitter Message	Feeling miffed ☹ victim of twishing
Facebook Post	LoL. C ya. BFN
Log Files	127.0.0.1 - Frank [10/oct/2000:13:55:36-0700]....
E Mail	Hey Joan, Possible to send across the first cut on the Hadoop chapter by Friday EOD.

Table: Few examples of disparate unstructured data.

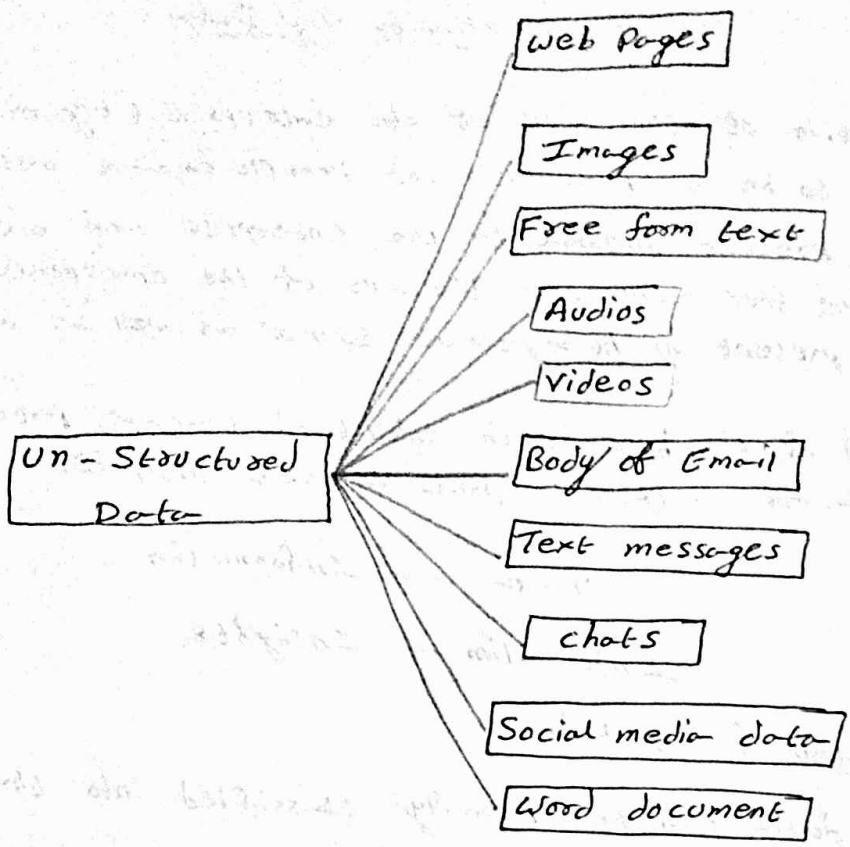


Fig:- Source of unStructured data

Issues:-

→ unstructured data is known not to conform to a process Pre-defined data model or be organized in a pre-defined manner, there are incidents where in the structure of the data can still be implied. (placed in the unstructured category)

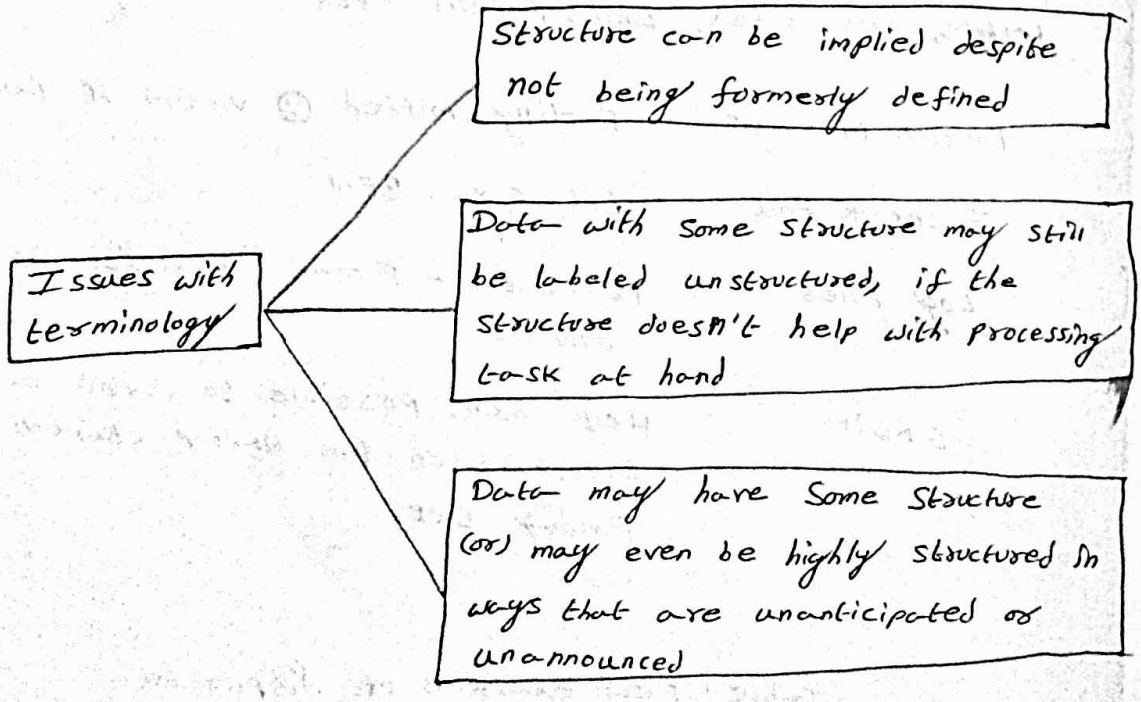


Fig:- Issues with terminology of unstructured data

→ There are situations where people argue that a text file should be in the category of semi-structured data and not unstructured data.

→ Let us look at where they are coming from well the text files does have a name, one can easily look at the properties to get information such as the owner of the file, the date on which the file was created, the size of the file etc.,

→ okay, we do not have little meta-data, but when it comes to analysis, we are more concerned with the content of the text file rather than the name or any of the other properties.

How to Deal with Un-Structured data?

→ unstructured data constitutes approximately 80% of the data that is being generated in any enterprise.

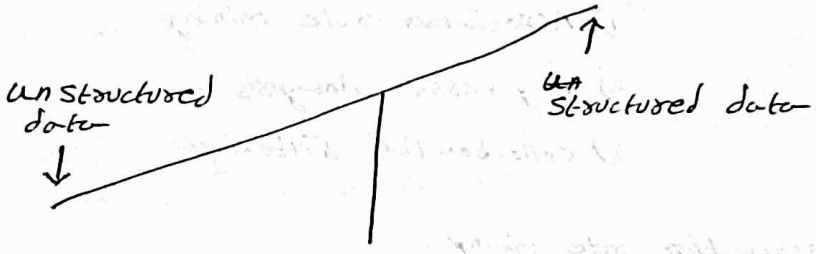


Fig:- unstructured data clearly constitutes a major percentage of enterprise data

→ The balance is clearly shifting in favor of unstructured data, such a big percentage cannot be ignored.

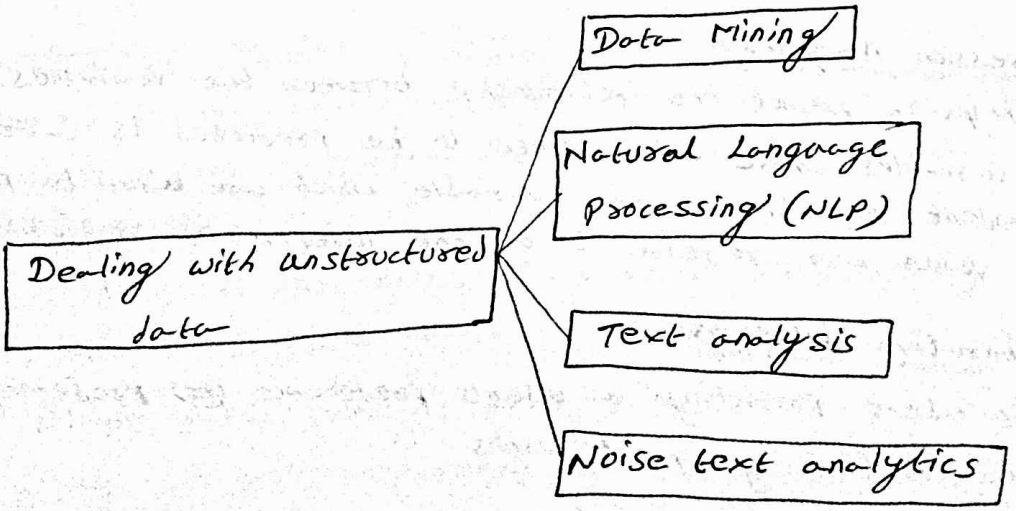


Fig:- Dealing with unstructured data

→ Following techniques are used to find patterns in unstructured data.

1) Data Mining
2) Natural Language processing
3) Text Analytics
4) Noise text analytics

Data Mining:-

→ First, we deal with large data sets
→ Second, we use Methods at the intersection of Artificial Intelligence, Machine Learning, Statistics, & data Systems to unearth consistent patterns in large data sets and/or Systemic relationships b/w variables.
→ It is the analysis step of the "Knowledge discovery in data-bases" process.
→ Few popular data Mining algorithms are as follows:

1) Association rule Mining
2) Regression Analysis
3) Collaborative filtering

a) Association rule mining:-

→ It is also called "Market Basket Analysis".
→ It is used to determine "what goes with what?"
→ It is about when you buy a product, what is other product that you are likely to purchase with it. for example, If you pickup a bread from the grocery, are you likely to pick eggs (or) cheese (or) JAM to go with it.

b) Regression Analysis:-

→ It helps to predict the relationship between two variables.
→ The variables whose value needs to be predicted is called the dependent variable, & the variable which are used to predict the value are referred to as the independent variables.

c) Collaborative filtering:-

→ It is about predicting a user's preference (or) preferences based on the group of users.

Users	Learning using Audio	Learning using videos	Textual Learners
1	Yes	Yes	No
2	Yes	Yes	Yes
3	Yes	Yes	No
4	Yes	?	?

Table :- Sample records depicting learner's for modes of Learning.

→ we are looking at predicting whether user 4 will prefer to learn using videos or it is a textual learner depending on one or a couple of his or her known preferences.

→ we analyze the preferences of similar user profile and on the basis of it, predict that user 4 will also like to learn using videos and is not a textual learner.

Natural Language Processing (NLP) :-

→ It is ~~realted~~ related to the area of human computer interaction.

→ It about enabling computers to understand human or natural language input.

Text analytics (or) Text Mining :-

→ Compared to the structured data stored in relational databases, text is largely unstructured and difficult to deal with algorithmically.

→ Text mining is the process of gleaning high quality & meaningful information from text.

→ It includes tasks such as text categorization, text clustering, Sentiment analysis, concept/entity extraction etc;

Noisy text analytics:-

→ It is the process of extracting structured or semi-structured information from noisy unstructured data such as chats, blogs, wiki's, emails, message-boards, text-messages, etc;

→ The noisy unstructured data usually comprises one or more of the following: spelling mistakes, abbreviations, acronyms, non-standard words, missing punctuation, missing letter case, filler words, such as "Uh", "Um", etc;

Semi-Structured Data:-

→ Semi structured data is also referred to as self-describing structure.

→ It has the following features:-

1) It does not conform to the data models that one typically associates with relational databases or any other forms of data tables.
2) It uses tags to segregate semantic elements.
3) Tags are also used to enforce hierarchies of records and fields within data.
4) There is no separation b/w the data and the schema. The amount of structure used is dictated by the purpose at hand.
5) In semi-structured data, entities belonging to the same class and also grouped together need not necessarily have the same set of attributes. And if at all, they have the same of attributes, the order of attributes may not be similar and for all partical purposes it is not important as well.

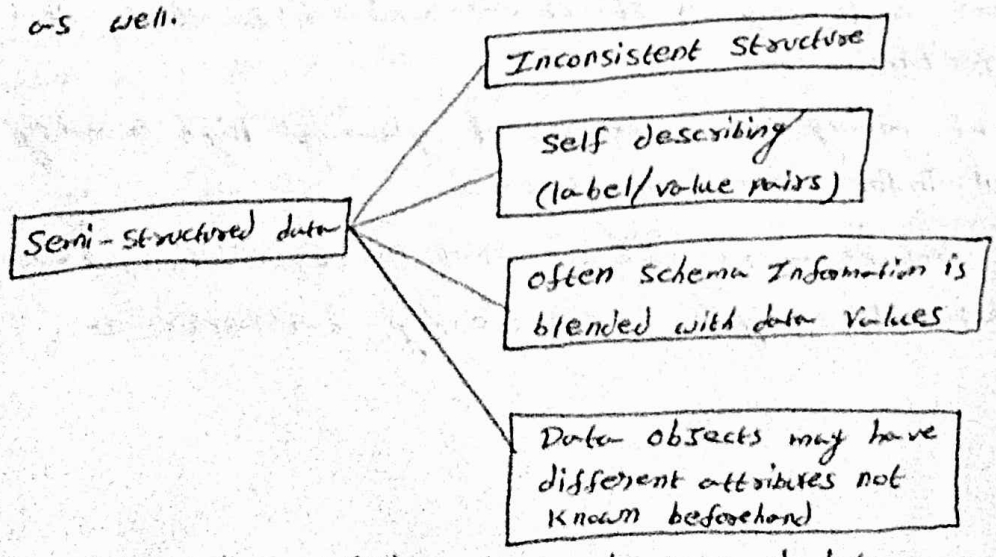


Fig: characteristics of semi-structured data

Sources of Semi-Structured Data:-

→ The Sources for Semi-Structured data, the front runners are "XML" and "JSON".

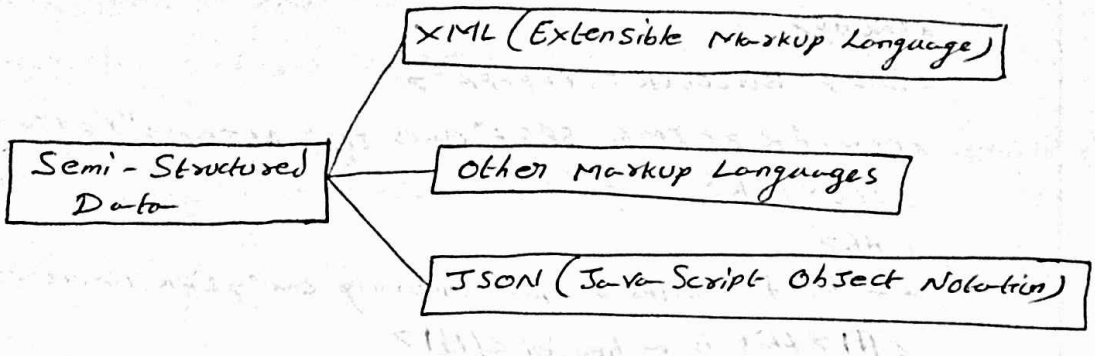


Fig:- Sources of Semi-Structured Data

1. XML:- extensible Markup Language (XML) is hugely popularized by web services developed utilizing the Simple Object Access Protocol (SOAP) principles.

2. JSON:- JavaScript is object Notation (JSON) is used to transmit data between a server and web application. JSON is popularized by web services developed utilizing the representational state Transfer (REST) - an architectural style for creating scalable web services. MongoDB (open-source, distributed, NoSQL, documented-oriented database) & couchbase (originally known as Membase, open-source, distributed, NOSQL, document-oriented database) store data natively in JSON format.

Example:-

* Sample JSON document

```
{
  _id: 9,
  BookTitle: "Fundamentals of Business Analytics",
  AuthorName: "Seema Acharya",
  Publisher: "wiley India",
  YearofPublication: "2011".
}
```

* An example of HTML is as follows:

```
<HTML>
<HEAD>
<TITLE> Place your title here </TITLE>
</HEAD>
<BODY BGCOLOR = "FFFFFF">
<CENTER><IMG SRC = "clouds.jpg" ALIGN = "BOTTOM">
</CENTER>
<HR>
<a href = "https://bigdata-university.com"> LINK Name </a>
<H1> this is a header </H1>
<P> a new Paragraph </P>
</BODY>
</HTML>
```

Structured Data:-

→ When do we say that the data is structured? The data conforms to a pre-defined Schema/Structure we say it is structured data.

→ Think structured data, and data model - A model of the types of business data that we intend to "store, process and access."

→ Most of the structured data is held in RDBMS. An RDBMS conforms to the relational data model wherein the data is stored in rows/columns.

	column 1	column 2	column 3	column 4
ROW 1				

Tabel :- A relation/table with rows & columns

→ The number of rows/records/tuples in a relation is called the cardinality of a relation and the number of columns is referred to as the degree of a relation.

1) The design of a relation/table, the fields/columns to store the data, the type of data that will be stored [numeric(integer or real), alphabets, etc.,]

2) The constraints that we would like our data to conforms to (constraints such as UNIQUE, NOT NULL or PRIMARY).

Example:-

→ A good structured table with absolute adherence to relational data Model.

Column Name	Data Type	Constraints
EMPNO	Varchar(10)	PRIMARY KEY
EMPNAME	Varchar(50)	
Designation	Varchar(25)	NOT NULL
DepNO	Varchar(5)	
ContactNo	Varchar(10)	NOT NULL

Table:- Schema of an "Employee" table in a RDBMS such as oracle.

→ That each record in the table will have exactly the same structure.

Example:-

EMPNO	EMPName	Designation	DeptNO	ContactNO
E101	Allen	Software Engineer	D1	9999-XXX
E102	Simon	Consultant	D1	989XXXXX

Table:- Sample records in the "Employee" table.

→ The above "Employee" table is related to the "Department" table on the basis of the common column "DeptNO".

→ It is not mandatory for the two tables that are related to have exactly the same name for the common column.

→ on the contrary, the two tables are related on the basis of values held within the column, "DeptNO".

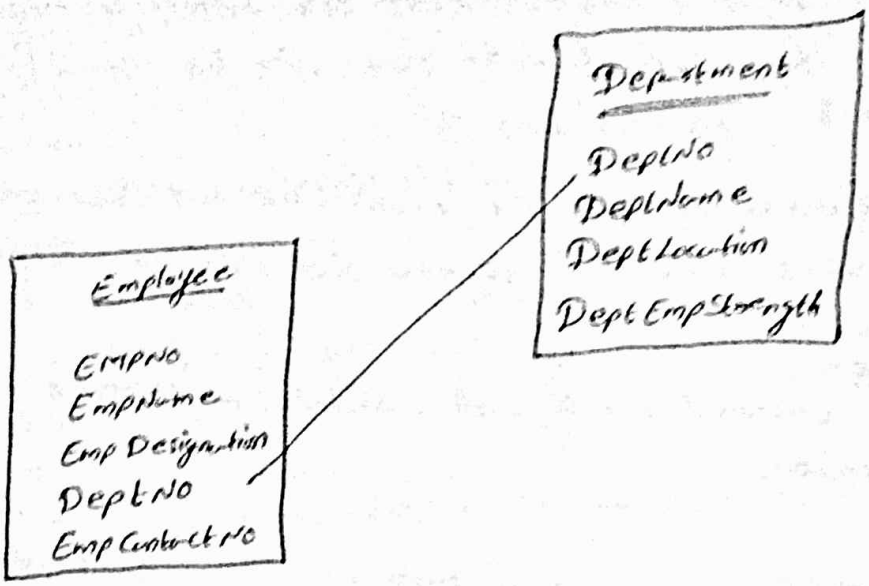


Fig:- Relationship b/w "Employee" and "Department" tables.

→ referential integrity constraints (primary - foreign key) wit the "Department" table and "Employee" table being the referenced table.

Sources of Structured Data:-

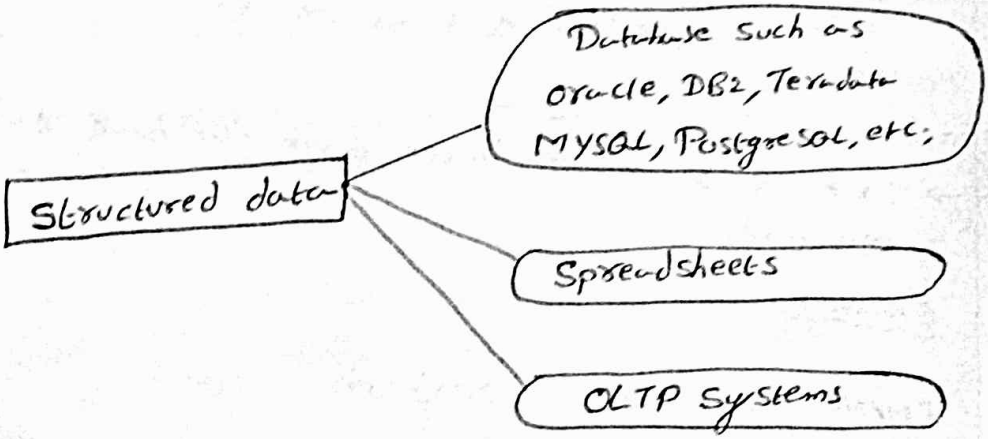


Fig:- Sources of Structured data

→ These databases are typically used to hold transaction/operational data generated and collected by day-to-day business activities.

→ In other words, the data of the OnLine Transaction Processing (OLTP) systems are generally quite structured.

Ease of working with Structured Data :-

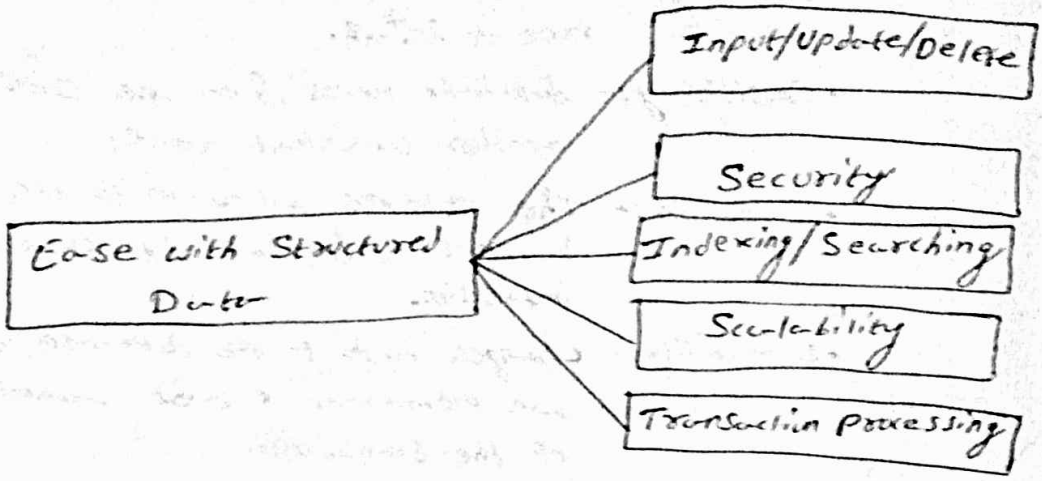


Fig :- Ease of working with structured data.

→ Structured data provides the ease of working with respect to the following :-

1) Insert/Update/delete :- The Data Manipulation Language (DML) operations provide the required ease with data input, storage, access, process, analysis etc;

2) Security :- These are available staunch encryption and tokenization solutions to warrant the security of information throughout its lifecycle. Organizations are able to retain control and maintain compliance adherence by ensuring that only authorized individuals are able to decrypt and view sensitive information.

3) Indexing :- It is a data structure that speed ups the data retrieval operations (Select statement) at the cost of additional writes and storage space, but the benefits that ensue in search operation.

4) Scalability :- The storage and processing capabilities of traditional RDBMS can be easily scaled up by increasing the horsepower of the database server.

5) Transaction Processing :- RDBMS has support for ACID properties of transaction.

- Atomicity :- means that either it happens in its entirety or none of it at all.
- Consistency :- database moves from one consistent state to another consistent state.
- Isolation :- The resource allocation to the transaction happens such that the transaction gets the impression.
- Durability :- changes made to the database during a transaction are permanent & that accounts for the durability of the transaction.

Characteristics of Data :-

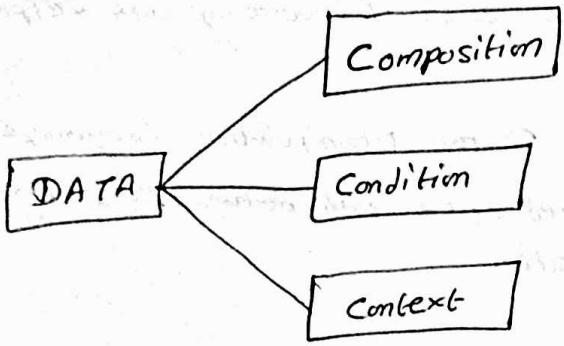


Fig :- Characteristics of data

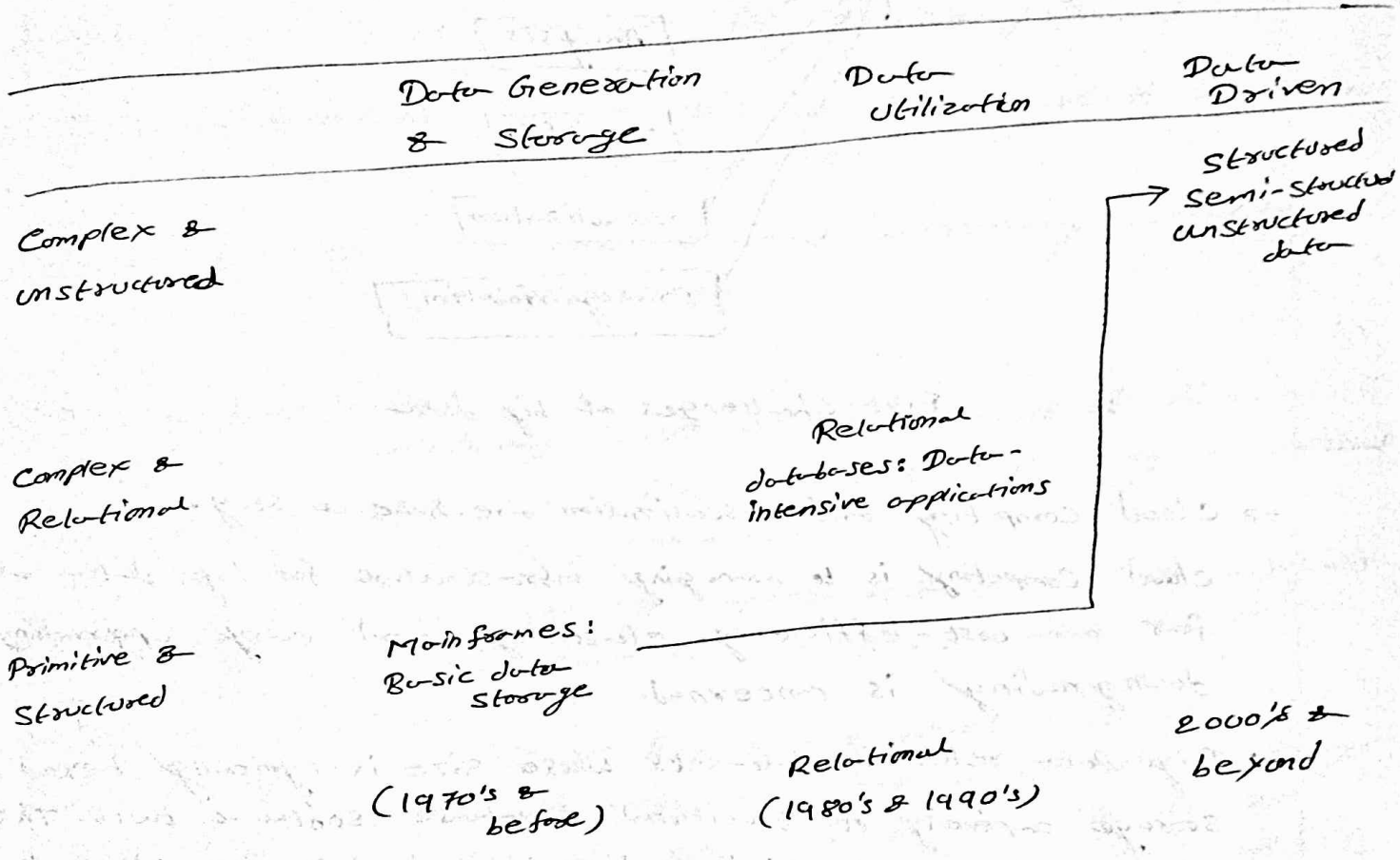
1) Composition :- It deals with structure of data, that is, the source of data, the granularity, the types, & the (quality) nature of data, whether it is static or real-time streaming.

2) Condition :- It deals with the state of data, that is, "can one use this data as is for analysis?" "Does it require cleansing for further enhancement and enrichment?"

3) Context :- It deals with, "Where has this data been generated?, "why was this data generated?", "How sensitive is this data?", "What are the events associated with this data?" & so on...

→ Big data is about complexity, in terms of multiple & unknown data-sets, in terms of exploding volume, in terms of the speed at which the speed data is being generated, & the speed at which it needs to be processed, & in terms of the variety of data.

Evolution of Big Data:-



Definition of Big Data:-

→ Big Data refers to high-volume, high-velocity, high-variety information assets that demand cost effective, ~~informa~~ innovative forms of information processing that enable enhanced insight and decision-making.

Challenges with Big Data:-

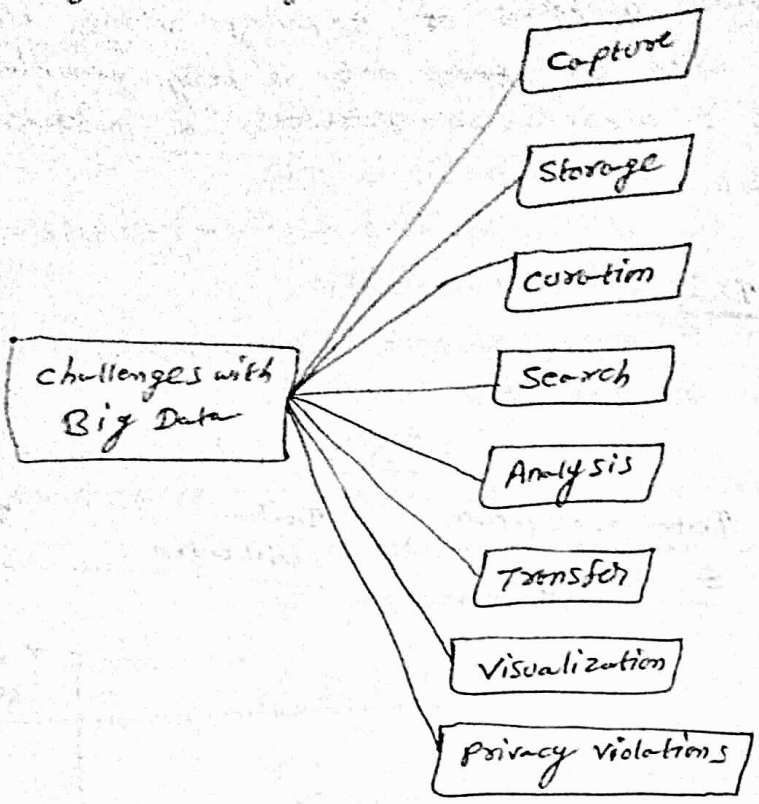


Fig:- Challenges of big data

→ Cloud computing and visualization are here to stay. Cloud computing is to managing infrastructure for big data as far as cost-efficiency, elasticity, and easy upgrading/downgrading is concerned.

→ Big data refers to data-sets whose size is typically beyond the storage capacity of traditional database software tools. The data changes are highly dynamic and therefore there is a need to ingest this as quickly as possible.

→ Data visualization is becoming popular as a separate discipline.

* Top challenges facing Big Data:-

1) Storage:- RDBMS (or NOSQL is one major concern that needs to be addressed to handle the need for scaling rapidly and elastically.

2) Security:- Most of the NOSQL big data platform have poor security mechanisms, when it comes to safeguarding big data. A sensitive information cannot be ignored by big data (credit card information, or personal information)

3) Schema :- We want the technology to be able to fit our big data and not the other way around. Here schema have no place. Here, the need of the hour is dynamic schema, pre-defined schema is passes.

4) Continuous availability :- 24/7 support is required, almost all RDBMS (or) NOSQL big data platforms have a certain amount of downtime built in.

5) Consistency :- Should one opt for consistency?

6) Partition tolerant : Systems that can take care of both h/w & s/w failures?

7) Data quality :- It require - data accuracy, completness, timeliness etc.

What is Big Data :-

→ Big Data is extremely large (or) complex set of data, and its so large that it's difficult to process it using traditional database & software techniques.

→ Every day we are creating approximately 2.5 quintillion bytes of data. So, where is this huge amount of data getting generated.

→ For example :- Earlier we had mobile phones with the functionality of calling and text messages (or) clicking some pictures. But with new technology like smart phones we have a lot of applications for music, sports, Social media like (facebook, twitter, LinkedIn) and many many more. Data is getting generated when we shop online also.

Why Big Data ? :-

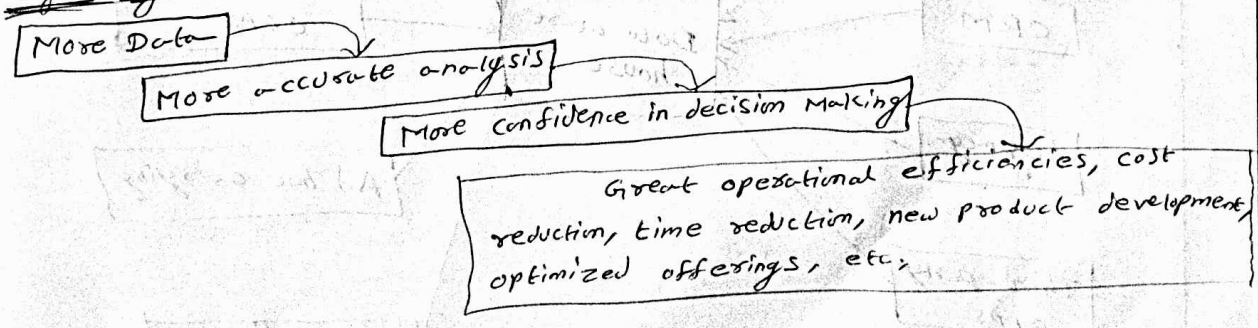


Fig :- Why Big Data

→ We have more data for analysis, It will be the analytical accuracy and also the greater would be the confidence in our decisions based on these analytical findings.

→ A great positive impact in terms of enhancing operational efficiencies, reducing cost and time, and innovating on new products, new services and optimizing existing services.

Why is Big Data Analytics Important?

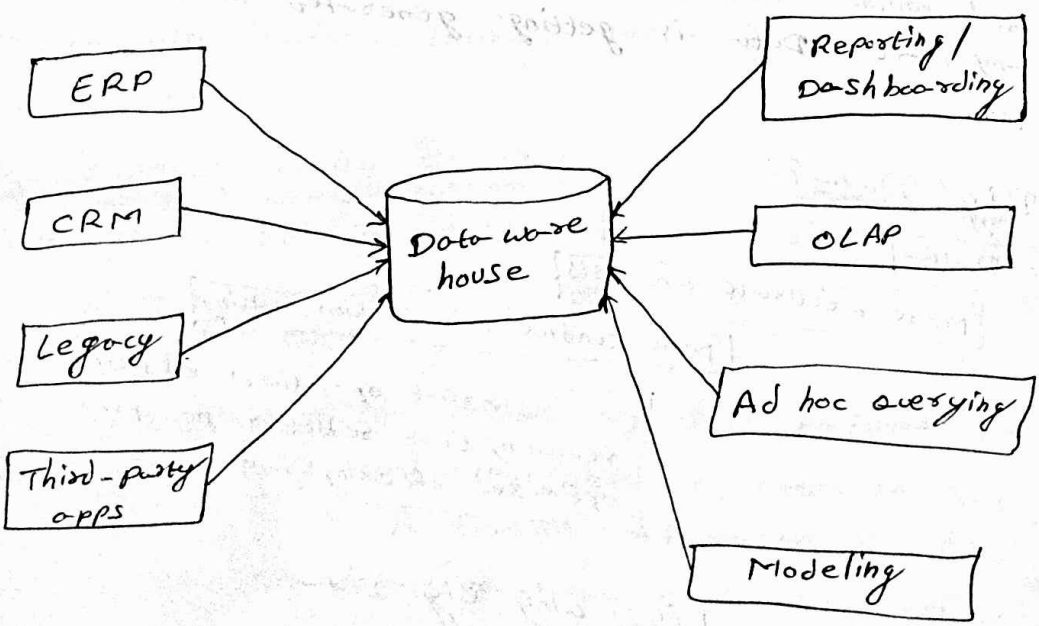
1) Reactive - Business Intelligence (BI):- It allows the business to make faster and better decision by providing the right information to the right person at the right time in the right format. It is about analysis of the past data and then displaying the finding of the analysis or reports in the form of enterprise dashboards, alerts, notifications etc;

2) Reactive - ~~Business~~ Big data Analytics:- Analysis is done on huge data-sets but the approach is still based on static data.

3) Proactive - Analytics:- This is to support futuristic decision making by the use of data mining, text mining and statistical analysis. This analysis is not on big data and still uses the traditional database management practices on big data.

4) Proactive - Big Data Analytics:- tera-bytes, peta-bytes, exa-bytes of information to filter out the relevant data to analyze. The ability to solve complex problems using more data.

A typical data warehouse Environment:-



→ Transactional (or) day-to-day business data is gathered from Enterprise Resource planning (ERP) systems, CRM, legacy systems, and several - Third party applications.

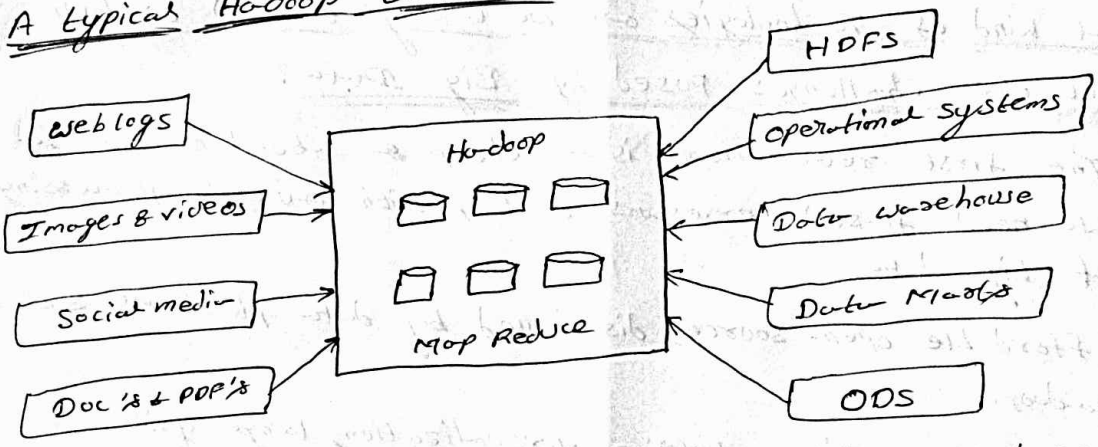
→ Data from these sources may differ in format [SQL, (or) in spreadsheet (or) in .csv, in .txt]. Source of the Data may be in same location (or) different location.

→ Data is then integrated, cleaned up, transformed and standardized through the process of ETL.

→ The transformed data is then loaded into the enterprise data warehouse (or) data marts.

→ Market leading business intelligence and analytics tools are then used to enable decision making.

A typical Hadoop Environment:-



→ Hadoop environment different from the data warehouse environment.

→ Data sources are quite different from web logs, to images, & videos to social media to the various docs, pdf's etc.,

→ Data is focused within the company's firewall but also data residing outside the company's firewall.

→ This data is placed with in Hadoop Distributed file system (HDFS). If needed, this can be repopulated back to operational systems (or) to the enterprise data warehouse (or) data marts (or) operational data source store to be picked for processing and analysis.

Traditional Business Intelligence V/S Big Data :-

Business Intelligence

1) Enterprise data is stored in a central server.

2) Typical database server that scales vertically.

3) Data is generally analyzed in an offline mode.

4) Traditional BI is about structured data & it is here that data is taken to processing functions

Big Data

1) Data resides in a Distributed file system.

2) Distributed file system scales out horizontally.

3) data is analyzed in both real time as well as offline mode.

4) Big data is about 3 V's and the processing functions are taken to the data.

What kind of Technologies are we looking toward to help Meet the challenges posed by Big Data?

1) The first requirement is of cheap & abundant Storage.

2) We need faster processors to help with quicker processing of big data

3) Affordable open-source, distributed big data platforms, such as Hadoop.

4) Parallel processing, clustering, virtualization, large grid environments (To distribute processing to a number of machine), high connectivity, and high throughputs rather than low latency.

5) cloud computing and other flexible resource allocation arrangements.