UNIT-I

Introduction to Big Data

__/__/____

L. A Lalitha

Classification of Digital Data

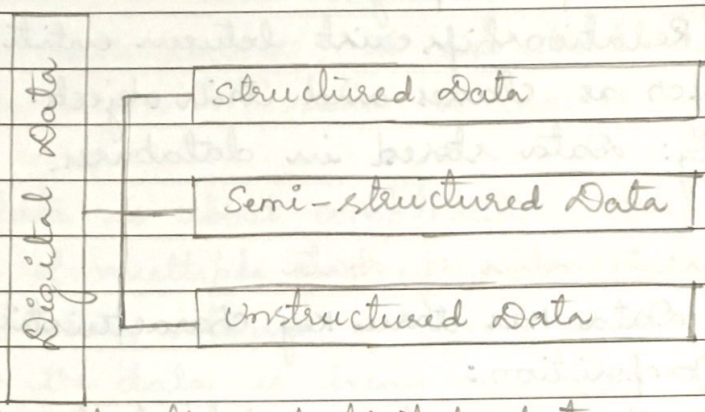


fig: classification of digital data

Digital data can be broadly classified into structured, semi-structured and unstructured data.

1. Unstructured data:

This is the data which doesnot conform to a data model or is not in a form which can be used easily by a computer program. About 80-90% data of an organization is in this format.

Eg: Memos, chat rooms, PowerPoint presentations, images, videos, letters, body of an email, etc.

2. Semi-structured data:

This is the data which doesnot conform to a data model but has some structure. However, it is not in a form which can be easily used by a computer program.

Eg: Emails, XML, markup languages like HTML, etc, Metadata for this data is available but is not sufficient.

3. Structured data:

This is the data which is in an organized form

(in rows and columns) and can be easily used by a computer program.

Relationships exist between entities of data, such as classes and their objects.

Eg: Data stored in databases.

Characteristics of Data

Data has three key characteristics

1. **Composition:**

The composition of data deals with the structure of data, that is, the sources of data, the granularity, the types, and the nature of data as to whether it is static or real-time streaming.

2. **Condition:**

The condition of data deals with the state of data, that is, "Can one use this data as it is for analysis?" or "Does it require cleaning for further enhancement and enrichment?".

3. **Context:**

The context of data 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", and so on...

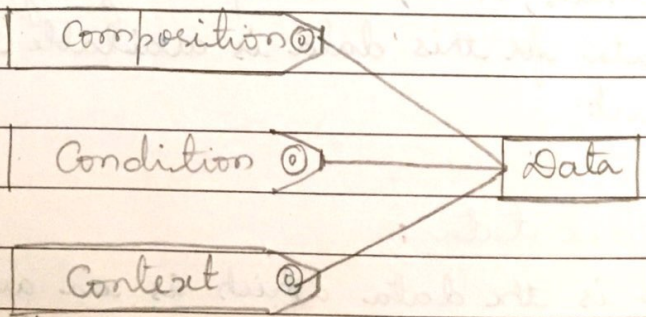


fig:- Characteristics of data

L. A Lalitha

* Small data (data as it existed prior to the big data revolution) is about certainty. It is about fairly known data sources; it is about no major changes to the composition or context of data.

* Big data is about complexity.... Complexity in terms of multiple and unknown datasets, in terms of exploding volume, in terms of the speed at which the data is being generated and the speed at which it needs to be processed, and in terms of the variety of data (internal or external, behavioral or social) that is being generated.

Evolution of Big Data

	Data generation and Storage	Data utilization	Data driven
Complex and Unstructured			Structured data, Unstructured data, Multimedia data
Complex and Relational		Relation databases, Data intensive applications	
Primitive & structured	Mainframes: Basic data storage		
	1970s and before	Relational (1980s and 1990s)	2000s and beyond

The Evolution of big data

1970s and before was the era of mainframes. The data was essentially primitive and structured. Relational databases evolved in 1980s and 1990s. The era was of data intensive applications. The World Wide Web (WWW) and the Internet of Things. (IOT) have led to an onslaught of structured, unstructured, and multi-media data

Definition of Big Data / What is Big Data?

Big data is often characterized by 3Vs
- the extreme Volume of data,
- the wide Variety of data types, and
- the Velocity at which the data must be processed.

Although big data doesn't equate to any specific volume of data, the term is often used to describe terabytes, petabytes, and even exabytes of data captured over time

Breaking down the 3Vs of big data

Volume: Such voluminious data can come from ten thousand different sources, such as business sales records, the collected results of scientific experiments or real time sensors used in the IOT. Data may be raw or preprocessed using separate software tools before analytics are applied.

Variety

Data may also exist in a wide variety of file types,
- including structured data such as SQL database stores
- Unstructured data such as document files, e

- streaming data from sensors.

Further, big data may involve multiple, simultaneous data sources which may not otherwise be integrated.

Eg:- a big data analytics project may attempt to gauge a product's success and future sales by correlating past sales data, return data and online buyer review data for that product.

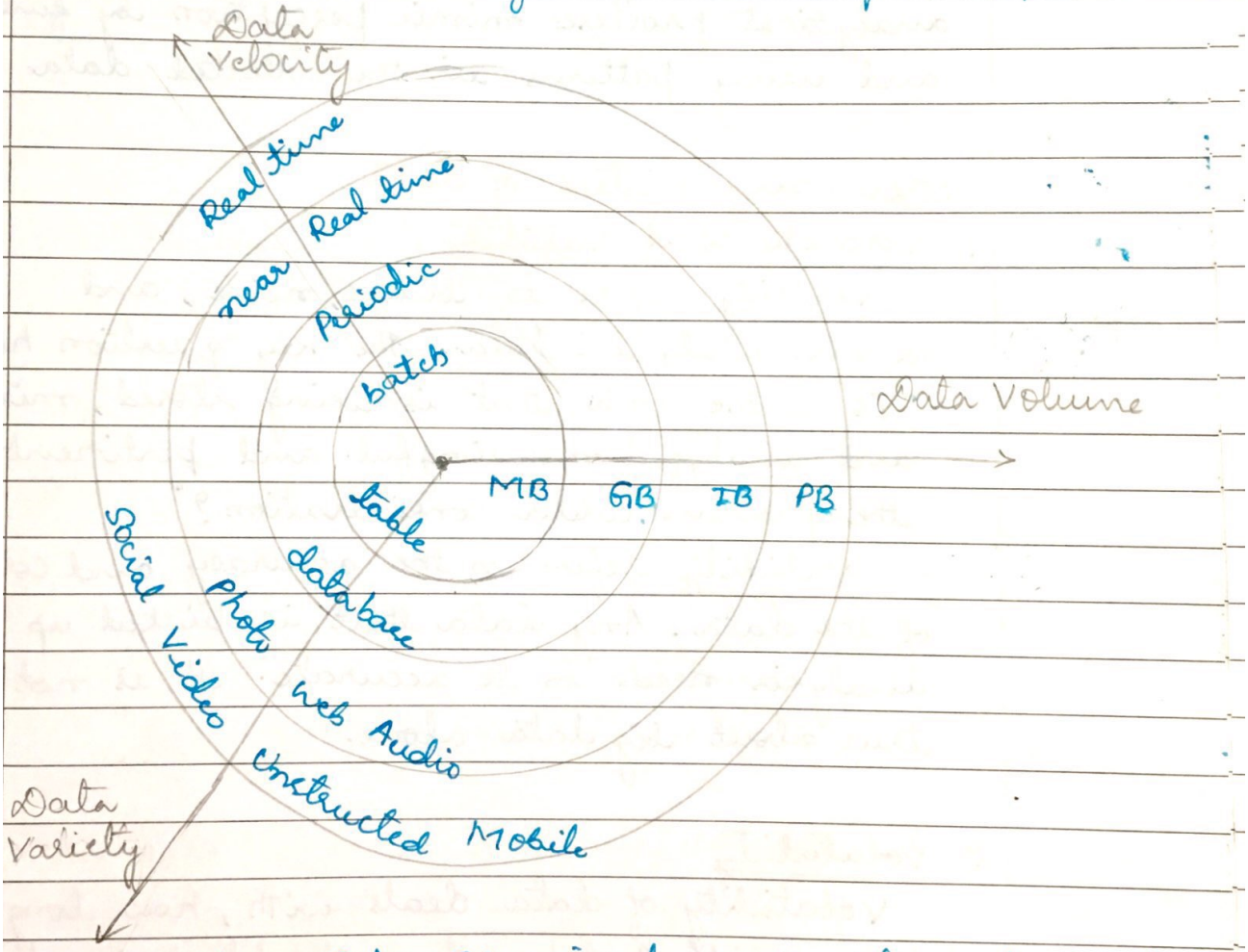


fig: Data: Big in volume, Variety and Velocity

Velocity:

Velocity refers to the speed at which big data must be analyzed. Every big data analytics project will ingest, correlate and analyze the data sources; and then render an answer or result based on an

overarching query. This means human analyst must have a detailed understanding of the available data and possess some sense of what answer they are looking for.

Velocity

Big data analysis expands into fields like Machine learning and Artificial Intelligence, where analytical processes mimic perception by finding and using patterns in the collected data.

Other characteristics of data

1. Veracity and Validity

Veracity refers to biases, noise, and abnormality in data. The key question here is: "Is all the data that is being stored, mined, and analyzed meaningful and pertinent to the problem under consideration?"

Validity refers to the accuracy and correctness of the data. Any data that is picked up for analysis needs to be accurate. It is not just true about big data alone.

2. Volatility

Volatility of data deals with, how long is the data valid? And how long should it be stored? There is some data that is required for long-term decisions and remain valid for longer period of time. However, there are also pieces of data that quickly becomes obsolete minutes after their generation.

3. Variability

Data flows can be highly inconsistent with periodic peaks.

Challenges with Big Data

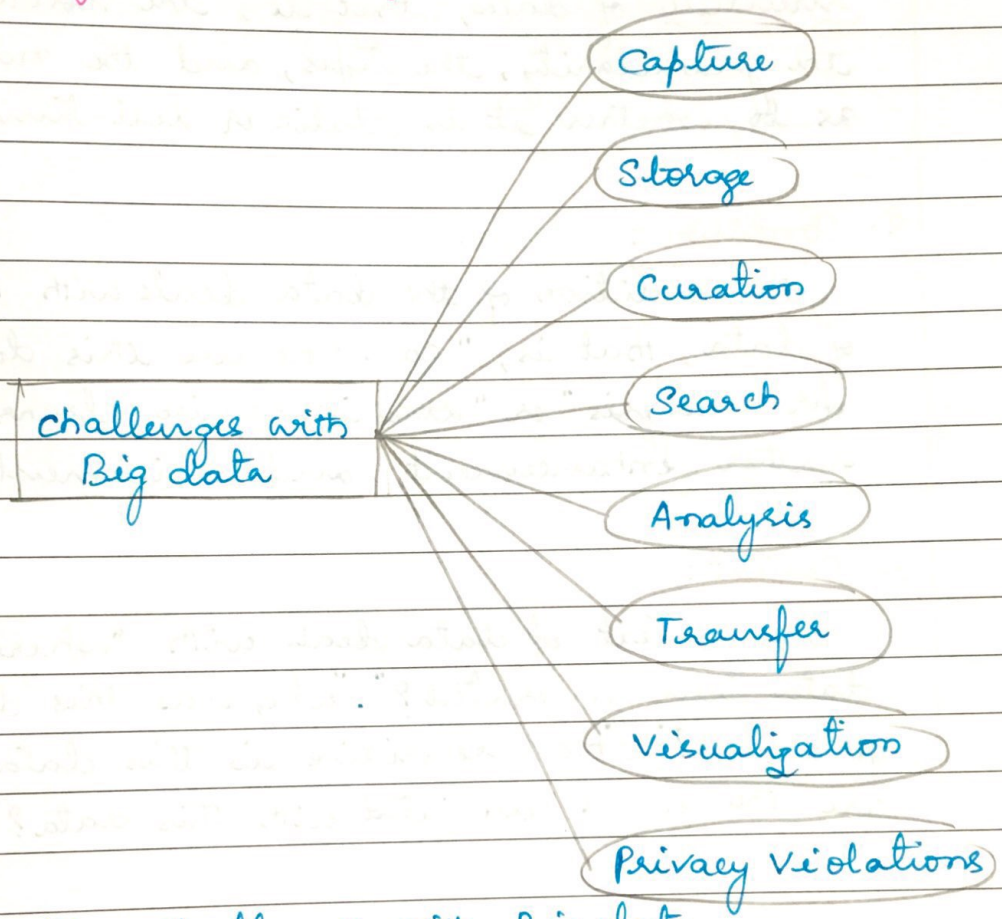


fig: Challenges with Big data.

Big data refers to databases/datasets whose size is typically beyond the storage capacity of traditional databases software tools. There is no explicit definition of how big the dataset should be for it to be considered "big data". Here we are to deal with data that is just too big, moves way to fast, and doesnot fit the structures of typical database systems. The data changes are highly dynamic and therefore there is a need to ingest this as quickly as possible.

Characteristics of Data

Data has three key characteristics:

1. **Composition:**

The composition of the data deals with the structure of data, that is, the sources of data, the granularity, the types, and the nature of data as to whether it is static or real-time streaming.

2. **Condition:**

The condition of the data deals with the state of data, that is, "can one use this data as is for analysis" or "does it require cleansing for further enhancement and enrichment?"

3. **Context:**

The context of data 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?" and so on

Structured Data:

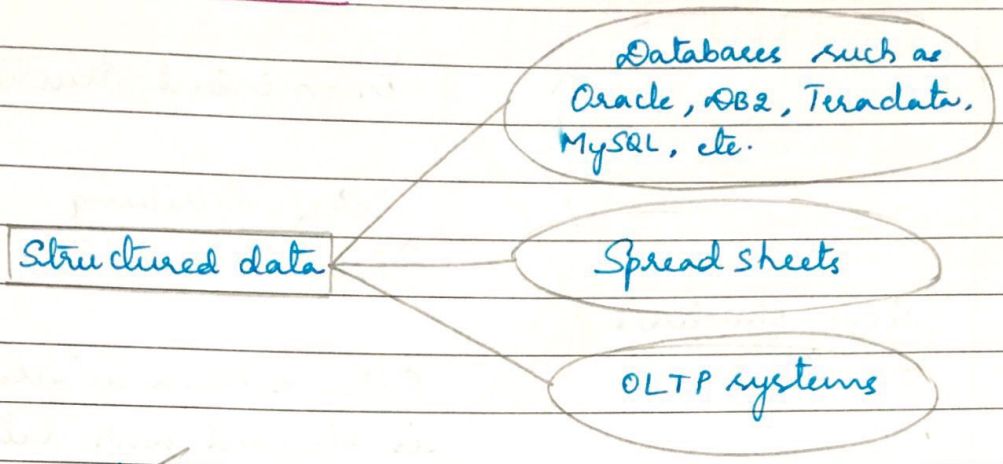


fig: Sources of structured data

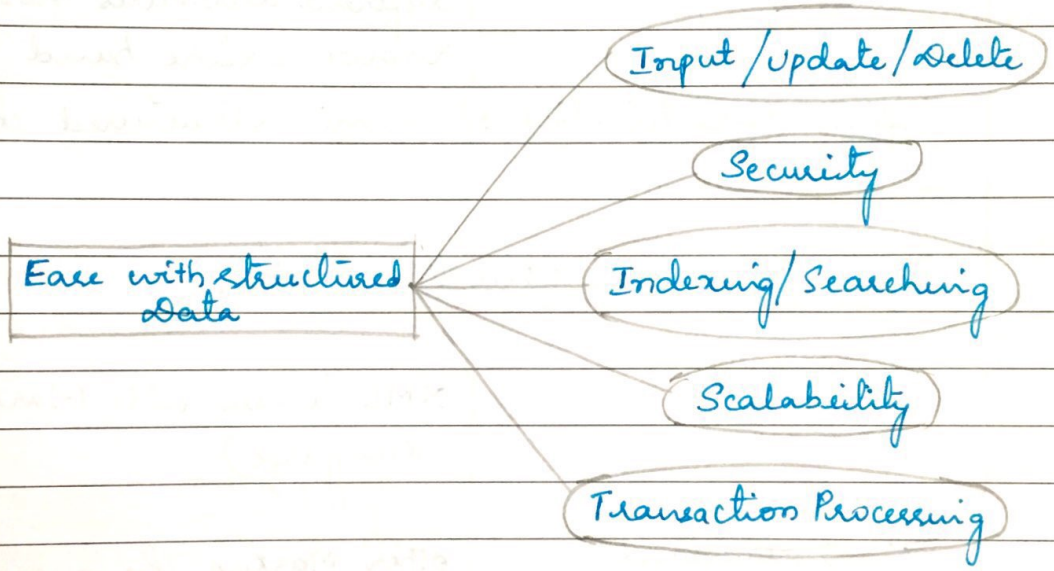


fig: Ease of working with structured data

ACID properties of Transaction

- Atomicity
- Consistency
- Isolation
- Durability

Semi-structured Data

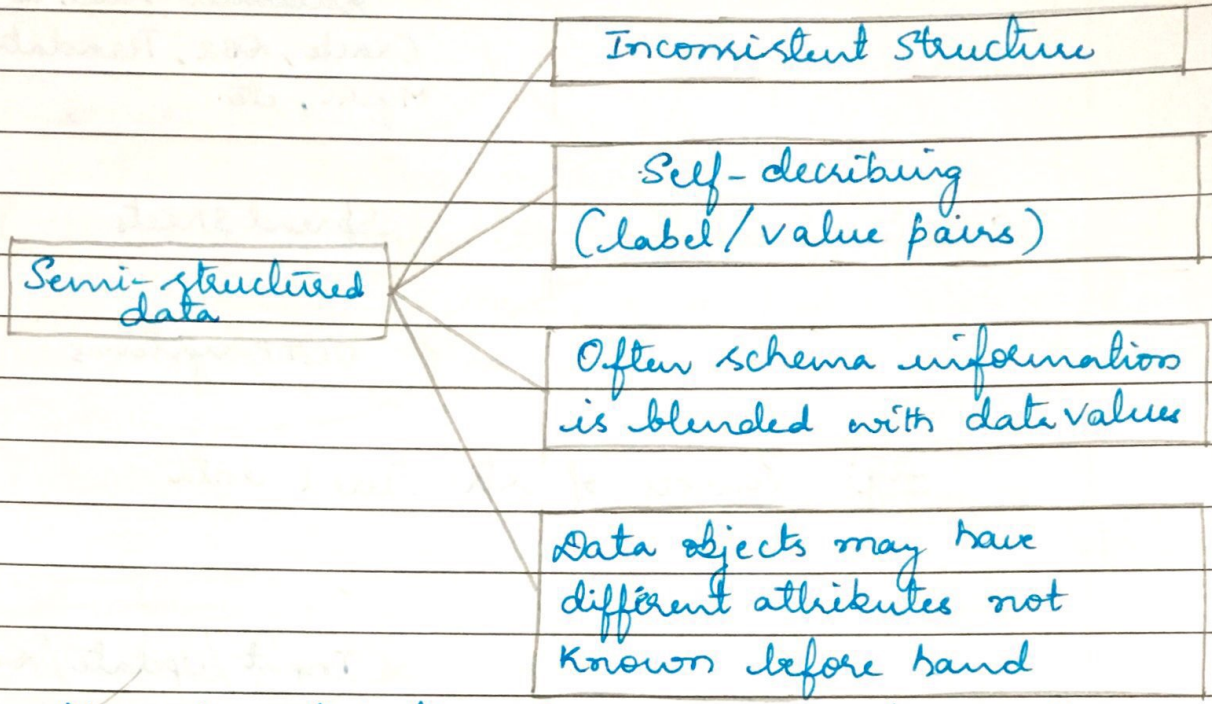


fig: characteristics of semi-structured data

Sources of Semi-Structured Data

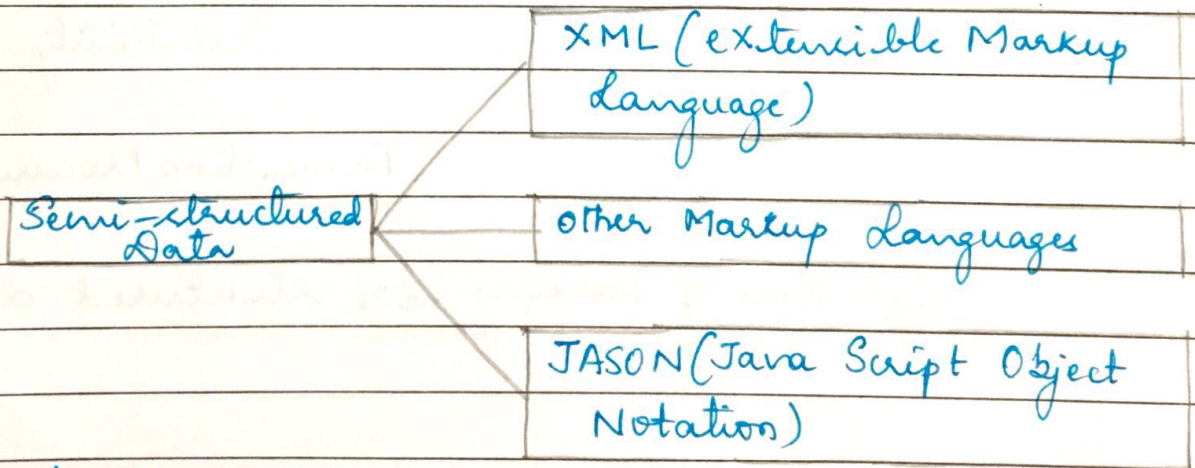


fig: Sources of semi-structured data

Unstructured Data

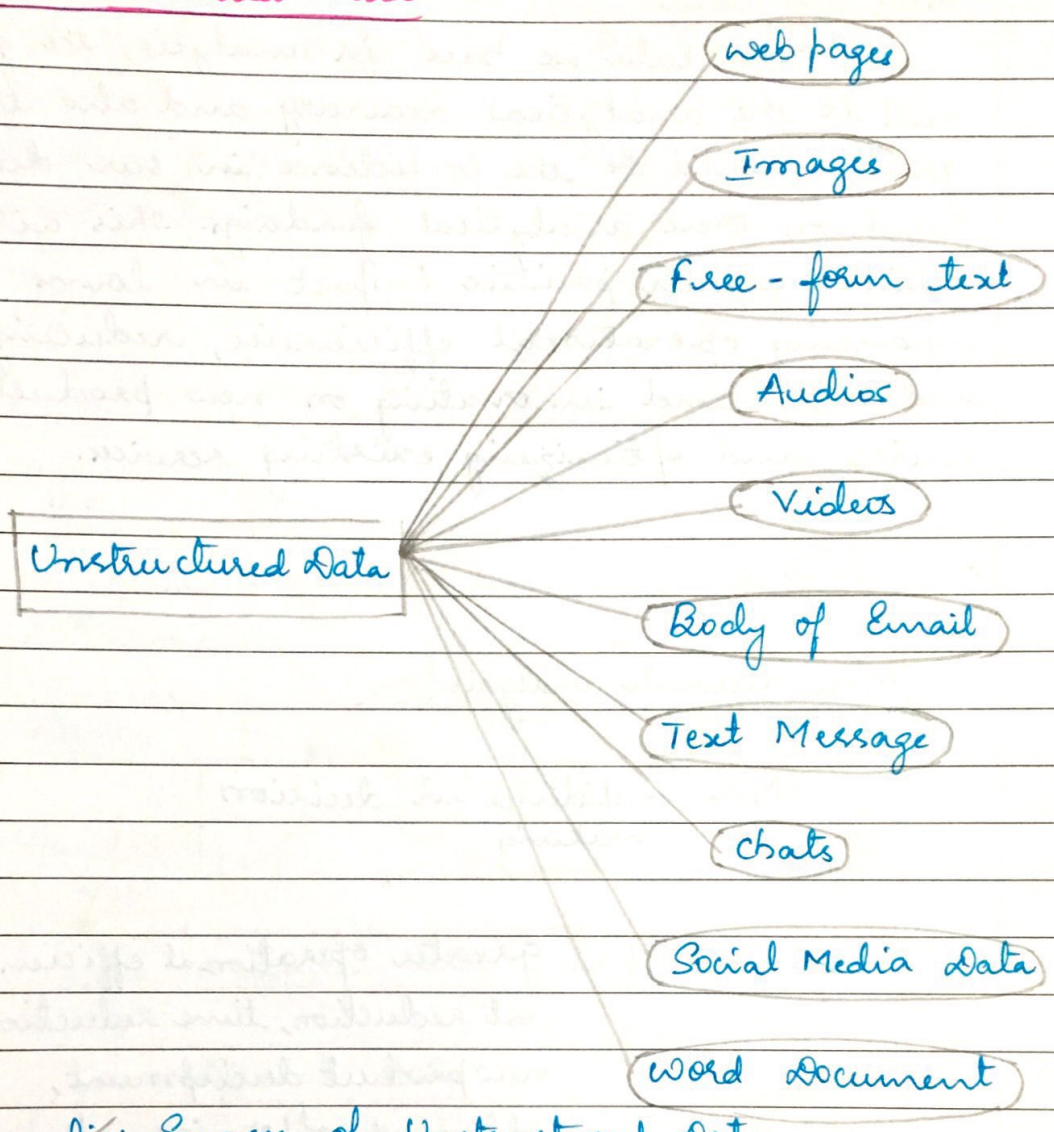


fig: Sources of Unstructured Data

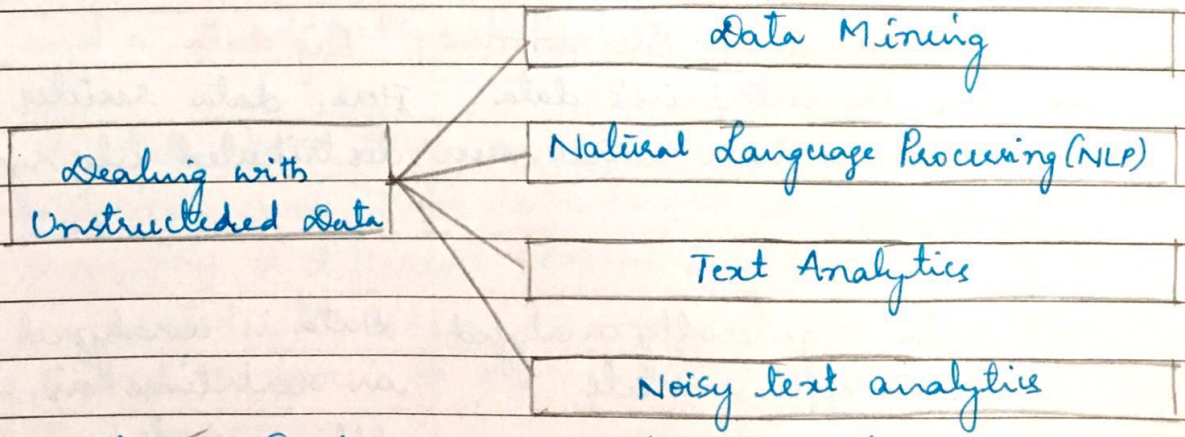
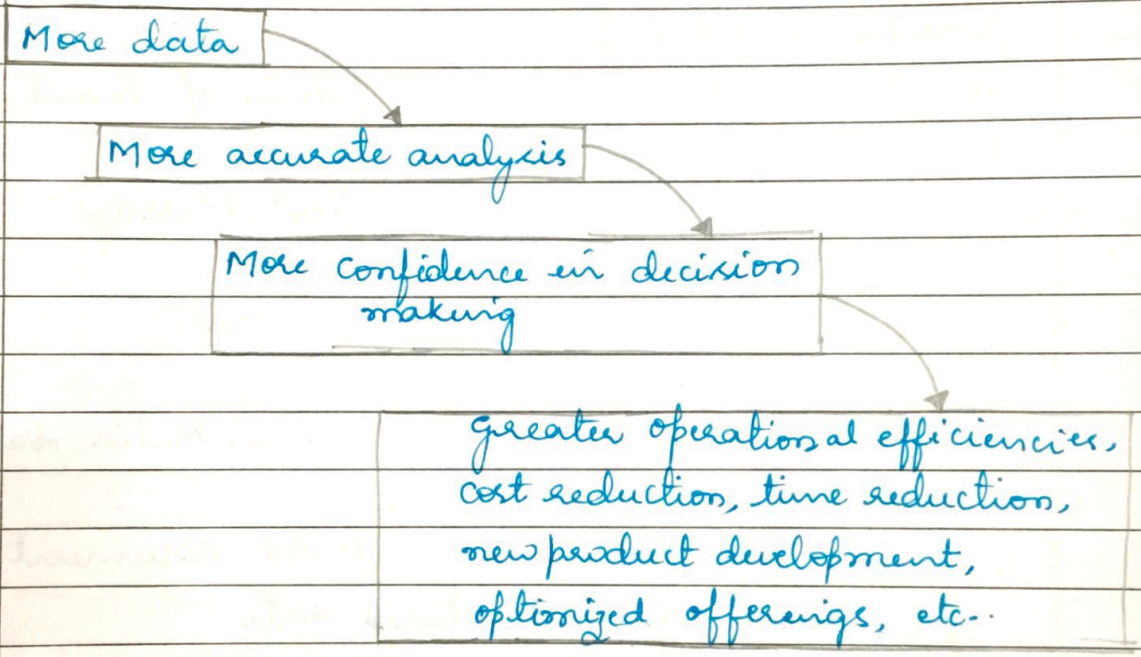


fig: Dealing with Unstructured data

Why Big Data?

The more data we have for analysis, the greater will be the analytical accuracy and also the greater would be the confidence in our decisions based on these analytical findings. This will entail a greater positive impact in terms of enhancing operational efficiencies, reducing cost and time, and innovating on new products, new services, and optimizing existing services.



Traditional Business Intelligence (BI) Versus Big Data

Traditional BI Environment	Big data
→ All the enterprise's data is housed in a central server (whereas)	Here, data resides in a distributed file system
→ Data is generally analyzed in an offline mode.	Data is analyzed both in real time as well as offline mode.

→ It is about structured data and it is here that data is taken to processing functions (move data to code)	It is about variety: Structured, semi-structured, and unstructured data and here the processing functions are taken to the data (move code to data)

A typical Data warehouse Environment

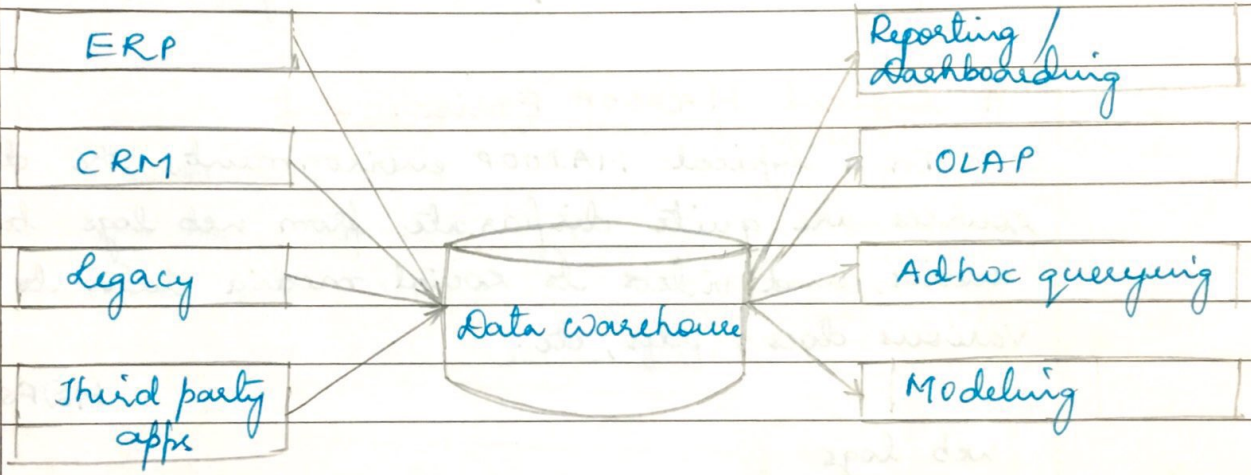


fig: A typical data warehouse environment

In a typical Data Warehouse (DW) environment, operational or transactional or day-to-day business data is gathered from Enterprise Resource Planning (ERP) systems, Customer Relationship Management (CRM), legacy systems, and a third party applications.

The data from these sources may differ in format. Data may come from data sources located in the same geography or different geographies. This data is then integrated, cleaned up, transformed and standardized through the process of extraction, Transformation, and loading (ETL).

The transformed data is then loaded into the enterprise data warehouse (available at enterprise level) or data marts (availabl at the business unit/ functional unit or business process level).

A host of market leading business intelligence and analytics tools are then used to enable decision making from the use of ad-hoc queries, SQL, enterprise dashboards, data mining, etc;

A typical HADOOP Environment

In a typical HADOOP environment, the data sources are quite disparate from web logs to images, audios, and videos to social media data to the various docs, pdfs, etc.

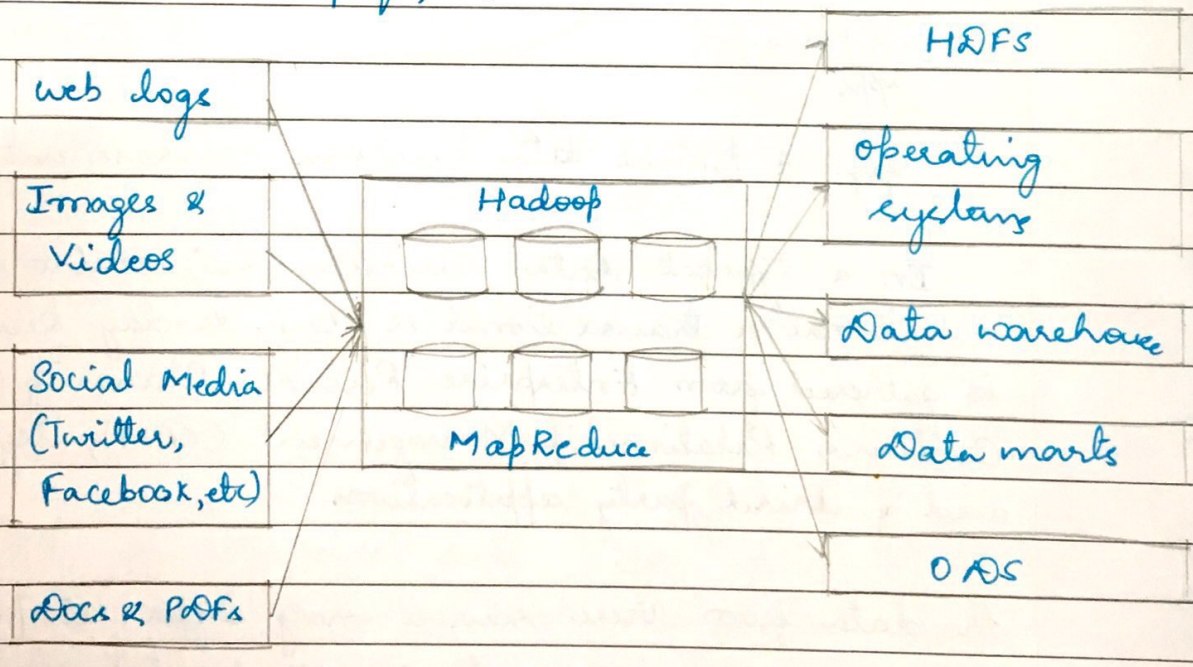


fig: A typical Hadoop Environment

Here, the data in focus is not just the data within the Company's firewall but also data residing outside the company's firewall. This data

is placed in Hadoop Distributed File System (HDFS). If need be, this can be repopulated back to operational systems or fed to the enterprise data warehouse or data marts or operational Data Store (ODS) to be picked for other processing & analysis.

Top challenges facing Big Data

1. Scale:

Storage (RDBMS) or NoSQL (Not only SQL) is one major concern that needs to be addressed to handle the need for scaling rapidly and elastically. The need of the hour is a storage that can best withstand the onslaught of large volume, velocity, and variety of big data? Should you scale vertically or should we scale horizontally?

2. Security:

Most of the NoSQL big data platform have poor security mechanisms (proper authentication and authorization mechanism) when it comes to safeguarding big data. A spot that cannot be ignored given that big data carries credit card information, personal information, and other sensitive data.

3. Schema:

Rigid schemas have no place. We want the technology to be able to fit our big data and not the other way around. The need of the hour is dynamic schema.

4. Continuous availability:

The big question here is how to provide 24/7 support because almost all RDBMS and NoSQL

big data platforms have a certain amount of downtime built in.

5. Consistency:

Should one opt for Consistency or eventual Consistency?

6. Partition Tolerant:

How to build partition tolerant systems that can take care of both hardware & software failures?

7. Data quality:

How to maintain data quality - data accuracy, completeness, timeliness, etc.? do we have appropriate meta data in place?

Why is Big Data Analytics important?

Various approaches to analysis of data and what is leads to.

1. Reactive - Business Intelligence:

Business Intelligence (BI) allows the business to make faster and better decisions by providing the right information to right person at the right time in the right format.

It is about analysis of the past or historical data and then displaying the findings of the analysis or reports in the form of enterprise dash boards, alerts, notifications, etc. It has support for both pre-specified reports as well as adhoc querying

2. Reactive - Big Data Analytics:

Here the analysis is done on huge datasets but the approach is still reactive as it is still based on static data.

3. Proactive - Analytics:

This is to support futuristic decision making by the use of datamining, predictive modeling, text mining, and statistical analysis. This analysis is not on big data as it still uses the traditional database management practices on big data and therefore has severe limitations on the storage capacity and the processing capability.

4. Proactive - Big Data Analytics:

This is sieving through terabytes, petabytes, exabytes of information to filter out the relevant data to analyze. This also includes high performance analytics to gain rapid insights from big data and the ability to solve complex problems using more data.

What kind of technologies are we looking toward to help meet the challenges posed by Big data?

1. Cheap and abundant storage
2. 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, high connectivity, and high throughputs rather than low latency.
5. Cloud computing & other flexible resource allocation arrangements.

- **Scalability:**
  It can process petabytes of data
- **Speed:**
  By means of parallel processing problems that take days to solve, it is solved in hours and minutes
- **Fault Tolerance:**
  MapReduce can take care of failures. If one copy of data is unavailable, another machine has a copy of the same key pair which can be used for solving the same subtask.