CHAPTER 01: TYPES OF DIGITAL DATA

Data

- Any data that can be processed by digital computer and stored in the sequences of 0's and 1's (Binary language) is knowns as digital data.
- Whenever you send an email, read a social media post, or take pictures with your digital camera, you are working with digital data.
- In general, **data** can be any character, text, numbers, voice messages, SMS, WhatsApp messages, pictures, sound, or video.

Data

- **Byte is** the basic unit of information in **computer** storage and processing, and is composed of eight bits; a kilobyte is 1,000 bytes; one megabyte is 1,000 kilobytes . (GB, TB, PB, EB, ZB, YB)
- **Digitizing** is the process of converting information into digital form and is necessary for a computer to be able to process and store the information.

Data

- It is an invaluable asset of any enterprise (big or small).
- Data is present internal to the enterprise and also exists outside the firewalls of the enterprise.
- Data may be in homogeneous or heterogeneous.
- Need of the hour is to
  - Understand, manage, process,
  - and take the data for analysis
  - to draw valuable insights.

Types of digital data

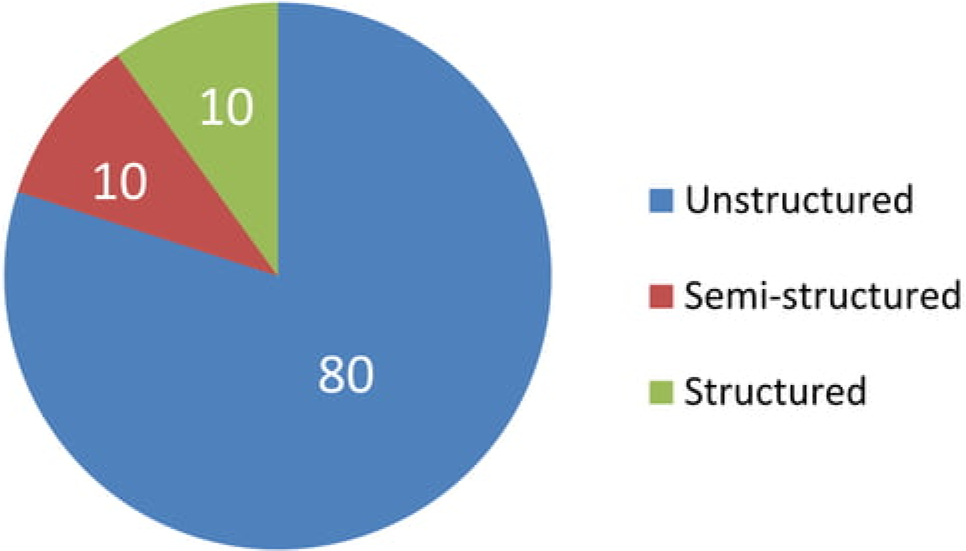
1. **Structured Data :** data stored in the form of rows and columns (databases, Excel)
2. **Un-structured Data**: No pre-defined schema (PPTs, images, Videos, pdfs)
3. **Semi-structured Data:** Hybrid schema (JSON, HTML, XML, Email, and so on),

Structured Data

0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

Unstructured Data

Distribution of digital data (in %)
(by Gartner)



Structured Data

- Data which is in an organized form (In rows & columns).
- Computer programs can use this data easily.
- Relationships exists between entities of data.
- Example
  - Data stored in databases
  - ERP
  - CRM
  - DW
  - Data Cube

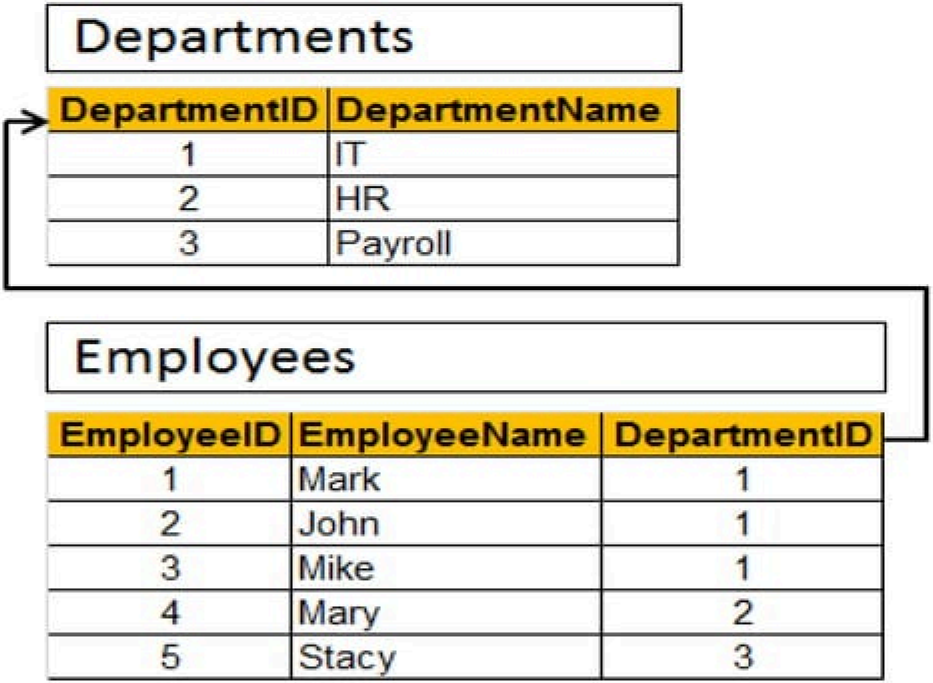
Structured Data

- The data conforms to a pre-defined schema or structure is known as structured data.
- The data can be processed, stored, and retrieved in a fixed format. This data can be processed easily by programs.
- Conforms to a relational data model.
- Structured data is organized in semantic chunks/entities with similar entities grouped together to form relations/tables.

structured Data

- Descriptions for all entities in a group
  - Have the same defined format
  - Have a predefined length
  - Follow the same order.

Example



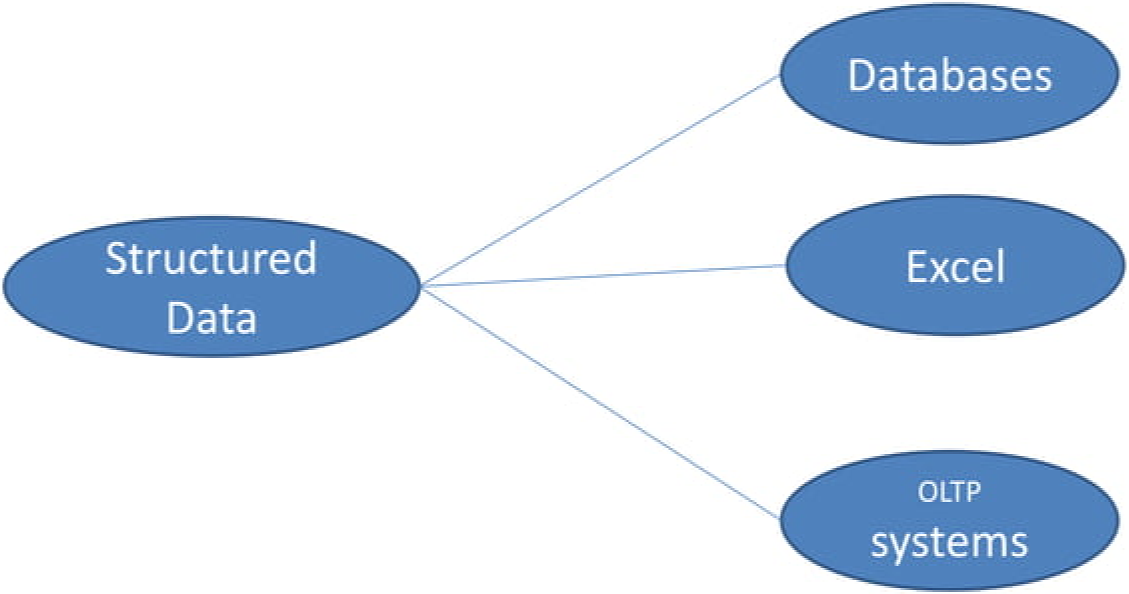
Departments

DepartmentID	DepartmentName
1	IT
2	HR
3	Payroll

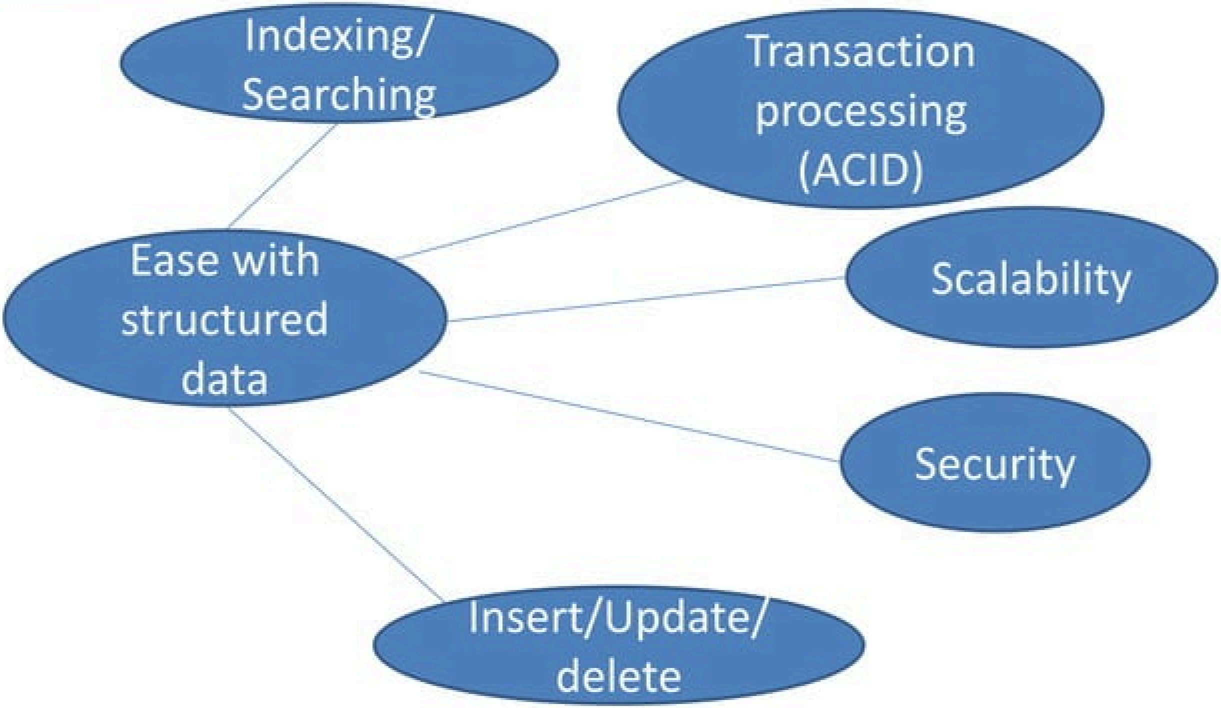
Employees

EmployeeID	EmployeeName	DepartmentID
1	Mark	1
2	John	1
3	Mike	1
4	Mary	2
5	Stacy	3

Sources of Structured Data



Ease with structured data



Database (RDBMS)

- Oracle Corp. – Oracle
- IBM – DB2, IBM-Informix
- Microsoft – SQL
- EMC – Greenplum
- Teradata – Teradata
- Open source- MySQL, PostgresSQL
- Sqlite
- Sequel Pro
- Amazon Aurora
- SAP SQL Anywhere, SAP IQ (Sybase)

Semi-structured Data

- Data which does not conform to a data model but has some structure.
- Computer programs can not use this data easily.
- Example
  - emails
  - XML
  - HTML
  - JSON, and so on.

Semi-structured data (SSD)

- It is referred to as self describing structure.
- It is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables.
- It uses metadata and tags to provide semantic information.

Characteristics of semi-structured data (SSD)

- Does not conform to a data model
- Cannot be stored in the form of rows and columns as in a database.
- The tags and elements are used to describe data.
- Attributes in a group may not be the same.
- Similar entities are grouped.
- Size of the same attributes in a group may differ
- Type of same attributes in group may differ.
- Evolving Schema
- Schema and data are tightly coupled.

Example (Names & Emails)

- One way is:

Name: Raju Patil

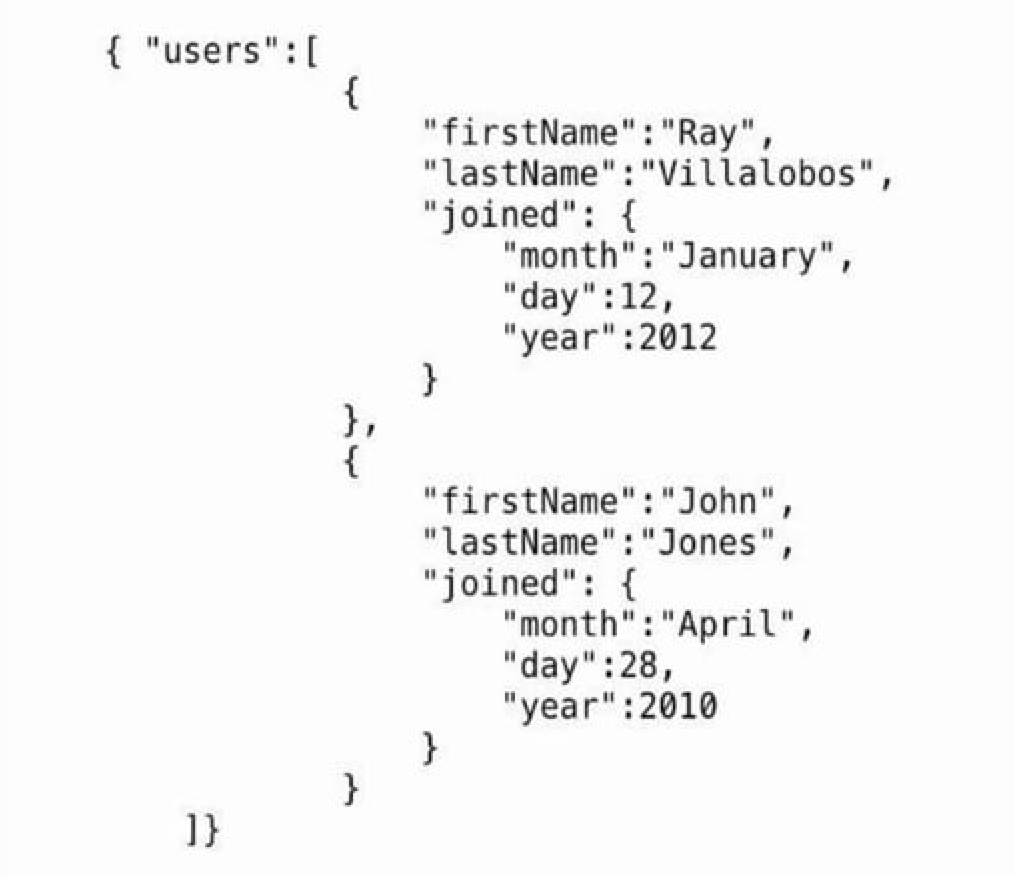
Email : rp@test.tcs.in, rp70@gmail.com

- Another way is:

First Name: Raju

Last Name :Patil

Email : rajup70@gmail.com



```
{ "users":[
            {
                "firstName":"Ray",
                "lastName":"Villalobos",
                "joined": {
                    "month":"January",
                    "day":12,
                    "year":2012
                }
            },
            {
                "firstName":"John",
                "lastName":"Jones",
                "joined": {
                    "month":"April",
                    "day":28,
                    "year":2010
                }
            }
    ]}
```

Sources of SSD

- Email
- XML
- TCP/IP
- Zipped files
- Mark-up languages
- Integration of data from heterogeneous sources.

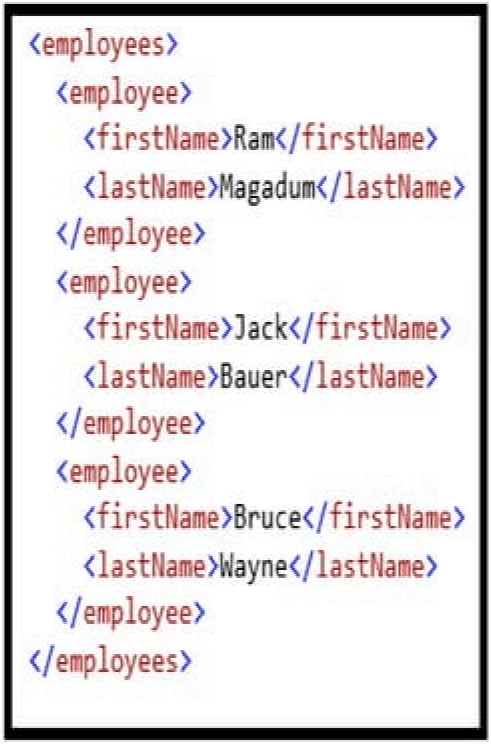
Example: Email format

To:	<Name>
From:	<Name>
Subject:	<Text>
CC:	<Name>
Body:	<Text, Graphics, Images, etc.><Name>

ABC Healthcare Blood Test Report

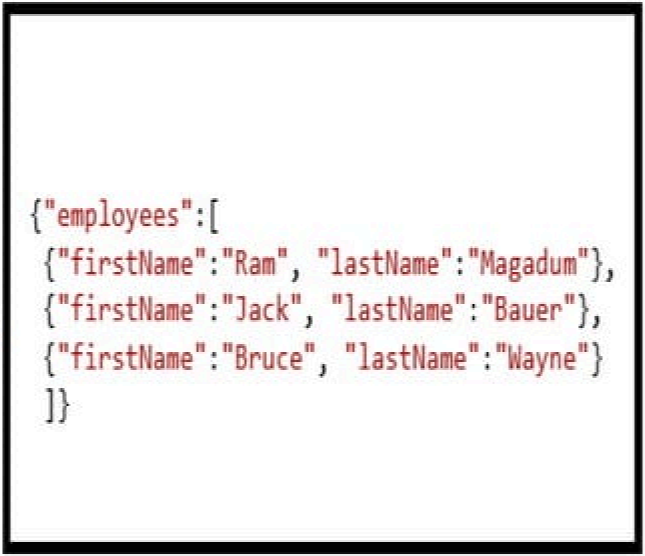
Date	<>	----	
Department	<>	-----	
Patient Name	<>	Attending Doctor	<>
Hemoglobin content	<>	Patient Age	<>
RBC count	<>		
WBC count	<>		
Platelet count	<>		
Diagnosis <notes>			
Conclusion <notes>			

XML



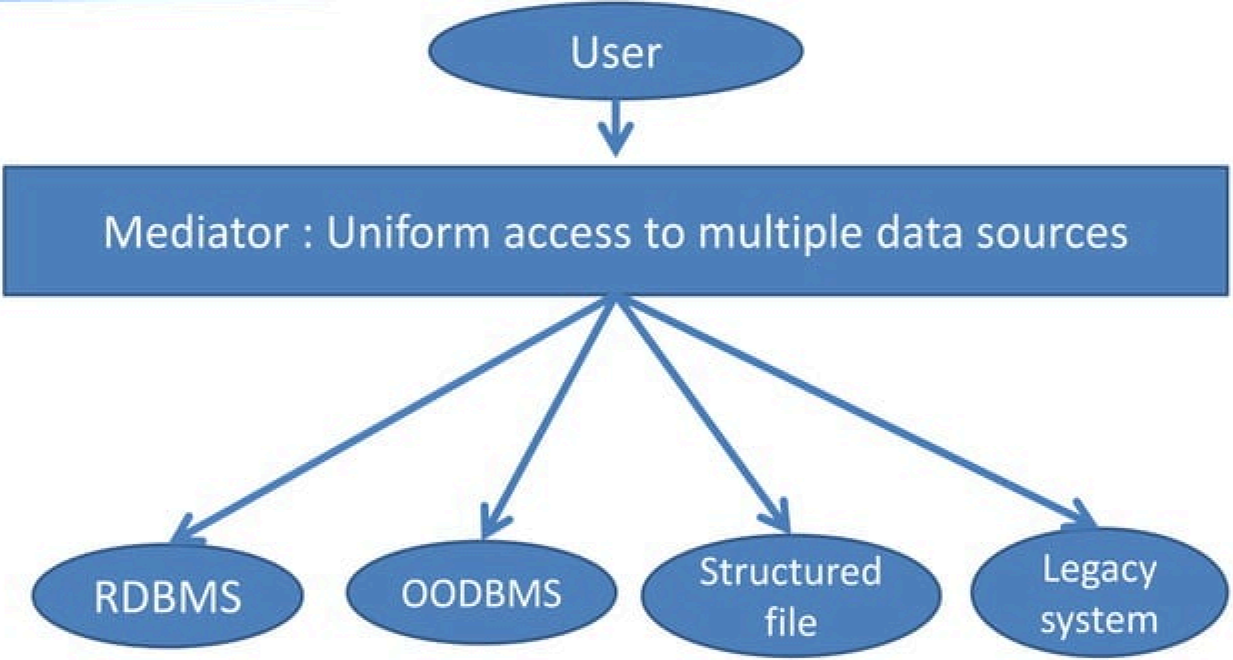
```
<employees>
  <employee>
    <firstName>Ram</firstName>
    <lastName>Magadum</lastName>
  </employee>
  <employee>
    <firstName>Jack</firstName>
    <lastName>Bauer</lastName>
  </employee>
  <employee>
    <firstName>Bruce</firstName>
    <lastName>Wayne</lastName>
  </employee>
</employees>
```

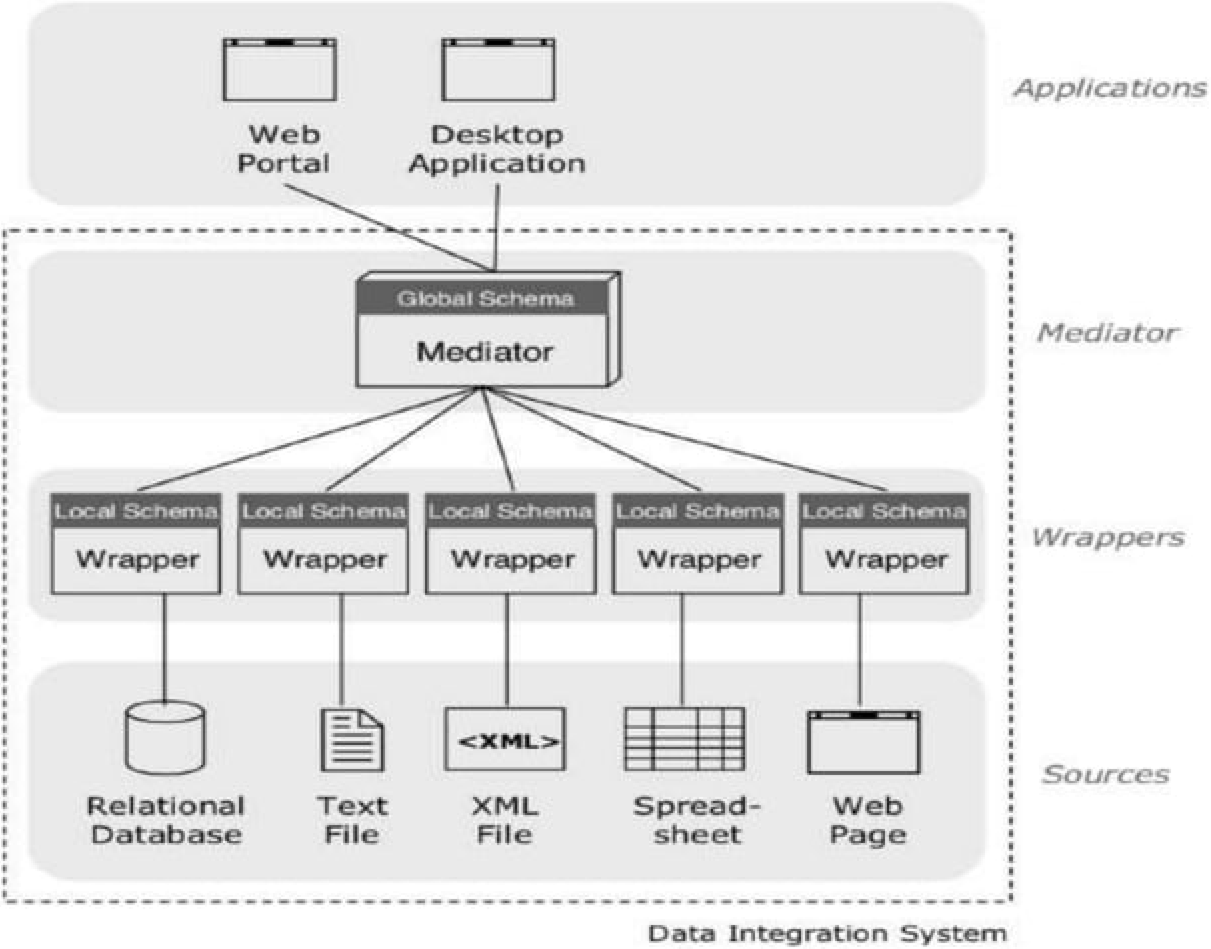
JSON



```
{"employees":[
 {"firstName":"Ram", "lastName":"Magadum"},
 {"firstName":"Jack", "lastName":"Bauer"},
 {"firstName":"Bruce", "lastName":"Wayne"}
 ]}
```

Integration of data from heterogeneous sources




Applications
Web Portal
Desktop Application
Global Schema
Mediator
Mediator
Local Schema
Wrapper
Local Schema
Wrapper
Local Schema
Wrapper
Local Schema
Wrapper
Local Schema
Wrapper
Wrappers
<XML>
Relational Database
Text File
XML File
Spread-sheet
Web Page
Sources
Data Integration System

Getting to know Unstructured data

- Over the past few days, Dr. Ben and Dr. Stanley had been exchanging long emails about a particular case of gastro-intestinal problem.
- Email contains procedure practiced by Dr. Stanley, about combination of drugs that has successfully cured gastro-intestinal disorders in patients.
- Dr. Mark has a patient in the “GoodLife” emergency unit with quite similar case of gastro-intestinal disorder.

Unstructured Data

- Unstructured data refers to the data that lacks any specific form or structure.
- This makes it very difficult and time-consuming to process and analyze unstructured data.
- Data which does not conform to any data model is USD.
- Computer programs can not use this data directly.
- About 80-90% data of an organization is in this format.
- An enormous amount of knowledge is hidden in this data.
- Hence finding useful knowledge/insight from USD is very crucial.

Unstructured Data

- Unstructured data is a generic label for describing data that is not contained in a database or some other type of data structure.
- Unstructured data can be textual or non-textual.
- Textual unstructured data is generated in media like email messages, PowerPoint presentations, Word documents, comments in social media, etc.
- Non-textual unstructured data is generated in media like images, CCTV footage, audio files and video files.
- Anything in a non-database form is unstructured data.

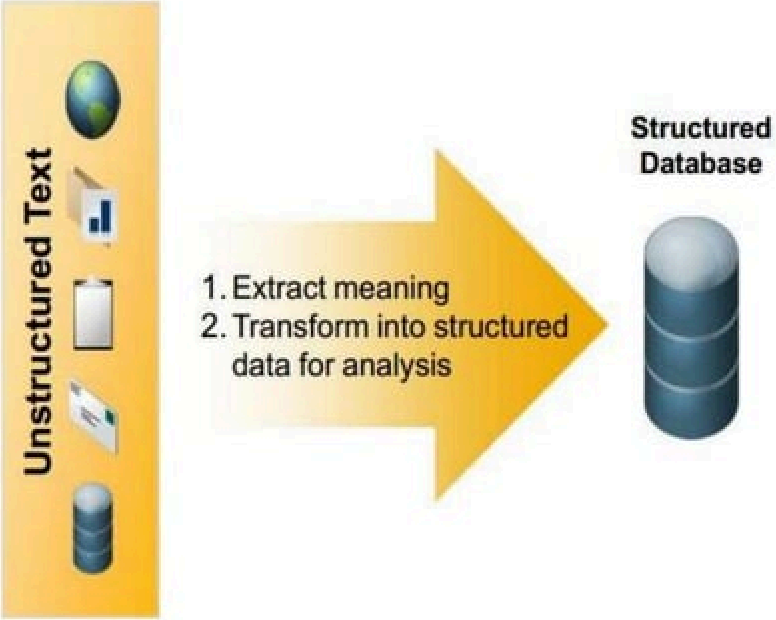
Unstructured Data

- Two types:
  1. Bitmap objects : image, video, or audio files
  2. Textual objects : word, emails, ppts and so on.

Unstructured Data

- Example
  - Memos, QR code (Quick Response), Blogs
  - Chat rooms, Tweets, Comments, likes, tags
  - PPTs, emoji's, emoticons (emotion icons)
  - Images, log files, social media posts
  - Videos, sensor data (raw), weather data
  - Doc files, geospatial data, surveillance data
  - Body of email , GPS data, sensor data, etc.
  - WhatsApp messages, CCTV footage and so on.

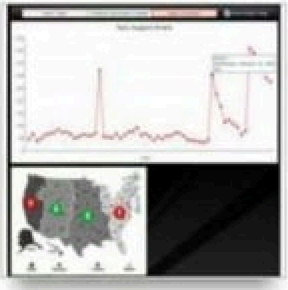
Getting to know Unstructured data



Unstructured Text

1. Extract meaning
2. Transform into structured data for analysis

Structured Database



Once structured it can be...

- Integrated
- Queried
- Analyzed
- Visualized
- Reported against

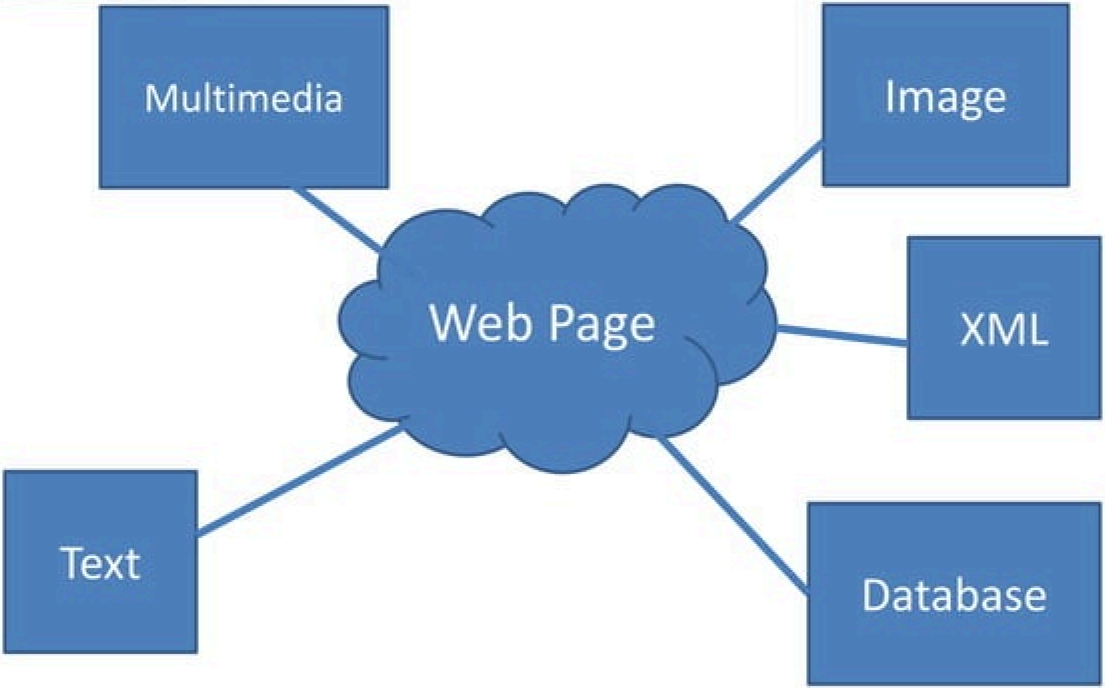
Characteristics of Unstructured data

- This data cannot be stored in the form of rows and columns as in a database and does not conform to any data model.
- It is difficult to determine the meaning of the data.
- It does not follow any rule or semantics, i.e. Not in any particular format or sequence.
- Not easily usable by a program.

Sources of Unstructured data

- Web pages
- Audio and Videos
- Images
- Body of an email
- Word document
- PPT and reports
- Chats and text messages
- Social media data
- White papers
- Surveys
- SMS
- Free form text
- Server Log files
- Product reviews

Web page is unstructured data



Challenges

- Storage space: A lot of space is required to store USD.
- Scalability: As the data grows, scalability becomes an issue and the cost of storing USD increases.
- Retrieve information: Difficult to retrieve required information from USD
- Security: Ensuring security is difficult due to varied sources of data. E.g. emails, web pages, etc.
- Indexing & searching: Very difficult and error-prone as the structure of the USD is not clear.

Challenges

- Interpretation : USD is not easily interpreted by conventional search algorithms.
- Classification : Different naming conventions followed across the organization make it difficult to classify data.
- Deriving meaning : Computer programs cannot automatically derive meaning or structure from USD.
- File formats : Increasing number of file formats makes it difficult to interpret data.

Portion of Unstructured data

Dealing with USD

1. **Data mining**
2. **Text mining /Text Analytics**
3. **NLP**
4. **Noisy text analytics**
5. **Manual tagging with meta data**
6. **Part of speech tagging**
7. **UIMA**
8. **Web Scraping**

Possible Solutions

Data Mining

- It is the computing process of discovering patterns in large data sets involving methods at the intersection of AI, machine learning & DL, statistics, and database systems.
- Popular algorithms:
  - Association rule mining (MBA)
  - Regression Analysis ($Y=mX+ c$)
  - Collaborative filtering

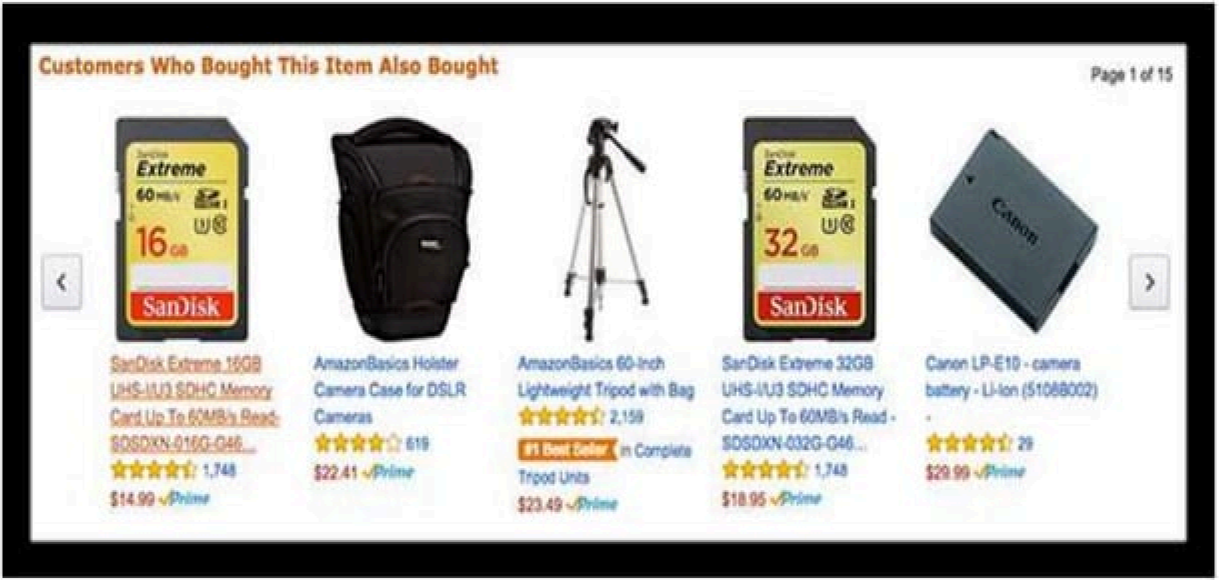
Collaborative filtering

- collaborative filtering uses *similarities between users and items simultaneously* to provide recommendations.
- It is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).
- Collaborative filtering works on a fundamental principle: **you are likely to like what someone similar to you likes.**

Collaborative filtering

- Collaborative filtering (CF) is a technique commonly used
- **Collaborative filtering** (**CF**) is a technique used by recommender systems to build personalized recommendations on the Web.
- Companies that employ CF model include **Amazon, Facebook, Twitter, LinkedIn, Spotify, Google News, Netflix, iTunes.**

Collaborative filtering



Customers Who Bought This Item Also Bought

Page 1 of 15

SanDisk Extreme 16GB UHS-I/U3 SDHC Memory Card Up To 60MB/s Read-SDSDXN-016G-G46...
★★★★½ 1,748
$14.99 ✓Prime

AmazonBasics Holster Camera Case for DSLR Cameras
★★★★☆ 619
$22.41 ✓Prime

AmazonBasics 60-Inch Lightweight Tripod with Bag
★★★★½ 2,159
#1 Best Seller in Complete Tripod Units
$23.49 ✓Prime

SanDisk Extreme 32GB UHS-I/U3 SDHC Memory Card Up To 60MB/s Read - SDSDXN-032G-G46...
★★★★½ 1,748
$18.95 ✓Prime

Canon LP-E10 - camera battery - Li-Ion (5108B002) -
★★★★½ 29
$29.99 ✓Prime

‹
SanDisk Extreme 60 MB/s 16 GB SanDisk
SanDisk Extreme 60 MB/s 32 GB SanDisk
Canon
›

Text analytics or text mining

- It is the process of converting unstructured **text** data into meaningful data for analysis, to measure customer opinions, product reviews, feedback and sentimental analysis to support fact based decision making.
- Uses many linguistic, statistical, and machine learning techniques such as clustering, pattern recognition, tagging, association analysis, predictive analytics, etc.

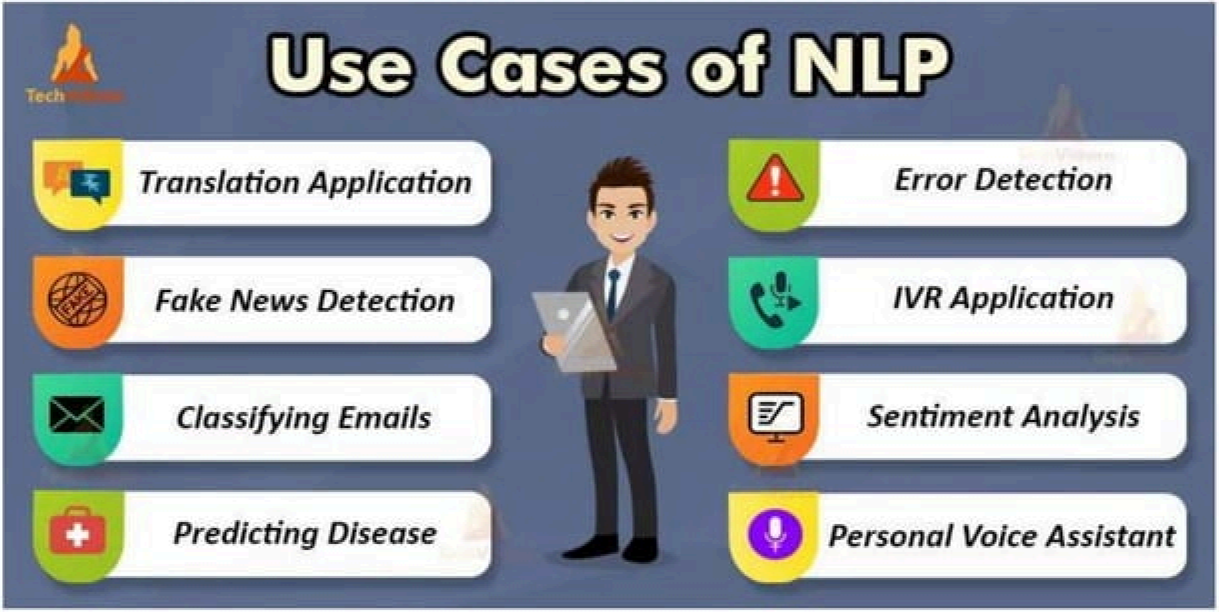
Text analytics or text mining

- It helps organizations to find potentially valuable business insights in corporate documents, customer emails, call center logs, survey comments, social network posts, medical records and other sources of text-based data.
- Text mining capabilities are also being incorporated into AI chatbots/virtual agents that companies deploy to provide automated responses to customers as part of their marketing, sales and customer service operations.

Natural Language Processing (NLP)

- Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken. NLP is a component of artificial intelligence (AI).
- It is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (natural) languages (HCI domain).
- NLP strives to build machines that understand and respond to text or voice data.

Natural Language Processing (NLP)



Use Cases of NLP

- Translation Application
- Fake News Detection
- Classifying Emails
- Predicting Disease
- Error Detection
- IVR Application
- Sentiment Analysis
- Personal Voice Assistant

Noisy text analytics

- It is the process of extracting structured or semi-structured information from noisy unstructured text data such as online chat, text messages, emails, message boards, blogs, wikis, etc.
- The noisy unstructured data comprises one or more of the followings:
  - Spelling mistakes,
  - Acronyms
  - Non-standard words (HBD, K, GN, GM, VGM, etc.)
  - Missing punctuations,
  - Missing letters and so on.

Manual tagging with metadata

- It is the process of tagging manually with adequate metadata to provide the semantics to understand unstructured data.

Road Accident

Image Credit: istockphoto.com/Lamina

Part of Speech Tagging

- It is also called as POS or POST or grammatical tagging.
- It is the process of reading text and tagging each word in the sentence as belonging to a particular part of speech such as “noun”, “verb”, “adjective”, “pronoun”, etc.

Unstructured Information Management Architecture(UIMA)

- It is an open source platform from IBM, which integrates different kinds of analysis engines to provide a complete solution for knowledge discovery from USD.
- It bridge the gap between structured and USD.

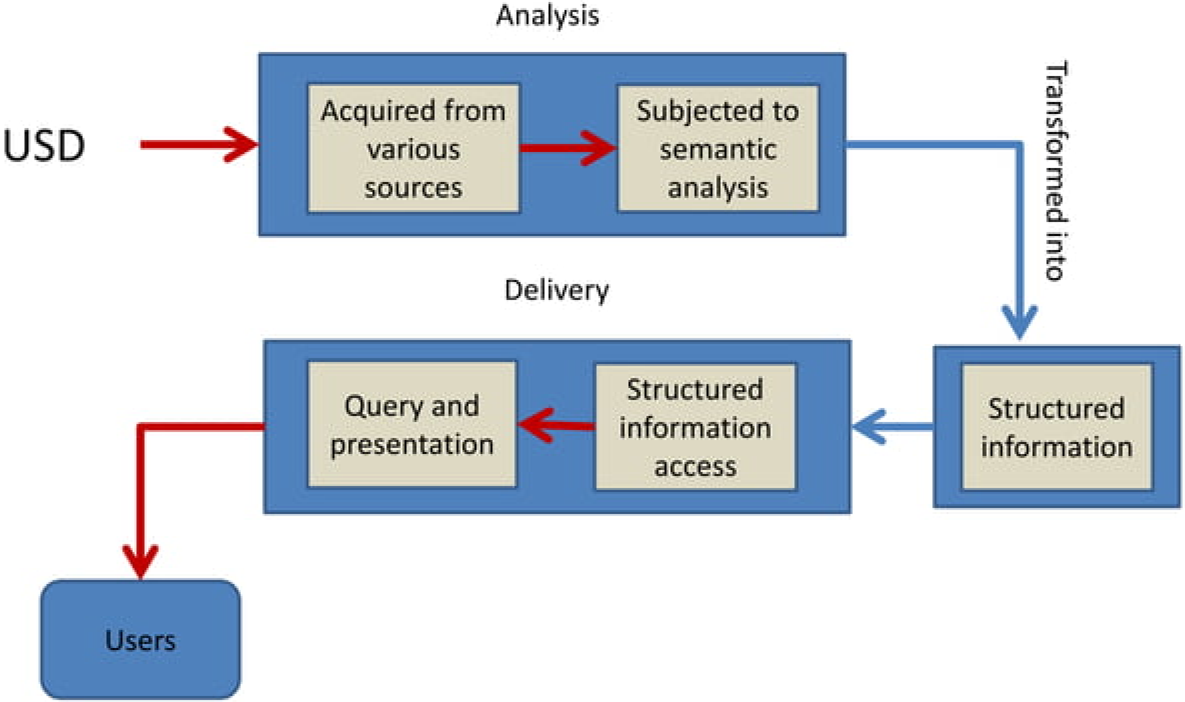
Uses of UIMA

- Used to convert unstructured data such as repair logs and service notes into relational tables.
- These tables can then be used by automated tools to detect maintenance or manufacturing problems.

Uses of UIMA

- Used in medical contexts to analyze clinical notes, such as the Clinical Text Analysis and Knowledge Extraction System (Apache CTAKES).
- CTAKES is an open-source Natural Language Processing (NLP) system that extracts clinical information from electronic health/medical record free-text (Users are **free** to type whatever they want in any form).

UIMA block diagram



Analysis

USD

Acquired from various sources

Subjected to semantic analysis

Transformed into

Delivery

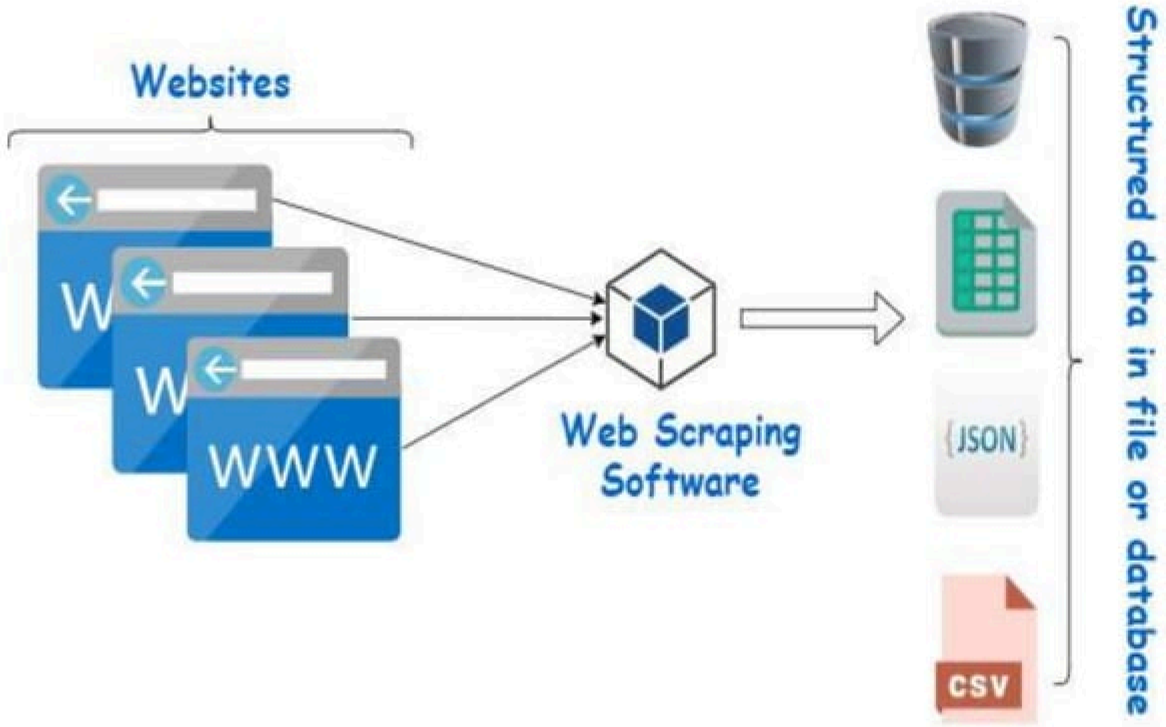
Structured information

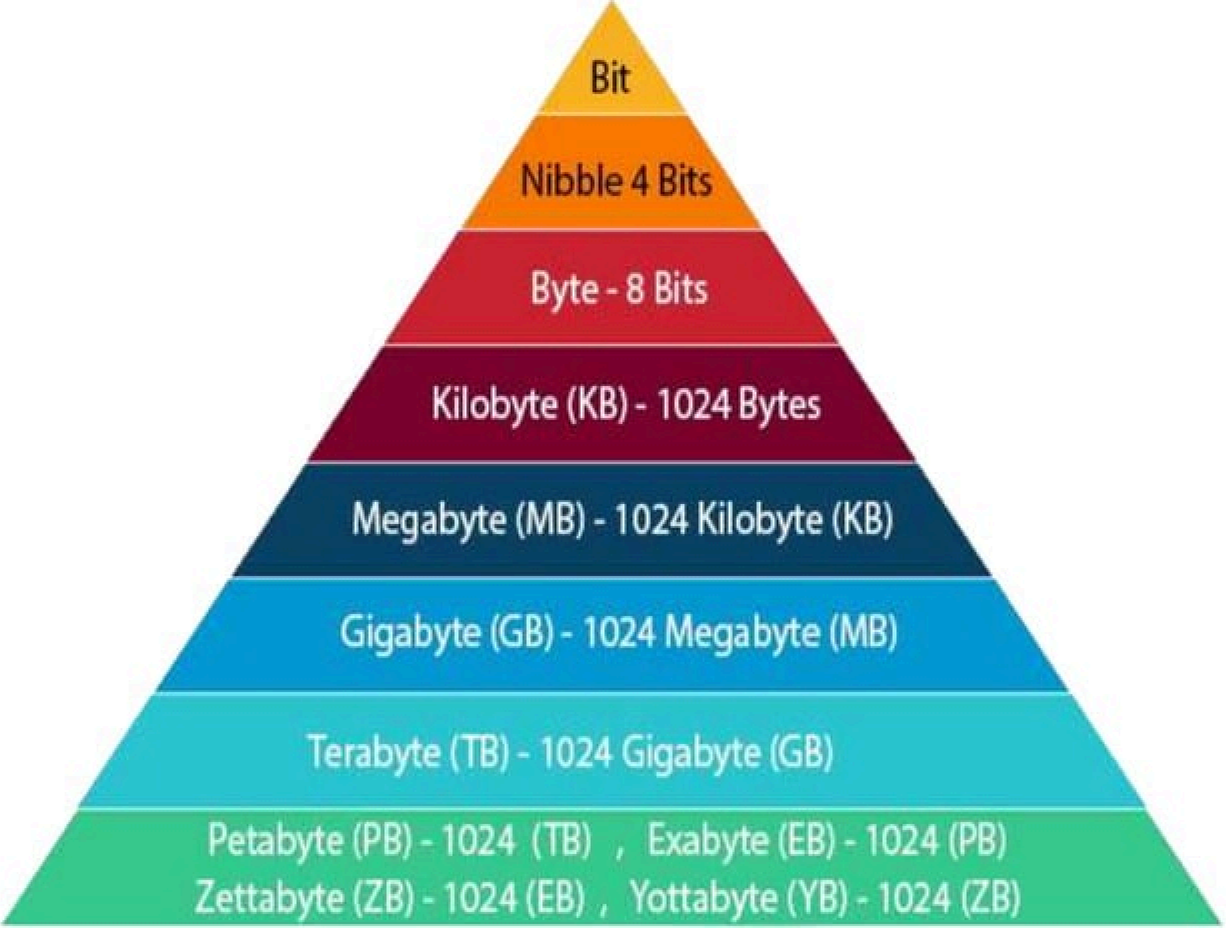
Structured information access

Query and presentation

Users

Web Scraping





Bit

Nibble 4 Bits

Byte - 8 Bits

Kilobyte (KB) - 1024 Bytes

Megabyte (MB) - 1024 Kilobyte (KB)

Gigabyte (GB) - 1024 Megabyte (MB)

Terabyte (TB) - 1024 Gigabyte (GB)

Petabyte (PB) - 1024 (TB) , Exabyte (EB) - 1024 (PB)
Zettabyte (ZB) - 1024 (EB) , Yottabyte (YB) - 1024 (ZB)

Big Data

- Big data is a term that **describes large, hard-to-manage volumes of data – both structured and unstructured -** none of traditional data management tools can store it or process it efficiently.
- experts now **predict that 74 zettabytes of data** will be in existence by 2021.

Big Data

- Every day, we create 2.5 quintillion(10^{18}) bytes of data —90% of the data in the world today has been created in the last two years alone.
- This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, WhatsApp, IOT and so on.

Characteristics of Data

- **Composition:** Deals with structure of data, i.e., sources of data, the granularity(Ex. Postal address), the types, nature of data (Static or real-time).
- **Condition:** Deals with the state of data, that is, “Can one use data as it is for analysis?” or “Does it require cleansing for further enhancement and enrichment?”.

Characteristics of Data

- **Context:** Deals with
  - Where, this data has been generated?
  - Why this data generated?
  - How sensitive is this data?
  - What are the events associated with this data?
  - And so on.

Gartner

- Is a global research and advisory firm providing insights, advice, and tools for leaders in IT, Finance, HR, Customer Service and Support

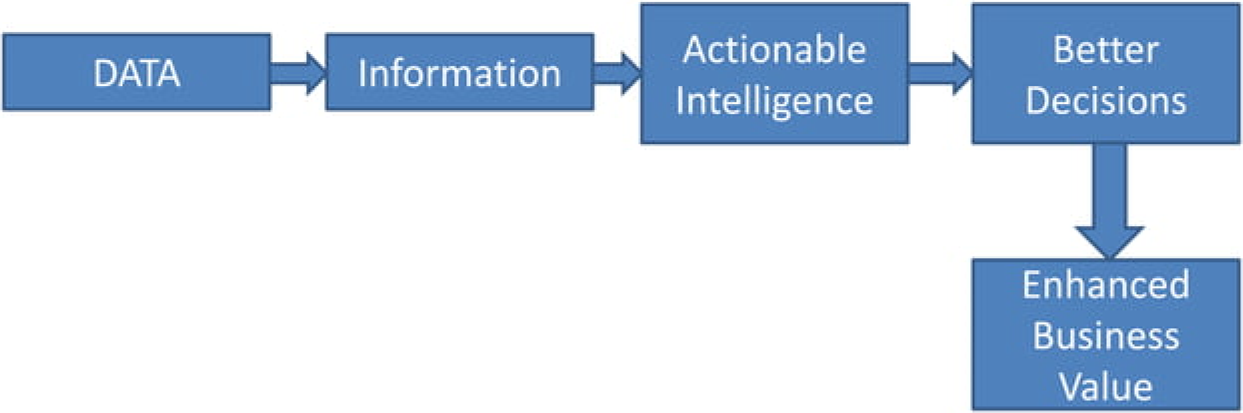
Big data definition- Gartner

- Big data is high-volume, high-velocity, and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.
- Cost effective and innovative forms of information processing: Talks about embracing new techniques and technologies to capture, store, process, persevere, integrate and visualize the big data(3vs).

Definition of Big data by Gartner

- Enhanced insight and decision making: Talks about deriving deeper, richer, and meaningful insights and then using these insights to make faster and better decisions to gain business value and thus a competitive edge.

Big data formula



DATA

Information

Actionable Intelligence

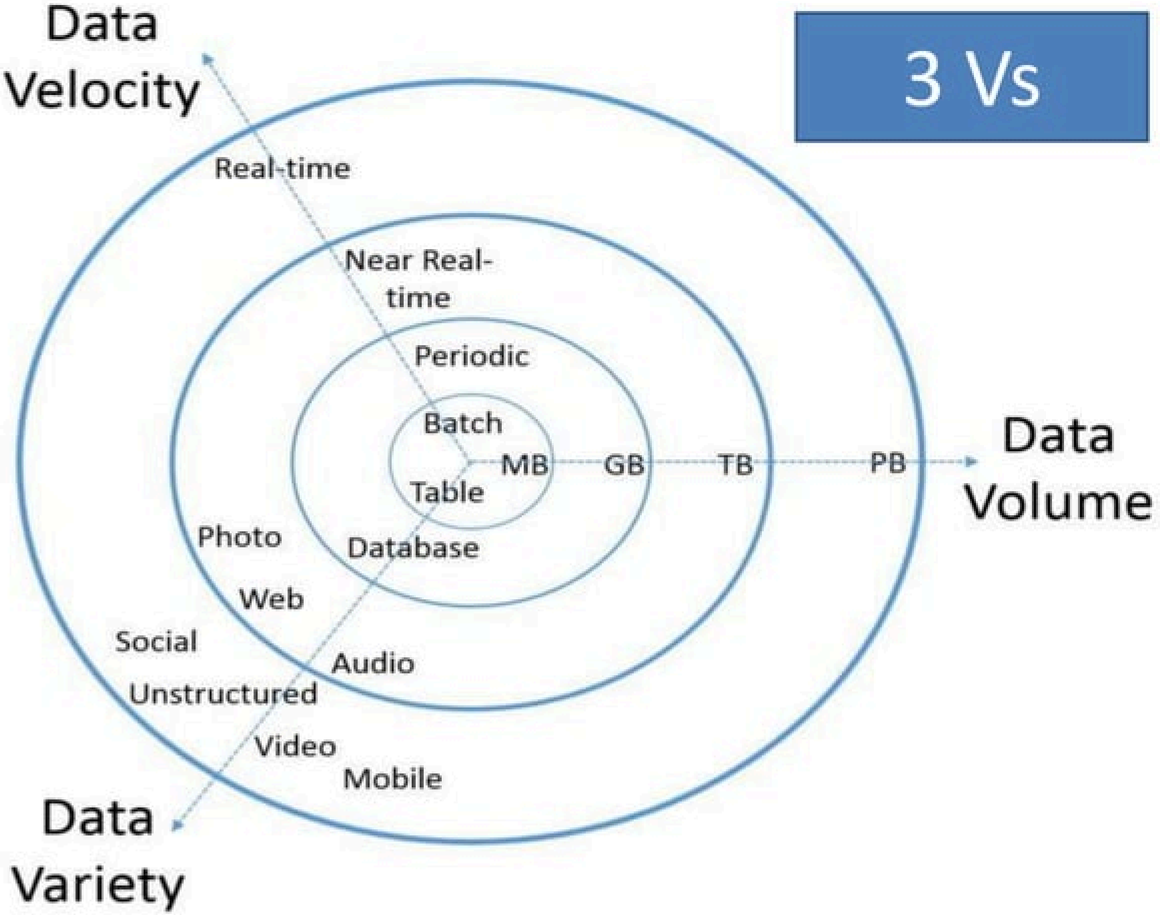
Better Decisions

Enhanced Business Value

Challenges with Big Data

- Capture
- Storage (Solution: Cloud Computing)
- Curation (Management of data + Data retention)
- Search
- Analysis
- Transfer
- Visualization
- Privacy violations

3 Vs



Data Velocity
Real-time
Near Real-time
Periodic
Batch
MB
GB
TB
PB
Data Volume
Table
Database
Photo
Web
Social
Audio
Unstructured
Video
Mobile
Data Variety

3 V’s of Big data

- The data that is big in Volume, Velocity and Variety is known as big data.

Sources of big data

- **Archives:** Archives of scanned documents, customer correspondence records, patient's health records, student's admission records, students' assessment records and so on.
- **Sensor data:** Car sensors, smart electric meters, office buildings, washing m/c, other electronic appliances and so on.
- **Machine log data:** Event logs, application logs, audit logs, server logs, etc.

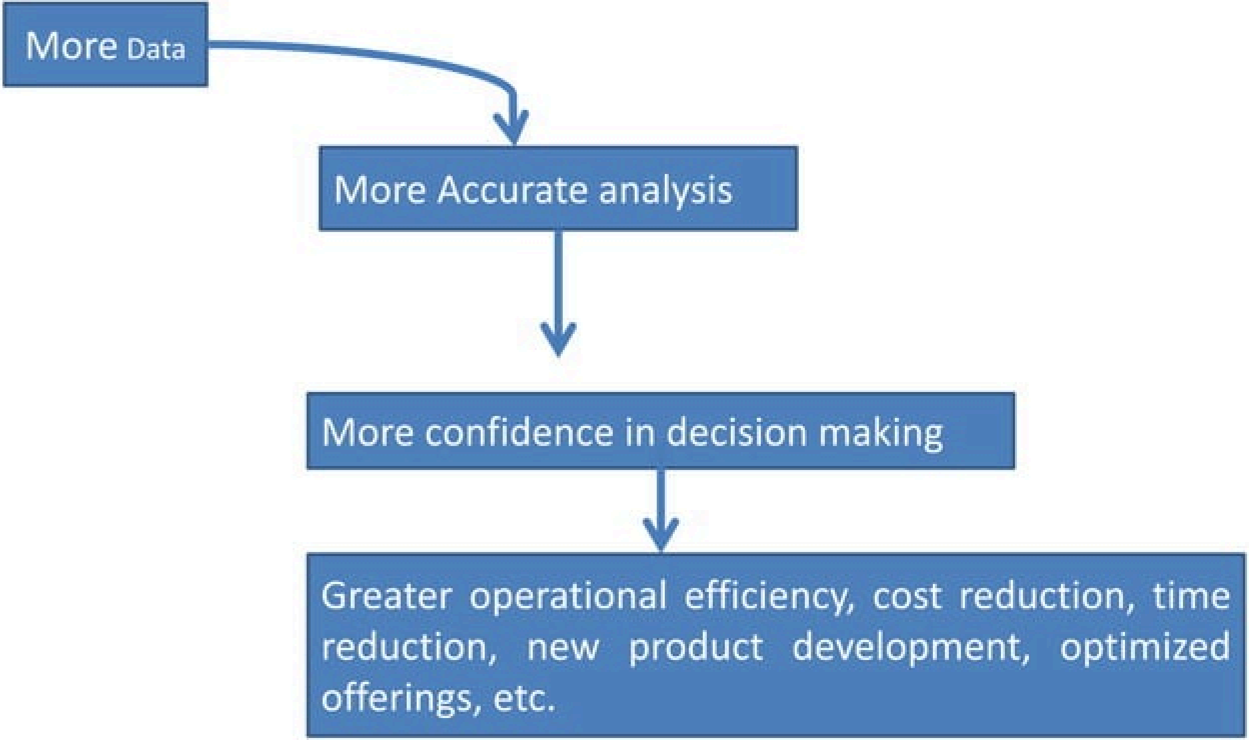
Sources of big data

- **Public web:** Wikipedia, Weather, regulatory, census, etc.
- **Data storage:** File systems, SQL database, NoSQL database (Mongo DB, Cassandra) and so on.
- **Media:** Audio, Video, image, etc.
- **Docs:** CSV, word docs, PDF, PPT, XLS, etc.
- **Business Apps:** ERP, CRM, HR, Google Docs, etc.
- **Social media:** Twitter blogs, Facebook, LinkedIn, YouTube, Instagram, etc.
- **IOT**

Other characteristics of big data

- **Veracity and Validity:** Refers to the accuracy (quality) and correctness of the data.
- **Volatility:** Deals with how long the data is valid?, and how long should it be stored?. (OTP, Aadhar No., PW)
- **Variability:** Data flows can be highly inconsistent with periodic peaks. **(In total 7V's of big data)**

Why Big data



More Data

More Accurate analysis

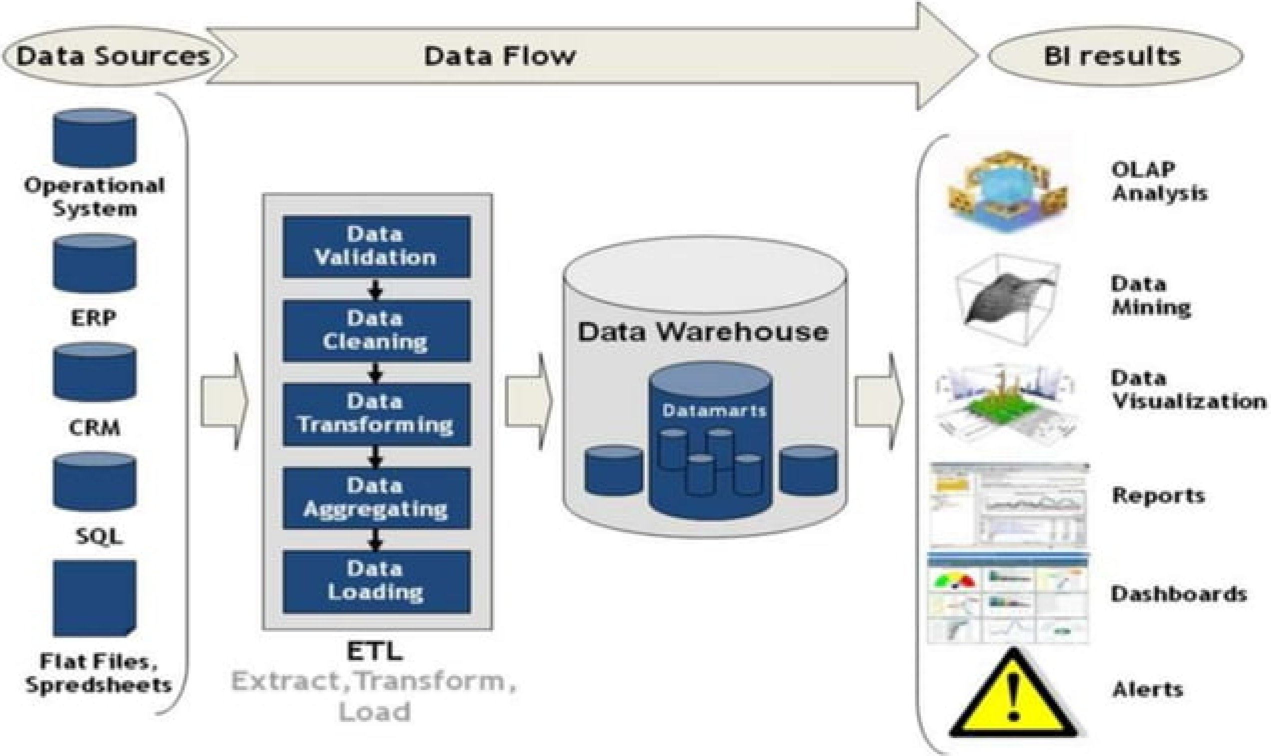
More confidence in decision making

Greater operational efficiency, cost reduction, time reduction, new product development, optimized offerings, etc.

Three reasons for leveraging big data

1. Competitive Advantage.
2. Decision making
3. To create new business value out of data.

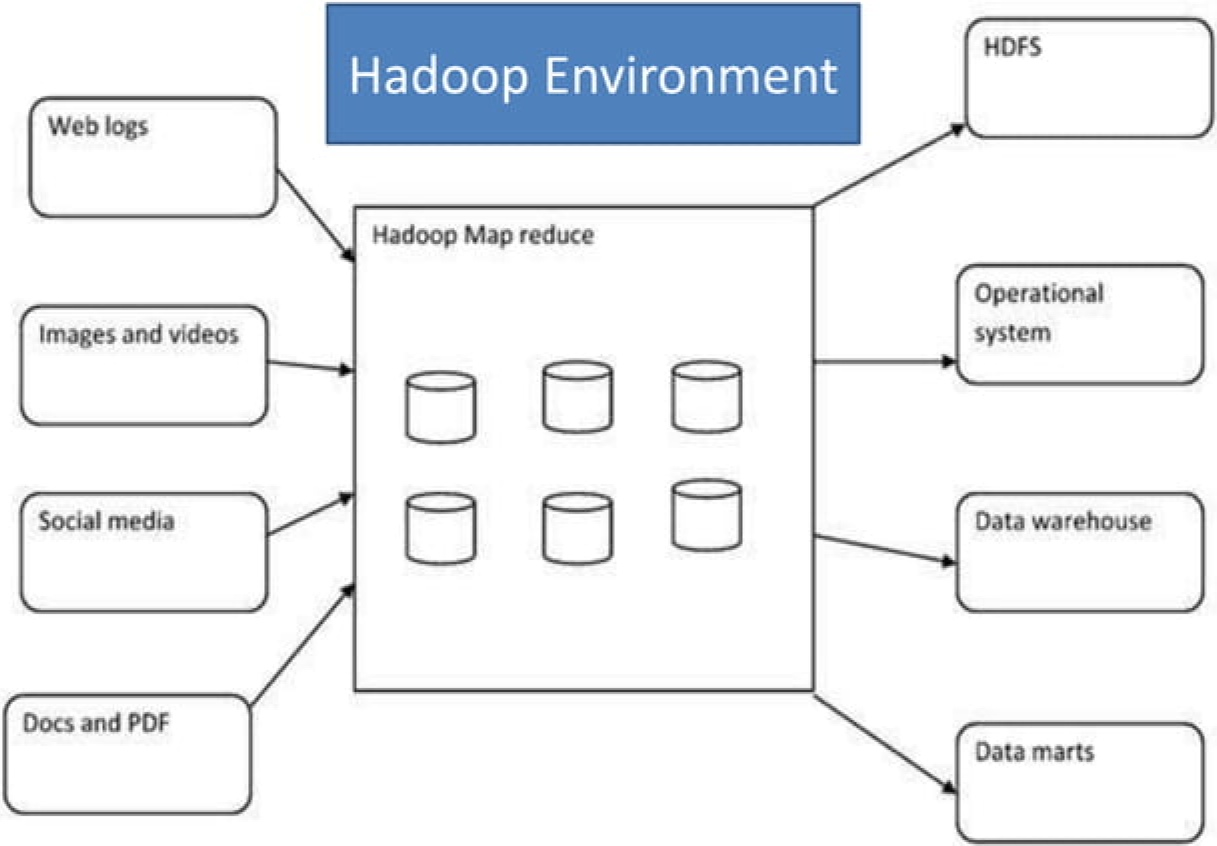
Typical data warehouse Environment



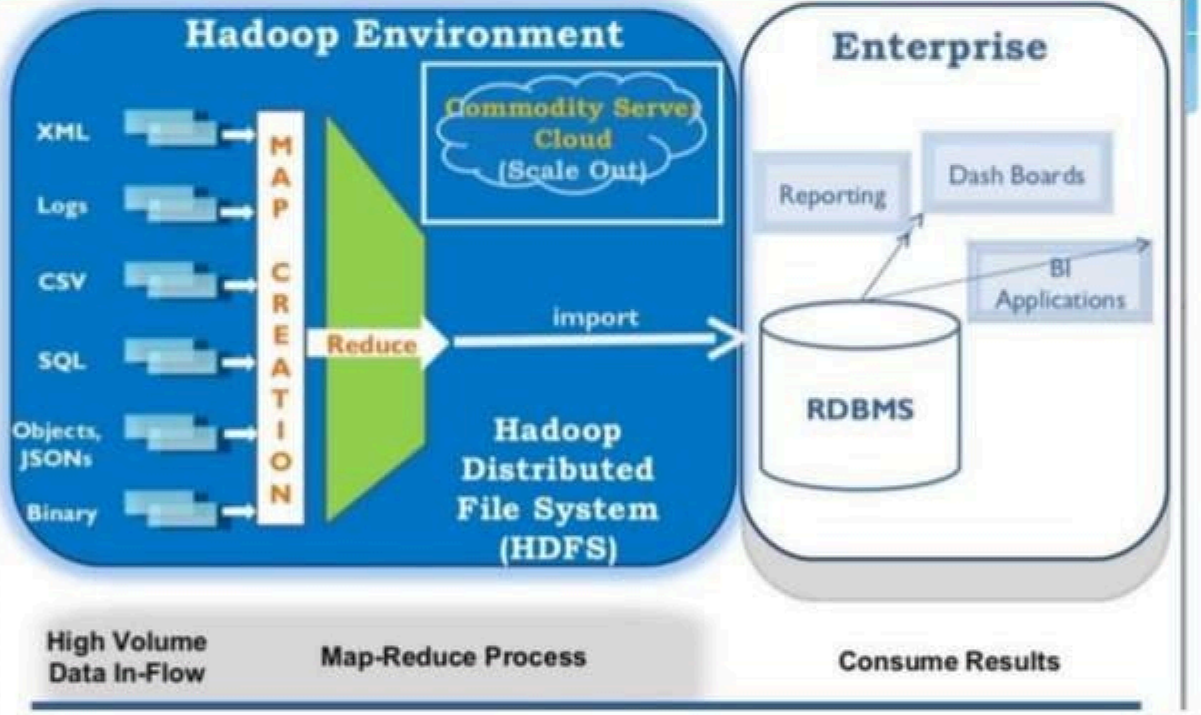
Typical Hadoop Environment

- It is different from DW environment.
- Here data sources are web logs, images, audios, videos, social media, doc files, pdfs, etc.

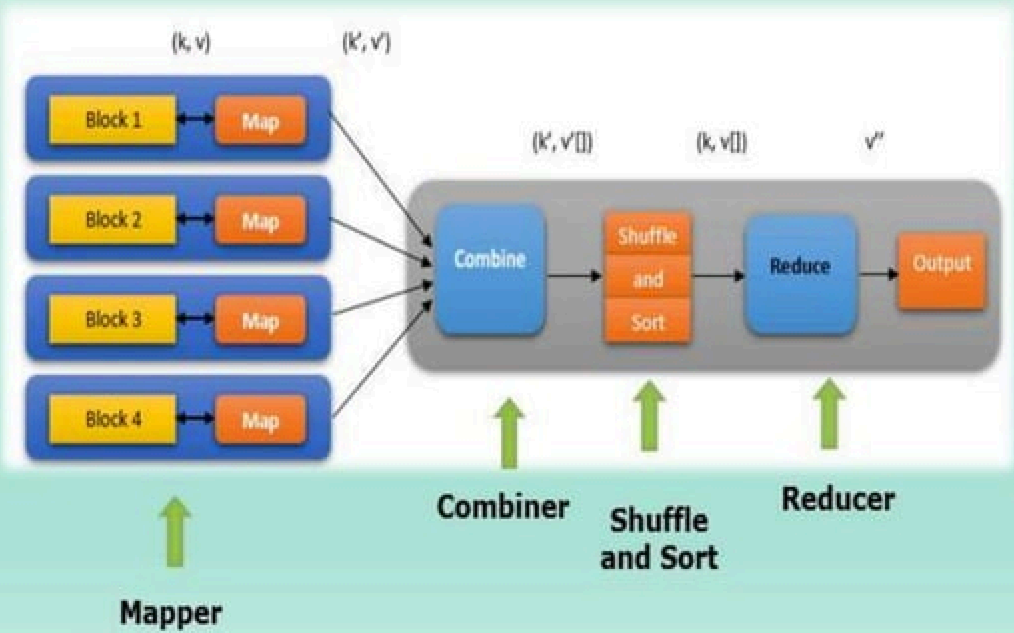
Hadoop Environment



A typical Hadoop System

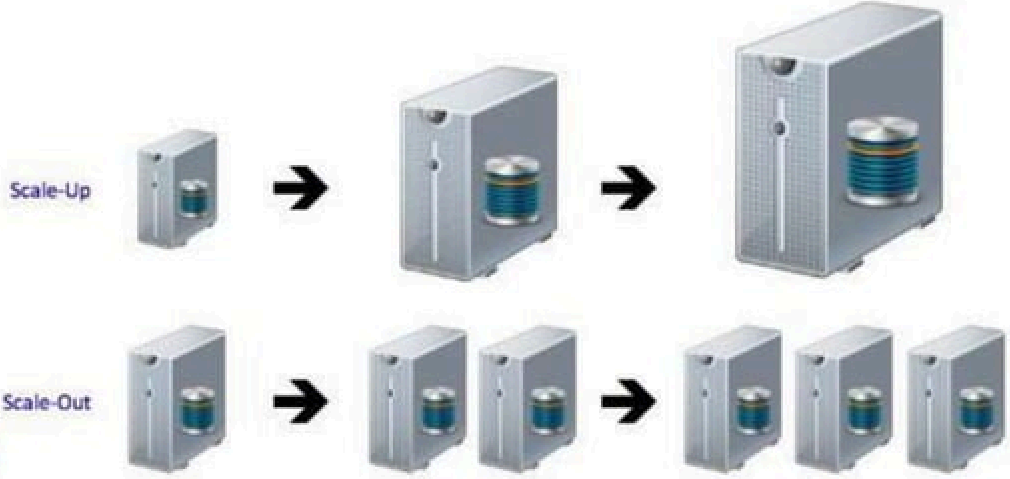


How MapReduce Works

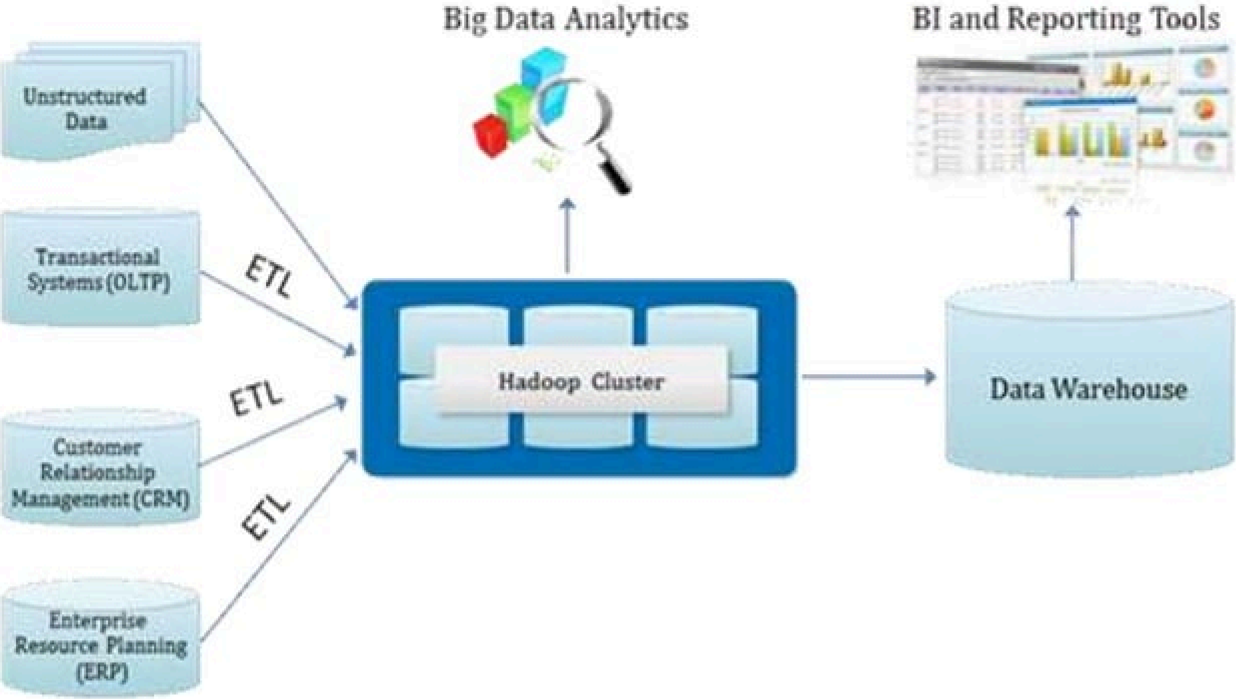


www.educba.com

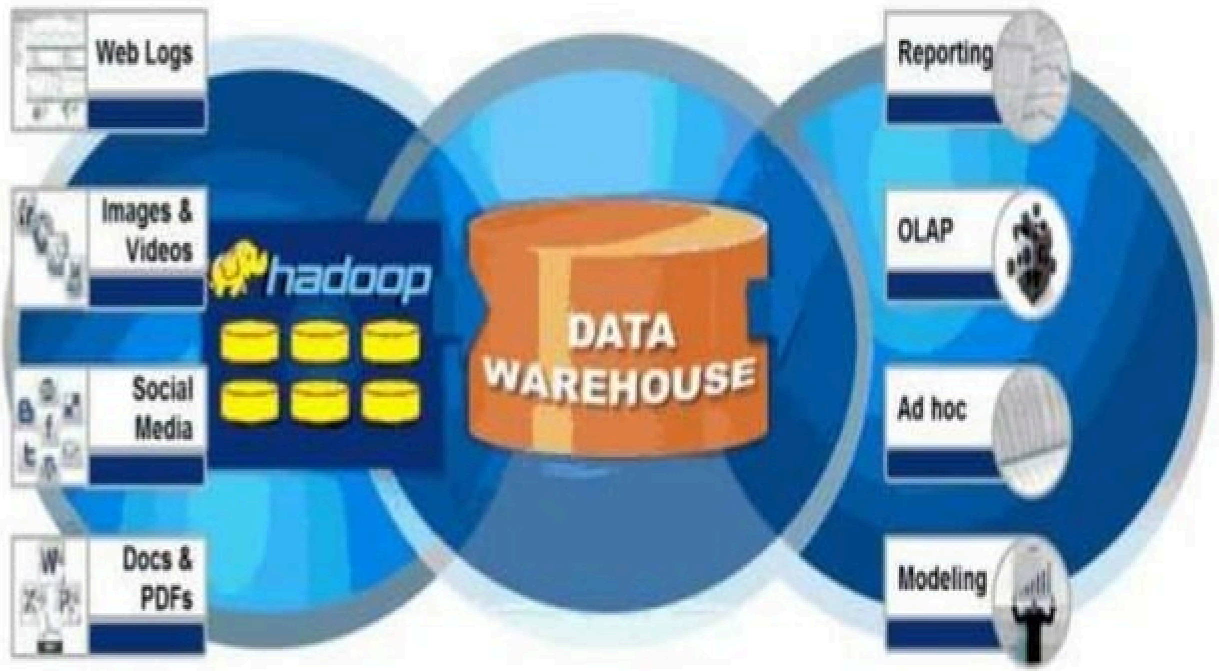
Scale Up vs Scale Out



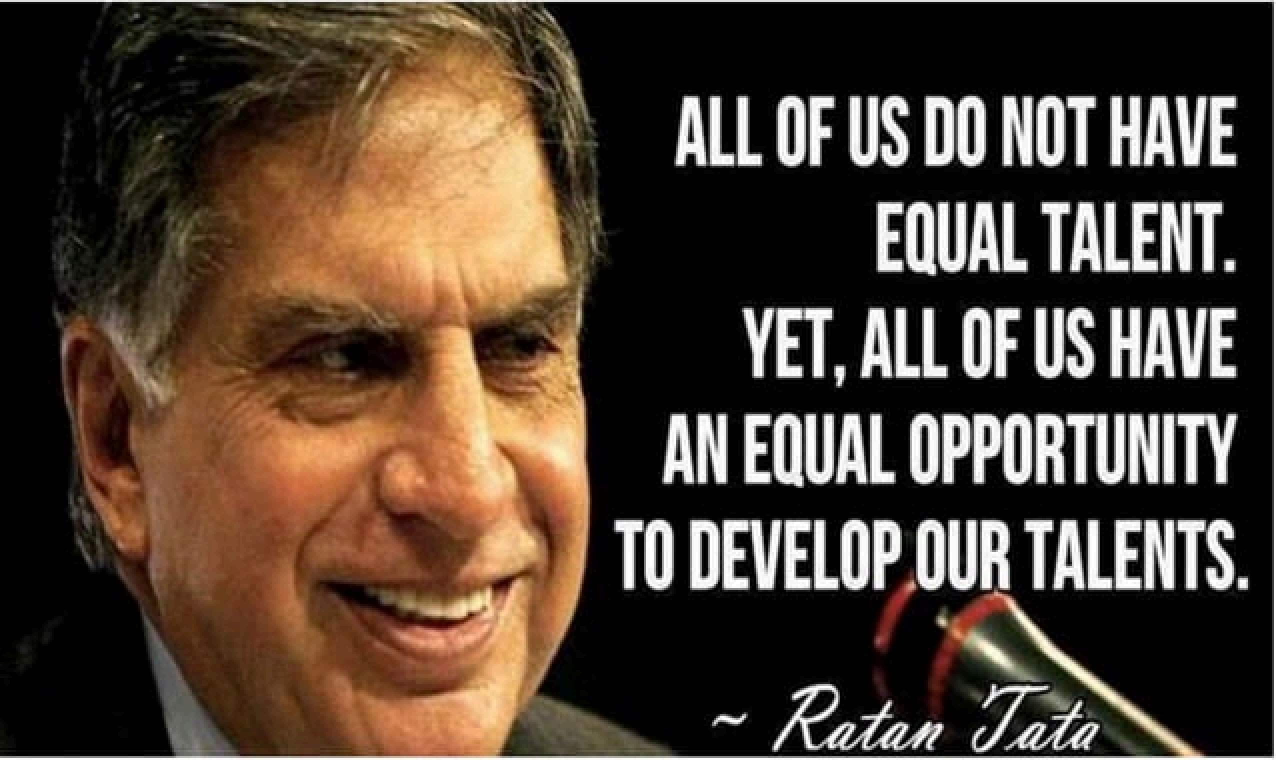
Big data & DW coexistence


Big Data Analytics
BI and Reporting Tools
Unstructured Data
Transactional Systems (OLTP)
Customer Relationship Management (CRM)
Enterprise Resource Planning (ERP)
ETL
ETL
ETL
Hadoop Cluster
Data Warehouse

Big data & DW coexistence



© 2013 SAS Best Practices



ALL OF US DO NOT HAVE
EQUAL TALENT.
YET, ALL OF US HAVE
AN EQUAL OPPORTUNITY
TO DEVELOP OUR TALENTS.

~ Ratan Tata