BIG DATA ANALYTICS (2180710)

Prepared By: Dr. Sheshang Degadwala
In charge Principal,
Head of Computer Engineering,
Sigma Institute of Engineering

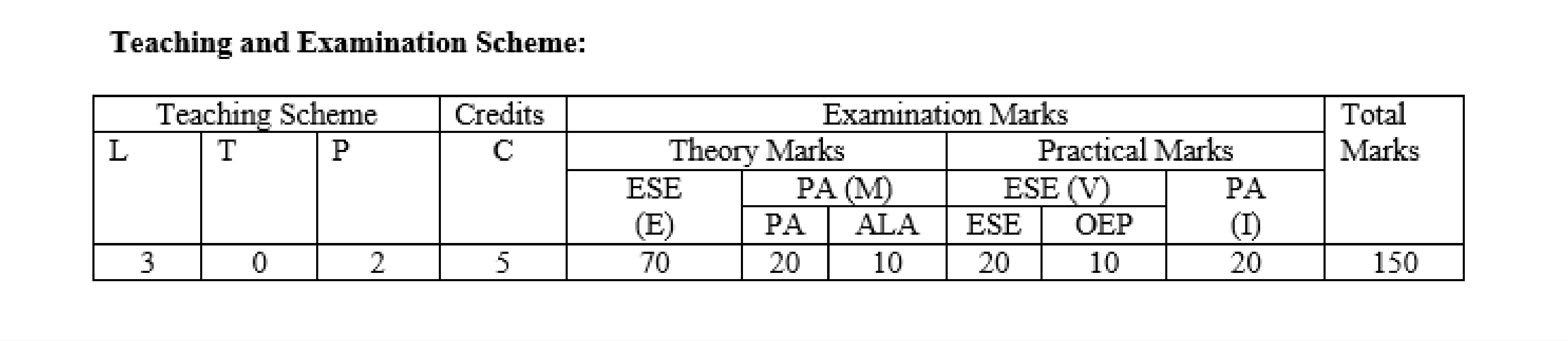
Reference Books

- BIG Data and Analytics , Sima Acharya, Subhashini Chhellappan, Willey

- MongoDB in Action, Kyle Banker,Piter Bakkum , Shaun Verch, Dream tech Press

- http://www.bigdatauniversity.com/

Teaching Scheme



Teaching and Examination Scheme:

Teaching Scheme			Credits	Examination Marks						Total Marks
L	T	P	C	Theory Marks			Practical Marks			
				ESE (E)	PA (M)		ESE (V)		PA (I)	
					PA	ALA	ESE	OEP		
3	0	2	5	70	20	10	20	10	20	150

Subject Content

- Unit - 1 INTRODUCTION TO BIG DATA
  - *Introduction– distributed file system–Big Data and its importance, Four Vs, Drivers for Big data, Big data analytics, Big data applications. Algorithms using map reduce*
- Unit- 2 INTRODUCTION TO HADOOP AND HADOOP ARCHITECTURE
  - *Big Data – Apache Hadoop & Hadoop EcoSystem, Moving Data in and out of Hadoop – Understanding inputs and outputs of MapReduce -, Data Serialization.*
- Unit – 3 HDFS, HIVE AND HIVEQL, HBASE
  - *HDFS-Overview, Installation and Shell, Java API; Hive Architecture and Installation, Comparison with Traditional Database, HiveQL Querying Data, Sorting And Aggregating, Map Reduce Scripts, Joins & Sub queries, HBase concepts, Advanced Usage, Schema Design, Advance Indexing, PIG, Zookeeper , how it helps in monitoring a cluster, HBase uses Zookeeper and how to Build Applications with Zookeeper.*

Subject Content

- Unit - 4 SPARK
  - *Introduction to Data Analysis with Spark, Downloading Spark and Getting Started, Programming with RDDs, Machine Learning with MLlib. 5*
- *Unit- 5 NoSQL*
  - *What is it?, Where It is Used Types of NoSQL databases, Why NoSQL?, Advantages of NoSQL, Use of NoSQL in Industry, SQL vs NoSQL, NewSQL*
- Unit – 6 Data Base for the Modern Web
  - *Introduction to MongoDB key features, Core Server tools, MongoDB through the JavaScript's Shell, Creating and Querying through Indexes, Document-Oriented, principles of schema design, Constructing queries on Databases, collections and Documents , MongoDB Query Language.*

INTRODUCTION OF BIG DATA

Big Data is EveryWhere!

- Lots of data is being collected and warehoused
  - *E-commerce, E-shopping*
  - *Purchases at Department/ Grocery stores*
  - *Bank/Credit Card transactions*
  - *Social Network*
  - *Web Data*

Data Units

Name	Equal to:	Size in Bytes
Bit	1 bit	1/8
Nibble	4 bits	1/2 (rare)
Byte	8bits	1
Kilobyte	1,024 bytes	1,024
Megabyte	1,024 kilobytes	1,048,576
Gigabyte	1,024 megabytes	1,073,741,824
Terrabyte	1,024 gigabytes	1,099,511,627,776
Petabyte	1,024 terrabytes	1,125,899,906,842,624
Exabyte	1,024 petabytes	1,152,921,504,606,846,976
Zettabyte	1,024 exabytes	1,180,591,620,717,411,303,424
Yottabyte	1,024 zettabytes	1,208,925,819,614,629,174,706,176

Type of Data

- Structured data Relational Data (Tables/Transaction)
- Unstructured data Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - *Social Network, Semantic Web , …*
- Streaming Data
  - *You can only scan the data once*

Need Analysis

- Structured data Relational Data (Tables/Transaction)
- More than 2 billion internet users
- Users spend more than 16 billion minutes on Facebook each day
- Each month more than 3 billion photos are uploaded to Facebook
- More than 10 TB data process by Facebook every day
- More than 7 TB data processed by Twitter every day
- New York Stock Exchange store >1 TB/day
- Storage requirement increase every year by rate of more than 10%

How much data?

- Google processes 20 PB a day
- Wayback Machine has 3 PB + 100 TB/month
- Facebook has 2.5 PB of user data + 15 TB/day
- eBay has 6.5 PB of user data + 50 TB/day

Why Big Data?

- **Growth of Big Data is needed**
  - *Increase of storage capacities*
  - *Increase of processing power*
  - *Availability of data(different data types)*
  - *Every day we create 2.5 quintillion bytes of data; 90% of the data in the world today has been created in the last two years alone*

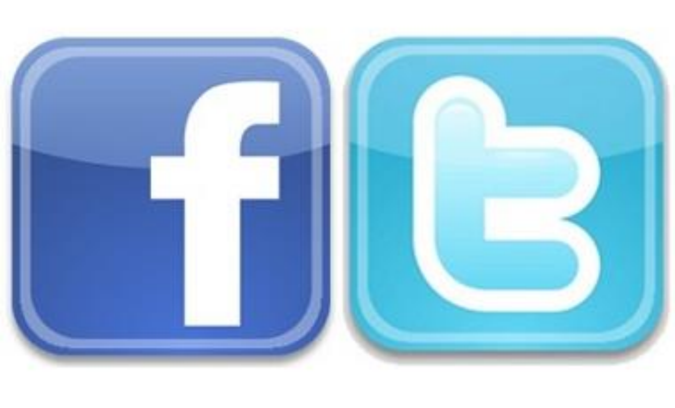
How Is Big Data Different?

1) Automatically generated by a machine

(e.g. Sensor embedded in an engine)

2) Typically an entirely new source of data

(e.g. Use of the internet)



3) Not designed to be friendly

(e.g. Text streams)

4) May not have much values

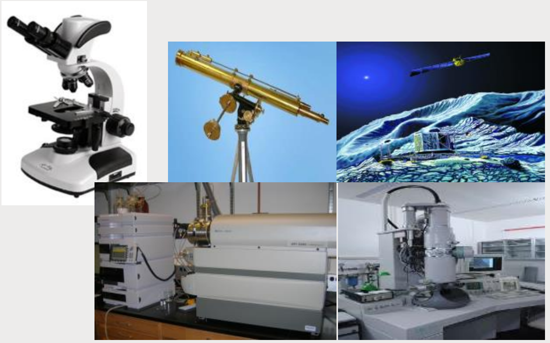
- *Need to focus on the important part*

Big Data Analytics

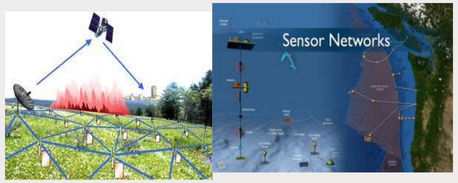
- Examining large amount of data
- Appropriate information
- Identification of hidden patterns, unknown correlations
- Competitive advantage
- Better business decisions: strategic and operational
- Effective marketing, customer satisfaction, increased revenue

Who’s Generating Big Data?

Social media and networks
(all of us are generating data)



Scientific instruments
(collecting all sorts of data)



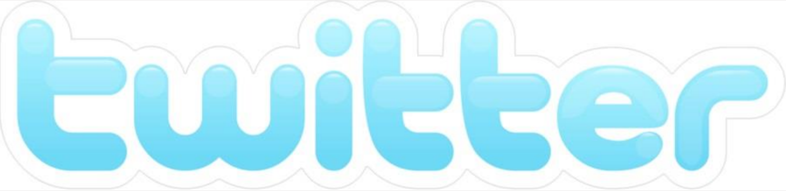
Sensor technology and networks
(measuring all kinds of data)

Mobile devices
(tracking all objects all the time)

The largest social network

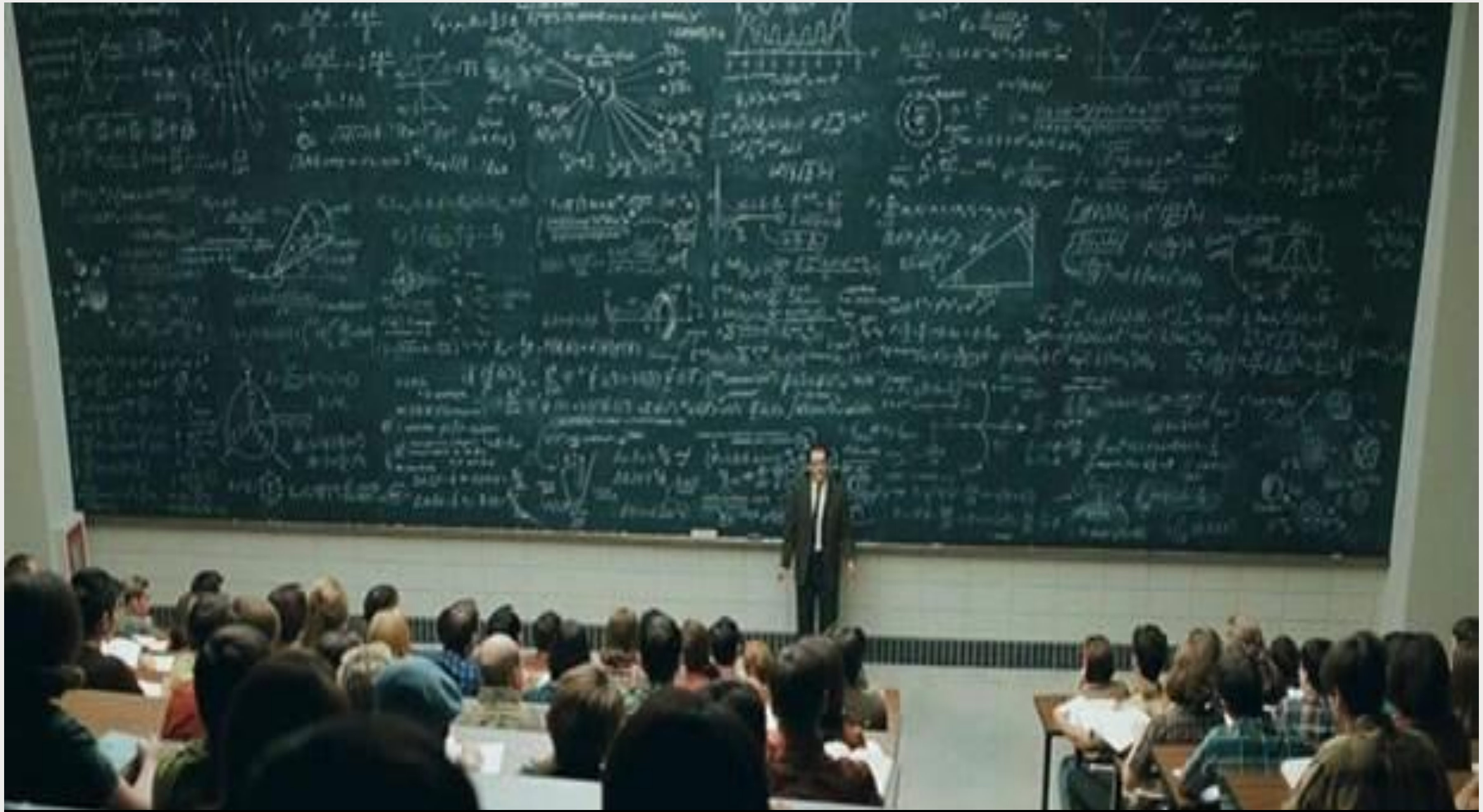
- Maintains the world‘s largest social network
  - *1110 million active users (Jan 2013)/(1.15 billion(March13)*
  - *69 billion friendship links, 2.7 billion likes per day.*
  - *Daily about 0.5 Petabytes of updates are being made into FACEBOOK including 40 millions photos.*
- Challenges
  - *Provide realtime updates of friends activities*
  - *suggest new friends (link prediction)*
  - *display content-related ads*

The world‘s fastest news medium



- Real-time communication via short messages
  - *2009: 2 million tweets per day*
  - *2010: 65 million tweets per day*
  - *2011: 200 million tweets per day*
  - *2012: 340 million tweets per day*
  - *2013:400 million tweets per day*
- Challenges
  - *Allow search in (near) realtime*
  - *Recommend interesting people (link prediction)*
  - *Find topics in the messages.*

What is Big Data?



“A massive volume of both structured and unstructured data that is so large & complex it's difficult to process with traditional database management tools & software techniques.”

Big Data: A Definition

- Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools.

- The challenges include capture, cleaning, storage, search, sharing, analysis, and visualization.

- Spot business trends, determine quality of research, prevent diseases, link legal citations, fight with crime, and determine real-time roadway traffic conditions.

What is Big Data?

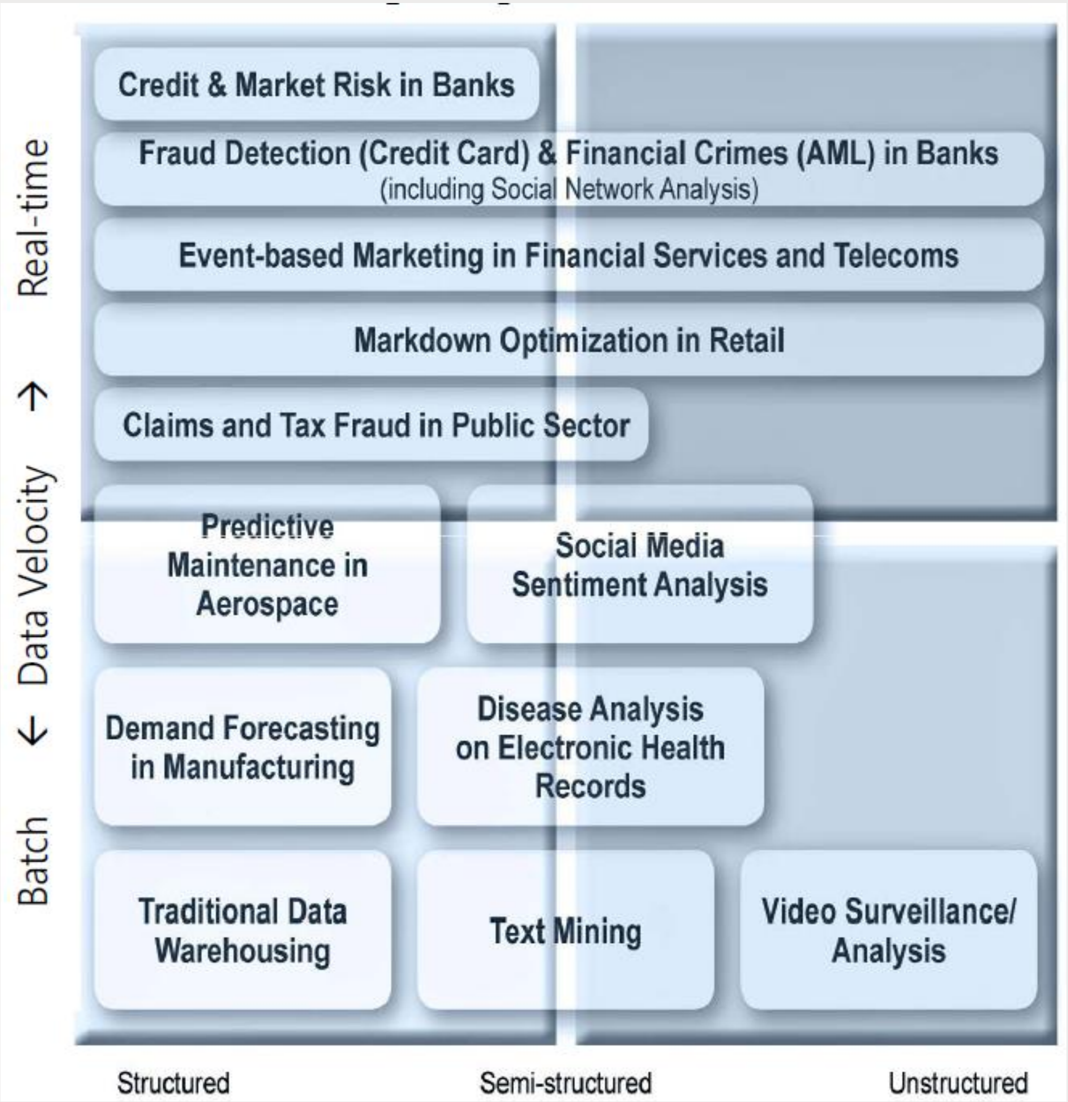
- Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day to day basis.

- But it’s not the amount of data that’s important. It’s what the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.
- While the term ‘big data is relatively new the act of gathering and storing large amounts of information for eventual analysis is ages old.

- “Big data simply means the huge volume of data, which is a collection of structured and unstructured data that cannot be processed using traditional computing techniques.”

The Structure of Big Data

- Structured
  - *Most traditional data sources*
- Semi-structured
  - *Many sources of big data*
- Unstructured
  - *Video data, audio data*



Real-time ← Data Velocity → Batch

Credit & Market Risk in Banks

Fraud Detection (Credit Card) & Financial Crimes (AML) in Banks (including Social Network Analysis)

Event-based Marketing in Financial Services and Telecoms

Markdown Optimization in Retail

Claims and Tax Fraud in Public Sector

Predictive Maintenance in Aerospace

Social Media Sentiment Analysis

Demand Forecasting in Manufacturing

Disease Analysis on Electronic Health Records

Traditional Data Warehousing

Text Mining

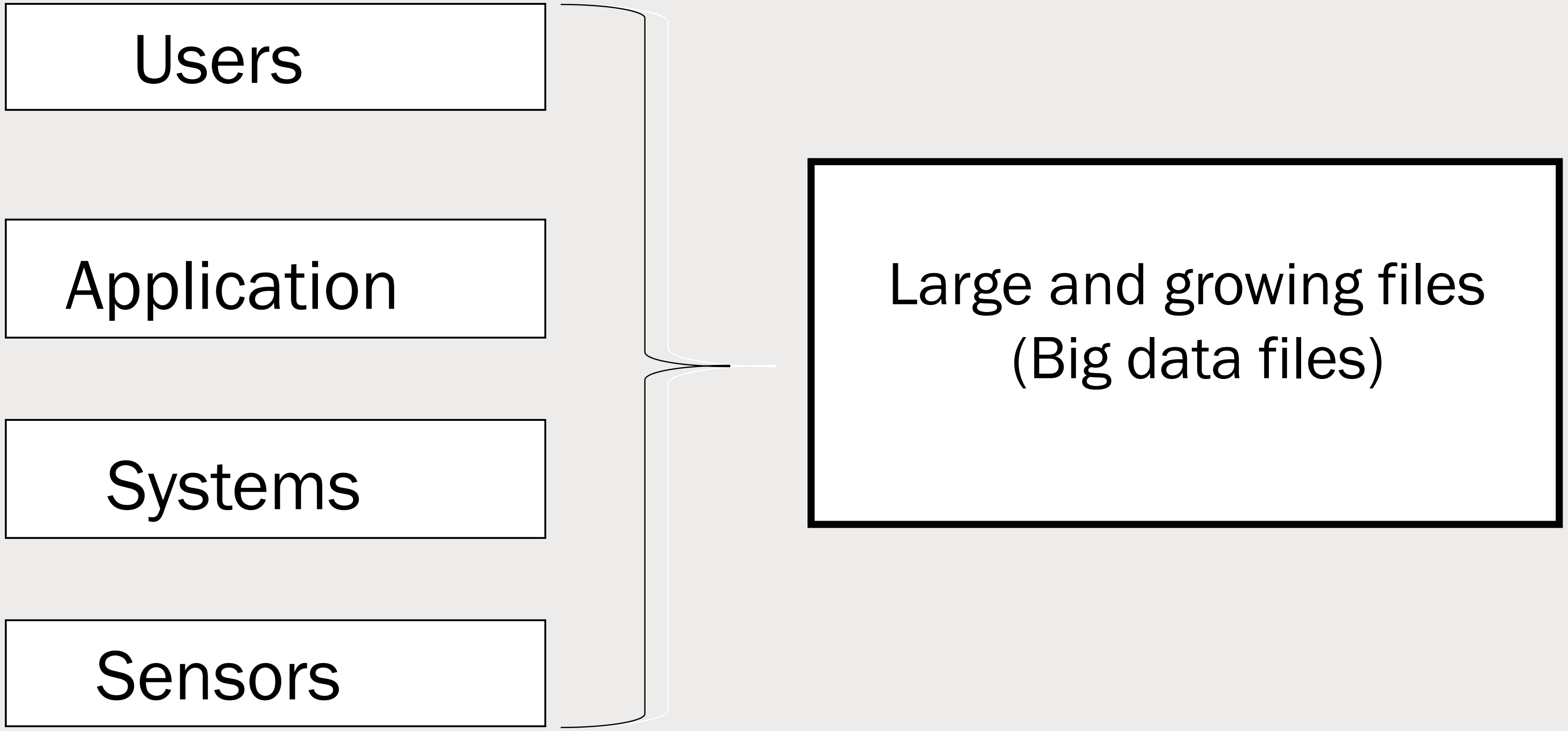
Video Surveillance/ Analysis

Structured

Semi-structured

Unstructured

Big Data sources



Types of tools used in Big-Data

- Where processing is hosted?
  - *Distributed Servers / Cloud (e.g. Amazon EC2)*
- Where data is stored?
  - *Distributed Storage (e.g. Amazon S3)*
- What is the programming model?
  - *Distributed Processing (e.g. MapReduce)*
- How data is stored & indexed?
  - *High-performance schema-free databases (e.g. MongoDB)*
- What operations are performed on data?
  - *Analytic / Semantic Processing*

Application Of Big Data analytics

Smarter Healthcare

Homeland Security

Traffic Control

Manufacturing

Multi-channel sales

Telecom

Trading Analytics

Search Quality

Risks of Big Data

- Will be so overwhelmed
  - *Need the right people and solve the right problems*

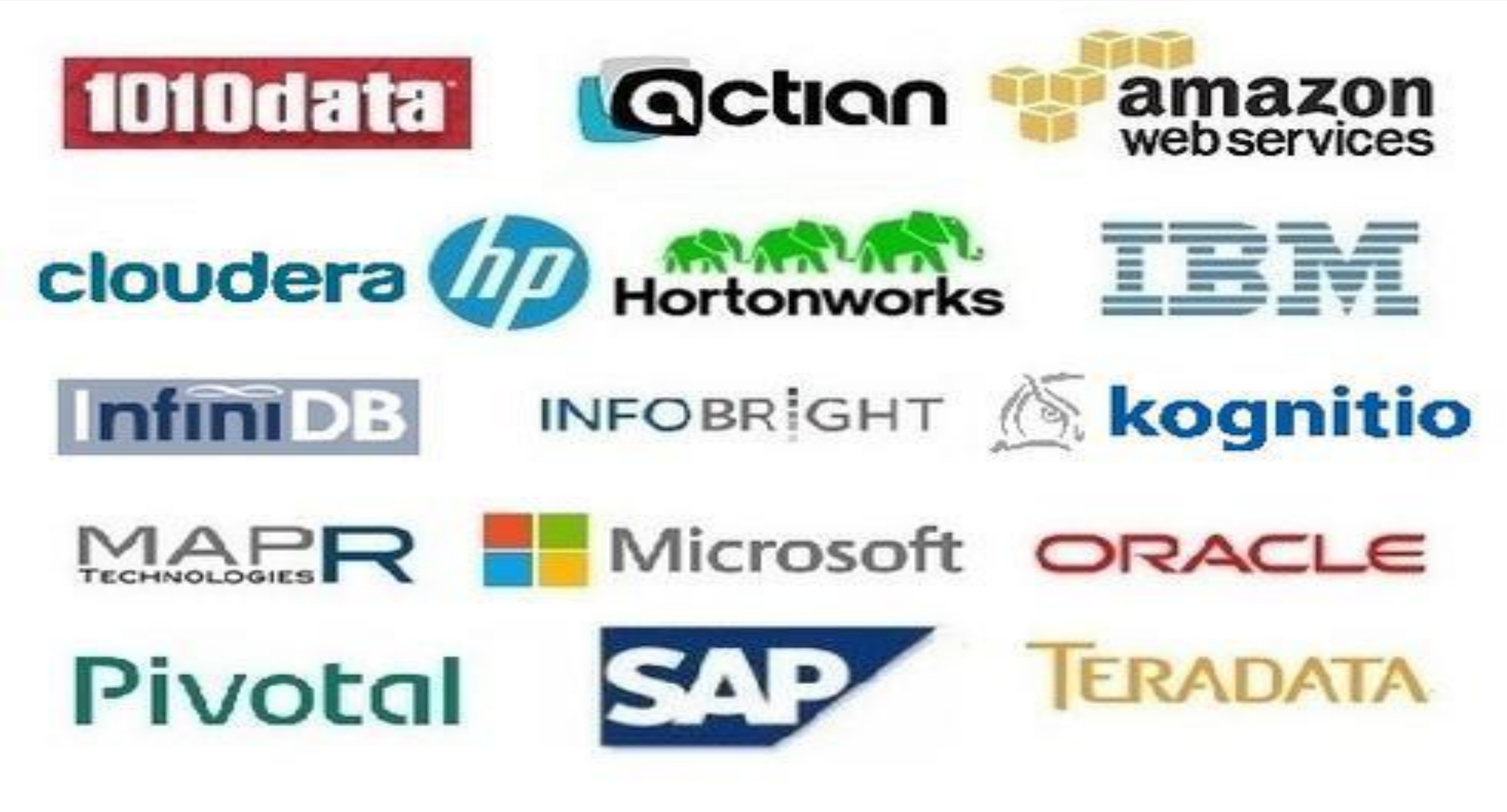
- Costs escalate too fast
  - *Isn’t necessary to capture 100%*

- Many sources of big data

  is privacy
  - *self-regulation*
  - *Legal regulation*

웹사이트 방문기록
SNS 기록
검색 통계
좋아하는 음식
위치 정보
ID
문자 메시지
구매 기록

Big Data Analytics Tools



CHAPTER 1

Introduction to Big Data

Topic Outlines

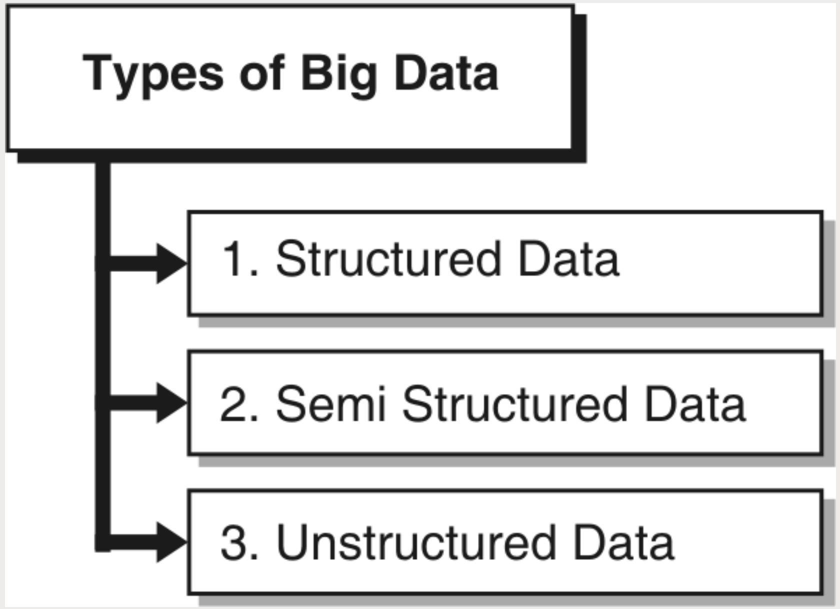
- Introduction
  - *What is Big Data?*
  - *Types of Big Data*
- Distributed File System
  - *HDFS Architecture*
  - *Goals of HDFS*
  - *Commands for HDFS*
- *Big Data and Its Importance*
  - *Advantages of Big Data*
  - *Use of big data across different sectors*
- *Four V's*
- *Drivers for Big Data*
- *Big Data Analytics*
- *Big Data Applications*
- *Algorithms Using Map Reduce*

What is Big Data?

- Big data simply means the huge volume of data, which is a collection of structured and unstructured data that cannot be processed using traditional computing techniques.

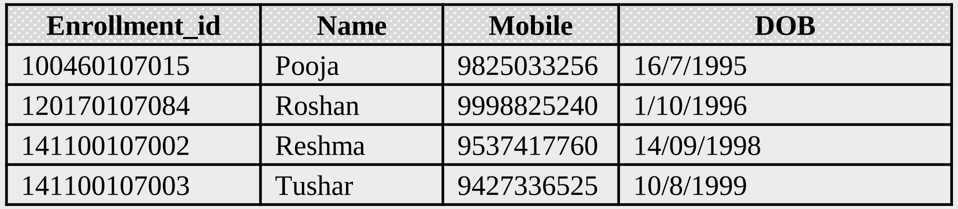
Types of Big Data

- Big data is not only a data, rather it has become a complete substance, which involves various tools, techniques and frameworks.
- Big Data includes immense volume, high velocity, and extensible variety of data.
- The data in it will be of three types.



Structured Data

- Structured data refers to any data that resides in a fixed field within a record or file. It includes data contained in relational databases and spreadsheets.
- Example of structured data can be any kind of relational database created in Main frame, SQL server, Sybase, DB2, Oracle, Excel, Access, Terradata, Neeteza or any other.
- Example of Structured Data



Enrollment_id	Name	Mobile	DOB
100460107015	Pooja	9825033256	16/7/1995
120170107084	Roshan	9998825240	1/10/1996
141100107002	Reshma	9537417760	14/09/1998
141100107003	Tushar	9427336525	10/8/1999

Semi Structured Data

- Semi-structured data 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, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data.
- The advantages of this model is it can represent the information of some data sources that cannot be affected by schema.
- Example of semi structured data can be any XML data.

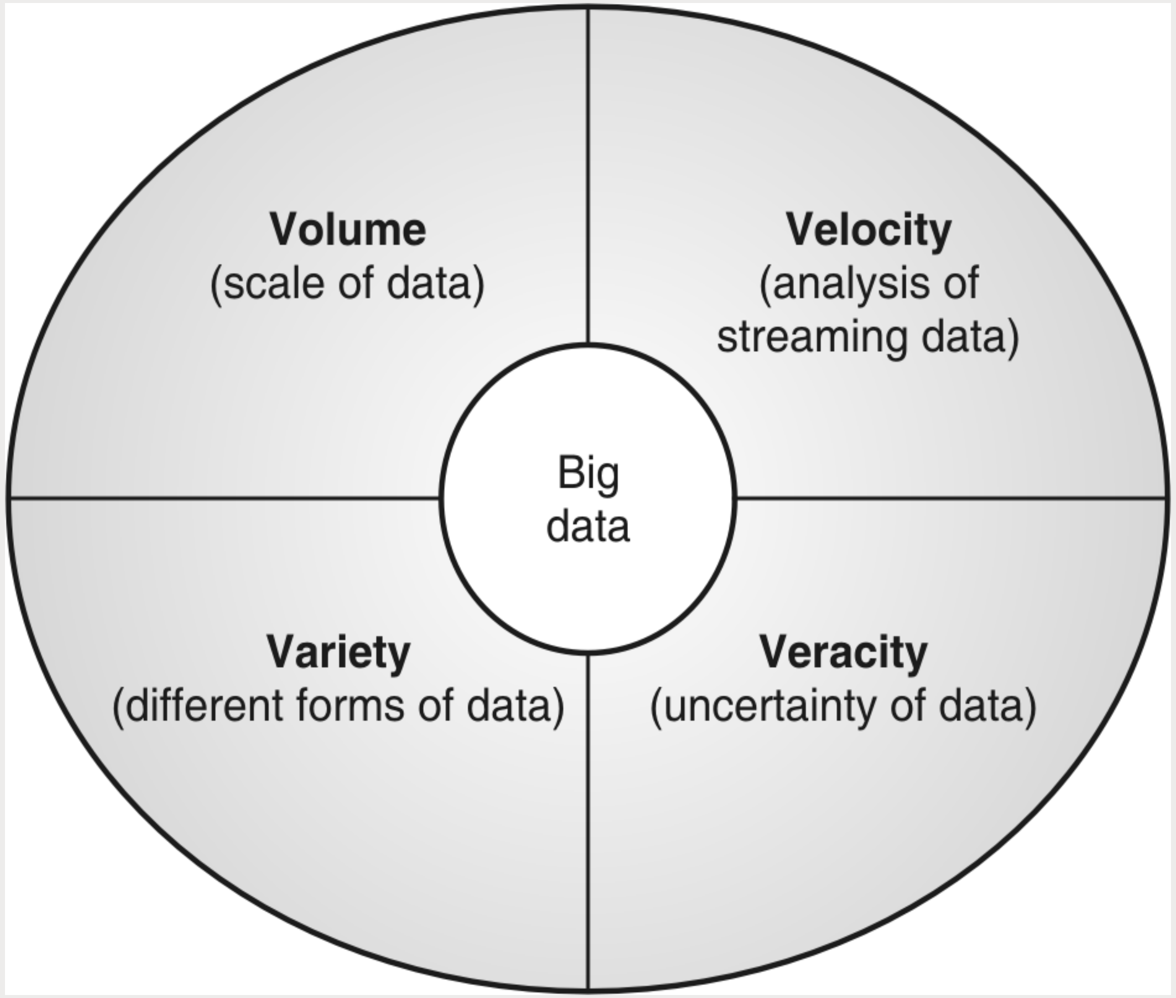
```
<?xml version='1.0'?>
<catalog>
<book id ="bk201">
<author>Rafael C. Gonzalez</author>
<title>Digital Image Processing</title>
<genre>Computer</genre>
<price>719</price>
</book>
```

Unstructured Data

- Unstructured data refers to information that either does not have a pre-defined data model or is not organized in a pre-schema defined manner.
- Unstructured data files often include text and multimedia data. Examples include videos, photos, e-mail messages, word processing documents, audio files, presentations, web pages and many other kinds of business documents.
- Experts estimate that in any organization, 80 to 90 percent of the data is unstructured. The amount of unstructured data in enterprises is growing than structured databases.
- Some examples of human-generated unstructured data are text internal to your company, Social media data, Mobile data, website content,

```
CREATE EXTERNAL TABLE IF NOT EXISTS access_log (log_line STRING)

PARTITIONED BY (hive_entry_timestamp STRING)

ROW FORMAT DELIMITED

   FIELDS TERMINATED BY ','

FIELDS TERMINATED BY '01'

STORED AS TEXTFILE

LOCATION '/usr/local/demo/access_logs';
```

Four V's



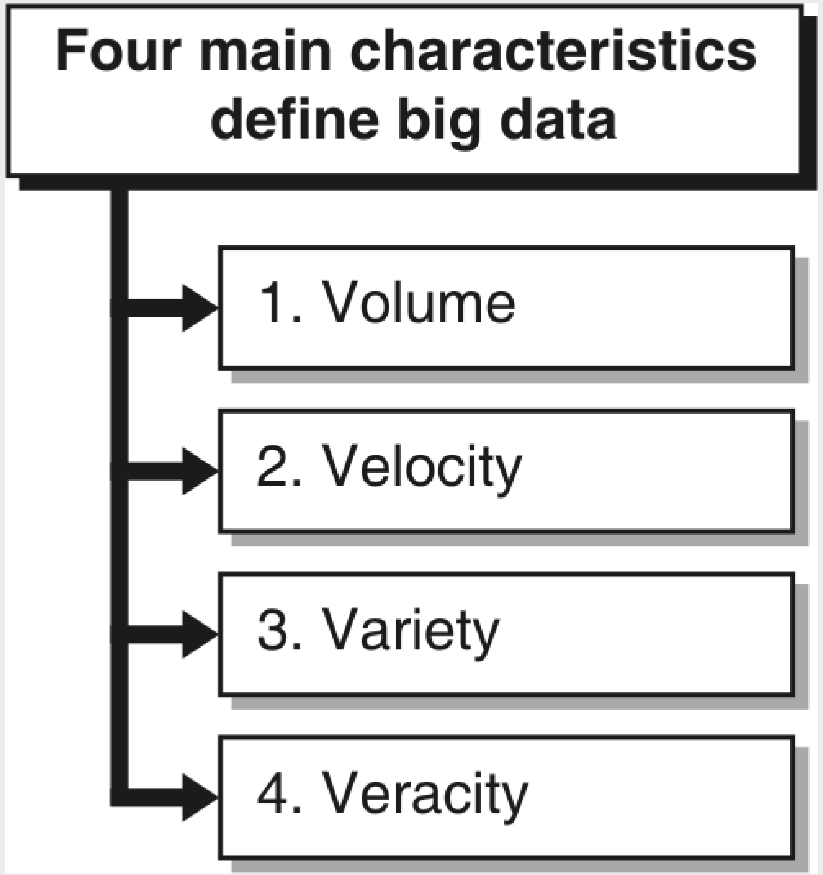
Volume
(scale of data)

Velocity
(analysis of streaming data)

Big data

Variety
(different forms of data)

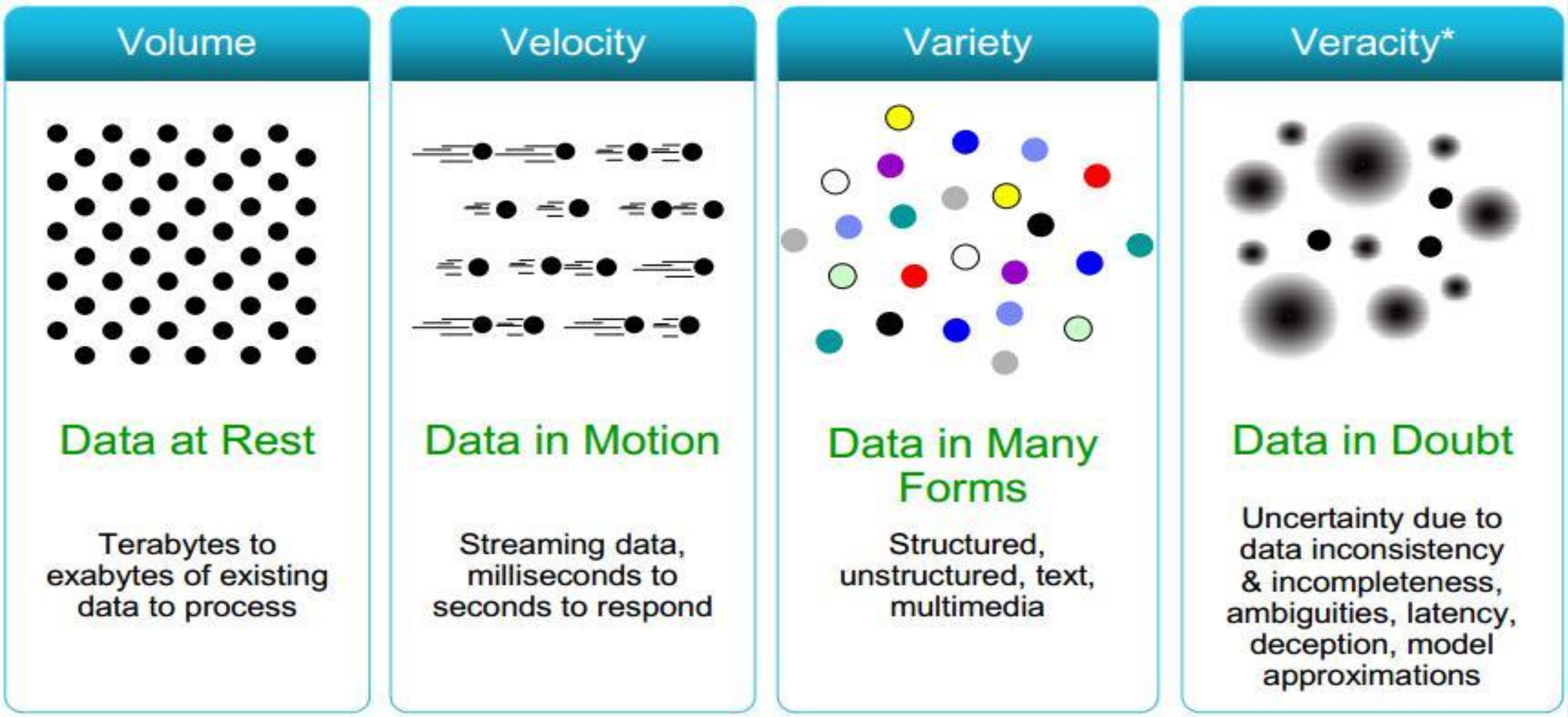
Veracity
(uncertainty of data)



Four main characteristics define big data

1. Volume
2. Velocity
3. Variety
4. Veracity

Four V's



Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*

Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Four main Characteristics

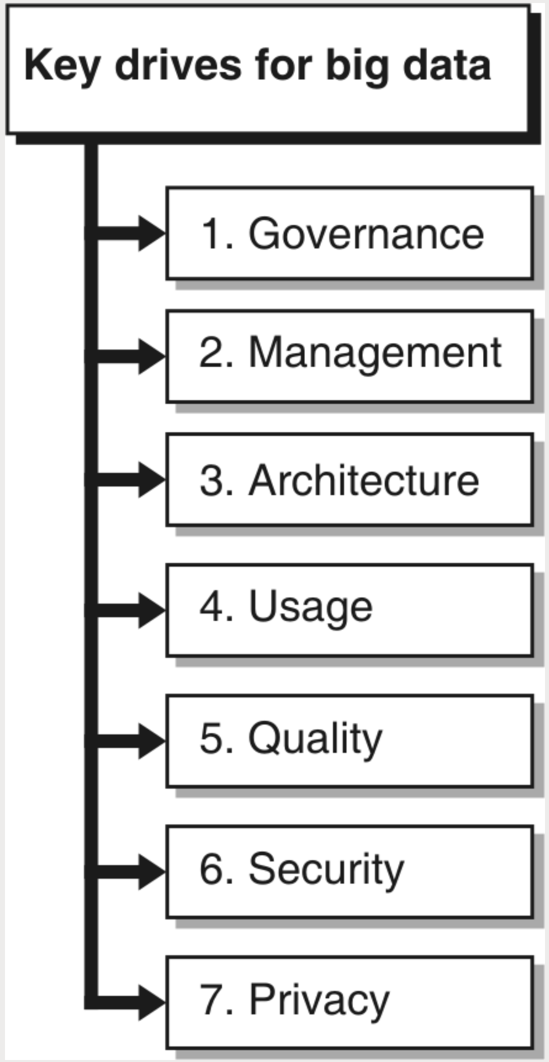
- Volume
  - *There is exponential growth in the data storage as the data is now becoming more than only text data. Most of the data found today is in the* ***format of videos, music and large images on social media channels****. The main characteristic that makes data "big" is its volume.*
  - ***Volume*** *refers to the vast amounts of data that is generated every second. It is very common to have Terabytes and Petabytes of the storage system for an organization.*
- Velocity
  - ***Velocity*** *refers to the speed at which new data is generated and the speed at which it moves around. Velocity is the frequency of incoming data that needs to be processed.*
  - *Just think about how many likes, SMS messages, Facebook status updates, or credit card swipes are being sent on a particular telecom carrier every minute of every day, and you will have a good idea of velocity.*
  - *A streaming application like Amazon Web Services and Kinesis and many more are example of an application that handles the velocity of data.*

Four main Characteristics (Cont.....)

- Variety
  - *Data generated from an organization are not only text data, but it includes audio file, video file, etc. This simply means that it refers to the different forms of data that we collect and use.*
  - ***Structured data** can be defined as a set of rules. For example, Name which can be of text type, mobile number will always be numbers; and date follow a specific pattern of date.*
  - ***unstructured data**, there are no rules. An image, a voice recording, a tweet or Facebook posts , they all can be of different type but express ideas and thoughts based on human understanding. One of the goals of big data is to use technology to take this unstructured data and make sense of it for better utilization.*
- *Veracity*
  - *Most of the company is losing a big amount in a year due to poor data management. Veracity refers to the uncertainty surrounding data, which is due to data inconsistency and incompleteness, which leads to another challenge, keeping big data organized.*
  - *Veracity refers to the biases, noise and abnormality in big data. It is used for data that is being stored, and mined meaningfully to the problem being analysed. In compare to volume and velocity, data analysis in Veracity is the biggest challenge.*

Drivers for Big Data

- Key drivers behind the big data market are listed below



Key drives for big data

1. Governance

2. Management

3. Architecture

4. Usage

5. Quality

6. Security

7. Privacy

Governance

- Good governance focuses on consistent guidance, procedures and clear management decision-making. Organizations need to ensure standard and exhaustive data capture, they need not protect all the data, but they need to start sharing data with in-built protections with the right levels and functions of the organization.

Management

- Integrating and moving data across the organization is traditionally forced by data storage platforms such as relational databases or batch files with partial ability to process huge amount of data, data with difficult structure or without structure at all, or data generated or received at very high speeds.

Architecture

- Data architecture should be prepared in such a manner that it can break down internal storage by enabling the sharing of key data sets across the organization. It should also ensure that knowledge are being captured and relayed across to the correct set of people in the organization in a timely and accurate manner.

Usage

- Big data can be beneficial to a wide range of users across the organization . Executive management and boards, business operations and risk professionals, including legal, internal auditor can use the big data to analyse the facts. Not even this persons but it is also used in finance and compliance as well as customer-facing departments like sales and marketing. The key challenge is to have the ability to interpret the big amount of data that can be collated from various sources.

Quality

- The quality of data sets and the assumption drawn from such data sets are increasingly becoming more crucial. Organizations need to build quality and monitoring functions and parameters for big data. It is obvious that correcting a data error can be much more costly than getting the data right in the first time. Wrong data can be disastrous and much more costly to the organization if not corrected.

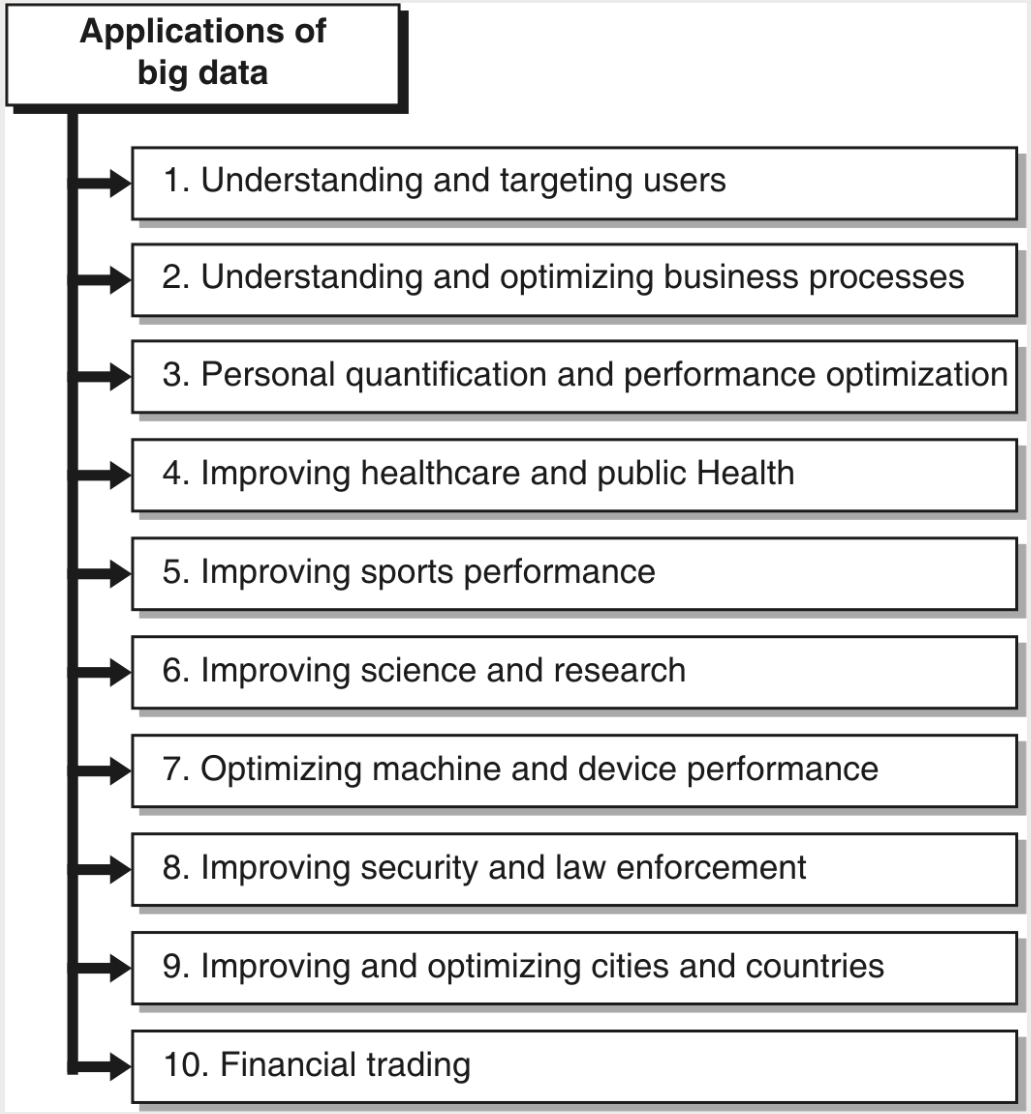
Security

- It will be definitely good for companies to start establishing security policies which are self-configurable. These policies must provide reliable relationships, and support data and resource sharing within the organizations, while ensuring that data analytics are optimized and not affected by such policies.

Privacy

- The increased use of big data also comes with big challenge like privacy protection. The traditional frameworks for protecting the privacy of personal information, forcing companies to audit the implementation of their privacy policies to ensure that privacy is being properly maintained in big data also.

Big Data Applications



Applications of big data

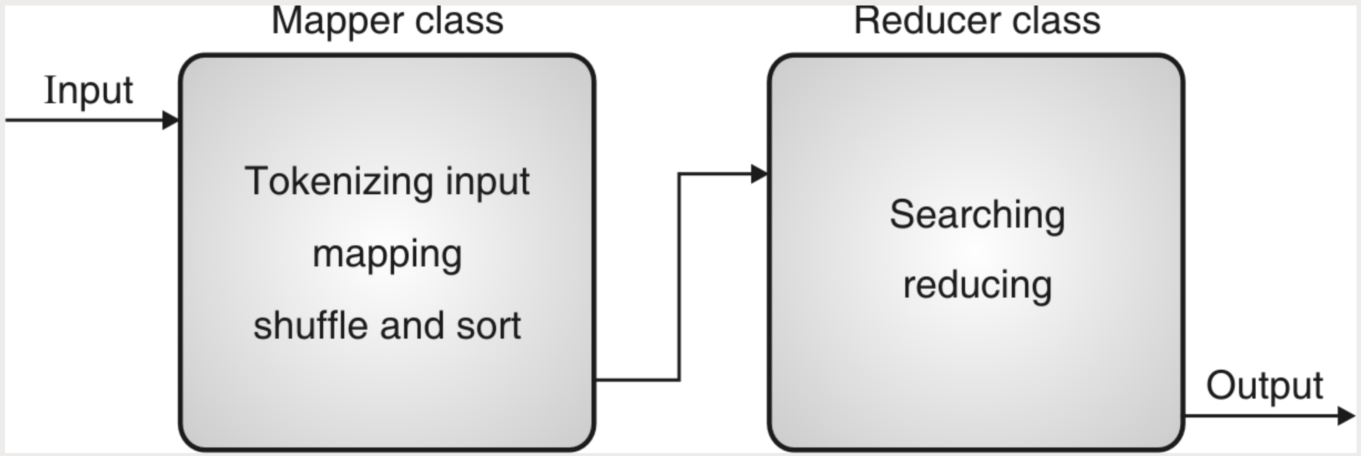
1. Understanding and targeting users
2. Understanding and optimizing business processes
3. Personal quantification and performance optimization
4. Improving healthcare and public Health
5. Improving sports performance
6. Improving science and research
7. Optimizing machine and device performance
8. Improving security and law enforcement
9. Improving and optimizing cities and countries
10. Financial trading

Algorithms Using Map Reduce

- There is a rising trend of applications that should handle big data. However, analyzing big data is a very difficult problem today. For such applications, the MapReduce framework has recently attracted a lot of attention.
- MapReduce or its open-source equivalent Hadoop is a powerful tool for building such applications.
- In the MapReduce framework, a distributed file system initially partitions data in multiple machines. Data is represented as key, value pairs. The computation is carried out using two user defined functions : map and reduce functions.
- The map function defined by a user is called on different partitions of input data in parallel. The key-value pairs output by each map function are next grouped and merged by each distinct key.
- Finally, a reduce function is invoked for each distinct key with the list of all values.

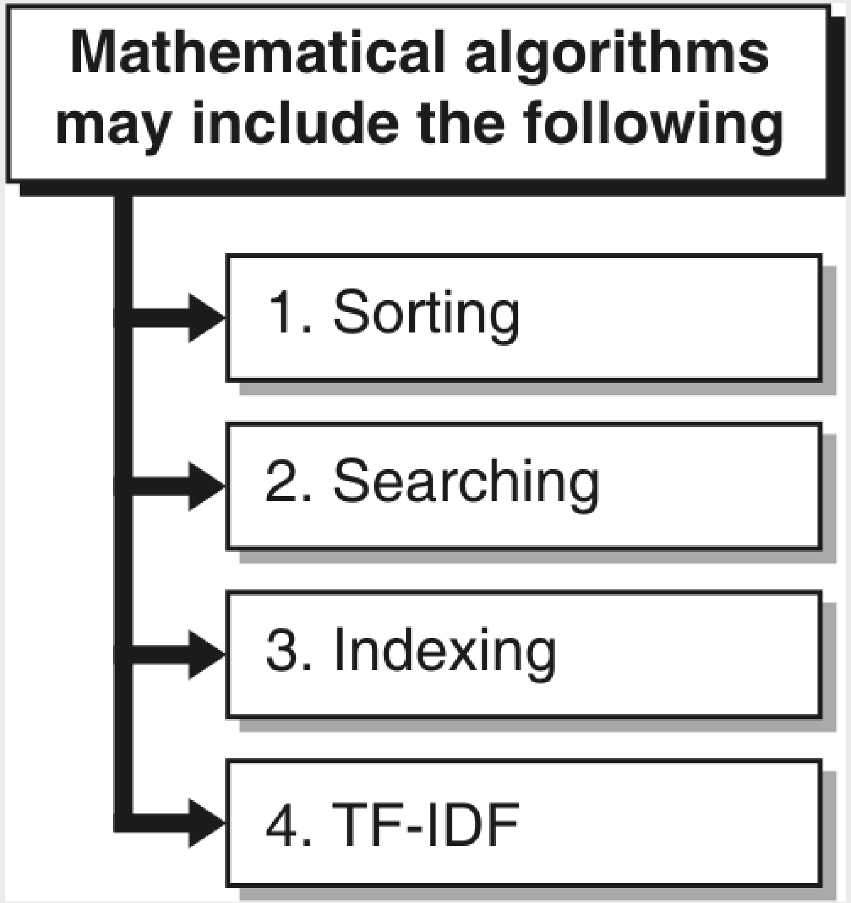
Map Reduce

- The MapReduce algorithm contains two important tasks, namely map and reduce.



- MapReduce implements various mathematical algorithms to divide a task into small parts and assign them to multiple systems. In technical terms, MapReduce algorithm helps in sending the Map and Reduce tasks to appropriate servers in a cluster.

MapReduce algorithm



Mathematical algorithms may include the following

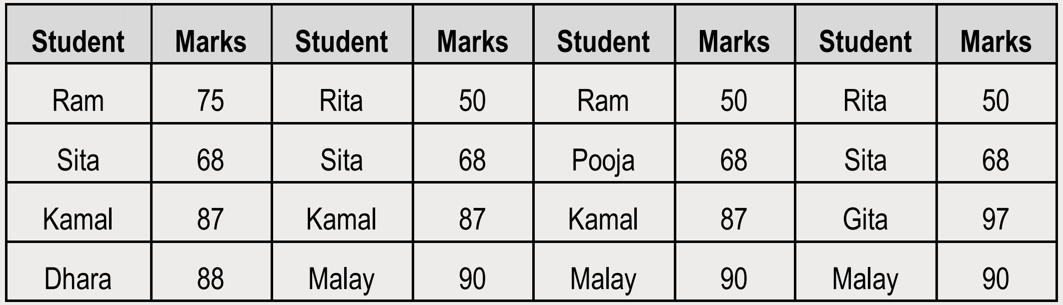
1. Sorting
2. Searching
3. Indexing
4. TF-IDF

Sorting

- It is used to process and analyze data.
- Sorting methods are implemented in the mapper class itself.
- To collect similar key-value pairs, the mapper class takes the help of RawComparator class to sort the key-value pairs.
- after tokenizing the values in the mapper class, the Context class collects the matching valued keys as a collection.
- The set of intermediate key-value pairs for a given Reducer is automatically sorted by Hadoop to form key-values (K2, {V2, V2, …}) before they are presented to the reducer.

Searching

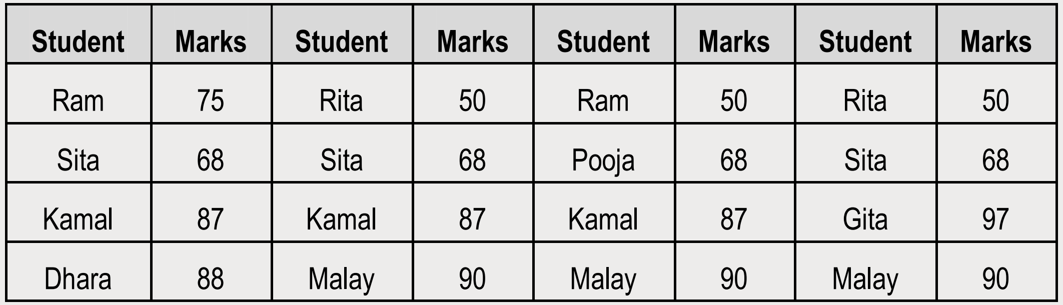
- Searching plays an vital role in MapReduce algorithm. It helps in the combiner and reducer phase. Let us try to understand how searching works with the help of an example.
- Example
- **Let us assume we have student data in four different files** : A, B, C, and D. Let us also assume there are duplicate student records in all four files because of importing the student data from all database tables repeatedly. See the following illustration.



Student	Marks	Student	Marks	Student	Marks	Student	Marks
Ram	75	Rita	50	Ram	50	Rita	50
Sita	68	Sita	68	Pooja	68	Sita	68
Kamal	87	Kamal	87	Kamal	87	Gita	97
Dhara	88	Malay	90	Malay	90	Malay	90

Searching

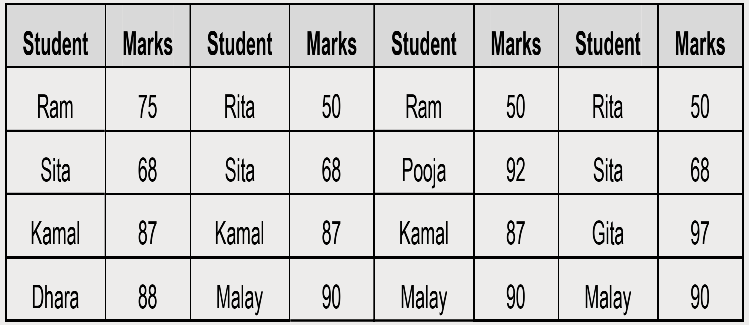
- Searching plays an vital role in MapReduce algorithm. It helps in the combiner and reducer phase. Let us try to understand how searching works with the help of an example.
- Example
- **Let us assume we have student data in four different files** : A, B, C, and D. Let us also assume there are duplicate student records in all four files because of importing the student data from all database tables repeatedly. See the following illustration.



Student	Marks	Student	Marks	Student	Marks	Student	Marks
Ram	75	Rita	50	Ram	50	Rita	50
Sita	68	Sita	68	Pooja	68	Sita	68
Kamal	87	Kamal	87	Kamal	87	Gita	97
Dhara	88	Malay	90	Malay	90	Malay	90

Searching

- **The Map phase** processes each input file and provides the employee data in key-value pairs



Student	Marks	Student	Marks	Student	Marks	Student	Marks
Ram	75	Rita	50	Ram	50	Rita	50
Sita	68	Sita	68	Pooja	92	Sita	68
Kamal	87	Kamal	87	Kamal	87	Gita	97
Dhara	88	Malay	90	Malay	90	Malay	90