Introduction to Big Data

ECAP456

Dr. Rajni Bhalla

Associate Professor

Learning Outcomes

After this lecture, you will be able to

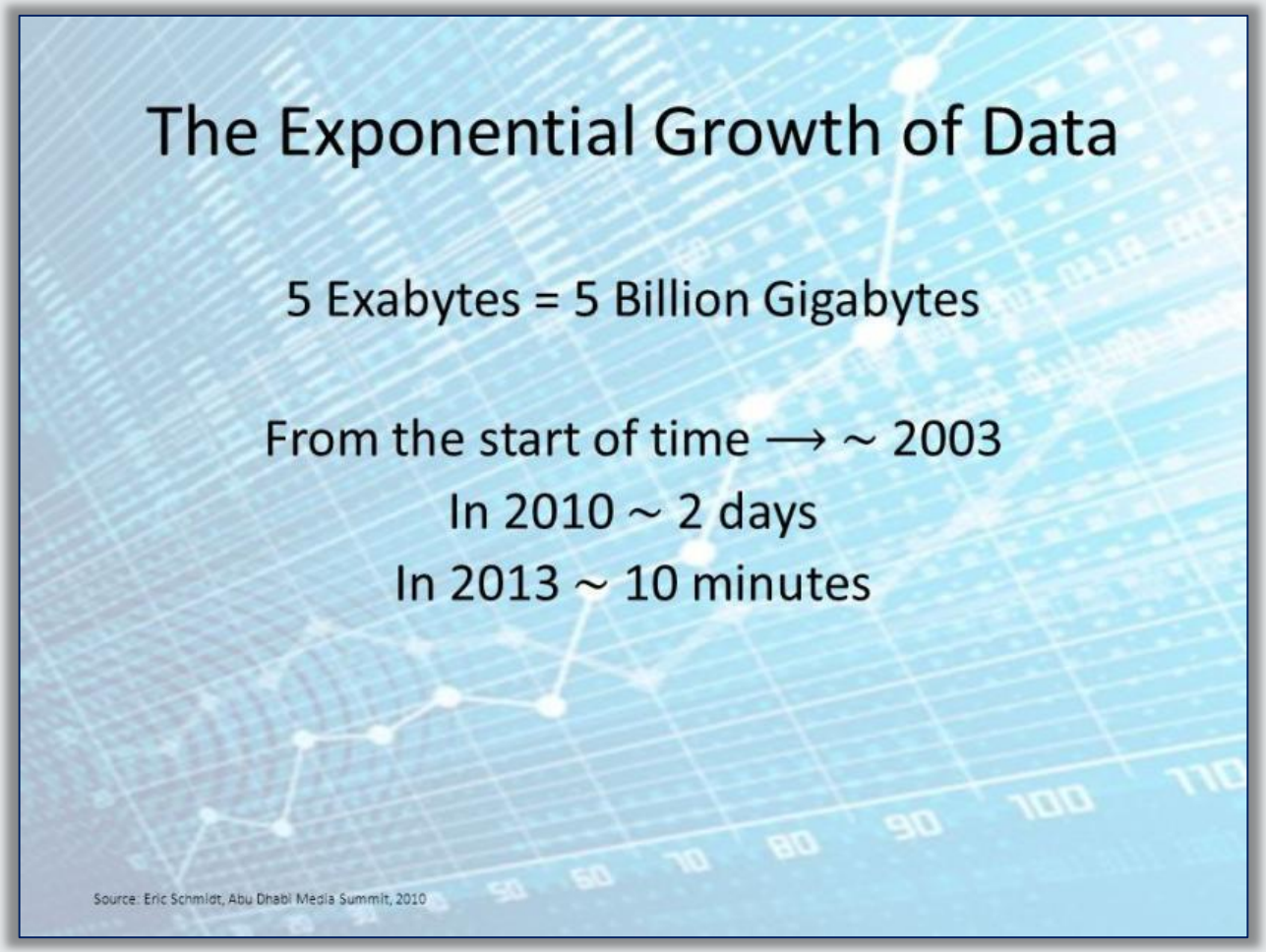
- hadoop - Big Data Overview
- hadoop - Big Data Solutions

Introduction

Introduction

Data produced by mankind is growing rapidly every year

Introduction



What is Big Data?

- Big data is a collection of large datasets that cannot be processed using traditional computing techniques.
- It is not a single technique or a tool, rather it has become a complete subject, which involves various tools, techniques and frameworks.

What Comes Under Big Data?

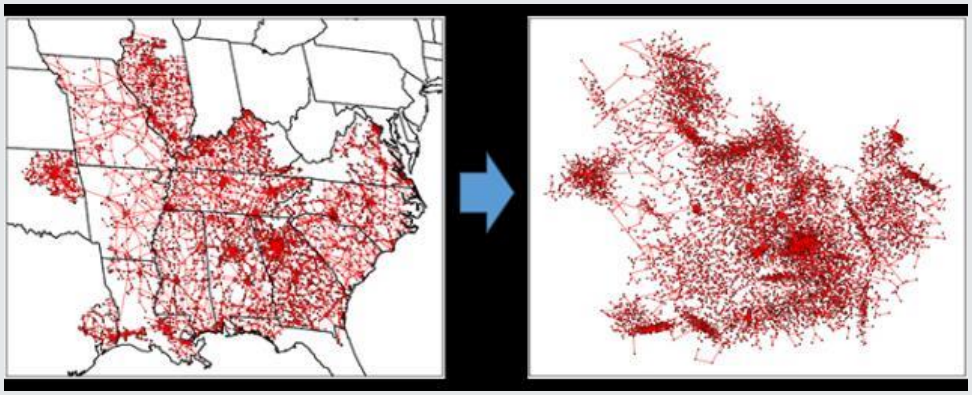
Black Box Data

Social Media Data

Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data.

What Comes Under Big Data?

Stock Exchange Data



Power Grid Data

What Comes Under Big Data?

Transport Data

Search Engine

What Comes Under Big Data?

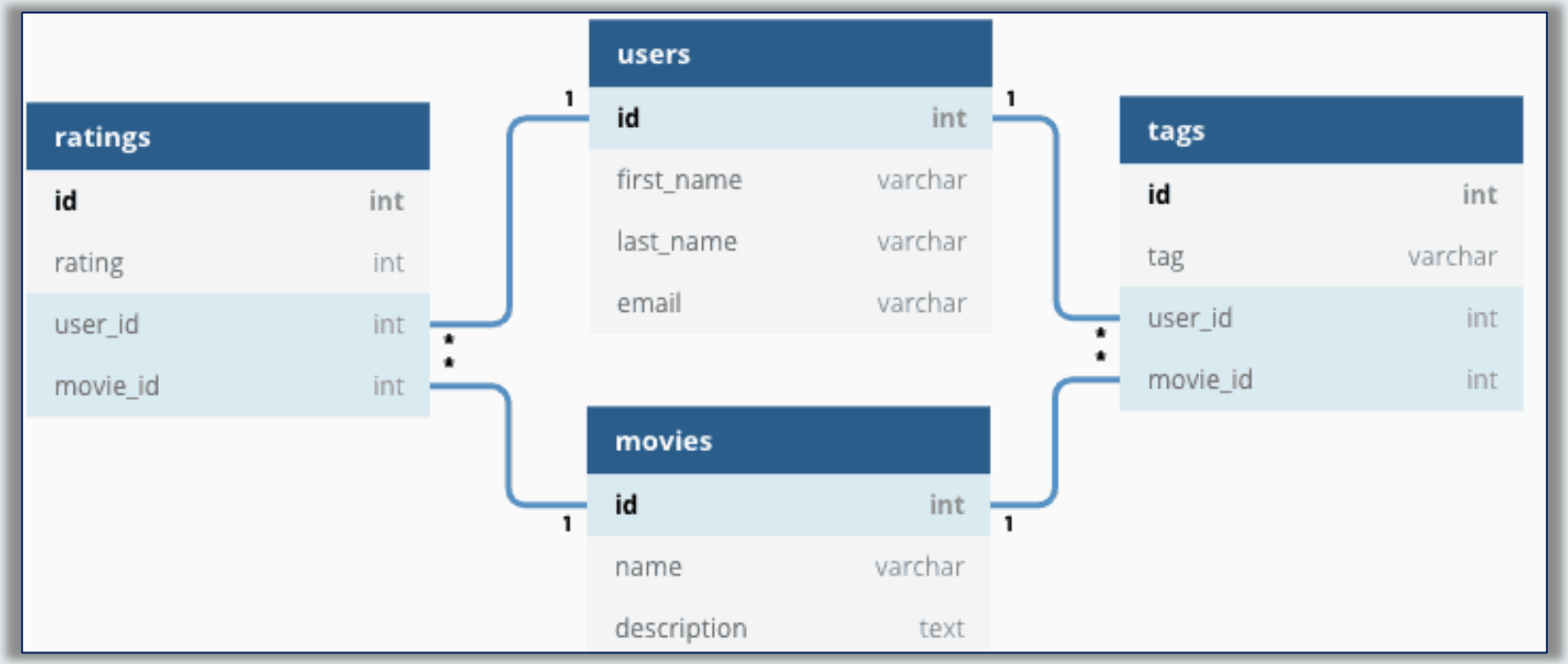
- Thus, Big Data includes huge volume, high velocity, and extensible variety of data..
  - Structured data – Relational data.
  - Semi Structured data – XML data.
  - Unstructured data – Word, PDF, Text, Media Logs.

What Comes Under Big Data?

- Structured data
- Semi Structured data
- Unstructured data

The data in it will be of three types

Structured data



Relational data

The data in it will be of three types

Semi Structured data

```
<?xml version="1.0"?>
- <ROOT>
  - <Customers>
    - <Customer CustomerName="Arshad Ali" CustomerID="C001">
      - <Orders>
        - <Order OrderDate="2012-07-04T00:00:00" OrderID="10248">
            <OrderDetail Quantity="5" ProductID="10"/>
            <OrderDetail Quantity="12" ProductID="11"/>
            <OrderDetail Quantity="10" ProductID="42"/>
          </Order>
        </Orders>
        <Address> Address line 1, 2, 3</Address>
      </Customer>
    - <Customer CustomerName="Paul Henriot" CustomerID="C002">
      - <Orders>
        - <Order OrderDate="2011-07-04T00:00:00" OrderID="10245">
            <OrderDetail Quantity="12" ProductID="11"/>
            <OrderDetail Quantity="10" ProductID="42"/>
          </Order>
        </Orders>
        <Address> Address line 5, 6, 7</Address>
      </Customer>
    - <Customer CustomerName="Carlos Gonzlez" CustomerID="C003">
      - <Orders>
        - <Order OrderDate="2012-08-16T00:00:00" OrderID="10283">
            <OrderDetail Quantity="3" ProductID="72"/>
          </Order>
        </Orders>
        <Address> Address line 1, 4, 5</Address>
      </Customer>
    </Customers>
</ROOT>
```

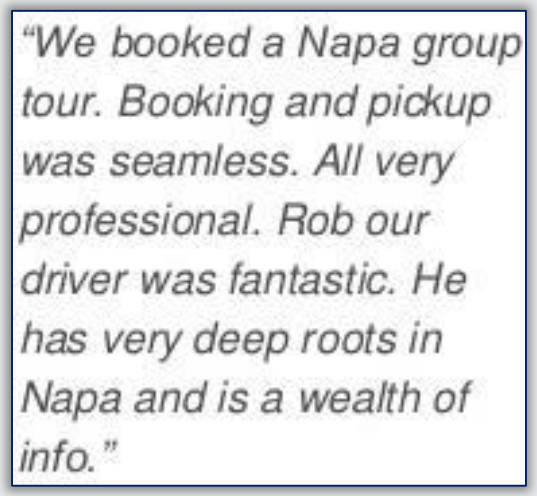
XML data

The data in it will be of three types

Unstructured data

Word

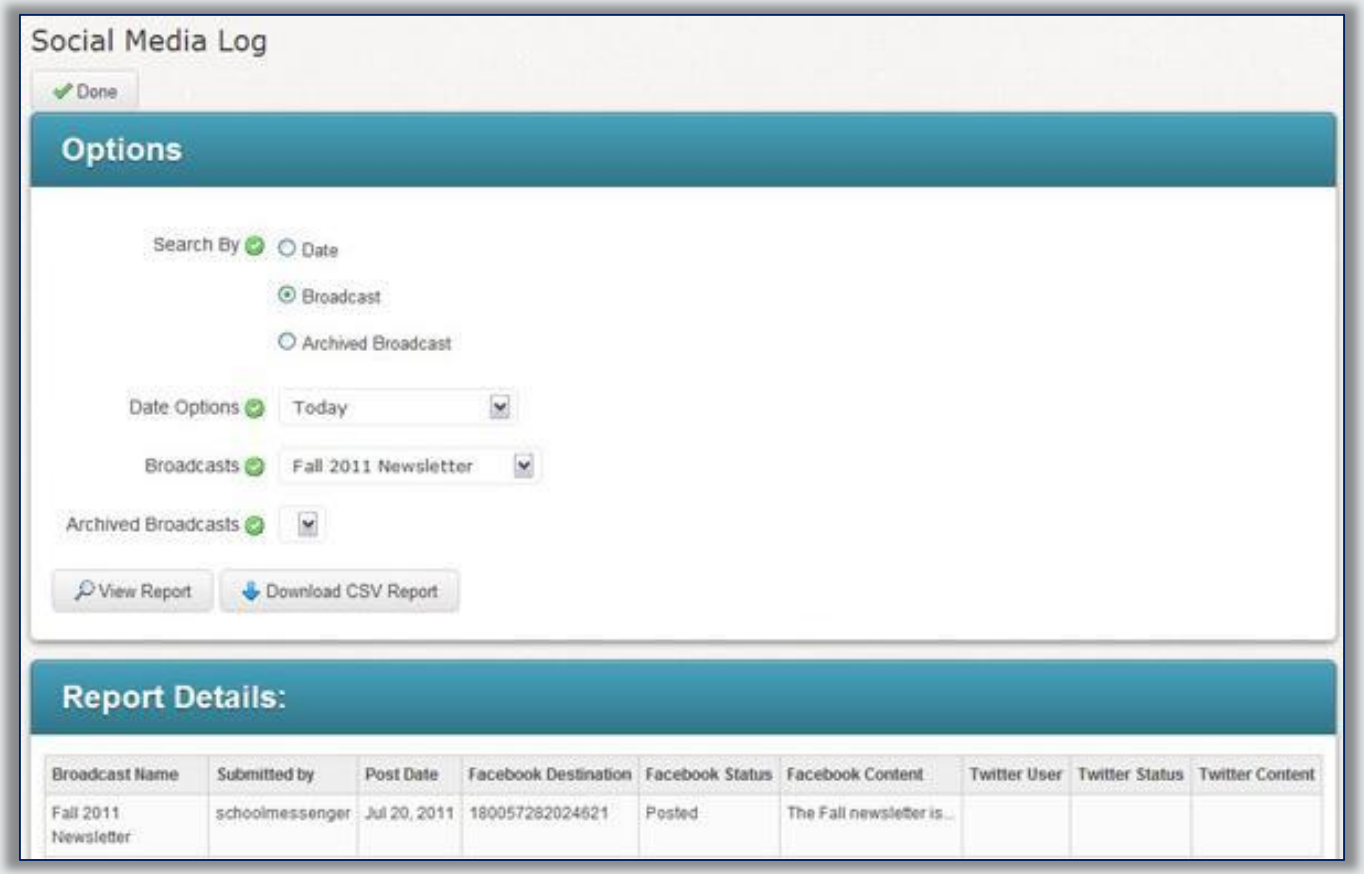
PDF



Text

The data in it will be of three types

Unstructured data



Social Media Log

Benefits of Big Data

Big Data Technologies

- Accurate analysis
- Greater operational efficiencies,
- cost reductions,
- and reduced risks for the business.

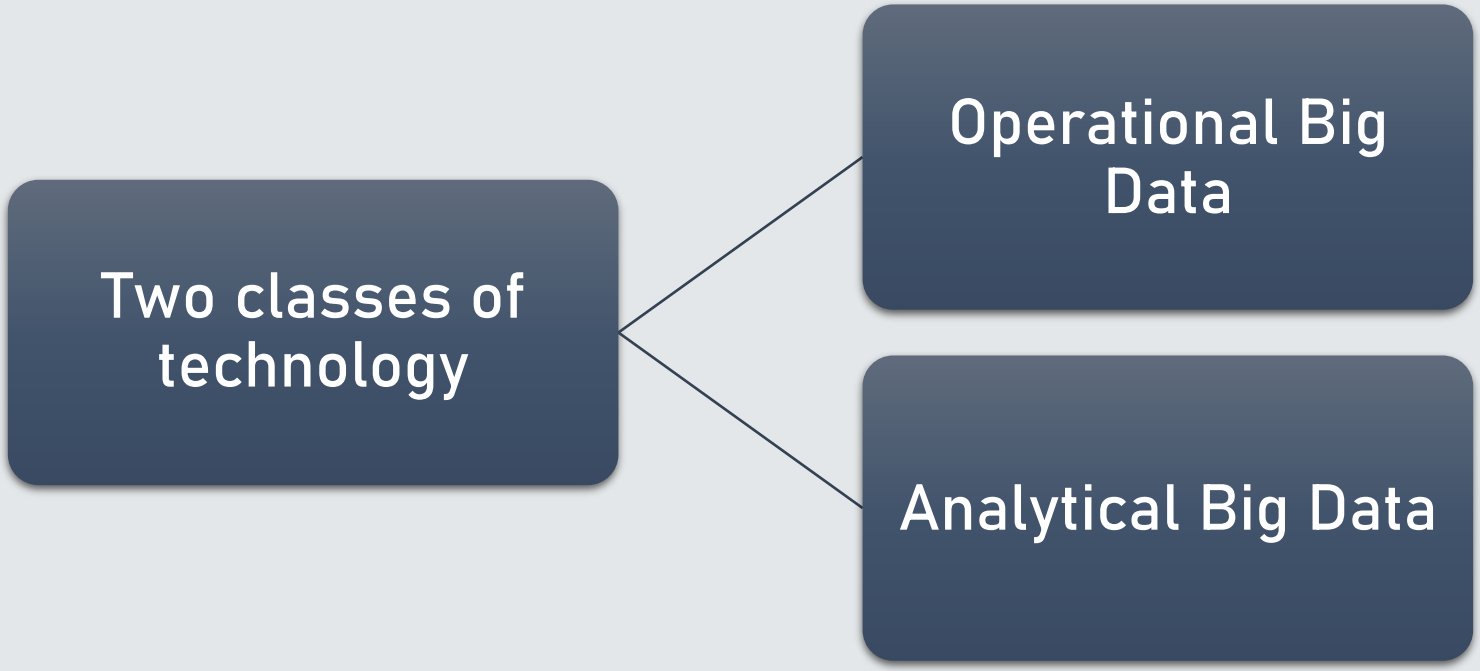
Big Data Technologies

Structure

Unstructured

Big Data Technologies

Big Data Technologies



Operational Big Data

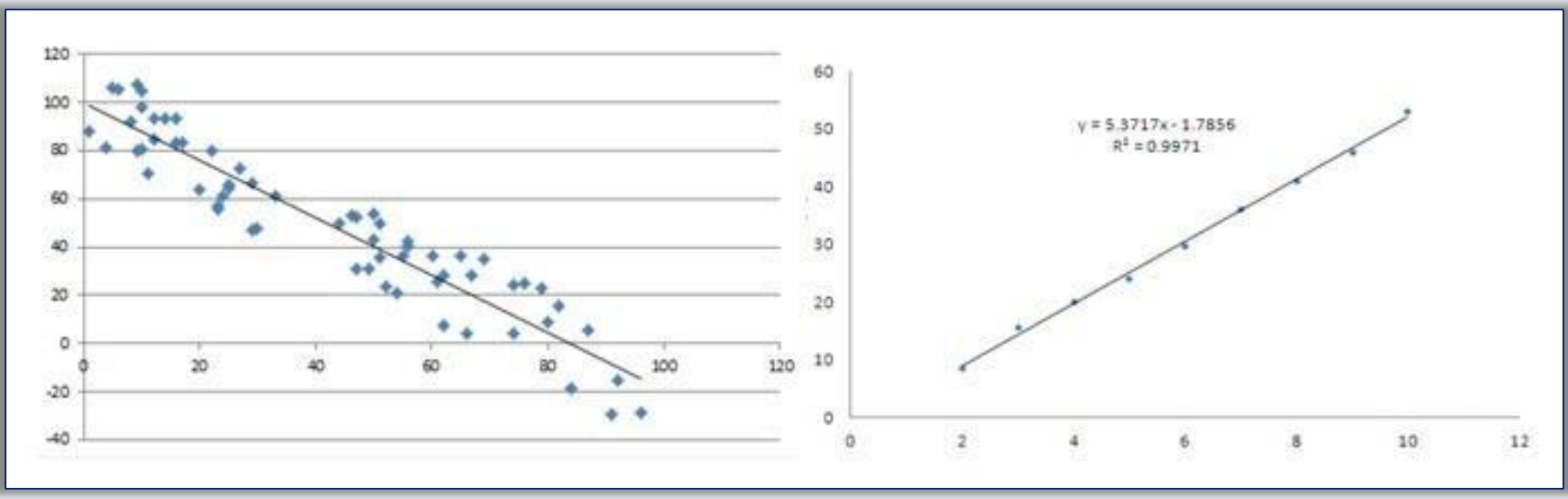
Operational Big Data

Manage

Cheaper

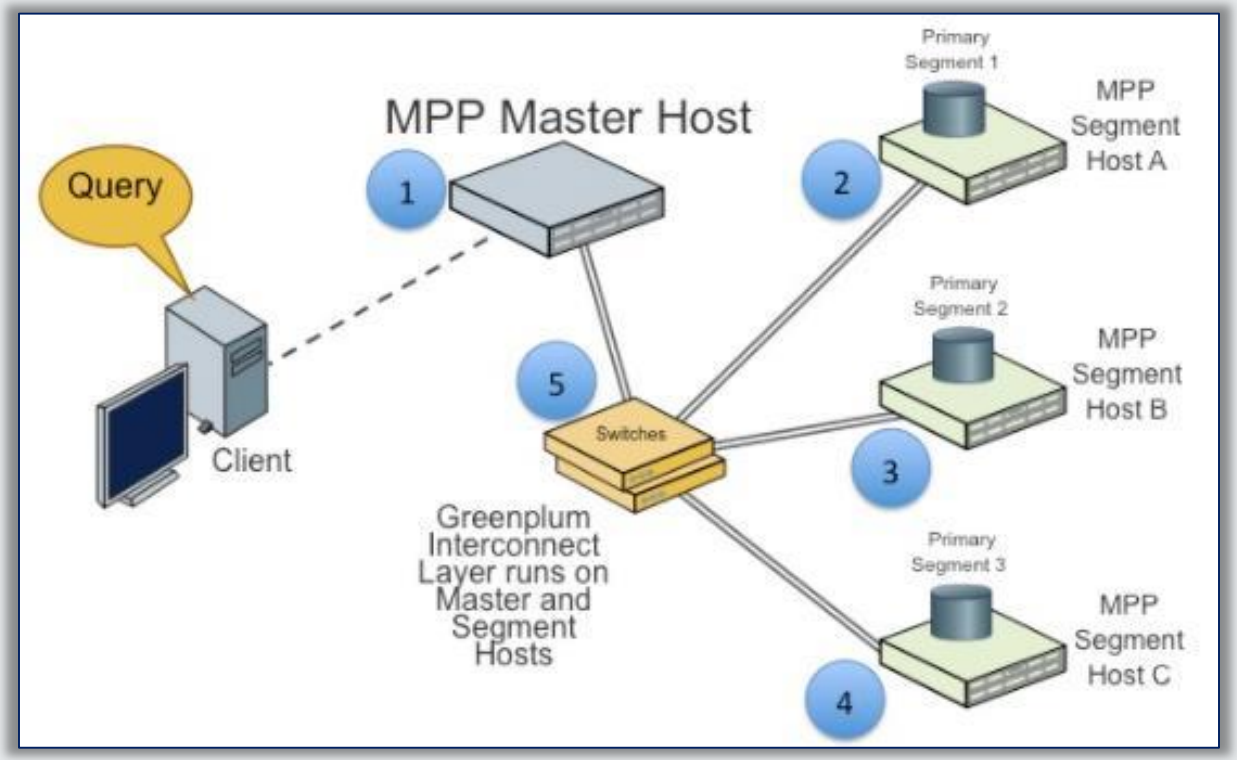
Faster

Operational Big Data



Patterns and Trends

Analytical Big Data



Massively Parallel Processing

Analytical Big Data

MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on MapReduce that can be scaled up from single servers to thousands of high- and low-end machines.

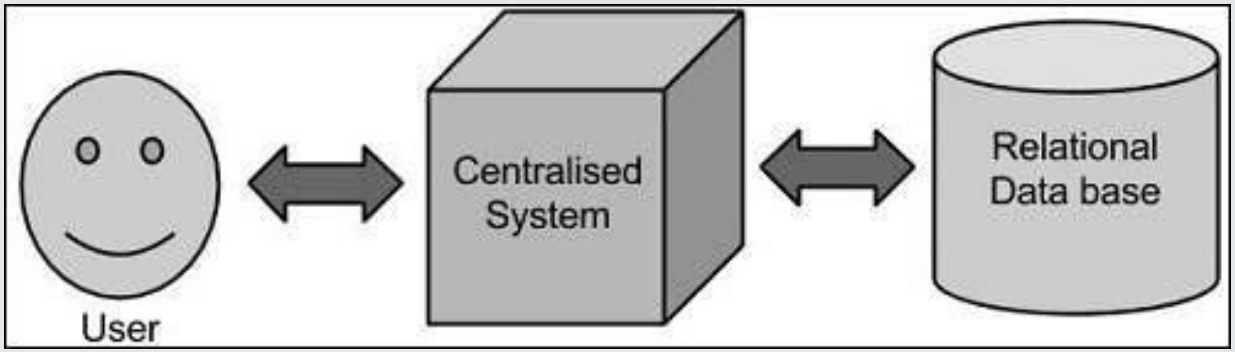
Operational vs. Analytical Systems

	Operational	Analytical
Latency	1 ms - 100 ms	1 min - 100 min
Concurrency	1000 - 100,000	1 - 10
Access Pattern	Writes and Reads	Reads
Queries	Selective	Unselective
Data Scope	Operational	Retrospective
End User	Customer	Data Scientist
Technology	NoSQL	MapReduce, MPP Database


BIG DATA SOLUTIONS

Traditional Approach

- Traditional Approach



Traditional Approach

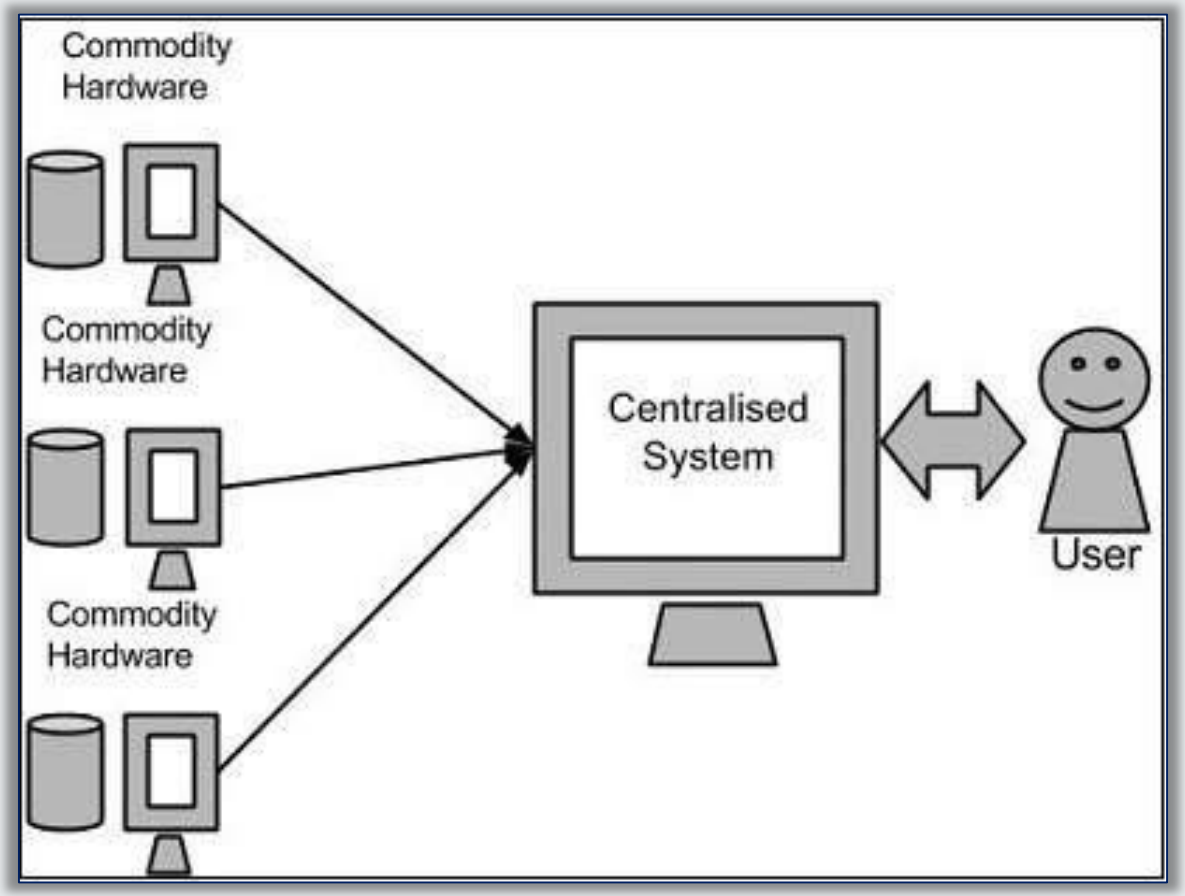
Limitation

- Process less voluminous data that can be accommodated by standard database servers
- Hectic task to process such data through a single database bottleneck.

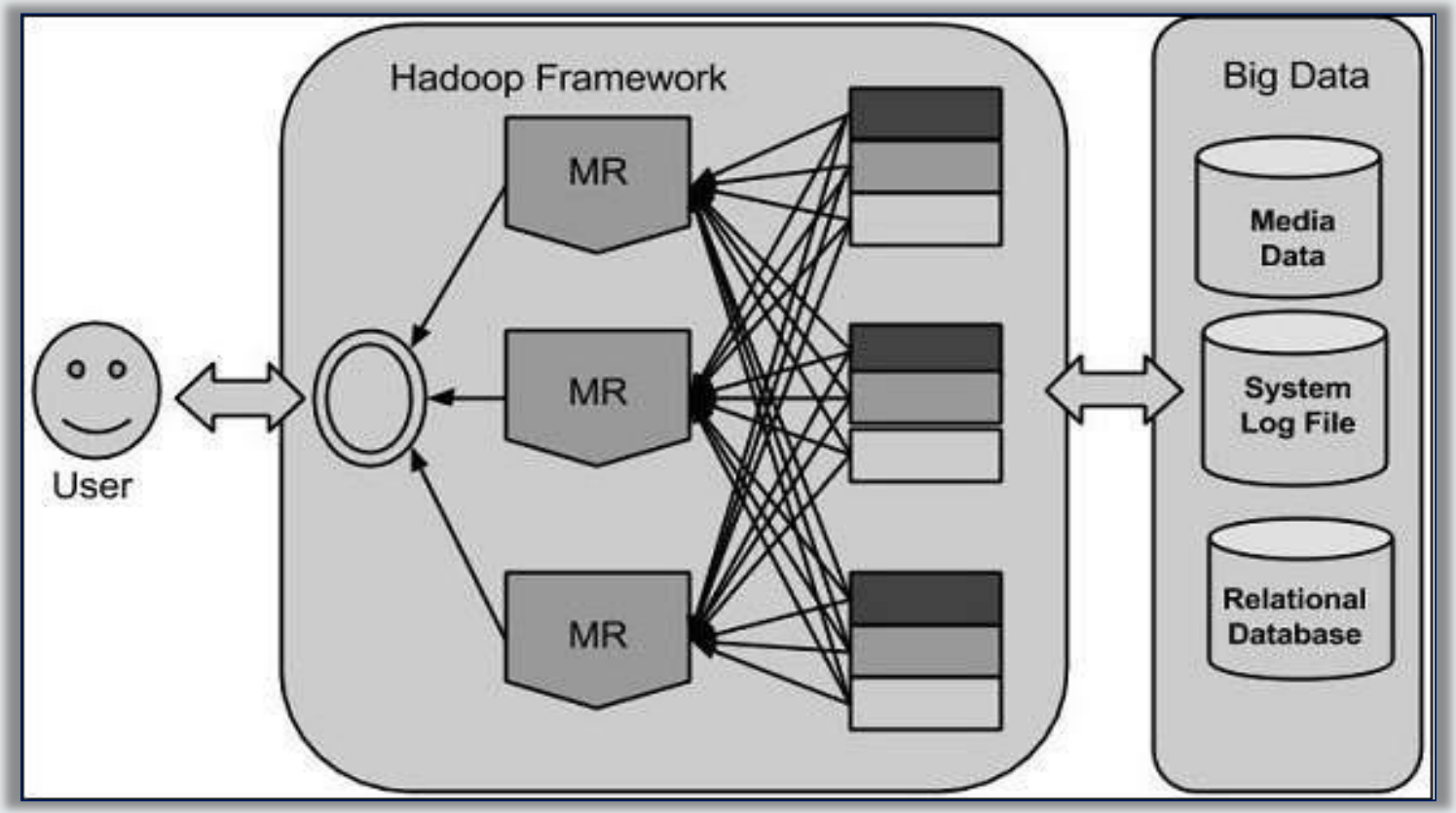
Google's Solution

- Google solved this problem using an algorithm called MapReduce.
- This algorithm divides the task into small parts and assigns them to many computers, and collects the results from them which when integrated, form the result dataset.

Google's Solution



Google's Solution




That's all for now...