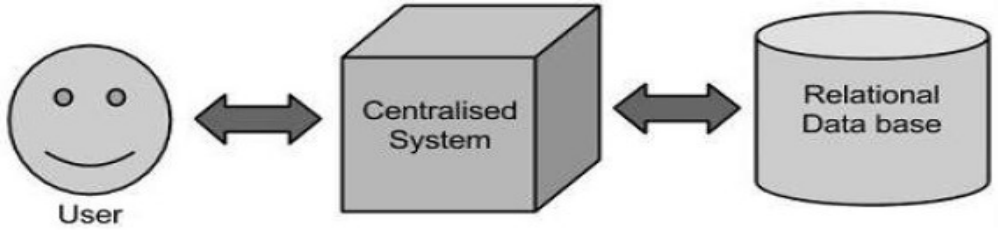
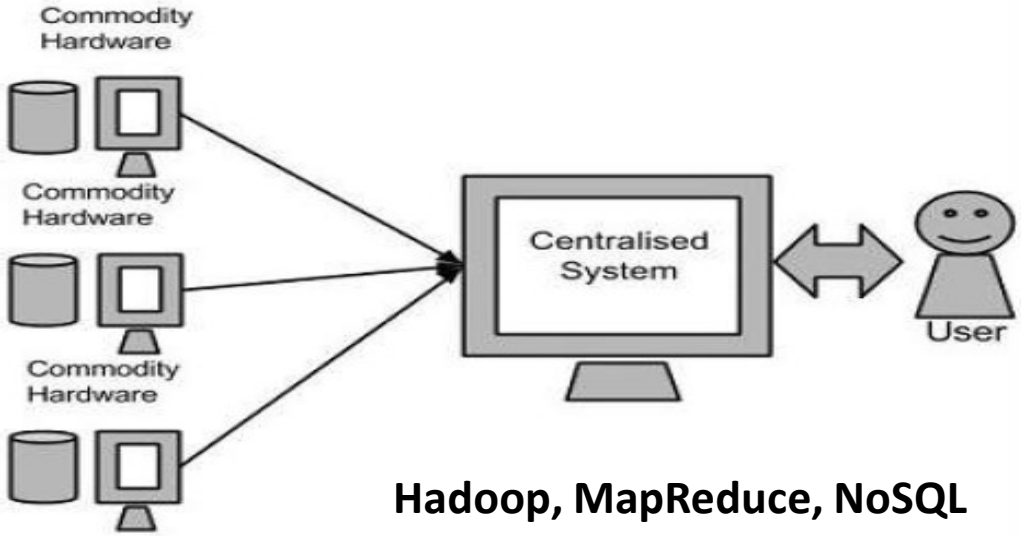
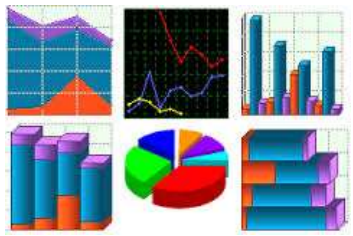


Copying data from different sources to centralized system . and keeping updated is not an easy task.
Sampling will not serve extracting information.





Hadoop, MapReduce, NoSQL

Advantages of Big data analytics

Scalability.
No Pre-processing.
Any Unstructured data from different data sources.
Protection against hardware failure.
A simple interface to carry out analysis.
Facility to view results in different ways as per user's needs.

	TRADITIONAL DATA WAREHOUSE	BIG DATA LAKE
Data Store	RDBMS – Static Schema ,OLTP	HDFS, NoSQL – Dynamic Schema
Volume Velocity	GBs , but not petabytes or zetabytes E. g.Billions of credit card transactions	Petabytes or Zetabytes. Constantly generated. Every second. e.g. Search query, Sensors, Satellite,StockPrice
Data Source Variety	Structured, processed and **Centralized**	Structured/Unstructured/Semi-structured, **Multi-Structured** **Distributed**
Processing	Repeated Read and Write	Write Once, Repeated Read
Integration	Easy	Difficult
Data Access	Interactive	Batch or near or Real time
Storage Cost	Expensive for large data volumes	Designed for low cost storage
Agility Flexibility	Less agile and fixed configuration, fixed schema	Highly agile, configure and reconfigure as needed, Flat schema, flexible schema
Security	Mature	Maturing
Users	Business Professional	Data Scientists
	Such project will take at least 1 year.	Few days/months
Scalability	Organizations with predictable & constant workloads are better served.	Servers can be added on demand to accommodate the growing workloads.
Relationship	Complex	Almost Flat with few

What do Big Data Professionals do?

The responsibilities of big data professional lies around dealing with huge amount of heterogeneous data, which is gathered from various sources coming in at a high velocity.

Architect distributed systems

Builds large scale data processing system

Process the data using various big data tools & ensure network connectivity

Roles of Big Data Professional

Big data professionals describe the structure and behavior of a big data solution and how it can be delivered using big data technologies such as Hadoop, Spark, Kafka etc. based on requirements.

Traditional Approach	**Big Data Approach**
Structured and Repeatable Analysis	Iterative and Exploratory Analysis
Business Users determine what questions to ask	IT delivers a platform to enable creative discovery
IT structures the data to answer that question.	Business users explores what questions could be asked
Monthly sales reports Profit analysis Customer surveys	Brand sentiments Product Strategy Maximum asset utilization Preventive care

Reference: https://www.edureka.co/blog/data-science-vs-big-data-vs-data-analytics/

Technologies of Big Data

Hadoop (Open src framework)	Huge volumes of data is fragmented into chunks. Theses chunks are stored and processed across thousands of servers.
NoSQL	DBMS for unstructured data. No need to model the data. BigTable, Hbase
MapReduce	Programming paradigm that assists massive scalability in Hadoop Cluster
HDFS Hadoop Distributed File System	Manages storage and retrieval of data and metadata required for computation.
Hbase Hadoop database	Hadoop Database. Non relational database. Real time data is stored in column oriented tables. It is a backend system for MapReduce jobs output.
Sqoop	(Data transfer tool)Hadoop to RDBMS/DWH or HDFS or Hbase or Hive
Hive (DWH platform)	Built on the top of Hadoop
HiveQL	Querying and Managing large datasets.
Mahout	Data Mining Library that can be used to implement data mining algorithms such as clustering, classification, frequent pattern

Advantages of Big data analytics

- **Scalability.**
- **No Pre-processing.**
- **Any Unstructured data from different data sources.**
- **Protection against hardware failure.**
- **A simple interface to carry out analysis.**
- **Facility to view results in different ways as per user's needs.**

Benefits

Insights about customer needs

Digital Information

Customer related dataset in timely fashion
+ Cutting Edge Technologies

Business Organization

Innovations in processes.
Improvement in productivity.
New opportunities to sustain in the market.

Insurance Company – Frauds by accessing internal and External claims.
-- Speed up the handling of simple claims.

Manufacturer/Distributers –
Supply – Chain issues
-- different logistic approaches to avoid additional cost

Material delay
Overstock
Stock-out conditions

Customer

Benefits

Insights about customer needs

Digital Information

Customer related dataset in timely fashion
+ Cutting Edge Technologies

Public Services – usage of services
Traffic -- Optimize their delivery mechanisms
Ambulance, Transportations

Smart City – Sensors, Crime, Emergency Services, Real estate, Energy, Financial transactions, astronomy etc..
-- make cities more efficient and sustainable to improve the lives of citizens.

Firms – Customer insights
-- Better Clarity to provide services and build strong customer loyalty
(Hotels, Telecomm Companies, Retailers..)

Customer

Case Study 1

How many times the goods are returned and refunded?

Company: can identify/find the performance of vendor or supplier

Feedback.txt

Sent to

Client Program

Simple Word Counting Exercise.

Customer Feedback.txt

Case Study 2 Click-Stream Analysis

Company: can identify/find the Loyal Customer

Visits Web site

Informs about products and services

Allows to do transactions

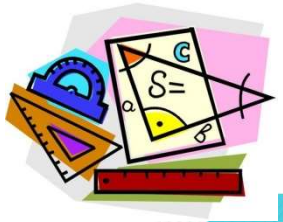
Millions Customer

Improve Contents, Navigation, Organization , Transaction Completion

Business Organization

ClickStream Data is generated for The customers/visitors

- Pages they load
- Time spent on every page
- Links clicked
- Frequency of their visits of pages



Mathematical Model

Loyal Customer

Sales and Marketing