BIG DATA ANALYTICS - 18CS72

Books

1. Raj Kamal and Preeti Saxena, “Big Data Analytics Introduction to Hadoop, Spark, and Machine-Learning”, McGraw Hill Education, 2018 ISBN: 9789353164966, 9353164966

2. Douglas Eadline, "Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem", 1st Edition, Pearson Education, 2016. ISBN13: 978-9332570351

Dr. Vaneeta M
Associate Professor
Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Vision

- **“To create competent professionals in Computer Science and Engineering with adequate skills to drive the IT industry”.**

Mission

- **Impart sound technical knowledge and quest for continuous learning.**
- **To equip students to furnish Computer Applications for the society through experiential learning and research with professional ethics.**
- **Encourage team work through inter-disciplinary project and evolve as leaders with social concerns.**

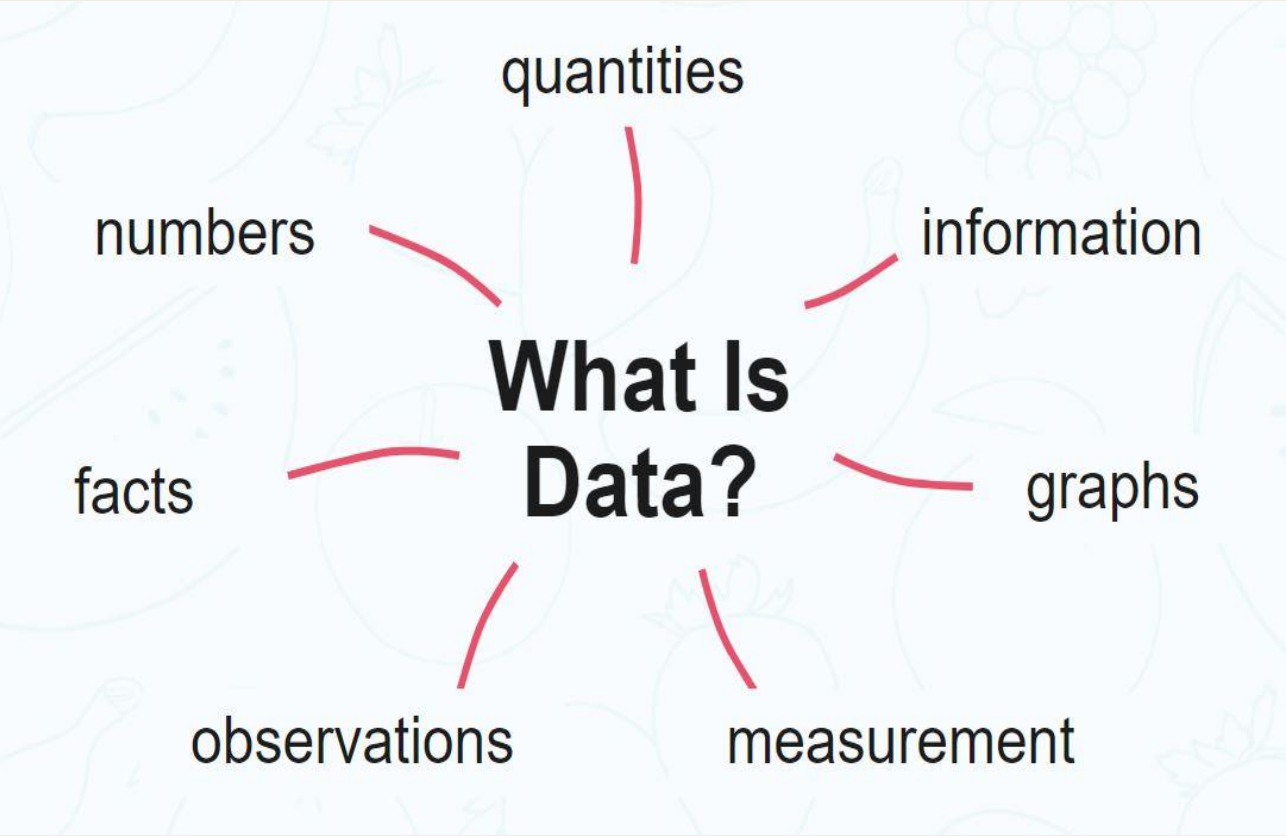
Module 1

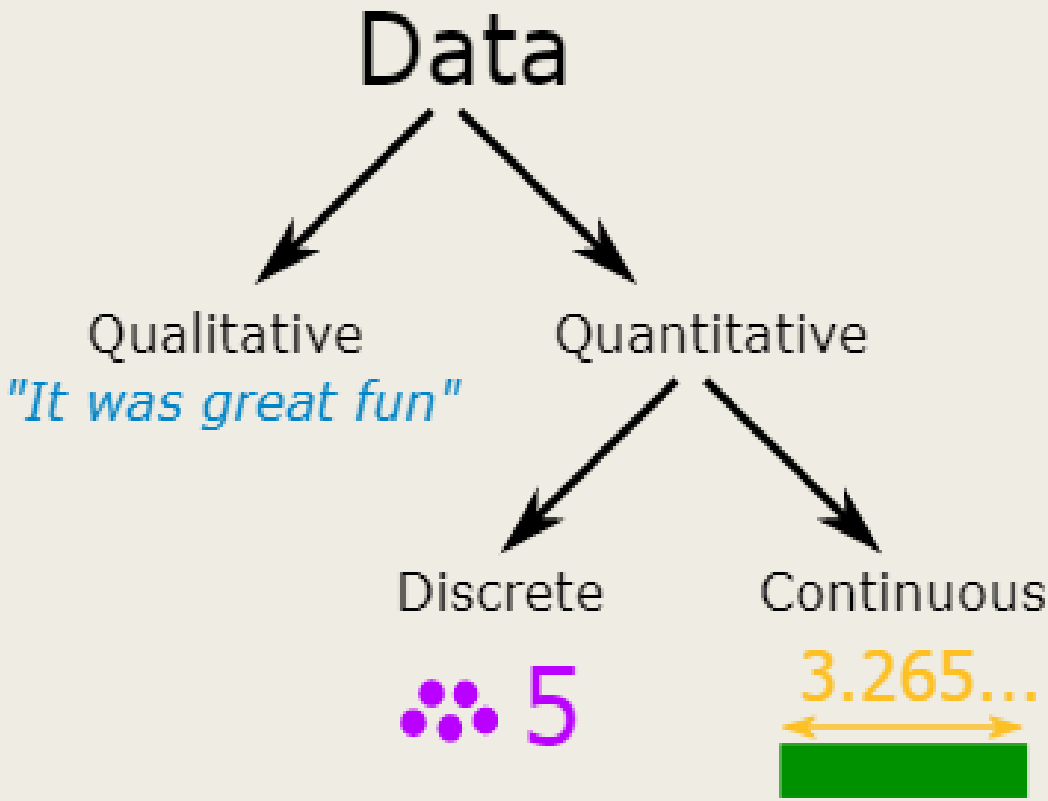
Chapter 1

Introduction to Big Data Analytics

CO1: Identify the fundamentals of Big Data analytics

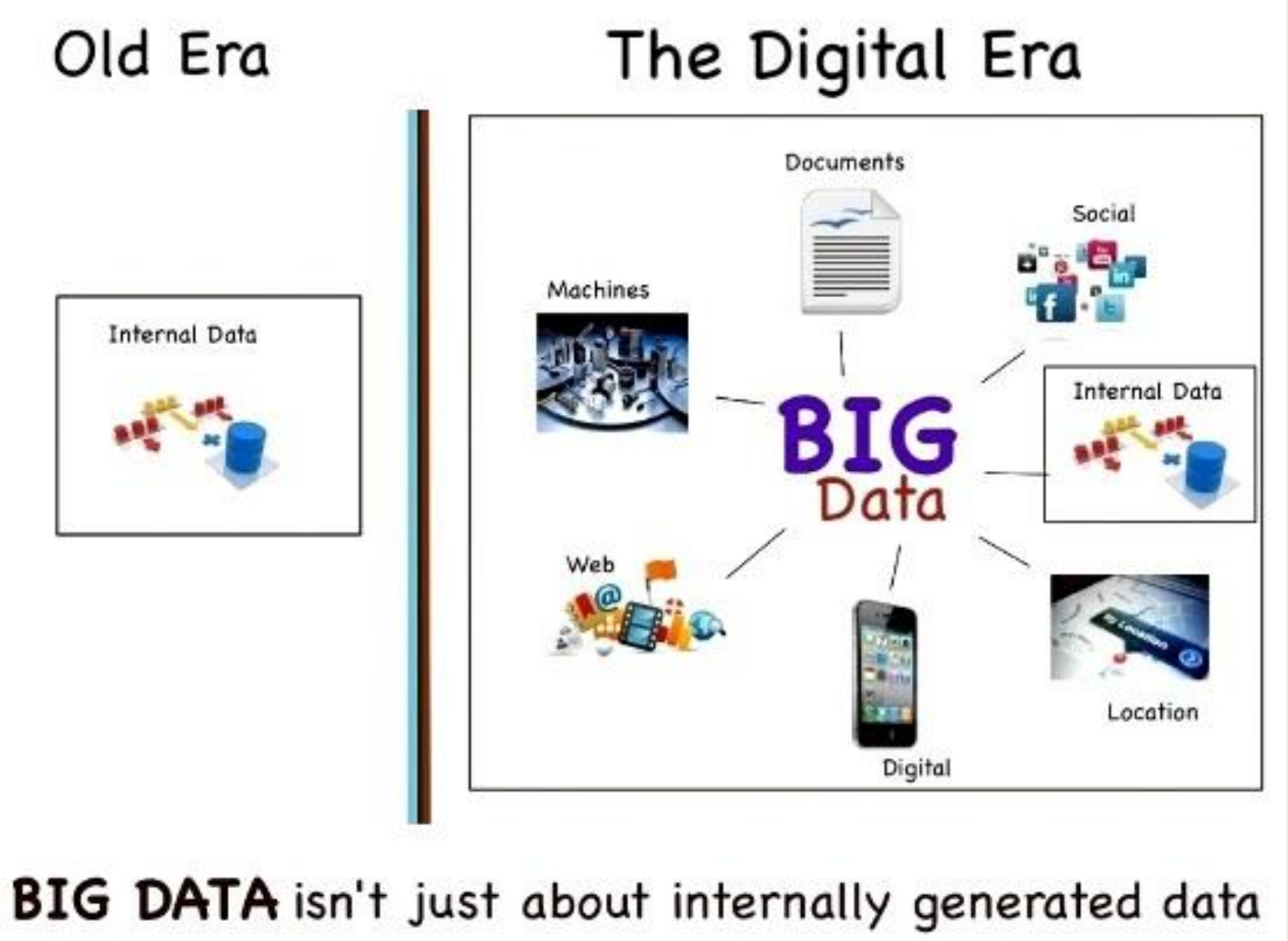
What is data?





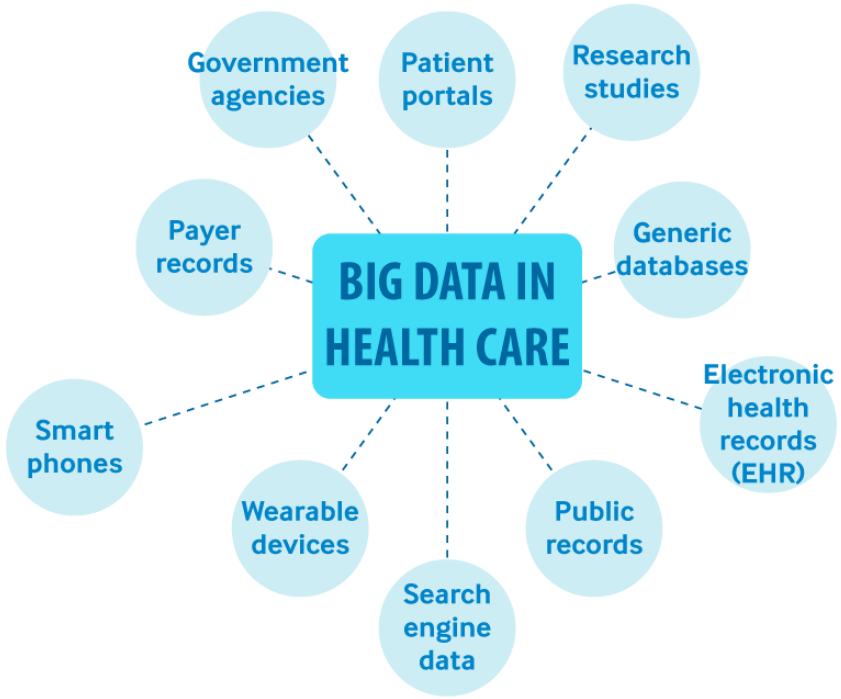
- Data are individual facts, statistics, or items of information, often numeric, that are collected through observation.
- Datum is a single value of a single variable
- Data are a set of values of qualitative or quantitative variables about one or more persons or objects
- Examples of data are **weights, prices, costs, numbers of items sold, employee names, product names, addresses, tax codes, registration marks** etc. Images, sounds, multimedia and animated data as shown.

What is Big Data?



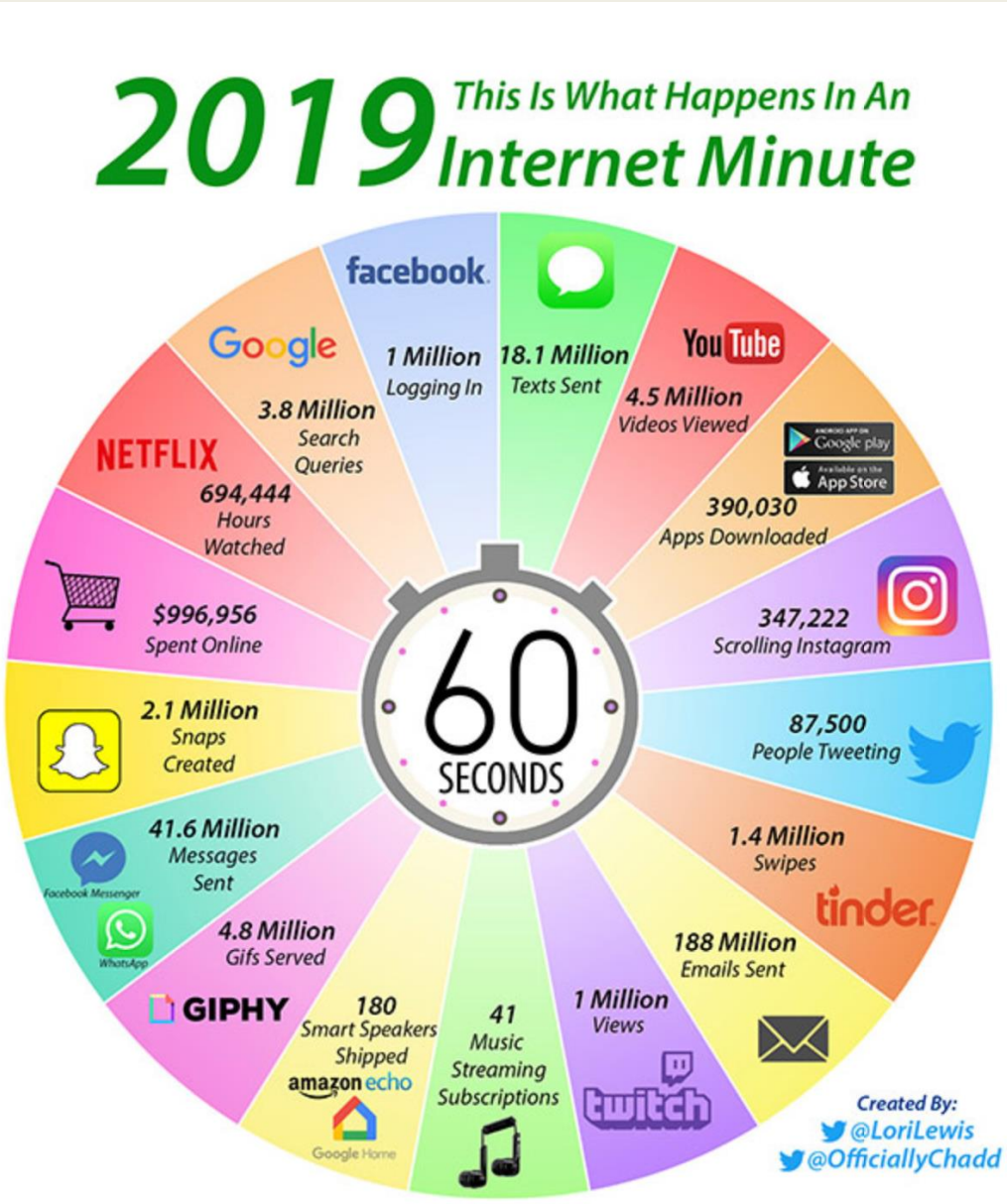
What is Big Data?

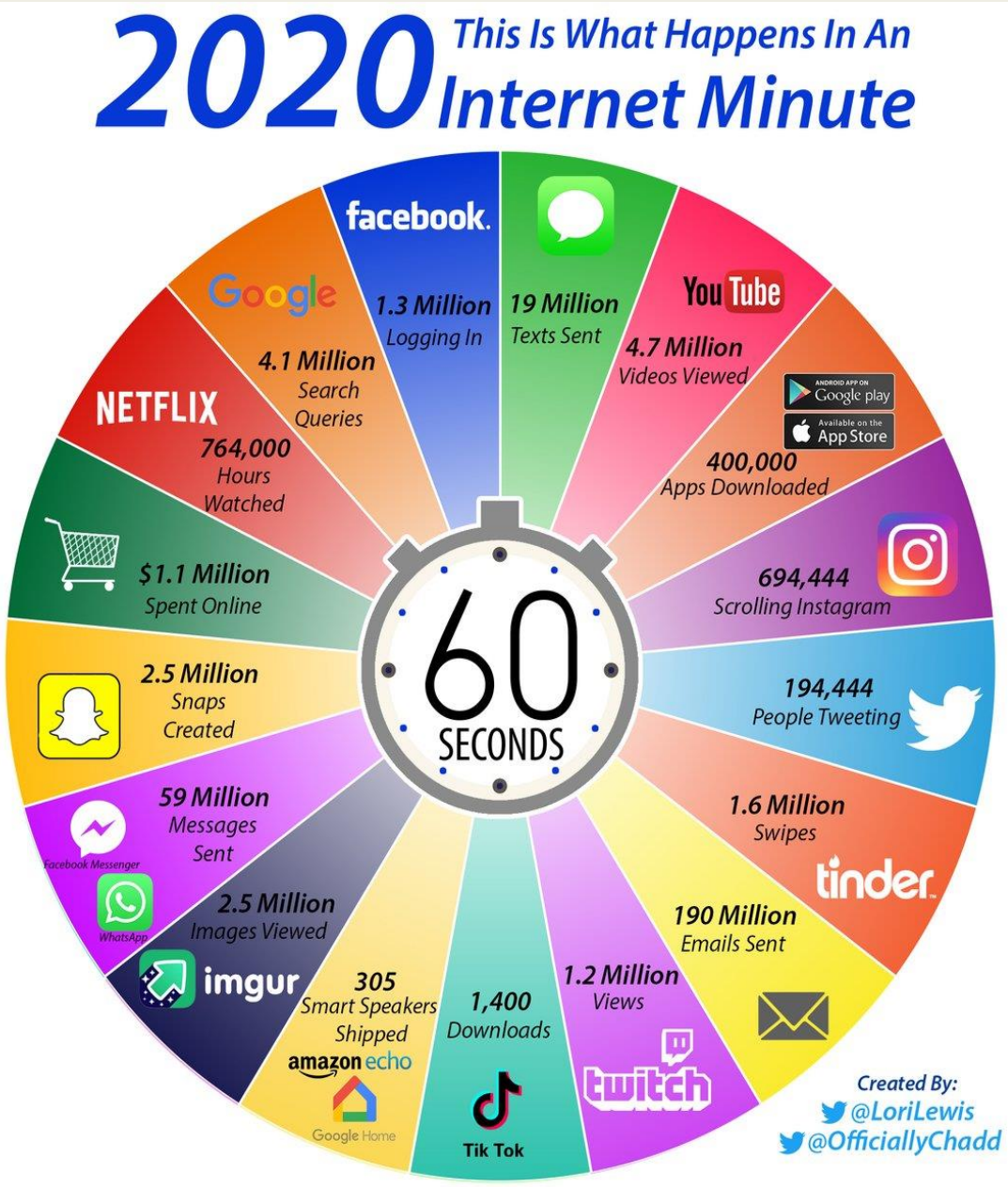
Sources of Big Data in Health Care



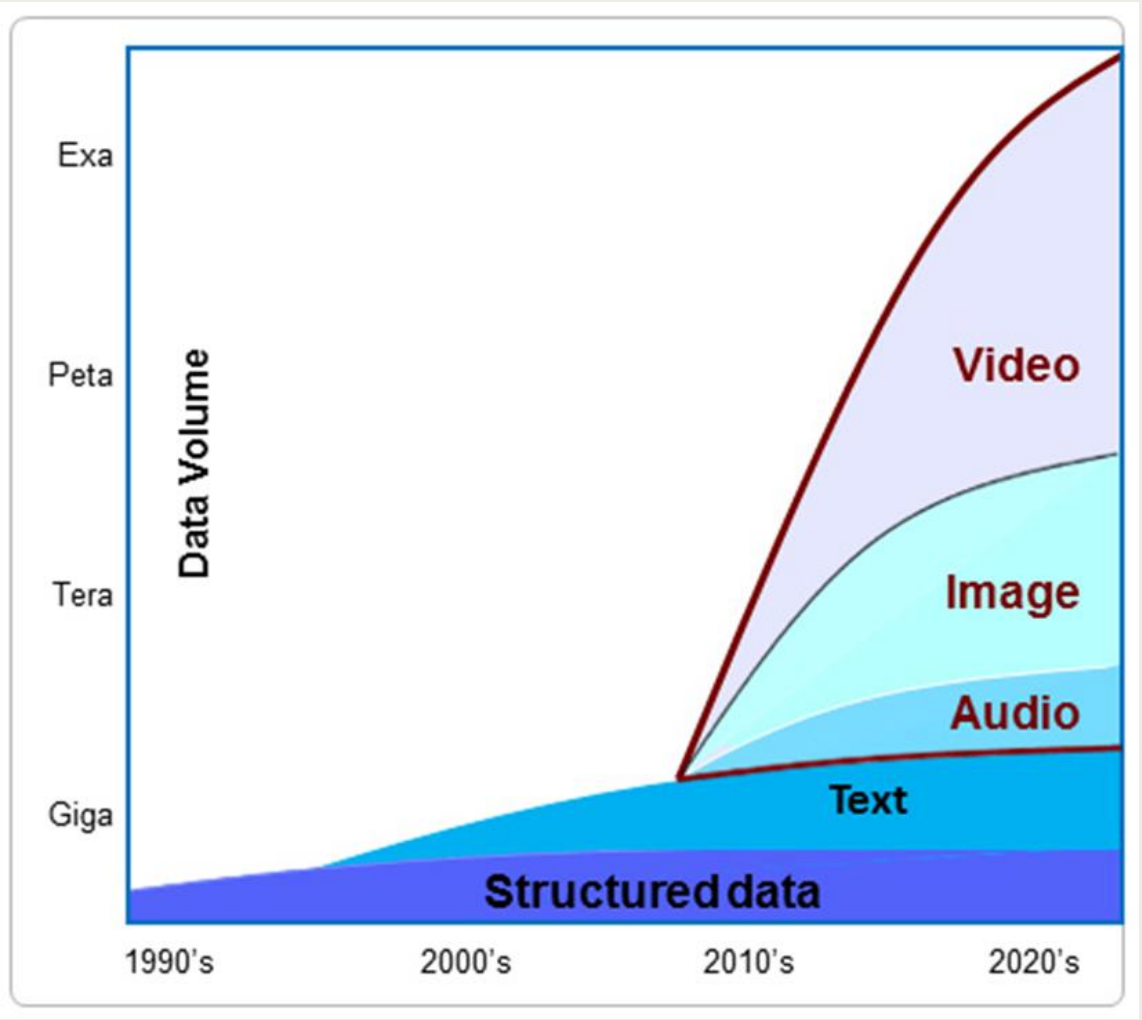
NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society

Big Data generated every minute?

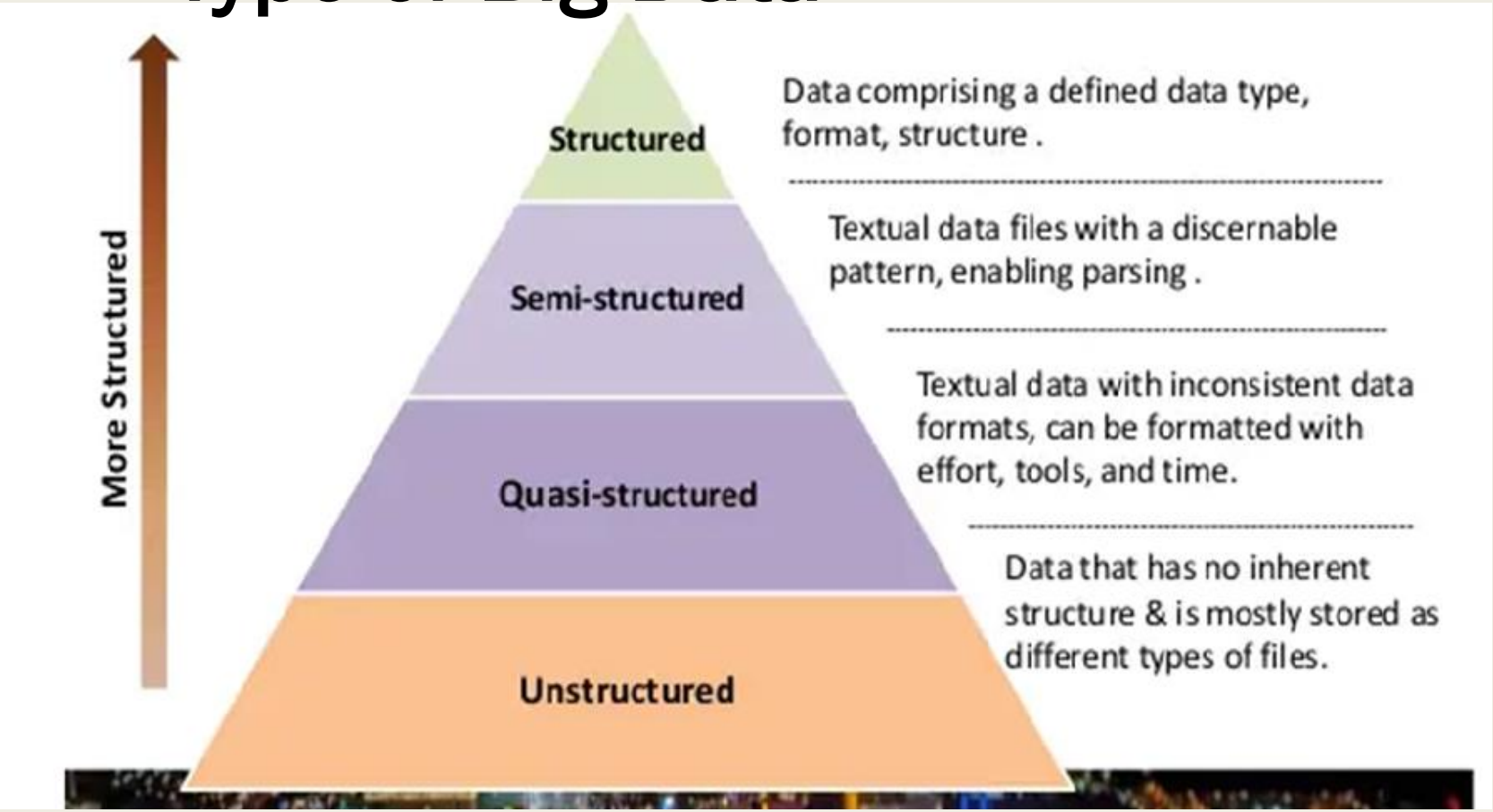




Multiples of Bytes		
Unit (Symbol)	Value (SI)	Value (Binary)
Kilobyte (kB)	10^{3}	2^{10}
Megabyte (MB)	10^{6}	2^{20}
Gigabyte (GB)	10^{9}	2^{30}
Terabyte (TB)	10^{12}	2^{40}
Petabyte (PB)	10^{15}	2^{50}
Exabyte (EB)	10^{18}	2^{60}
Zettabyte (ZB)	10^{21}	2^{70}
Yottabyte (YB)	10^{24}	2^{80}



Type of Big Data



- Structured examples are **names, addresses, credit card numbers, geolocation, and so on**.
- Semi-structured sources are **emails, XML and other markup languages**, binary executables, TCP/IP packets, zipped files, data integrated from different sources, and web pages.
- Quasi-structured data is **more of a textual data with erratic data formats**. This data type includes web clickstream data such as Google searches.
- Unstructured data are: Media and entertainment data, surveillance data, geo-spatial data, audio, weather data. Invoices, records, emails, productivity applications. Sensor data. Analytics ML and AI

What is Information?

Data vs. Information

Data

- raw facts
- no context
- just numbers and text

Information

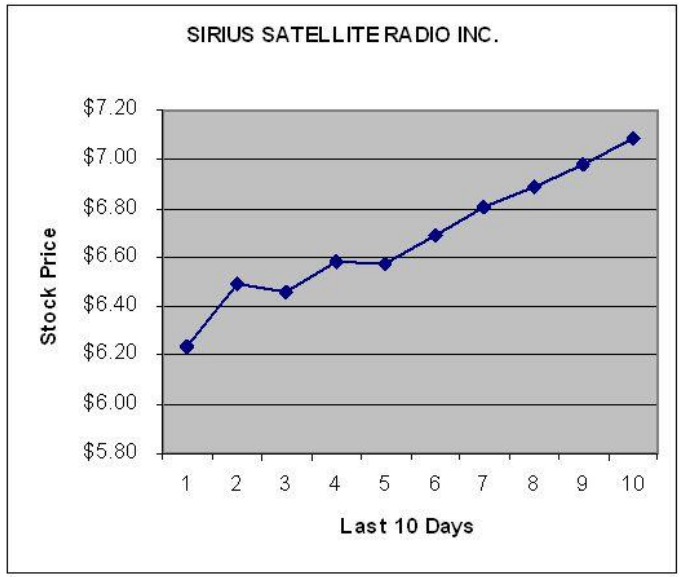
- data with context
- processed data
- value-added to data
  - summarized
  - organized
  - analyzed

Data vs. Information

Data

- 6.34
- 6.45
- 6.39
- 6.62
- 6.57
- 6.64
- 6.71
- 6.82
- 7.12
- 7.06

Information



What is Big Data Analytics?

Big Data Analytics is “**the process of examining large data sets containing a variety of data types, to uncover hidden patterns, unknown correlations, market trends, customer preferences, and other useful information**.”

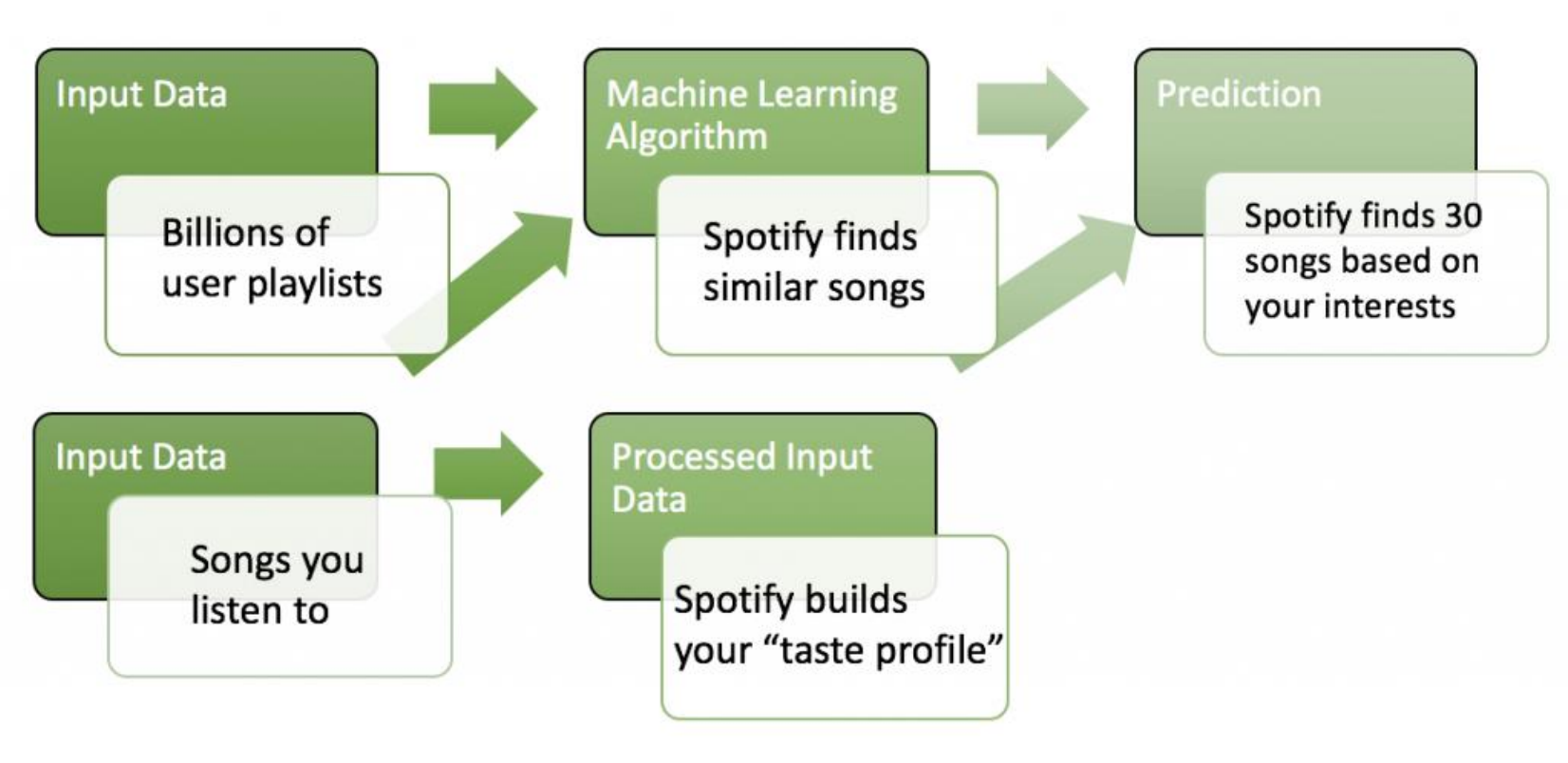
There are several benefits of using big data:

- The major role of big data in any company is to make **better business decisions**.
- It will encourage companies to **amass better market** and consumer intelligence.
- It can **enhance internal efficiency and operations** for nearly any type of business.
- Modern big data analytics and operations anticipate the patterns of consumers. After that, they use those **patterns to motivate brand** loyalty as they can collect more data to observe more trends and also the ways to make consumers satisfied.
- It helps in delivering **smarter services and products**.


Top 10 Companies using Big Data
Amazon
Netflix
American Express
Starbucks
LinkedIn
McDonald's
General Electric
Swiggy
Miniclip
Spotify

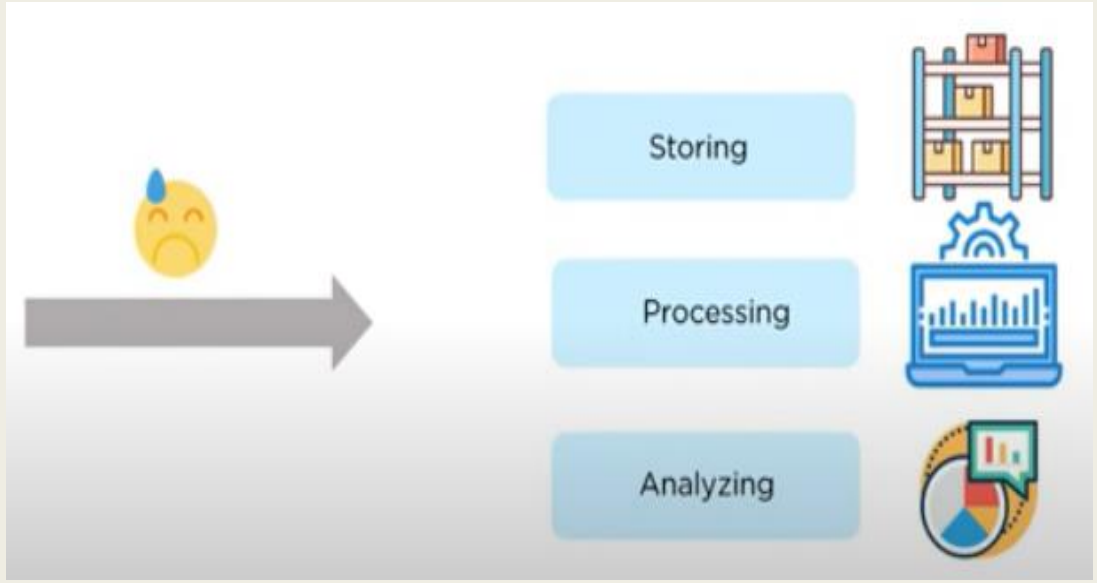
How Spotify uses data to keep You Listening.

- Spotify has **96 million** user
- All users **generate tremendous** amount of data
  Songs Played,
  Repeatedly used playlist,
  likes, shares, and search history
- Spotify analyses big data for **suggesting** songs for users
- Use Recommendation system to do predictions.



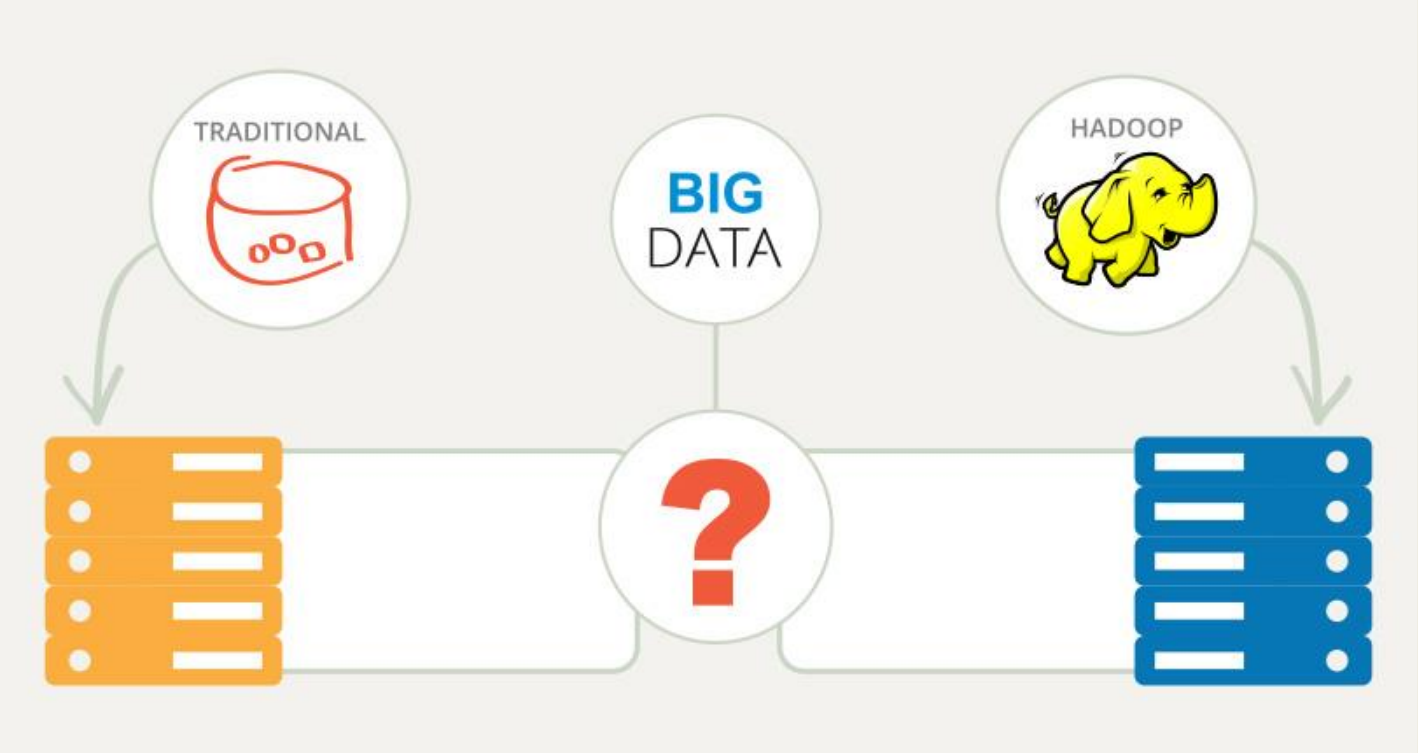
Handling Big Data?

Difficulty with RDBMS



Solution

What is HADOOP?



- The Apache Hadoop software library is a big data framework.
- It allows **distributed processing** of large data sets across **clusters** of computers.
- It is one of the best big data tools designed to scale up from **single servers to thousands of machines.**

Comparison between RDBMS and HADOOP

RDBMS		HADOOP
Structured	Data Types	Multi and Unstructured
Limited, No Data Processing	Processing	Processing coupled with Data
Standards & Structured	Governance	Loosely Structured
Required On Write	Schema	Required On Read
Reads are Fast	Speed	Writes are Fast

SQL and NoSQL

- In summary, the five key differences between SQL vs. NoSQL are:

1. SQL databases are **relational**, NoSQL databases are **non-relational**.
2. SQL databases use structured query language and have a **predefined schema**. NoSQL databases have **dynamic schemas** for unstructured data.
3. SQL databases are **vertically scalable**, while NoSQL databases are **horizontally scalable.**
4. SQL databases are **table-based**, while NoSQL databases are **document, key-value, graph, or wide-column stores.**
5. SQL databases are better for **multi-row transactions**, while NoSQL is better for unstructured data like **documents or JSON**.

mongoDB

- MongoDB is a **document** oriented NoSQL database.
- MongoDB stores data in flexible **JSON** like document format.
- The fields can vary from document to document, and it gives you the flexibility to **change the schema** any time.
- MongoDB is a distributed database, so it **provides high availability & horizontal scalability.**

Cassandra

- Apache Cassandra is an **open source, distributed and decentralized/distributed storage system** (database), for managing very large amounts of structured data spread out across the world.
- **Cassandra** is the modern version of the **relational database**
- Data is grouped by **column** instead of row, for fast retrieval.
- It provides highly available service with no single point of failure.
- It is scalable, fault-tolerant, and consistent.

Components of HADOOP

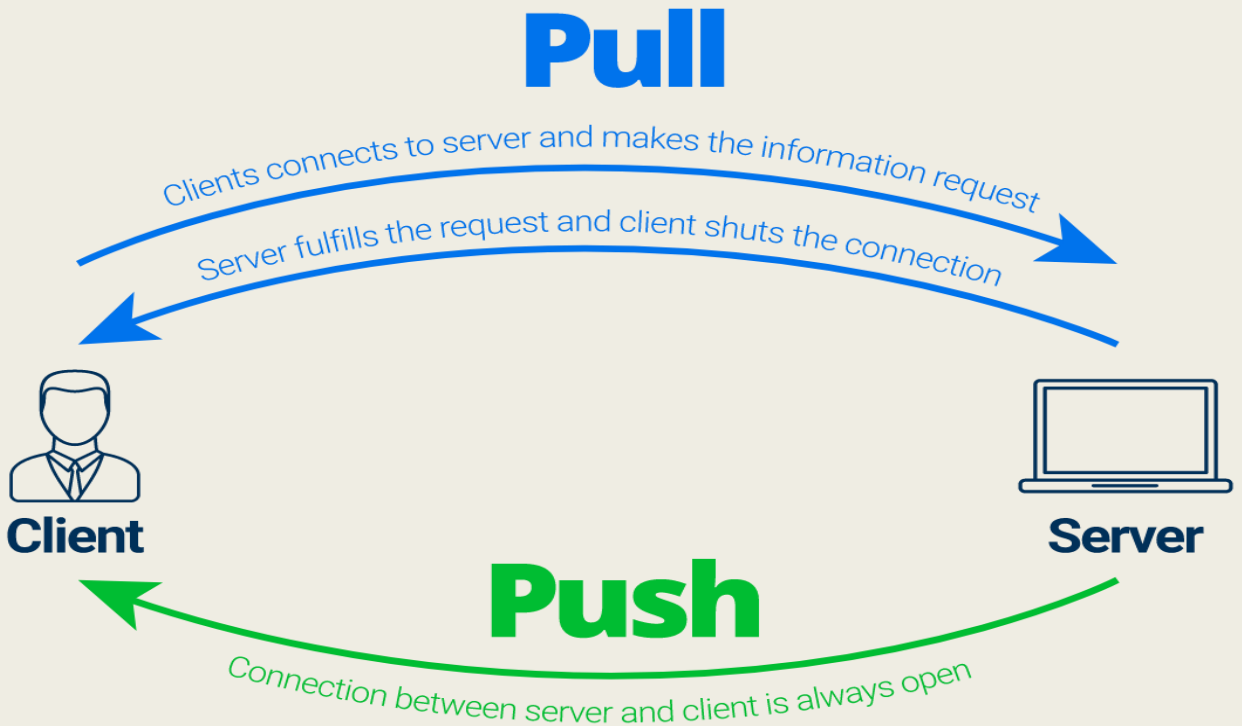
- **MapReduce**
  - MapReduce is a **programming model or pattern** within the Hadoop framework that is used **to access big data stored in the Hadoop File System (HDFS)**.
  - It is a core component, integral to the functioning of the Hadoop framework.
  - A MapReduce job usually **splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner**.
  - MapReduce consists of two distinct tasks – Map and Reduce.
- **Hive**
  - Hive **allows users to read, write, and manage petabytes of data using SQL**. Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets.
- **Pig**
  - Pig is a high level scripting language that is used with Apache Hadoop. Pig **enables data workers to write complex data transformations without knowing Java**. Pig's simple SQL-like scripting language is called Pig Latin, and appeals to developers already familiar with scripting languages and SQL.

Definitions of Data

- “Data is information usually in form of facts or statistics that one can analyze or use for further calculations”[**English Dictionary**]
- “Data is information that can be stored and used by a computer program”[**Computing**]
- “Data is information presented in numbers, letters or other form”[**Electrical Engineering]**
- “Data is information from series of observation, measurement or facts”[**Science**]
- “Data is information from series of behavioral observations, measurements or facts”[**Social Science**]

Definition of Web Data

- Web is **large scale integration and presence** of data on web servers.
- Data as **documents and resources.**
- **URL** enables access to web data resources
- **Web data** is present on web server in form of:
  - *Text*
  - *Image*
  - *Videos*
  - *Audios*
  - *Multimedia files*
- Access Data
  - *Pull (Request Data)*
  - *Push (Publish/Post Data)*



Web Data

- Internet Applications Provide and consume web data :
  - *Websites, web portals, online business applications, emails, chats, tweets and social networks*
- **Example of Web Data**

Web based free content based encyclopedia project supported by Wikimedia Foundations.

Provider of real time navigation, traffic, public transport and nearby places by Google Inc.

Digital Teaching and learning environment that saves students and instructor time .

book store

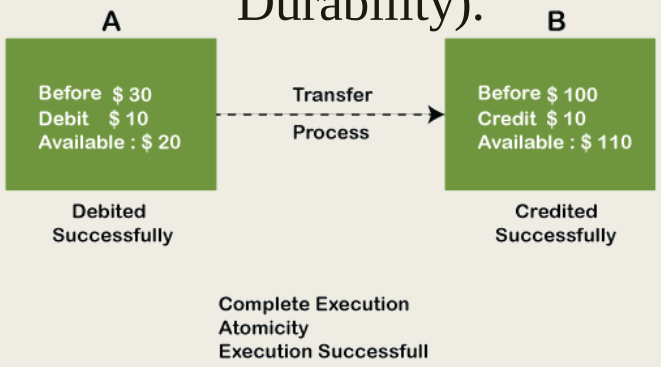
Allows billions of people to discover, watch and share originally created videos by Google Inc.

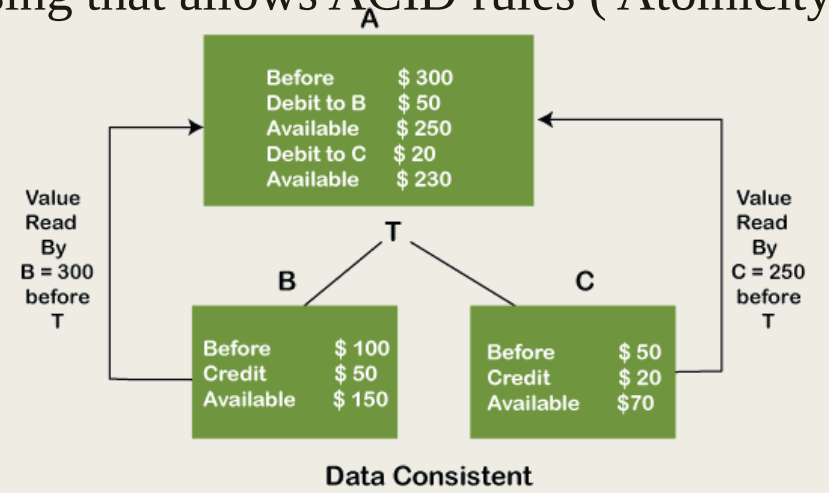
Classification of Data

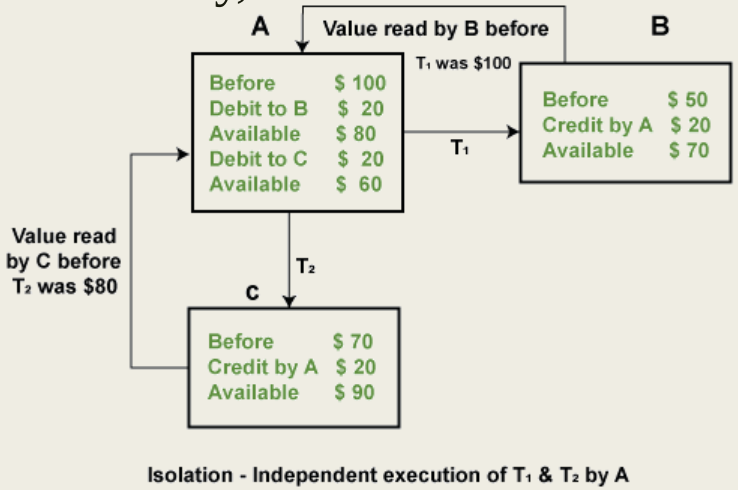
- Data can be classified as :
  1. Structured
  2. Semi-Structured
  3. Multi-Structured
  4. Unstructured

Using Structured Data

- Structured data enables the following:
  - Table Oriented
  - Perform Data insert, delete, update, and append
  - Indexing to enable **faster** data retrieval
  - **Encryption** and **decryption** for data security.
  - Highly organized, clearly defined, Easy to access and analyse
  - **Scalability** – Enables increasing or decreasing capacities and data processing operations such as storing, processing and analytics.
  - **Transactions** processing that allows ACID rules (Atomicity, Consistency, Isolation and Durability).







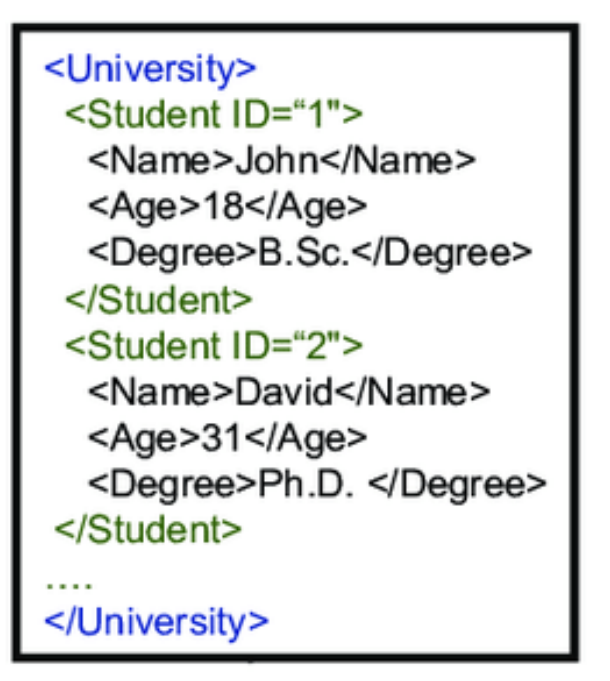
Using Structured Data

- **Examples**: Name, Age, Address, Currency etc.
- **Sources of Structured Data:**
  - SQL Database
  - Spreadsheets
  - Online Forms
  - Point of Sales
  - Web and Server Logs

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Using Semi-Structured Data

- No rigid fixed schema
- Does not conform and associate with formal data model structure.(Ex: RDBMS)
- Self describing [data itself carries information about the structure] and flexible structure.
- Contain tags or other markers, which separate semantic elements and enforce hierarchies of record and fields within data.
- Sources : Social Media: Tweets, Blogs, Likes, Follower



Using Multi-Structured Data

- Data has multiple formats.
- Found in non-transactional systems.
- Ex: Streaming data on customer interaction
- Data of multiple sensors
- Data at web or enterprise server or data warehouse.
- Large scale integrated system are required to aggregate and use the widely distributed resource efficiently.

Using Unstructured Data

- Data does not poses any feature such as table or database.
- Ex: .TXT, .CSV files
- Data may be as key-value pair such as hash key value pairs.
- Data may have internal structure such as e-mail.
- The relationships, schema and features need to be separately established.
- Ex: Mobile data: text messages, chat messages, tweets,, blogs and comment.
- Website data
- Social media
- Satellite images, atmospheric data, surveillance, traffic video etc.

Big Data Definition

- Big Data is high-volume, high-velocity and/or high-variety information asset that requires new forms of processing for enhanced decision making, insight discovery and process optimization**. (Gartner1 2012)**
- "Data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges." **[Oxford English Dictionary (traditional database of authoritative definitions)]**
- "A collection of data sets so large or complex that traditional data processing applications are inadequate."**— Wikipedia**
- **"**Big Data refers to data sets whose size is beyond the ability of typical database software tool to capture, store, manage and analyze".**[The McKunsey Global Institute 2011]**

Big Data Characteristics

- Industry analyst Doug Laney described the ‘3Vs’, i.e. volume, variety and/or velocity as the key “data management challenges” for enterprises
- Analytics also describe the ‘4Vs’, i.e. volume, velocity, variety and veracity as the characteristics
- **Big Data Volume**
  - Term big relates to size of the data and hence the characteristic
  - Size defines the amount or quantity of data, which is generated from an application(s)
  - The size determines *the processing considerations needed for handling that data*

Big Data Velocity

- Term velocity refers to the speed of generation of data
- Velocity is a measure of how fast the data generates and processes
- To meet the demands and challenges of processing Big Data, the velocity of generation of data plays a crucial role.

Big Data Variety

- Data is generated by multiple sources in a system.
- So there is variety in data and introduce complexity .
- Data consists of various forms and formats.
- Term refers to a variety of data, due to the availability of a large number of heterogeneous platforms in the industry.
- Type to which big data belongs, is important characteristics that needs to be known for proper processing of data.

Big Data Veracity

- It is important characteristic to take into account the quality of data captured, which can vary greatly, affecting its accurate analysis.
- Data veracity, is **how accurate or truthful a data set may be**.
- An example of a high veracity data set would be **data from a medical experiment or trial**. Data that is high volume, high velocity and high variety must be processed with advanced tools.

4Vs (i.e. volume, velocity, variety and veracity) needs tools

- For mining, discovering patterns, business intelligence, artificial intelligence (AI), machine learning (ML), text analytics, descriptive and predictive analytics, and the data visualization

Big Data Types

- **Social Networks and web data**: Facebook, Twitter, e-mails, blogs, YouTube.
- **Transactions and Business Processes Data**: Credit card transactions, flight booking, public agencies data such as medical records, insurance business data.
- **Customer master data**: facial recognition, name, age date of birth etc…
- Machine generated data: M2M of IoT data:
  - ***Computer generated data*** *: Logs, weblogs, security/surveillance, videos/images etc.*
  - ***Fixed sensor****: Home Automation, weather sensor, pollution sensor, traffic sensors etc..*
  - ***Mobile sensor*** *(Tracking) and location data.*
- Human generated data: Biometric data, Human machine interaction data, e-mail record with mail server and MySQL database of student grades.
  - *Human record their experience, audio, video, photos etc.*

Usage of Big Data generated from multiple types of data sources for optimizing the services offered, products, schedules and predictive tasks.

- Manufacturing and retail marketing company: LEGO Toys
- Use several big data sources such as
  - Machine generated data from sensors at the toy packaging
  - Transaction data of sales stored as web data for automated reordering by the retail stores
  - Tweets, Facebook, posts, emails, messages and web data for messages and reports.
  - Understand the theme, demand of toys in present day by children and predict the future types and needs.
  - Send messages to retailers and children using social media on the arrival of new and popular toys.

Example of features of 3V's in Big Data and Applications

Satellite images of the earth atmosphere and its regions:

- **Volume**: Large data generated by KALPANA, INSAT-1A and INSAT-3D, Foreign satellites. Record images of full disk and east and west sectors and regions.
- **Velocity**: Satellite collects images round the clock. Big data analytic helps in drawing of maps of wind velocities, temperature and other whether parameters.
- **Variety**: Images can be in visible range such as IR-1, IR-2, Shortwave Infrared SWIR, medium range IR MIR and color composite.
- **Veracity:** uncertain or imprecise data arises due to poor resolution used for recording or noise in images due to signal impairment.

Big Data generated from multiple types of data sources Examples

(i) **Chocolate Marketing Company with large number of installed Automatic Chocolate Vending Machines (ACVMs)**

Sells five flavors of chocolates. The company uses big data types as:

(i) Machine Generated data: sale of chocolate

(ii) Reports of filled and unfilled machine transaction data.

(iii) Human generated data of buyer machine interaction at ACVM

(iv) Social networks and web data on feedback and personalized message based on interactions.

(v) Human generated data on facial recognition of the buyers.'

(vi) Use big data for efficient and optimum planning of fill service for chocolate

(vii) Sentiment analysis of buyers for specific flavors.

(viii) ACVM's location and periods of higher sales analysis, additions and relocation of machines and predictions, strategies and planning for festival sales.

(ii) **Automotive Components and Predictive Automotive Maintenance Services (ACPAMS) rendering customer services for maintenance and servicing of (Internet) connected cars and its components**

- *Machine generated data: from sensors placed at brakes, steering and engine from each car*
- *Transactions data stored at service website*
- *Social networks and web data in form of messages, feedback and reports from customers.*
- *Message for scheduled and predictive maintenances.*
- *Service generates reports on social networks and updates the web data for the manufacturing plant.*

iii) **Weather data Recording, Monitoring and Prediction (WRMP) Organization**

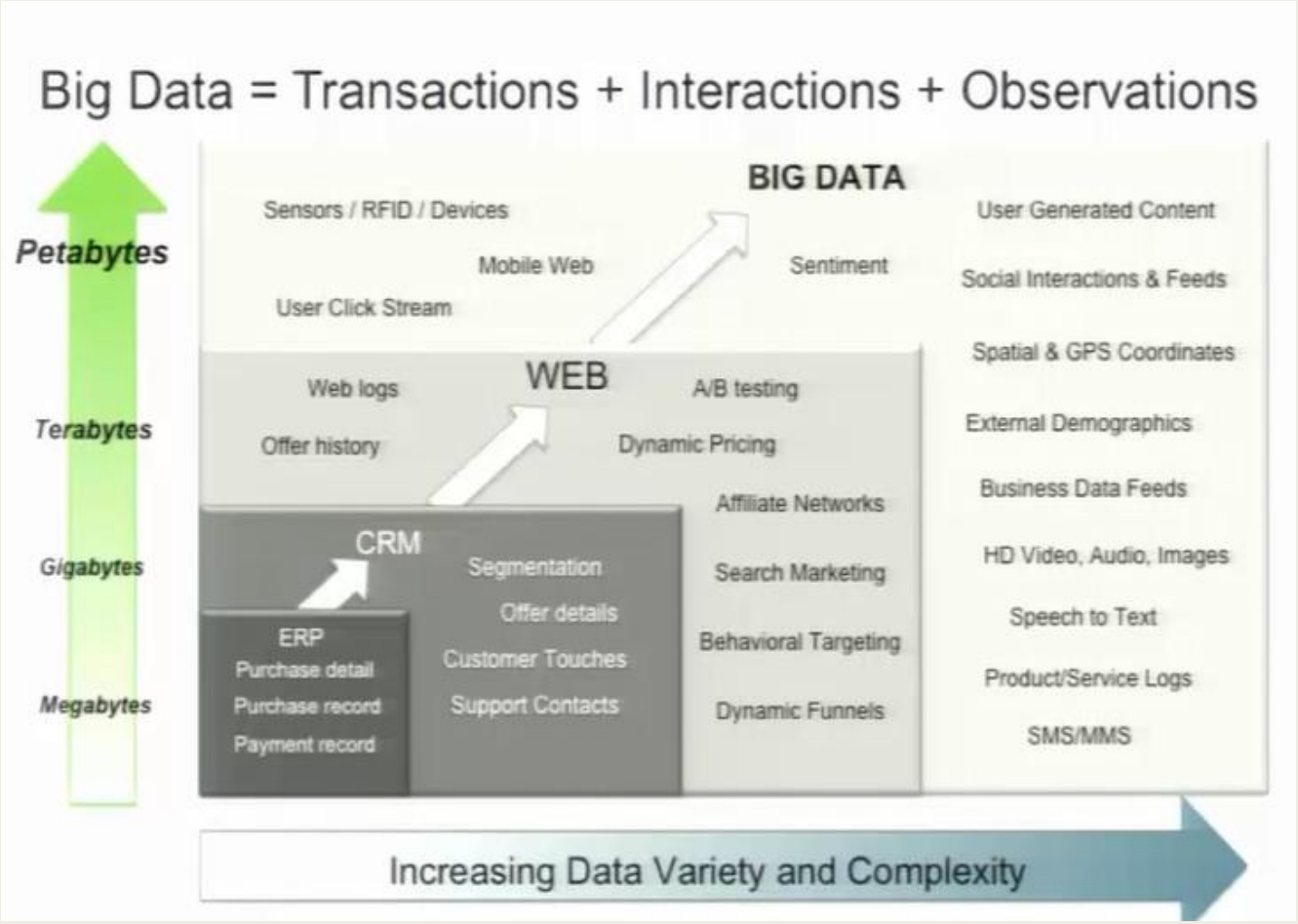
- ***Machine generated data****: from sensors placed at weather stations and satellites, social networks and web data and the* ***reports*** *and* ***alerts*** *issued by many centers around the world.*
- *Stores and processes social and web data collected from* ***other centers****.*
- *Organizations issues maps and weather warnings, predicts weather, rainfall in various regions, expected date of arrival of monsoon.*

Basis of Big Data Classification

- Big data can be classified on the **basis of its characteristics** that are used for designing data architecture for processing and analytics.

1. **Data sources (Traditional):** Data storage such as records, RDBMS, distributed databases, row oriented in-memory data tables, column oriented in-memory data tables, data warehouse, server, machine-generated data, human sourced data, Business Process data, Business Intelligence data.
2. **Data formats (Traditional)** : Structured and Semi-Structured.
3. **Big Data Source:** NoSQL database(MongoDB, Cassandra), Sensors data, audit trial of financial transactions, web, social media, weather data, health record.
4. **Big Data Format:** Unstructured, Semi-Structured and multi-structured data
5. **Data Stores Structure:** Web, Enterprise or cloud server, data warehouse, row oriented OLTP, column oriented OLAP, graph database hashed entries.
6. **Processing rates**

8. **Processing data rates** : Batch, real time, near real time, streaming.
9. **Processing Big Data rates** : High volume, velocity, variety and veracity, Batch, real time, near real time, streaming.
10. **Analysis types**: Batch, real time, near real time dataset analytics.
11. **Big Data Processing method**: Batch processing using MapReduce, real time processing using SparkStreaming and SparkSQL.
12. **Data Analysis Methods**: Statistical, predictive, regression Analysis, machine learning
13. **Data Usage**: Human, Business Process, Knowledge discovery, enterprise applications, Data Stores.


Big Data = Transactions + Interactions + Observations
Petabytes
Terabytes
Gigabytes
Megabytes
BIG DATA
Sensors / RFID / Devices
Mobile Web
User Click Stream
Sentiment
User Generated Content
Social Interactions & Feeds
Spatial & GPS Coordinates
External Demographics
Business Data Feeds
HD Video, Audio, Images
Speech to Text
Product/Service Logs
SMS/MMS
WEB
Web logs
Offer history
A/B testing
Dynamic Pricing
Affiliate Networks
Search Marketing
Behavioral Targeting
Dynamic Funnels
CRM
Segmentation
Offer details
Customer Touches
Support Contacts
ERP
Purchase detail
Purchase record
Payment record
Increasing Data Variety and Complexity

Big Data Handling Techniques

- **Following are techniques deployed for Big Data Storage, applications, data management, mining and analytics:**

1. Huge data volume storage, data distribution, high speed networks and high performance computing.
2. **Applications scheduling** using open source, reliable, scalable, distributed file system, distributed database, parallel and distributed computing system such as Hadoop.
3. **Open source tools** which are scalable, elastic and provide virtualized environment, clusters of data nodes, task an thread management.
4. **Data management using NoSQL**, document database, column oriented, graph database and other form of databases used a per needs of applications and in-memory data management using columnar .
5. **Data mining and analytics**, data retrieval, data visualization and machine learning Big data tools.

Scalability and Parallel Processing

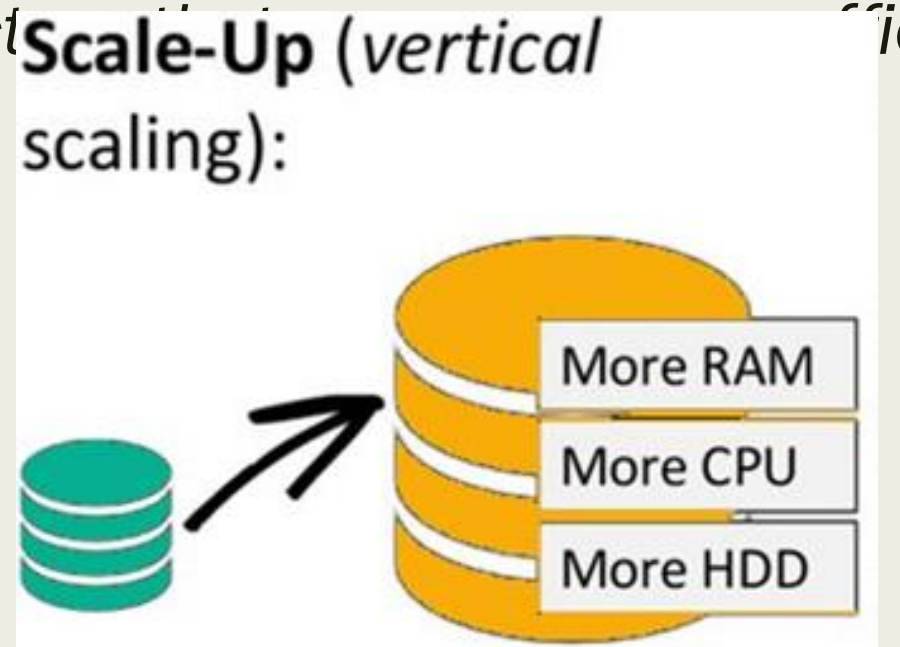
- **Big Data needs:**
  - Process large volume of data (Terabyte and Petabyte)
  - Intensive processing
  - Needs many computing nodes
  - Within short time
  - Minimum Cost
- **Convergence of data environment and Analytics**
  - Big data can co-exist with traditional data store.(RDBMS or Data Warehouse)
  - Scaling up and scaling out, both vertical and horizontal computing resources.

Need of Scaling

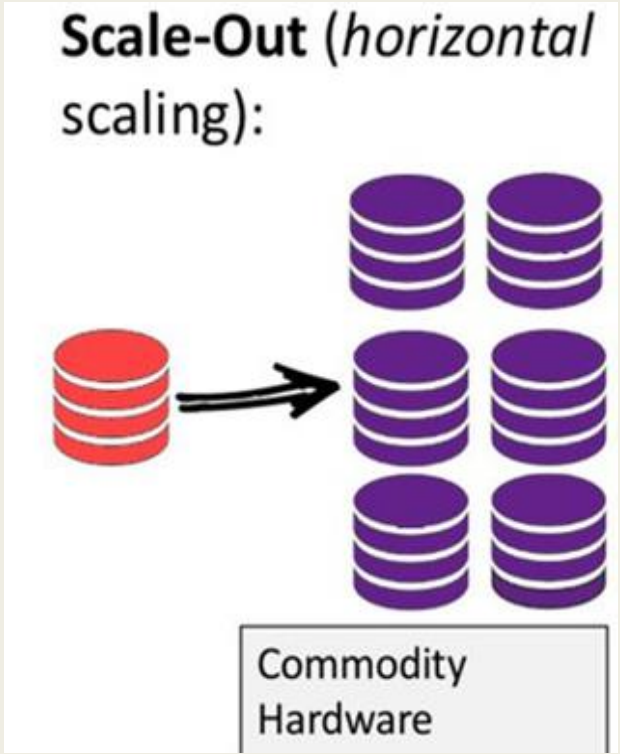
- Scaling can be difficult, but absolutely necessary in the growth of a successful data-driven company.
- **Few signs indications to implement a scaling platform**.
  - When users begin complaining about **slow performance, or service outages**, it's time to scale.
  - In addition to this, increased **application latency**, slow read queries rises and database writes are also important indicators that a scale is needed.
  - Don't wait for the problem to turn into major source of contention in the minds of your customers. This can have a massively negative impact on retaining those customers.
  - If possible, try to anticipate the problem before it becomes severe.

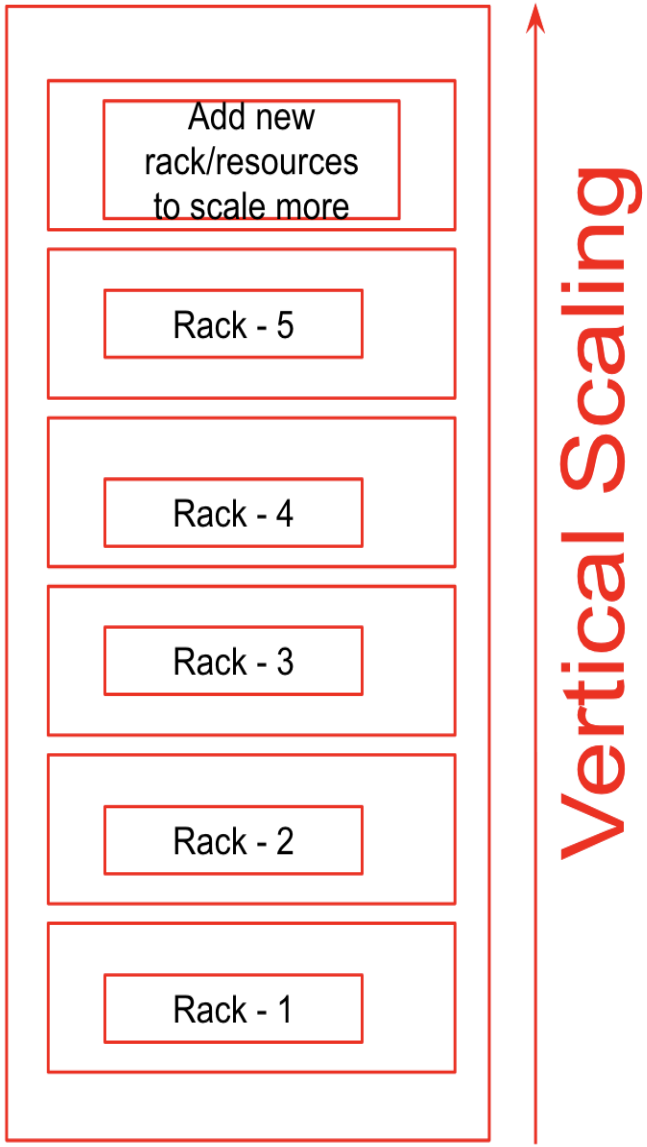
Analytics Scalability to Big Data

- **Scaling up / Vertical Scalability:**
  - *Scaling up the given system resources and increasing the system analytics reporting and visualization capabilities.*
  - *Design Architect[illegible]ficiently.*
  - *Ex:*



- **Scaling out / Horizontal Scalability:**
  - Increasing the number of systems working in **coherence** and scaling out the workload.
  - Use more resources and distribute the processing and storage tasks in parallel.
  - If r resources in a system process x terabyte of data in time t, then p*x terabyte process on parallel distributed nodes such that time taken up remains t or slightly more then t.



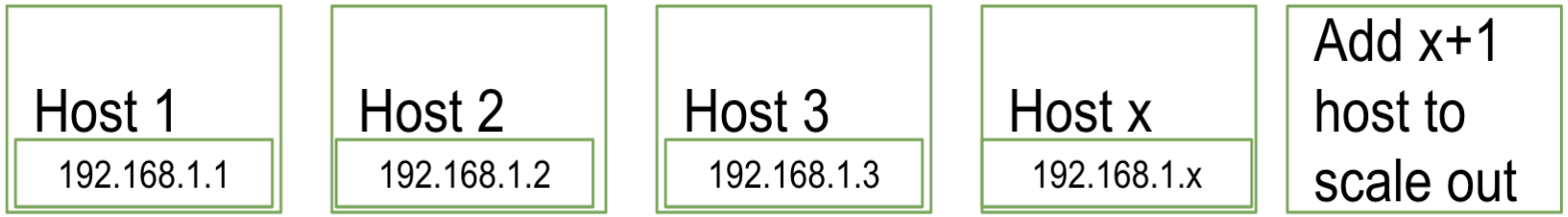


Vertical Scaling

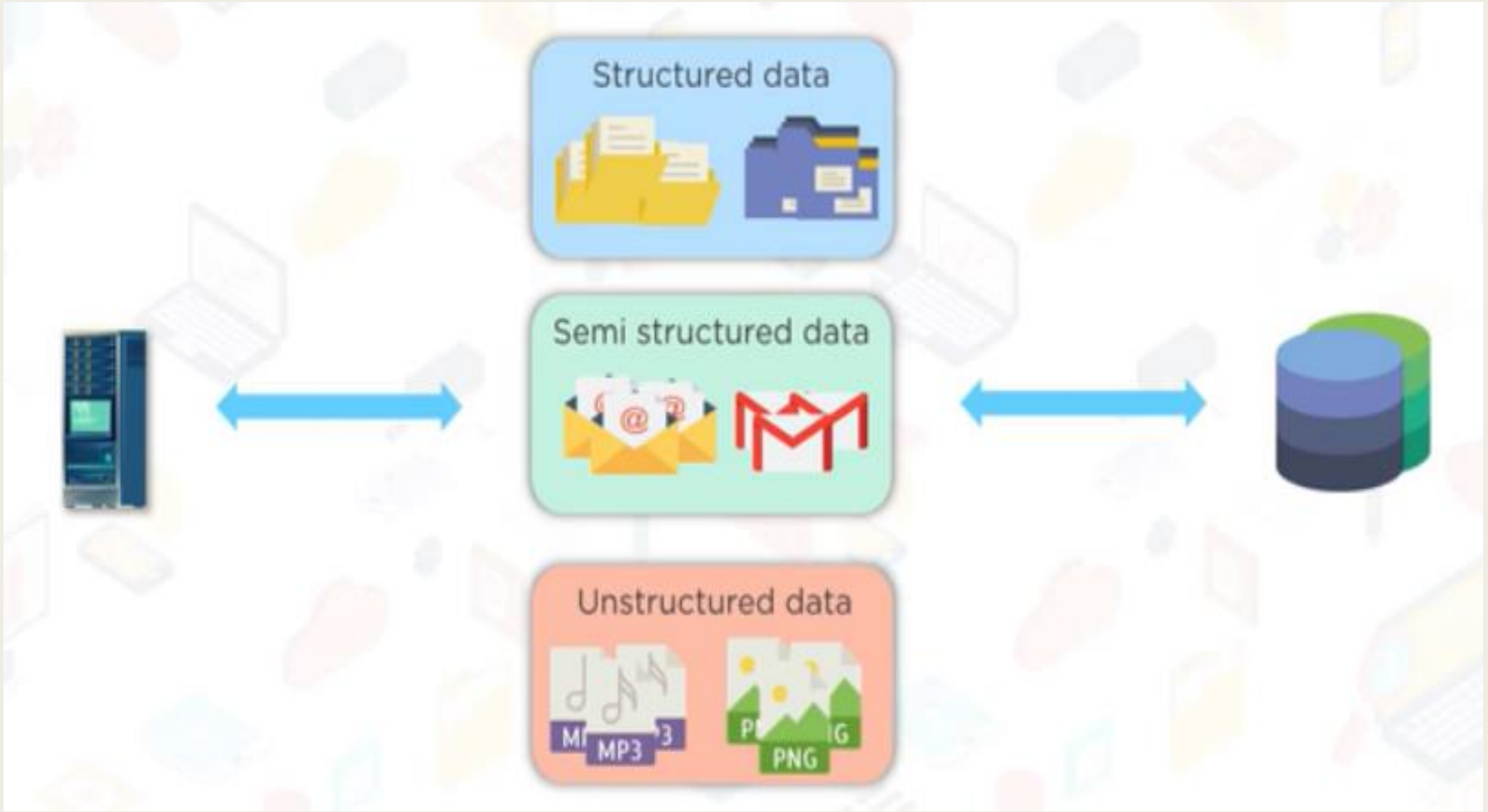
To scale more, Add more RAM, CPU, Memory to the **one existing machine**

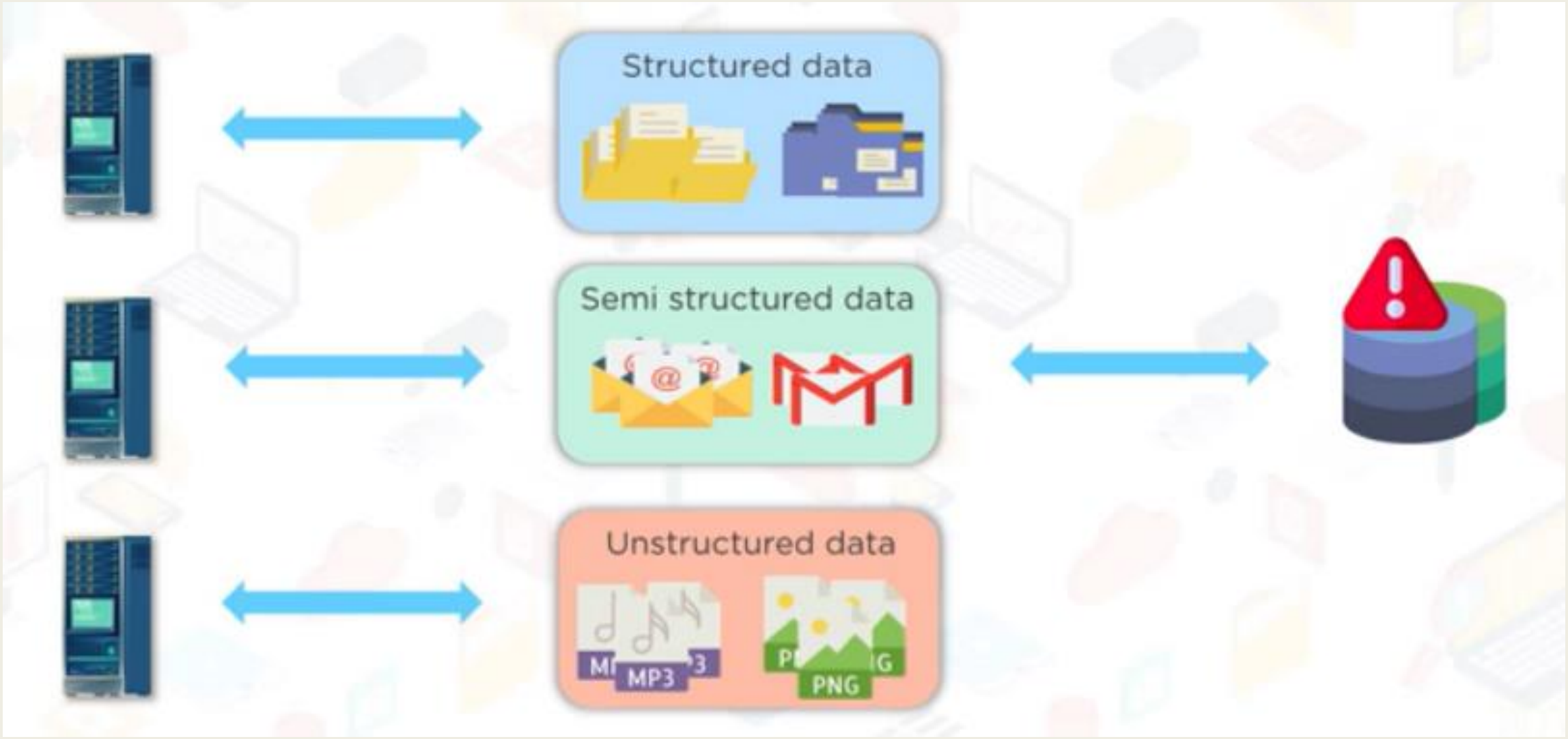
Horizontal Scaling

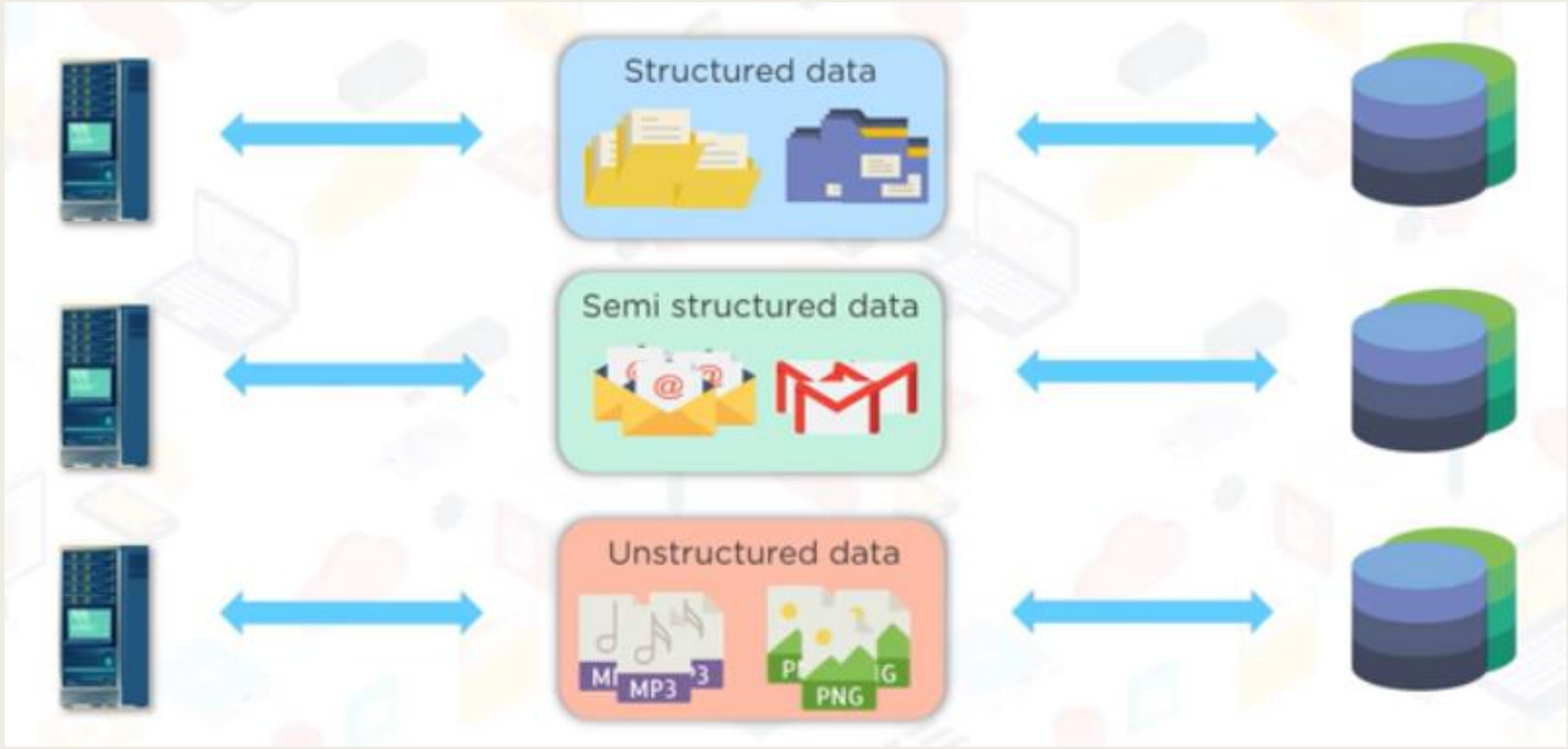
To scale more: Add more machines to existing **group of distributed system**

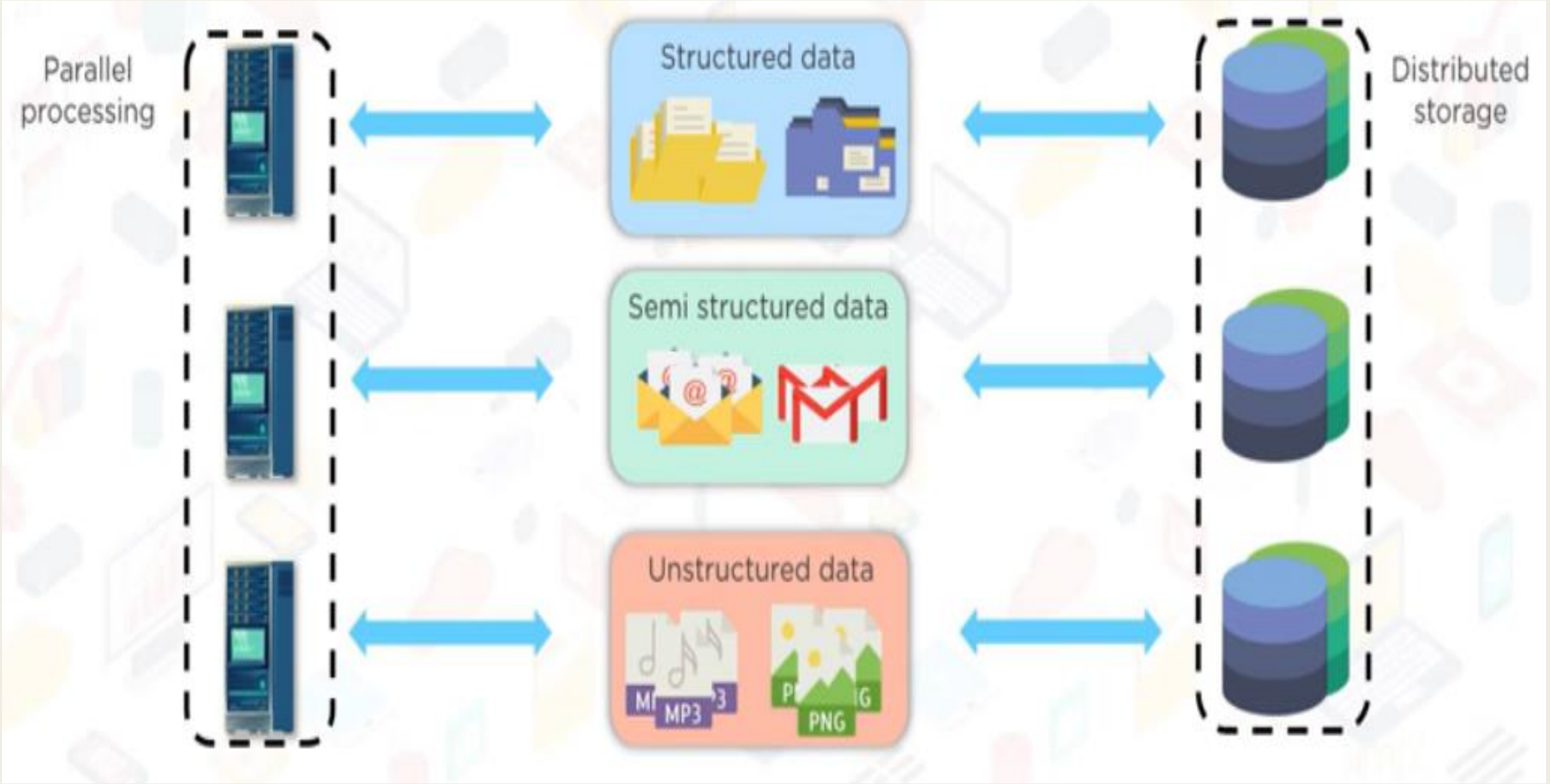


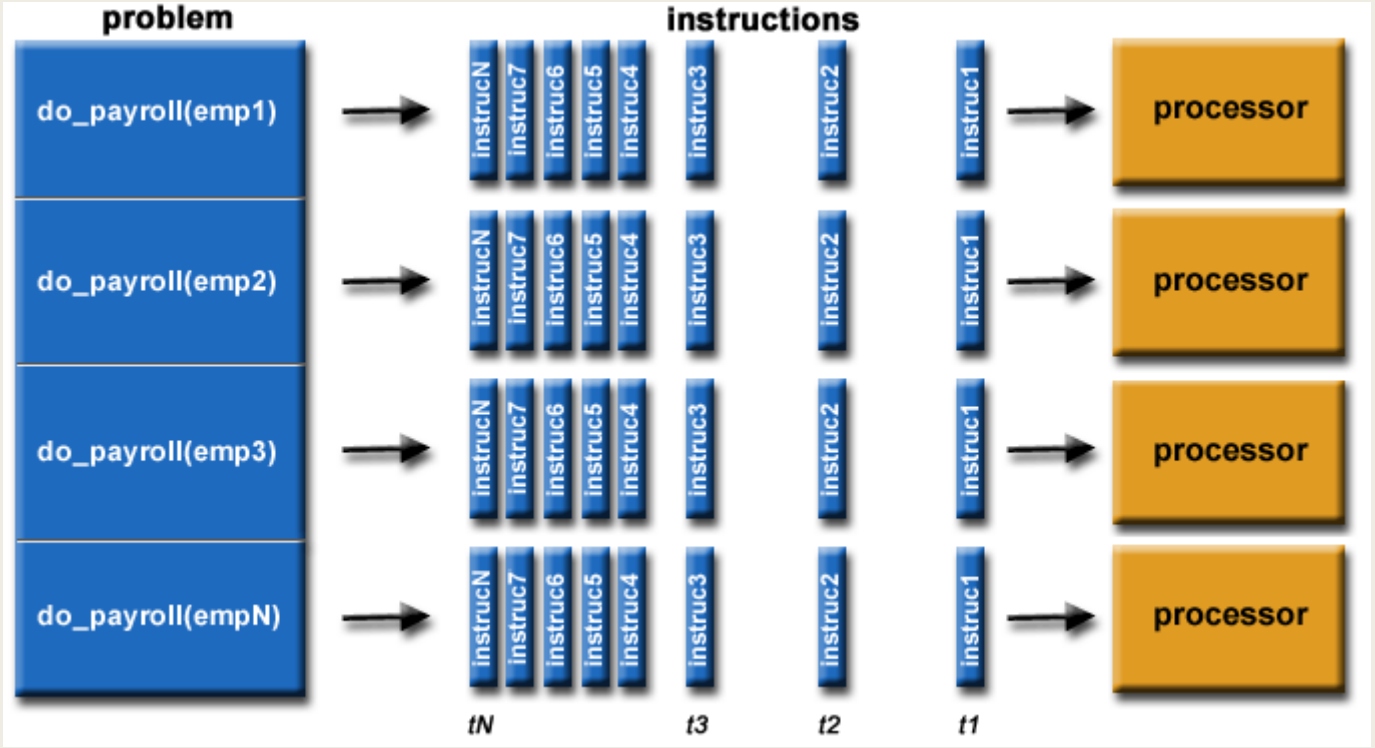
- The **easiest way** to scale up and scale out execution of analytics software is to implement it on a bigger machine with more CPU's for greater 3V's and complexity of data
- Software will perform better on bigger machines.
- Expensive compared to extra performance achieved by efficient design of algorithms.
- If more CPU's add in a computer, but the software does not exploit the advantage of them, then that will not get any increased performance out of the additional CPU.
- Alternatively use Massively Parallel Platforms, Cloud, Grid. Clusters and distributed computing and analysis.


Structured data
Semi structured data
Unstructured data
MP3
PNG

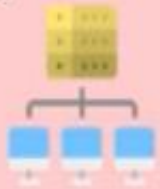
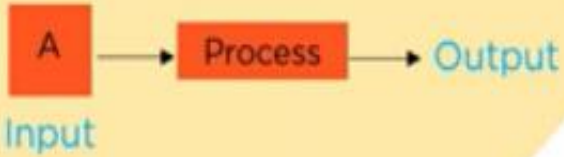
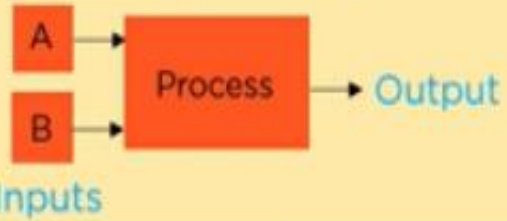

Structured data
Semi structured data
Unstructured data
MP3
PNG

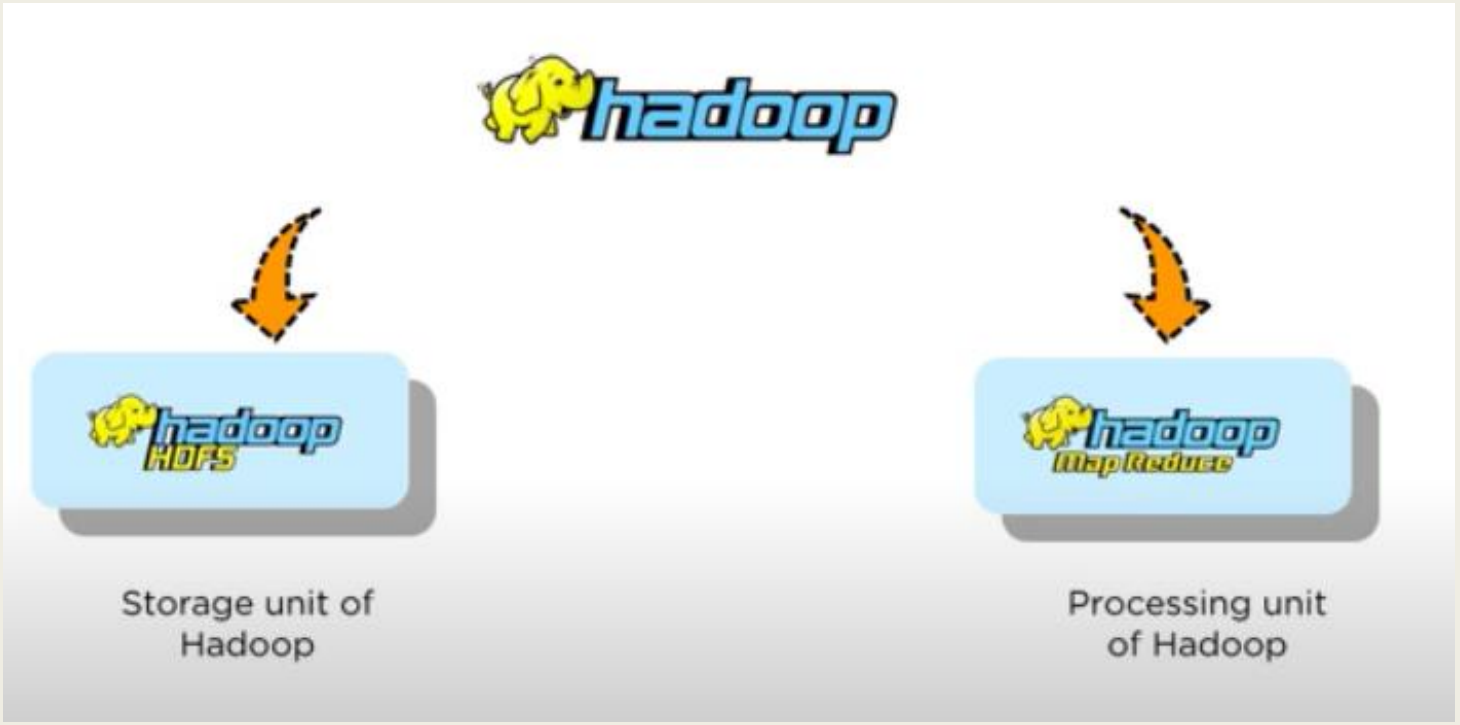

Structured data
Semi structured data
Unstructured data
MP3
PNG


Parallel processing
Structured data
Semi structured data
Unstructured data
MP3
PNG
Distributed storage


problem
instructions
do_payroll(emp1)
do_payroll(emp2)
do_payroll(emp3)
do_payroll(empN)
instrucN
instruc7
instruc6
instruc5
instruc4
instruc3
instruc2
instruc1
processor
tN
t3
t2
t1

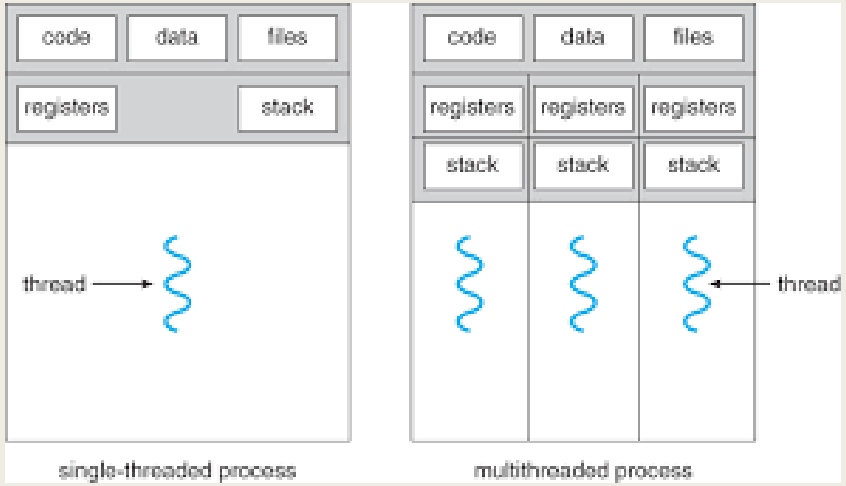
Big Data challenges and solution

Challenges	Solutions
Single central storage 	Distributed storage 
Serial processing 	Parallel processing 
Lack of ability to process unstructured data 	Ability to process every type of data 

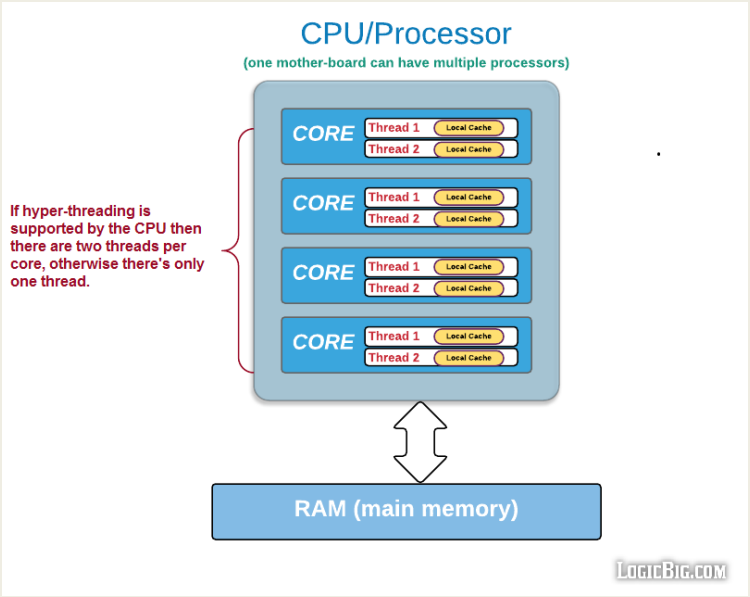

hadoop
hadoop HDFS
hadoop Map Reduce
Storage unit of Hadoop
Processing unit of Hadoop

Massively Parallel Processing Platforms

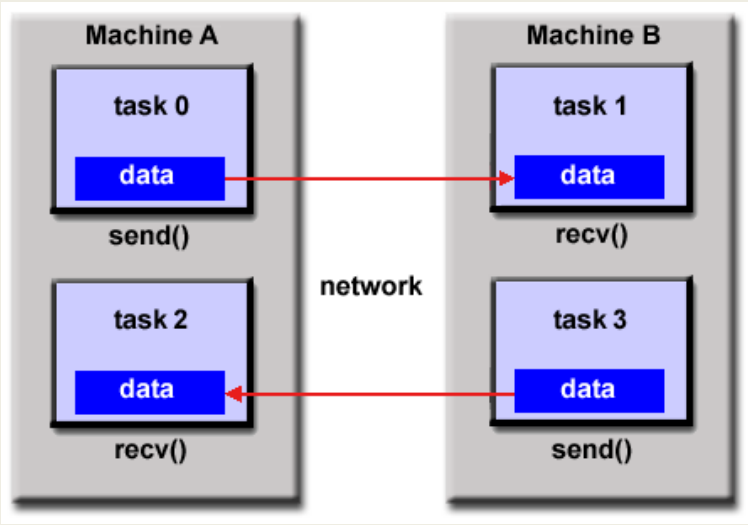
- Large or complex programs are impractical or impossible to execute on single computer with limited memory.
- So enhance or scale up the computer or use massive parallel processing(MPP) platforms.
- Parallelization of tasks can be done at several levels:



Distributing separate task onto separate thread on the same CPU



Distributing separate task onto separate CPU on same Computer

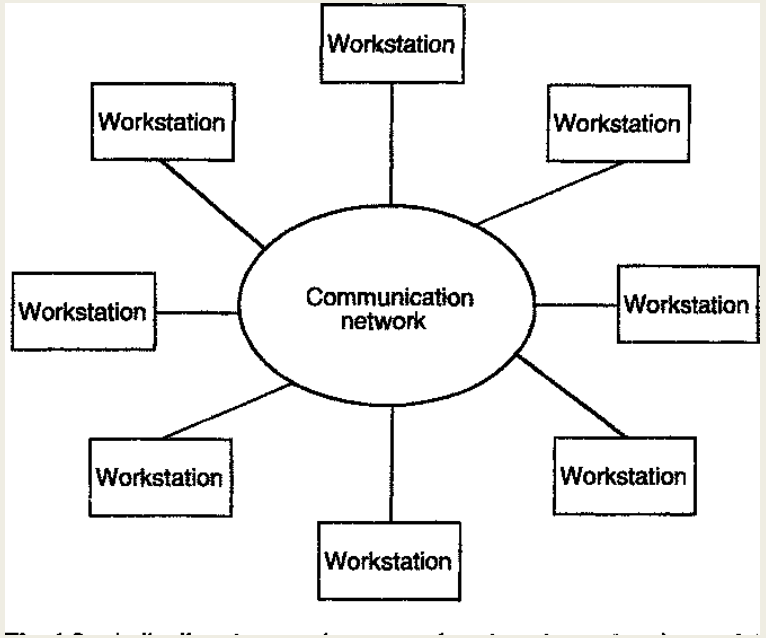


Distributing separate tasks onto separate computers.

- The software must draw the advantage of multiple computers and should be able to parallelize tasks.
- Computational problem is broken into discrete pieces of sub-tasks that can be processed simultaneously.
- Total time taken will be much less than with a single compute resource.

Distributed Computing Model

- Distributed computing is a **model in which components of a software system are shared among multiple computers**. Even though the components are spread out across multiple computers, they are run as one system. This is done in order to improve efficiency and performance.
- They process and analyze big and large datasets on distributed computing nodes connected by high-speed networks.
- Distributed computing model uses cloud, grid or clusters.



Cloud Computing

- Cloud Computing is a type of Internet-based computing that provides shared processing resources and data to the computers and other devices on demand.
- Cloud computing is named as such because **the information being accessed is found remotely in the cloud or a virtual space**. Companies that provide cloud services enable users to store files and applications on remote servers and then access all the data via the Internet.
- Approach– Parallel and distributed computing in a cloud environment.
- Multiple nodes perform automatically and interchangeably to offer high data security compared to other distributed technologies.

Cloud Computing Features:

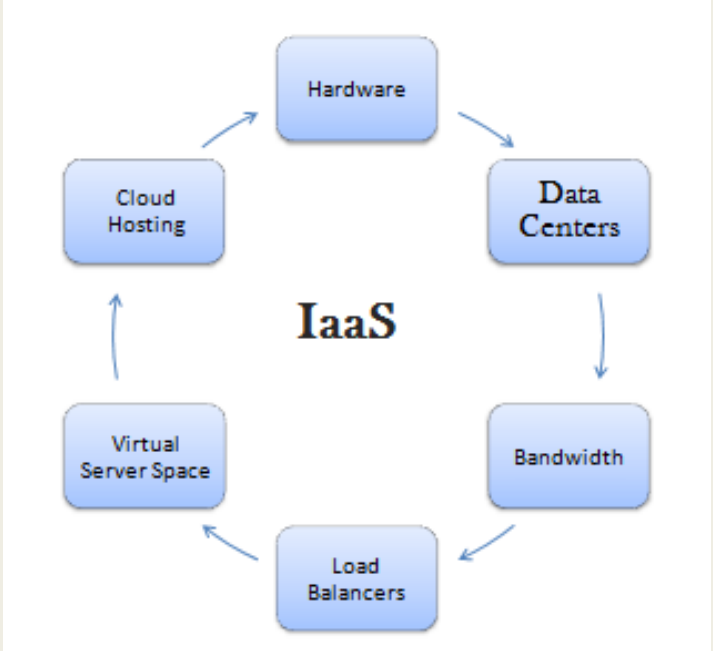
- On-demand Service
- Resource Pooling
- Scalability
- Accountability
- Broad Network Access
- Cloud services can be accessed from anywhere and at any time through Internet.
- Local private cloud can be setup on local cluster of computers.

Cloud Resources

1. **Amazon Web Services or AWS** as an abbreviation is a popular **Cloud Service Provider** that enables on-demand services like compute, storage, networking, security, databases, etc which can be accessed through the internet across the globe and the user is not required to manage or monitor these resources.
2. Elastic Compute Cloud EC2 : Amazon Elastic Compute Cloud is a part of Amazon.com's cloud-computing platform, Amazon Web Services, that allows users to rent virtual computers on which to run their own computer applications.
3. Microsoft Azure or Apache CloudStack : Microsoft Azure, commonly referred to as Azure, is a cloud computing service created by Microsoft for building, testing, deploying, and managing applications and services through Microsoft-managed data centers.
4. Amazon Simple Storage Service (S3) : Amazon S3 or Amazon Simple Storage Service is a service offered by Amazon Web Services that provides object storage through a web service interface. Amazon S3 uses the same scalable storage infrastructure that Amazon.com uses to run its global e-commerce network.

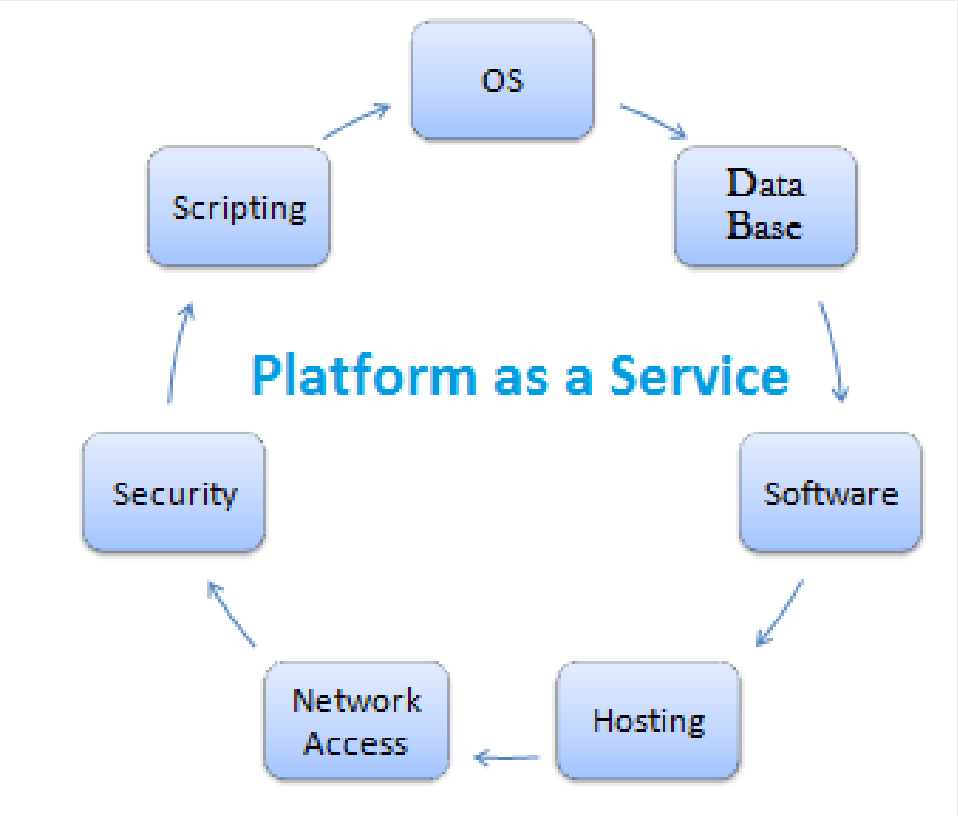
Classification of Cloud Services:

- **Infrastructure as a Service (IaaS):**
  - Provide access to hard disk, network connections, databases storage, data center and virtual server space.
  - Ex: Tata Communications : Tata Communications is **a digital ecosystem enabler that powers today's fast-growing digital economy**. Amazon Data Centers
  - Virtual Servers.
  - Apache Cloudstack: Open Source software for deploying and managing a large network of virtual machines.



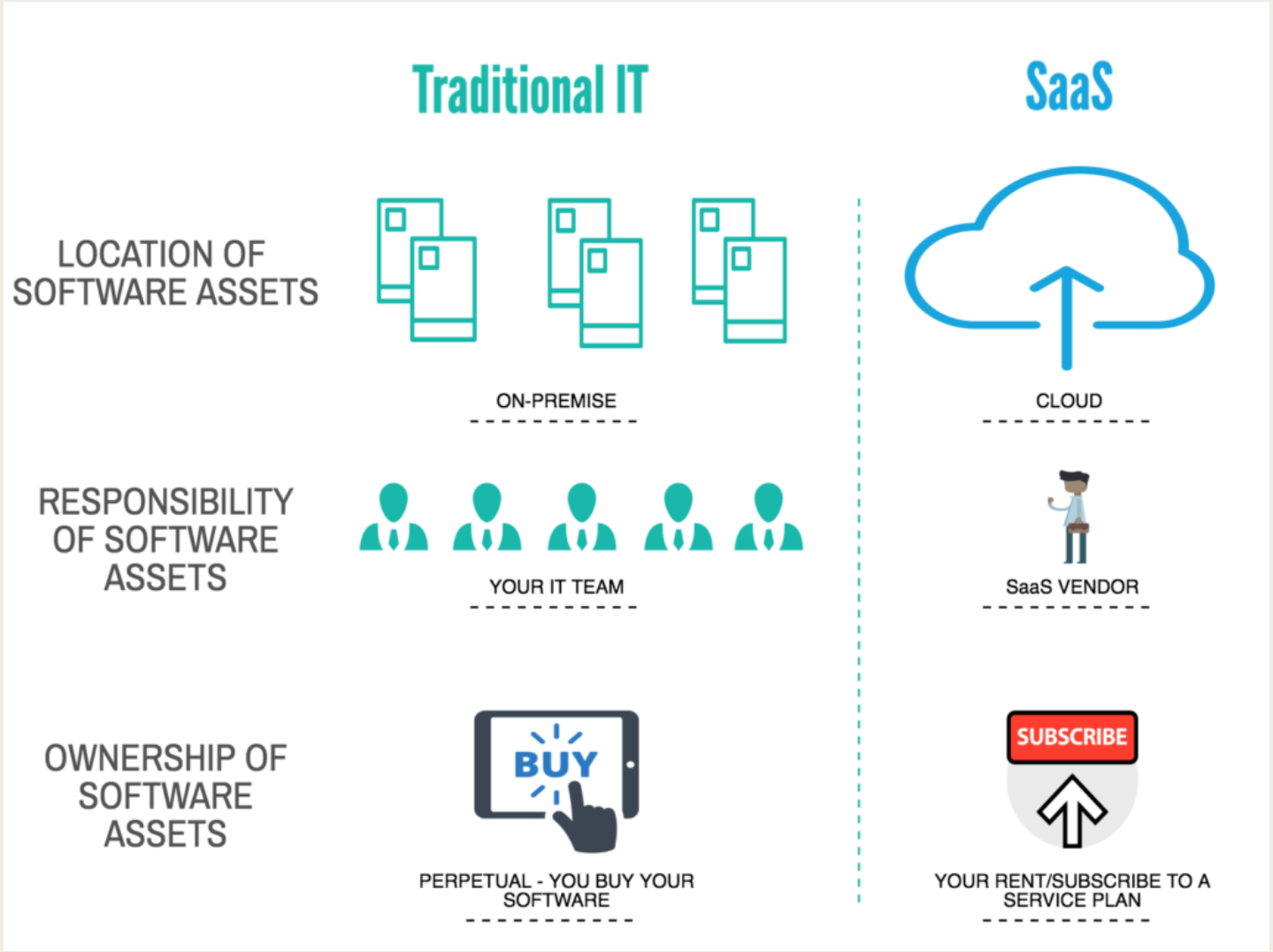
Platform as a Service (PaaS)

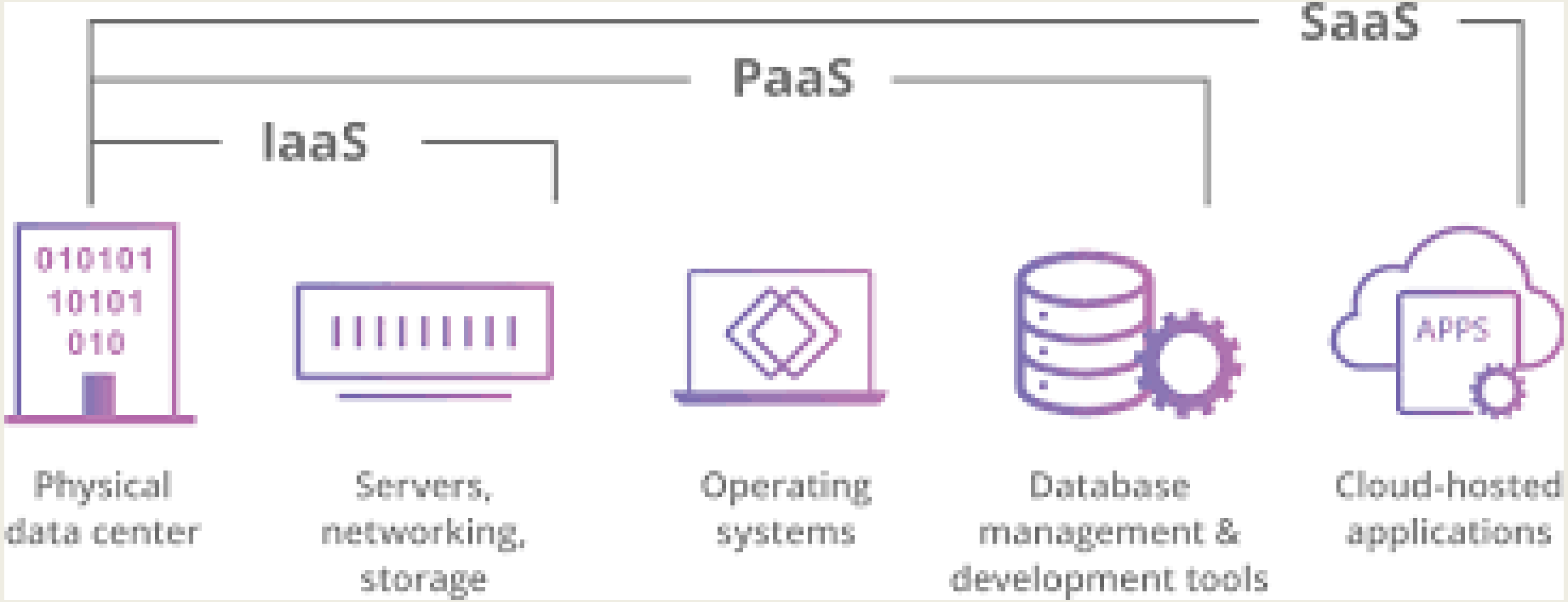
- Providing the runtime environment to allow developers build applications and services.
- Software at cloud support and manage the services, storage, networking, deploying, testing, collaborating, hosting and maintaining applications.
- Ex: Hadoop Cloud Service

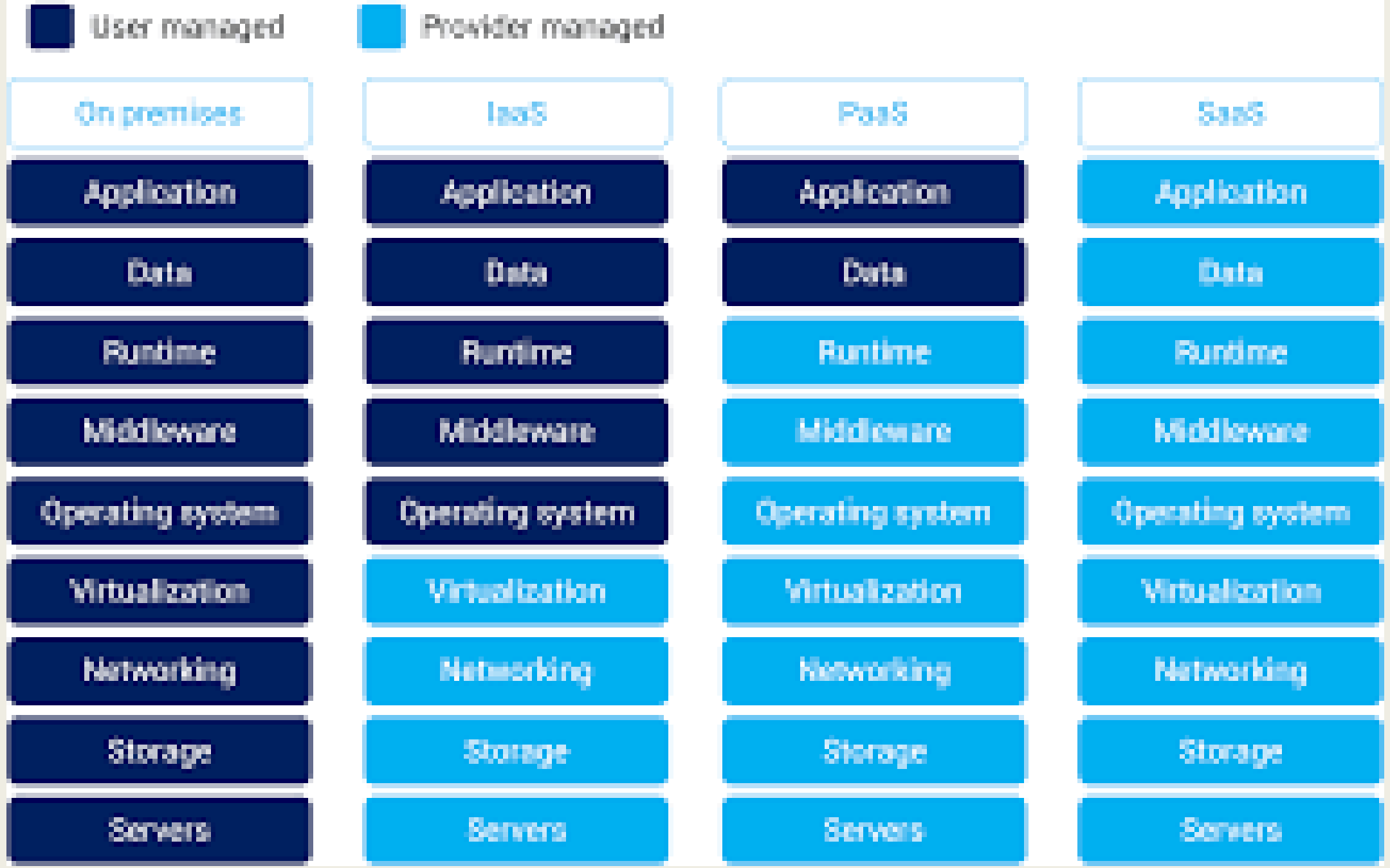


Software as a Service (SaaS)

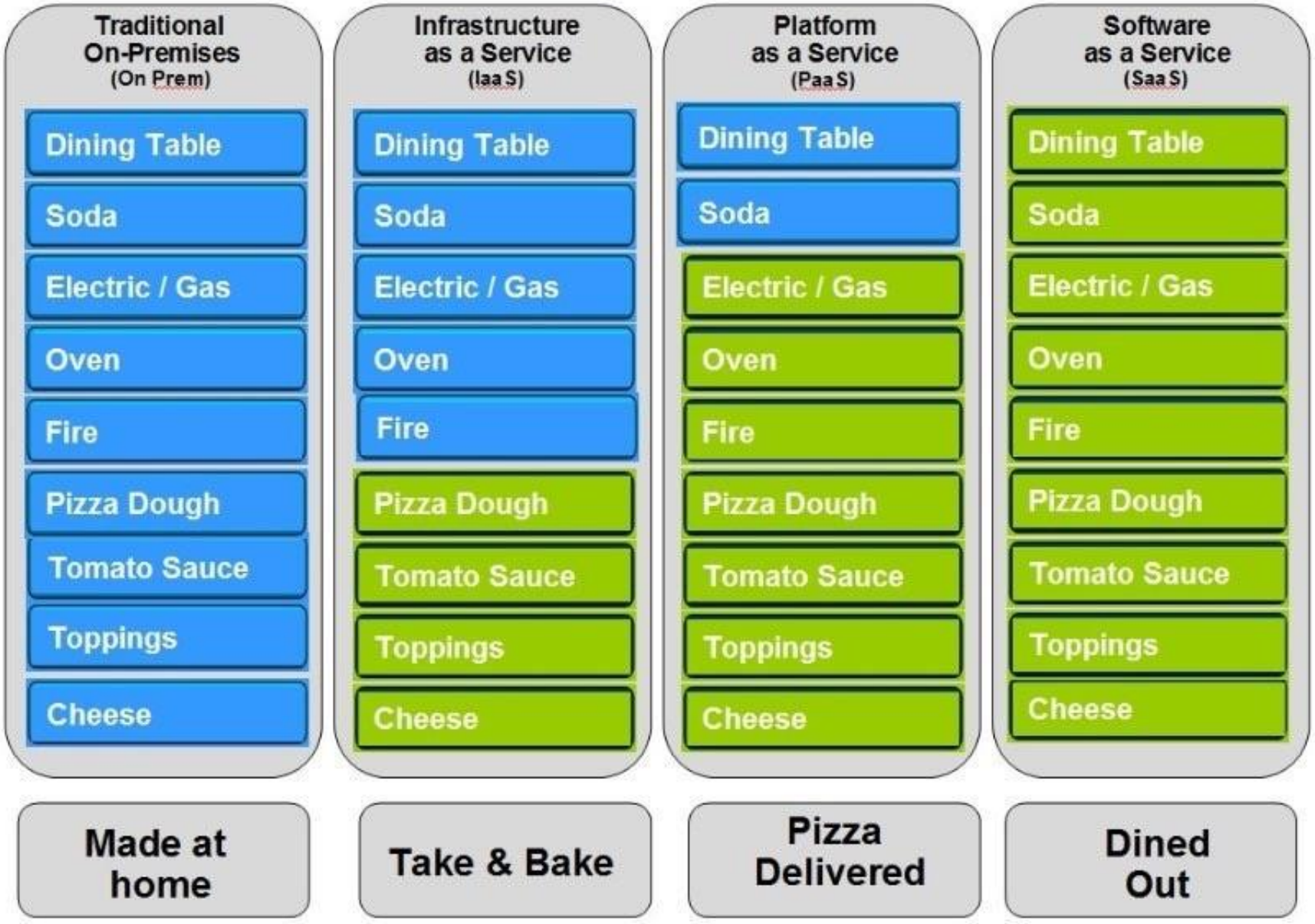
- Provide software applications as service to end-user.
- Software applications are hosted by service providers and made available to customer over Internet.
- Ex: GoogleSQL, IBM BigSQL

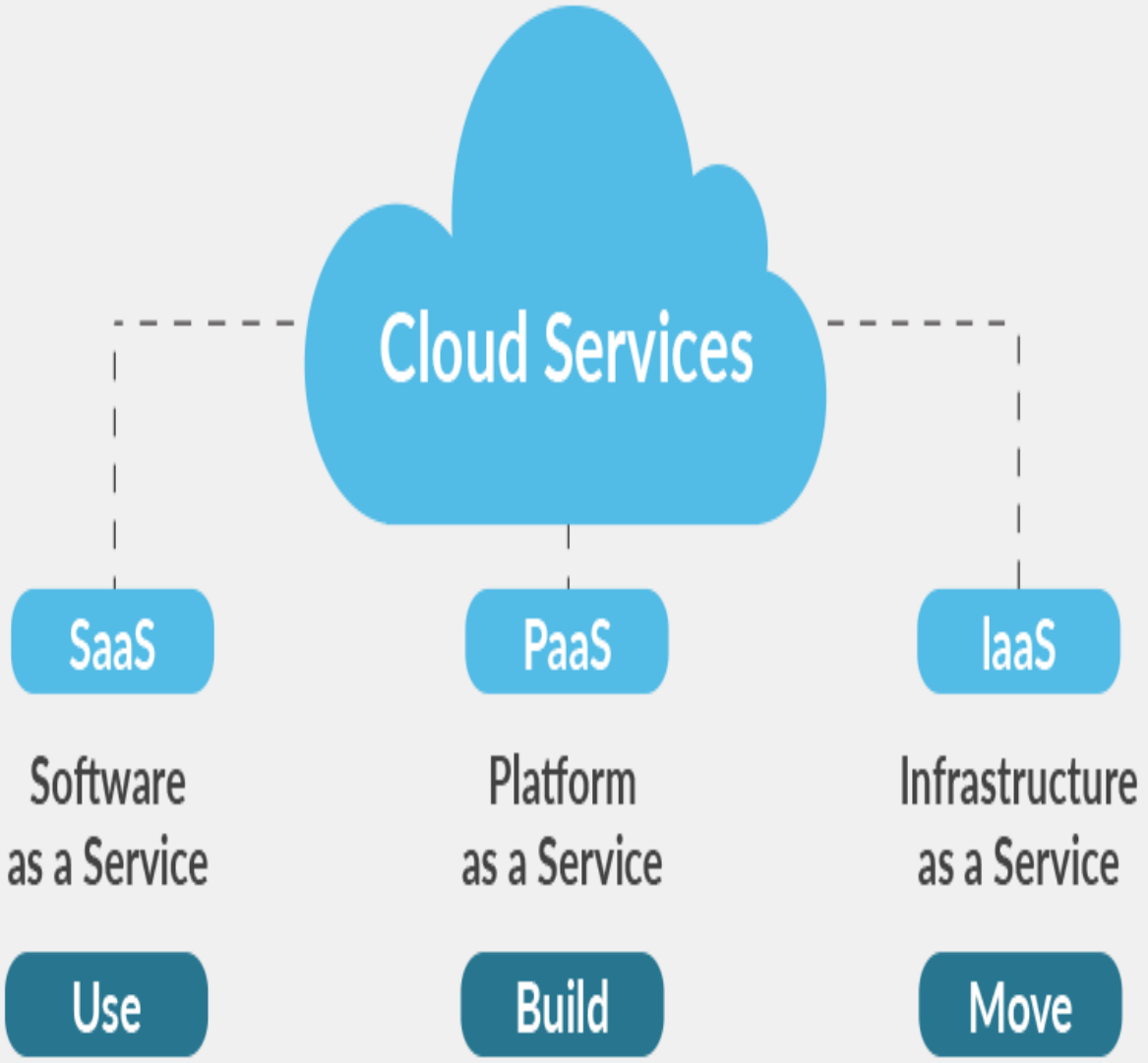


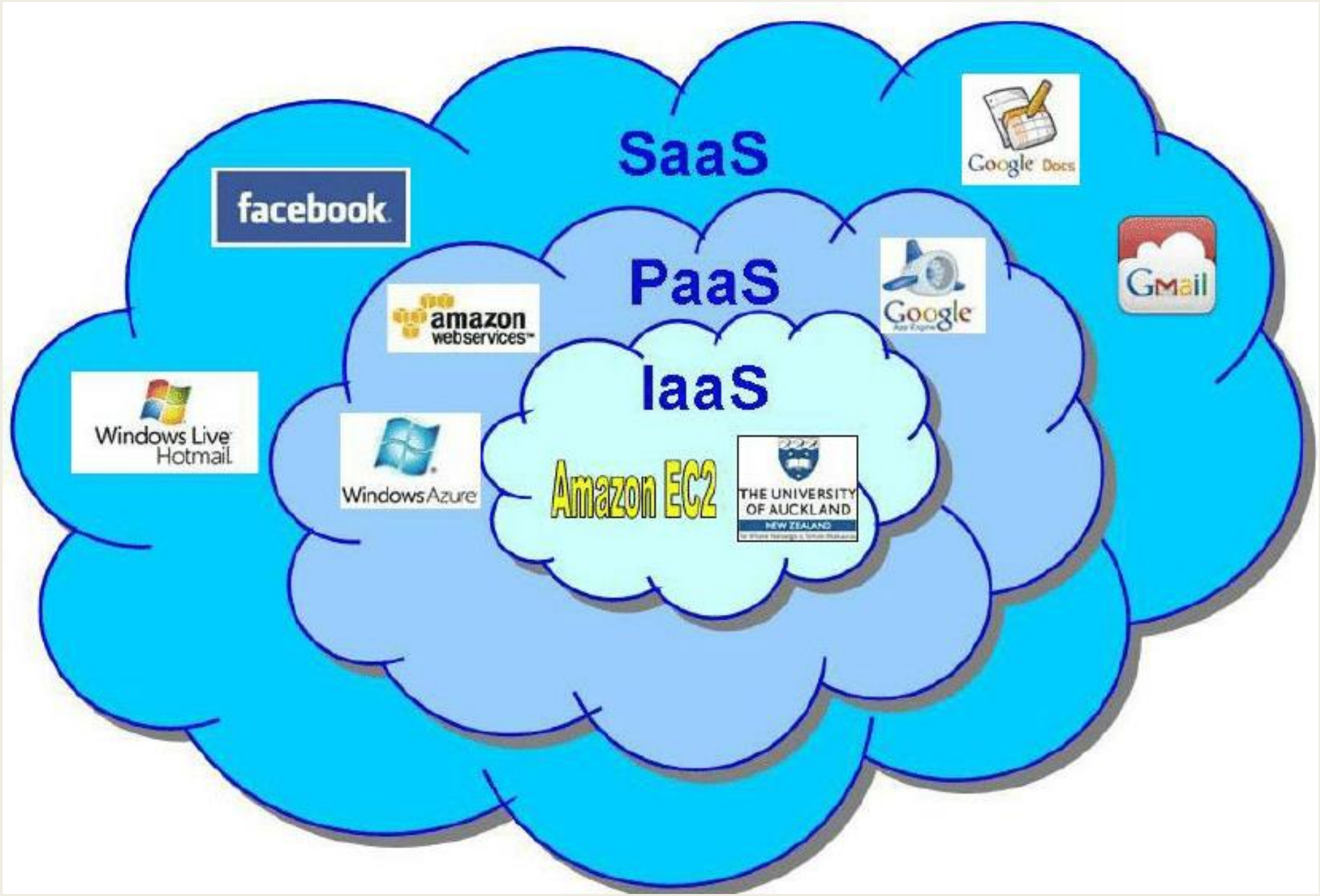

SaaS
PaaS
IaaS
010101
10101
010
APPS
Physical data center
Servers, networking, storage
Operating systems
Database management & development tools
Cloud-hosted applications


User managed
Provider managed
On premises
IaaS
PaaS
SaaS
Application
Data
Runtime
Middleware
Operating system
Virtualization
Networking
Storage
Servers

Pizza as a Service

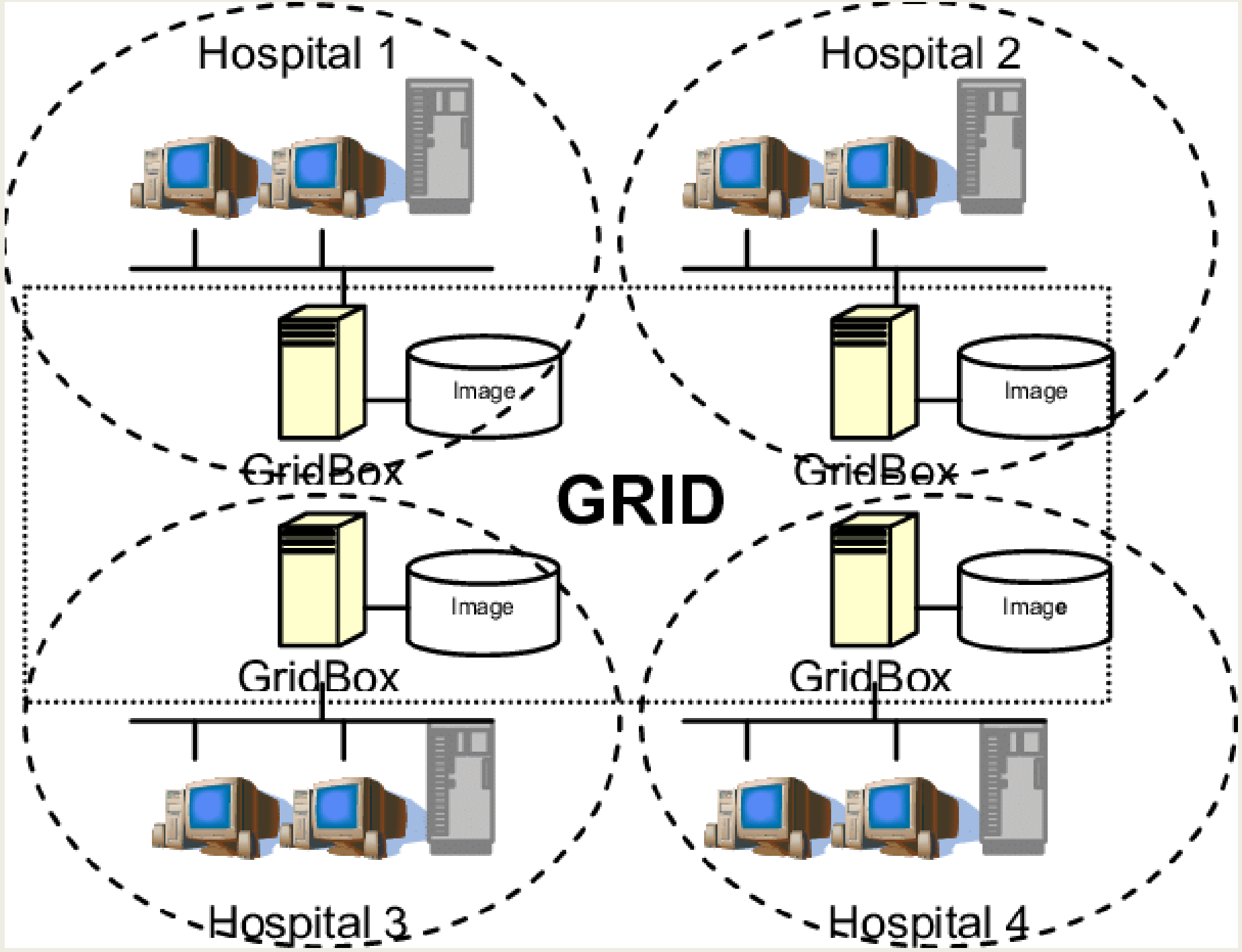



Cloud Services
SaaS
PaaS
IaaS
Software as a Service
Platform as a Service
Infrastructure as a Service
Use
Build
Move


SaaS
PaaS
IaaS
facebook
Google Docs
Gmail
amazon web services
Google
Windows Live Hotmail
Windows Azure
Amazon EC2
THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

1.3.4 Grid and Cluster Computing

- Grid Computing
  - Group of computers from several locations are **connected** with each other to achieve a common task.
  - Distributed computing
  - Resources are **heterogenous** and **geographically** disperse.
  - Provide **large scale resource sharing** which is flexible, coordinated and secure among its users.
  - User consists of individual, organization and resources.
  - Suitable for **data-intensive storage.**
  - **Computational grid** focuses on computationally intensive operations.


Hospital 1
Hospital 2
Image
Image
GridBox
GridBox
GRID
Image
Image
GridBox
GridBox
Hospital 3
Hospital 4

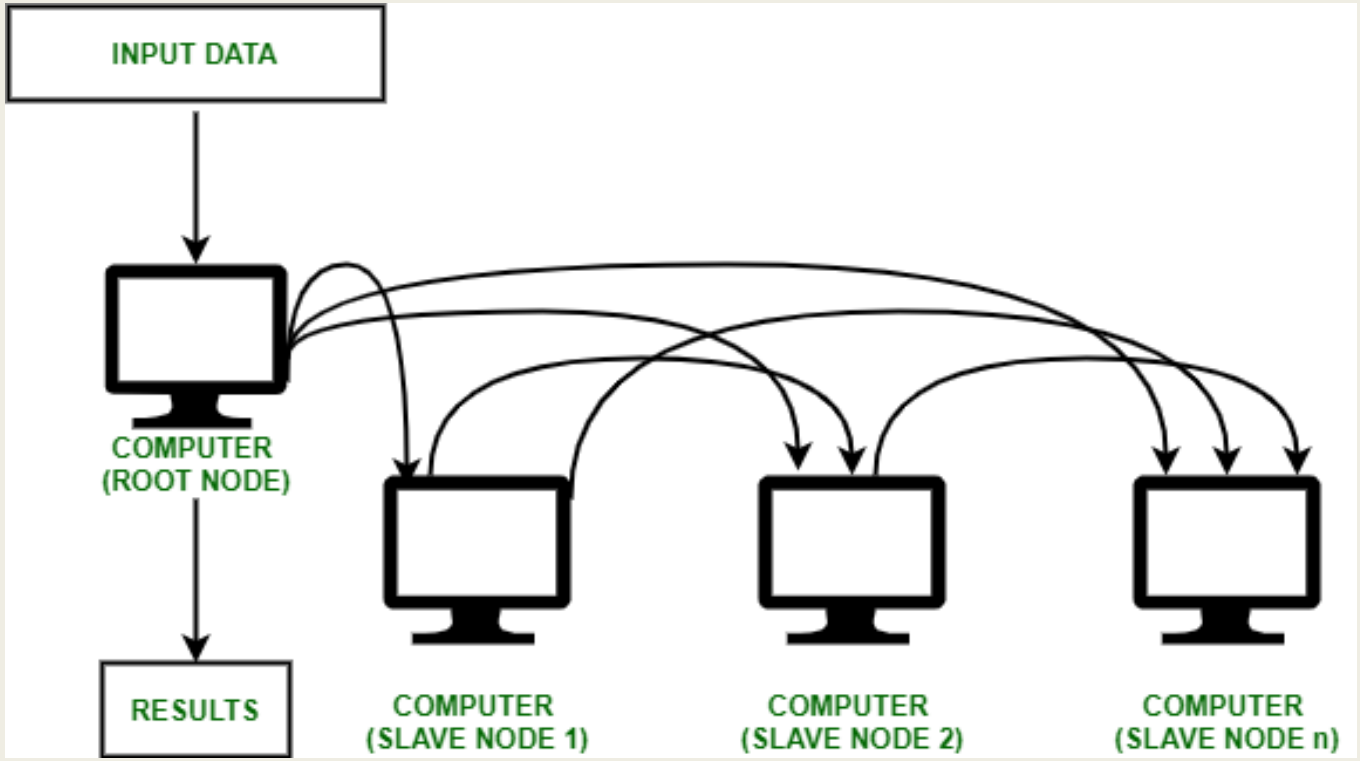
- **Grid Features**
  - It is **scalable**
  - Resource sharing to attain **coordination** and **coherence** among resources
  - Forms **distributed** network for resource sharing
- **Drawbacks**
  - Failure in case of **underperformance** or failure of any of the participating nodes.
  - Systems **storage capacity varies** with the number of users, instances and amount of data transferred at a given time.
  - Sharing resources among large number of users helps in **reducing infrastructure costs and raising load capacities**.

Grid Computing Vs Cloud Computing

Cloud Computing	Grid Computing
Cloud Computing follows client-server computing architecture.	Grid computing follows a distributed computing architecture.
Scalability is high.	Scalability is normal.
Cloud Computing is more flexible than grid computing.	Grid Computing is less flexible than cloud computing.
Cloud operates as a centralized management system.	Grid operates as a decentralized management system.
In cloud computing, cloud servers are owned by infrastructure providers.	In Grid computing, grids are owned and managed by the organization.
Cloud computing uses services like Iaas, PaaS, and SaaS.	Grid computing uses systems like distributed computing, distributed information, and distributed pervasive.
Cloud Computing is Service-oriented.	Grid Computing is Application-oriented.

Cluster Computing

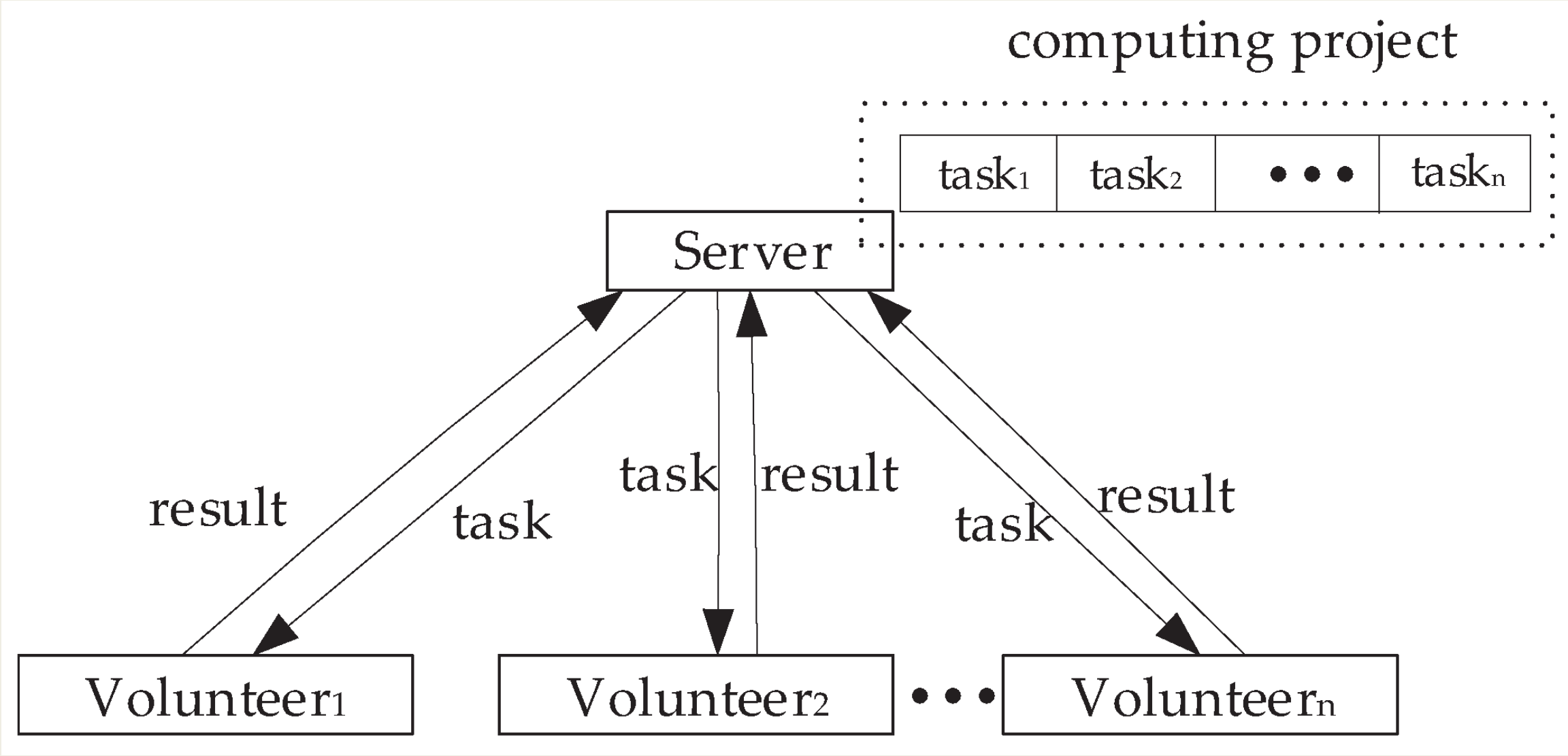
- Cluster computing is **a collection of tightly connected computers that work together so that they act as a single entity**
- A group of computers connected by a network
- The group works together to accomplish the same task.
- Used for **load balancing.**
- Shift processes between nodes to keep an even load on the group of connected computers.



Distributed Computing	Cluster Computing	Grid Computing
Loosely Coupled	Tightly Coupled	Large Scale
Heterogeneous	Homogeneous	Cross Organizational
Single Administration	Cooperative working	Geographical distribution and Distributed management

Volunteer Computing

- Volunteer computing is distributed computing paradigm which uses computing resources of volunteers.
- Volunteers are organizations or members who own personal computers.
- Provide computing resources to projects of importance that use resources to do distributed computing and storage.
- Projects examples are science-related projects executed by universities or academia in general.
- Volunteer computing is **a type of distributed computing in which people donate their computers' unused resources to a research-oriented project**.


computing project
task1
task2
• • •
taskn
Server
result
task
task
result
task
result
Volunteer1
Volunteer2
• • •
Volunteern

1.4 DESIGNING DATA ARCHITECTURE

- The designing of Big Data architecture layers and how to manage data for analytics

1.4.1 Data Architecture Design

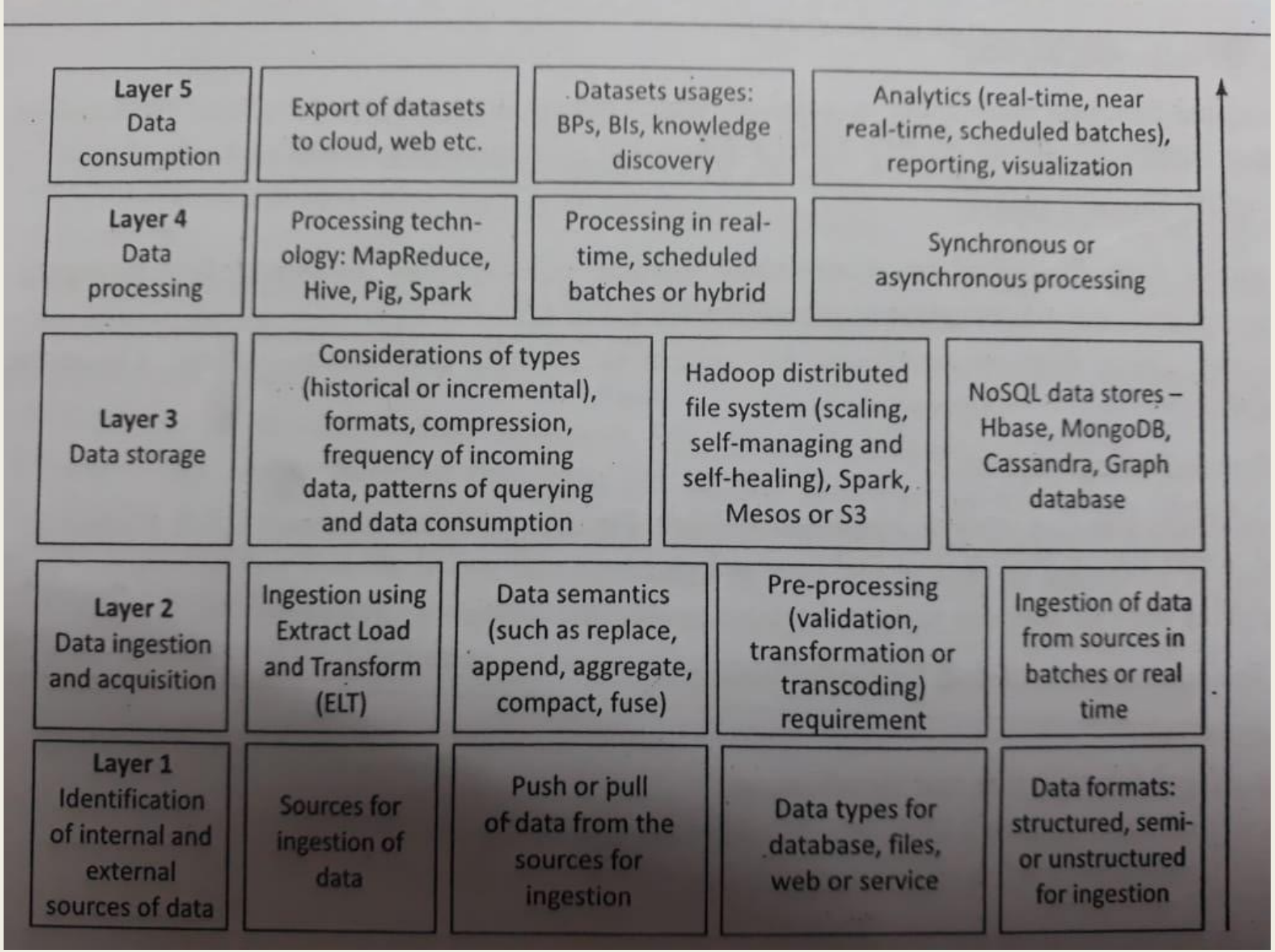
Techopedia defines Big Data architecture as follows:

"Big Data architecture is the logical and/or physical layout/structure of how Big Data will be stored, accessed and managed within a Big Data or IT environment. Architecture logically defines how Big Data solution will work, the core components (hardware, database, - software, storage) used, flow of information, security and more."

Challenges in designing Big Data Architecture

- Characteristics of Big Data make designing Big Data architecture a complex process.
- Faster additions of new technological innovations increase the complexity in design.
- The requirements for offering competing products at lower costs in the market make the designing task more challenging for a Big Data architect.

- Data analytics need the number of **sequential steps.**
- Big Data architecture design task simplifies when using the **logical layers approach.**
- Figure shows the **logical layers and the functions** which are considered in Big Data architecture.



Layers

- Data processing architecture consists of five layers:

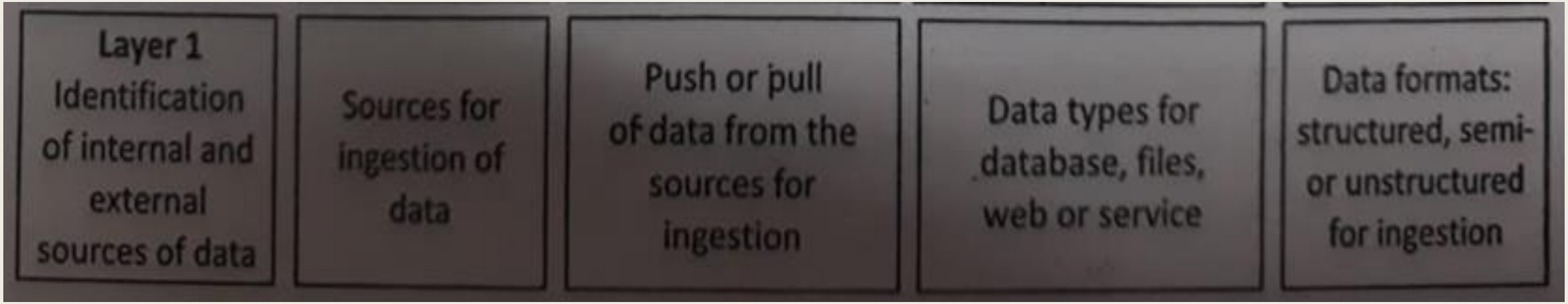
1. Identification of data sources,
2. Acquisition, ingestion, extraction, pre-processing, transformation of data,
3. Data storage at files, servers, cluster or cloud,
4. Data-processing, and
5. Data consumption in the number of programs and tools.

Data consumed for applications, such as

- Business intelligence
- Data mining
- Discovering pattern/clusters
- Artificial intelligence (AI)
- Machine learning (ML)
- Text analytics
- Descriptive and predictive analytics
- Data visualization.

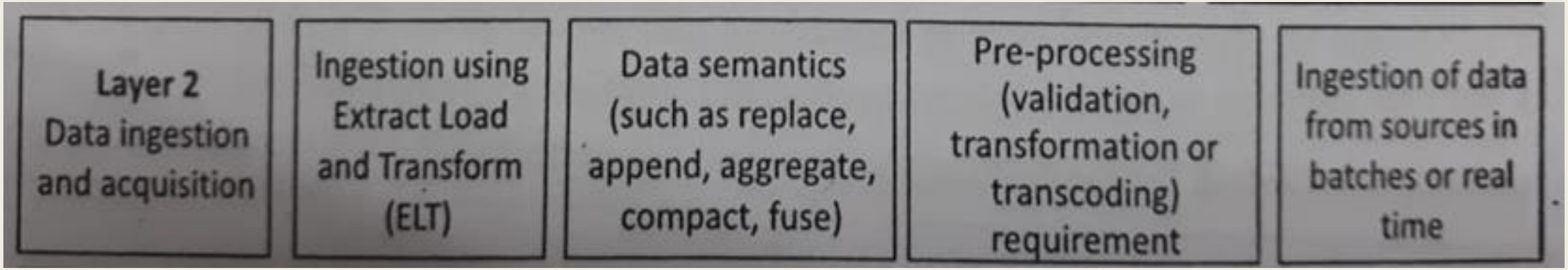
- Logical layer 1 (L1) is for identifying data sources, which are external, internal or both. The layer 2 (L2) is for data-ingestion.
- **L1 considers the following aspects in a design:**

1. Amount of data needed at ingestion layer 2 (L2)
2. Push from L1 or pull by L2 as per the mechanism for the usages
3. Source data-types: Database, files, web or service
4. Source formats, i.e., semi-structured, unstructured or structured



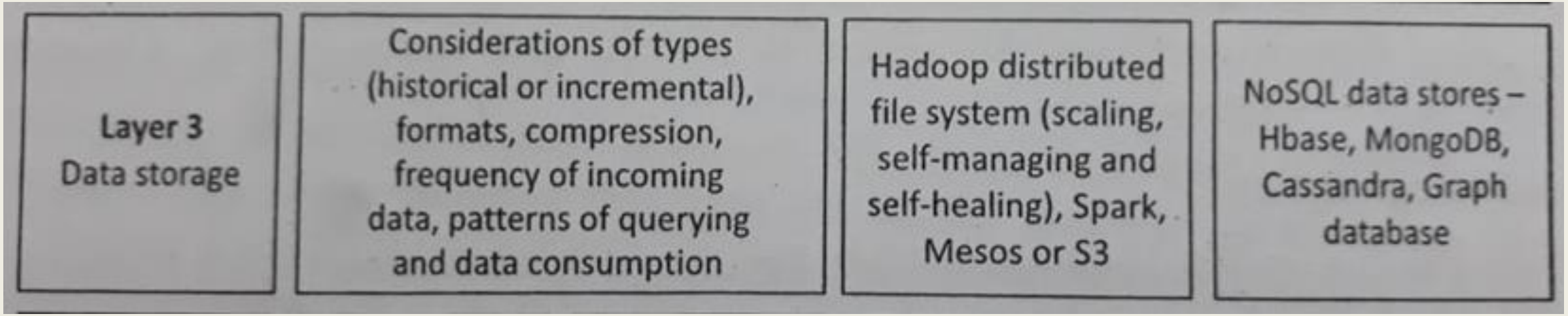
- Data ingestion means a process of absorbing information, just like the process of absorbing nutrients and medications into the body by eating or drinking them (Cambridge English Dictionary). Ingestion is the process of obtaining and importing data for immediate use or transfer. Ingestion may be in batches or in real time using pre-processing or semantics.
- **L2 considers the following aspects:**

- Ingestion and ETL processes are either in real time, which means store and use the data as generated, or in batches. Batch processing is using discrete datasets at scheduled or periodic intervals of time.



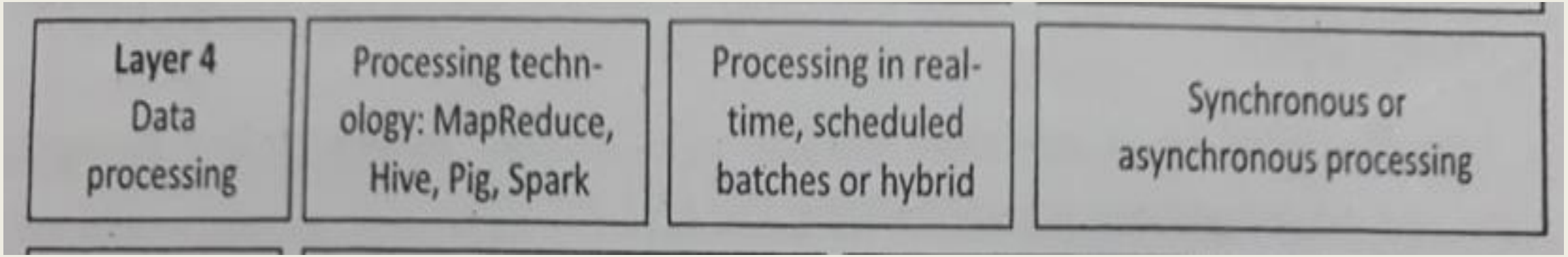
- **L3 considers the followings aspects:**

- Data storage type (historical or incremental), format, compression, incoming data frequency, querying patterns and consumption requirements for L4 or L5
- Data storage using Hadoop distributed file system or NoSQL data stores- HBase, Cassandra, MongoDB.



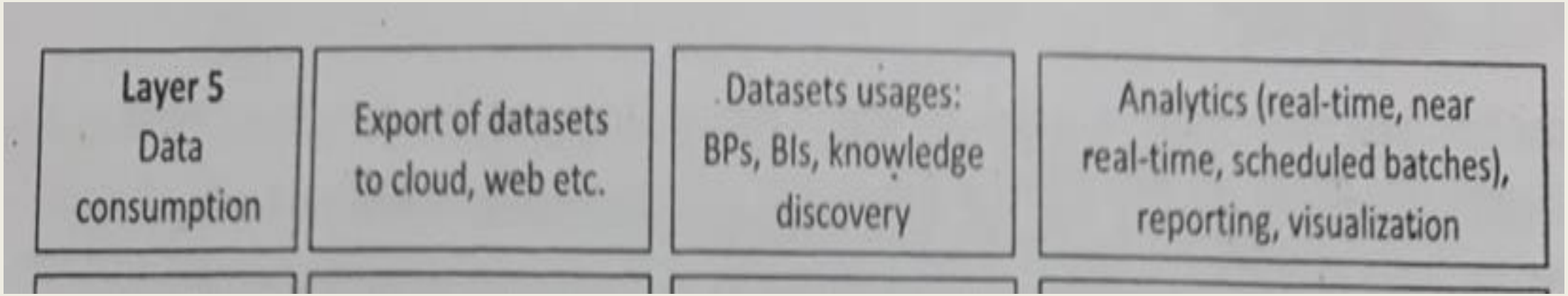
- **L4 considers the followings aspects:**

1. Data processing software such as MapReduce, Hive, Pig, Spark, Spark Mahout, Spark Streaming
2. Processing in scheduled batches or real time or hybrid
3. Processing as per synchronous or asynchronous processing requirements at L5.



- **L5 considers the consumption of data for the following**:

1. Data integration
2. Datasets usages for reporting and visualization
3. Analytics (real time, near real time, scheduled batches), BPS, BIs, knowledge discovery
4. Export of datasets to cloud, web or other systems



1.4.2 Managing Data for Analysis

- Data managing means **enabling, controlling, protecting, delivering and enhancing the value of data and Information asset**.
- Reports, analysis and visualizations need **well-defined data**.
- Data management also enables **data usage in applications**.
- The **process for managing needs to be well defined** for fulfilling requirements of the applications.

Data management functions include:

1. Data assets **creation, maintenance and protection**.
2. Data **governance**, which includes establishing the processes for ensuring the **availability, usability, integrity, security and high-quality of data**. The processes enable trustworthy data availability for analytics, followed by the decision making at the enterprise.
3. Data **architecture** creation, modelling and analysis
4. **Database administration and management system**. For example, RDBMS (relational database management system), NoSQL
5. Managing **data security**, data access control, deletion, privacy and security
6. Managing the **data quality**

7. **Data collection** using the ETL process
8. Managing **documents, records** and contents
9. Creation of **reference and master data**, and data control and supervision.
10. Data and **application integration**.
11. **Integrated data management**, enterprise-ready data creation, fast access and analysis, automation and simplification of operations on the data
12. Data **warehouse** management
13. Maintenance of **business intelligence**
14. **Data mining and analytics algorithms.**

1.5 DATA SOURCES, QUALITY, PRE-PROCESSING AND STORING

1.5.1 Data Sources

- **External Sources**:
  - *Sensors, trackers, web logs, computer systems logs and feeds.*
  - *Sources can be machines, which source data from data-creating programs.*

- **Internal Source :**
  - *Data repositories, such as database, relational database, flat file, spreadsheet, mail server, web server, directory services, even text or files such as comma-separated values (CSV) files.*
  - *Source may be a data store for applications.*

1.5.1.1 Structured Data Sources

- The source may be on the **same computer running a program or a networked computer**.
- Examples of structured data sources are SQL Server, MySQL, Microsoft Access database, Oracle DBMS, IBM DB2, Informix, Amazon SimpleDB or a file-collection directory at a server.
- A data source name implies a defined name, which a process uses to **identify the source**.
- A **data dictionary** enables references for accesses to data.
- The dictionary consists of a set of **master lookup tables**.
- The dictionary stores at a **central location**.
- The central location enables **easier access** as well as administration of changes in sources.
- The name of the dictionary can be UniversityStudents_DataPlusGrades. A master-directory server can also be called NameNode.

- The applications consider data sources as the ones where the database tables reside and where the software runs logic objects for an enterprise. Data sources can point to:

1. A database in a **specific location or in a data library** of OS
2. A specific machine in the enterprise that **processes logic**
3. A data source **master table** which stores data source definitions.

1.5.1.2 Unstructured Data Sources

- Unstructured data sources are distributed over **high-speed networks**.
- The data need **high velocity processing**.
- Sources are from distributed file systems.
- The sources are of **file types**, such as .txt (text file), .csv (comma separated values file).
- Data may be as **key-value pairs**, such as hash key-values pairs.
- Data may have **internal structures**, such as in e-mail, Facebook pages, twitter messages etc.
- The data **do not model**, reveal relationships, hierarchy relationships or object-oriented features, such as extensibility.

1.5.1.3 Signals and GPS

- The **data sources** can be sensors, sensor networks, signals from machines, devices, controllers and intelligent edge nodes of different types in the industry M2M communication and the GPS systems.
- Sensors are electronic devices that sense the physical environment. Sensors are devices which are used for **measuring** temperature, pressure, humidity, light intensity, traffic in proximity, acceleration, locations, object(s) proximity, orientations and magnetic intensity, and other physical states and parameters. Sensors play an active role in the **automotive industry.**
- RFIDs and their sensors play an active role in RFID based **supply chain management**, and tracking parcels, goods and delivery.

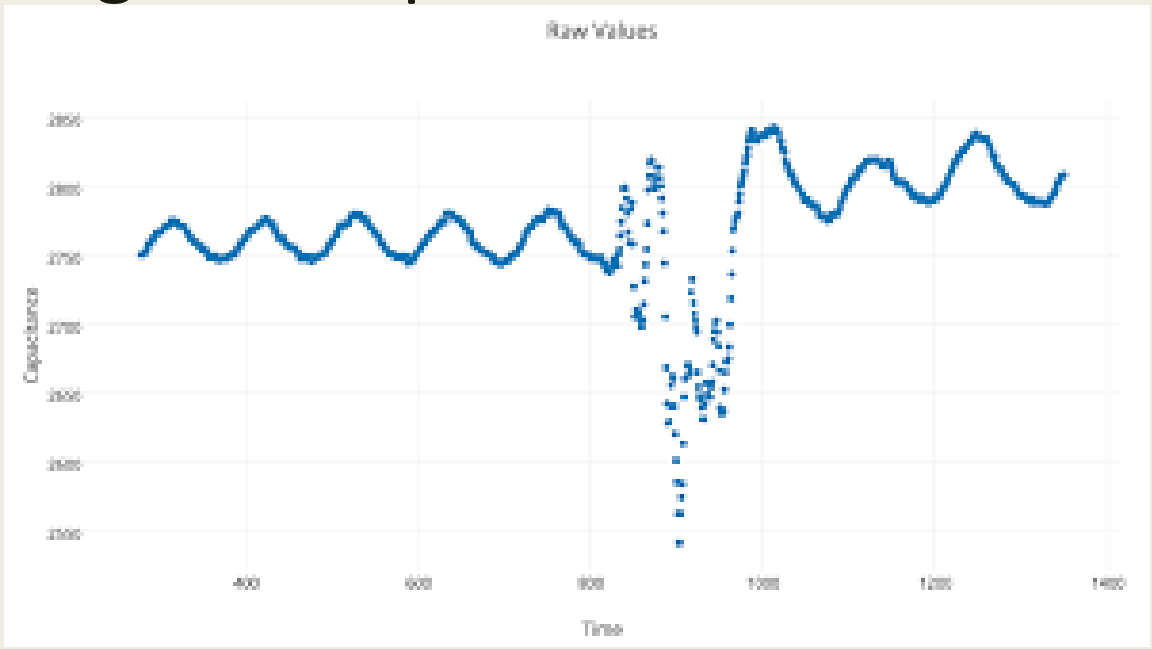
1.5.2 Data Quality

- Data quality is high when it represents the real-world construct to which references are taken.
- High quality means data, which enables all the required operations, analysis, decisions, planning and knowledge discovery correctly.
- A definition for high quality data, especially for artificial intelligence applications, can be data with five R's as follows: Relevancy, recency, range, robustness and reliability. Relevancy is of utmost importance.
- A uniform definition of data quality is difficult. A reference can be made to a set of values of quantitative or qualitative conditions, which must be specified to say that data quality is high or low.

1.5.2.1 Data Integrity

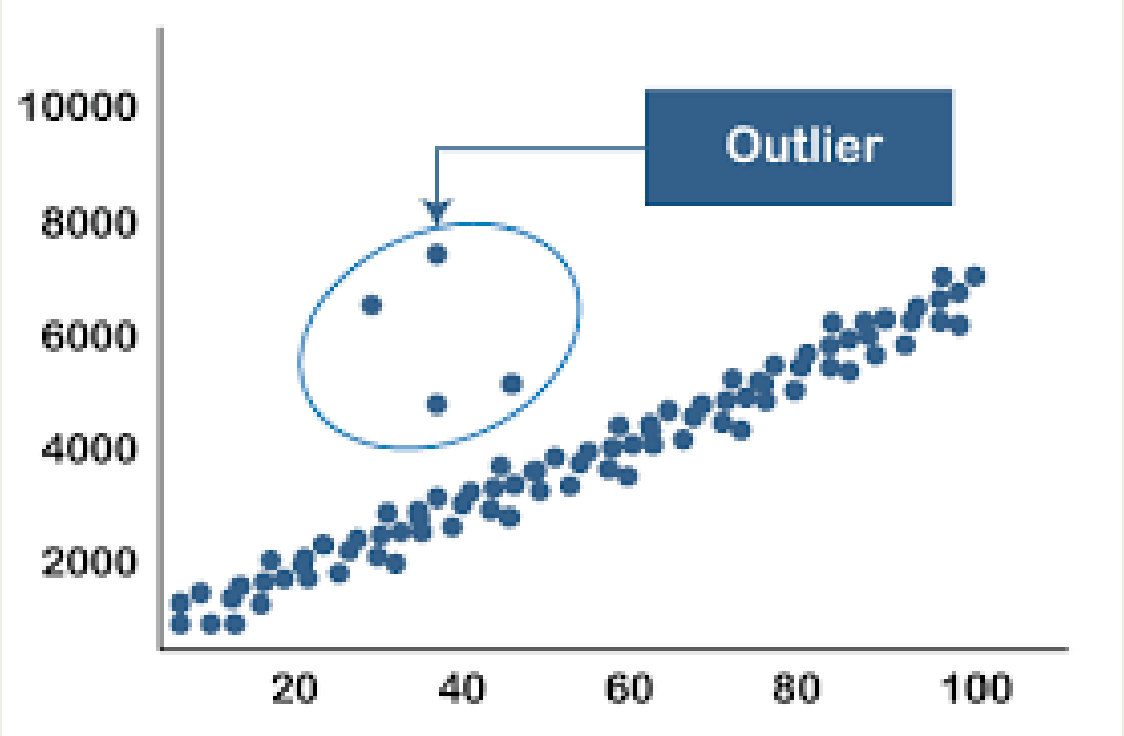
- Data integrity refers to the maintenance of consistency and accuracy in data over its usable life.
- Software, Which store, process, or retrieve the data, should maintain the integrity of data. Data should be incorruptible.

1.5.2.2 Data Noise, Outliers, Missing and Duplicate Values



■ Noise

- *Data giving additional meaningless information besides true (actual/required) information.*
- *Noise refers to **difference in the value** measured from true value due to additional influences.*
- *Result of data analysis is **adversely affected** due to noisy data.*
- *Noise is **random in character**, which means frequency with which it occurs is variable over time.*
- *The values show nearly equal **positive and negative deviations**.*
- *A **statistical analysis** of deviation can select the noise in data and true values can be retrieved.*



Outliers

- *A factor which effects **quality** is an outlier.*
- *An outlier in data refers to data, which appears to **not belong to the dataset**.*
- *For example, data that is **outside an expected range**.*
- *Actual outliers need to be **removed** from the dataset, else the result will be effected by a small or large amount.*
- *Alternatively, if **valid data is identified as outlie**r, then also the results will be affected.*
- *The outliers are a result of **human data-entry errors, programming bugs, some** transition effect or phase lag in stabilizing the data value to the true value.*

- **Missing Values**
  - *Missing value implies data not appearing in the data set.*

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

Duplicate Values

Duplicate value implies the same data appearing two or more times in a dataset

id	first_name	last_name	email
1	Carine	Schmitt	carine.schmitt@verizon.net
4	Janine	Labrune	janine.labrune@aol.com
6	Janine	Labrune	janine.labrune@aol.com
2	Jean	King	jean.king@me.com
12	Jean	King	jean.king@me.com
5	Jonas	Bergulfsen	jonas.bergulfsen@mac.com
10	Julie	Murphy	julie.murphy@yahoo.com
11	Kwai	Lee	kwai.lee@google.com
3	Peter	Ferguson	peter.ferguson@google.com
9	Roland	Keitel	roland.keitel@yahoo.com
14	Roland	Keitel	roland.keitel@yahoo.com
7	Susan	Nelson	susan.nelson@comcast.net
13	Susan	Nelson	susan.nelson@comcast.net
8	Zbyszek	Piestrzeniewicz	zbyszek.piestrzeniewicz@att.net

1.5.3 Data Pre-processing

- Data pre-processing is an important step at the **ingestion layer**.
- The **outlier** needs to be removed.
- Pre-processing is a must **before data mining and analytics**
- Pre processing is also a must **before running a Machine Learning (ML) algorithm.**
- Analytics needs prior **screening of data quality** also.
- Data when being exported to a cloud service or data store needs pre-processing.

- Pre-processing needs are:

(I)Dropping out of range, inconsistent and outlier values.

(II)Filtering unreliable, irrelevant and redundant information.

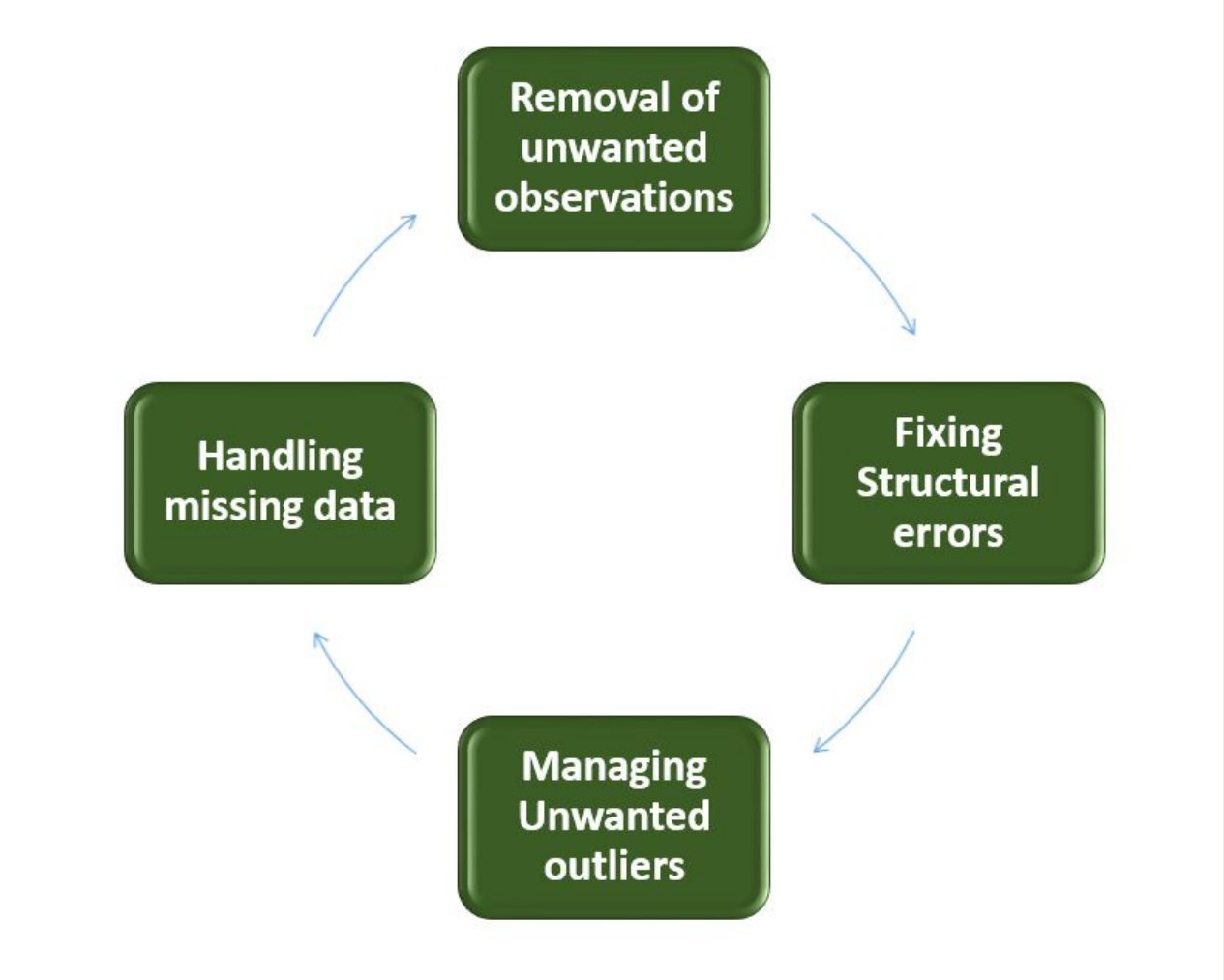
(III)Data cleaning, editing, reduction and/or wrangling.

(IV)Data validation, transformation or transcoding.

(V)ELT processing.

■ Data Cleaning

– *Data cleaning refers to the process of removing or correcting incomplete, incorrect, inaccurate or irrelevant parts of the data after detecting them.*

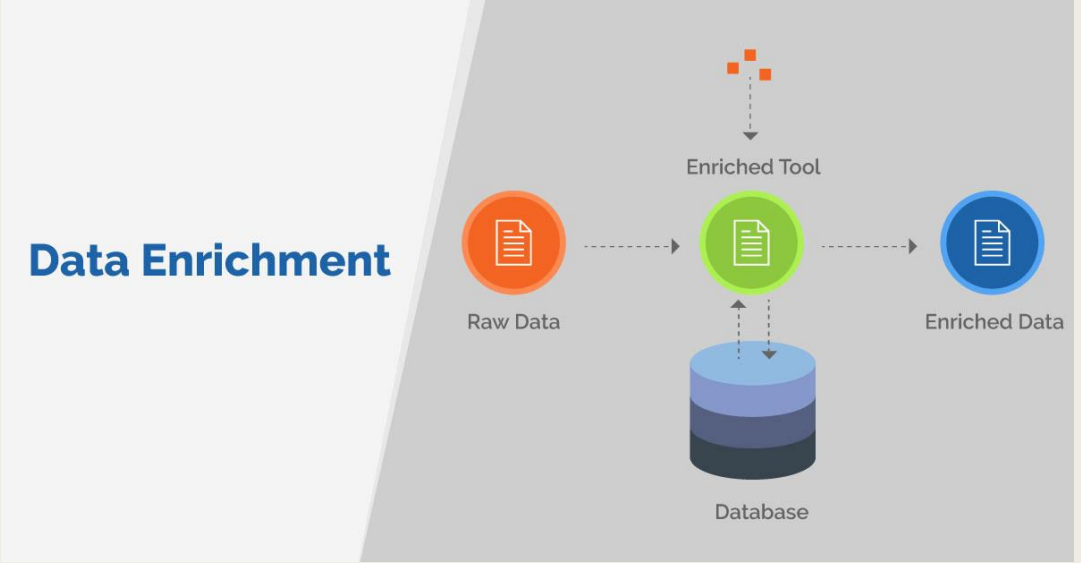


Data Cleaning Tools

- Data can generate in a system in many formats when it is obtained from the web.
- Data cleaning tools help in refining and structuring data into usable data.
- *OpenRefine :* OpenRefine, previously known as GoogleRefine, is a powerful, **open source software which visualizes and manipulates large quantities of data all at once**.
- **OpenRefine** looks like a spreadsheet, but operates like a database, allowing for increased discovery capabilities beyond programs like Microsoft Excel.
- **DataCleaner:** DataCleaner is a Data Quality toolkit that allows you to profile, correct and enrich your data.

Data Enrichment

- Data enrichment refers to operations or processes which refine, enhance or improve the raw data.



Data Editing

Data editing refers to the process of reviewing and adjusting the acquired datasets.

The editing controls the data quality.

Editing methods are (1) interactive, (ii) selective, (iii) automatic, (iv) aggregating and (v) distribution.

Table (1): A hypothetical example of numerical data

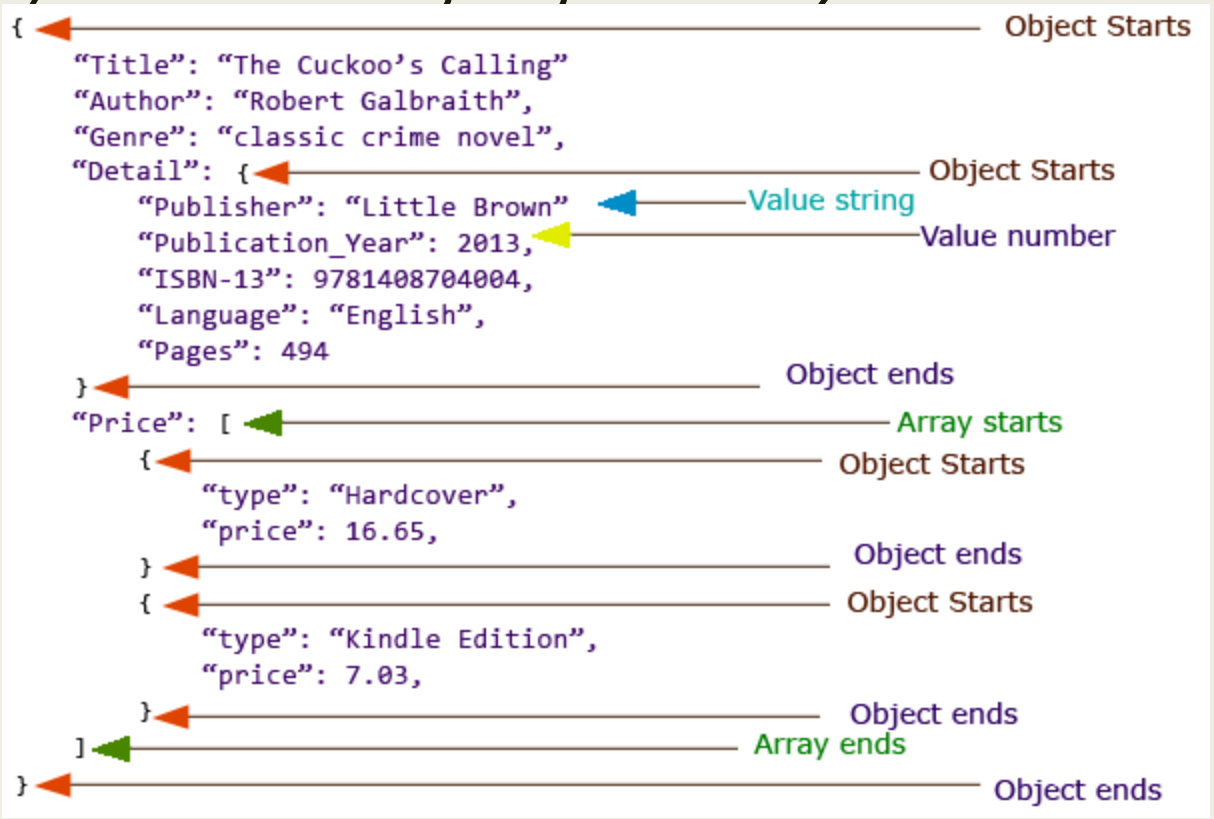
	Column 1	Column 2	Column 3
Row 1	26	22	12
Row 2	Green	8	7
Row 3	84	60	-

- Data Reduction
  - *Data reduction enables the transformation of acquired information into an ordered, correct and simplified form.*
  - *The reductions enable ingestion of meaningful data in the datasets.*
  - *The basic concept is the **reduction of multitudinous amount** of data, and **use of the meaningful parts**.*
  - *The reduction uses editing, sealing, coding, sorting, collating, smoothening, interpolating and preparing tabular summaries.*

- **Data Wrangling**
- Data wrangling refers to the process of transforming and mapping the data.
- Results from analytics are then appropriate and valuable.
- For example, mapping enables data into another format, which makes it valuable for analytics and data visualizations.

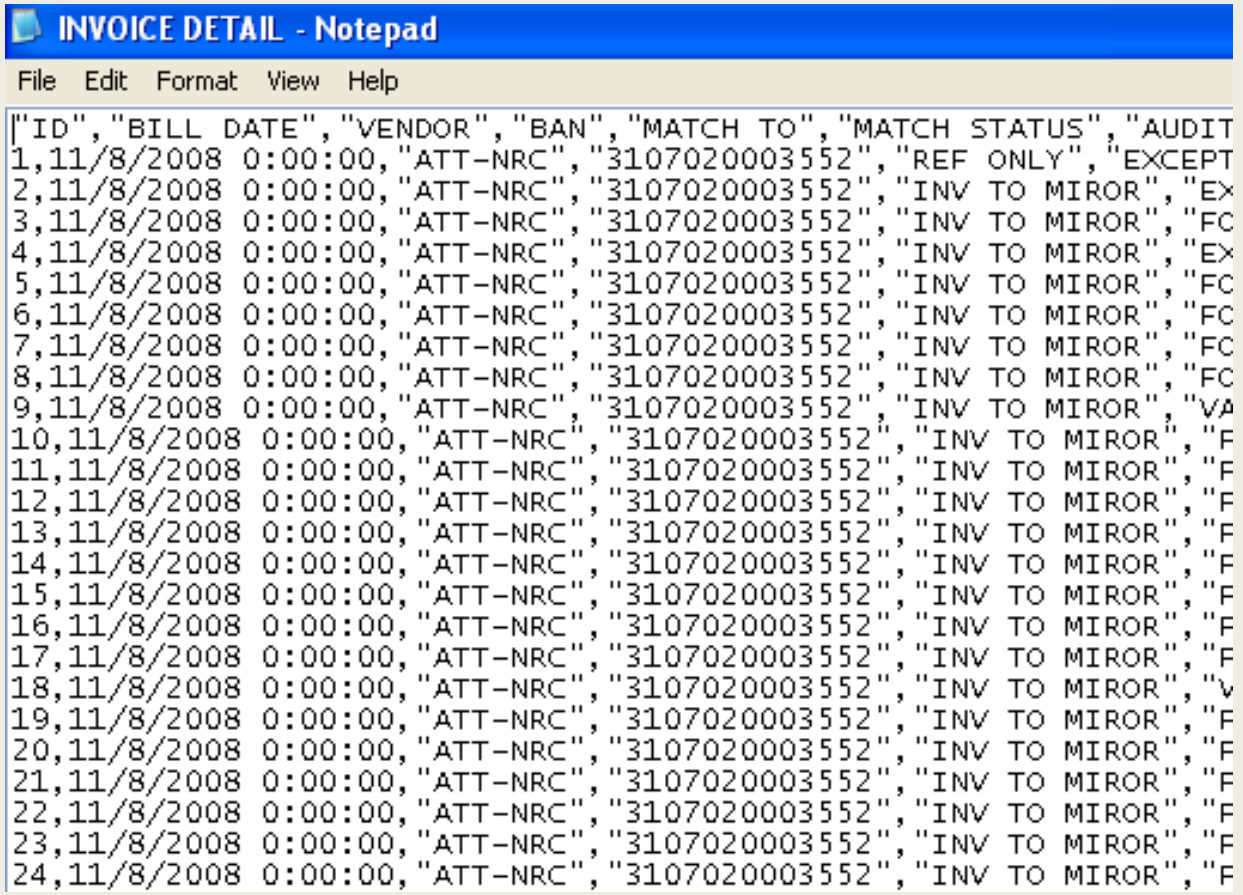
Data Format used during Pre-Processing

- Examples of formats for data transfer from (a) data storage, (b) analytics application, (b) service or (d) cloud can be:
  - *Java Script Object Notation (JSON) as batches of object arrays or resource arrays*



CSV Format

- An example is a table or Microsoft Excel file which needs conversion to CSV format.
- Each CSV file line is a data record.
- Each record consists of one or more fields, separated from each other by commas.
- RFC 4180 standard specifies the various specifications.
- A CSV file may also use space, tab or delimiter tab-separated formats for the values in the fields.



- ***Tag Length Value***
- ***Example : 02 02 05***

TLV Encoding

Idea: transmitted data is self-identifying

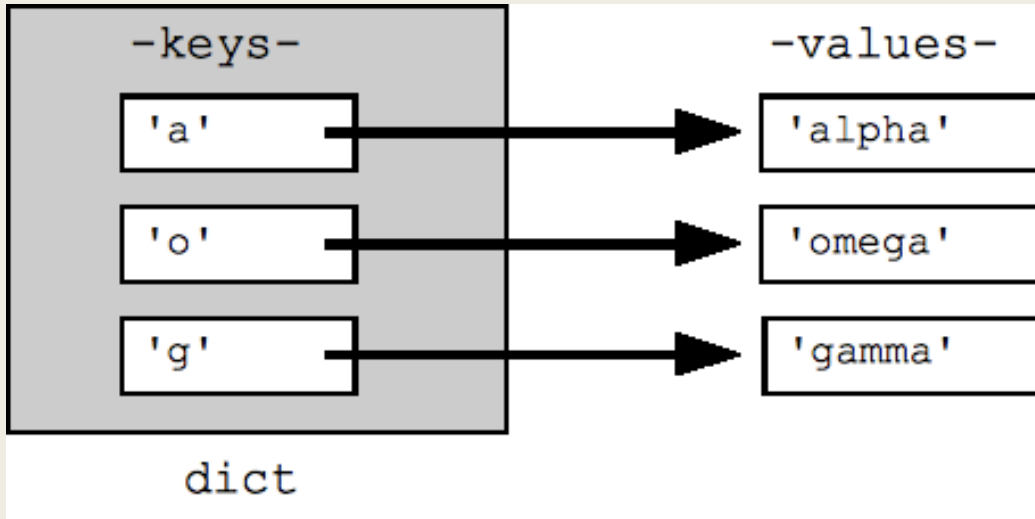
- T: data type, one of ASN.1-defined types
- L: length of data in bytes
- V: value of data, encoded according to ASN.1 standard

Tag Value	Type
1	Boolean
2	Integer
3	Bitstring
4	Octet string
5	Null
6	Object Identifier
9	Real

Key-value pairs

Hash-key-value pairs

Key	Value
Name	Joe Bloggs
Age	42
Occupation	Stunt Double
Height	175cm
Weight	77kg



1.5.4 Data Store Export to Cloud

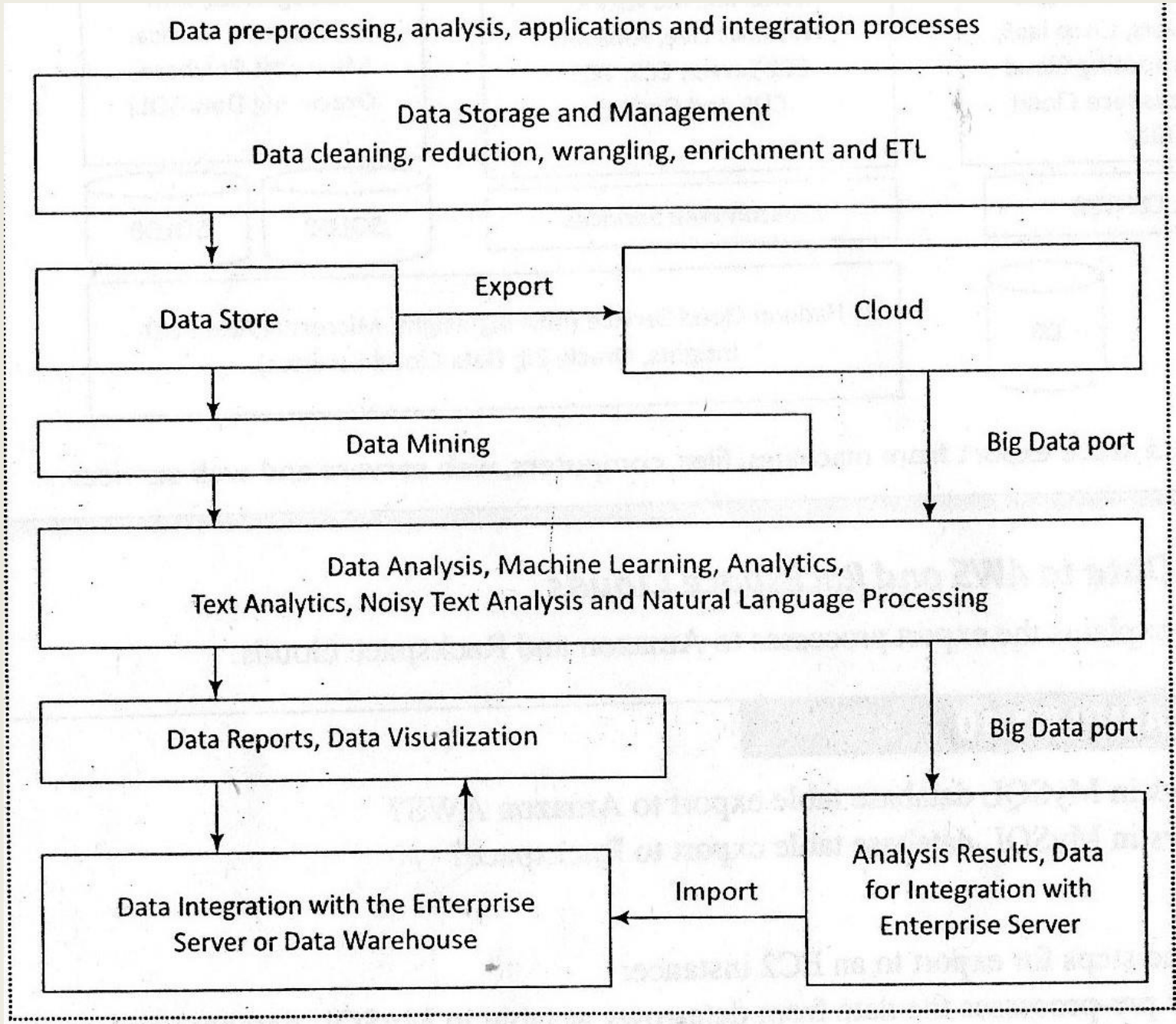
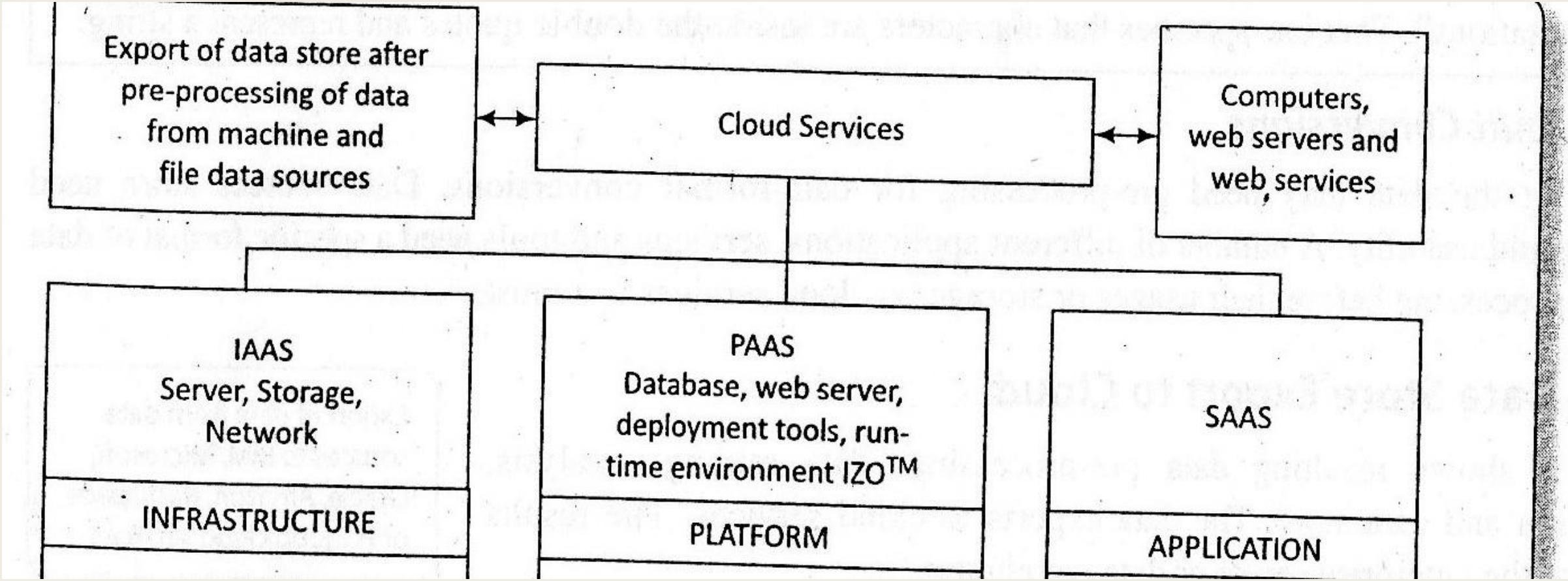


Figure 1.3 Data pre-processing, analysis, visualization, data store export

1.5.4.1 Cloud Services

Cloud offers various services. These services can be accessed through a cloud client (client application), such as a web browser, SQL or other client.



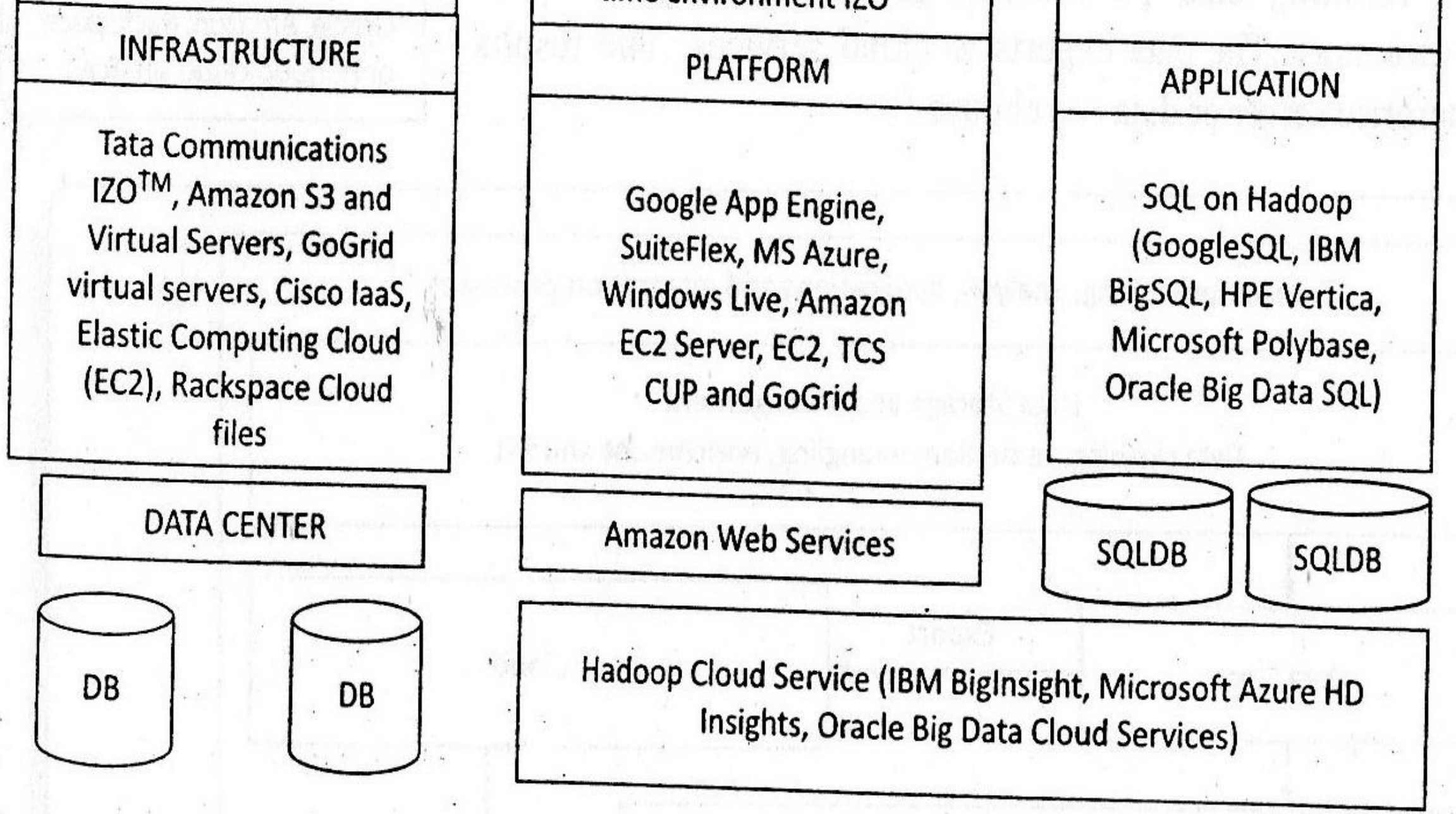


Figure 1.4 Data store export from machines, files, computers, web servers and web services

1.5.4.2 Export of data to AWS and Rackspace Clouds

- Google Cloud Platform, offered by Google, is a suite of cloud computing services that runs on the same infrastructure that Google uses internally for its end-user products, such as Google Search, Gmail, Google Drive, and YouTube.

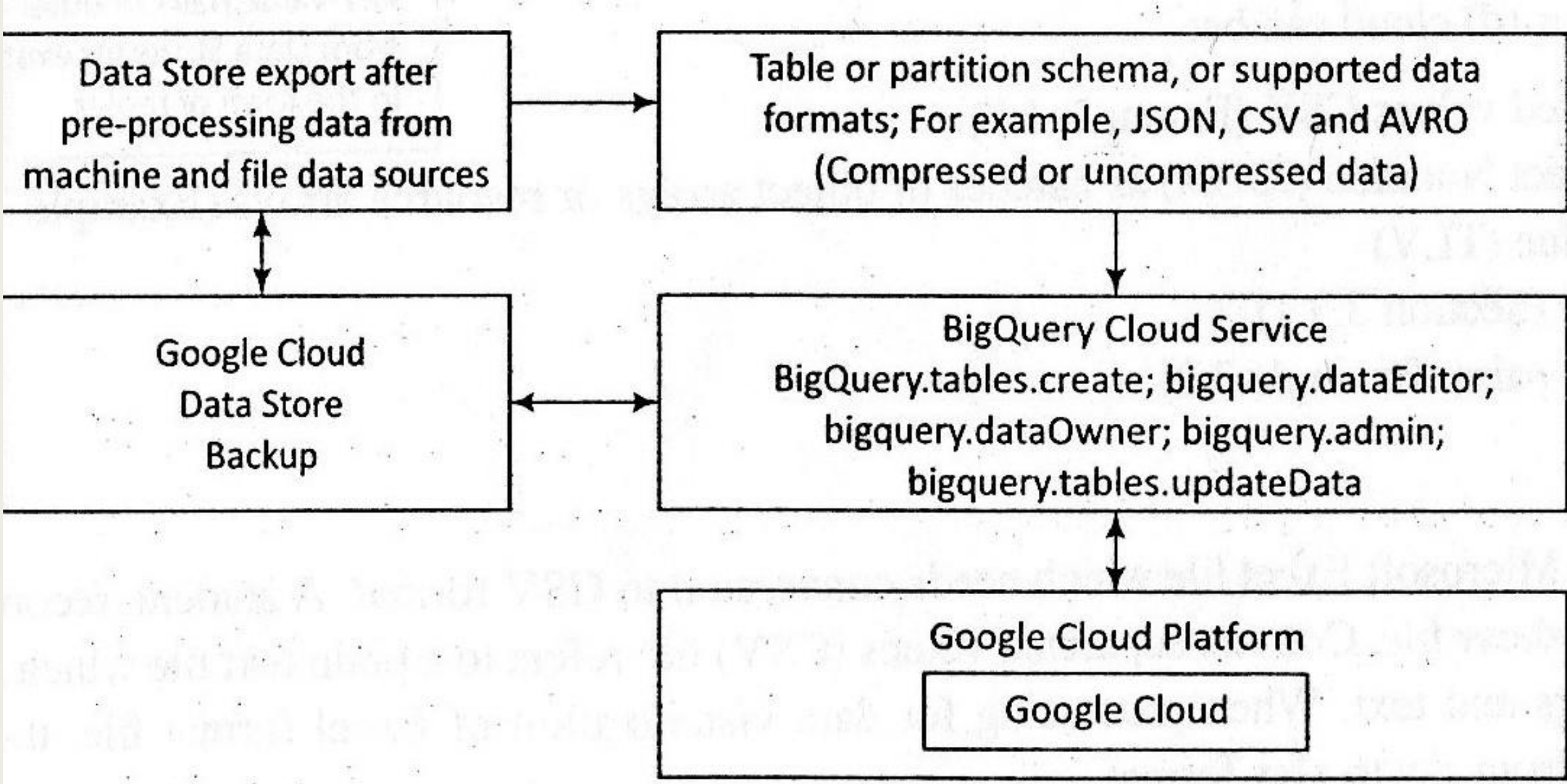


Figure 1.5 BigQuery cloud service at Google cloud platform

1.6 DATA STORAGE AND ANALYSIS

- The following subsections describe **data storage and analysis, and comparison** between Big Data management and analysis with **traditional database management** systems.
- **1.6.1 Data Storage and Management: Traditional Systems**
- **1.6.1.1 Data Store with Structured or Semi-Structured Data**
  - *Traditional systems use structured or semi-structured data.*
  - *The following example explains the sources and data store of structured data.*

1.6.1.2 SQL

- An RDBMS uses SQL (Structured Query Language).
- SQL is a **language for viewing or changing (update, insert or append or delete) databases**.
- SQL was originally based on the **tuple relational calculus and relational algebra**.
- **SQL does the following**:

1. **Create schema :Base tables, views, constraints, describe the data and define the data in the database.**

2 **Create catalog : A set of schemas that** describe the database.

3. Data Definition Language (DDL) : creating. altering and dropping of tables and establishing the constraints. Drop databases and tables, establish **foreign keys, create view, stored procedure, functions in the database etc.**

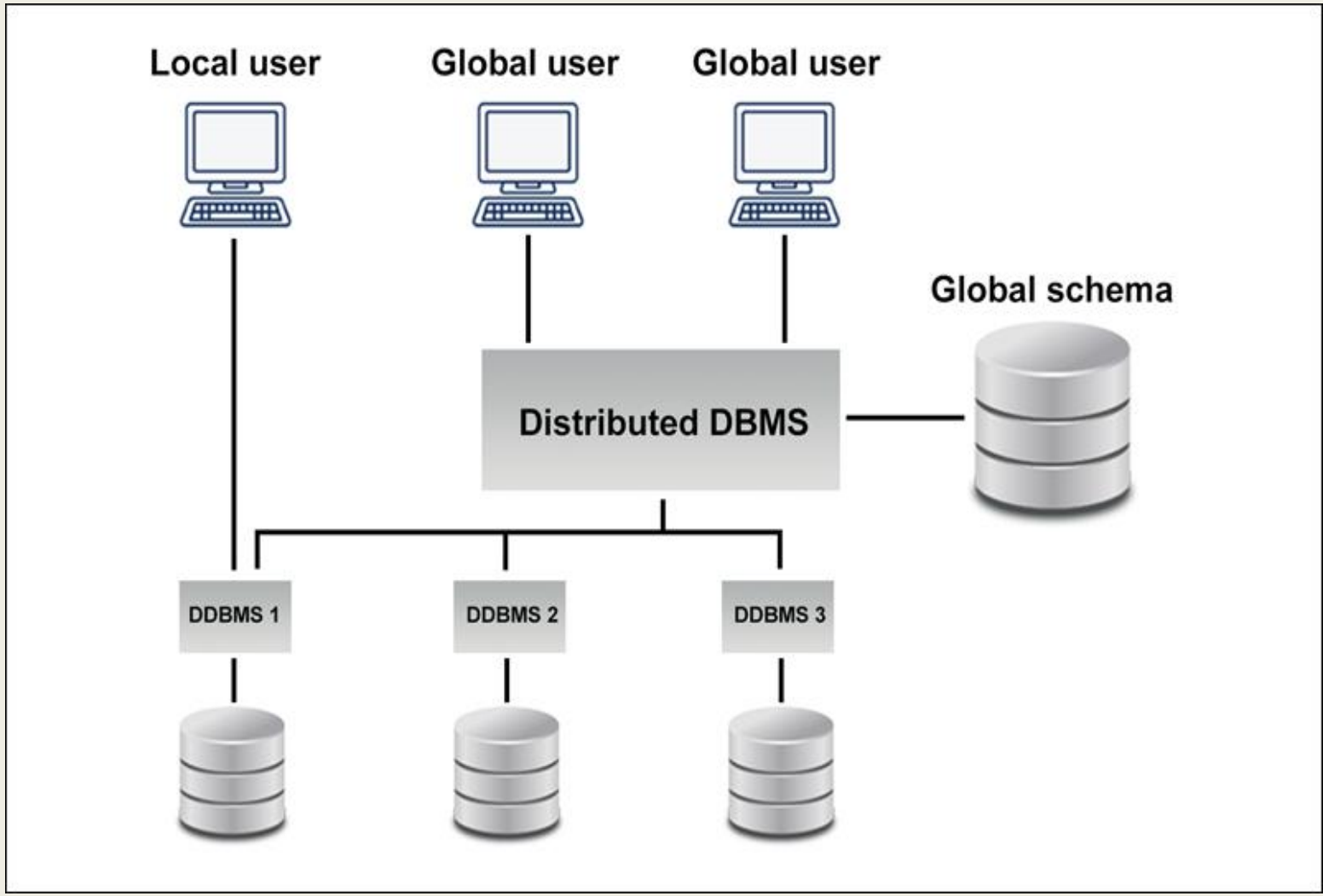
4. **Data Manipulation Language (DML)** for commands that maintain and **query the database. A user can manipulate (INSERT/UPDATE) and access (SELECT) the data.**

5. **Data Control Language (DCL)** for commands that control a database, and include **administering of privileges and committing**. A user can **set (grant, add or revoke) permissions on tables, procedures and views**

1.6.1.4 Distributed Database Management System

- A distributed DBMS (DDBMS) is a **collection of logically interrelated databases at multiple system over a computer network.**
- The **features**

1. L**ogically related databases**.
2. **Cooperation; transparency**
3. L**ocation independent, move**



1.6.1.5 In-Memory Column Formats Data

- Data in a column are kept together in-memory in columnar format. A single memory access loads many values at the column.
- Faster data retrieval.
- An address increment to a next memory address for the next value is fast.
- Online Analytical Processing (OLAP) in real-time transaction processing is fast.
- The CPU accesses all columns in a single instance of access
- Online viewing of analyzed data and visualization
- Granularity
- Summarized information

Column-oriented: each column is stored in a separate file

Each column for a given row is at the same offset.

Key	Fname	Lname	State	Zip	Phone	Age	Sales
1	Bugs	Bunny	NY	11217	(123) 938-3235	34	100
2	Yosemite	Sam	CA	95389	(234) 375-6572	52	500
3	Daffy	Duck	NY	10013	(345) 227-1810	35	200
4	Elmer	Fudd	CA	04578	(456) 882-7323	43	10
5	Witch	Hazel	CA	01970	(567) 744-0991	57	250

1.6.1.6 In-Memory Row Format Databases

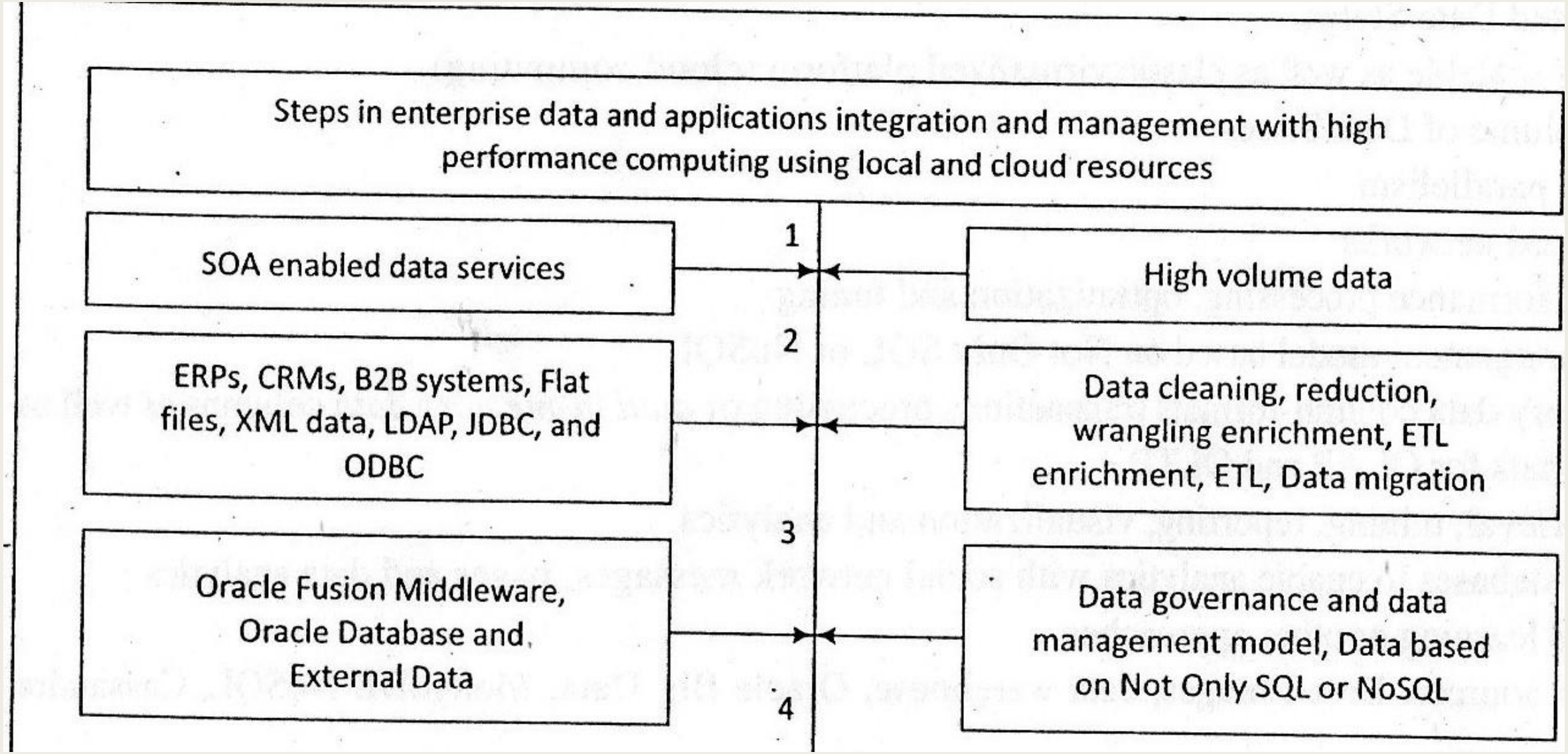
- Faster data processing during OLTP (online transaction processing).
- Each row record has corresponding values in multiple columns and the on-line values store at the consecutive memory addresses in row format.
- Example: A specific day's sale
- Continuous computation that happens as data is flowing through the system.

Row-oriented: rows stored sequentially in a file

Key	Fname	Lname	State	Zip	Phone	Age	Sales
1	Bugs	Bunny	NY	11217	(123) 938-3235	34	100
2	Yosemite	Sam	CA	95389	(234) 375-6572	52	500
3	Daffy	Duck	NY	10013	(345) 227-1810	35	200
4	Elmer	Fudd	CA	04578	(456) 882-7323	43	10
5	Witch	Hazel	CA	01970	(567) 744-0991	57	250

1.6.1.7 Enterprise Data-Store Server and Data Warehouse

- Enterprise data, after data cleaning process, integrate with the server data at warehouse.
- Enterprise data server **use data from several distributed sources** which store data using various technologies.
- All data **merge using an integration tool**. Integration enables **collective viewing** of the datasets at the data warehouse.
- Enterprise data integration may also include **integration with application(s), such as analytics, visualization, business intelligence and knowledge discovery.**


Steps in enterprise data and applications integration and management with high performance computing using local and cloud resources
SOA enabled data services
1
High volume data
ERPs, CRMs, B2B systems, Flat files, XML data, LDAP, JDBC, and ODBC
2
Data cleaning, reduction, wrangling enrichment, ETL enrichment, ETL, Data migration
Oracle Fusion Middleware, Oracle Database and External Data
3
Data governance and data management model, Data based on Not Only SQL or NoSQL
4

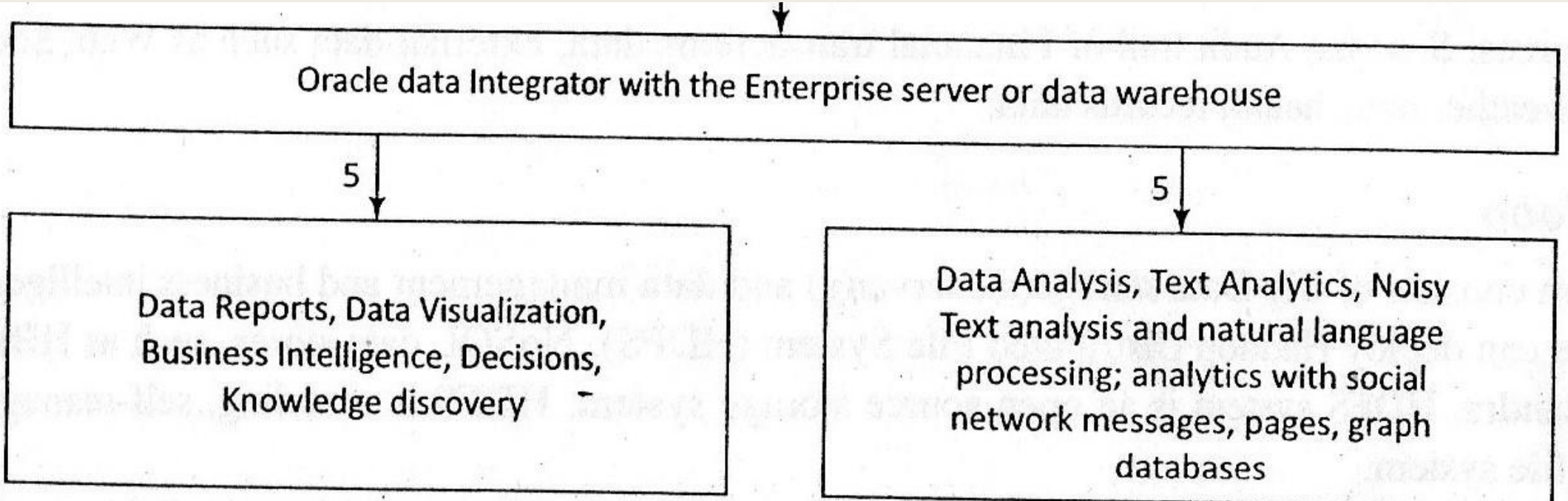


Figure 1.6 Steps 1 to 5 in Enterprise data integration and management with Big-Data for high performance computing using local and cloud resources for the analytics, applications and services

1.6.2 Big Data Storage

1.6.2.1 Big Data NoSQL or Not Only SOL

- NoSQL databases are considered as semi-structured data.
- Big Data Store uses **NoSQL**.
- NOSQL stands for **No SQL or Not Only SQL**.
- The stores **do not integrate with applications using SQL**.
- NoSQL is also used in **cloud data store**.

Features of NoSQL are as follows:

It is a class of non-relational data storage systems, and the flexible data models and multiple schema:

*(i) Class consisting of **uninterrupted key/value** or big hash table [Dynamo (Amazon S3)]*

*(ii) Class consisting of **unordered keys and using JSON** (PNUTS)*

*(iii) Class consisting of **ordered keys and semi-structured data storage systems** [BigTable, Cassandra (used in Facebook/Apache) and HBase]*

*(iv) Class consisting of **JSON** (MongoDB)*

*(v) Class consisting of **name/value in the text (**CouchDB)*

*(vi) May **not use fixed table schema***

*(vii) Do not use the **JOINS***

*(viii) Data written at one node can replicate at multiple nodes, therefore Data storage is **fault tolerant,***

*(ix) May **relax the ACID** rules during the Data Store transactions.*

*(x) Data Store can be partitioned and follows **CAP theorem** (out of three properties, consistency, availability and partitions, at least two must be there during the transactions)*

1.6.2.2 Coexistence of Big Data, NoSQL and Traditional Data Stores

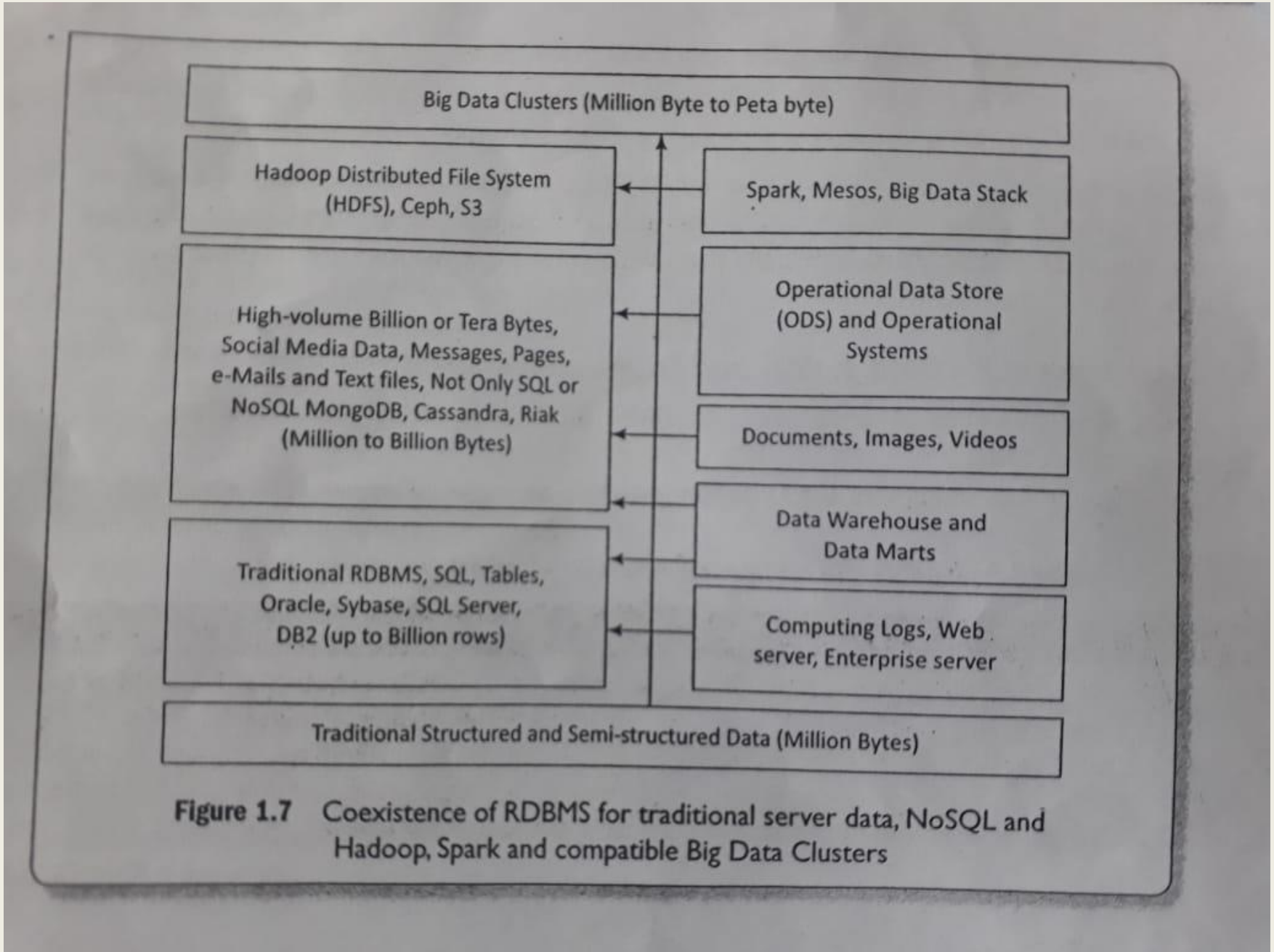


Figure 1.7 Coexistence of RDBMS for traditional server data, NoSQL and Hadoop, Spark and compatible Big Data Clusters

1.6.3 Big Data Platform

- The data generate at a higher velocity, in more varieties or in higher veracity, Managing Big Data requires large resources of MPPs, cloud, parallel processing and specialized tools.

Bigdata platform should provision tools and services for:

1.Storage, Processing and analytics

2. Developing, deploying, operating and **managing Big Data environment**,

3. Reducing the **complexity** of multiple data sources and **integration** of applications into one cohesive solution

4. Custom development, querying and **integration with other systems**

- Data management, storage and analytics of Big **data captured at the companies and services** require the following:

1. **New innovative non-traditional methods** of storage, processing and analytics

2.**Distributed** Data Stores

3. Creating **scalable** as well as elastic virtualized platform (cloud computing)

4. **Huge volume** of Data Stores

5. The traditional as well as **Big Data techniques**.

6. **High speed networks**

7. **High performance** processing, optimization and tuning

8. **Data management model** based on Not Only SQL or NoSQL

9. In-memory data column-formats transactions processing or dual in-memory data columns as row formats for **OLAP and OLTP**

10. **Data retrieval, mining, reporting, visualization and analytics**

11. **Graph databases to enable analytics** with social network messages, pages and data analytics

12. Machine learning or other approaches

13. **Big data sources**: Data storages, data warehouse Oracle Big Data, MongoDB NoSQL, Cassandra NoSQL.

14. **Data sources**: Sensors, Audit trail of Financial transactions data, external data such as Web, Social Media, weather data, health records data.

1.6.3.1 Hadoop

- Big Data platform consists of Big Data storage(s), server(s) and data management and business intelligence software.

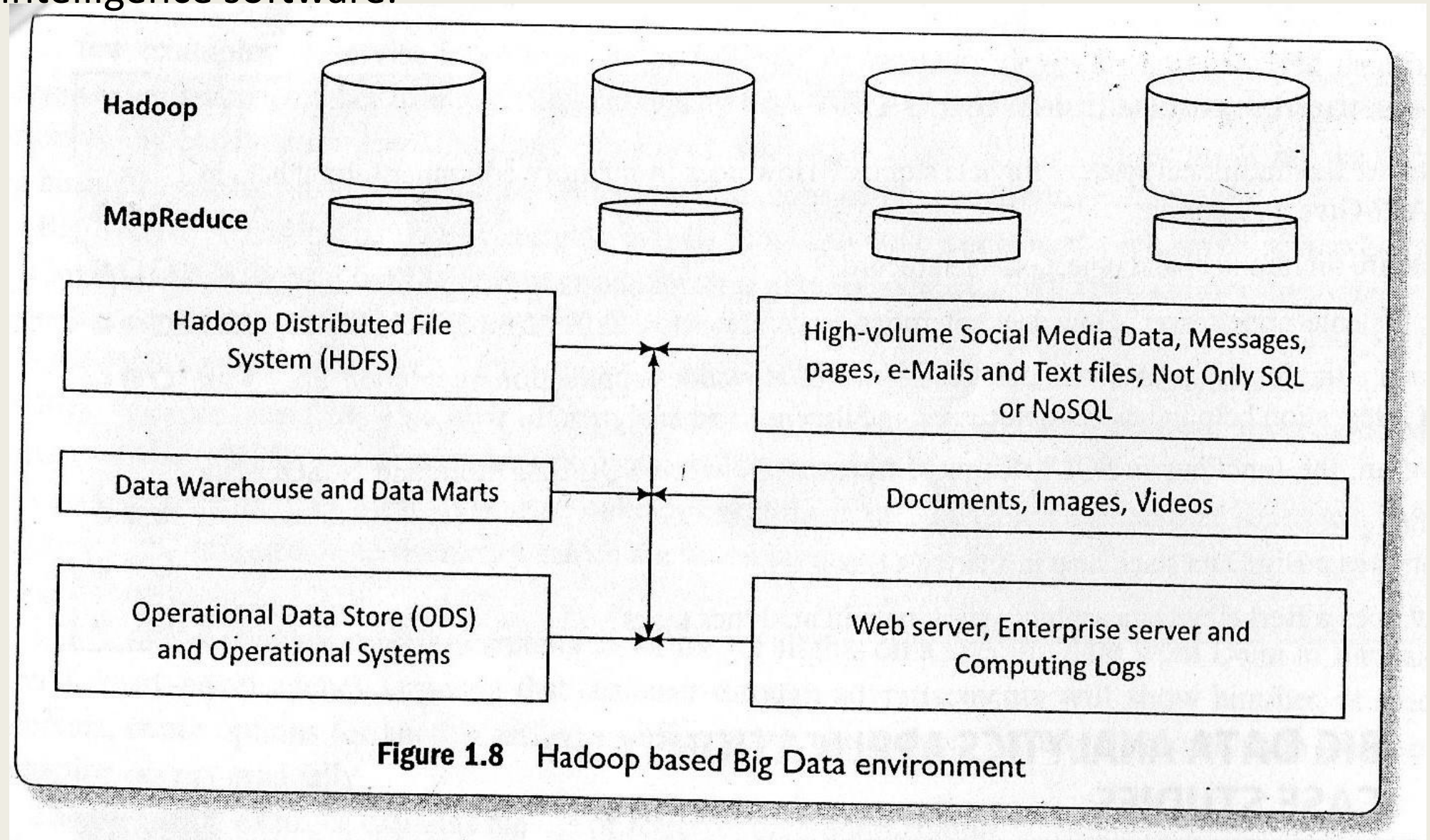


Figure 1.8 Hadoop based Big Data environment

1.6.3.2 Mesos

- Mesos v0.9 is a resources management platform which enables **sharing of cluster of nodes by multiple frameworks and which has compatibility with an open analytics stack** [data processing (Hive, Hadoop, HBase, Storm), data management (HDFS)].

1.6.3.3 Big Data Stack

- A stack consists of a **set of software components and data store units**.
- Applications, machine-learning algorithms, analytics and visualization tools **use Big Data Stack (BDS) at a cloud service, such as Amazon EC2, Azure or cloud**.
- The stack uses **cluster of high performance machines**

Table 1.5 Tools for Big Data environment

Types	Examples
MapReduce	Hadoop, Apache Hive, Apache Pig, Cascading, Cascalog, mrjob (Python MapReduce library) Apache S4, MapR, Apple Acunu, Apache Flume, Apache Kafka
NoSQL Databases	MongoDB, Apache CouchDB, Apache Cassandra, Aerospike, Apache HBase, Hypertable
Processing	Spark, IBM BigSheets, PySpark, R, Yahoo! Pipes, Amazon Mechanical Turk, Datameer, Apache Solr/Lucene, ElasticSearch
Servers	Amazon EC2, S3, GoogleQuery, Google App Engine, AWS Elastic Beanstalk, Salesforce Herok
Storage	Hadoop Distributed File System, Amazon S3, Mesos

1.6.4 Big Data Analytics

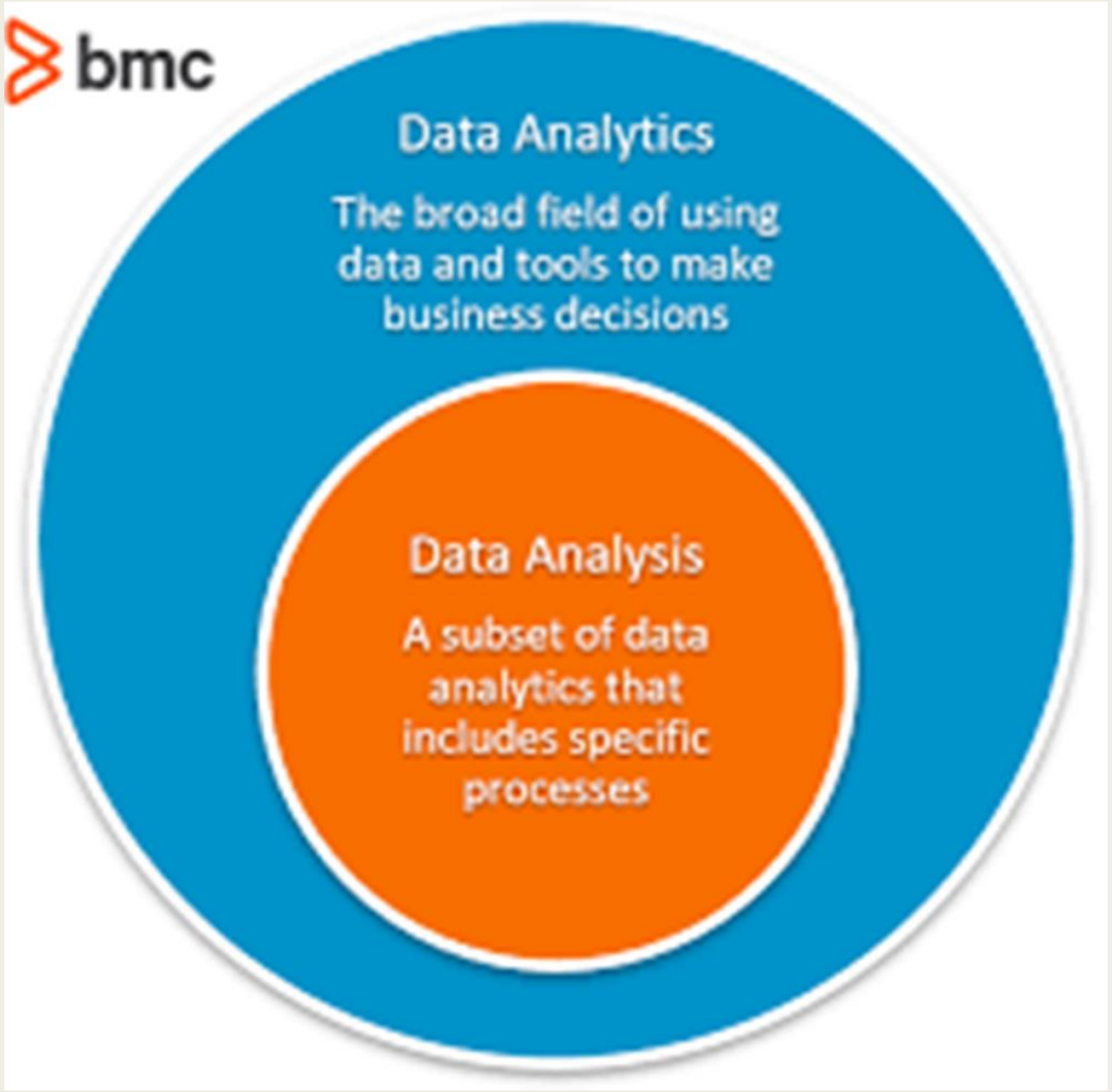
- **Data analysis can be defined as**.
- "Analysis of data is a process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information, suggesting conclusions and supporting decision making." (Wikipedia)
- Data analysis helps in finding **business intelligence and helps in decision making**.

Preliminary Data Analysis

- **Tabulation**
  - Simple Counts
  - For example
    - 74 families in the study own 1 car
    - 2 families own 3
  - **Missing data (9)**
    - 1 Family did not report
    - Not useful for further analysis

Number of Cars	Number of Families
1	75
2	23
3	2
9	1
Total	101

- **Data Analytics** can be formally defined as the **statistical** and **mathematical** data analysis that **clusters, segments, ranks and predicts** future possibilities.
- An important feature of data analytics is **predictive forecasting and prescriptive capability.**
- Analytics uses **historical data and forecasts new values or results**. Analytics suggests techniques which will provide the most **efficient and beneficial results** for an enterprise.



1.6.4.2 Phases in Analytics

Analytics has the following phases before deriving the new facts, providing business intelligence and generating new knowledge.

1. **Descriptive analytics** enables deriving the additional value from visualizations and reports

2. **Predictive analytics** is advanced analytics which enables extraction of new facts and knowledge, and

3. **Prescriptive analytics** enable derivation of the additional value and undertake better decisions for new option(s) to maximize the profits

4. **Cognitive analytics** enables derivation of the additional value and undertake better decisions. Analytics integrates with the enterprise server or data warehouse.


Business Analytics
Descriptive
Predictive
Prescriptive
Questions
What happened?
What is happening?
What will happen?
Why will it happen?
What should I do?
Why should I do it?
Enablers
• Business reporting
• Dashboards
• Scorecards
• Data warehousing
• Data mining
• Text mining
• Web/media mining
• Forecasting
• Optimization
• Simulation
• Decision modeling
• Expert systems
Outcomes
Well defined business problems and opportunities
Accurate projections of the future states and conditions
Best possible business decisions and transactions

Figure 1.9 shows an overview of a reference model for analytics architecture. The figure also shows on the right-hand side the Big Data file systems, machine learning algorithms and query languages and usage of the Hadoop ecosystem.

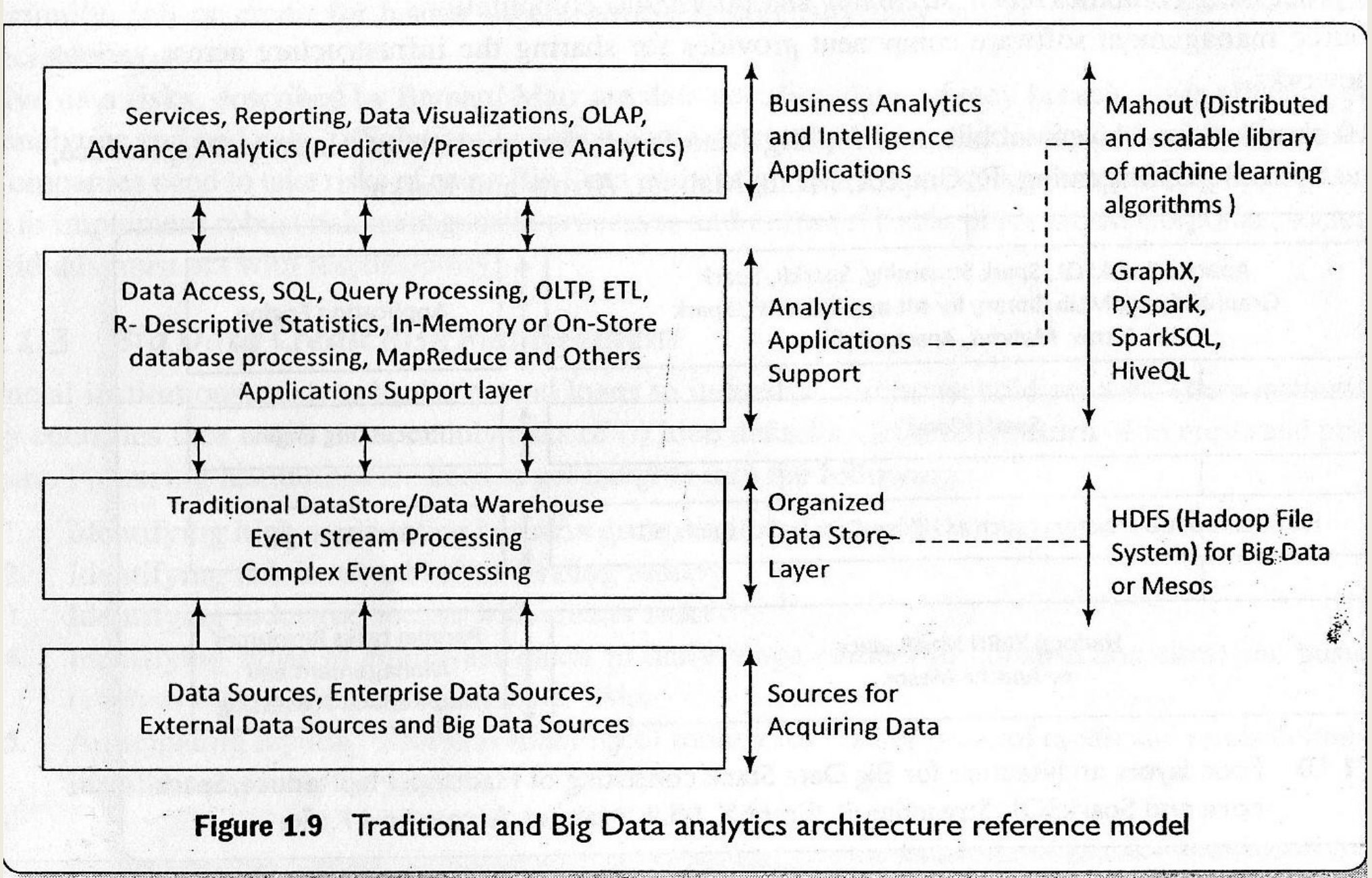


Figure 1.9 Traditional and Big Data analytics architecture reference model

1.6.4.3 Berkeley Data Analytics Stack (BDAS)

- The importance of Big Data lies in the **fact that what one does with it rather than how big or large it is.**
- Identify whether the gathered data is able to help in obtaining the following findings:
- 1) cost reduction, 2) time reduction, 3) new product planning and development, 4) smart decision making using predictive analytics and 5) knowledge discovery.
- Big Data analytics need innovative as well as cost effective techniques.
- BDAS is an open-source data analytics stack for complex computations on Big Data." It supports efficient, large-scale in-memory data processing, and thus enables user applications achieving three fundamental processing requirements; accuracy, time and cost

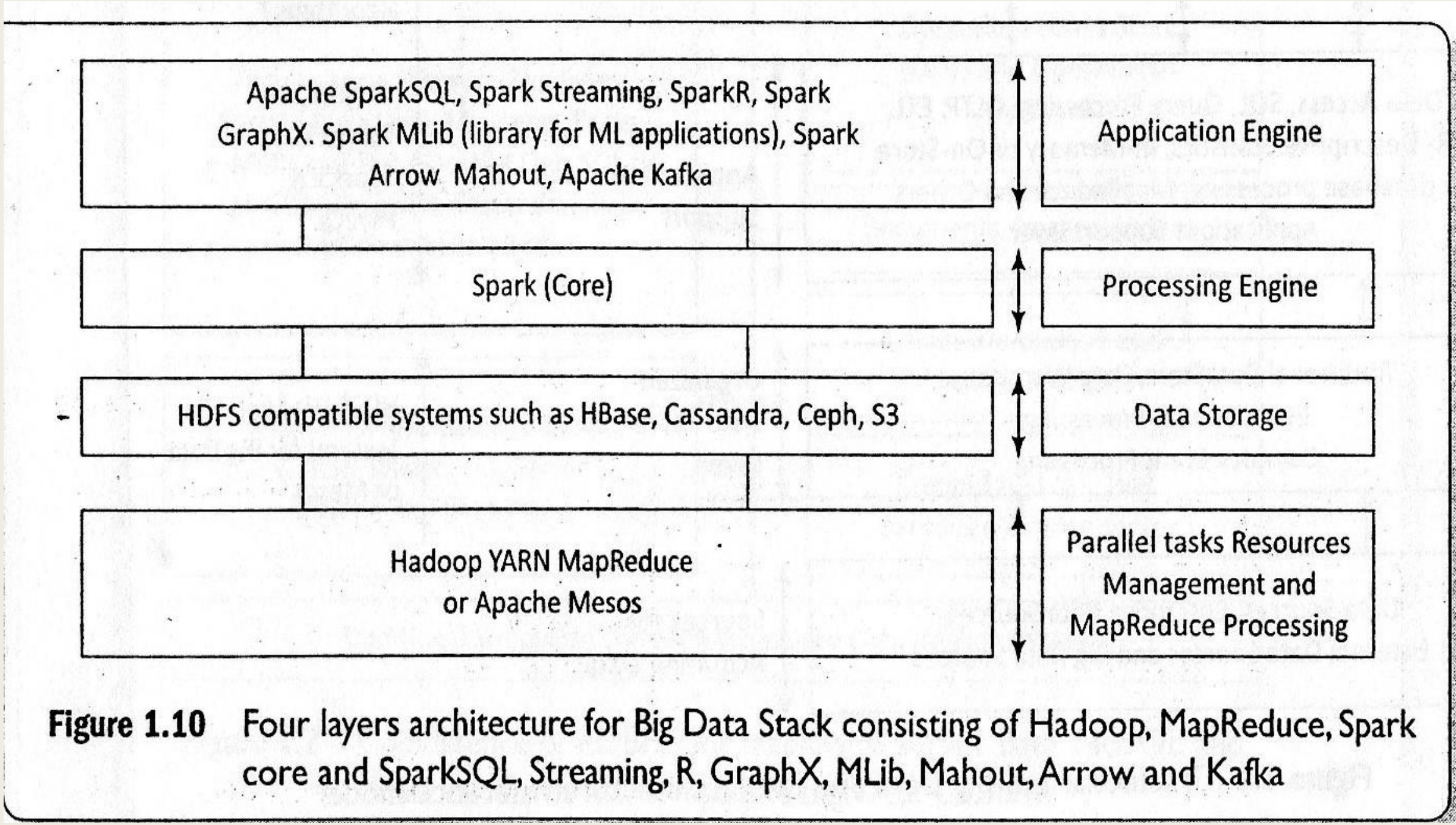


Figure 1.10 Four layers architecture for Big Data Stack consisting of Hadoop, MapReduce, Spark core and SparkSQL, Streaming, R, GraphX, MLib, Mahout, Arrow and Kafka

1.7 Big Data use cases, applications and case studies in various fields, such as marketing, sales and healthcare

- Many applications such a social network and social media, cloud applications, public and commercial web sites, scientific experiments, simulators and e-government services generate Big Data.
- Big Data analytics find applications in many areas.
- Some of the popular ones are **marketing, sales, health care, medicines, advertising etc.**

1.7.1 Big Data in Marketing and Sales

- Data are **important** for most aspect of **marketing and sales**
- **Advertising Customer Value (CV)** depends on three factors –
  - *Quality*
  - *Service*
  - *Price.*
- Big data analytics deploy large volume of data to **identify** and **derive** intelligence using **predictive** models about the **individuals**.
- The facts enable marketing companies to decide what **products to sell**.

- A **definition** of marketing is the **creation, communication and delivery of value to customers.**
- Customer (**desired**) value means what a customer desires from a product.
- Customer (**perceived**) value means what the customer believes to have received from a product after purchase of the product.
- **Customer Value Analytics (CVA)** means analyzing what a customer really needs.
- CVA makes it possible for leading marketers, such as Amazon to deliver the consistent customer experiences.

Following are the **five application areas in order of the popularity of Big Data use cases**:

1. CVA using the inputs of evaluated purchase patterns, preferences, quality, price and post

2. Operational analytics for optimizing company operations

3. Detection of frauds and compliances

   An example of fraud is borrowing money on already mortgage assets. Example of timely compliances means returning the loan and interest installments by the borrowers.

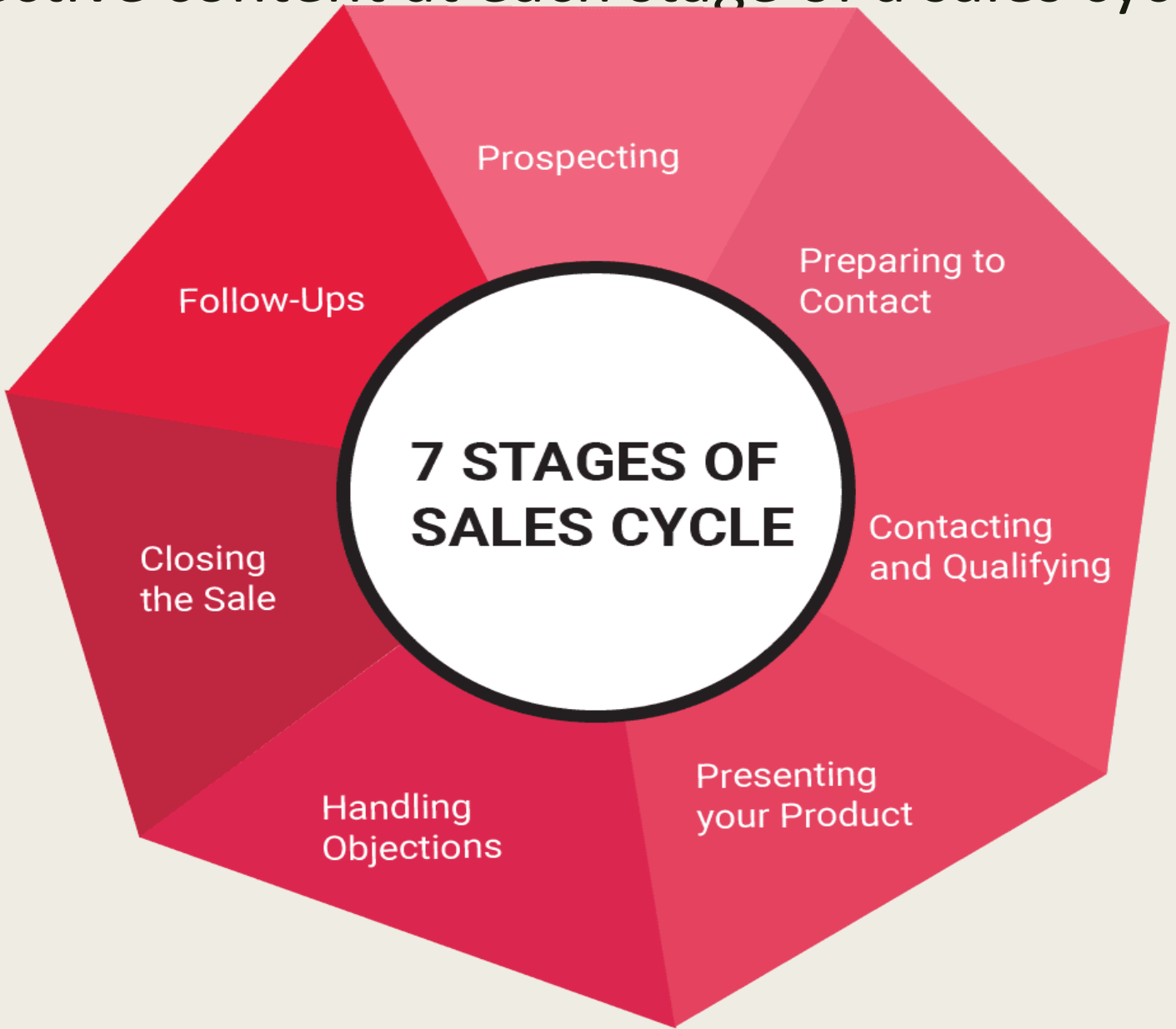
4. New products and innovations in service

   A company develops software and then offers services like Uber.

5. Enterprise data warehouse optimization.

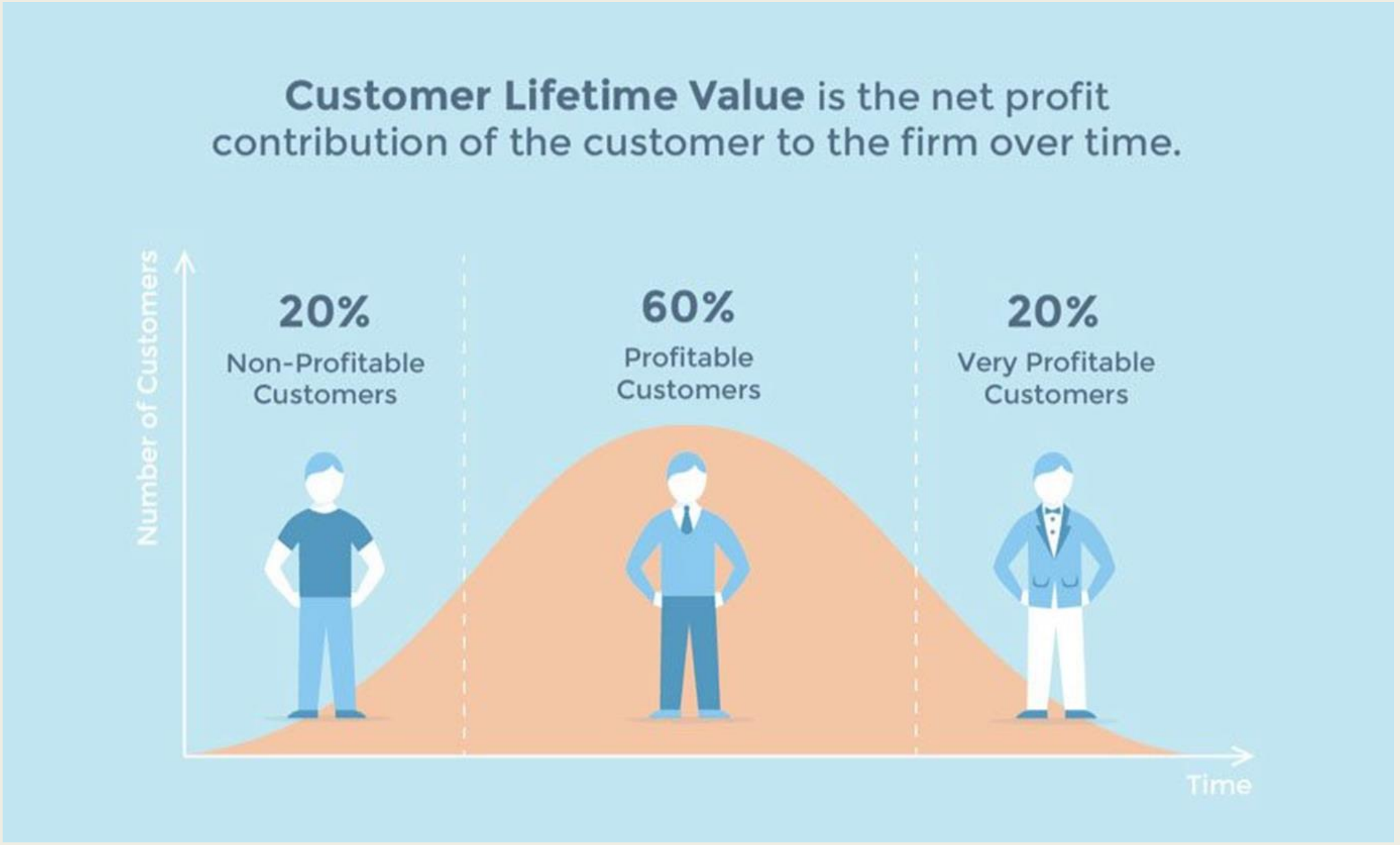
- Big data is providing **marketing insights into**

(i) most effective content at each stage of a sales cycle.

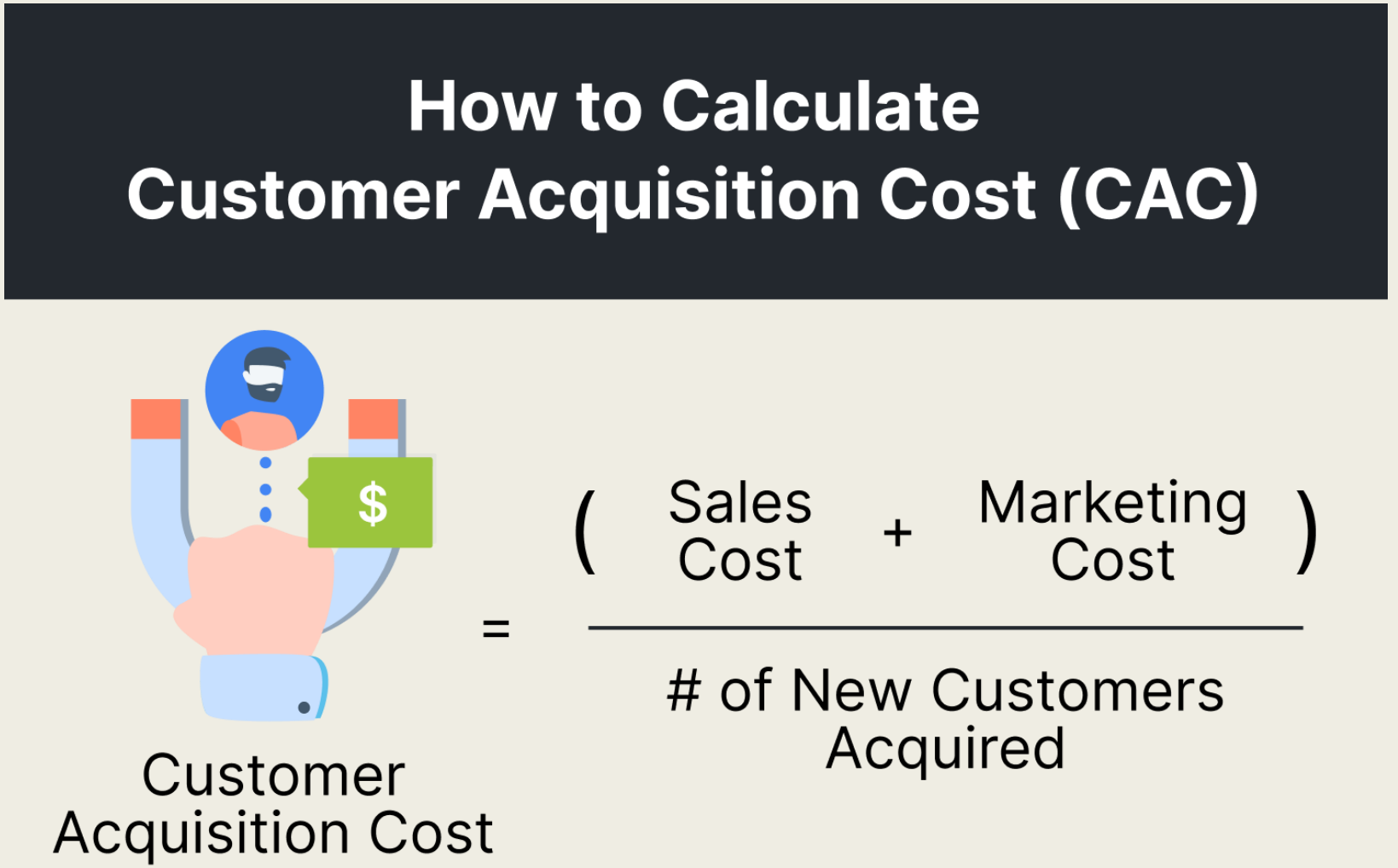


(ii) investment in improving the Customer Relationship Management (CRM),

(iii) addition to strategies for increasing Customer Lifetime Value (CLTV),



(iv) lowering of Customer Acquisition Cost (CAC).



- **Contextual marketing** means using an online marketing model in which a marketer sends to potential customers the targeted advertisements, which are based on the search terms during latest browsing patterns usage by customers.
- For example, if a customer is searching an airline for flights on a specific date from Delhi to Bangalore, then a smart travel agency targeting that customer through advertisements will show him/her, at specific intervals, better options for another airline or different but cheap dates for travel or options in which price reduction occurs gradually.

1.7.1.1 Big Data Analytics in Detection of Marketing Frauds

- Fraud detection is vital to prevent financial loses to users.
- Fraud means someone deceiving deliberately.
- For example,
  - *mortgaging the same assets to multiple financial institutions,*
  - *compromising customer data and transferring customer information to third party,*
  - *falsifying company information to financial institutions,*
  - *marketing product with compromising quality,*
  - *marketing product with service level different from the promised,*
  - *stealing intellectual property, and much more.*

- Big Data usages has the **following features-for** enabling detection and prevention of frauds:

1. **Fusing** of existing data at an enterprise data warehouse with data from sources such as social media, websites, blogs, e-mails, and thus enriching existing data

2. Using **multiple sources** of data and connecting with many applications

3. Providing **greater insights using** querying of the multiple source data

4. Analyzing data which **enable structured reports and visualization**

5. Providing high volume **data mining, new innovative applications and thus leading to new business intelligence and knowledge discovery**

6. Making it **less difficult and faster detection of threats**, and predict likely frauds by using various data and information publicly available.

1.7.1.2 Big Data Risks

- Large volume and velocity of Big Data provide greater insights but also associate risks with the data used, Data included may be **erroneous**, less **accurate** or far from **reality**.
- Analytics introduces **new errors** due to such data.
- Big Data can cause **potential harm to individuals.**
- For example, when **someone puts false or distorted data about an individual in a blog. Facebook post, WhatsApp groups or tweets**, the individual may suffer loss of educational opportunity, job or credit for his/her urgent needs. A company may suffer financial losses.
- Five data risks, described by Bernard Marr are **data security, data privacy breach, costs affecting profits, bad analytics and bad data**."
- Companies need to take risks of using Big Data and **design appropriate risk management procedures.**

1.7.1.3 Big Data Credit Risk Management

- Credit risk is **the possibility of a loss resulting from a borrower's failure to repay a loan or meet contractual obligations**.
- Financial institutions, such as banks, extend loans to industrial and household sectors. These institutions in many countries face credit risks, mainly risks of (i) loan defaults, (ii) timely return of interests and principal amount.

- Financing institutions are keen to get **insights into the following**:

1. Identifying high credit rating **business groups and individuals**,

2. Identifying risk involved **before lending money**

3. Identifying **industrial sectors with greater risks**

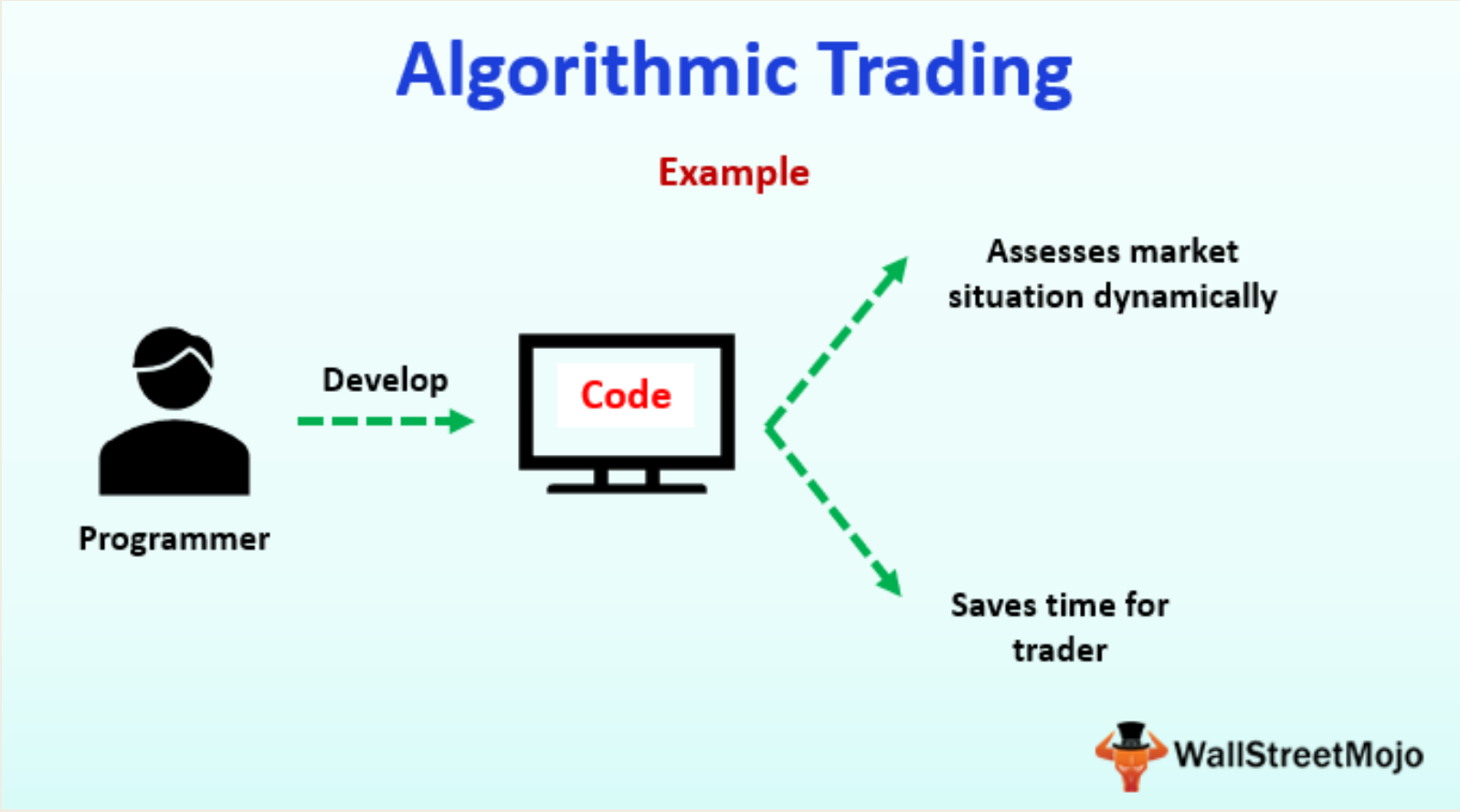
4. Identifying **types of employees** (such as daily wage earners in construction sites) and businesses (such as oil exploration) with greater risks

5. Anticipating **liquidity issues** (availability of money for further issue of credit and rescheduling credit installments) over the years.

- One innovative way to manage credit risks and liquidity risks is use of available data and Big Data.
- Big **Data analytics monitors social media, data,** contact addresses, mobile numbers, website, financial status, activities or job changes to find the emerging credit risk that may affect a customer loan returning capacity.
- The data companies assist in **rating the customer in application** processing and also during the period of repayment of a loan.
- **Friends on Facebook** and their credit rating, comments and assets posted also help in determining the risks.

1.7.1.4 Big Data And Algorithmic Trading

Wikipedia gives a definition of algorithm trading as follows: "**Algorithmic trading is a method of executing a large order (too large to fill all at once) using automated pre-programmed trading instructions accounting for variables such as time, price and volume.**" Complex mathematics computations enable algorithmic trading and business investment decisions to buy and sell.



1.7.2 Big Data and Healthcare

- Big Data analytics in health care use the following data sources:

(i) clinical records

(ii) pharmacy records

(iii) electronic medical records

(iv) diagnosis logs and notes and

(v) additional data, such as deviations from person usual activities, medical leaves from job, social interactions.

- **Healthcare analytics using Big Data can facilitate the following:**

1. **Provisioning of value-based and customer-centric healthcare,**

   Cost effective patient care by improving healthcare quality using latest knowledge, usages of electronic health and medical records and improving coordination among the healthcare providing agencies

2. **Utilizing the 'Internet of Things' for health care**

   Data enables the monitoring of the devices data for patient parameters

3. **Preventing fraud, waste, abuse in the healthcare industry and reduce healthcare costs**

   Examples of frauds are excessive or duplicate claims for clinical and hospital treatments. Example of waste is unnecessary tests. Abuse means unnecessary use of medicines, such as tonics and testing facilities.

4. Improving outcomes

Accurately diagnosing patient conditions, early diagnosis, predicting problems such as congestive heart failure, anticipating and avoiding complications, matching treatments with outcomes and predicting patients at risk for disease or readmission.

5. Monitoring patients in real time.

Machine learning algorithms which process real-time events. They provide physicians the insights to help them make life-saving decisions and allow for effective interventions.

1.7.4 Big Data in Advertising

- The impact of Big Data is tremendous on the digital advertising industry. The digital advertising industry sends advertisements **using SMS, e-mails, WhatsApp, LinkedIn, Facebook, Twitter and other mediums.**

- Big Data technology and analytics provide insights, patterns and models, which relate the **media exposure of all consumers to the purchase activity** of all consumers using multiple digital channels. Big Data help in identity management and can provide an advertising mix for **building better branding exercises**.

- Big Data captures data of multiple sources in large volume, velocity and variety of data unstructured and enriches the structured data at the enterprise data warehouse. Big data real time analytics provide emerging trends and patterns, and gain actionable insights **for facing competitions from similar** products. The data helps digital advertisers to **discover new relationships, lesser competitive regions and areas.**