



Big Data Storage and Processing

MSc in Data Analytics

CCT College Dublin

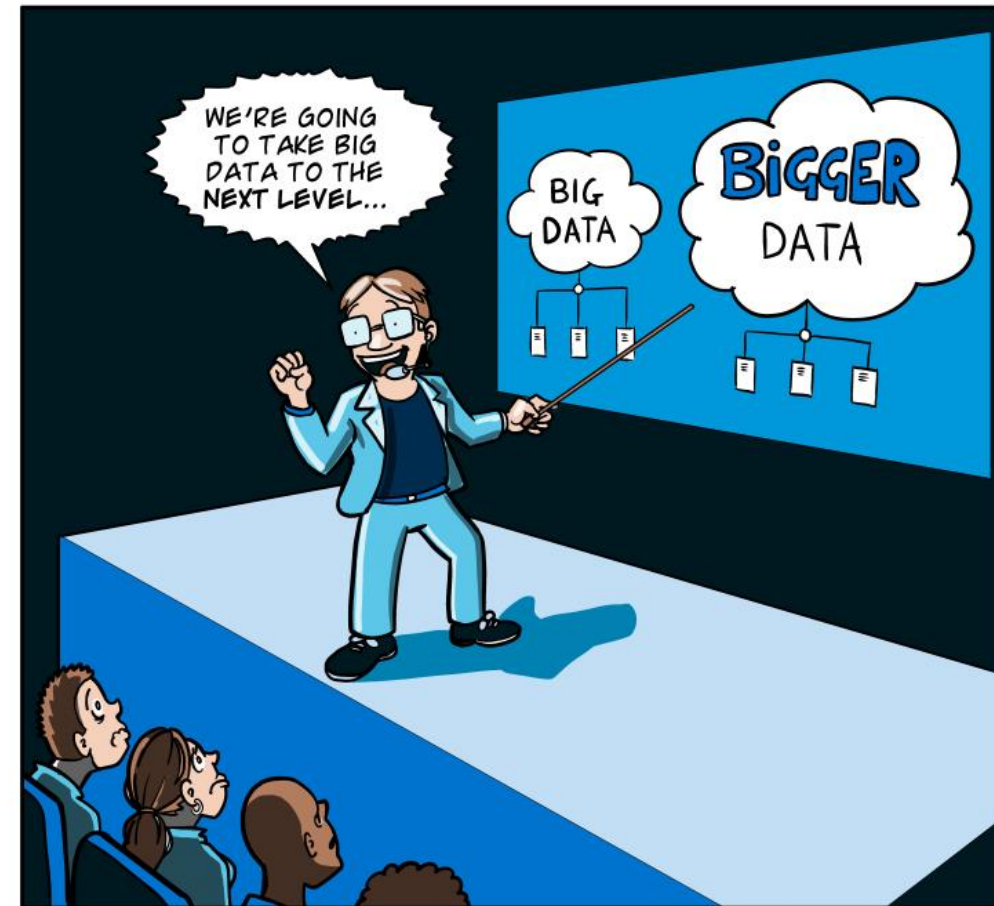
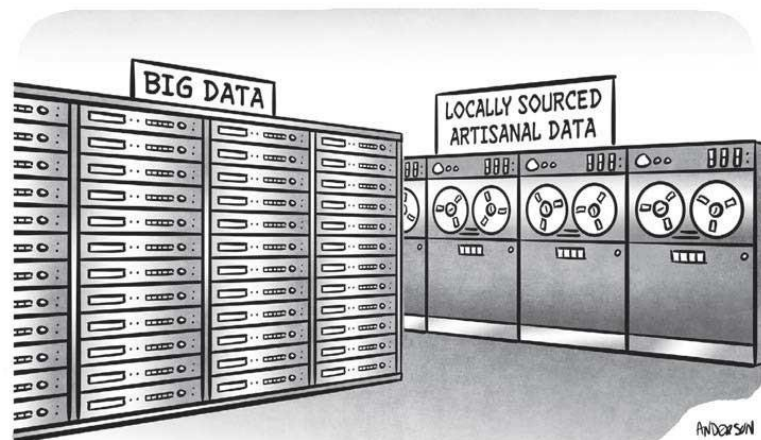
Introduction to Big Data Storage & Processing (BDSP)

Week 1

Lecturer: Dr. Muhammad Iqbal*
Email: miqbal@cct.ie

- Big Data Introduction
- Characteristics and Understanding of Big Data
- Types of Data: Transactional or Analytical
- ACID vs BASE (Relational (Legacy approach) and Non-relational (NoSQL))
- Big Data Architectures & Processing
- Data and Storage Paradigm
- Distributed Computing Overview
- Requirements and Challenges: BIG DATA
- **MLO1:** Critically assess the **data storage and management requirements** of a given data project from a modern perspective and evaluate limitations of legacy approaches to Big Data. (Linked to PLO 3)

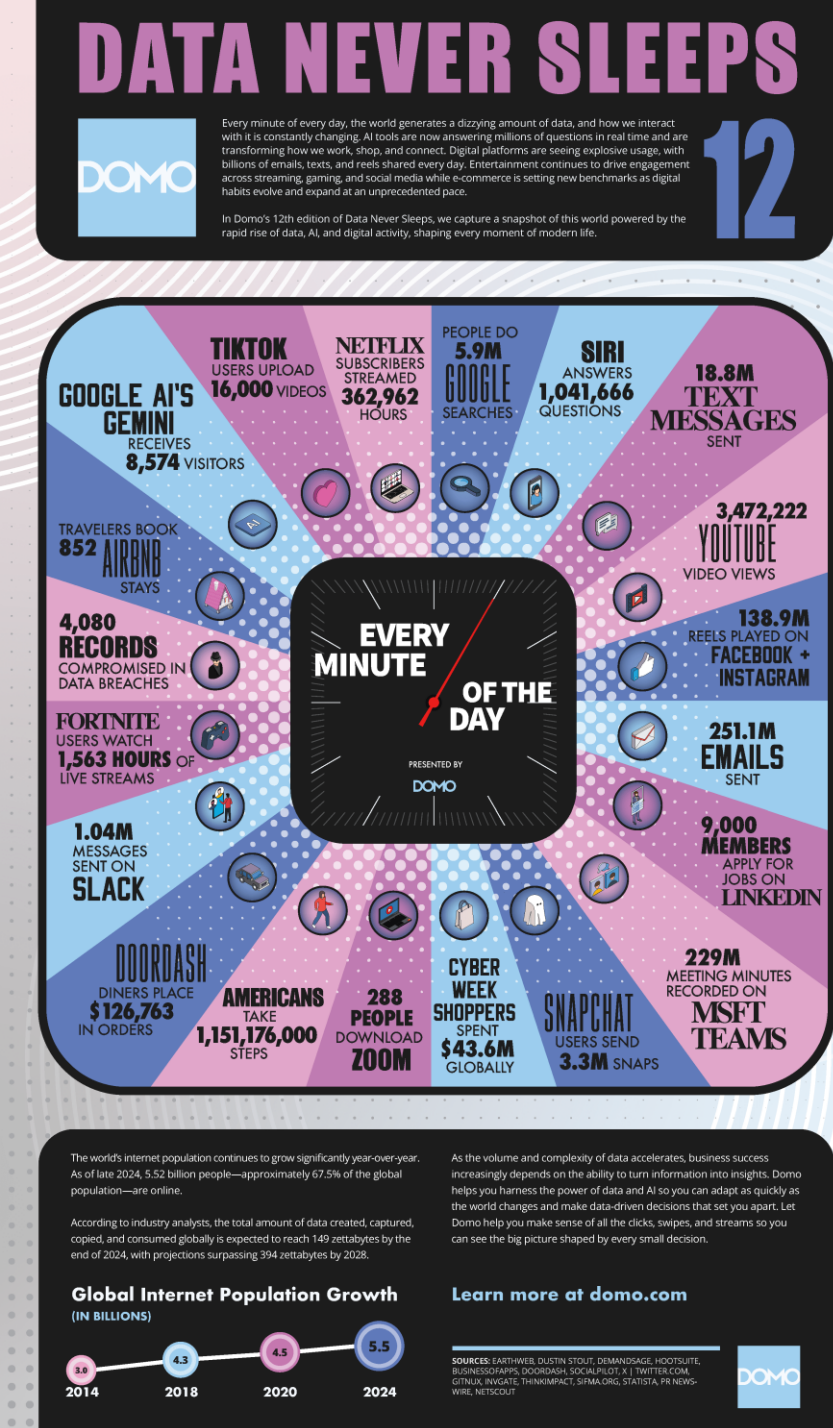
Big Data



12th edition of Data Never Sleep

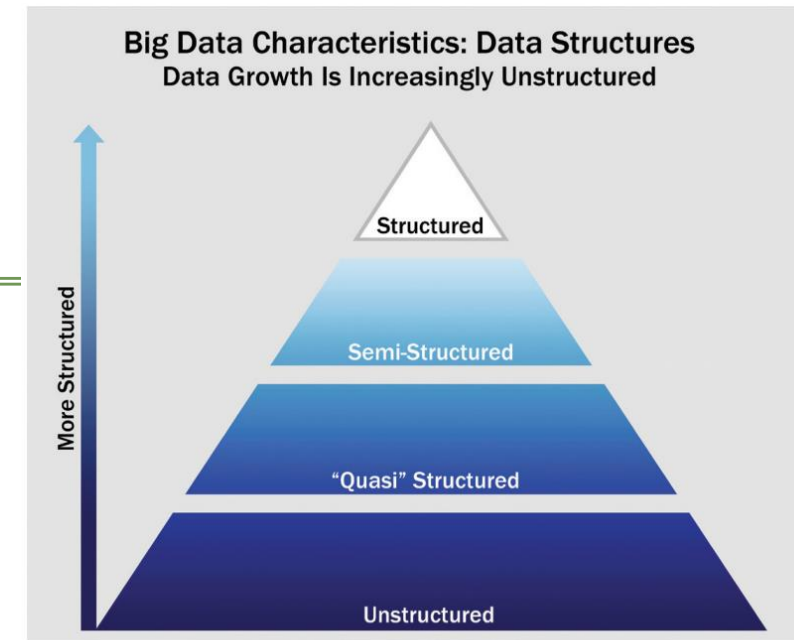
- **Big Data Context**
 - More data is generated and consumed than ever before
 - Increased demands for processing, storage, bandwidth and I/O
 - Expectations increasing and the data will be available whenever and wherever required
 - Technology advancements
 - More data can be stored
 - Less energy required for storage

<https://www.domo.com/learn/infographic/data-never-sleeps-12>

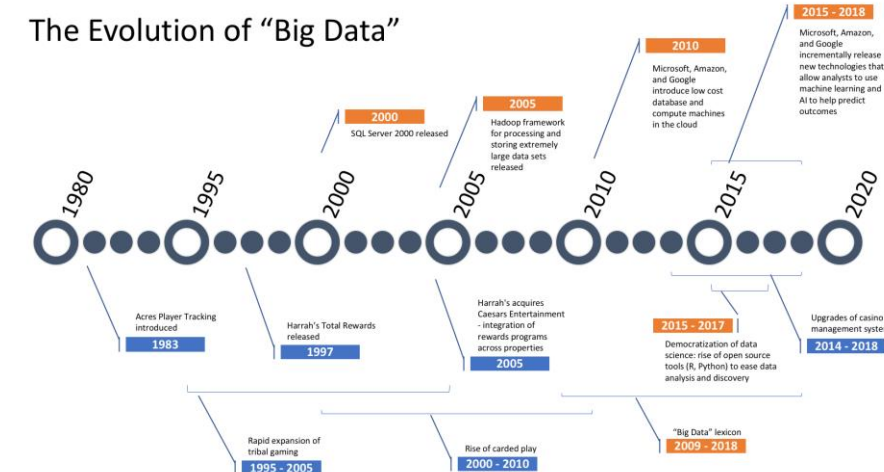


Introduction to Big Data

- **BIG DATA** has been increasingly used in our daily lives. From social networks to mobile applications, and internet search, a huge amount of data is being **generated, collected, and processed**.
- The term "**Big data**" refers to a huge and complicated data collection that is challenging to analyze with typical data processing software or readily available **database management systems (RDBMS)**.
- The data is evolved from the relational database management system storage to NoSQL storage.

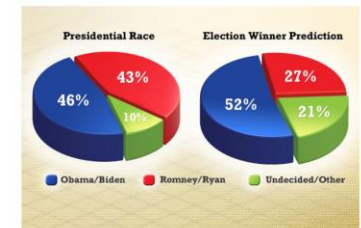


Big Data Growth is increasingly unstructured

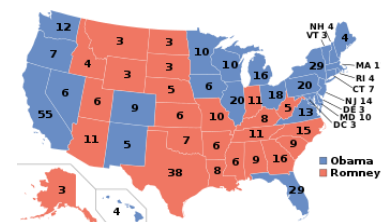


Introduction to Big Data

- In 2006, **LinkedIn**, the social networking giant, started analyzing profiles of its users by suggesting people they may know.
- The aim behind this concept was to motivate users to broaden their social networks based on their interests and provide them with useful ideas. By putting this idea into practice, **LinkedIn** discovered that the majority of its recommendations for inviting people were fruitful.
- In 2012, for US presidential elections, President Obama's campaign experienced massive boost and success through predictive analysis using a Big dataset consisting of voter's profiles, their likes, and their patterns.
- **BIG DATA** has a huge potential for information extraction and it only possible thorough understanding and implementation of available systems that are used for storing, processing, linking, and analyzing.



Presidential Race Question: "Thinking about the presidential election in November, will you vote for Republicans Mitt Romney and Paul Ryan, Democrats Barack Obama and Joe Biden, someone else, or haven't you decided yet?" (includes leavers) Election Prediction Question: "Regardless of who you are voting for, which candidate do you think is most likely to win the 2012 presidential election?" Source: The University of Connecticut/Hartford Courant survey of 1,188 likely voters, Sept. 11-Sept. 18, 2012.



Characteristics of Big Data

- Big data can be characterized by identifying some important characteristics and these are referred to as **five V's of Big data**.
1. **Volume:** Big data refers to the massive volume of data such that the amount of data challenges the storage and processing requirements. From Terabytes (10^{12}) to exabytes (10^{18}).
 2. **Velocity:** Data is being generated at a very fast pace. The high rate of data generation signifies the importance of data.
 3. **Variety:** Data under consideration could be obtained from numerous sources, such as web logs, user tweets, and search patterns etc. Similarly, data could have different formats such as CSV, tables, text documents, and graphs.
 4. **Veracity:** Veracity refers to the trustworthiness, accuracy, or authenticity of data.
 5. **Value:** Data must be of high value; i.e., stale data has limited value.

- The term Big data refers to the huge amount of data such that we could organize, manage, analyze, and understand trends and patterns in data at **volumes** and **rates** that push the frontiers of current technologies.
- One of the fundamental questions in Big data is that for a given Big data problem in consideration
 - **How much data is enough?**
 - **How much data is needed to be analyzed in order to compute the result?**
- The answer to this question is not trivial. For example, opinion polls are based on data samples. In a similar context, gender-wise assessment and population are based on data sampling. Sampling increases the chance of error.

Questions and Discussion

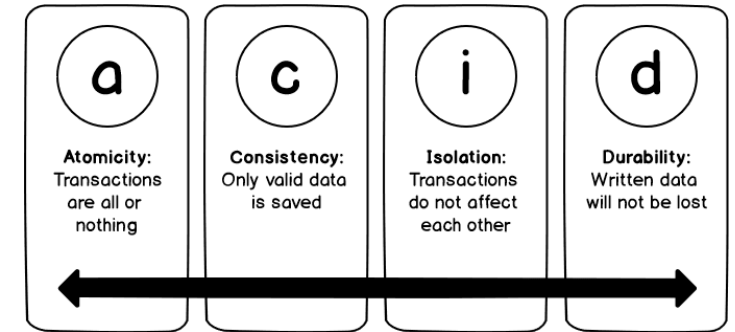
- On Data 12th edition of Never Sleeps illustration, In every minute, how many users sent text messages?
- On Facebook and Instagram, how many reels played on?
- How many (3 or 5 or 7) characteristics a Big data has?
- What is your understanding on Big Data now?

Types of Data

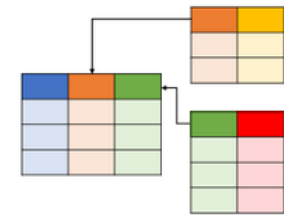
Transactional or Analytical



- 1. Transactional Systems:** These are the types of systems which support transaction processing. Such kind of systems adhere to **ACID (Atomicity, Consistency, Isolation, and Durability)** properties. They have proper **schema** and data for each transaction is uniquely identified.
 - 2. Analytical Systems:** Data does not necessarily adhere to a proper schema. It may have duplicates and missing values etc. Such systems are more appropriate for analyzing data.
- The term Big data has been associated for analytical systems because such systems do not require strong consistency and have schema-less data with duplicates, multi-formatting, and missing values.



SQL DATABASES



Relational

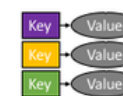
NoSQL DATABASES



Column



Graph



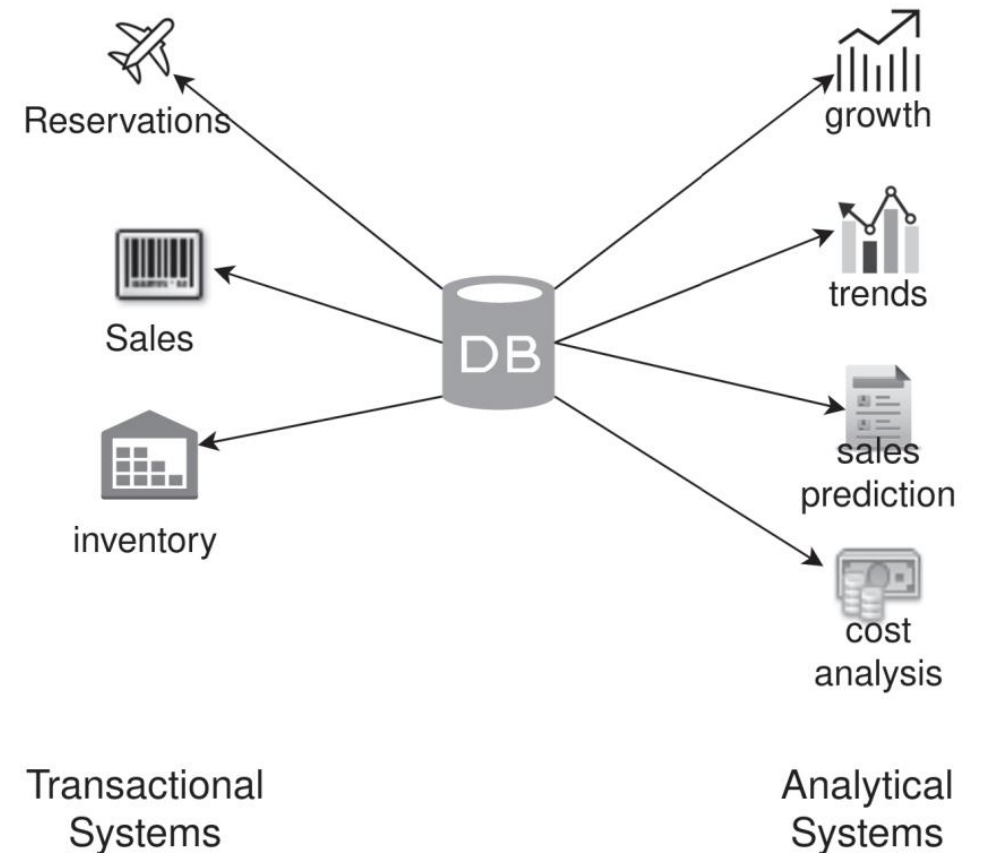
Key-Value



Document

ACID vs BASE

- For distributed systems, meeting **ACID** guarantees is really challenging. Therefore, many Big data systems employ **BASE** properties.
- **BASE** is an acronym for **Basically Available Soft state Eventual consistency**.
- **BASE** implies that in case of network failure, Big data systems tend to compromise on **consistency** in order to provide **availability**.
- The main focus of such systems is to ensure **availability**, whereas **eventual consistency** model is followed.



Transactional systems vs. analytical systems

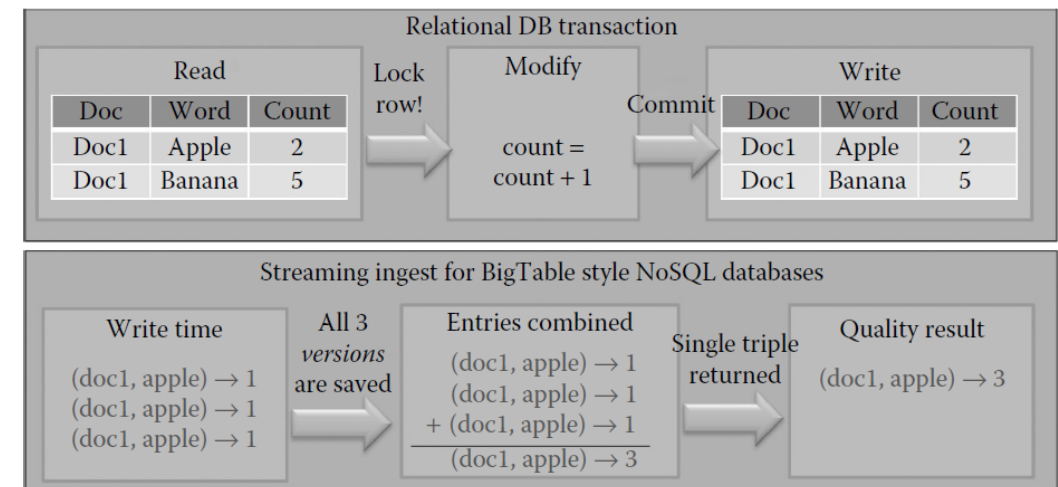
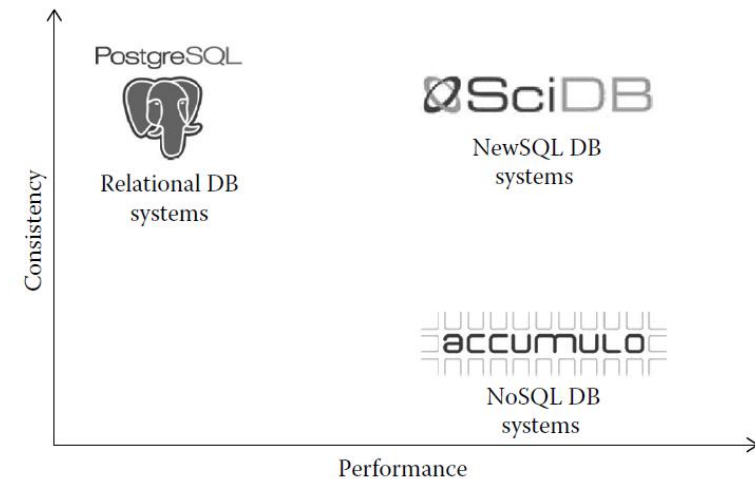
Databases

Relational and Non-relational

	Relational Databases	NoSQL	NewSQL
Examples	MySQL, PostgreSQL, Oracle	HBase, Cassandra, Accumulo	SciDB, VoltDB, MemSQL
Schema	Typed columns with relational keys	Schema-less	Strongly typed structure of attributes
Architecture	Single-node or sharded	Distributed, scalable	Distributed, scalable
Guarantees	ACID transactions	Eventually consistent	ACID transactions (most)
Access	SQL, indexing, joins, and query planning	Low-level API (scans and filtering)	Custom API, JDBC, bindings to popular languages

A simple guide to differentiate between SQL, NoSQL, and NewSQL style databases.

- Traditional relational databases provide high **consistency**.
- NoSQL** databases provide high **performance** at the cost of **consistency**, and **NewSQL** databases attempt to bridge the gap.



Relational update transaction compared with nonrelational database update transaction.

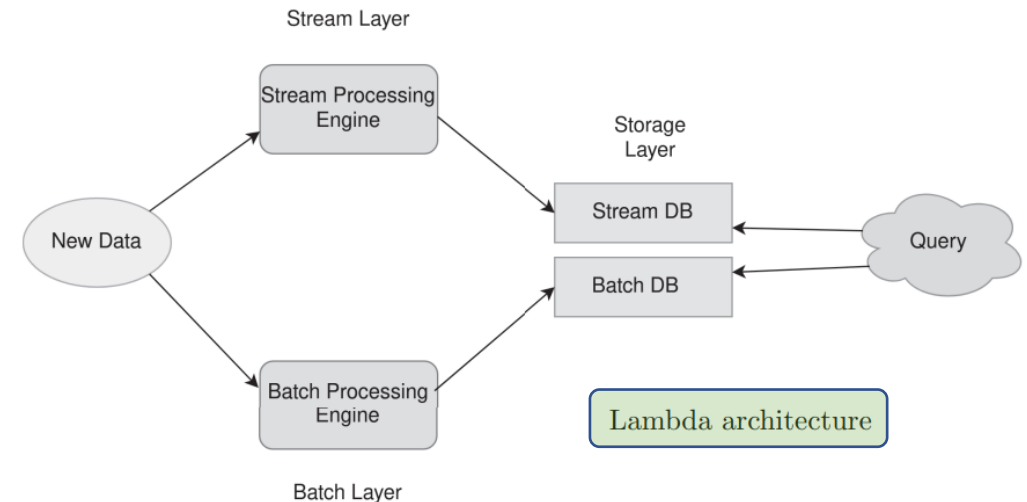
Questions and Discussion

- How many kinds of major databases you have identified so far?
- What are the differences between ACID and BASE?

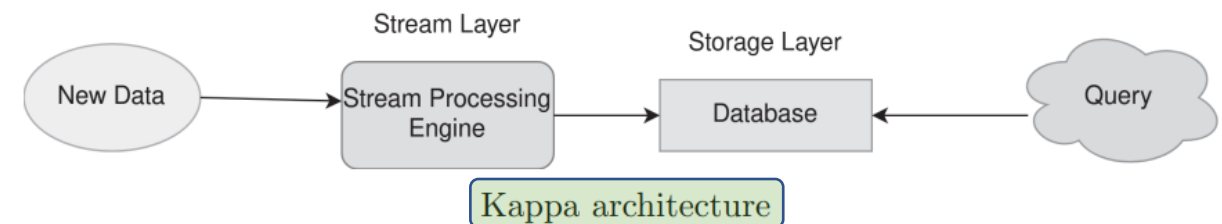
Big Data Architectures

- Different architectures can be used for Big data processing.
- We introduce two major architectures used in Big data.
- An efficient real time data processing architecture needs to be **scalable** and **fault-tolerant** and it should support batch and incremental updates. The kinds of architectures are mentioned below

1. Lambda Architecture



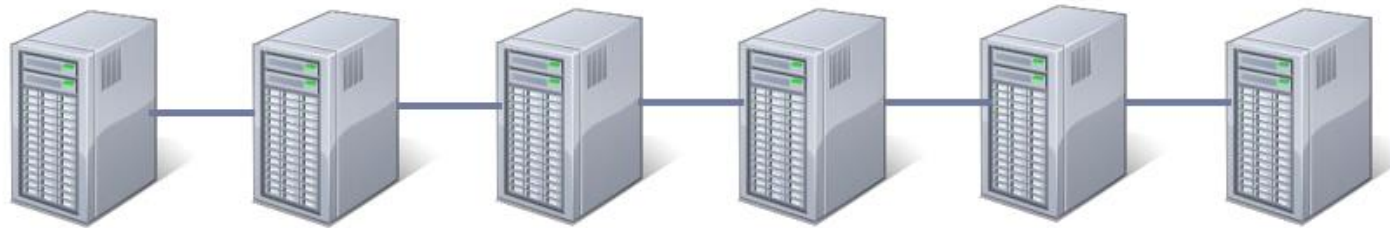
2. Kappa Architecture



- **How can we effectively address the challenges associated with BIG Data?**
- Build Bigger and more powerful machines
- Although technological advancements have made possible for the storage of vast amounts of data, there are still some technological limitations that impact on how that data can be used.
- **Typical Hard Drive from 1990**
- 1370 MB capacity
 - 4.4 MB/s transfer speed
 - **Five minutes required to read a full drive**
- **2022**
- 1 TB capacity
 - 100 MB/s transfer speed or 200 MB/s
 - **2.9 hours or 1.45 hours required to read full drive**

Big Data Processing

- How can we effectively address the challenges associated with **BIG Data**?
- Use a collection of reasonably powerful machines in concert



- Various Cloud platforms are available to help with processing BIG DATA



Big Data Processing

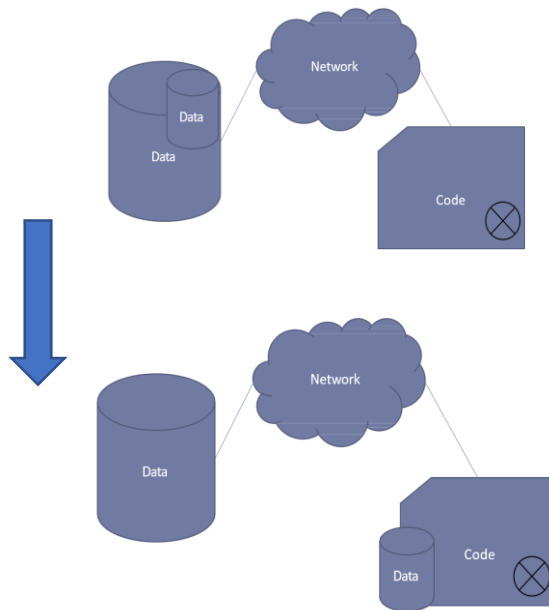
Distributed Systems

- The most popular Big data processing technique using clusters of computers is **MapReduce**.
 - Is a **programming model** that can be implemented using **HADOOP** framework and others.
 - Has two steps: **Map & Reduce**
 - Divide the data into chunks and split them by the computers into the clusters.
- **Expected characteristics of Distributed Systems**
 - Resource sharing
 - Openess
 - Concurrency
 - Scalability
 - Fault tolerance
 - Transparency

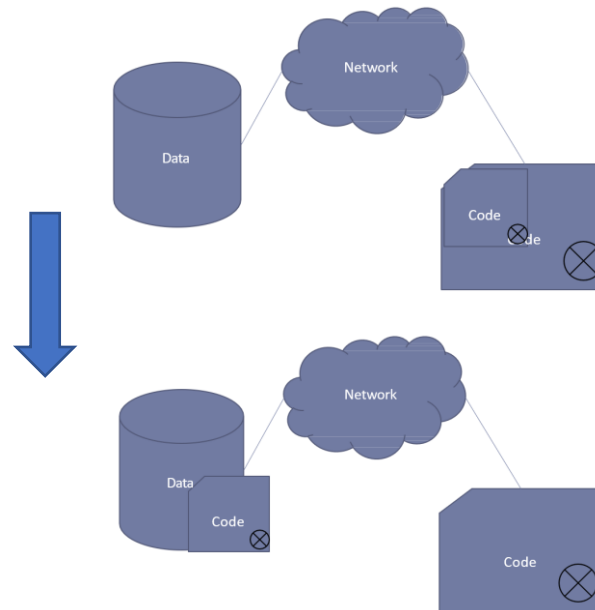
Data and Storage Paradigm

Storage / Compute Locality

- **Scenario 1**



- **Scenario 2**



- **Which Scenario is preferable?**
- Move the processing elements (e.g., execution of code) to the locality of the data rather than moving the data for processing.

- Co-locate the data and processing (a.k.a. compute)
- Minimise data transfer
- Limit bulk transfer of data

Distributed Computing Overview

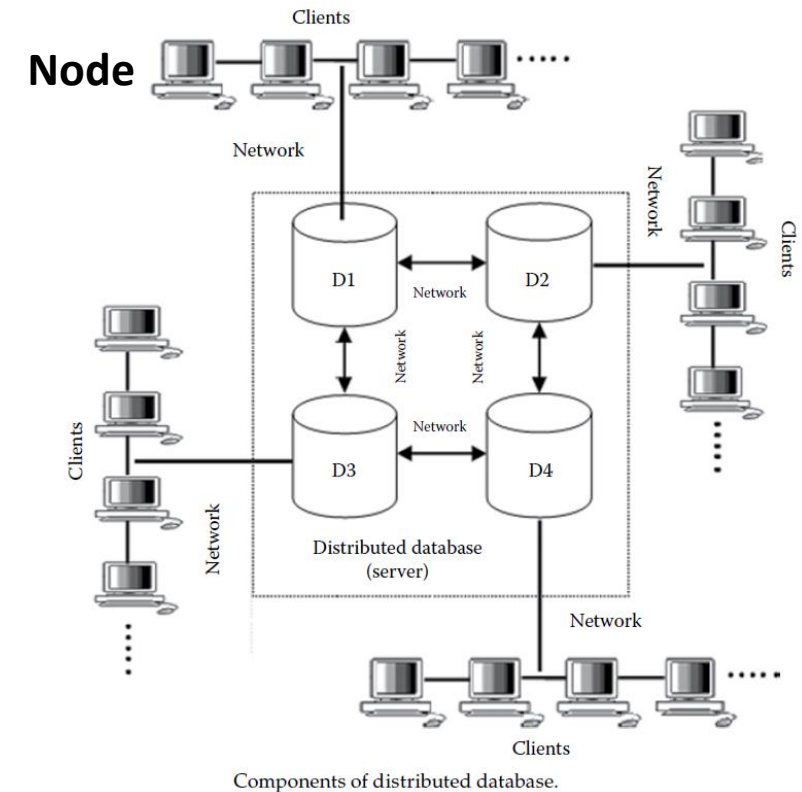
**Parallel
Computing**

**Distributed
Computing**

- Parallel computing uses multiple compute resources acting simultaneously to solve a computational problem
- There are overlaps between **parallel** computing and **distributed** computing
- **For example,**
 - The processors in the computers in a distributed system use parallelism
 - Distributed systems can be used to solve parts of problems in parallel

Distributed Databases

- A database server is the software that administers a database, and a client is an application that requests information and seeks services from a server.
- **Distributed processing (DP)** refers to the use of more than one computer (or processor) to run an application and perform the processing for an individual task.
- **DP** refers to local area networks (LANs) designed so that a single program can run simultaneously at various sites.
- **DP** is composed of distributed databases, wherein the data are stored across different computer systems.



- The main goal of **DP** system is to connect users and resources in a transparent, open, and scalable way.

Parallel DBMS and DDBMS

- **Parallel Database Management Systems** refers to the management of data in a tightly coupled multiprocessor computer and is done by a full-fledged DBMS.

Parallel DBMS versus DDBMS

Parallel DBMS	DDBMS
Machines are physically located close to each other, for example, same server room.	Machines can be located far-off from each other, for example, in diverse continent.
Machines connect with dedicated high-speed LANs and switches.	Machines can be connected using public-purpose network, for example, Internet.
Communication expenditure is assumed to be small.	Communication expenditure and predicaments cannot be ignored.
Can employ shared-memory, shared-disk, or shared-nothing architecture.	Usually employs shared-nothing architecture.

Requirements and Challenges

BIG DATA

1. Scalability
2. Availability and Fault Tolerance
3. Efficient Network Setup
4. Flexibility
5. Privacy and Access Control
6. Elasticity
7. Batch Processing and Interactive Processing
8. Efficient Storage
9. Multi-tenancy
10. Efficient Processing
11. Efficient Scheduling



Big Data Research
Volume 27, 28 February 2022, 100289



Big Data in Forecasting Research: A Literature Review

Ling Tang ^{a, b}, Jieyi Li ^a, Hongchuan Du ^c, Ling Li ^d , Jun Wu ^a, Shouyang Wang ^e

Show more

Add to Mendeley Share Cite

<https://doi.org/10.1016/j.bdr.2021.100289>

[Get rights and content](#)

Highlights

- A review on full-scale big data in forecasting research is presented.
- Big data in forecasting research fell into UGC data, device data and log data.
- Three steps were taken: data collection, data process and prediction improvement.
- For each data type, what, where and how big data improved prediction are detailed.

- **Research Article (pdf copy available on Moodle)**
- <https://www.sciencedirect.com/science/article/pii/S2214579621001064>

Questions and Discussion

- How many architectures are present in Big Data processing?
- Which data and storage paradigm (Storage/ Compute) is considered as useful for Big Data Storage and Processing applications?
- Express your understanding in your own words for parallel and distributed computing. Which programming model is useful for distributed processing using a Big data framework?

- Big Data: Concepts, Technology, and Architecture, Balamurugan Balusamy, Nandhini Abirami R, Seifedine Kadry, Amir H. Gandomi, ISBN: 978-1-119-70187-3, March 2021, 368 pages.
- Big Data Systems: A 360-degree Approach, Jawad Ahmed Shamsi, Muhammad Ali Khojaye, CRC Press, 2021.
- NoSQL Database for Storage and Retrieval of Data in Cloud, Ganesh Chandra Deka, CRC Press, Taylor & Francis Group, 6000 Broken Sound Parkway NW, 2017.
- Big Data: Principles And Best Practices Of Scalable Real-time Data Systems, Nathan Marz, James Warren, Manning Publications, 2015.
- Greg Schulz, 2012, Cloud and Virtual Storage Networking, CRC Press.
- G. Somasundaram, A Shrivastava (Editors), 2009, Information Storage and Management: Storing, Managing, and Protecting Digital Information, Wiley Publishing
- Some images are used from Google search repository.