Data Storage & Processing
Data Analysis
Data Warehouse
Big Data Tools
No sequel Database
Realtime Processing
Messaging System

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

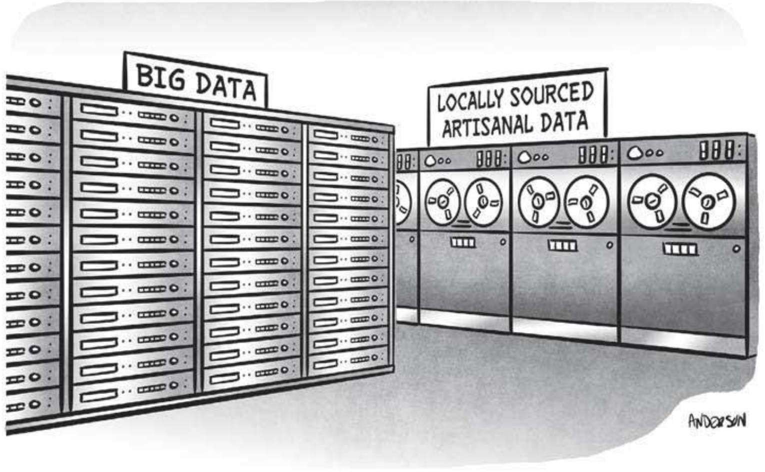
Agenda

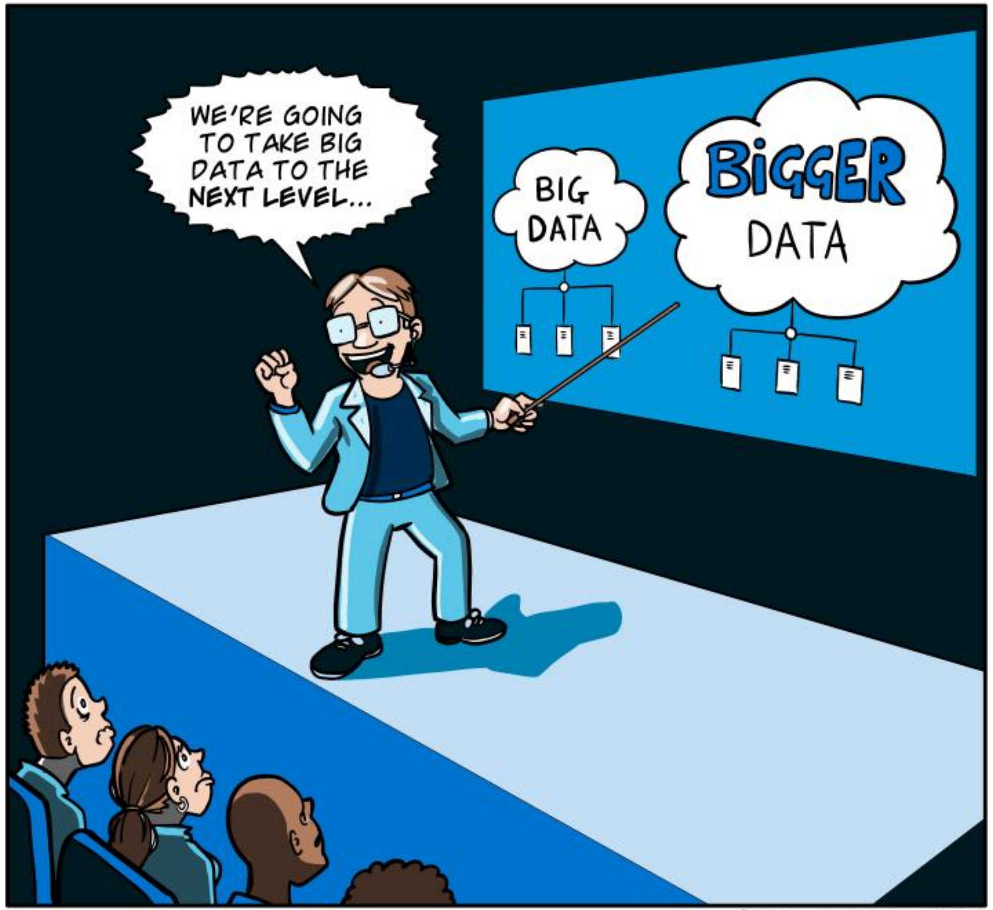
- Big Data Introduction
- Characteristics and Understanding of Big Data
- Types of Data: Transactional or Analytical
- ACID vs BASE (Relational (Legacy approach) and Non-relational (NoSQL))
- Big Data Architectures & Processing
- Data and Storage Paradigm
- Distributed Computing Overview
- Requirements and Challenges: BIG DATA

Big Data

LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISH BURNE

© marketoonist.com



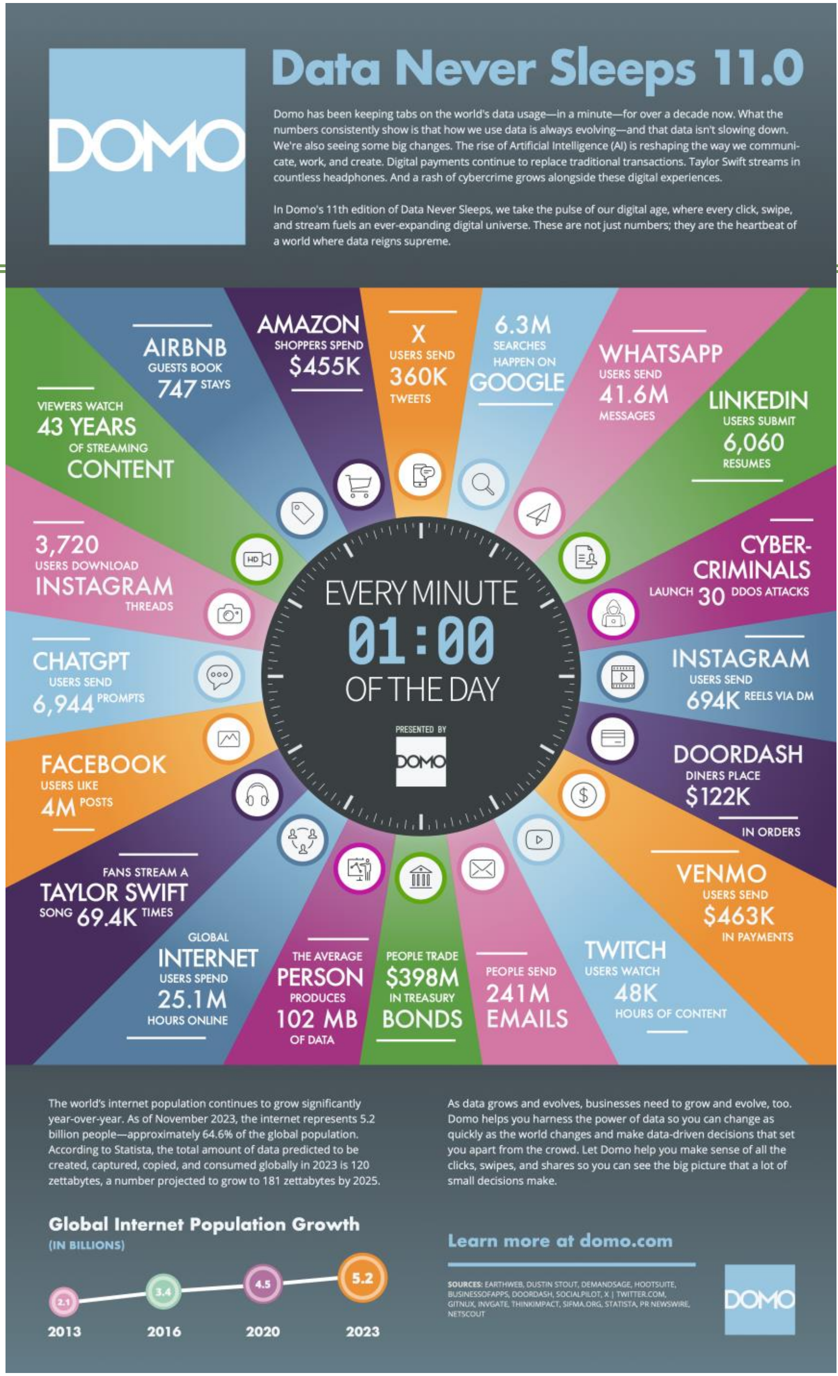


© CLOUDTWEAKS.COM

11th 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-11



Data Never Sleeps 11.0

Domo has been keeping tabs on the world's data usage—in a minute—for over a decade now. What the numbers consistently show is that how we use data is always evolving—and that data isn't slowing down. We're also seeing some big changes. The rise of Artificial Intelligence (AI) is reshaping the way we communicate, work, and create. Digital payments continue to replace traditional transactions. Taylor Swift streams in countless headphones. And a rash of cybercrime grows alongside these digital experiences.

In Domo's 11th edition of Data Never Sleeps, we take the pulse of our digital age, where every click, swipe, and stream fuels an ever-expanding digital universe. These are not just numbers; they are the heartbeat of a world where data reigns supreme.

DOMO
AIRBNB GUESTS BOOK 747 STAYS
AMAZON SHOPPERS SPEND $455K
X USERS SEND 360K TWEETS
6.3M SEARCHES HAPPEN ON GOOGLE
WHATSAPP USERS SEND 41.6M MESSAGES
LINKEDIN USERS SUBMIT 6,060 RESUMES
VIEWERS WATCH 43 YEARS OF STREAMING CONTENT
3,720 USERS DOWNLOAD INSTAGRAM THREADS
CYBER-CRIMINALS LAUNCH 30 DDOS ATTACKS
CHATGPT USERS SEND 6,944 PROMPTS
INSTAGRAM USERS SEND 694K REELS VIA DM
EVERY MINUTE 01:00 OF THE DAY
PRESENTED BY DOMO
FACEBOOK USERS LIKE 4M POSTS
DOORDASH DINERS PLACE $122K IN ORDERS
FANS STREAM A TAYLOR SWIFT SONG 69.4K TIMES
VENMO USERS SEND $463K IN PAYMENTS
GLOBAL INTERNET USERS SPEND 25.1M HOURS ONLINE
THE AVERAGE PERSON PRODUCES 102 MB OF DATA
PEOPLE TRADE $398M IN TREASURY BONDS
PEOPLE SEND 241M EMAILS
TWITCH USERS WATCH 48K HOURS OF CONTENT

The world's internet population continues to grow significantly year-over-year. As of November 2023, the internet represents 5.2 billion people—approximately 64.6% of the global population. According to Statista, the total amount of data predicted to be created, captured, copied, and consumed globally in 2023 is 120 zettabytes, a number projected to grow to 181 zettabytes by 2025.

As data grows and evolves, businesses need to grow and evolve, too. Domo helps you harness the power of data so you can change as quickly as the world changes and make data-driven decisions that set you apart from the crowd. Let Domo help you make sense of all the clicks, swipes, and shares so you can see the big picture that a lot of small decisions make.

Global Internet Population Growth
(IN BILLIONS)

Learn more at domo.com

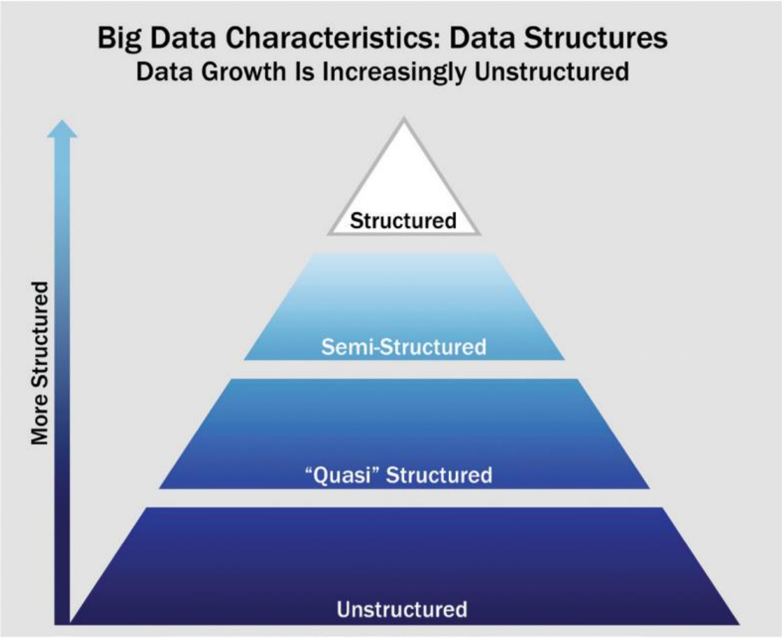
SOURCES: EARTHWEB, DUSTIN STOUT, DEMANDSAGE, HOOTSUITE, BUSINESSOFAPPS, DOORDASH, SOCIALPILOT, X | TWITTER.COM, GITNUX, INVGATE, THINKIMPACT, SIFMA.ORG, STATISTA, PR NEWSWIRE, NETSCOUT

Introduction to Big Data

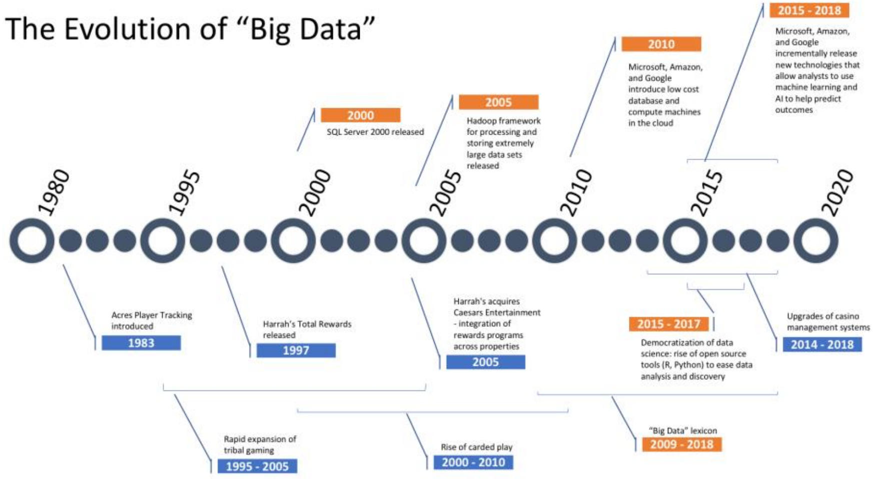
- **BIG DATA** has been increasingly used in our daily lives. From social networks to mobile applications, and internet search, a huge amount of data is being **generated**, **collected**, and **processed**.

- 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 Characteristics: Data Structures
Data Growth Is Increasingly Unstructured
More Structured
Structured
Semi-Structured
"Quasi" Structured
Unstructured

Big Data Growth is increasingly unstructured


The Evolution of "Big Data"
1980
1995
2000
2005
2010
2015
2020
2000
SQL Server 2000 released
2005
Hadoop framework for processing and storing extremely large data sets released
2010
Microsoft, Amazon, and Google introduce low cost database and compute machines in the cloud
2015 - 2018
Microsoft, Amazon, and Google incrementally release new technologies that allow analysts to use machine learning and AI to help predict outcomes
Acres Player Tracking introduced
1983
Harrah's Total Rewards released
1997
Harrah's acquires Caesars Entertainment - integration of rewards programs across properties
2005
2015 - 2017
Democratization of data science: rise of open source tools (R, Python) to ease data analysis and discovery
Upgrades of casino management systems
2014 - 2018
Rapid expansion of tribal gaming
1995 - 2005
Rise of carded play
2000 - 2010
"Big Data" lexicon
2009 - 2018

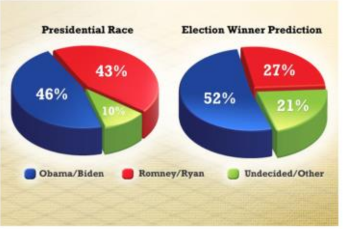
Introduction to Big Data

- In 2006, **LinkedIn,** the social networking giant, started analyzing profiles of its users by suggesting people they may know.

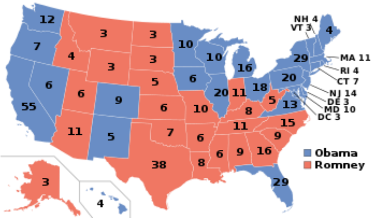
- The aim behind this concept was to motivate users to broaden their social networks based on their interests and provide them with useful ideas. By putting this idea into practice, **LinkedIn** discovered that the majority of its recommendations for inviting people were fruitful.

- 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 leaners) 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 1,186 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})**. For example, traffic monitoring, Weather and Environmental Sensors etc.

2. **Velocity:** Data is being generated at a very fast pace. The high rate of data generation signifies the importance of data. For example, Social Media Real-time Analytics.

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. For example, health care data quality, sensor data reliability etc.

5. **Value:** Data must be of high value; i.e., stale data has limited value. For example, Retail customer data: Personalized Marketing, Demand Forecasting etc.

Understanding Big Data

- The term big data refers to the huge amount of data such that we could organize, manage, analyze, and understand 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 11th edition of Never Sleeps illustration, In every minute, how many users posts on Twitter?

- On Facebook, how many users share photographs on Facebook?

- How many (3 or 5 or 7) characteristics a Big data has?

- What is your understanding on Big Data now?

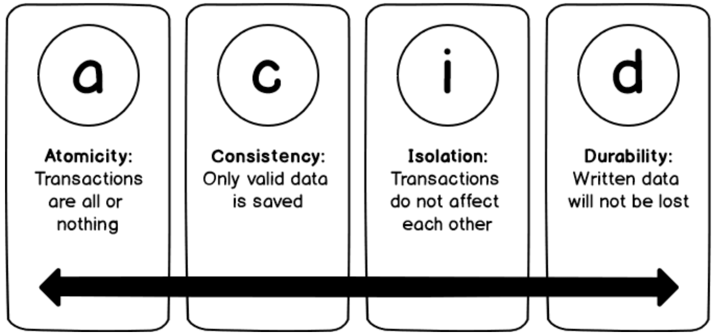
Types of Data

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.



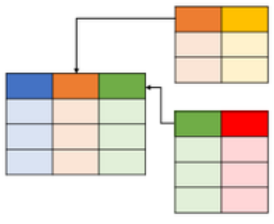
Atomicity: Transactions are all or nothing

Consistency: Only valid data is saved

Isolation: Transactions do not affect each other

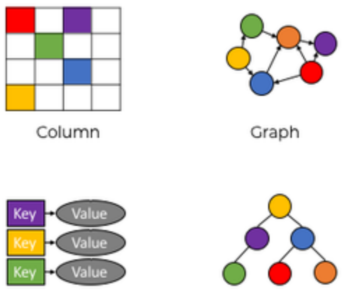
Durability: Written data will not be lost

SQL DATABASES



Relational

NoSQL DATABASES



Column

Graph

Key-Value

Document

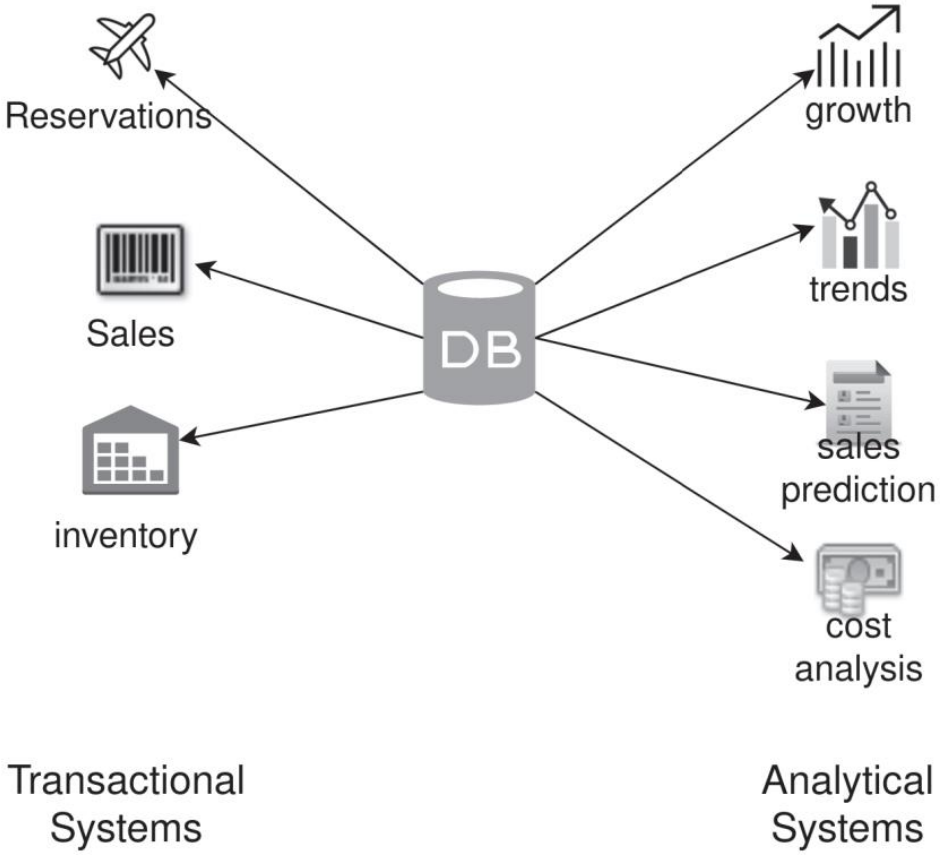
ACID vs BASE

- 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**.

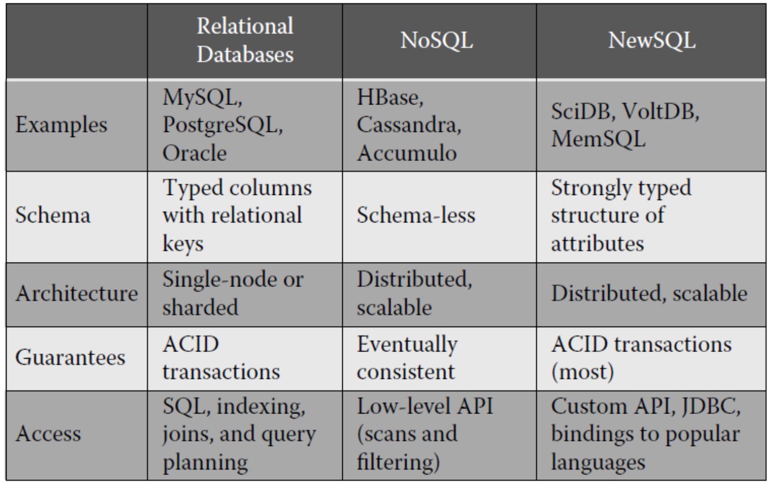
- The main focus of such systems is to ensure **availability**, whereas **eventual consistency** model is followed.



Transactional systems vs. analytical systems

Databases

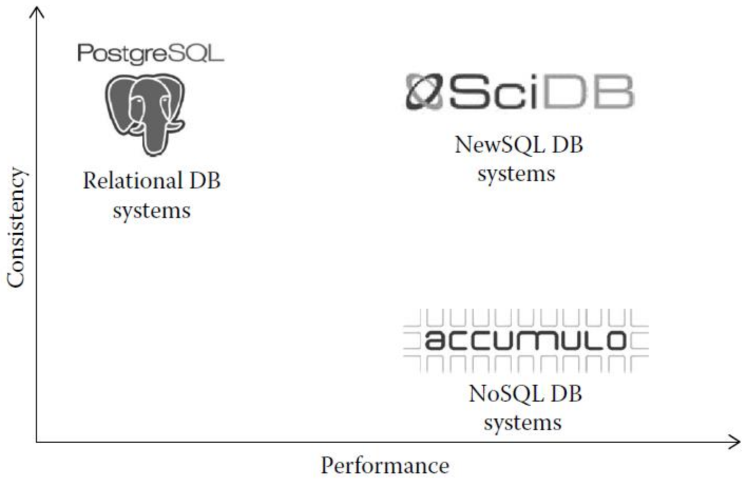
Relational and Non-relational

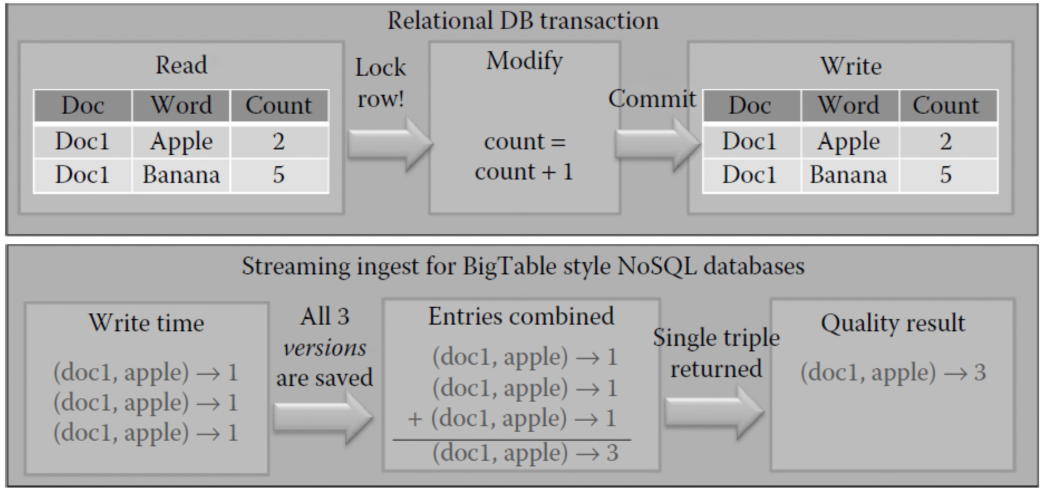


	Relational Databases	NoSQL	NewSQL
Examples	MySQL, PostgreSQL, Oracle	HBase, Cassandra, Accumulo	SciDB, VoltDB, MemSQL
Schema	Typed columns with relational keys	Schema-less	Strongly typed structure of attributes
Architecture	Single-node or sharded	Distributed, scalable	Distributed, scalable
Guarantees	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.





Read:

Doc	Word	Count
Doc1	Apple	2
Doc1	Banana	5

Write:

Doc	Word	Count
Doc1	Apple	2
Doc1	Banana	5

Write time:
(doc1, apple) → 1
(doc1, apple) → 1
(doc1, apple) → 1

Entries combined:
(doc1, apple) → 1
(doc1, apple) → 1
+ (doc1, apple) → 1
(doc1, apple) → 3

Quality result:
(doc1, apple) → 3

Relational update transaction compared with nonrelational database update transaction.

Questions and Discussion

- How many kinds of major databases you have identified so far?

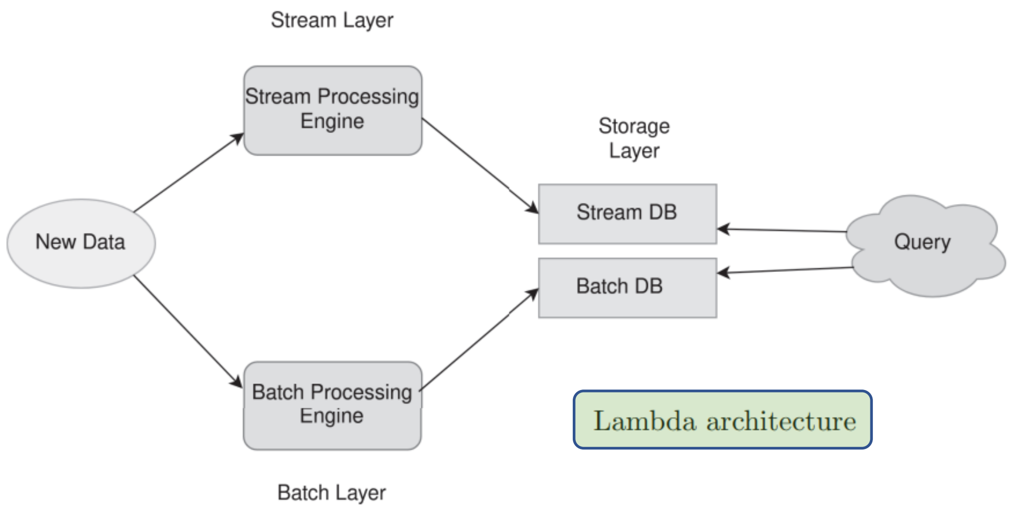
- What are the differences between ACID and BASE?

Big Data Architectures

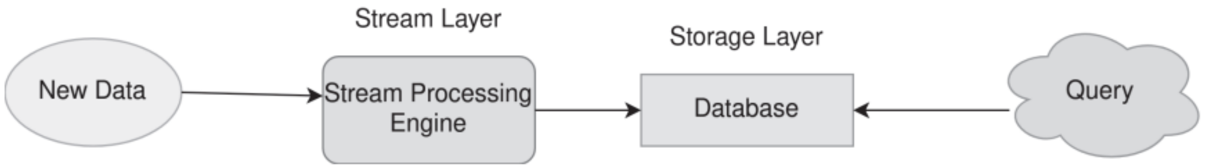
- Different architectures can be used for big data processing.

- 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



Lambda architecture



Kappa architecture

1. **Lambda Architecture**

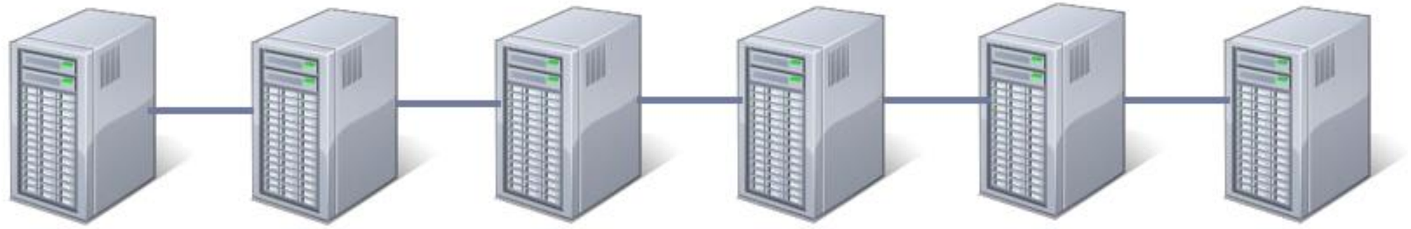
2. **Kappa Architecture**

Big Data Processing

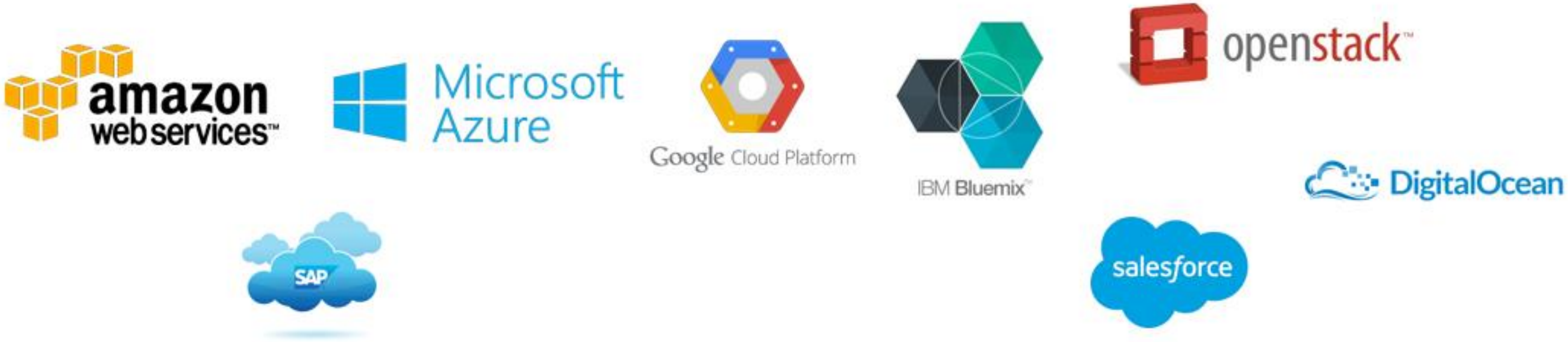
- **How can we effectively address the challenges associated with BIG Data?**
- Build bigger and more powerful machines
- Although technological advancements have made possible for the storage of vast amounts of data, there are still some technological limitations that impact on how that data can be 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

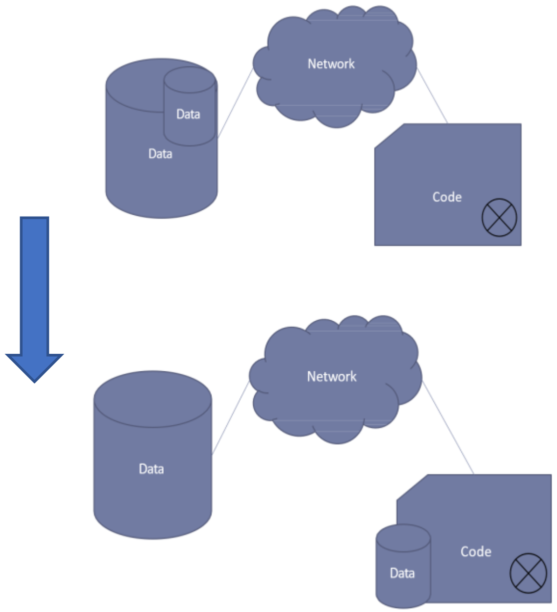
- The most popular big data processing technique using clusters of computers is **MapReduce.**
  - Is a **programming model** that can be implemented using **HADOOP** framework and others.
  - Has two steps: **Map & Reduce**
  - Divide the data into chunks and split them by the computers into the clusters.

- **Expected characteristics of Distributed Systems**
  - Resource sharing
  - Openess
  - Concurrency
  - Scalability
  - Fault tolerance
  - Transparency

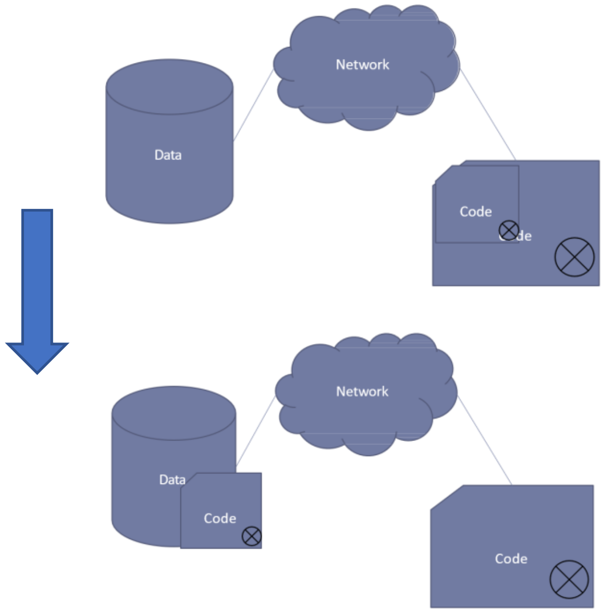
Data and Storage Paradigm

Storage / Compute Locality

- **Scenario 1**



- **Scenario 2**



- **<u>Which Scenario is preferable?</u>**
- 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 transferal 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

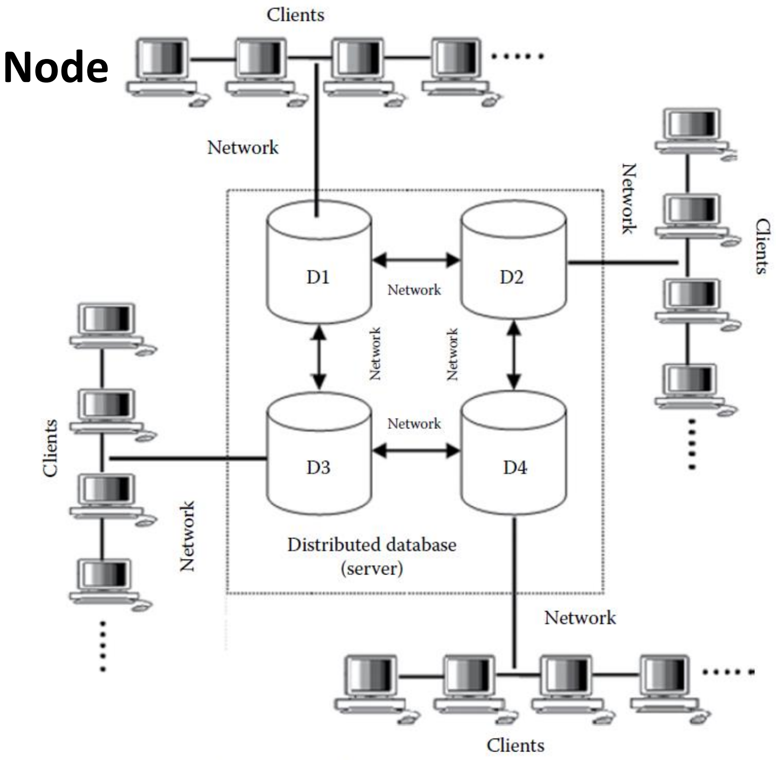
- A database server is the software that administers a database, and a client is an application that requests information and seeks services from a server.

- **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.

Node

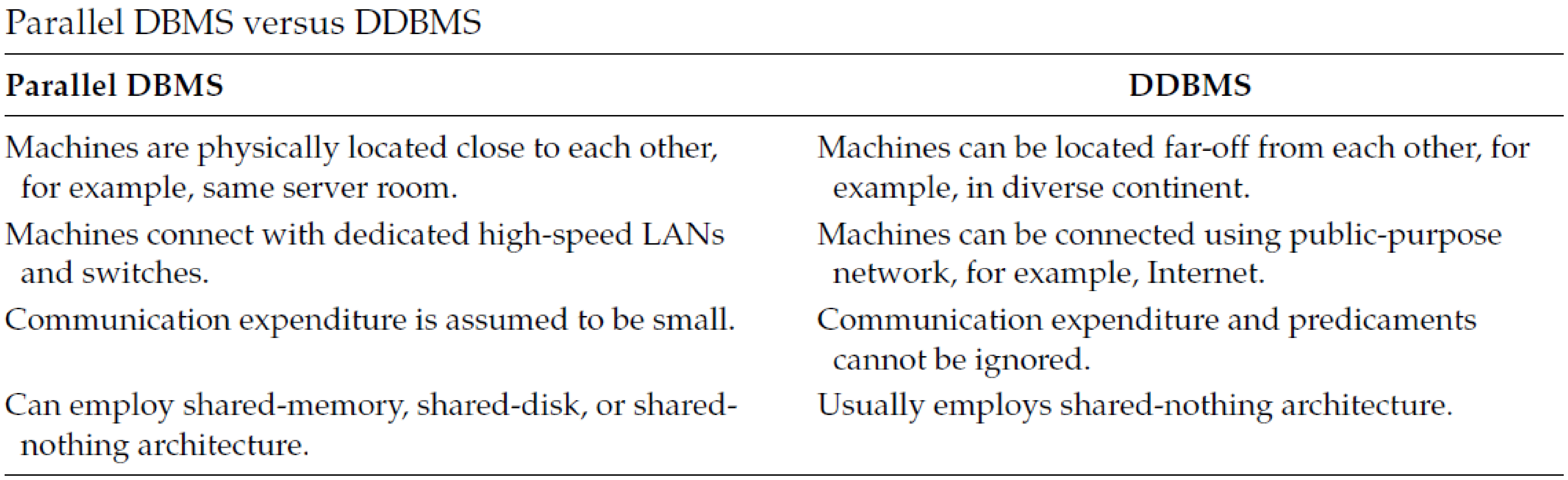


Components of distributed database.

- 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

- **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?

Resources/ References

- 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.