LECTURER: DR. TAI LE QUY

ANALYTICAL SOFTWARE AND FRAMEWORKS

TOPIC OUTLINE

Introduction to Analytical Software and Frameworks	1
Data Storage	2
Statistical Modeling Frameworks	3
Machine Learning and Artificial Intelligence Frameworks	4
Cloud Computing Platforms, On-Premise Solutions, Distributed Computing	5
Database Technology	6

UNIT 2

DATA STORAGE

STUDY GOALS

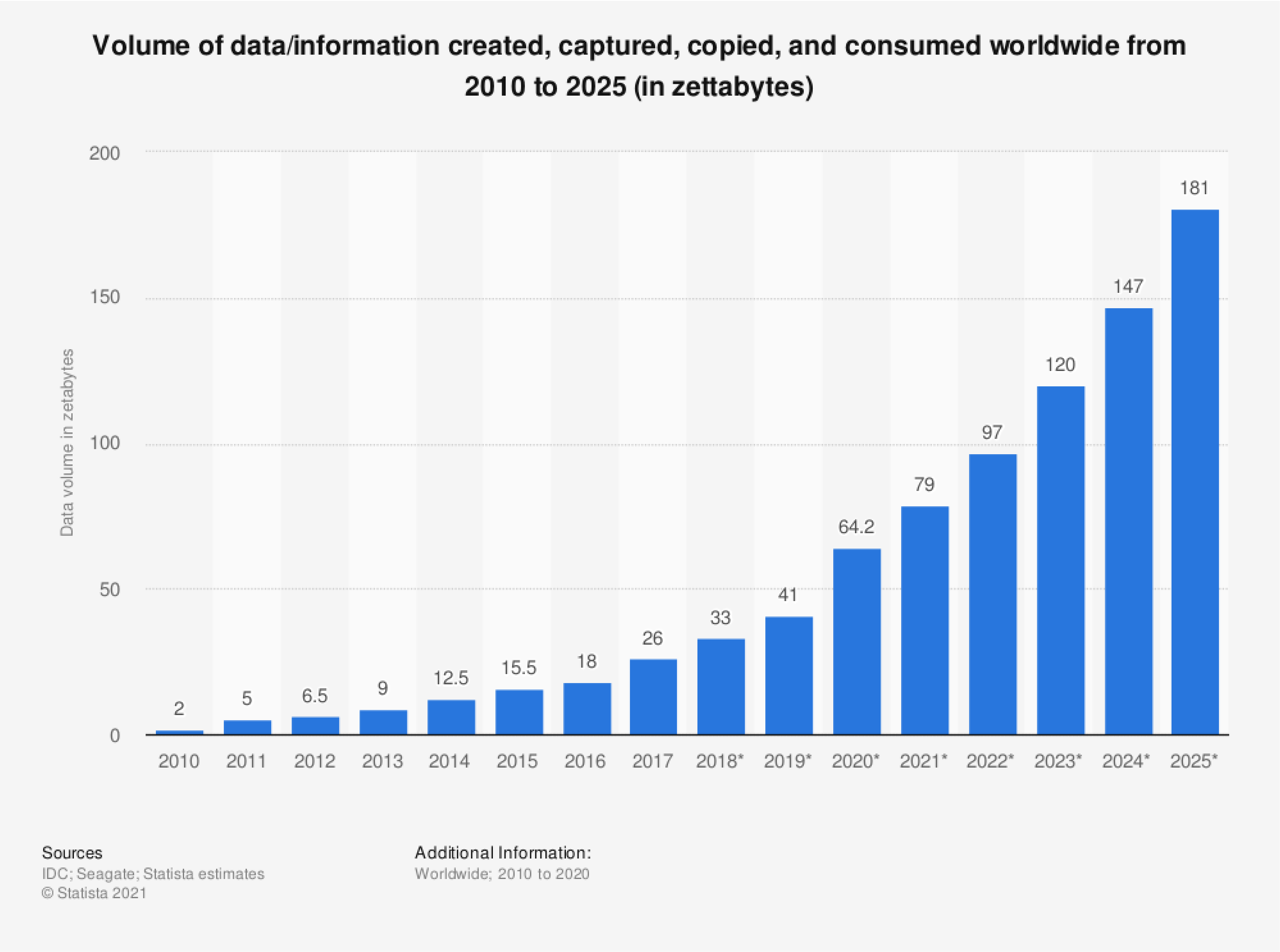
- Understand what is meant by data storage
- Know different forms of data storage
- Understand what is meant by data clustering
- Know different types and architectures of replica
- Understand what is meant by data indexing
- Know types, components, and usage of Data Warehousing

EXPLAIN SIMPLY

1. What is data storage?
2. Why are data clustering and data indexing important?
3. What are the components of a data warehouse and what are their responsibilities?

DATA STORAGE

Preserving digital data and information for ongoing or future operations on storage devices ensuring its accessibility when needed



Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025 (in zettabytes)

Sources
IDC; Seagate; Statista estimates
© Statista 2021

Additional Information:
Worldwide; 2010 to 2020

Source of the image: IDC, & Statista, 2021.

DATA STORAGE DEVICES AND FORMS OF DATA STORAGE

Storage Devices or Services

- **Direct-attached storage** (DAS) or
- **Network-based** (NAS, SAN, HCI)

- Flash and SSD storage
- Internal and external hard disks (HDDs)
- Cloud storage
- Hybrid cloud storage
- Optical media
- Enterprise storage

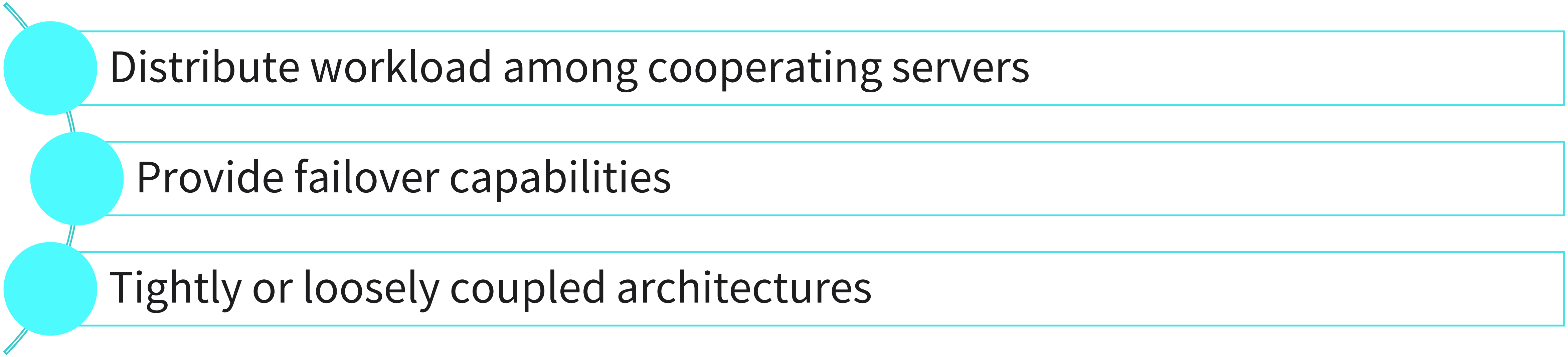
Forms of Data Storage

- **File** storage (hierarchical structure of directories and files)
- **Block** storage (raw storage for server operating systems, databases, filesystems of virtual machines, or containers)
- **Object** storage (for large amounts of unstructured data and their customizable metadata)

Source of the text: Course book DLMBDSA02, p. 30f.; IBM, 2021a; IBM, 2021b.

DATA CLUSTERING

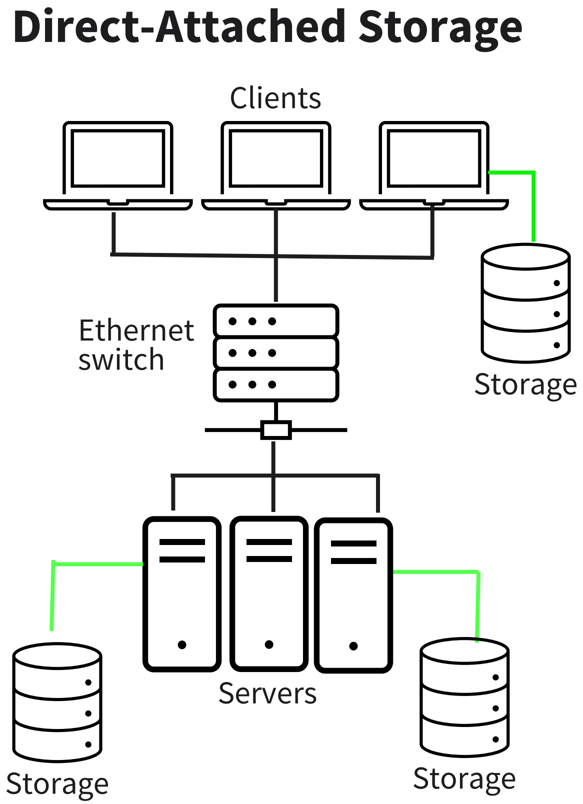
Clustering generally refers to architecture in which multiple resources, like servers, network devices, or storage arrays, work together to **increase reliability, scalability, performance, and capacity**.



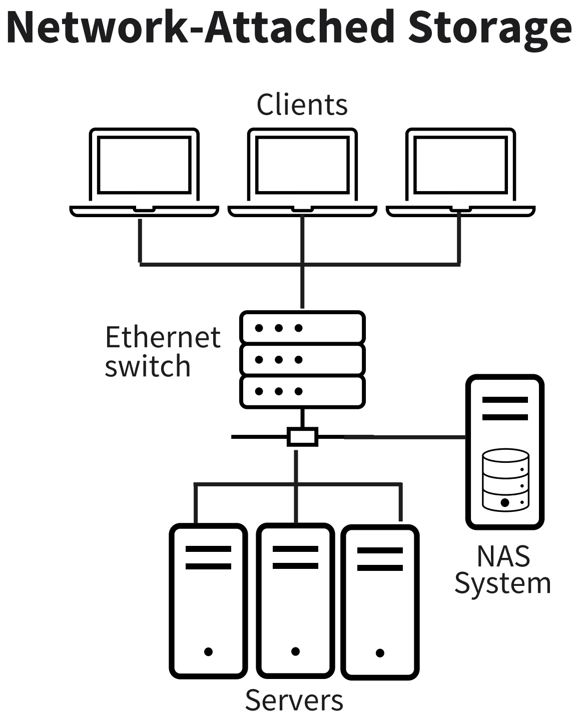
- Distribute workload among cooperating servers
- Provide failover capabilities
- Tightly or loosely coupled architectures

Source of the text: : Course book DLMBDSA02, p. 32.

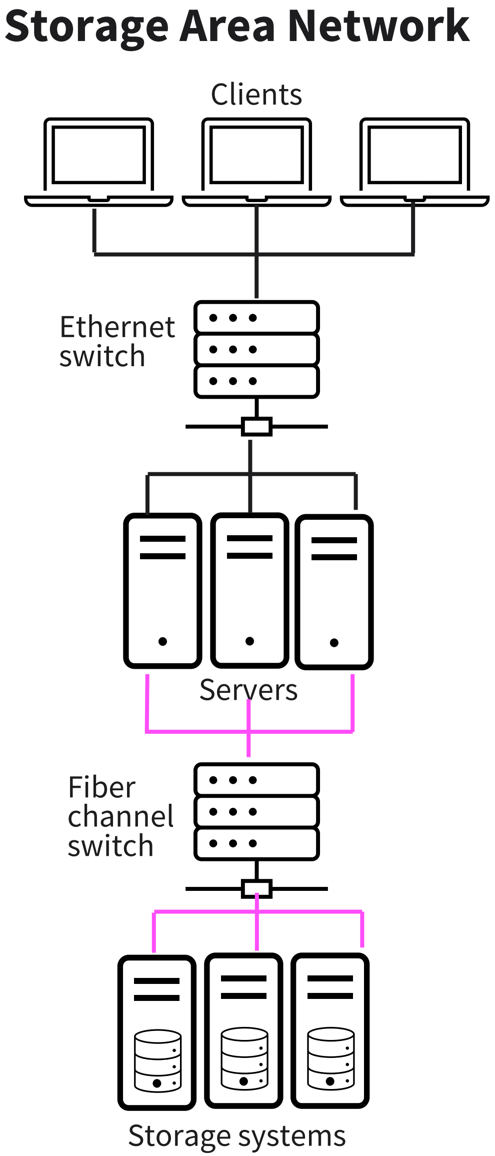
DATA CLUSTERING — SHARED STORAGE SYSTEMS



Direct-Attached Storage



Network-Attached Storage



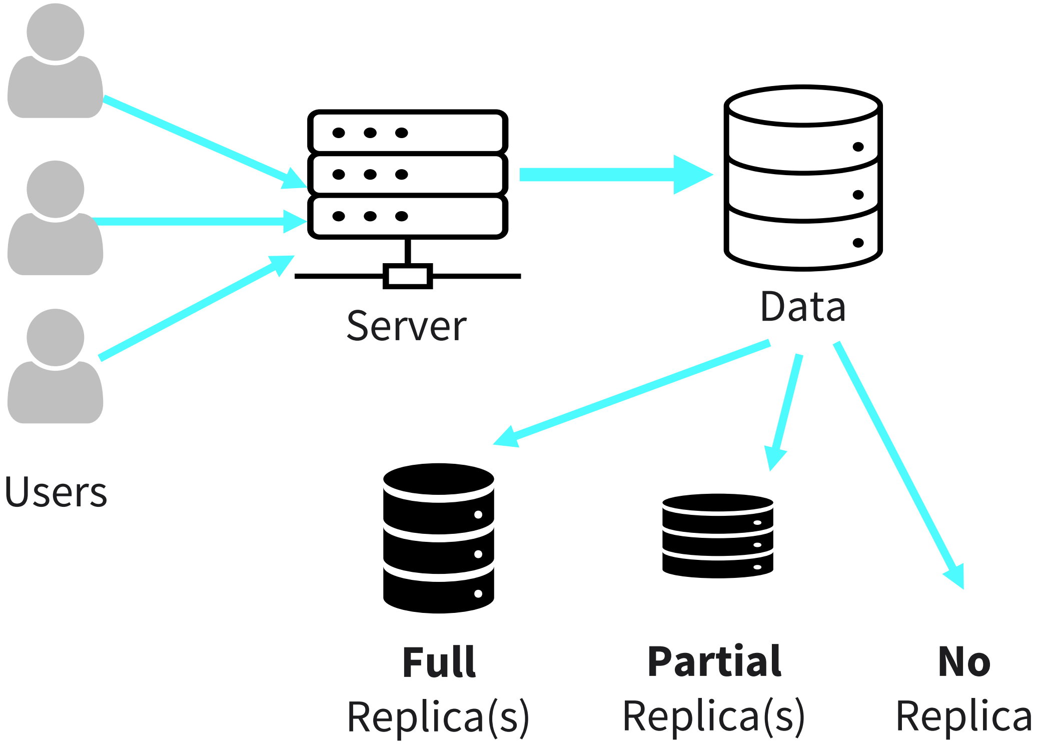
Storage Area Network

Source of the image: Own creation, based on Krenn, 2020.

DATA REPLICATION

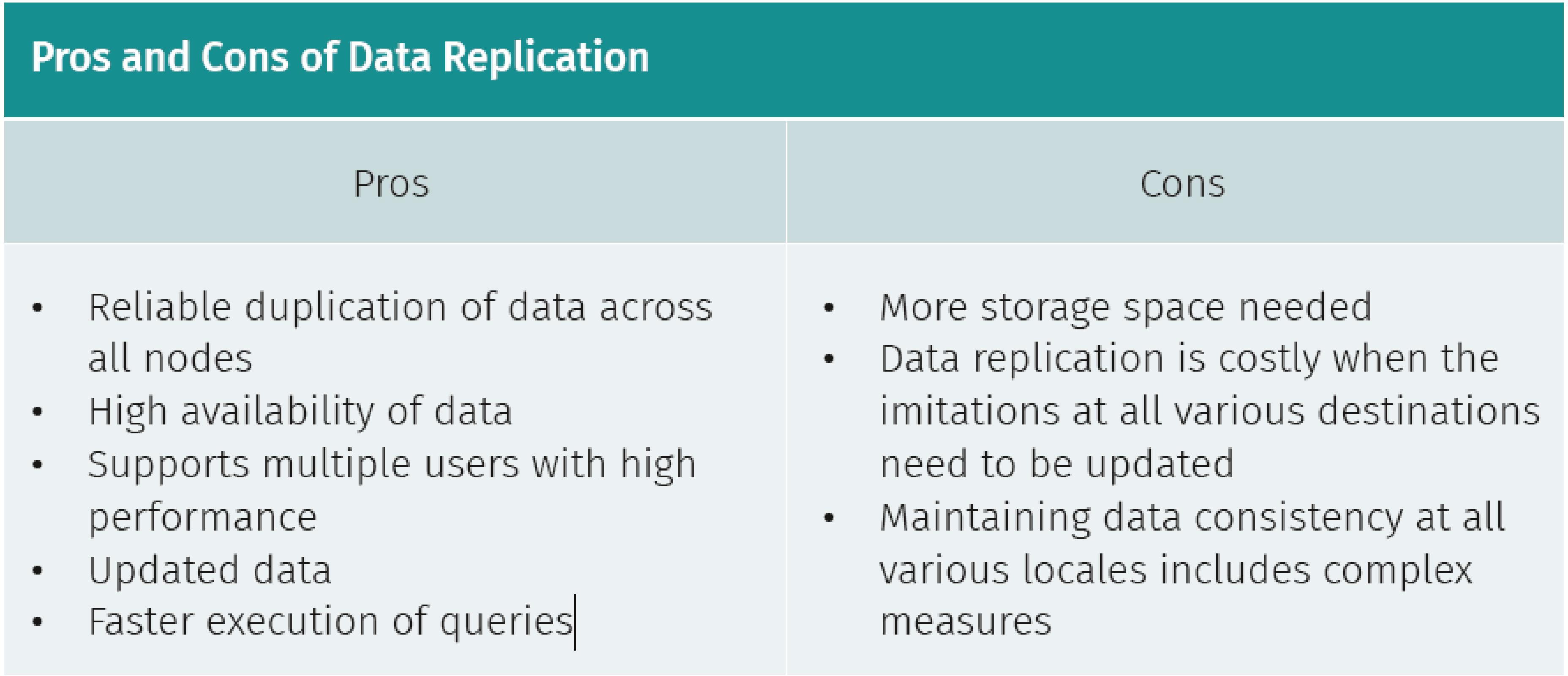
Consistently copying data from one computing resource to one or more others.

- **improve availability, query performance, and fault tolerance**
- Unstructured and structured data
- Synchronous or asynchronous
- Snapshot or ongoing



Source of the text: Course book DLMBDSA02, pp. 33—38.

DATA REPLICATION

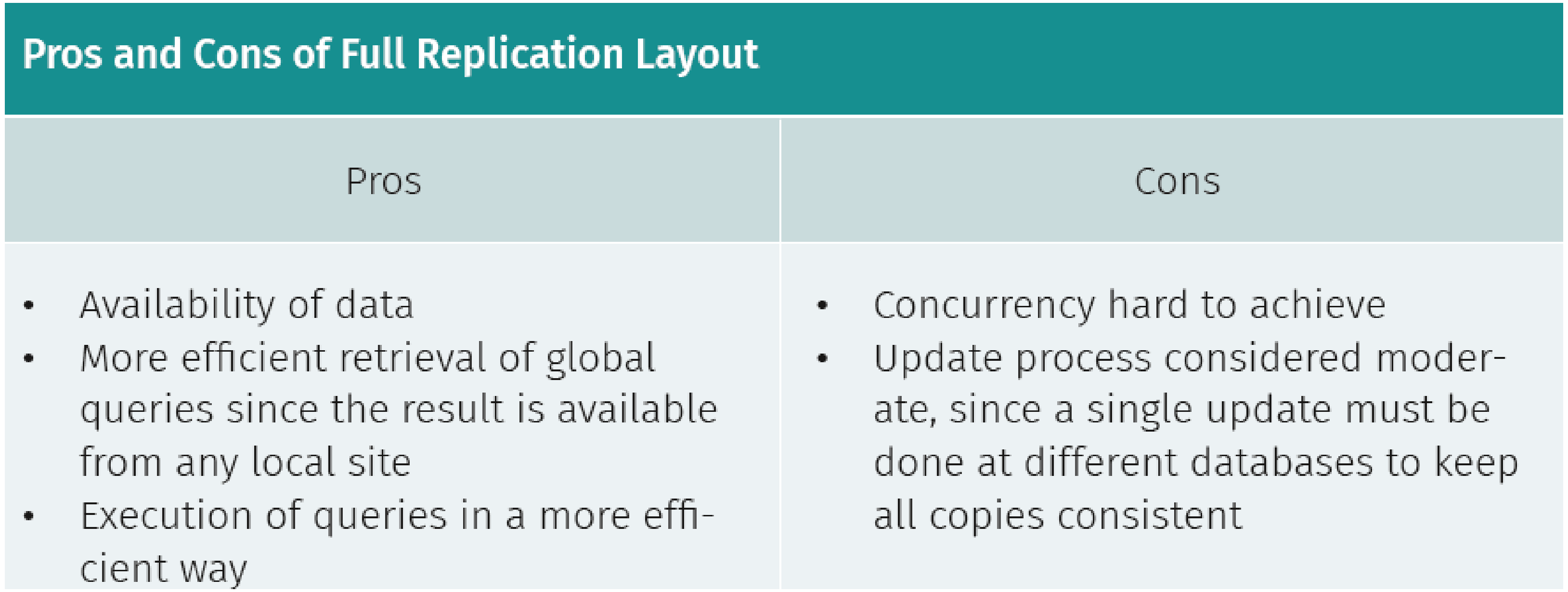


Pros and Cons of Data Replication

Pros	Cons
• Reliable duplication of data across all nodes • High availability of data • Supports multiple users with high performance • Updated data • Faster execution of queries	• More storage space needed • Data replication is costly when the imitations at all various destinations need to be updated • Maintaining data consistency at all various locales includes complex measures

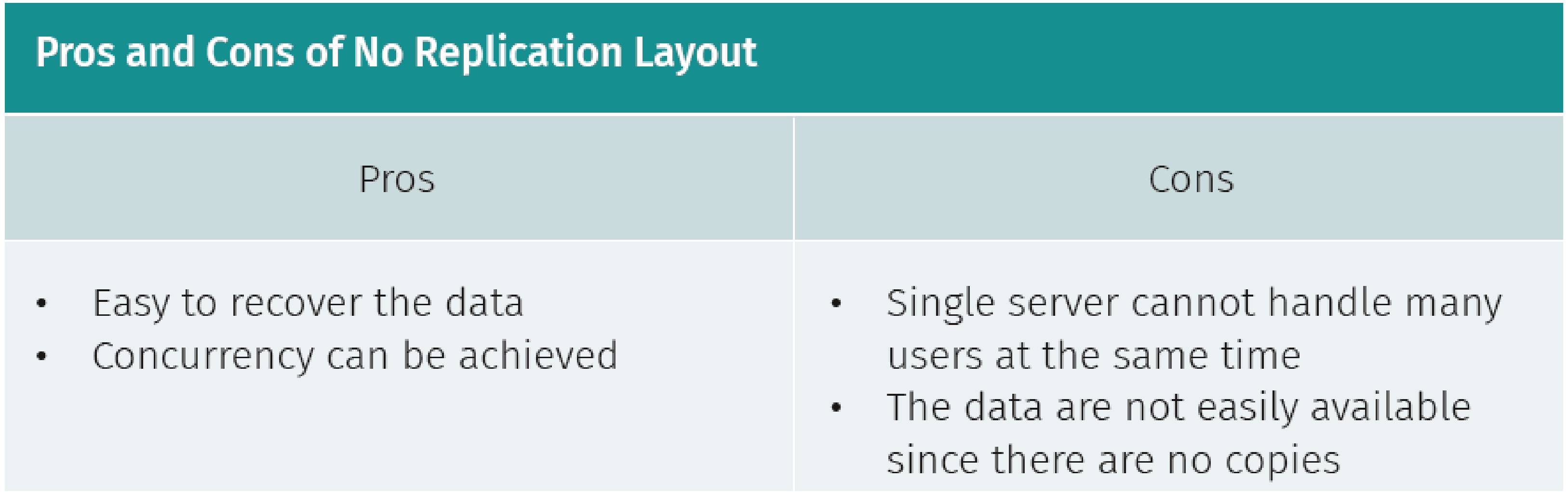
Source: Course book

DATA REPLICATION



Pros and Cons of Full Replication Layout

Pros	Cons
• Availability of data • More efficient retrieval of global queries since the result is available from any local site • Execution of queries in a more efficient way	• Concurrency hard to achieve • Update process considered moderate, since a single update must be done at different databases to keep all copies consistent



Pros and Cons of No Replication Layout

Pros	Cons
• Easy to recover the data • Concurrency can be achieved	• Single server cannot handle many users at the same time • The data are not easily available since there are no copies

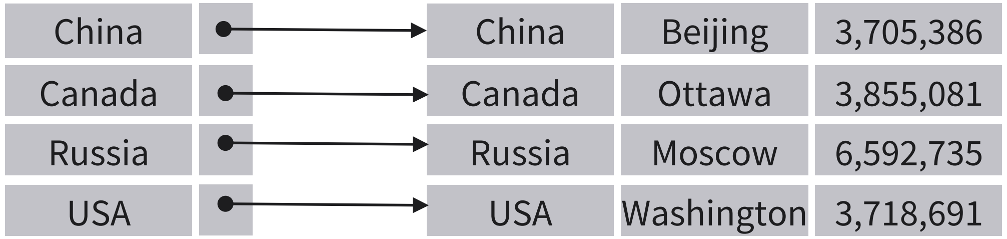
Source; Course book

DATA INDEXING

Speeding up data retrieval from storage by creating additional data structures holding indexed values or hashes thereof as well as pointers to the full indexed data/record.

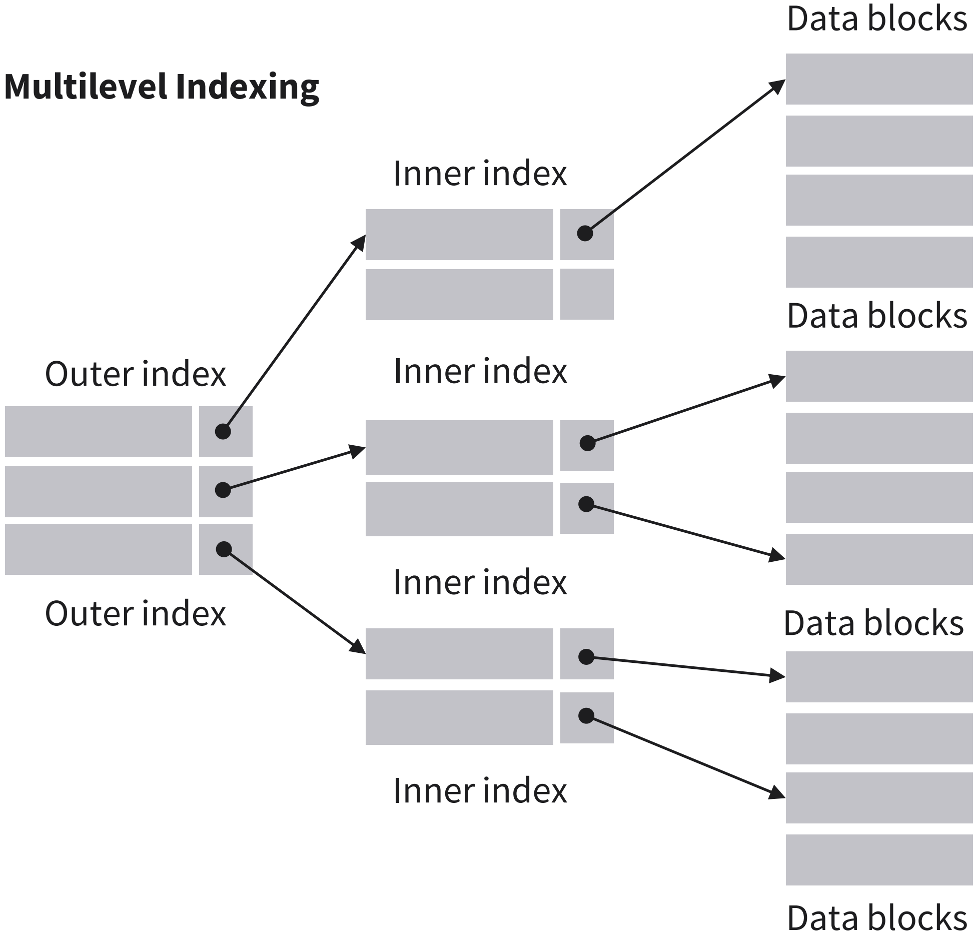
- Dense indexing
- Sparse indexing
- Multi-level indexing

Dense Indexing



China	●	China	Beijing	3,705,386
Canada	●	Canada	Ottawa	3,855,081
Russia	●	Russia	Moscow	6,592,735
USA	●	USA	Washington	3,718,691

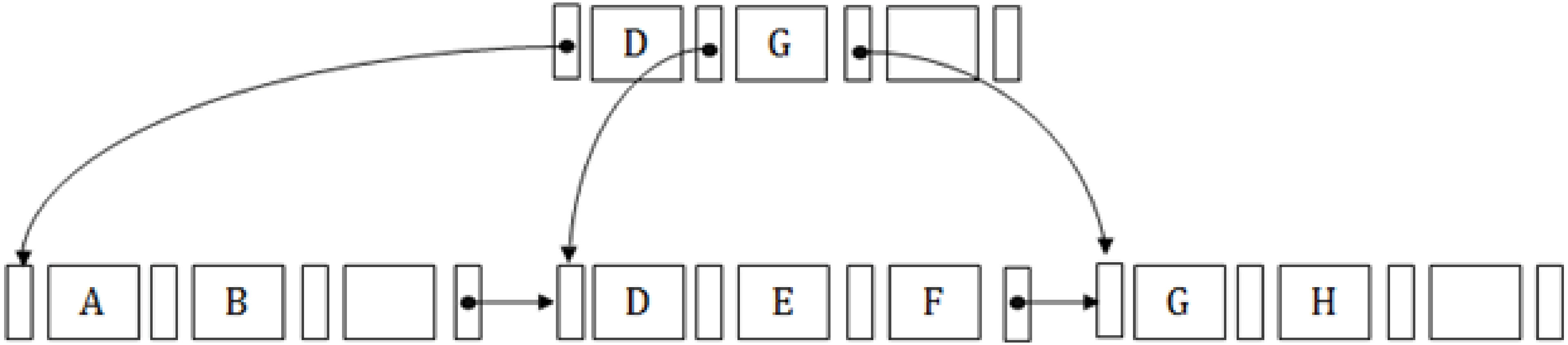
Multilevel Indexing



Source of the text: Course book DLMBDSA02, pp. 38—42.
Source of the images: Course book DLMBDSA02, pp. 39—41.

B+ TREE

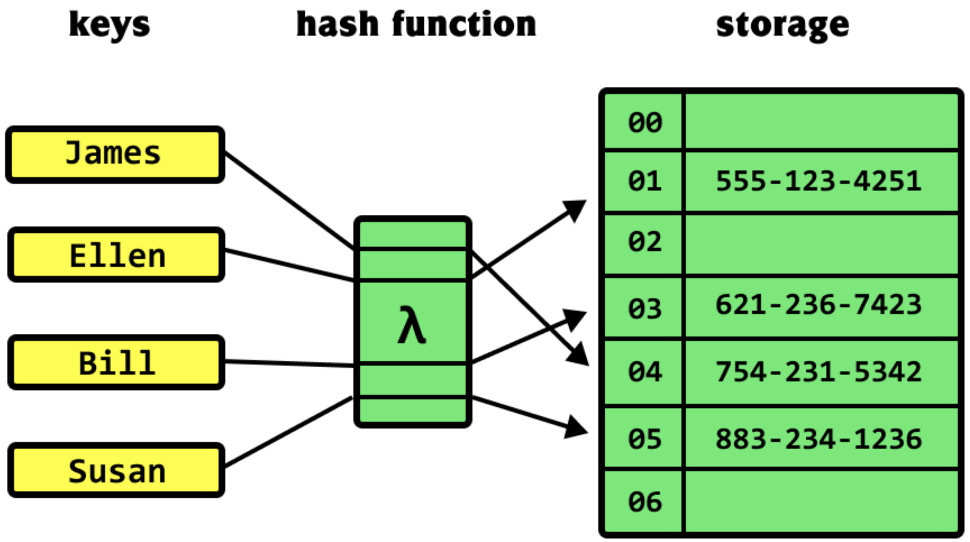
- The B+ tree is a balanced binary search tree. It follows a multi-level index format.
- In the B+ tree, leaf nodes denote actual data pointers. B+ tree ensures that all leaf nodes remain at the same height.
- In the B+ tree, the leaf nodes are linked using a link list. Therefore, a B+ tree can support random access as well as sequential access.



Source of the text and image: https://www.javatpoint.com/dbms-b-plus-tree

HASH TABLE

- Hashing is the transformation of a string of characters into a shorter fixed-length value named as the index that represents the string value.
- A hash table (hash map) is a data structure that maps indexes to values.



Source of image: https://khalilstemmler.com/blogs/data-structures-algorithms/hash-tables/

DATA WAREHOUSE (DWH)

“A data warehouse is a […] ***data architecture*** *that* ***tracks integrated, consistent, and detailed data over time, establishing relationships*** *between them* ***using metadata and schema****.”*

SUBJECT-ORIENTED
Reflecting business entities and processes of the organization

INTEGRATED AND CONSISTENT
Standardized formats and values, complete, accurate, and integer

TIME-VARIANT AND NON-VOLATILE
Captures and tracks data changes over time

METADATA, SCHEMA, AND THE DATA DICTIONARY
describing context of data and their structure

Source of the text: Teradata, 2021a.

DATA WAREHOUSE TYPES

ENTERPRISE DATA WAREHOUSE

The one single, integrated, and connected data source that brings together all the data from multiple (100s) data sources, usually > 4000 tables and > 100 GB

DATA MART

Subject-oriented, single-focus area, slice of the data warehouse logical model serving a narrow group of users, can be huge but is a narrow selection of all the data available in the DWH, usually < 20 tables and < 100 GB

OPERATIONAL DATA STORE

Complements the non-volatile nature and aggregated data in DWH with near-real time data, as a staging area receiving and aggregating operational data from transactional sources to provide query capabilities without interfering their performance

Source of the text: Teradata, 2021b, Teradata, 2021c, Huskin, 2021.

DATA WAREHOUSE COMPONENTS

LOAD MANAGER

Extract from various transactional sources, **Transform** and cleanse to follow unified formats and using a common set of enterprise definitions, **Load** and transferred into the DWH

WAREHOUSE MANAGER

Check data consistency, create indexes, de-normalization, aggregations, transformations, merging of data sources, archiving and backups

END-USER ACCESS MANAGER

Data reporting, query tools, application development tools, EIS tools, OLAP and data-mining tools

QUERY MANAGER

Management of user queries, scheduling, execution of queries

Source of the text: Course book DLMBDSA02, p. 46.

REVIEW STUDY GOALS

- Understand what is meant by data storage
- Know different types of data storage
- Understand what is meant by data clustering
- Know different types and architectures of replica
- Understand what is meant by data indexing
- Know types, components, and usage of Data Warehousing

SESSION 2

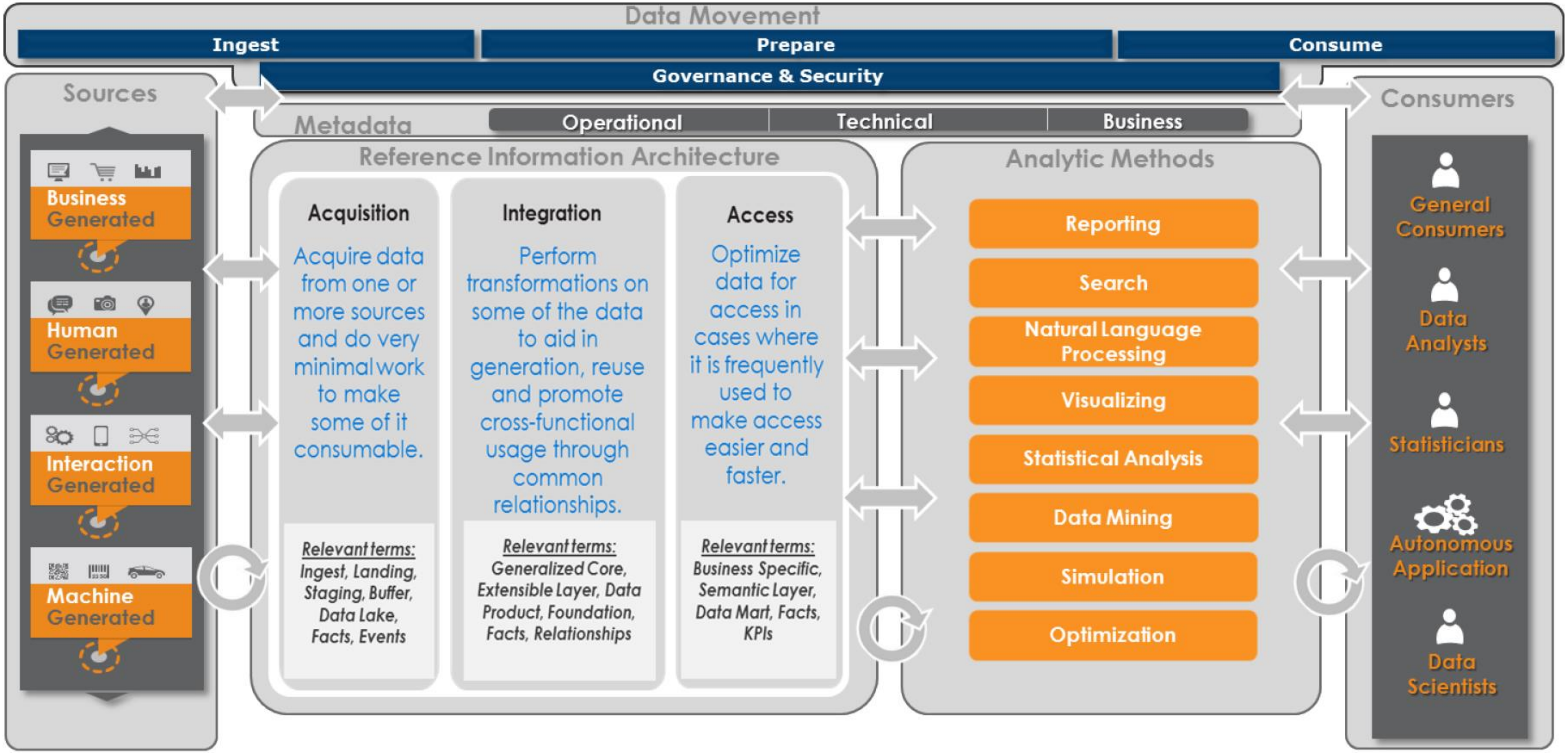
TRANSFER TASK

TRANSFER TASK

Gather in groups of 2—3 to work on the following question.

What does the following illustration represent?
Which elements of our session today do you recognize?

TRANSFER TASK



Data Movement

Ingest

Prepare

Consume

Governance & Security

Sources

Business Generated

Human Generated

Interaction Generated

Machine Generated

Metadata

Operational

Technical

Business

Reference Information Architecture

Acquisition

Acquire data from one or more sources and do very minimal work to make some of it consumable.

Relevant terms: Ingest, Landing, Staging, Buffer, Data Lake, Facts, Events

Integration

Perform transformations on some of the data to aid in generation, reuse and promote cross-functional usage through common relationships.

Relevant terms: Generalized Core, Extensible Layer, Data Product, Foundation, Facts, Relationships

Access

Optimize data for access in cases where it is frequently used to make access easier and faster.

Relevant terms: Business Specific, Semantic Layer, Data Mart, Facts, KPIs

Analytic Methods

Reporting

Search

Natural Language Processing

Visualizing

Statistical Analysis

Data Mining

Simulation

Optimization

Consumers

General Consumers

Data Analysts

Statisticians

Autonomous Application

Data Scientists

TRANSFER TASK

Gather in groups of 2—3 to work on the following questions.

Imagine you are working in an automotive organization! Your company would like to collect position data of cars describing where a vehicle was situated at a specific moment in time. What do you have to consider if you would have to design a data architecture for this purpose? Make assumptions and justify them! What if the company also would like to analyze image data to feed a Deep Learning algorithm?

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

LEARNING CONTROL QUESTIONS

1. Which of the following is not a widely used form of data storage?

    a) Block storage

    b) Object storage

    c) File storage

    d) Group storage

2. Which of the following is not a storage replica scheme?

a) Partial replica
b) Mirror replica
c) Full replica
d) No replica

3. Which of the following is not a component of a data warehouse?

   a) End-user access tools
   b) Import manager
   c) Query manager
   d) Load manager

LIST OF SOURCES

Huskin, E. (2021). *Operational Data Store vs. Data Warehouse.* https://data-science-blog.com/blog/2021/01/25/operational-data-store-vs-data-warehouse/

IDC, & Statista. (June 7, 2021). *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025 (in zettabytes) [Graph].* https://www-statista-com/statistics/871513/worldwide-data-created/

IBM (2021a). *What is data storage?* https://www.ibm.com/topics/data-storage

IBM (2021b). *Block Storage.* https://www.ibm.com/cloud/learn/block-storage

Teradata (2021a). *What is Data Warehousing?* https://www.teradata.de/Glossary/What-is-Data-Warehousing

Teradata (2021b). *Data Warehouses vs. Data Marts.* https://www.teradata.de/Trends/Data-Warehouse/Data-Warehouses-vs-Data-Marts-Infographic

Teradata (2021c). *What is a Data Mart?* https://www.teradata.de/Glossary/What-is-a-Data-Mart