DATA SCIENCE FOUNDATIONS

MANAGEMENT, SYSTEMS, AND STORAGE

– Ashika Ashok (D081)
Rudra More (D097)

DATA MANAGEMENT AND ORGANIZATION

Introduction to Data Management

- Involves collecting, storing, organizing, and maintaining data to ensure its accessibility, accuracy, and usability
- Backbone of data-driven decision-making in data science and business analytics

Importance:

- Ensures high-quality data for analysis and insights
- Reduces redundancy, enhances data security, and improves operational efficiency

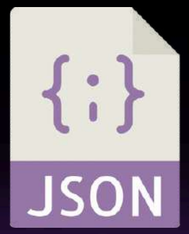
Types Of Data

Structured Data

- Organized in rows/columns
- Example : databases, excel sheets, etc.

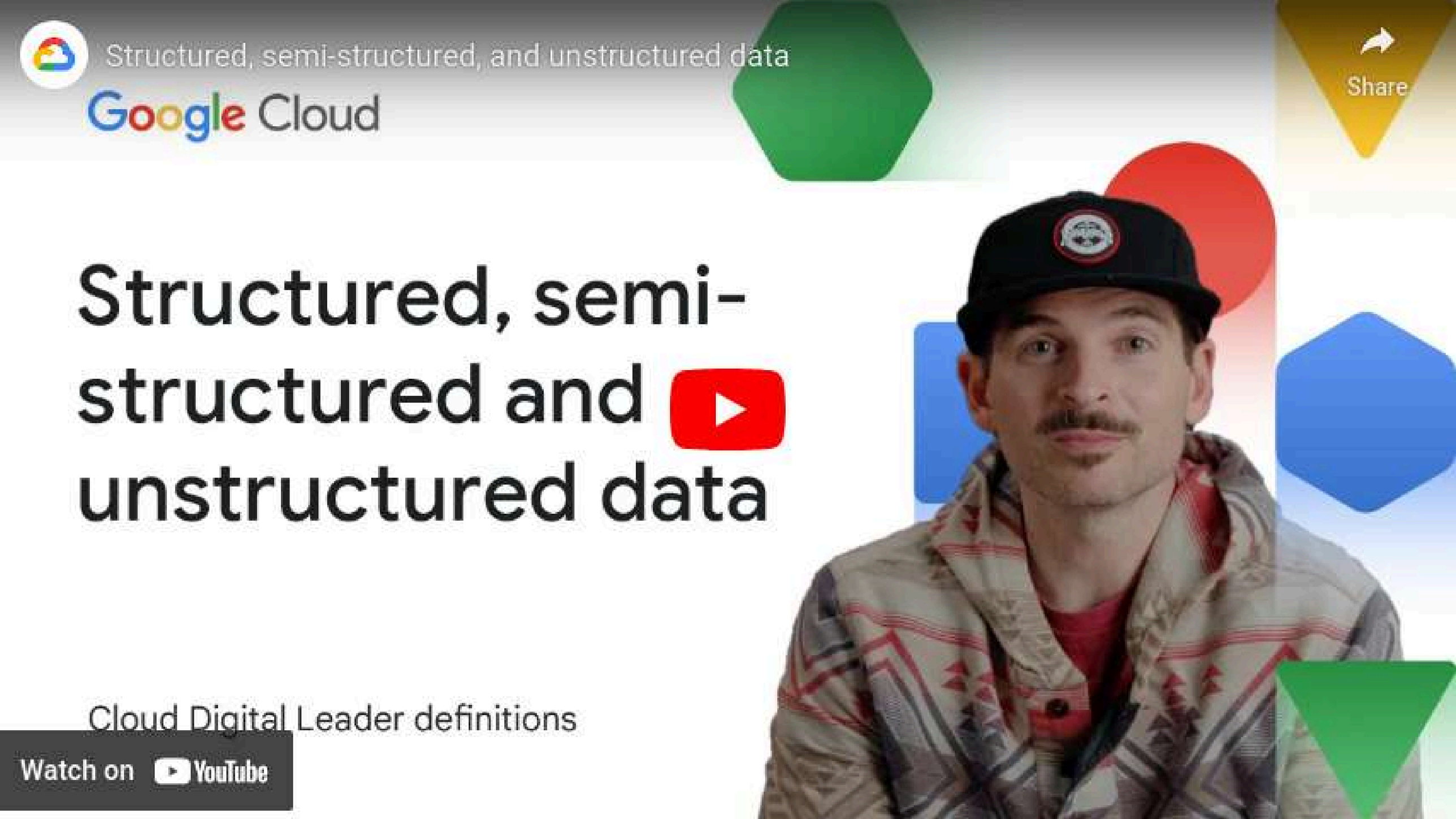
Semi – Structured Data

- Lacks a fixed schema
- Example : JSON, XML, etc.



Unstructured Data

- No predefined format
- Example : videos, images, etc.


Structured, semi-structured, and unstructured data
Google Cloud
Share
Structured, semi-structured and unstructured data
Cloud Digital Leader definitions
Watch on YouTube

Data Management Lifecycle

Generation

Collection

Storage

Processing

Management

Analysis

Visualization & Interpretation

Destruction

Key Practices in Data Organization

METADATA MANAGEMENT

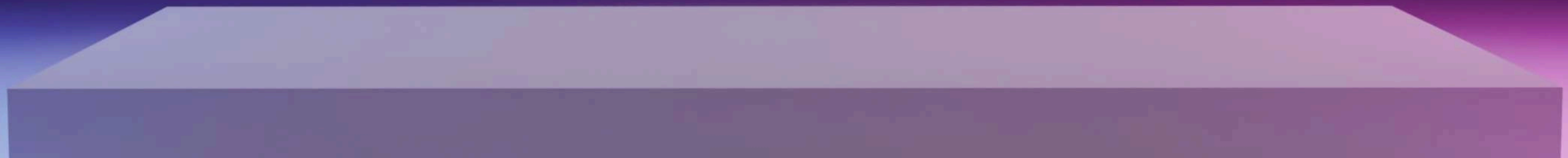
- Provides context to data by describing its structure and origin.
- Example: Dataset attributes like column names and types.

DATA VALIDATION AND INTEGRITY

- Ensures data is accurate and meets quality standards.
- Example: Setting constraints in databases (e.g., unique keys).

VERSION CONTROL

- Tracks changes to datasets over time.
- Example: Using Git for collaborative data projects.




Netflix ux case study | Netflix data analytics case study - Disruptive innovation
Share
NETFLIX
Netflix Data Analytics
A Case Study
5 MINUTES LEARNING
Watch on YouTube

LARGE SCALE DATA SYSTEMS

Introduction to Large-Scale Data Systems

LARGE-SCALE DATA

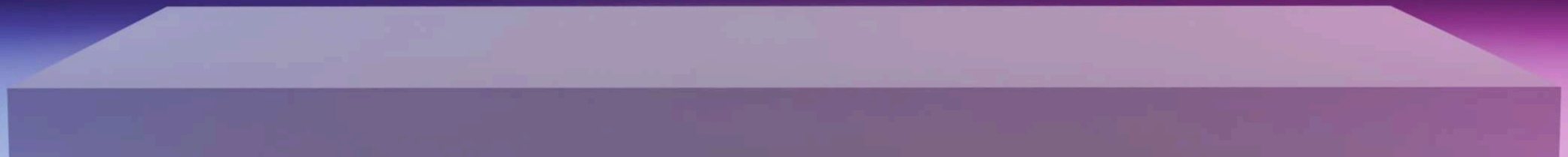
Refers to datasets that are too big, fast, or complex to be processed and managed using traditional methods

3V'S

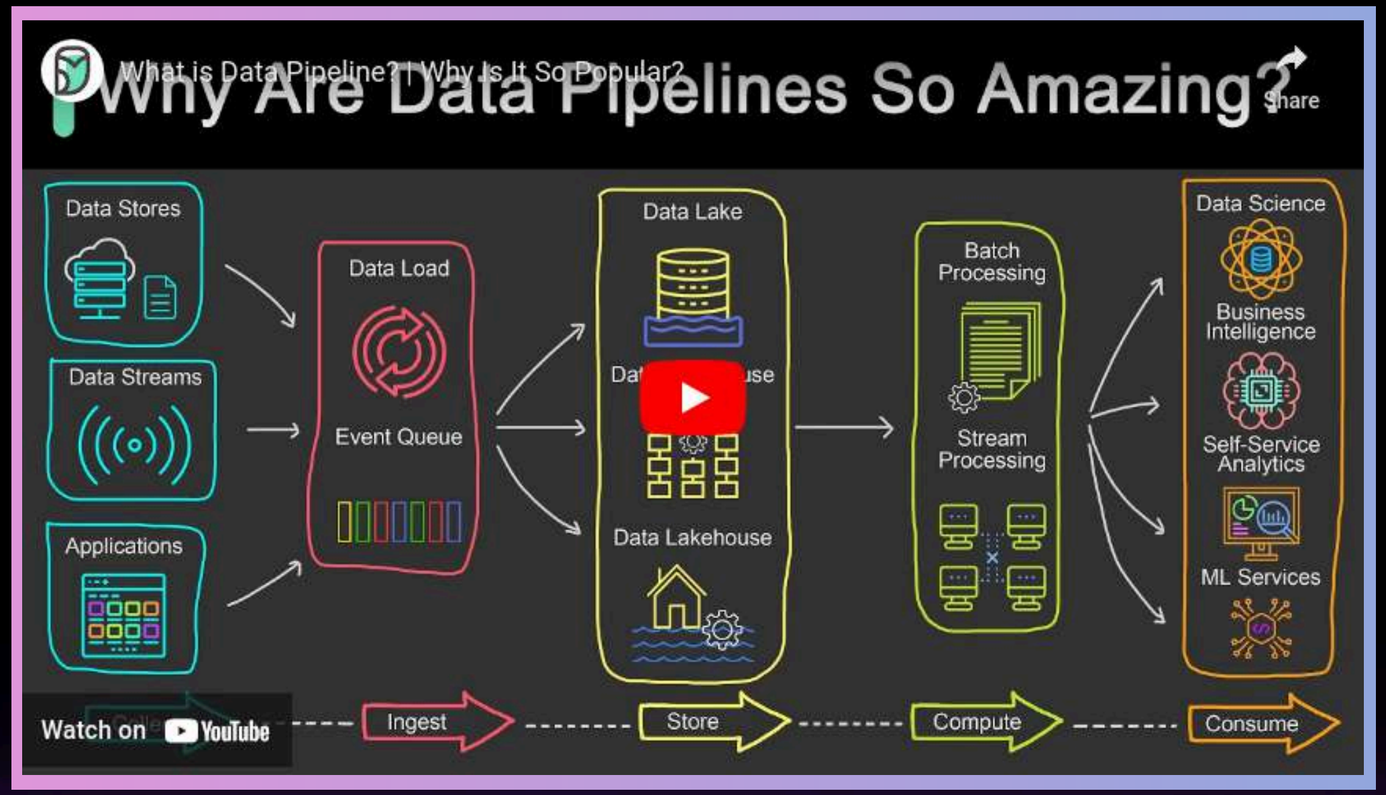
Volume: Size of the data
Velocity: Speed at which data is generated & processed
Variety: Different formats and types (structured, semi-structured, unstructured)

EXAMPLES

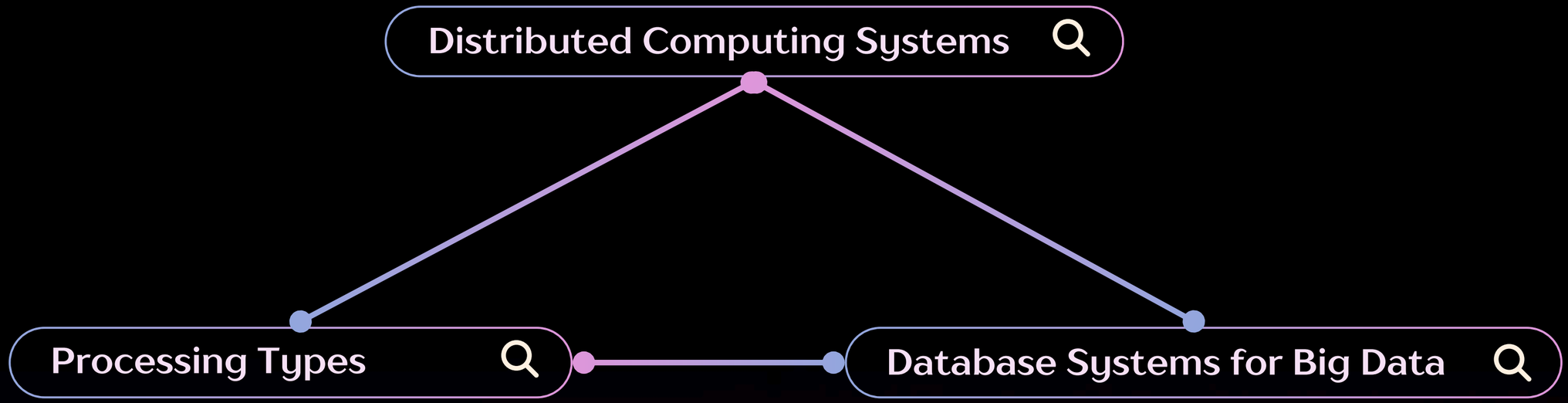
Social media platforms, IoT devices , E-commerce websites



Architecture



Core Technologies



Distributed Computing Systems

Definition

- Use clusters of machines to process data in parallel.
- Divide large tasks into smaller, manageable tasks executed across multiple nodes.

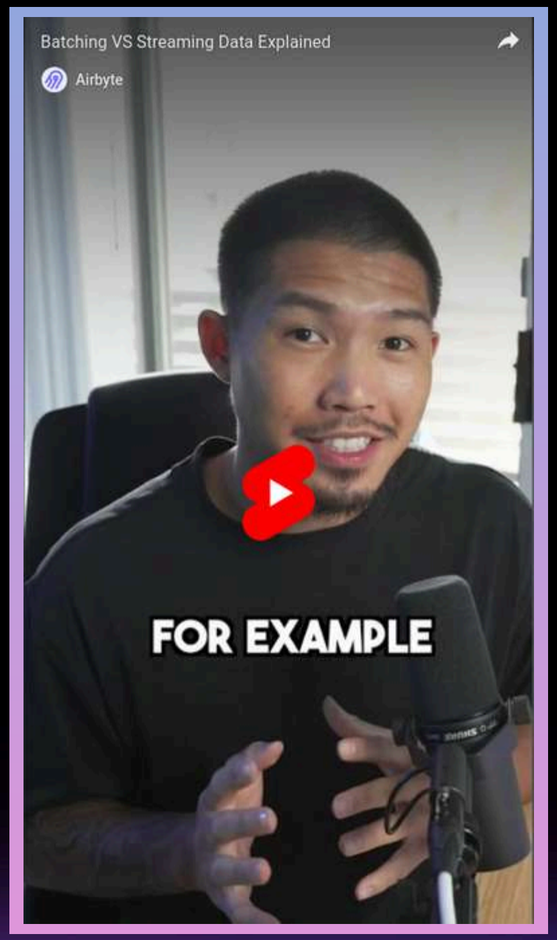
Importance

- Use clusters of machines to process data in parallel.
- Divide large tasks into smaller, manageable tasks executed across multiple nodes.

Examples

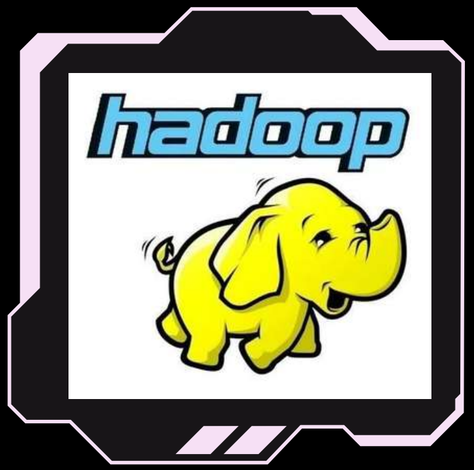
- Apache Hadoop: Framework for batch processing & distributed storage.
- Apache Spark: Faster in-memory processing for large-scale data analysis.

Processing Types



Database Systems For Big Data

- HDFS (Hadoop Distributed File System): Designed for distributed storage.
- Cassandra: A NoSQL database for handling massive amounts of structured and semi-structured data.



Use Cases

Recommendation Systems

Platforms like Netflix, Spotify, and Amazon

Fraud Detection

Banks and fintech companies

Predictive Analytics

Retailers forecast demand, inventory, and market trends

Challenges Faced

Scalability

Difficulty in expanding systems to handle rapidly growing data volumes

Performance and Latency

Challenges in ensuring fast data access and real-time processing

Fault Tolerance

Struggles with maintaining operations during hardware or software failures

Cost

Managing the rising expenses of storage and computation efficiently.

PARADIGMS FOR DISTRIBUTED DATA STORAGE

Distributed Data Storage

Definition

Refers to storing data across multiple servers, often located in different geographical locations, to ensure scalability, reliability, and availability.

Centralized vs. Distributed Storage

- Centralized: Single point of storage; prone to bottlenecks and failures.
- Distributed: Data spread across multiple nodes; increases redundancy and performance.

Key Paradigms and Techniques

Replication

Duplicates data across multiple servers for fault tolerance and data redundancy & ensures data availability even if some nodes fail.

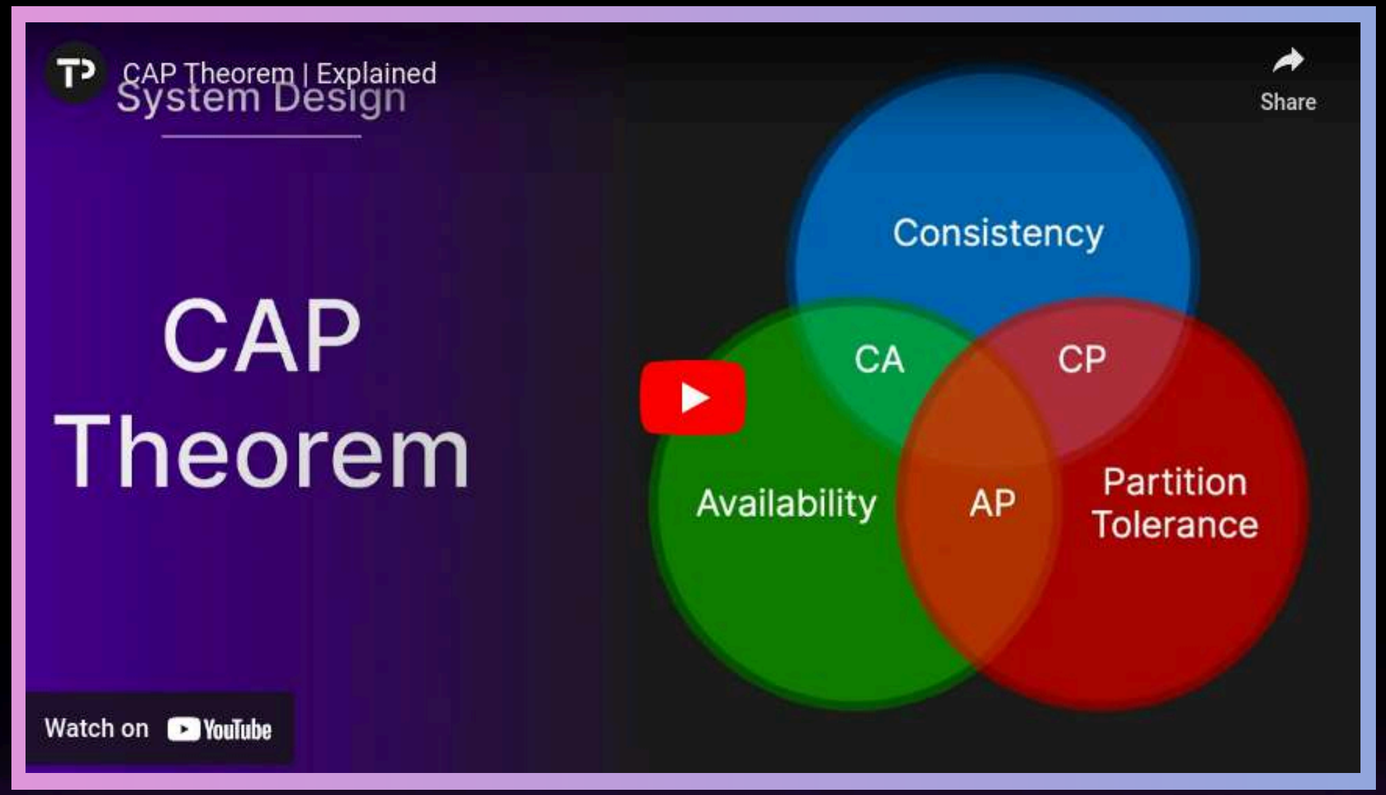
Sharding

Divides data horizontally across servers (e.g., by user IDs or regions) & improves performance by distributing workloads.

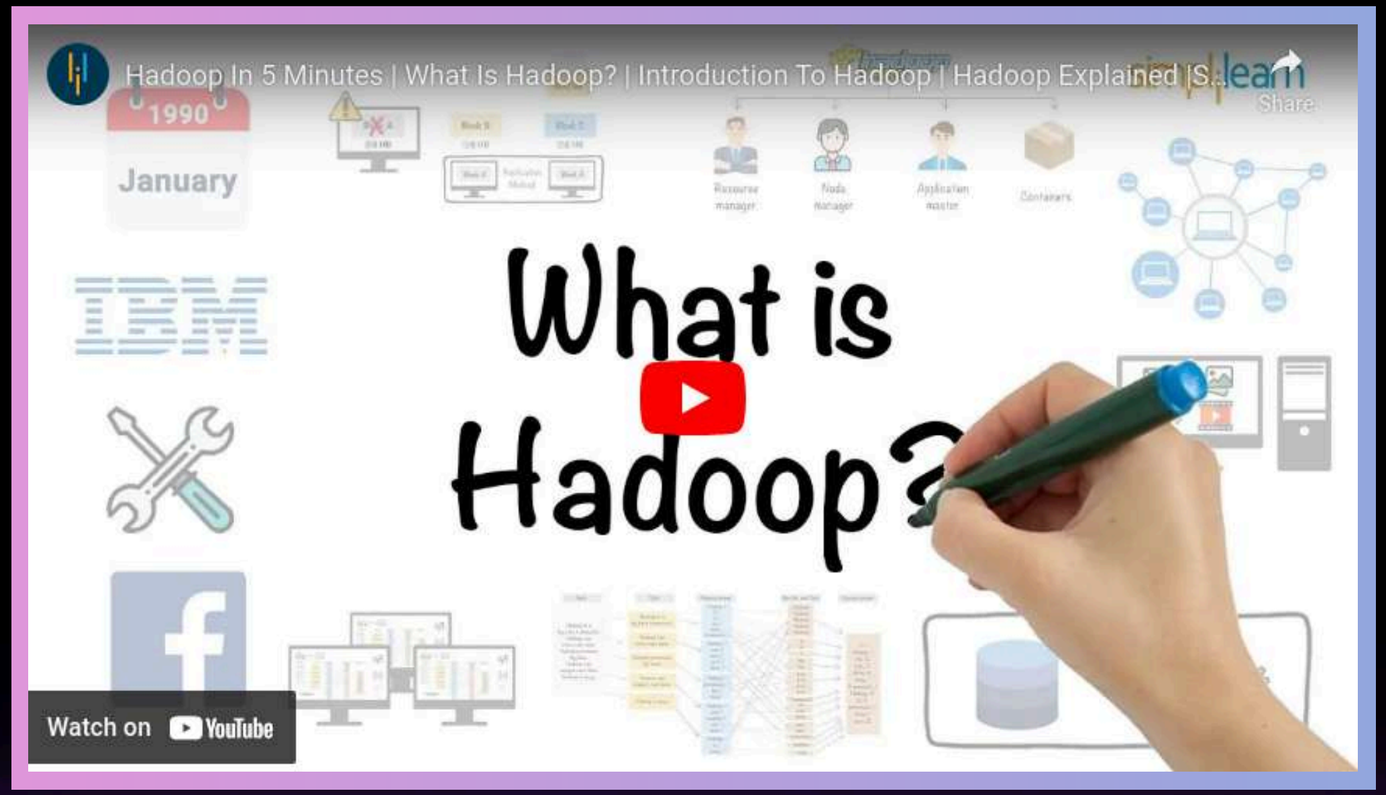
Consistency Models

- Strong Consistency: All users see the same data immediately.
- Eventual Consistency: Data updates propagate eventually, allowing better performance.
- Causal Consistency: Updates maintain causal order.

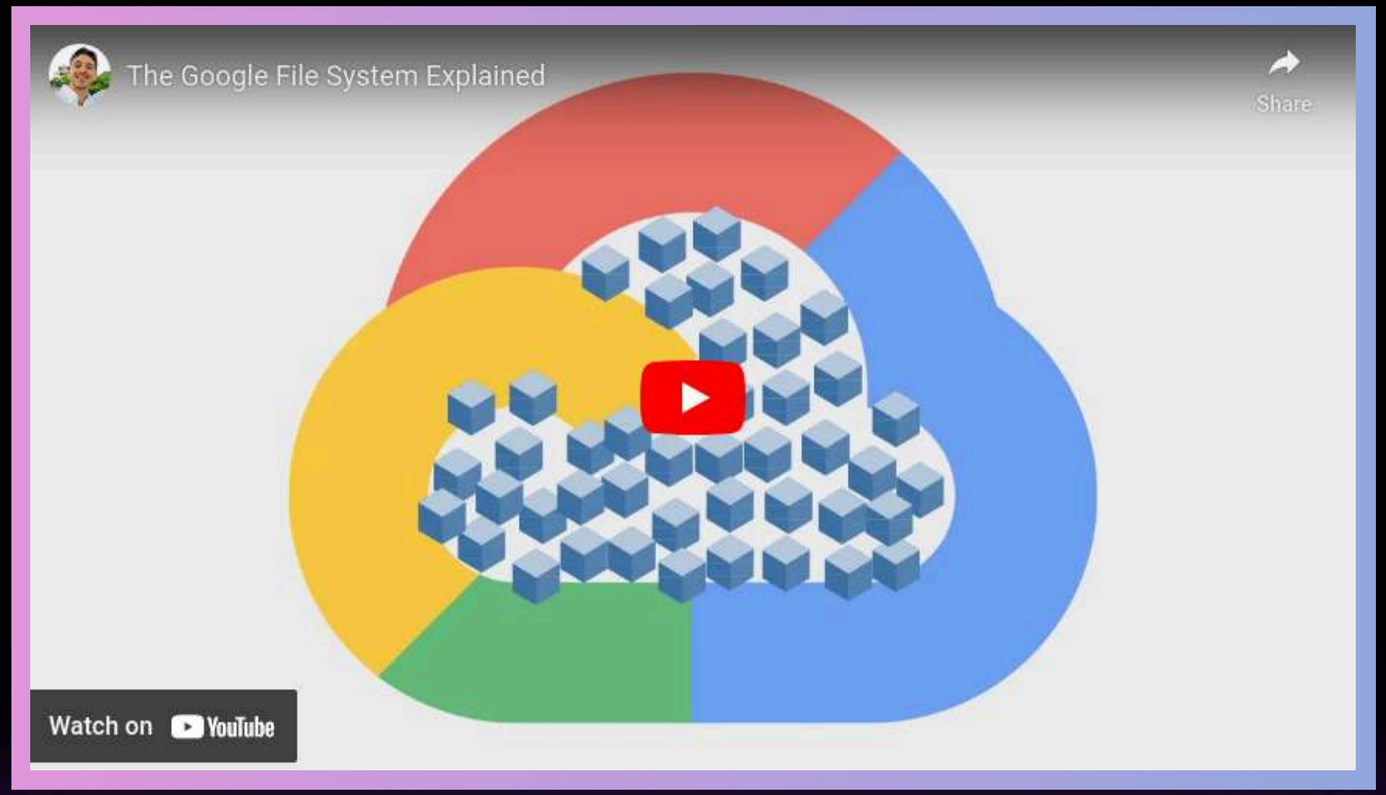
CAP Theorem


CAP Theorem | Explained
System Design
Share
CAP
Theorem
Consistency
CA
CP
Availability
AP
Partition
Tolerance
Watch on YouTube

Distributed File System - HDFS



Distributed File System - GFS



Databases for Distributed Data Storage

MongoDB

Document-oriented; supports sharding and replication

Cassandra

Column-family-based; optimized for scalability and fault tolerance

Databases for Distributed Data Storage

Amazon S3

Highly durable and scalable object storage

Google Bigtable

Ideal for high-throughput applications

Challenges Faced

Network Latency

Increased response times due to data synchronization across nodes

Data Security

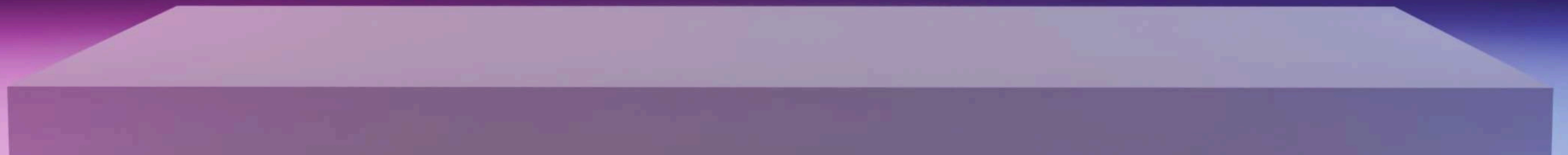
Higher risk of breaches when data is distributed across multiple locations

Synchronization Issues

Difficulties in maintaining consistency between replicas, especially in eventual consistency models

Complexity

More challenging to design, implement, and manage than centralized systems



MULTIPLE CHOICE QUESTIONS

WHAT IS THE PRIMARY GOAL OF DATA MANAGEMENT?

a) Data encryption
b) Efficient storage, processing, and analysis
c) Visualization of data
d) Reducing latency

WHAT IS THE PURPOSE OF METADATA MANAGEMENT?

a) To clean the data
b) To describe and provide context for data
c) To ensure data is encrypted
d) To visualize the data

WHAT QUALIFIES AS "LARGE-SCALE DATA"?

a) High volume, velocity, and variety of data
b) Data stored in relational databases
c) Only structured data
d) Data with minimal redundancy

WHICH IS AN EXAMPLE OF A USE CASE FOR LARGE-SCALE DATA SYSTEMS?

a) Running small-scale surveys
b) Building recommendation systems like Netflix
c) Analyzing individual spreadsheets
d) Visualizing static graphs

WHAT IS THE PRIMARY GOAL OF DISTRIBUTED DATA STORAGE?

a) Centralized data storage
b) Scalability, reliability, and fault tolerance
c) Visualizing data insights
d) Reducing encryption overhead

WHICH DISTRIBUTED FILE SYSTEM IS COMMONLY USED FOR BIG DATA?

a) MongoDB
b) SQL Server
c) Hadoop Distributed File System (HDFS)
d) PostgreSQL

THANK YOU