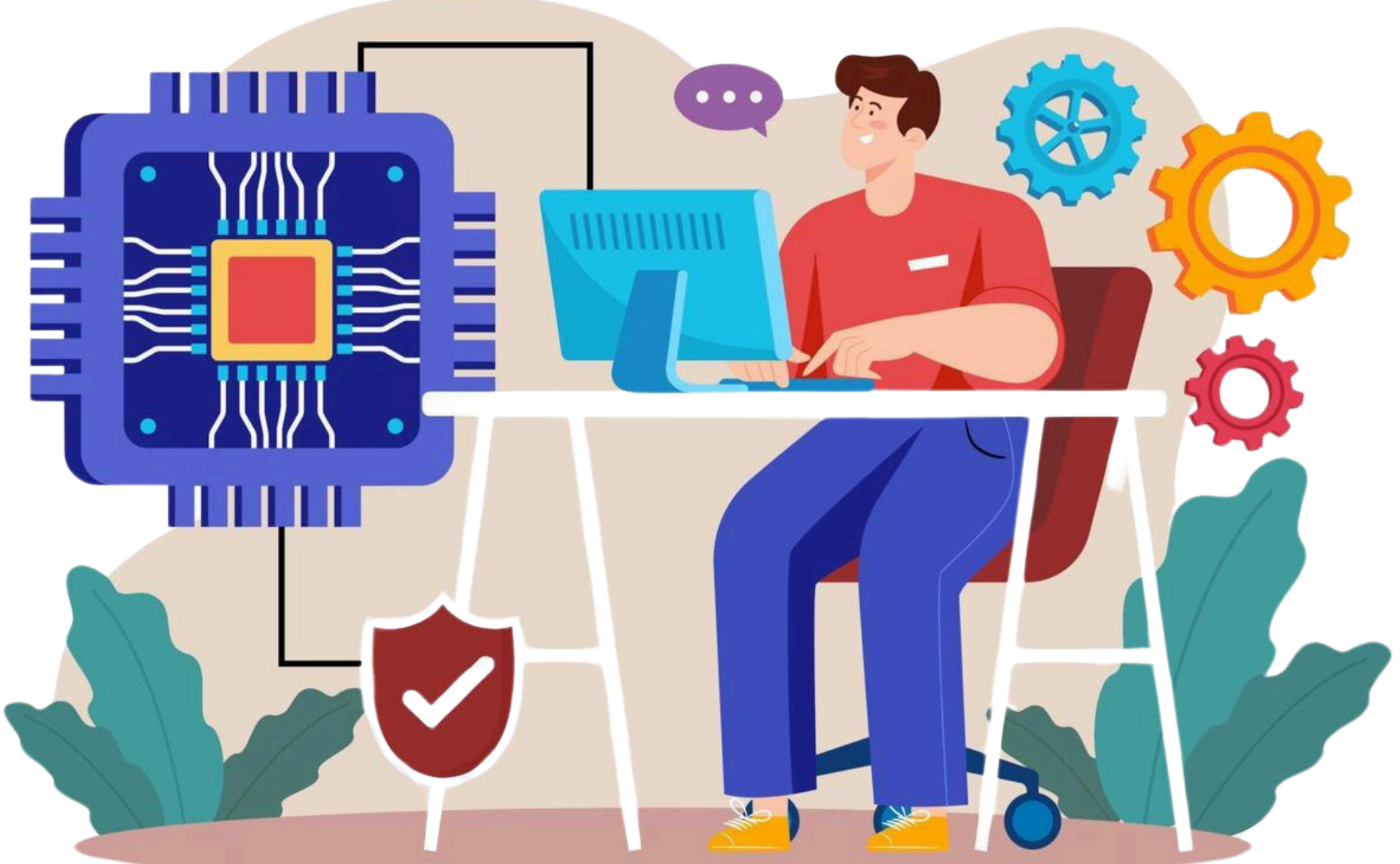
DATA ENGINEERING

TERMS YOU NEED TO KNOW

PART - 2



THE RAVIT SHOW

Don't Forget to Save For Later

21. Dimensional Modeling

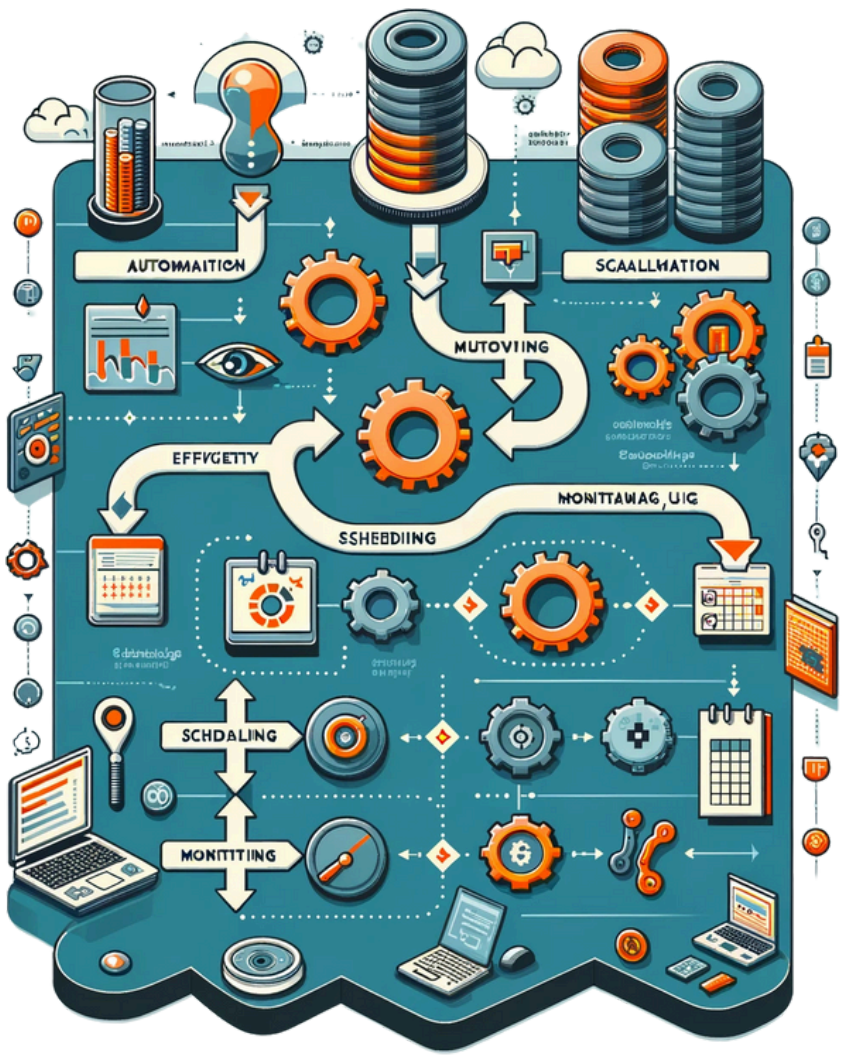
Dimensional modeling is a data modeling technique used in data warehousing to organize data into facts and dimensions. It simplifies querying by structuring data into easily understandable categories, such as sales (fact) and time (dimension), which are commonly used for reporting and analysis.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

22. Data Pipeline Orchestration



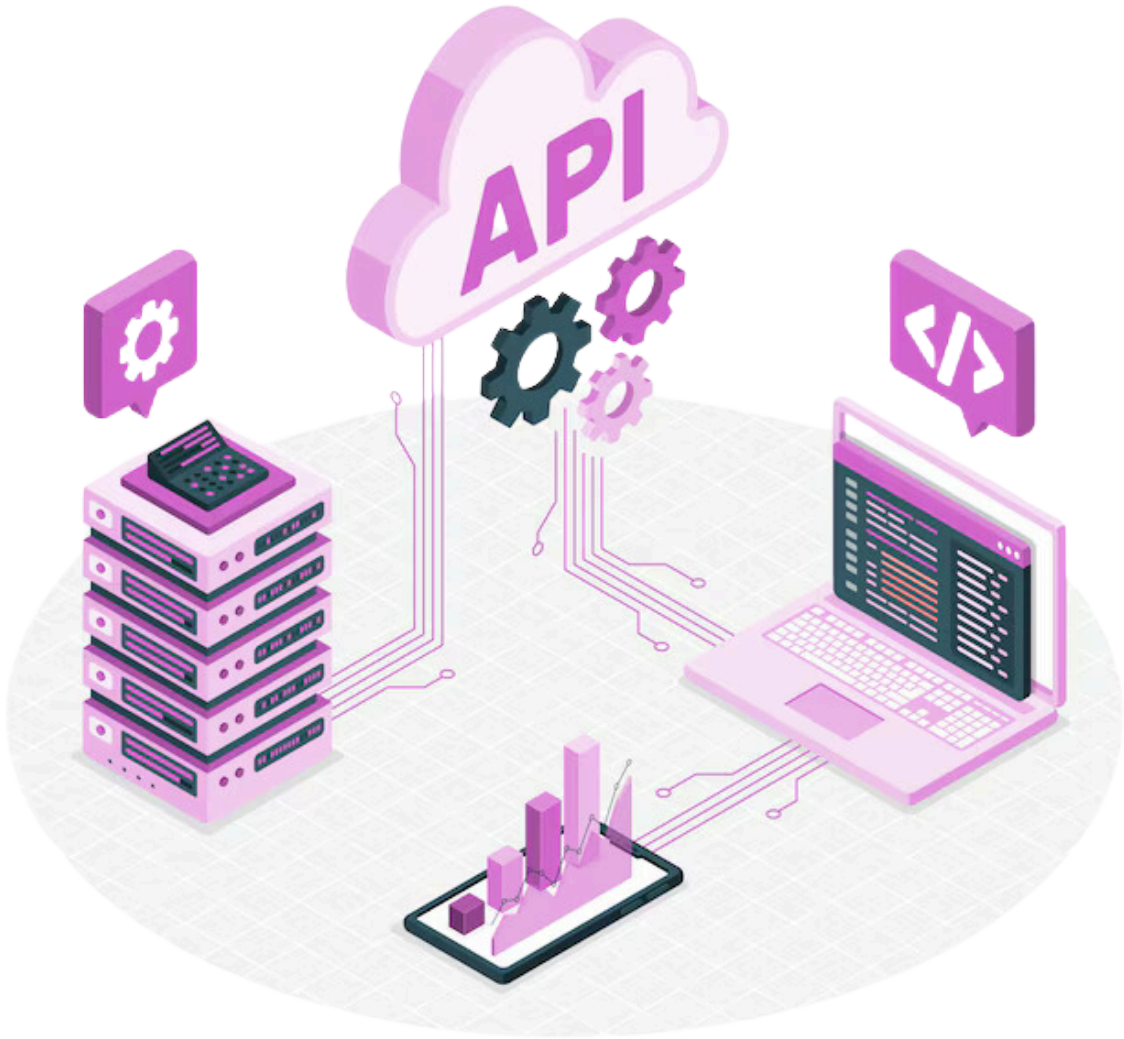
Data pipeline orchestration refers to managing and automating the execution and scheduling of tasks across a data pipeline. It involves coordinating various data processing steps (ETL, data transformations) to ensure seamless, efficient, and error-free operations.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

23. APIs
(Application Programming Interfaces)



Data pipeline orchestration refers to managing and automating the execution and scheduling of tasks across a data pipeline. It involves coordinating various data processing steps (ETL, data transformations) to ensure seamless, efficient, and error-free operations.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

24. Data Security

Data security involves implementing policies and technologies to protect data from unauthorized access, corruption, or loss. It includes encryption, access control, monitoring, and compliance with privacy regulations to ensure the confidentiality and integrity of data.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

25. Data Lineage

Data lineage refers to the tracing and visualization of data's lifecycle, from its origin (source) through various stages of processing and transformation to its final destination. Understanding data lineage is crucial for tracking data quality, compliance, and auditing purposes.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

26. Data Virtualization

Data virtualization is the process of creating a unified, abstract view of data from multiple sources without physically moving or replicating the data. It enables real-time access to data from disparate systems, making it easier to query and analyze.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

27. Streaming Data

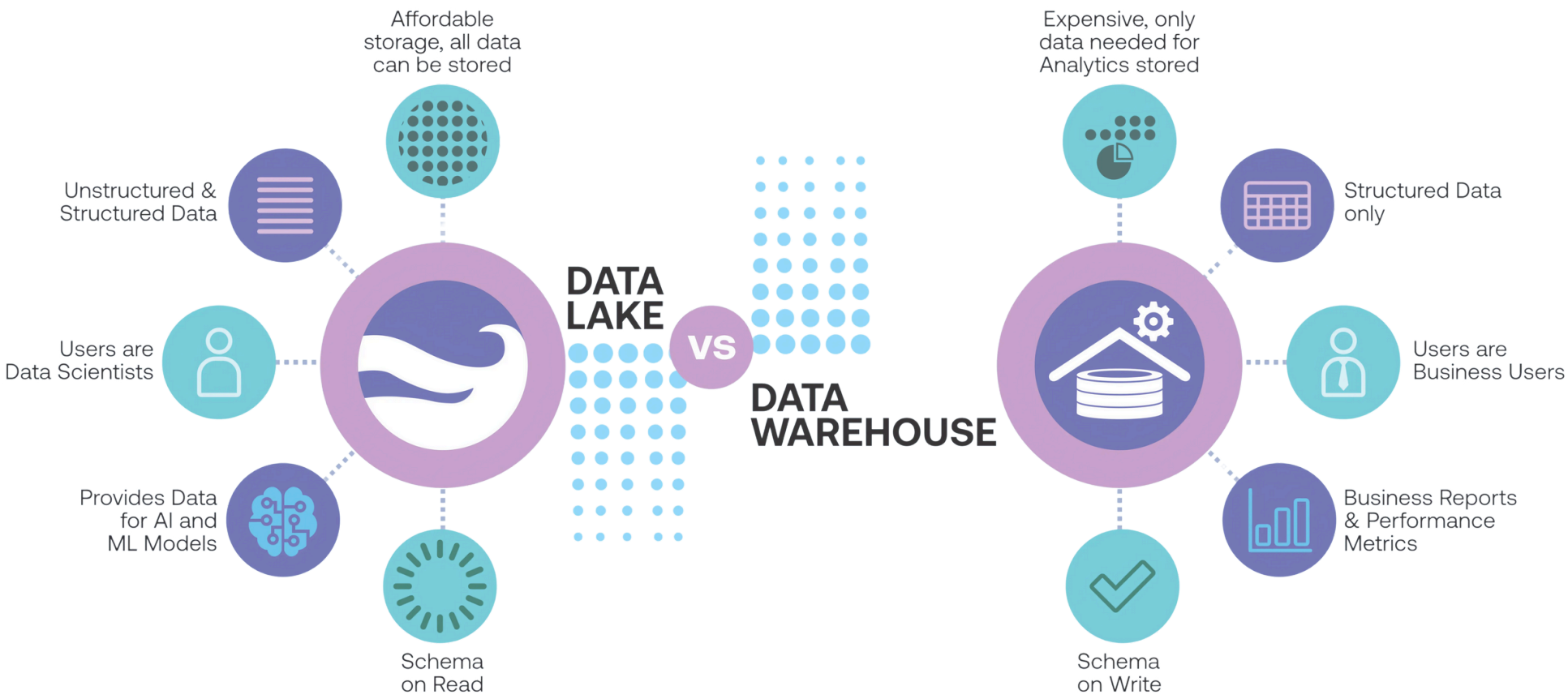
Streaming data refers to continuously generated data that is processed and analyzed in real time, often in systems like social media feeds, sensor networks, and financial markets. It requires specialized technologies like Apache Kafka or Apache Flink to process and analyze data as it is produced.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

28. Data Warehouse vs Data Lake



A data warehouse stores structured data that has been pre-processed and is optimized for querying, while a data lake holds raw data (structured and unstructured) in its native form, providing a scalable and flexible environment for future processing, machine learning, and advanced analytics.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

29. Data Federation

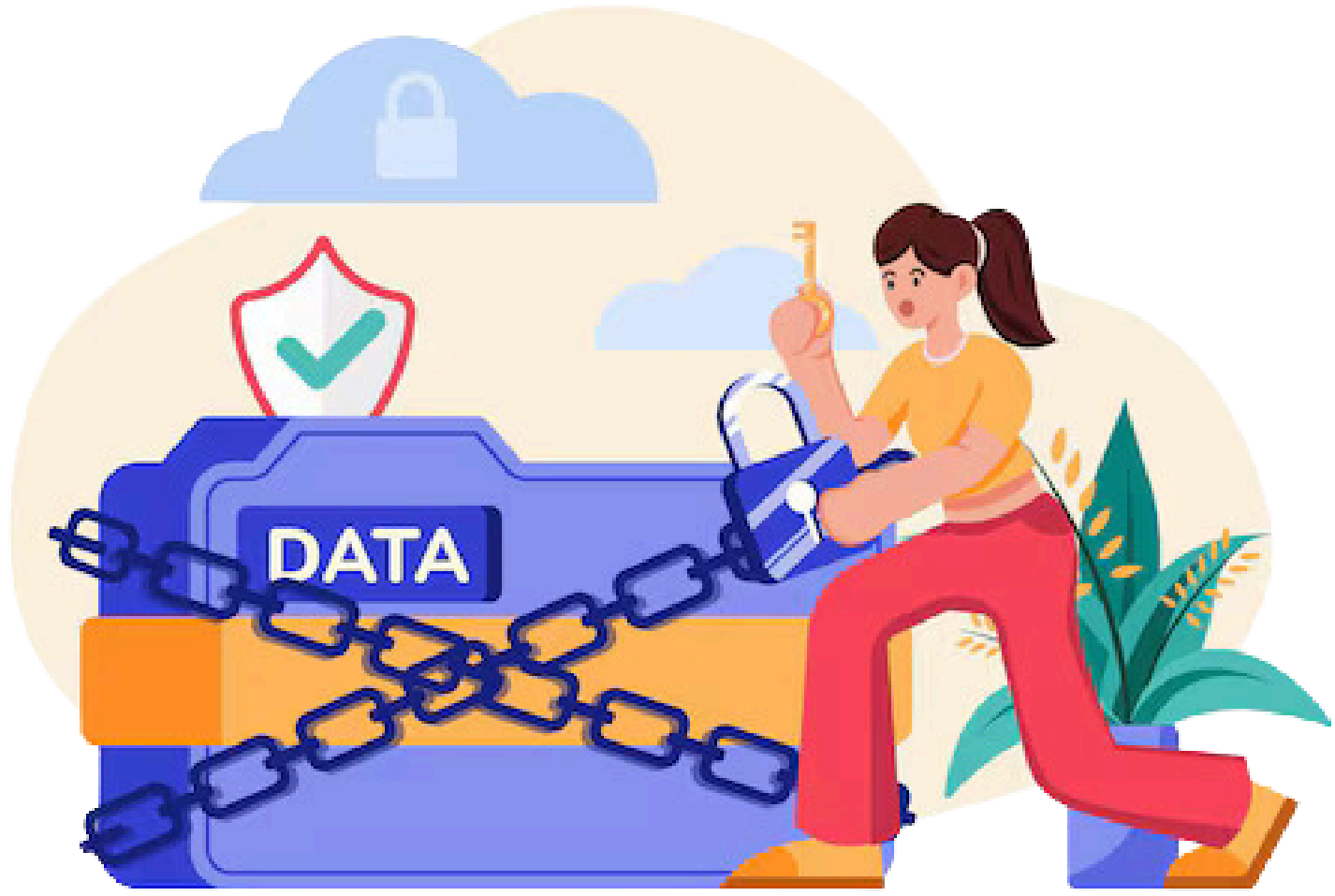
Data federation allows for the creation of a unified data view by accessing data from multiple systems or sources without the need to move or replicate the data. It simplifies querying across disparate data systems, providing a single interface for data access.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

30. Data Encryption



Data encryption is the process of converting data into a coded form to prevent unauthorized access. It is commonly used during data transmission (in transit) or while the data is stored (at rest) to ensure confidentiality and security.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

31. Data Architecture

Data architecture refers to the design of data systems, processes, and technologies used to collect, store, manage, and analyze data. A strong data architecture ensures that data is organized, accessible, and scalable while meeting performance and security requirements.

Swipe

32. Data Processing Engine

A data processing engine is a software system or platform designed to process large volumes of data, often in parallel, using tools like Apache Spark, Apache Flink, or Google BigQuery. These engines are optimized for speed and scalability to handle complex data processing tasks.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

33. NoSQL Databases

NoSQL databases are non-relational databases designed to handle unstructured, semi-structured, and highly scalable data. They use flexible data models such as key-value pairs, graphs, or documents, and are often used for big data and real-time applications where traditional SQL databases may fall short.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

34. SQL Databases

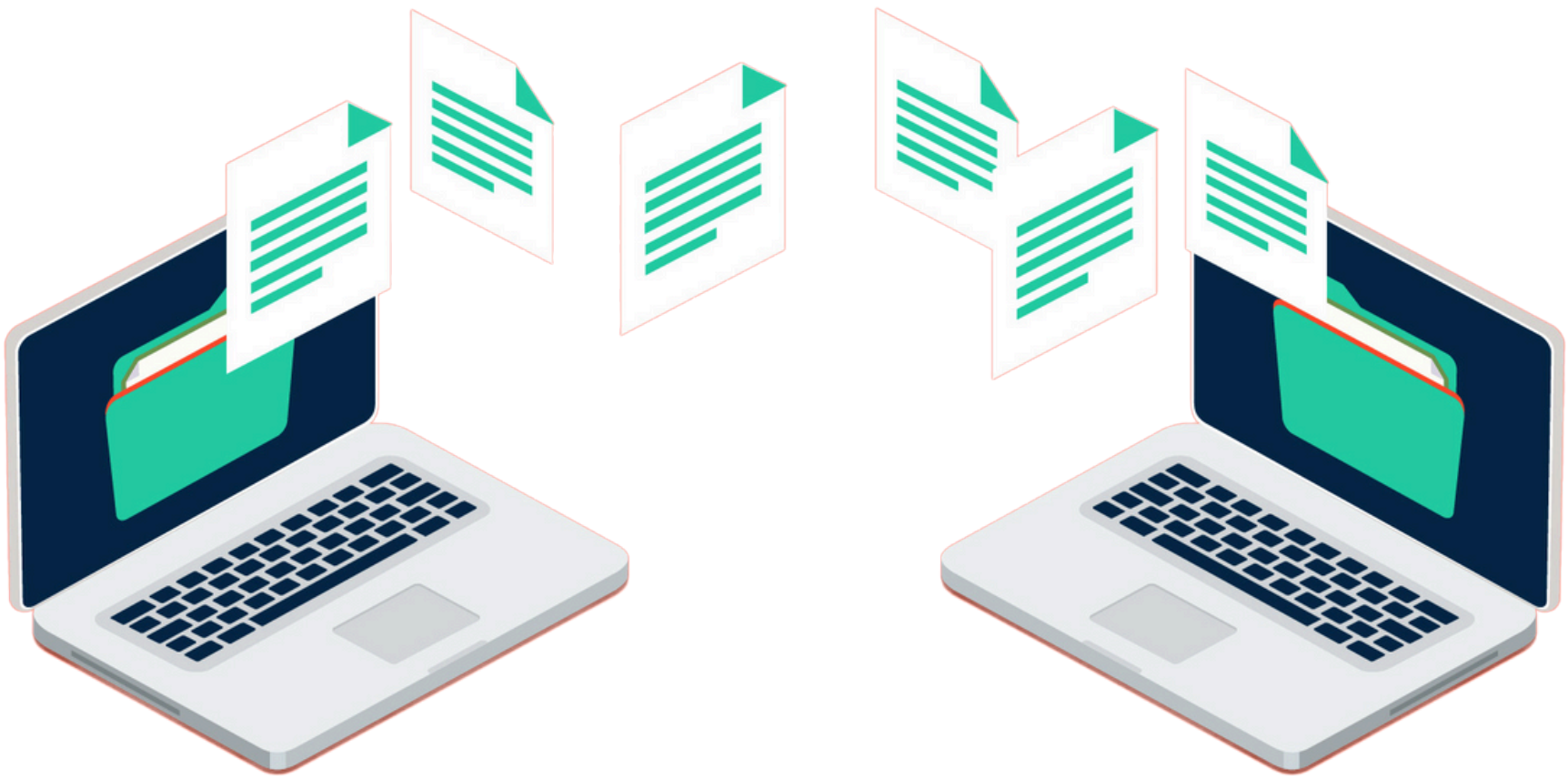
SQL databases are relational databases that store data in tables with predefined relationships between them. They use Structured Query Language (SQL) to manage and query structured data, typically suited for transaction-oriented applications like e-commerce or financial systems.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

35. Data Replication



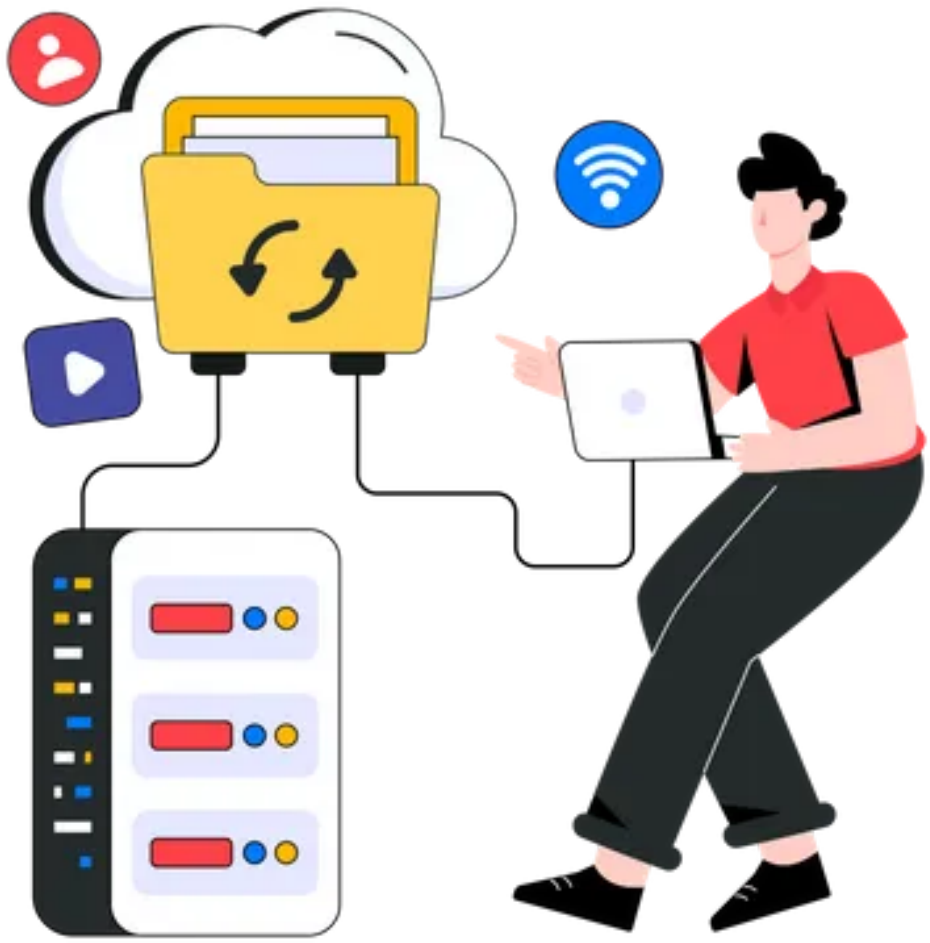
Data replication is the process of copying data from one system to another to ensure data availability, reliability, and fault tolerance. This can be done in real-time (synchronous) or in batches (asynchronous) depending on the use case.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

36. Data Synchronization



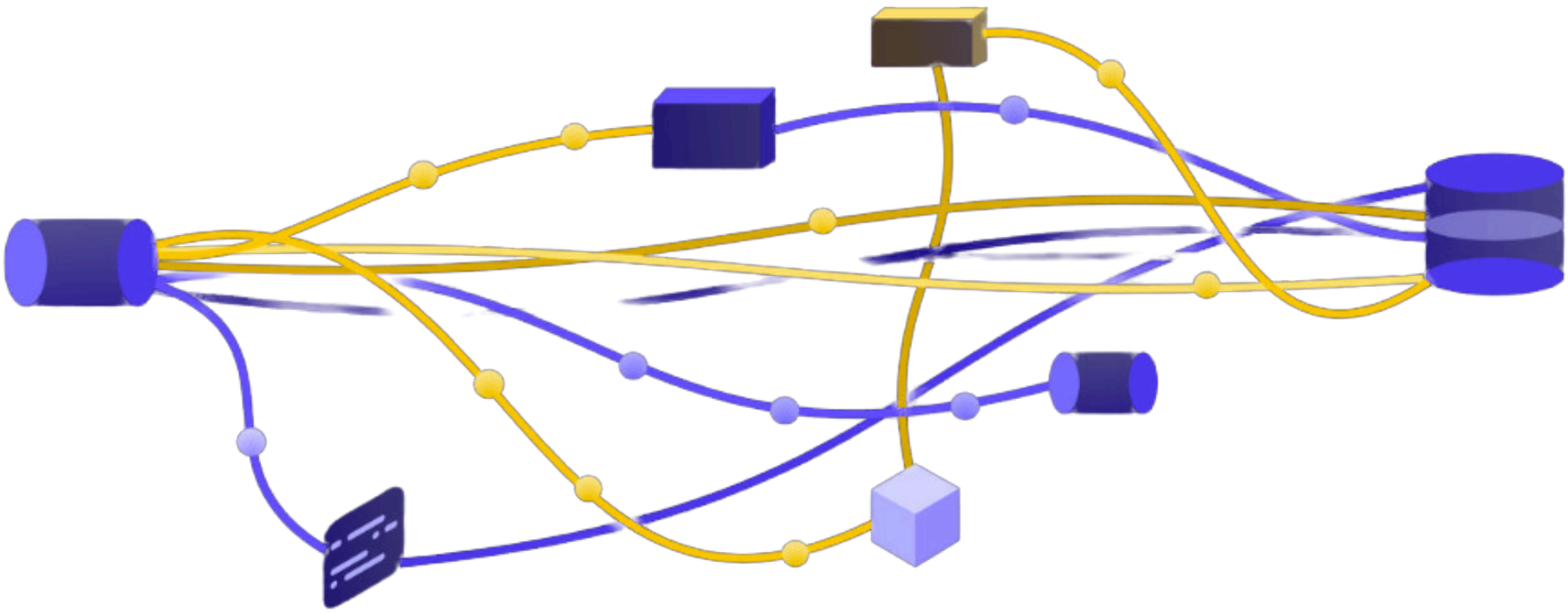
Data synchronization ensures that data across multiple systems or locations remains consistent and up-to-date. This is especially important when data is distributed across different databases, applications, or cloud platforms.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

37. Data Fabric



Data fabric is an integrated layer of data and technologies designed to provide seamless access to data across the organization. It enables efficient data management, governance, and analysis by connecting disparate data sources, both on-premises and in the cloud.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

38. Data Mart

A data mart is a subset of a data warehouse, focusing on a specific business area or department (e.g., finance, marketing). It simplifies querying by providing a specialized, smaller data repository that is tailored to the needs of a particular team or function.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

39. OLTP
(Online Transaction Processing)

OLTP refers to a type of data processing used in systems that manage real-time transactions, such as banking or e-commerce. OLTP databases are optimized for fast insert, update, and delete operations and are used for managing day-to-day transactional data.

Swipe

THE RAVIT SHOW

Don't Forget to Save For Later

40. OLAP
(Online Analytical Processing)

OLAP refers to systems optimized for complex querying and data analysis. OLAP databases allow users to interactively analyze large datasets from multiple dimensions, often used in business intelligence tools for creating reports, dashboards, and data visualizations.

Swipe

Don't Forget to
Save For Later

Was it useful?

Let me know in the comments

THE RAVIT SHOW

@theravitshow