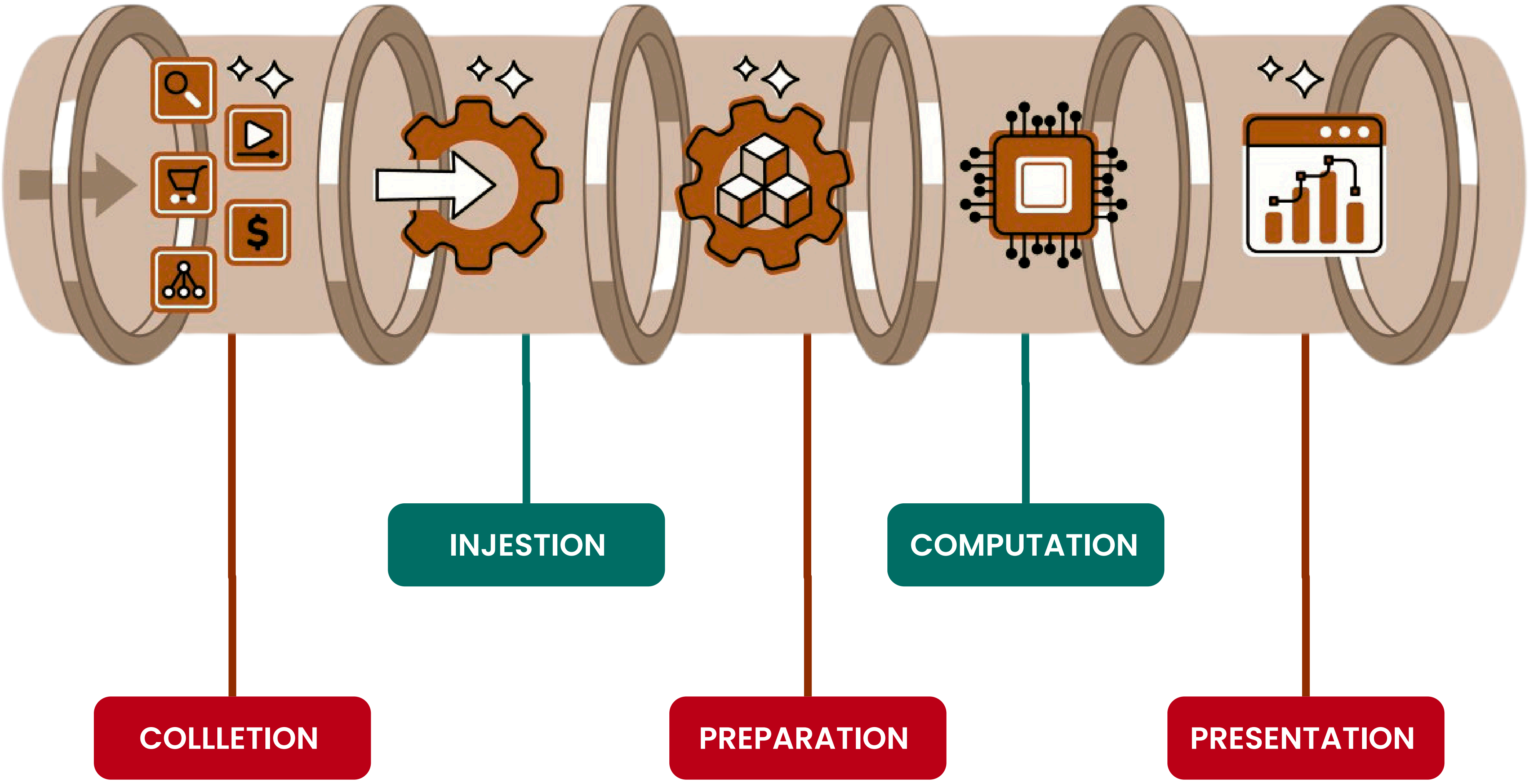
DATA ENGINEERING

TERMS YOU NEED TO KNOW

Don't Forget to
Save For Later

Data Pipeline

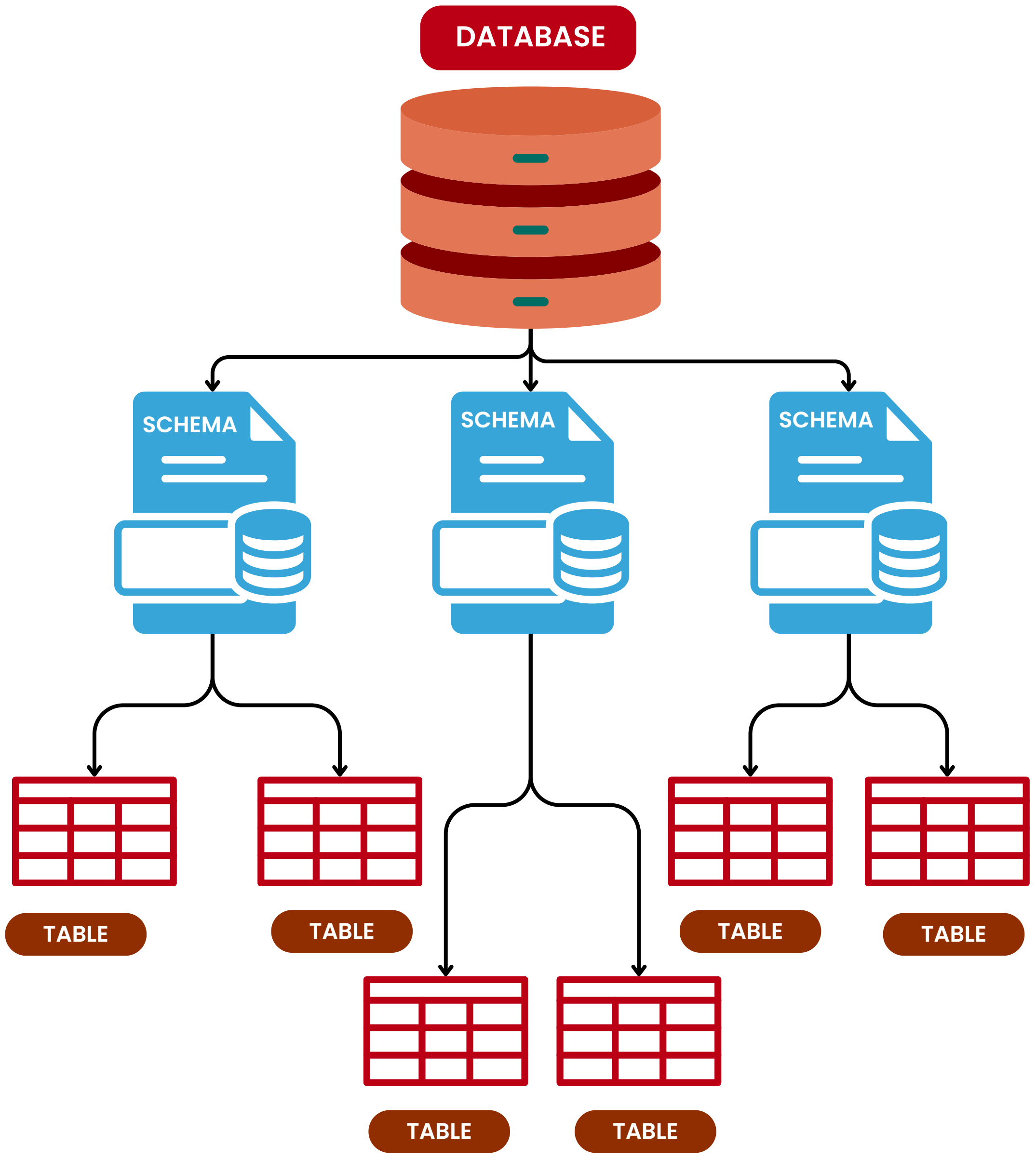


Data pipeline: **Mapping data flow from collection to analysis for organizational insights. Source, process, deliver for analytics efficiency.**

Swipe

Don't Forget to Save For Later

Database vs Schema vs Table

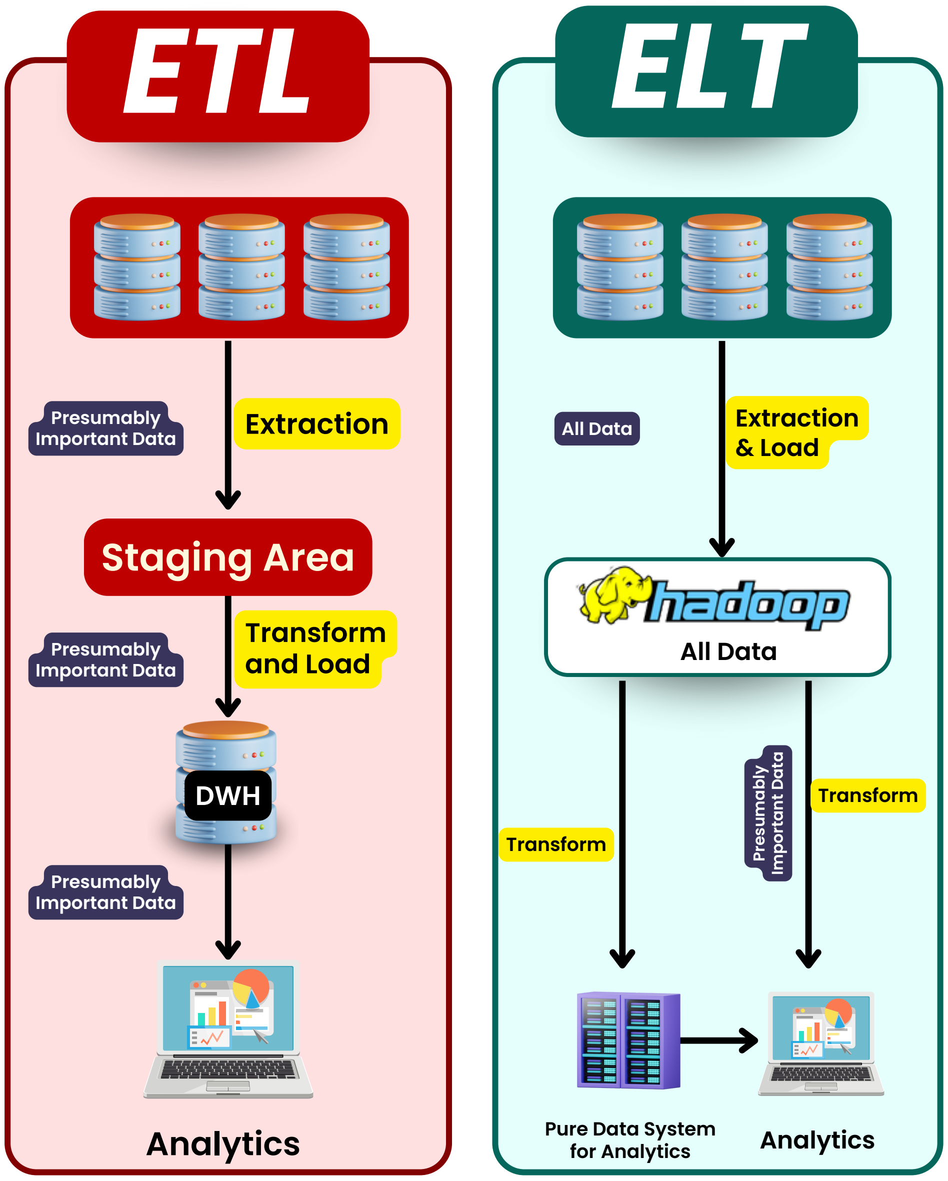


Database: Organized collection of related data, like a file cabinet for storing information.
Schema: Blueprint defining structure and organization within a database.
Table: Grid-like structure within a schema, where data is stored in rows and columns.

Swipe

Don't Forget to
Save For Later

ETL And ELT

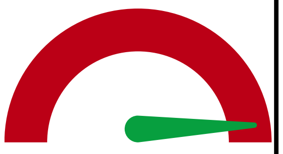
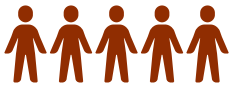
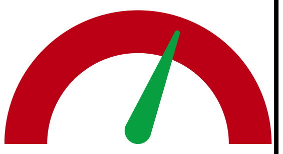
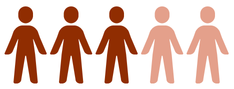
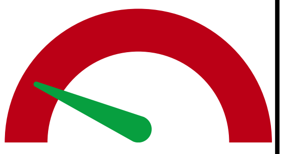
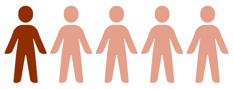

ETL
Presumably Important Data
Extraction
Staging Area
Presumably Important Data
Transform and Load
DWH
Presumably Important Data
Analytics
ELT
All Data
Extraction & Load
hadoop
All Data
Transform
Presumably Important Data
Transform
Pure Data System for Analytics
Analytics

ETL (Extract, Transform, Load): Traditional approach where data is extracted from various sources, transformed to fit into the target schema, and then loaded into the destination.

ELT (Extract, Load, Transform): Newer approach where data is first loaded into the target system, then transformed within the system itself, often leveraging the power of modern data warehouses.

Swipe

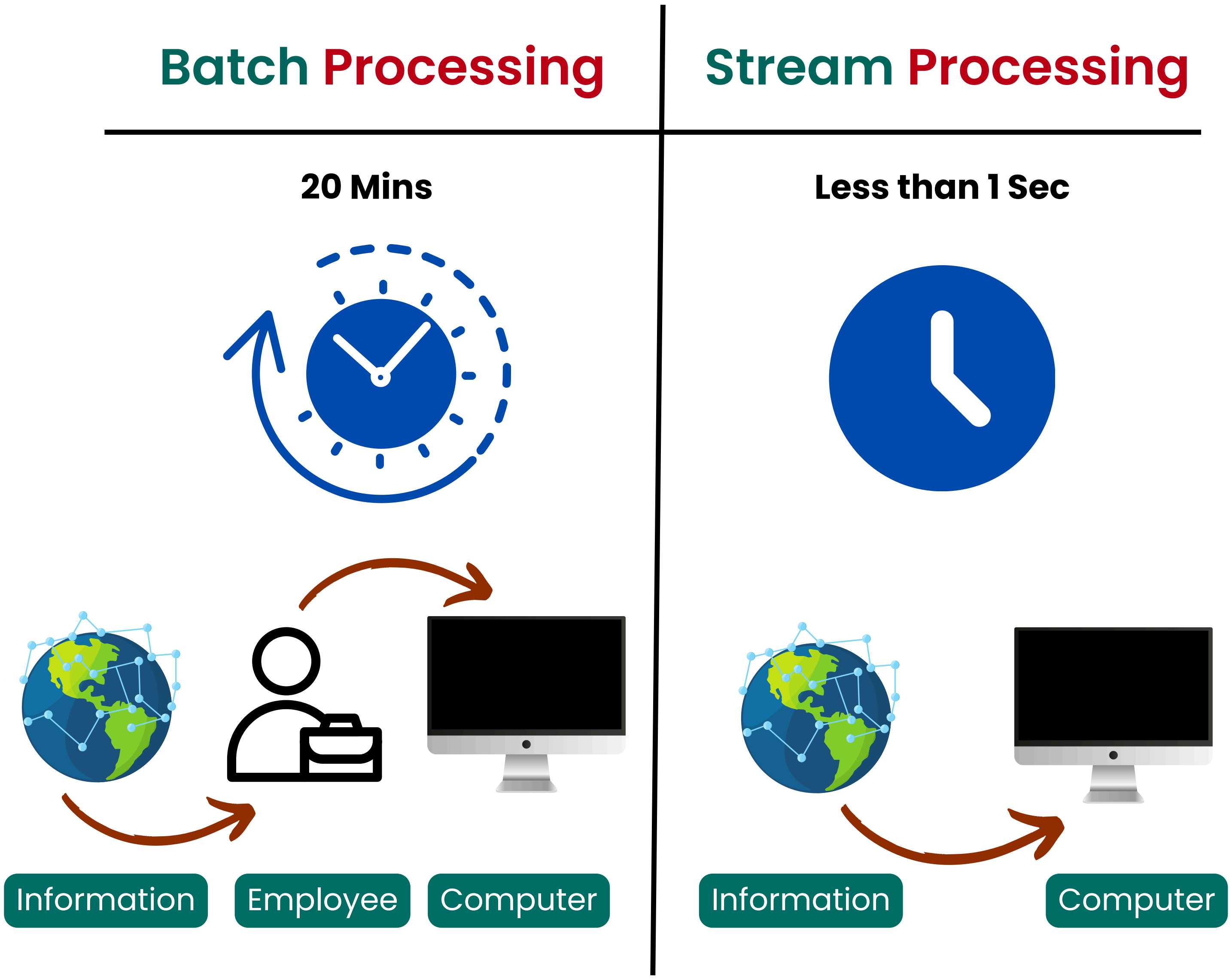
Data Lake vs Data warehouse vs Data Mart

	Important Use	Time To Market	Cost	Data Growth
Data Lake	Predective & advanced Analytics	Weeks to Months		
Data Ware house	Operational & Performance Analytics	hours to Days		
Data Mart	Business specific Reporting & Analytics	Minutes to Hours		

Data Lake: Unstructured repository for diverse data types, preserving raw format.
Data Warehouse: Centralized, structured storage optimized for analytics and reporting.
Data Mart: Tailored subsets of data warehouse, focused on specific business areas.

Don't Forget to
Save For Later

Batch vs Stream Processing

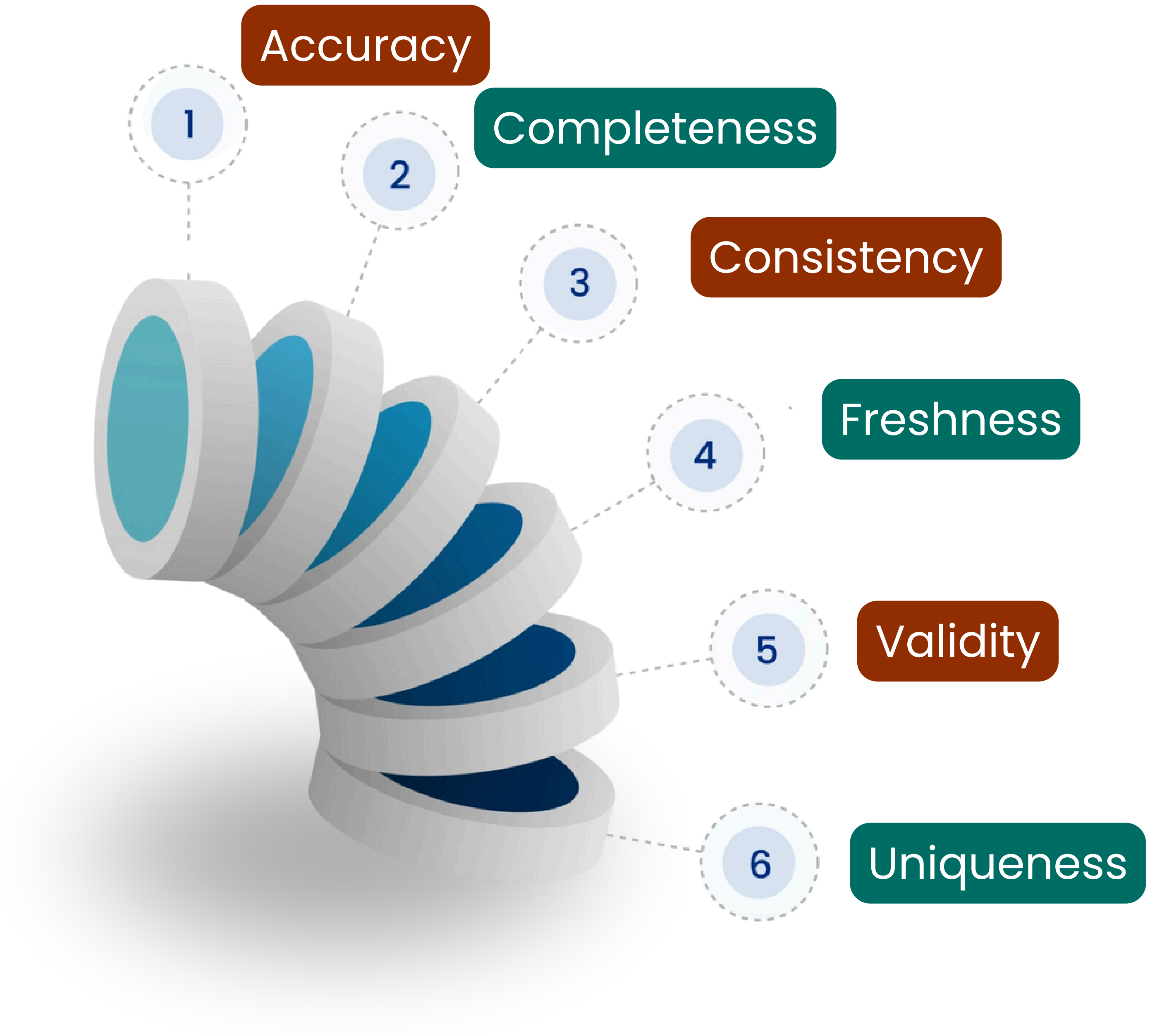


Batch Processing: Handling large volumes of data at scheduled intervals, ideal for complex computations and historical analysis.
Stream Processing: Real-time data processing, analyzing data as it arrives, suitable for immediate insights and actions.

Swipe

Don't Forget to Save For Later

Data Quality

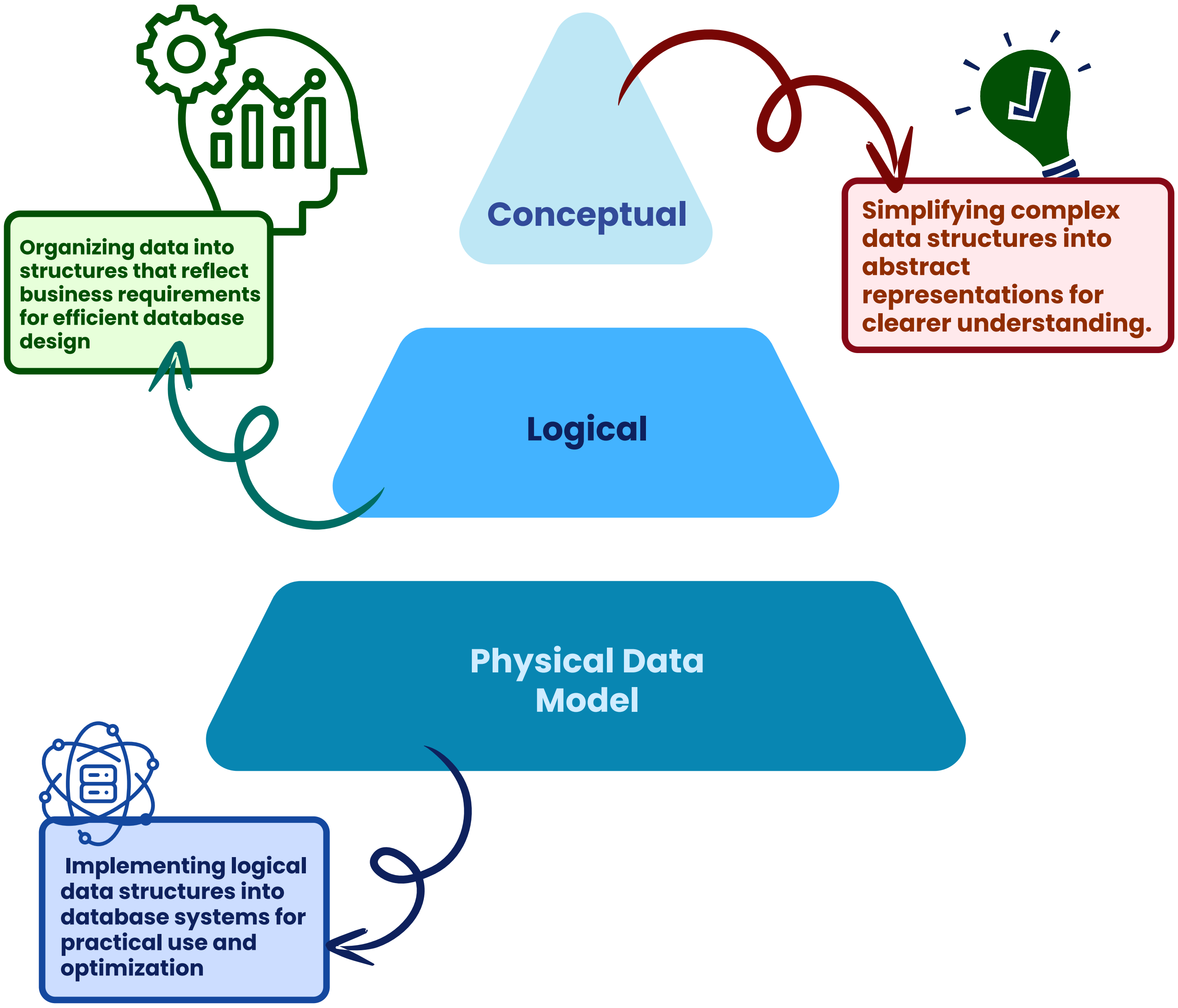


Data quality ensures accuracy, completeness, and reliability of data for informed decision-making. It involves maintaining consistency, relevance, and timeliness within datasets through governance, management, and cleansing processes.

Swipe →

Don't Forget to
Save For Later

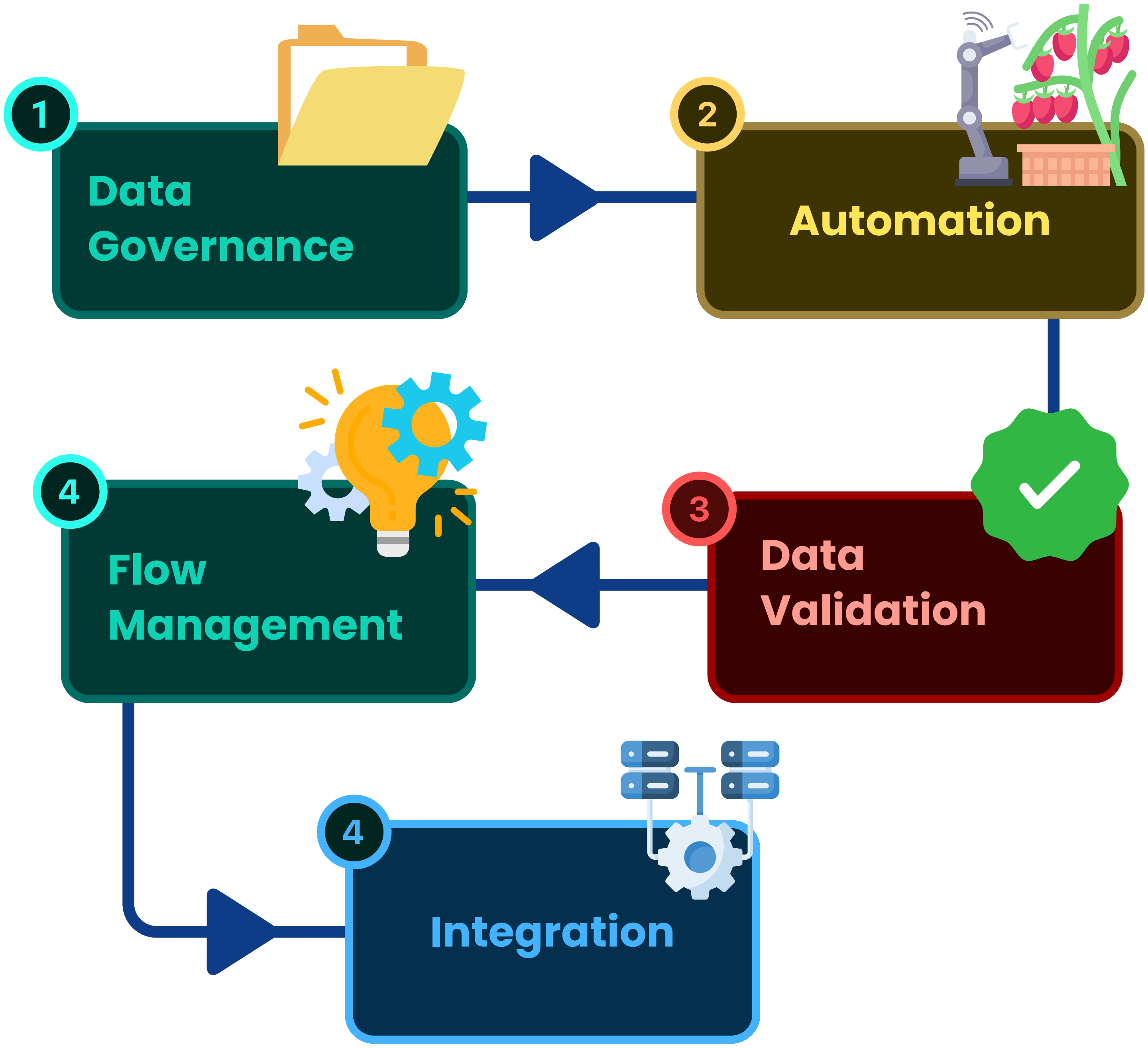
Data Modelling


Conceptual
Logical
Physical Data Model
Organizing data into structures that reflect business requirements for efficient database design
Simplifying complex data structures into abstract representations for clearer understanding.
Implementing logical data structures into database systems for practical use and optimization

Data modeling: Designing the structure and relationships of data to facilitate efficient storage, retrieval, and analysis, typically represented through diagrams or schemas.

Swipe →

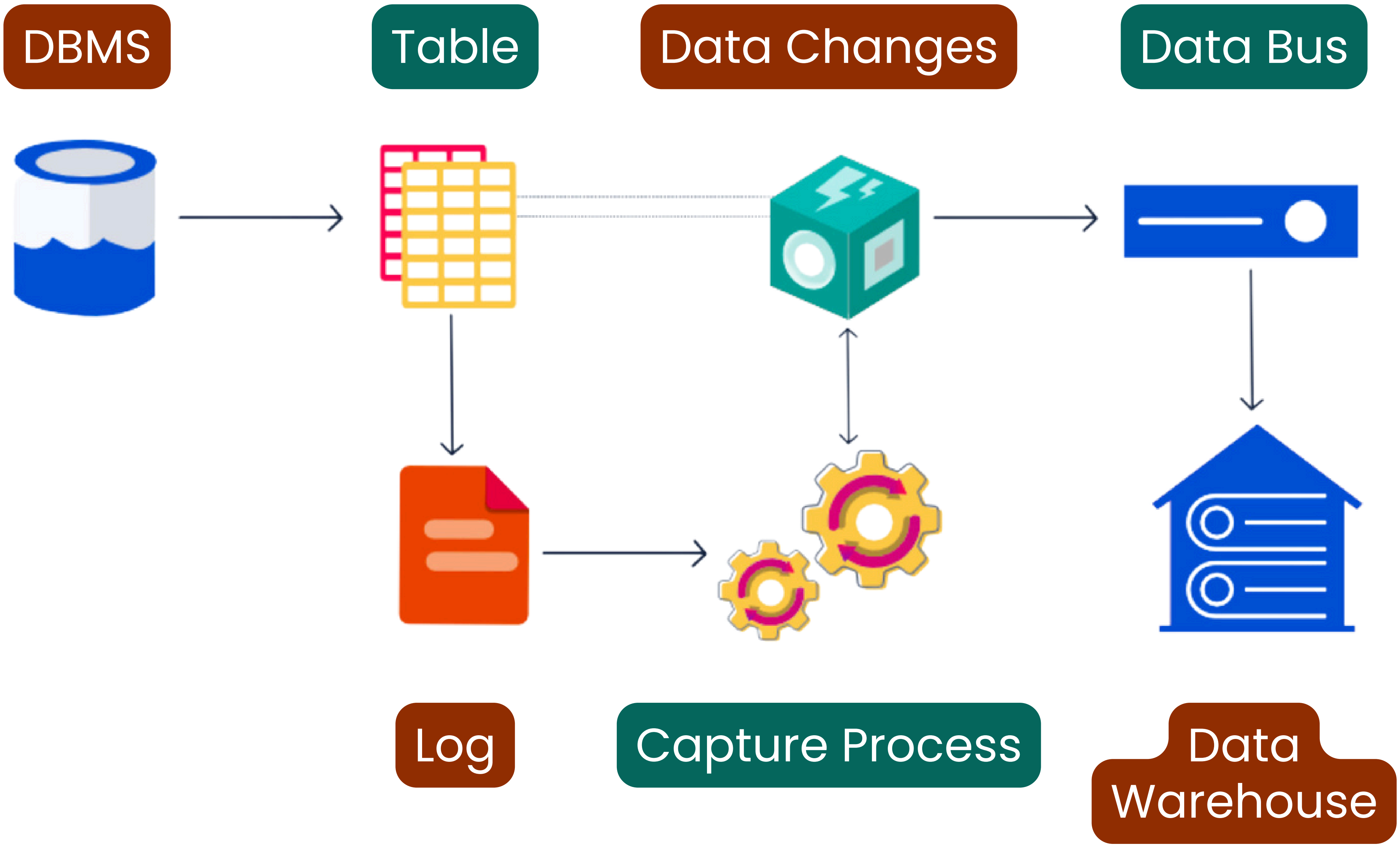
Data Orchestration



Data Orchestration: Coordinating and automating the flow of data across systems, processes, and environments to ensure seamless integration, transformation, and delivery.

Swipe

Data Lineage



DBMS

Table

Data Changes

Data Bus

Log

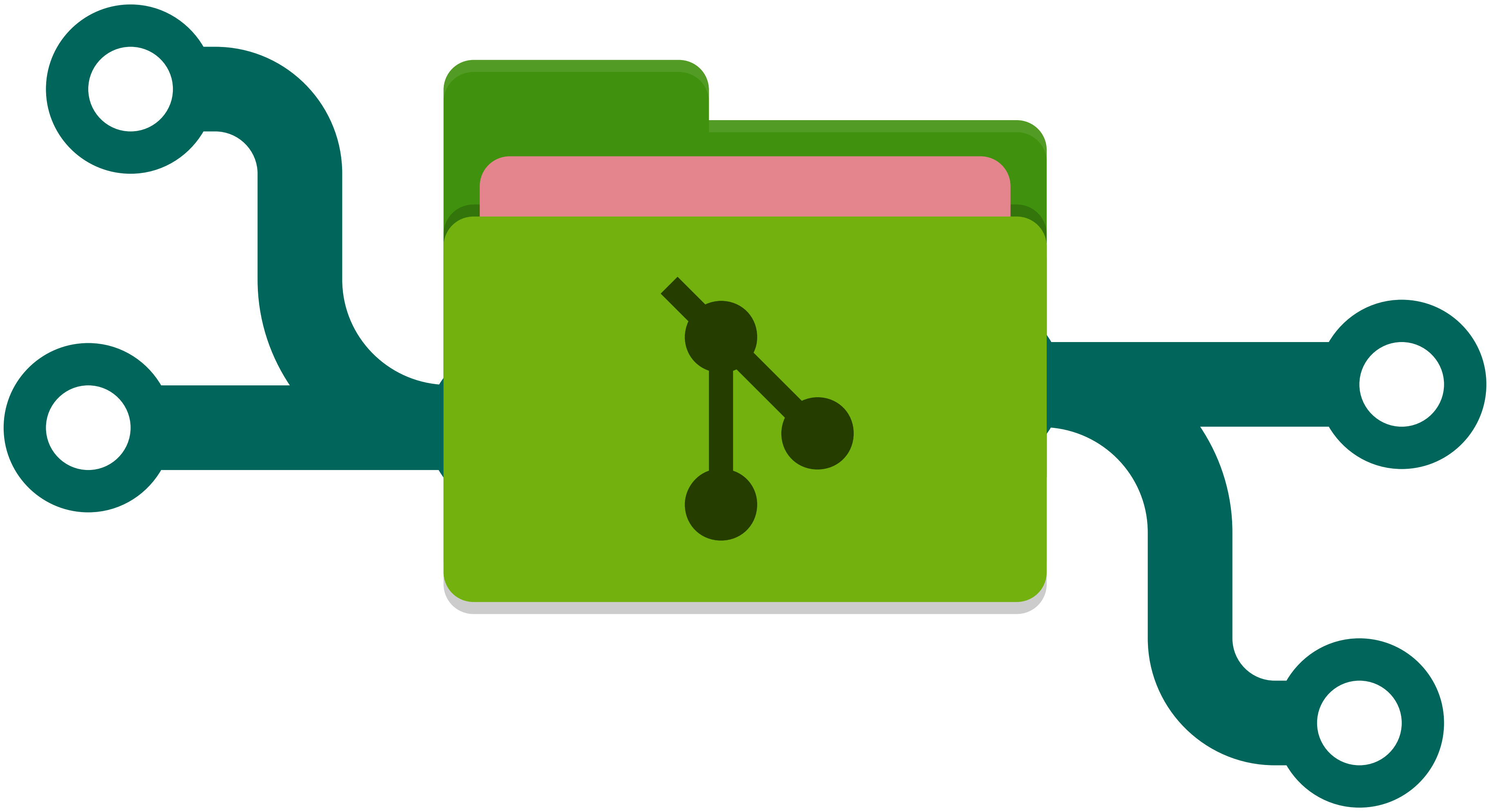
Capture Process

Data
Warehouse

Data lineage traces the path of data from its origin to its destination, including any transformations along the way. It provides transparency and insight into data's journey, helping ensure accuracy, compliance, and trust in data processes.

Don't Forget to
Save For Later

Git



Git is a Version control system for tracking changes in code, facilitating collaboration and project management. Utilizes branches and commits to manage code history and team workflows.

Swipe

Don't Forget to
Save For Later

Was it useful?

Let me know in the comments

THE
RAVIT
SHOW

@theravitshow