

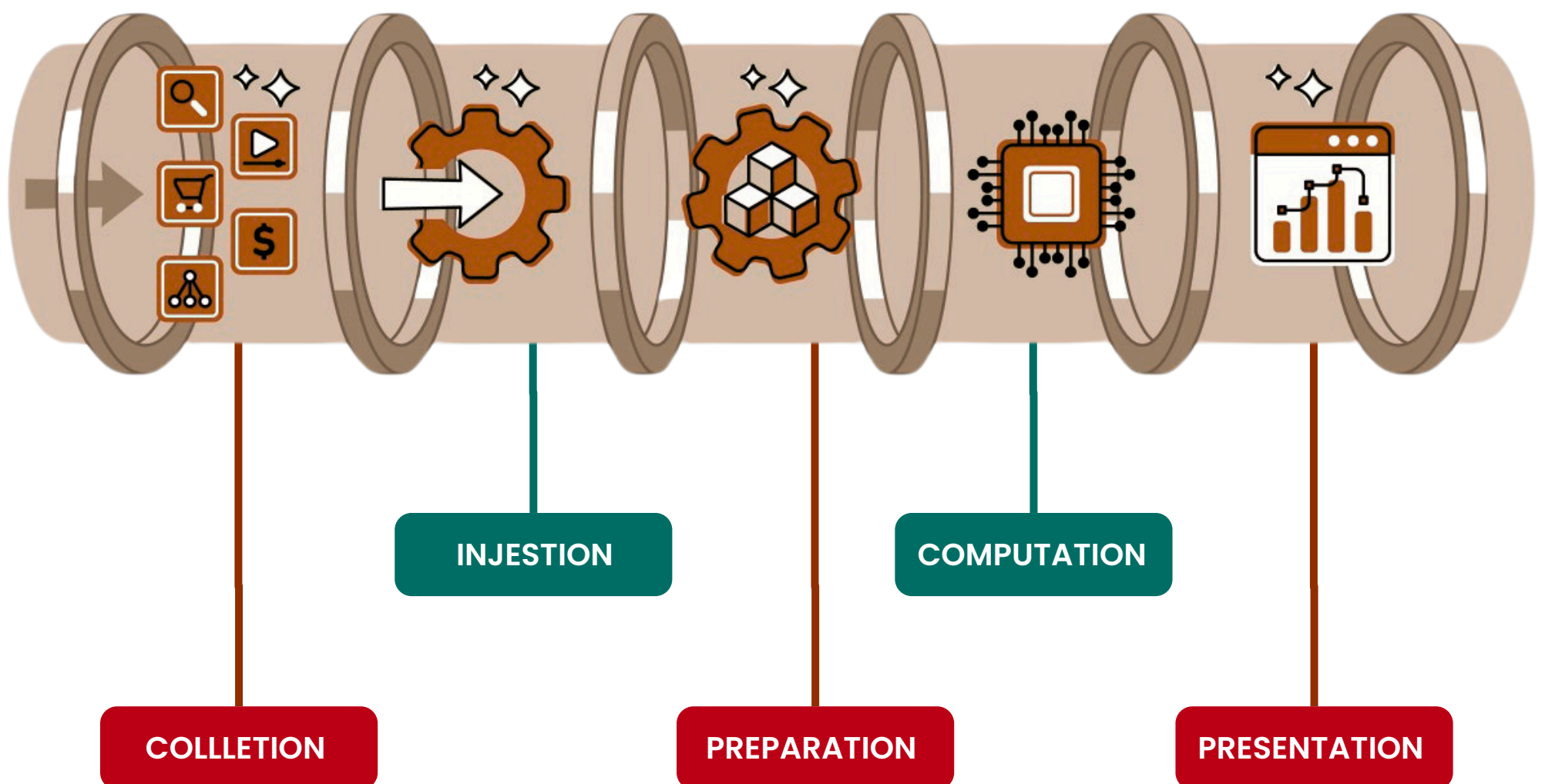
DATA ENGINEERING

TERMS YOU NEED TO KNOW





Data Pipeline



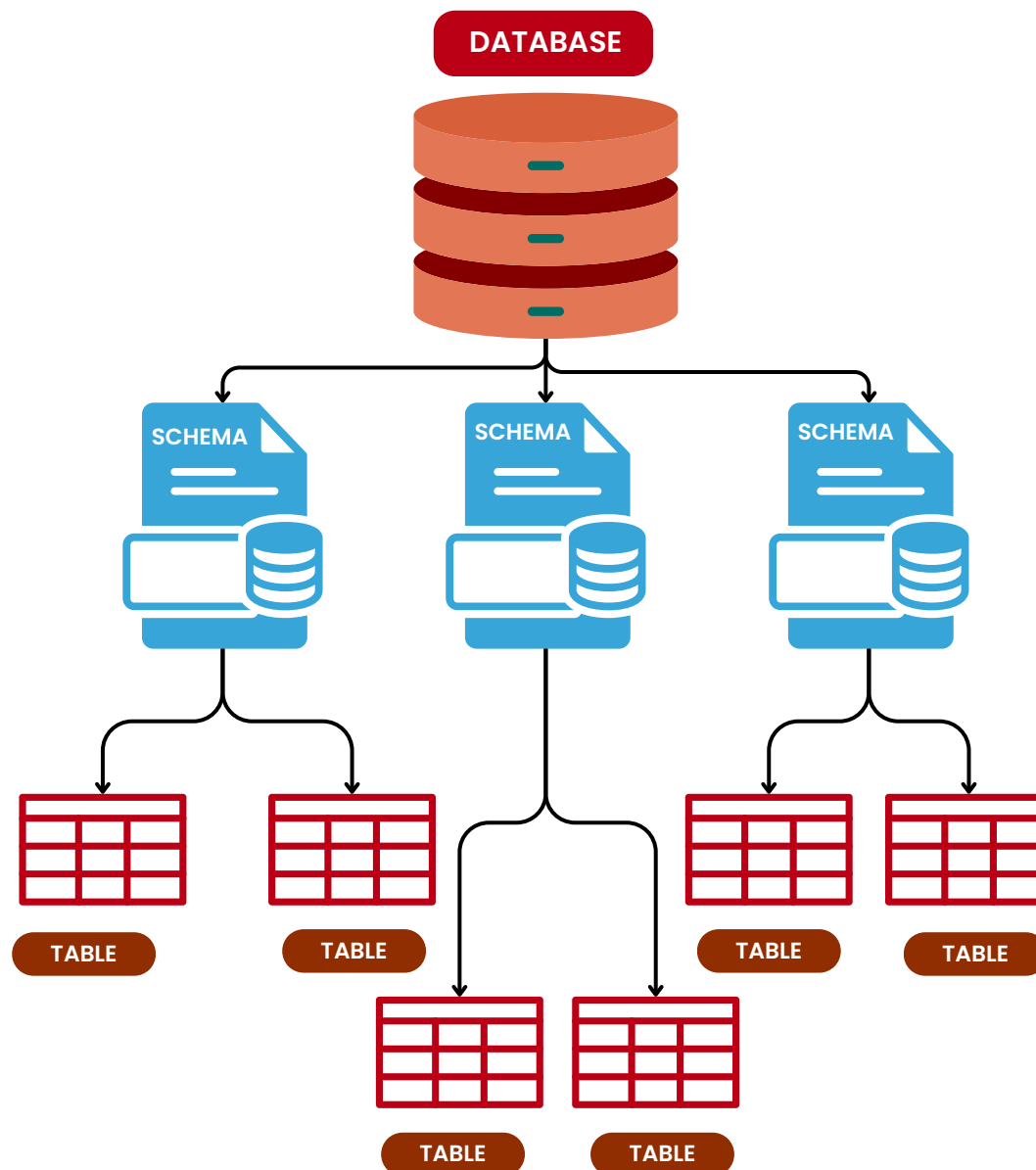
Data pipeline: Mapping data flow from collection to analysis for organizational insights. Source, process, deliver for analytics efficiency.

Swipe





Database vs Schema vs Table



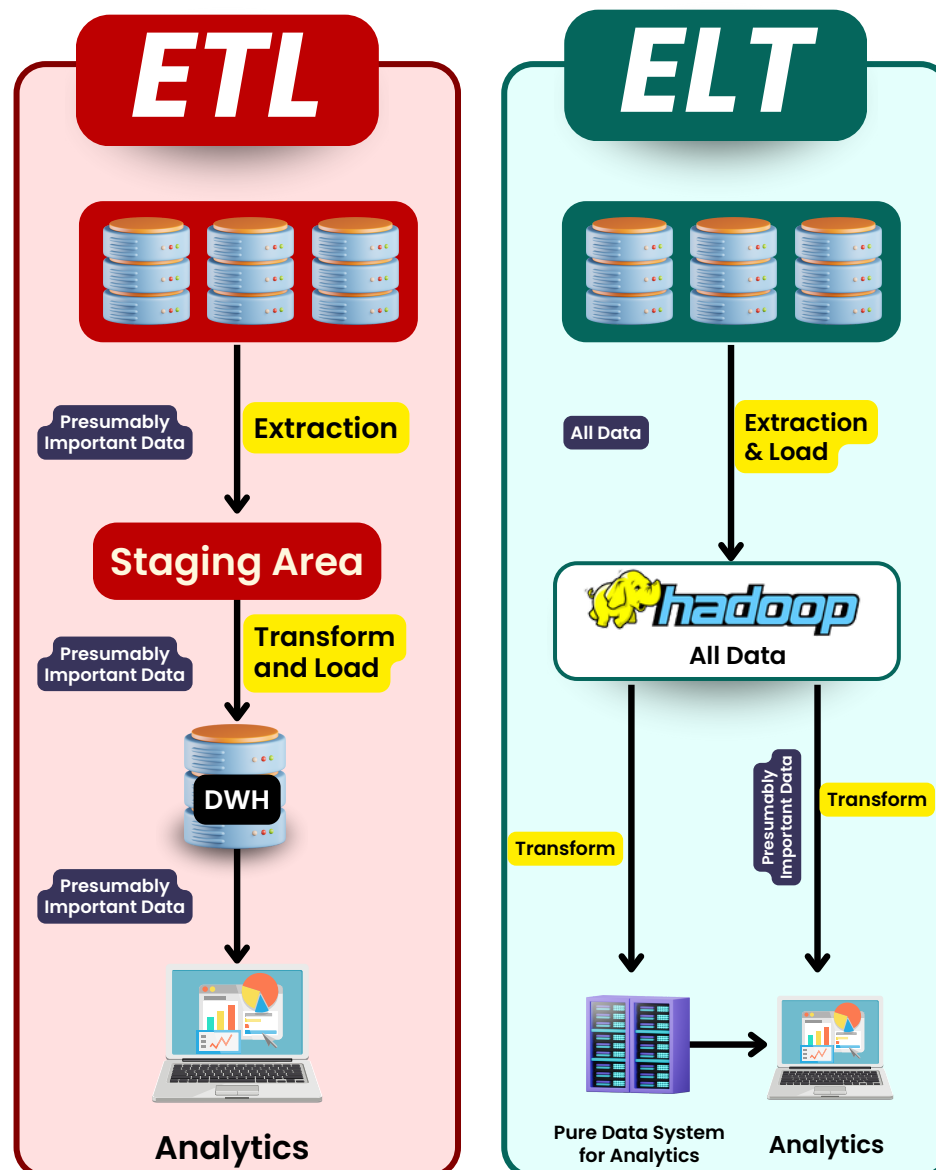
Database: Organized collection of related data, like a file cabinet for storing information.

Schema: Blueprint defining structure and organization within a database.

Table: Grid-like structure within a schema, where data is stored in rows and columns.



ETL And ELT





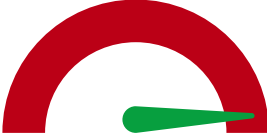







ETL (Extract, Transform, Load): Traditional approach where data is extracted from various sources, transformed to fit into the target schema, and then loaded into the destination.

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.





Data Lake vs Data warehouse vs Data Mart

	Important Use	Time To Market	Cost	Data Growth
Data Lake	Predictive & advanced Analytics	 Weeks to Months		
Data Warehouse	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.

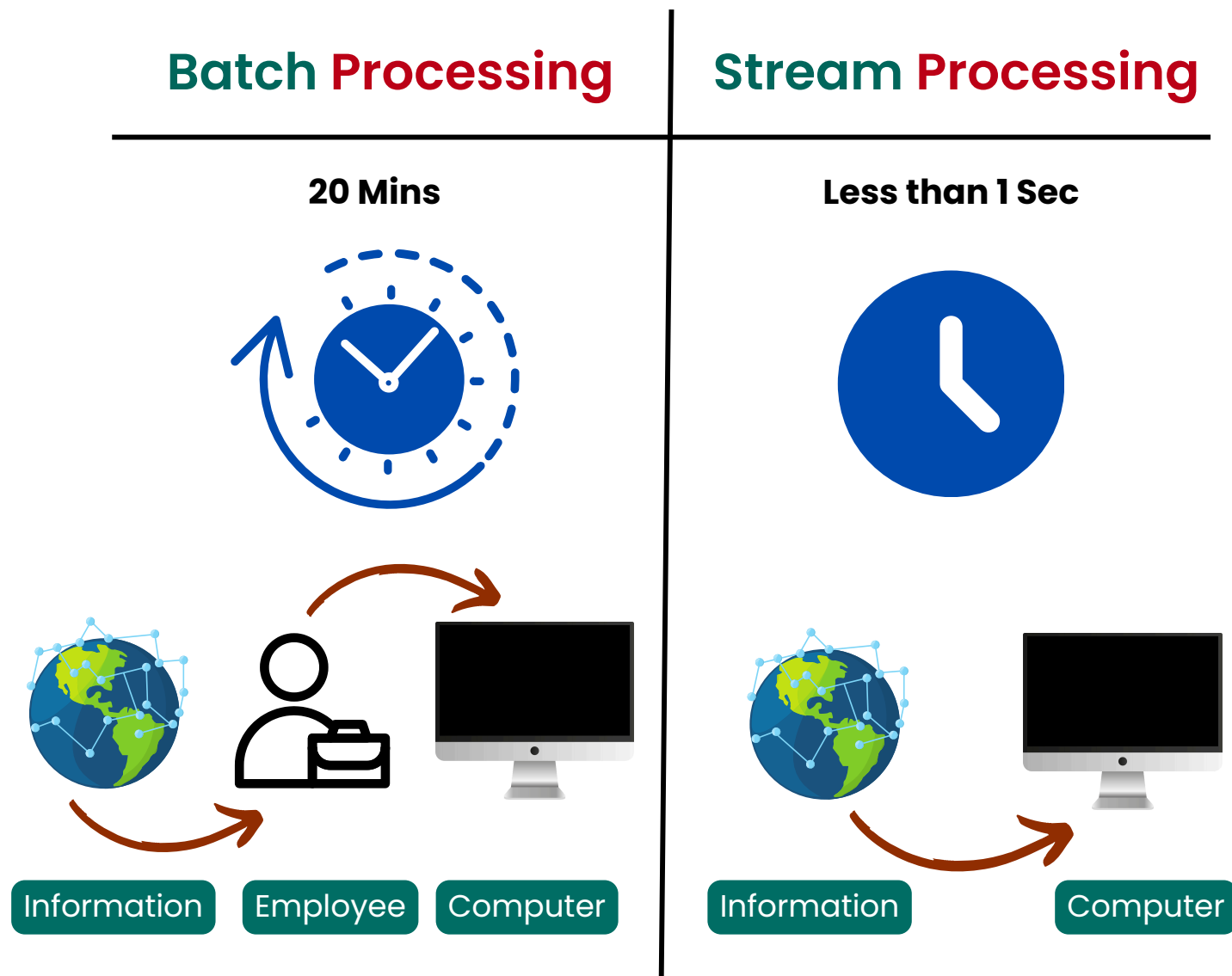
Data Mart: Tailored subsets of data warehouse, focused on specific business areas.

Swipe





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





Data Quality



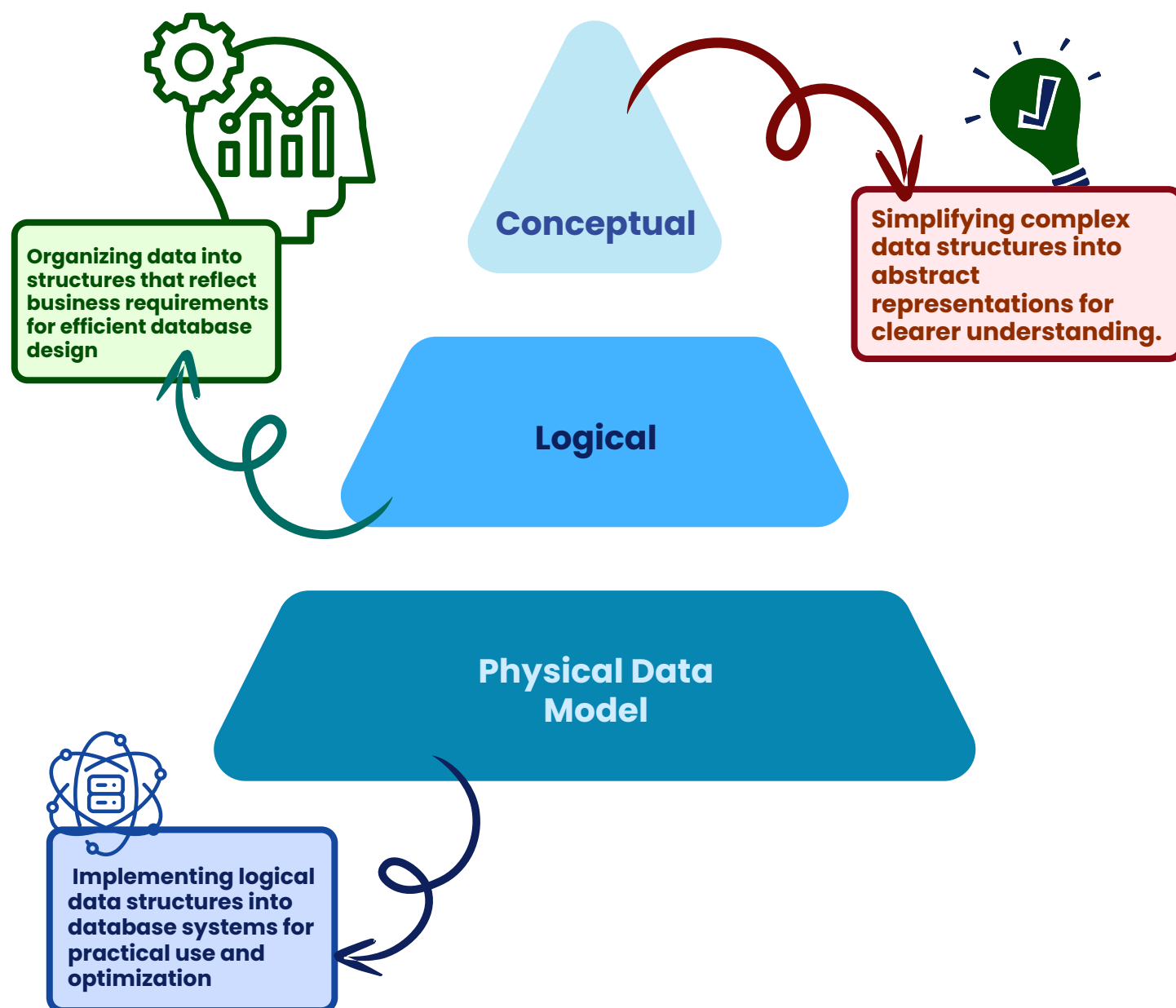
Data quality ensures accuracy, completeness, and reliability of data for informed decision-making. It involves maintaining consistency, relevance, and timeliness within datasets through governance, management, and cleansing processes.

Swipe





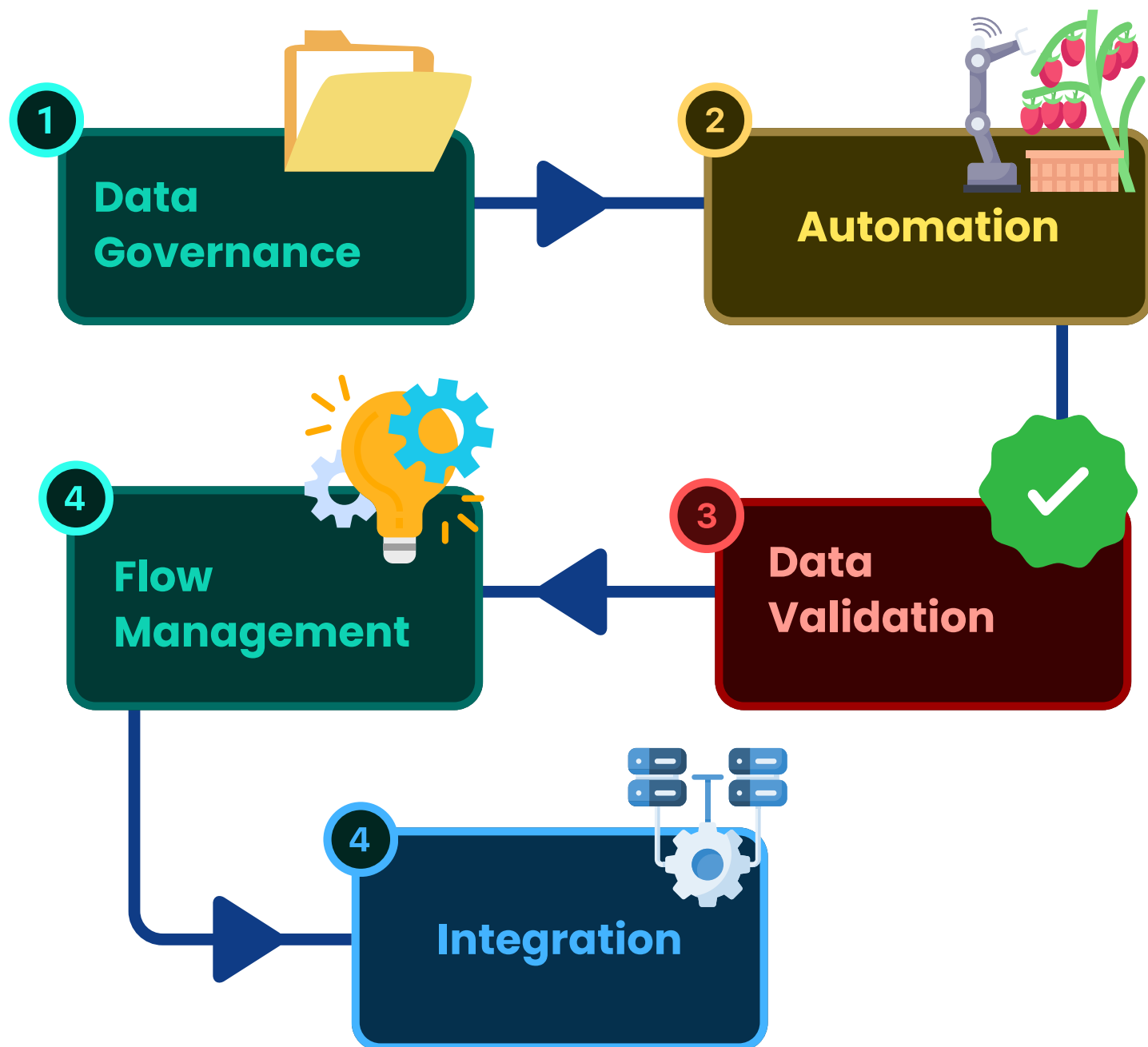
Data Modelling



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



Data Orchestration



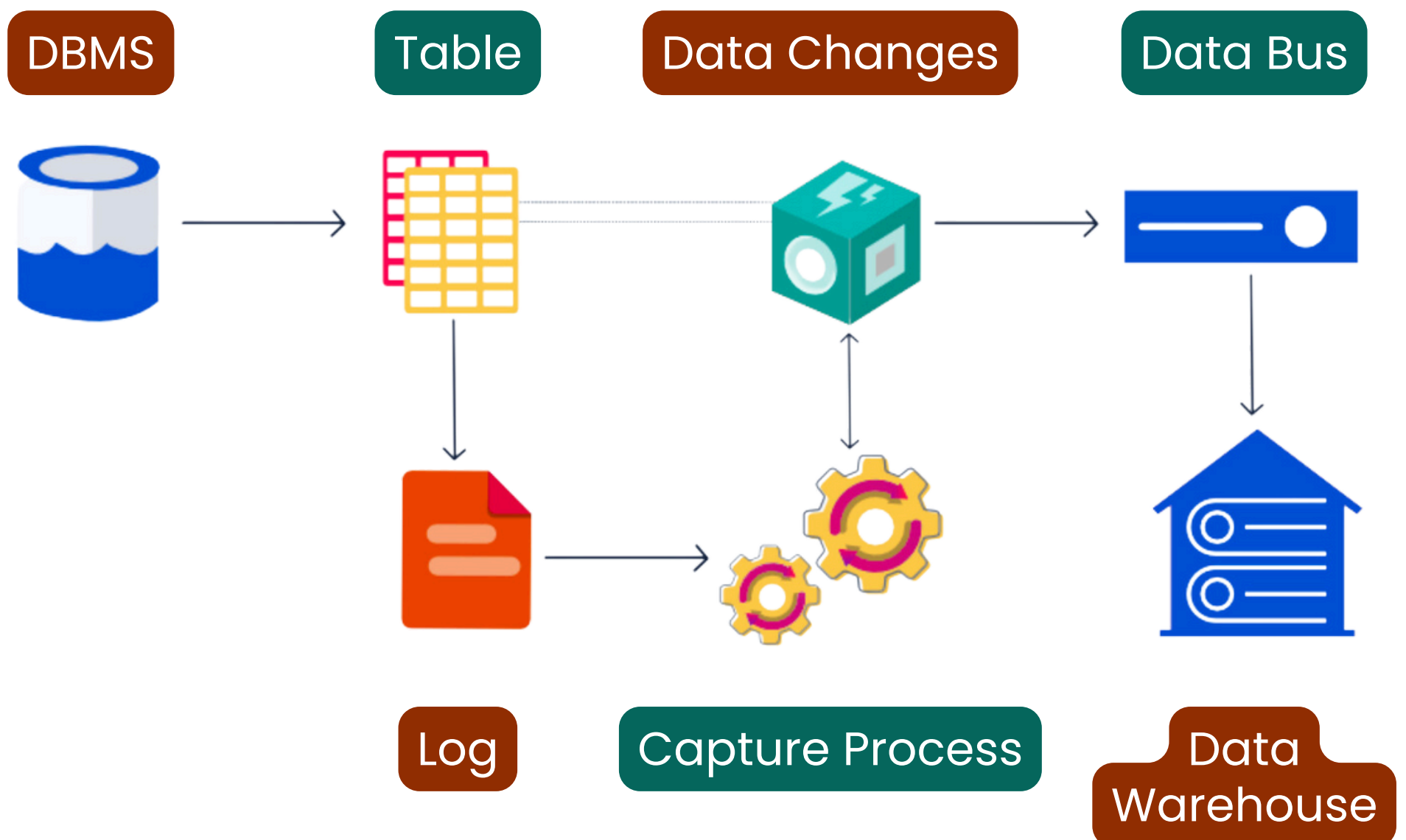
Data Orchestration: Coordinating and automating the flow of data across systems, processes, and environments to ensure seamless integration, transformation, and delivery.

Swipe





Data Lineage



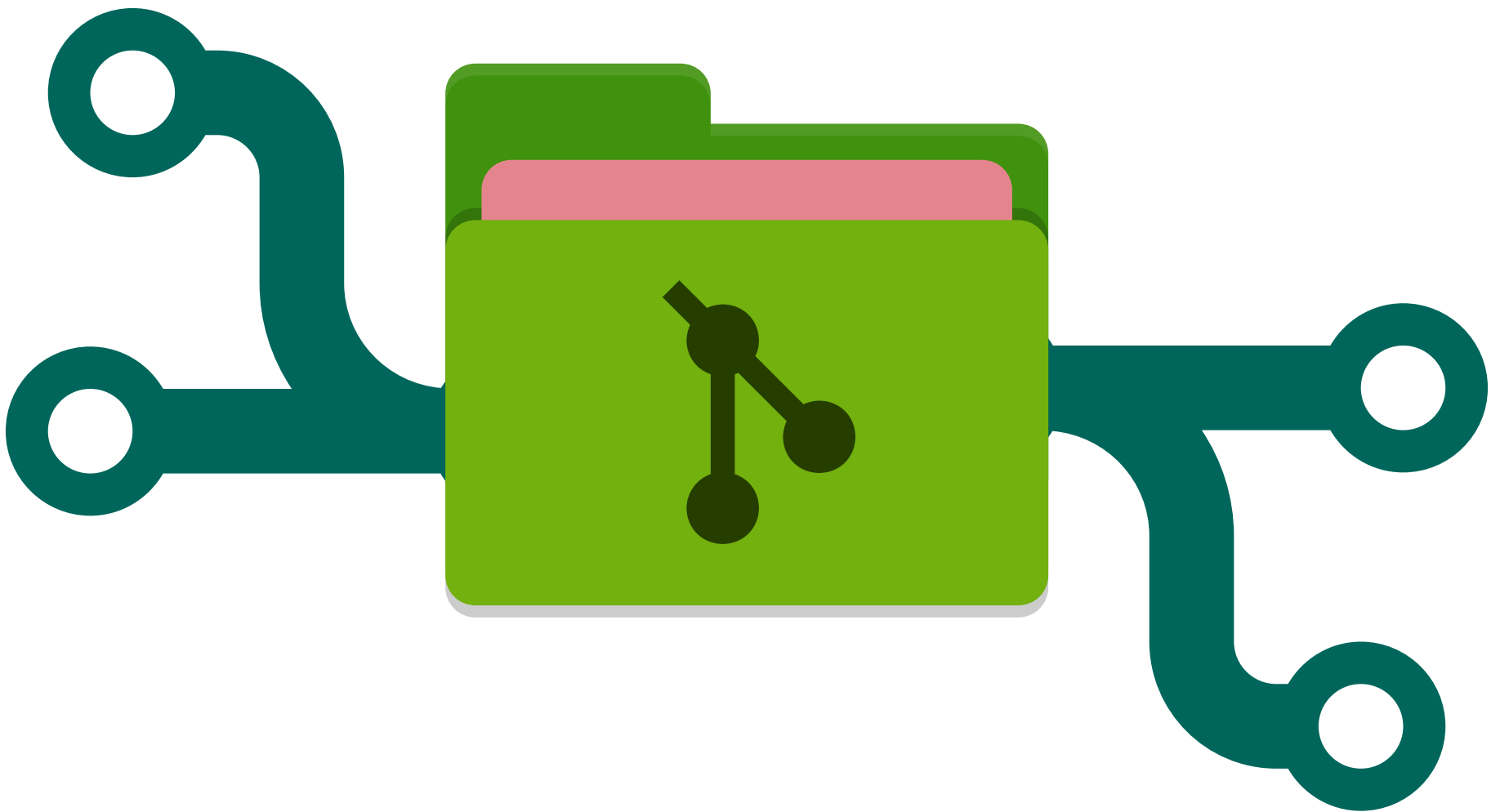
Data lineage traces the path of data from its origin to its destination, including any transformations along the way. It provides transparency and insight into data's journey, helping ensure accuracy, compliance, and trust in data processes.

Swipe





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





Was it useful?

Let me know in the comments



@theravitshow