ETL Process

(Extraction, Transformation, Load)

What is[1] Extraction

This is the first stage, where data is pulled from various sources.

Extraction Methods

- **Full Extraction**: Extract all data at once (complete dataset).
- **Push Extraction**: The source pushes the data to the system.
- **Pull Extraction**: The system pulls the data from the source.

Extract Types

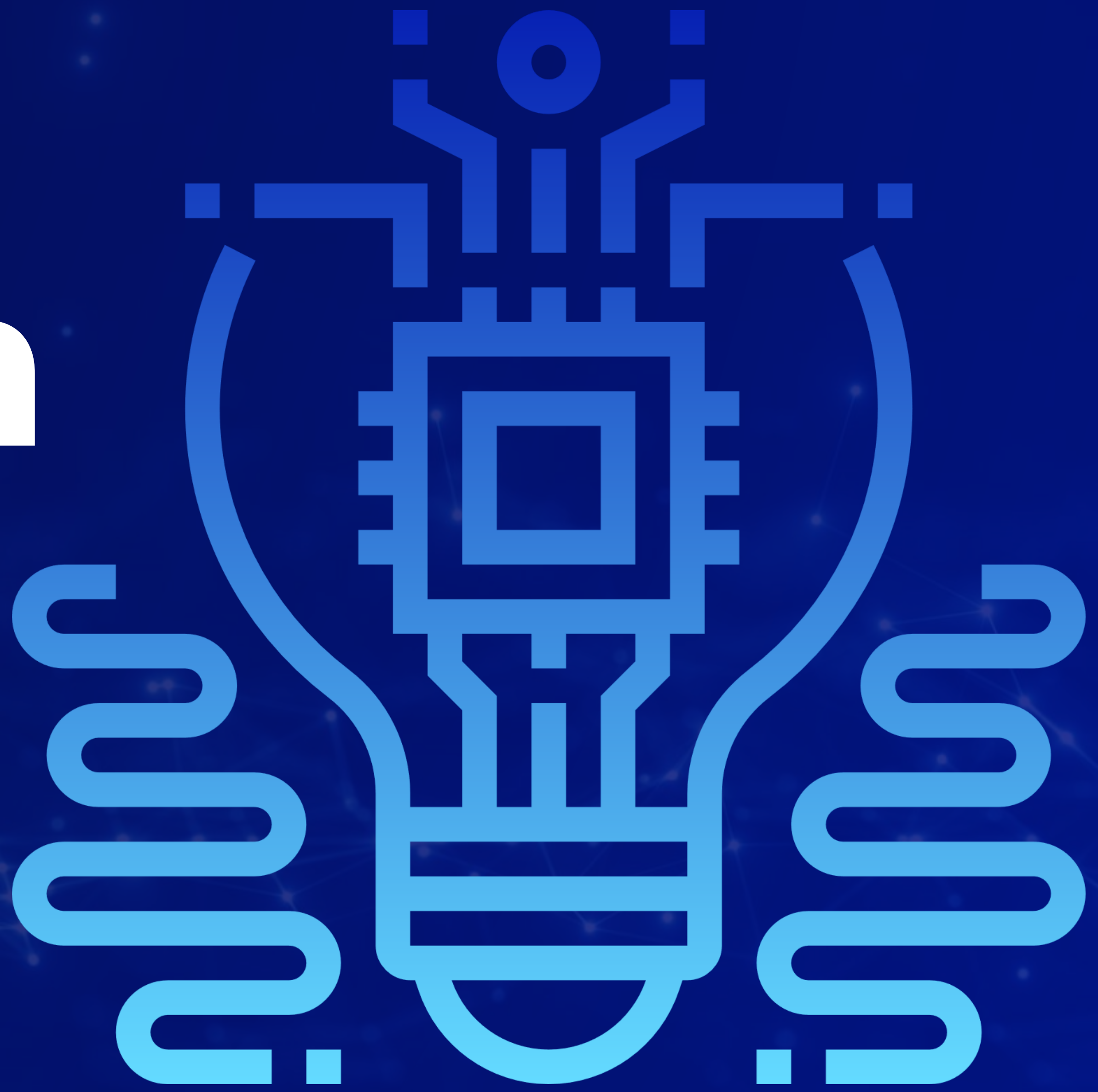
- **Full Extraction**: All data is extracted again from scratch.
- **Incremental Extraction**: Only new or modified data is extracted.

Extract Techniques

- **Manual Data Extraction**: Manual export (e.g., Excel file).
- **Database Querying**: Using SQL queries from databases.
- **File Parsing**: Reading files like CSV, JSON, or XML.
- **API Calls**: Using APIs to fetch data.
- **Event-Based Streaming**: Real-time data streaming (e.g., Kafka).
- **CDC (Change Data Capture)**: Tracking and capturing only changes.
- **Web Scraping**: Extracting data from websites.

What is[1] Transformation

This is the stage where data is cleaned, processed, and prepared before loading.



Data Cleansing

- **Data Remove Duplicates**: Eliminate duplicate records.
- **Data Filtering**: Filter out irrelevant values.
- **Handling Missing Data**: Deal with missing values.
- **Handling Invalid Values**: Correct invalid data.
- **Outlier Detection**: Identify and handle unusual values.
- **Data Type Casting**: Change data types (e.g., text → number).
- **Handling Unwanted Spaces**: Remove extra spaces.

Business Rules & Logic

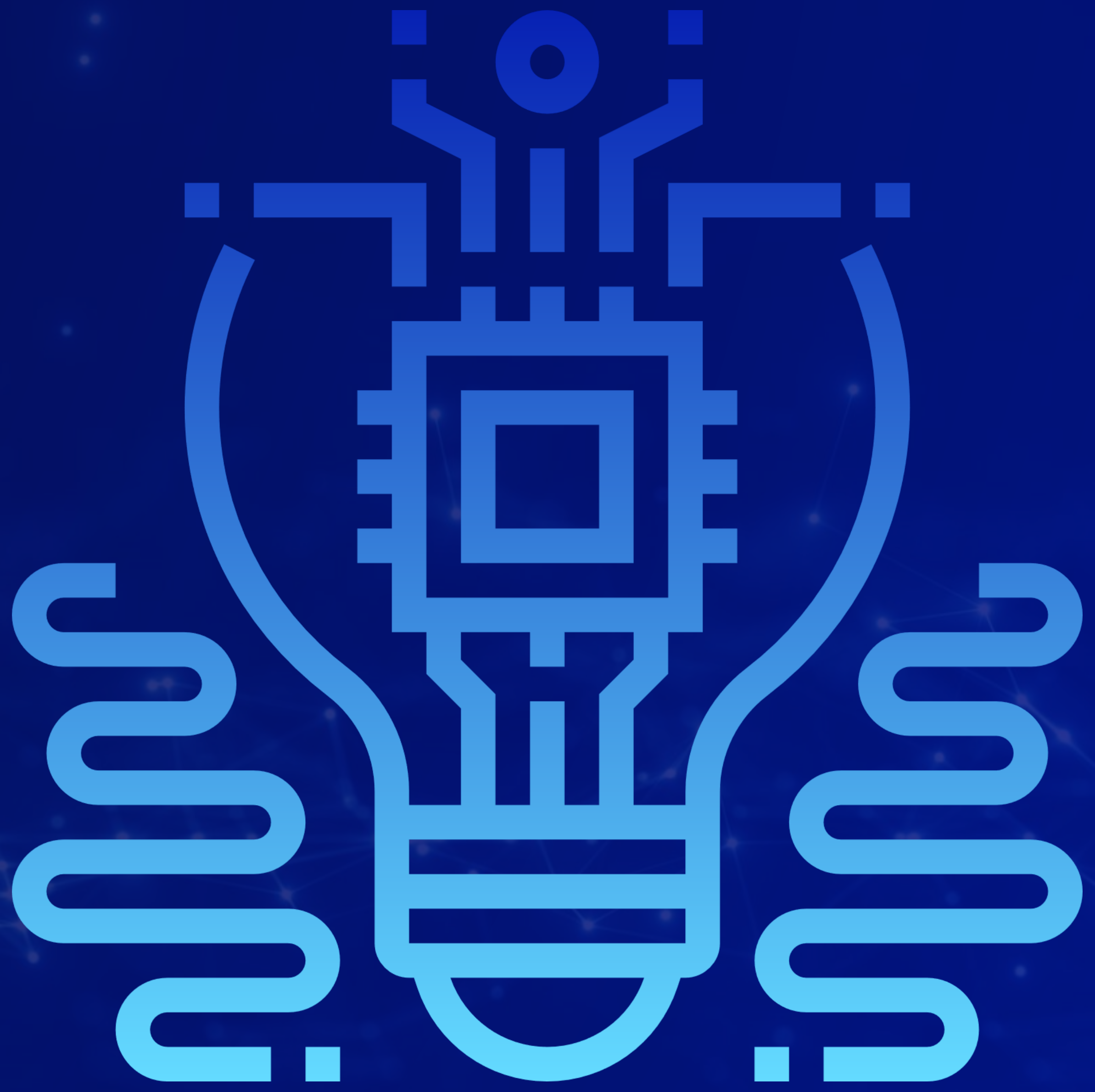
Apply business-specific rules, e.g., if sales < 1000 → VIP.

Data Normalization & Standardization

Format data into a consistent structure (e.g., unify all date formats).

Data Aggregations

Summarize data (e.g., calculate averages or totals).



Derived Columns

Create new columns from existing ones (e.g., price × quantity = total)

And Surrogate Key.

Data Integration

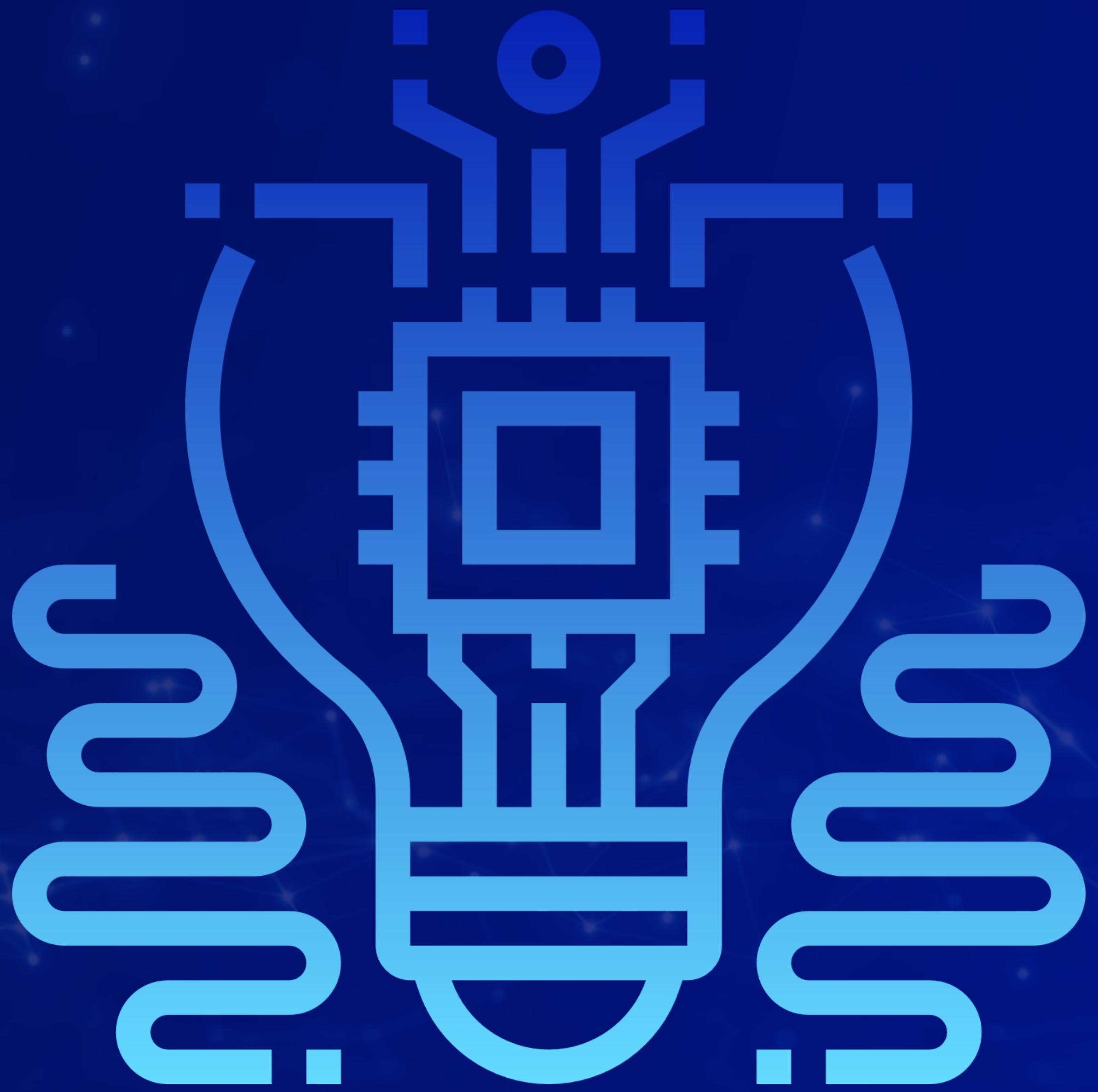
Combine data from multiple sources into one dataset.

Data Enrichment

Apply business-specific rules, e.g., if sales < 1000 → VIP.

What is[1] Load

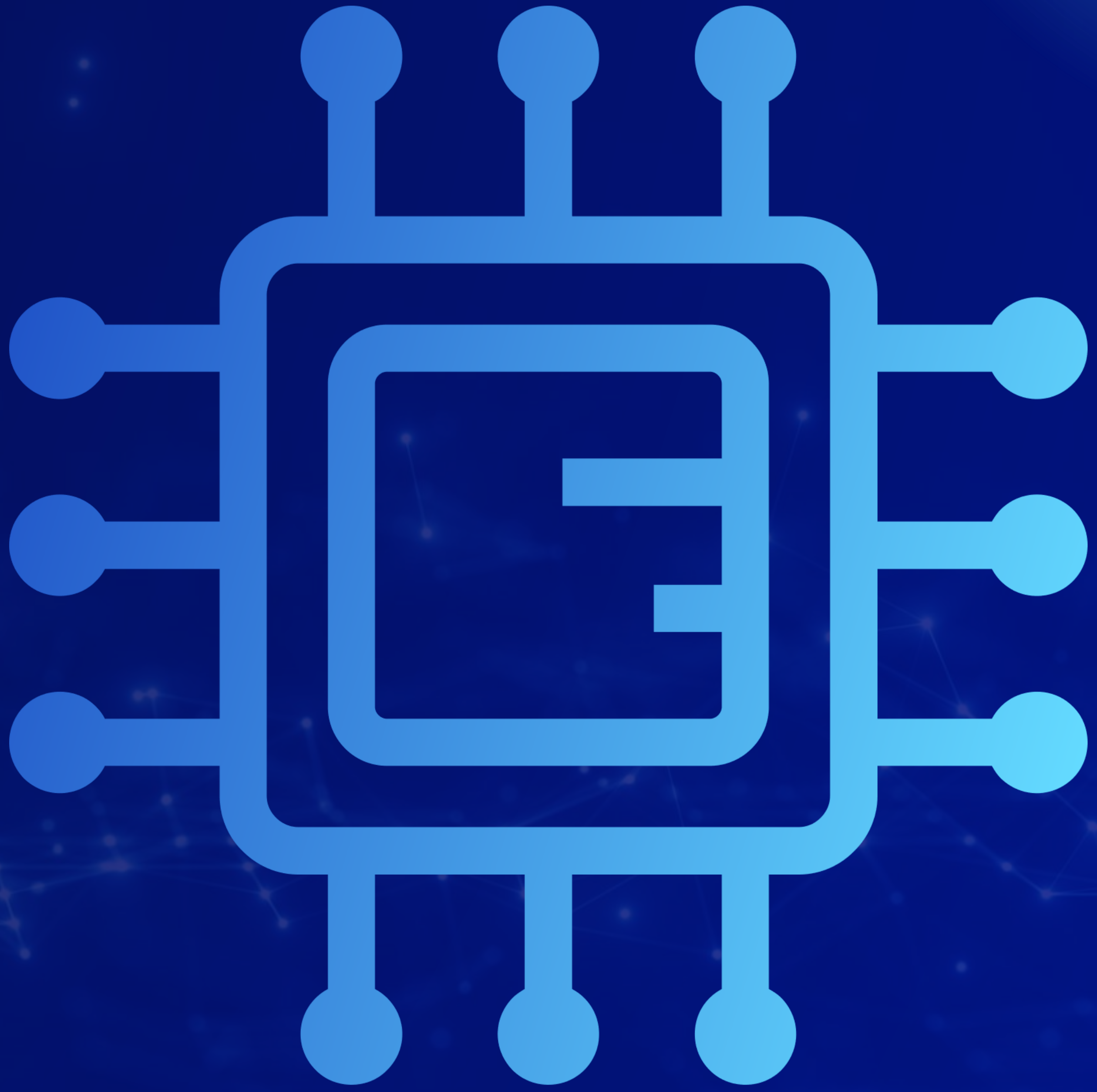
the final stage, loading the clean, transformed data into the target system (e.g., a Data Warehouse).



Processing Types

- **Batch Processing**: Load data periodically in batches.
- **Stream Processing**: Load data in real-time.

Load Methods



1.Full Load

- **Truncate & Insert**: Delete all rows, then insert new data.
- **Drop & Create & Insert**: Drop the table, recreate it, then insert new data.
- **Upsert**: Update existing records or insert new ones (sometimes used).

2. Incremental Load

- **Upsert**: Update existing data and add new records.
- **Append**: Add only new data without modifying old data.
- **Merge**: Compare new and old data to update, insert, or delete as needed.

Slowly Changing Dimensions (SCD)

- **SCD 0**: No updates (fixed).
- **SCD 1**: Overwrite old data directly.
- **SCD 2**: Keep both old and new data with change history.

Quick Summary

Stage	Function	Examples
Extraction	Pull data from sources	SQL, APIs, Files
Transformation	Clean and transform data	Remove duplicates, Add columns
Load	Load data into final system	Data Warehouse, Data Lake