DATA INTEGRATION

WHAT IS DATA INTEGRATION?

Refers to the **process of bringing together data from multiple sources** across an organization to **provide a complete, accurate, and up-to-date dataset for BI, data analysis and other applications and business processes.**

1. **Definition**
   "Data integration is the process of combining data from multiple sources into a unified, accurate, and up-to-date dataset."

- **What it means:** Gathering data from different systems and ensuring consistency.
- **Why it's important:** Provides a single source of truth for analysis and decision-making.

2. **Purpose**
   "For BI, analysis, and business processes."

- **BI (Business Intelligence):** Helps generate reports, dashboards, and insights.
- **Data Analysis:** Supports decision-making with reliable data.
- **Business Processes:** Ensures smooth operations across departments.

3. Key Processes
"It involves data replication, ingestion, and transformation to standardize data."

- **Data Replication:** Copying data from one system to another.
- **Data Ingestion:** Importing raw data from various sources.
- **Data Transformation:** Converting data into a consistent format.

4. Target Storage
"For storage in repositories like data warehouses, data lakes, or lakehouses."

- **Data Warehouse:** Structured, processed data for analysis (e.g., SQL-based).
- **Data Lake:** Raw, unstructured, and structured data for flexible use.
- **Data Lakehouse:** A mix of both, offering structure and flexibility.

KEY COMPONENTS OF DATA INTEGRATION

Visual: A simple flowchart of ETL (Extract → Transform → Load)

1. **Data Sources** – Where data comes from (databases, APIs, files, etc.)
2. **ETL (Extract, Transform, Load)** – A process for moving and converting data
3. **Data Storage** – Where integrated data is stored (data warehouses, cloud, etc.)
4. **Data Consumers** – Applications, reports, or AI that use the integrated data

FIVE APPROACHES TO DATA INTEGRATION

Data integration uses five main patterns: ETL, ELT, streaming, API, and data virtualization. It's implemented via manual coding (SQL) or automated tools.

1. ETL Pipeline

Is a traditional type of data pipeline which **converts raw data to match the target system** via three steps: **Extract, Transform and Load.** Data is transformed in a **staging area before it is loaded** into the target repository (typically a data warehouse). This allows for **fast and accurate data analysis in the target system** and is most **appropriate for small datasets** which require complex transformations.

a. Definition

"ETL is a traditional type of data pipeline which converts raw data to match the target system via three steps: extract, transform, and load."

- **ETL (Extract, Transform, Load):** A process for preparing data for storage and analysis.
- **Purpose:** Ensures data is structured and usable in the target system.
- **Traditional approach:** Used in data warehouses for structured analytics.

b. Three Key Steps

- **Extract:** Collect raw data from various sources (databases, APIs, files).
- **Transform:** Clean, filter, and format data in a **staging area** before storage.
- **Load:** Move the processed data into the final storage (e.g., data warehouse).

c. Staging Area & Transformation

"Data is transformed in a staging area before it is loaded into the target repository (typically a data warehouse)."

- **Staging Area:** A temporary space where data is cleaned and modified.
- **Why this step?** Prevents errors and ensures consistency before final storage.

d. Benefits & Best Use Case

"This allows for fast and accurate data analysis in the target system and is most appropriate for small datasets which require complex transformations."

- **Fast & Accurate Analysis:** Pre-processed data enables quick queries.
- **Best for Small Datasets:** Works well for structured data needing heavy transformation.
- **Example Use Case:** Financial reporting, customer analytics, or HR data processing.

2. ELT Pipeline

The data is loaded first and then transformed within the target system, typically a cloud-based data lake, data warehouse or data lakehouse.
This approach is **more appropriate when datasets are large and timeliness is important**, since loading is often quicker.

Operates:

- **Micro-batch ("Delta load")** - only loads the data modified since the last successful load.
- **CDC** - continually loads data as and when it changes on the source.

a. Definition

"In the more modern ELT pipeline, the data is immediately loaded and then transformed within the target system."

- **ELT (Extract, Load, Transform):** A modern data pipeline approach.
- **Key Difference from ETL:** Data is loaded first, then transformed within the target system.

b. Target Storage Systems

"Typically a cloud-based data lake, data warehouse, or data lakehouse."

- **Data Lake:** Stores raw data for flexible analysis.
- **Data Warehouse:** Stores structured data optimized for queries.
- **Data Lakehouse:** A hybrid approach combining both.

c. When is ELT Preferred?

"This approach is more appropriate when datasets are large and timeliness is important, since loading is often quicker."

- **Best for Large Datasets:** ELT handles high-volume data efficiently.
- **Faster Loading:** Immediate ingestion allows real-time or near-real-time analytics.

d. ELT Processing Methods

"ELT operates either on a micro-batch or (CDC) timescale."

- **Micro-Batch ("Delta Load")** – Loads only the modified data since the last update.
- **Change Data Capture (CDC)** – Continuously updates data as changes occur at the source.

e. Key Takeaway

- **ELT is more efficient for cloud-based systems** where processing power is scalable.
- **CDC within ELT ensures real-time updates**, while micro-batch is useful for periodic updates.

3. DATA STREAMING

It **moves data continuously in real-time from source to target**.
Modern data integration (DI) platforms **can deliver analytics-ready data** into streaming and cloud platforms, data warehouses, and data lakes.

a. Definition

"Instead of loading data into a new repository in batches, streaming data integration moves data continuously in real-time from source to target."

- **Data Streaming:** A real-time data integration method.
- **Key Difference from Batch Processing:** No waiting for scheduled loads—data flows instantly.

b. How It Works

- **Continuous Data Flow:** Data is processed as it is generated.
- **No Staging Area:** Unlike ETL, it doesn't rely on intermediate storage.

c. Where It's Used

"Modern data integration (DI) platforms can deliver analytics-ready data into streaming and cloud platforms, data warehouses, and data lakes."

- **Streaming Platforms:** Kafka, Apache Flink, Spark Streaming.
- **Cloud Platforms:** AWS, Google Cloud, Azure.
- **Data Warehouses & Lakes:** Enables real-time analytics in big data environments.

d. Key Benefits

- **Real-time insights** – Faster decision-making.
- **Handles high-velocity data** – Ideal for IoT, finance, and social media analytics.
- **Scalable & flexible** – Works well in cloud environments.

Application Integration

It allows **separate applications to work together by moving and syncing data between them.**
The most typical use case is to support operational needs such as ensuring that your HR system has the same data as your finance system. Therefore, the application integration **provides consistency between the data sets.**

1. Definition

"Application integration (API) allows separate applications to work together by moving and syncing data between them."

- **Application Integration:** Connecting different software systems to function as one.
- **Role of APIs:** APIs (Application Programming Interfaces) enable seamless data exchange.

2. Common Use Case

"The most typical use case is to support operational needs such as ensuring that your HR system has the same data as your finance system."

- **Example:** If an employee's details are updated in the HR system, the finance system should reflect those changes for payroll accuracy.
- **Why It Matters:** Prevents data inconsistencies across different departments.

3. Importance of Data Consistency

"Therefore, the application integration must provide consistency between the data sets."

- **Ensures Accuracy:** No duplicate or outdated data across systems.
- **Improves Efficiency:** Reduces manual data entry and errors.

4. Role of SaaS Automation Tools

"These various applications usually have unique APIs for giving and taking data, so SaaS application automation tools can help you create and maintain native API integrations efficiently and at scale."

- **Different Apps, Different APIs:** Each system may have its own way of sending/receiving data.
- **SaaS Automation Tools:** Platforms like Zapier, MuleSoft, and Boomi simplify integration.
- **Scalability:** Makes it easier to manage integrations as businesses grow.

Key Takeaway

- **Application integration via APIs ensures smooth data flow** across business tools.
- **Automation tools simplify API management,** reducing development effort.

4. Data Virtualization

Delivers data in real time, but only when it is requested by a user or application.
This can create a unified view of data and makes data available on demand by **virtually combining data from different systems.**

a. Definition

"Like streaming, data virtualization also delivers data in real time, but only when it is requested by a user or application."

- **Data Virtualization:** A real-time data integration method that retrieves data **only when needed** (on demand).
- **Key Difference from Streaming:** Data isn’t continuously moved; instead, it's accessed dynamically.

b. How It Works

"Creates a unified view of data and makes data available on demand by virtually combining data from different systems."

- **No Physical Copying:** Data remains in its original location.
- **Unified Data View:** Combines different data sources without actual integration.
- **On-Demand Access:** Applications and users query data when needed.

c. Where It’s Used

"Virtualization and streaming are well suited for transactional systems built for high-performance queries."

- **Transactional Systems:** Banking, e-commerce, or any system needing **fast, real-time queries**.
- **BI & Reporting:** Provides instant insights without data duplication.

d. Key Benefits

- **Faster access to real-time data** – No need to move data.
- **Reduced storage costs** – No redundant data copies.
- **Seamless integration** – Works with multiple systems without complex ETL.

THREE USE CASES

USE CASE #1: DATA INGESTION

Is the **process of moving data from multiple sources to a storage system** (e.g., data warehouse, data lake). It can occur in real-time (streaming) or batches and typically **involves cleaning and standardizing data for analytics.**

Breakdown of Data Ingestion

- **Definition:** Moves data from different sources to a storage system.

- **Processing Methods:**
  - **Streaming (Real-time):** Continuous data flow.
  - **Batch Processing:** Data is moved at scheduled intervals.

- **Key Steps:**
  - Extract data
  - Clean & standardize
  - Store in a target system

- **Use Cases:**
  - Cloud migration
  - Data warehouse/lake setup for analytics

USE CASE#2: DATA REPLICATION

It involves **copying and moving data between systems** (e.g., from an on-premises database to a cloud data warehouse). It ensures backup, synchronization, and operational availability.
Replication can **occur in bulk, batches, or real-time across data centers or the cloud.**

Breakdown of Data Replication

- **Definition:** Copies and transfers data between systems.

- **Purpose:**
  - Backup and disaster recovery
  - Synchronization for real-time access
  - High availability and performance

- **Types of Replication:**
  - **Bulk:** Copies large datasets at once
  - **Batch:** Scheduled updates
  - **Real-time:** Continuous synchronization

- **Use Cases:**
  - Cloud migration
  - Disaster recovery solutions
  - Multi-data center synchronization

USE CASE #3: DATA WAREHOUSE AUTOMATION

It involves **copying and moving data between systems** (e.g., from an on-premises database to a cloud data warehouse). It ensures backup, synchronization, and operational availability. Replication can **occur in bulk, batches, or real-time across data centers or the cloud.**

Breakdown of Data Warehouse Automation

- **Definition:** Automates the creation and management of a data warehouse.

- **Key Processes:**
  - **Data modeling** – Structuring data for storage & use
  - **Real-time ingestion** – Moving data continuously
  - **Data marts** – Organizing data for analysis
  - **Governance** – Ensuring data security & compliance

- **Benefits:**
  - Faster data availability
  - Reduced manual effort
  - Improved data quality & compliance