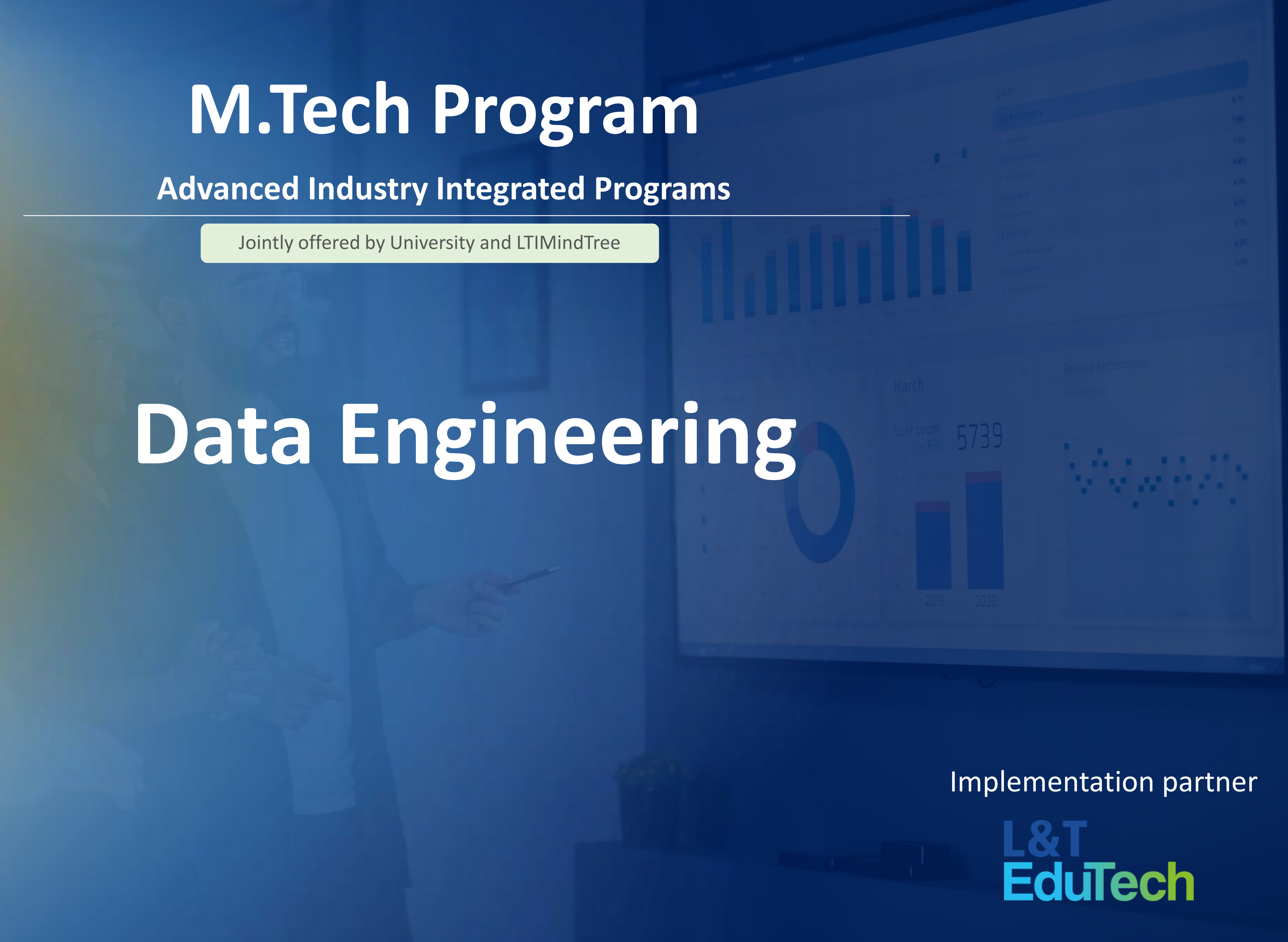
M.Tech Program

Advanced Industry Integrated Programs

Jointly offered by University and LTIMindTree

Data Engineering



Knowledge partner

LTIMindtree

Implementation partner

L&T EduTech

Course Objective

- Recognize data types and structures.
- Grasp big data fundamentals and analytics.
- Master data ingestion processes and tools.
- Understand exploratory data analysis techniques.
- Learn storage methods and data flow.

Modules

- Data Types & Formats
- Data Ingestion techniques
- Data Profiling & Visual Representation via various tools (Pandas)
- Storage and retrieval methods
- Data Lineage Analysis

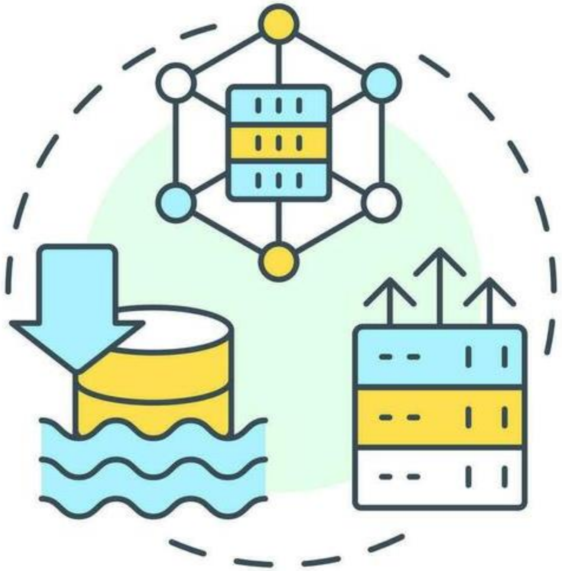
Data Ingestion Techniques

Streaming

Streaming

Data Ingestion

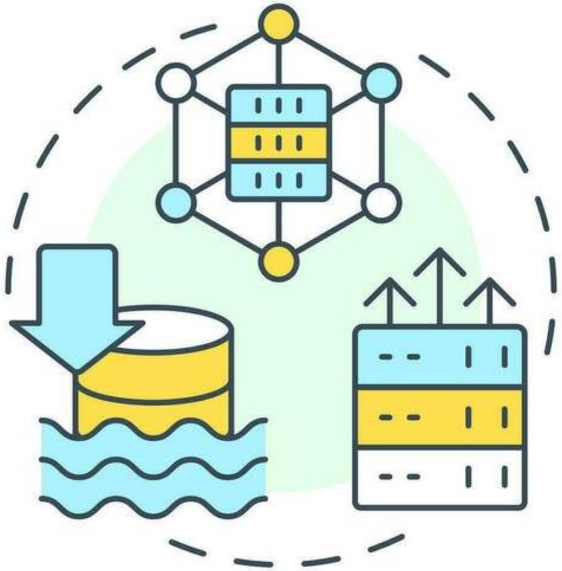
- **Data Ingestion:** Moving and replicating data from various sources to destinations like cloud data lakes or warehouses.
- **Data Sources:** Databases, files, streaming data, change data capture (CDC), applications, IoT devices, machine logs.
- **Destination:** Data is ingested into the landing or raw zone.
- **Purpose:** Prepares data for business intelligence and downstream transactions.
- **Outcome:** Enables advanced analytics readiness.



Streaming

Data Ingestion

- **Data Ingestion:** Moving and replicating data from various sources to destinations like cloud data lakes or warehouses.
- **Data Sources:** Databases, files, streaming data, change data capture (CDC), applications, IoT devices, machine logs.
- **Destination:** Data is ingested into the landing or raw zone.
- **Purpose:** Prepares data for business intelligence and downstream transactions.
- **Outcome:** Enables advanced analytics readiness.



Batching

Batching

Data Integration

- **Data Integration:** Combines data from different sources into a unified view.
- **Role in Data Management:** Ensures data is consistent, accurate, and accessible across platforms.
- **Complexity Comparison:** Data integration is more complex than data ingestion.
- **Sources for Intégration:** APIs, applications, files, etc.
- **Objective:** Provides a comprehensive and coherent view of data, regardless of source.

Batching

Key Differences Between Data Integration and Data Ingestion

Aspect	Data Ingestion	Data Integration
Purpose	Importing or ingesting data for immediate use or storage.	Providing a consistent view of data from multiple sources.
	Collecting raw data and making it available for further processing and analysis.	Integrating data into a single, coherent system for better understanding of operations, customers, and market trends.
Process	Can use batch or streaming methods	Involves ETL (Extract, Transform, Load) procedure
	Batch ingestion collects and processes data at periodic intervals.	Extract: Retrieving data from source systems.
	Streaming ingestion processes data almost instantaneously as it arrives.	Transform: Cleaning and converting data into a suitable format.
		Load: Transferring transformed data into a target data warehouse or database.

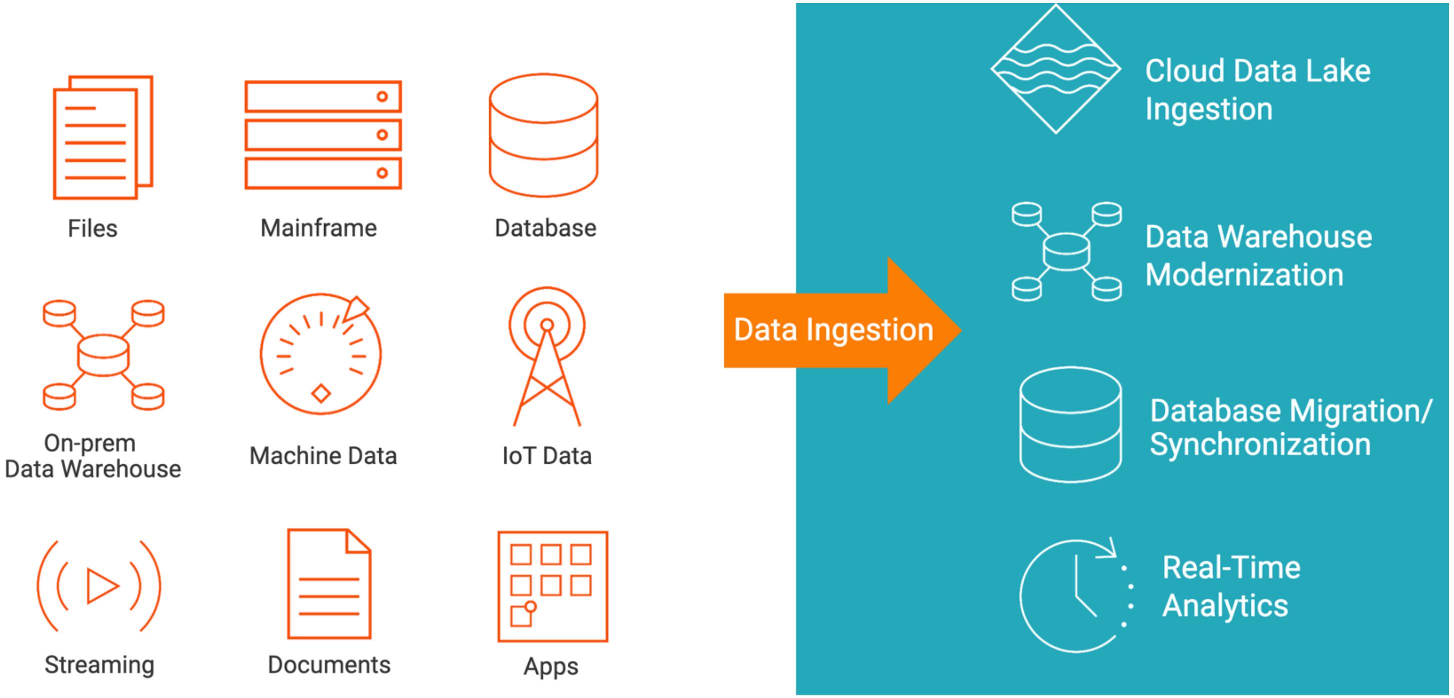
Batching

Key Differences Between Data Integration and Data Ingestion

Aspect	Data Ingestion	Data Integration
Scope	Initial stage in the overall data pipeline.	Embraces a broader perspective including ingestion, harmonizing, and consolidating data.
	Focused on collection and immediate processing or storage of incoming data.	Ongoing process ensuring all integrated data stays updated and aligned with source systems.
Common Use Cases	Real-Time Analytics in E-commerce: Personalized recommendations, dynamic pricing.	Customer Relationship Management (CRM): Consolidated view of customer interactions.
	IoT Devices and Sensor Data: Predictive maintenance, monitoring equipment health.	Sales and Marketing Analytics: Comprehensive insights into customer journey and campaign effectiveness.

Batching

Data Ingestion from Various Sources to Cloud Modernization



Batching

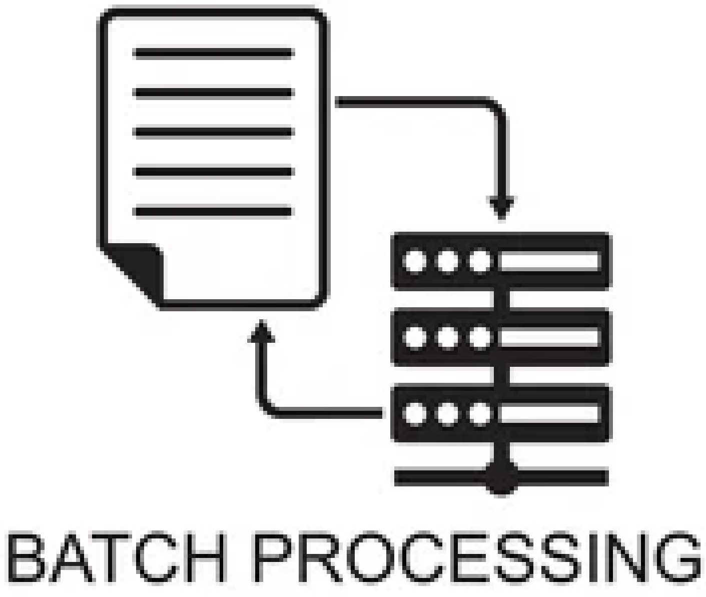
Data Ingestion Types

- **Real-Time Ingestion:**
  - Processes and stores data as soon as it is generated.
  - Minimizes delay between data generation and processing.
  - Example: Monitoring power grid data for preventive maintenance.

Batching

Data Ingestion Types

- **Batch Ingestion:**
    - Collects and moves data in scheduled batches or triggered by events.
    - Suitable for large data volumes.
    - Techniques include file-based ingestion (e.g., CSV, JSON, XML).



Batching

Data Ingestion Types

- **Lambda Architecture:**
  - Combines batch and real-time processing layers.
  - Batch Layer: Processes large volumes of data using frameworks like Apache Hadoop.
  - Speed Layer: Handles real-time data processing with low-latency results using technologies like Apache Storm or Spark Streaming.
  - Serving Layer: Provides a unified view of data from both batch and speed layers.

Batching

Data Ingestion Types

- **Micro-Batching:**
  - Processes data in small, fixed-size batches at regular intervals (milliseconds to seconds).
  - Bridges the gap between traditional batch processing and real-time streaming.

Batching

Data Ingestion Framework

- System/platform for collecting, importing, and processing large volumes of data from various sources into centralized storage or processing environment.

Batching

Data Ingestion Framework

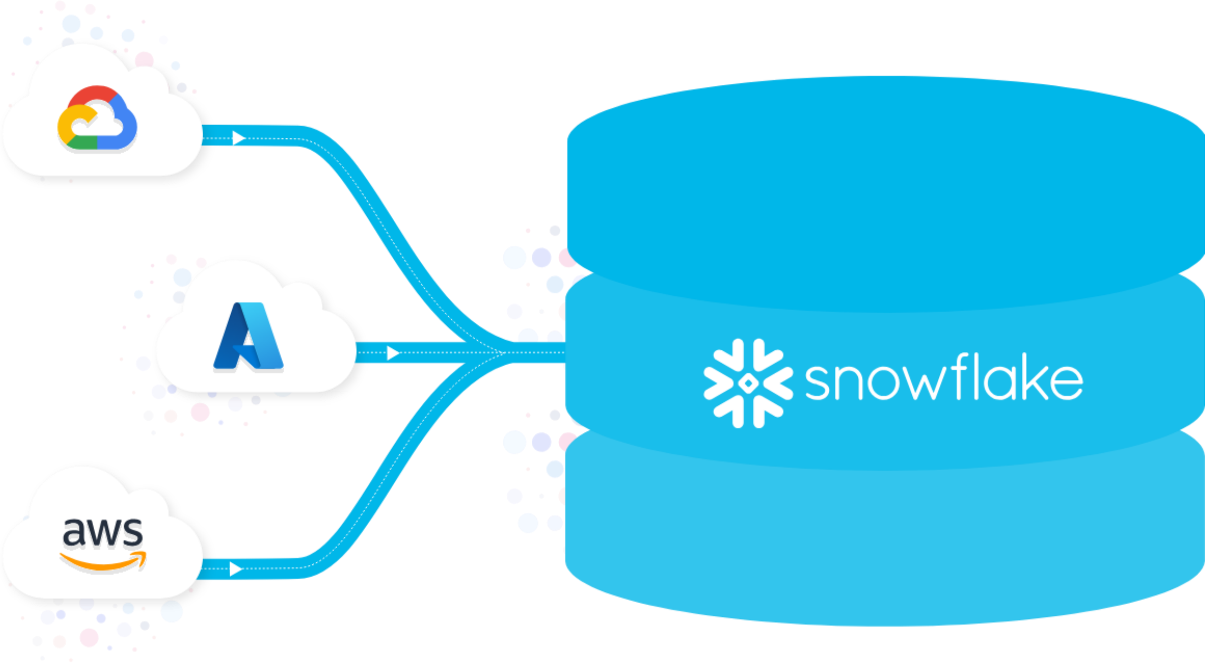
Key Components:

- **Data Sources:** Databases, files, streams, APIs, sensors, etc.
- **Data Connectors:** Adapters to interface with different data sources.
- **Data Transport:** Supports batch processing, real-time streaming, or both.
- **Error Handling and Monitoring:** Mechanisms for handling errors and ensuring data integrity.
- **Scalability and Performance:** Can handle large data volumes and scale horizontally.
- **Security:** Features for authentication, authorization, encryption, and compliance with data protection regulations.

Batching

Data Ingestion by Platform

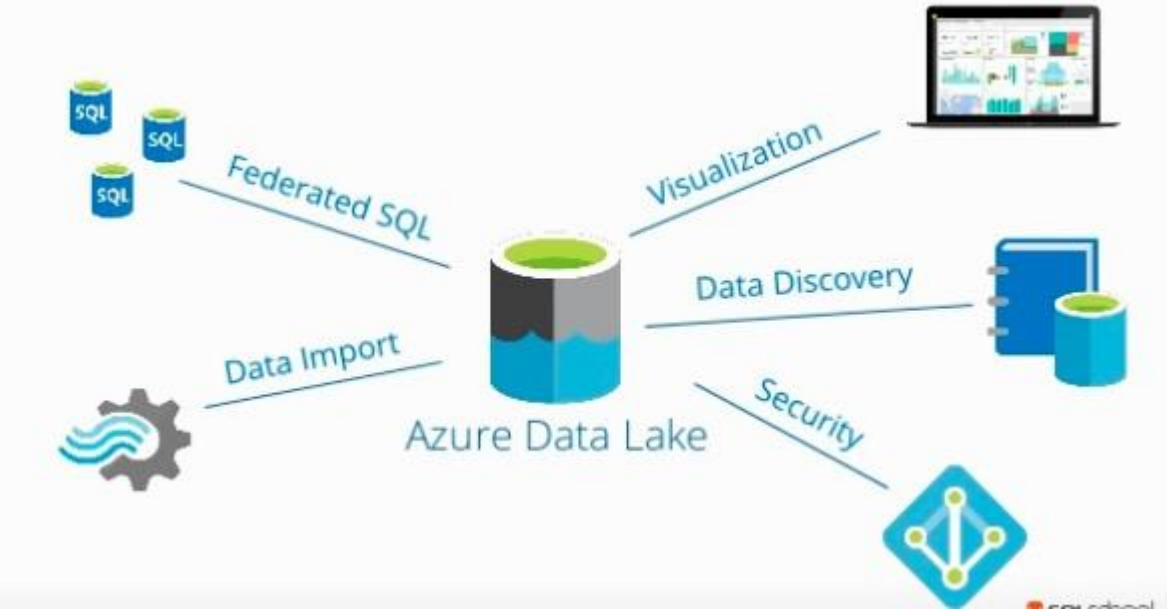
- **Snowflake:**
  - Ingest and replicate large volumes of data at scale.
  - Sources: Application data, mainframes, databases, data warehouses, machine data, IoT, streaming data, logs, files.
  - Apply simple data transformations during ingestion for analytics readiness.



Batching

Data Ingestion by Platform

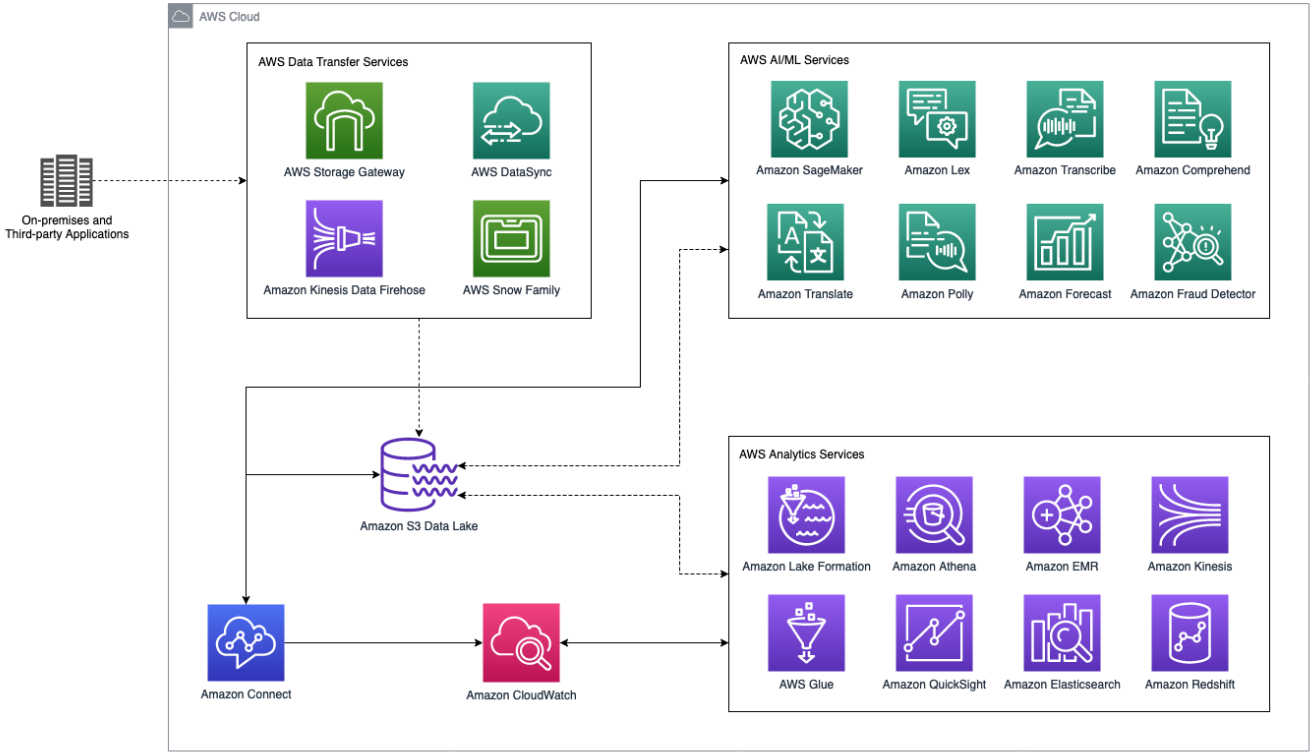
- **Microsoft Azure:**
  - Ingest and replicate data at scale from on-premises databases.
  - Automatically capture changed data into Azure Synapse for synchronization and replication.



Batching

Data Ingestion by Platform

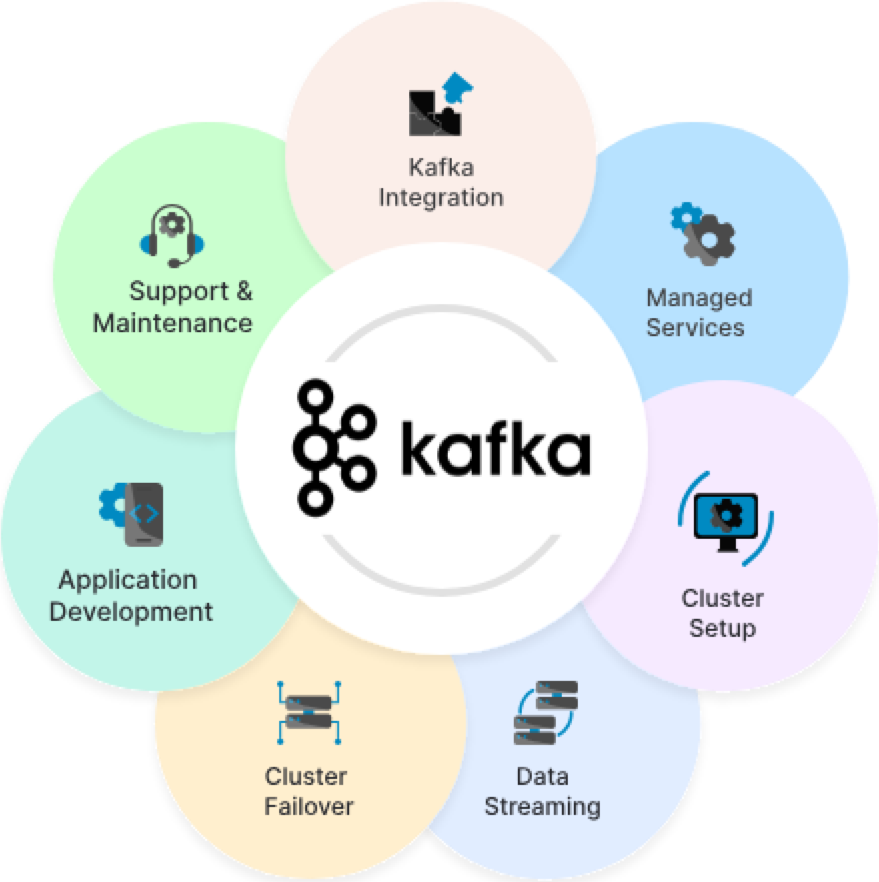
- **Amazon Web Services(AWS):**
  - Accelerate analytics, AI, and machine learning by moving data from SaaS or on premises sources to Amazon S3 or AWS Redshift.



Batching

Data Ingestion by Platform

- **Kafka:**
  - Ingest streaming and IoT data with low latency for real-time analytics and streaming CDC use cases.



Batching

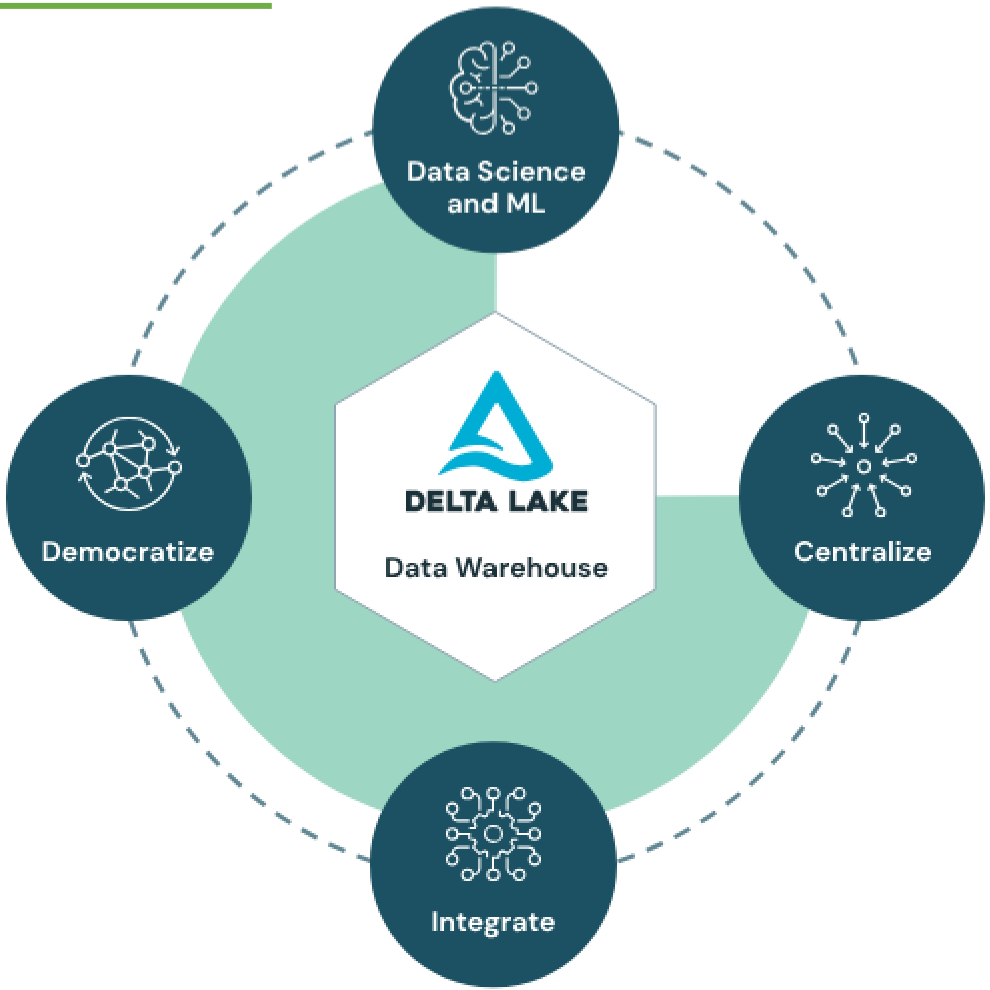
Data Ingestion by Platform

- **Google BigQuery:**
  - **Ingest and replicate large data volumes** from on-premises sources to Google Cloud Storage and BigQuery.
  - **Sources:** Oracle, SQL Server, MySQL, Teradata, Netezza, DB2.
  - Supports schema drift to detect and replicate schema changes.

Batching

Data Ingestion by Platform

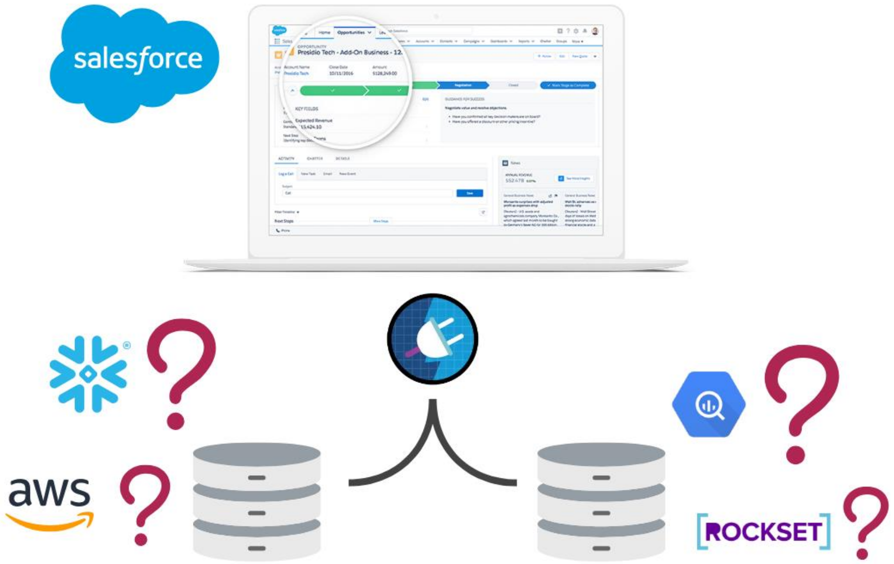
- **Databricks Delta Lake:**
  - Rapidly load data from databases to Databricks Delta Lake.
  - Move massive amounts of data in minutes with a user-friendly interface.



Batching

Data Ingestion by Platform

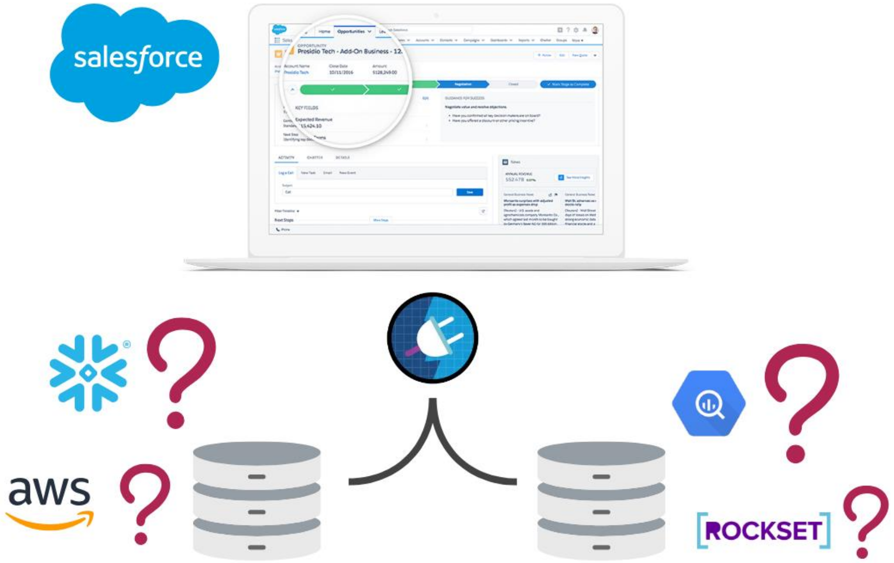
- **Salesforce:**
  - Ingest data from Salesforce to multiple destinations including Snowflake, AWS Redshift, Microsoft Azure Synapse, Google BigQuery, AWS S3, ADLS Gen2, Google Cloud Storage, Databricks, and Kafka.
  - Use Informatica Cloud Mass Ingestion for application synchronization.



Batching

Data Ingestion by Platform

- **Salesforce:**
  - Ingest data from Salesforce to multiple destinations including Snowflake, AWS Redshift, Microsoft Azure Synapse, Google BigQuery, AWS S3, ADLS Gen2, Google Cloud Storage, Databricks, and Kafka.
  - Use Informatica Cloud Mass Ingestion for application synchronization.



Batching

Essential Data Ingestion Capabilities

- **Unified Experience for Data Ingestion:**
  - Single solution to ingest data from multiple sources.
  - Apply simple transformations (e.g., filtering bad records) before ingestion.
- **Handle Unstructured Data and Schema Drift:**
  - Parse unstructured data for downstream use.
  - Intelligent handling of schema drift and automatic propagation of changes.

Batching

Essential Data Ingestion Capabilities

- **Versatile Out-of-the-Box Connectivity:**
  - Connect to various sources: files, databases, mainframes, IoT, applications, and streaming sources.
  - Persist enriched data to cloud data lakes, warehouses, and messaging systems.
- **High Performance:**
  - Ensure continuous data availability with no downtime.
  - Real-time ingestion with Kappa architecture or batch processing with Lambda architecture.
  - High availability and recovery from ingestion job failures with exactly one delivery guarantee.

Batching

Essential Data Ingestion Capabilities

- **Wizard-Based Data Ingestion:**
  - Efficient, no-code ingestion with a wizard-based tool.
  - Ingest data into a cloud data warehouse with CDC capability for current, consistent data.
- **Real-Time Data Ingestion:**
  - Accelerate real-time log, CDC, and clickstream data ingestion into Kafka, Azure Event Hub, Amazon Kinesis, and Google Cloud Pub/Sub.
  - Enable real-time analytics.

Batching

Essential Data Ingestion Capabilities

- **Cost-Efficient:**
  - Save money by automating costly and time-consuming processes.
  - Reduce costs by avoiding infrastructure and skilled technical resource expenses.

Batching

Data Ingestion Benefits

- **Efficient Data Collection:** Enables efficient collection of raw data from diverse sources.
- **Data Centralization:** Facilitates centralization into a single repository for easier management and consumption.
- **Real-time Insights:** Supports real-time ingestion for timely insights and faster data-driven decisions.
- **Integration with Analytics Tools:** Allows seamless integration with analytics and visualization tools for advanced analytics, reporting, and business intelligence.

Batching

Data Ingestion Benefits

- **Operational Efficiency:** Automates data ingestion, reducing manual effort and improving operational efficiency, freeing resources for strategic tasks.

Different tools and solutions available for Data Ingestion

Different tools and solutions available for data ingestion

Data Ingestion Challenges

Sluggish Processes

The Cost Factor

Increased Complexity

The Risk to Data Security

Unreliability

Different tools and solutions available for data ingestion

Data Ingestion Challenges

Manual Processes

- Traditional methods are inefficient for diverse data.

Need for Automation

- Advanced tools needed for faster ingestion.

Cost Factor

- High costs for infrastructure and skilled team.

Data Security Risk

- Ensuring security during multiple ingestion stages.

Unreliability of Bad Data

- Difficult to maintain clean, accurate data.

Different tools and solutions available for data ingestion

Types of Data Ingestion Tools

Hand Coding

- Offers greatest control but requires coding skills and extensive time for modifications.

Single-purpose Tools

- Drag-and-drop with pre-built connectors, limited scalability, and difficult team collaboration.

Data Integration Platforms

- Comprehensive features for all steps, but slow adaptation and requires specialized developers.

DataOps Approach

- Uses agile methodologies, automates processes, and lets engineers focus on business data needs.

Different tools and solutions available for data ingestion

Best Data Ingestion Tools

Airbyte: Open-source tool with 120+ connectors, providing raw and normalized data for analysis.

Hevo: Fully automated, no-code platform with 150+ integrations across databases, SaaS, and streaming services.

Amazon Kinesis: Cloud-based service for scalable data ingestion and processing from numerous distributed sources.

Apache Flume: Handles large data volumes, focuses on ingestion into Hadoop Distributed File System (HDFS).

Apache Gobblin: Ingests large data volumes into HDFS, includes ETL, data quality management, and error correction.

Apache Kafka: Ideal for high-volume real-time streaming, known for high throughput and low latency.

Different tools and solutions available for data ingestion

Best Data Ingestion Tools

Apache NiFi: Automates data flow between systems, providing high throughput, low latency, and robust loss tolerance.

Drop base: Transforms offline data into live databases, supports real-time collaboration on data projects..

Integrate.io: Drag-and-drop tool with 100+ connectors, offering data ingestion and transformation capabilities.

Matillion: ETL tool with 70+ connectors, ideal for SMBs migrating data to cloud-based databases. Apache **Gobblin**: Ingests large data volumes into HDFS, includes ETL, data quality management, and error correction.

Pentaho Data Engineering involves using the Pentaho platform for data integration, data analysis, and data visualization tasks.

Different tools and solutions available for data ingestion

Pentaho

- **Pentaho Reporting:** A collection of tools for creating relational and analytical reports.
- **Purpose:** Transforms data into meaningful information.
- **Output Formats**: Can generate reports in: HTML, Excel, PDF, Text, CSV, XML

Different tools and solutions available for data ingestion

Pentaho

- **Pentaho Reporting:** A collection of tools for creating relational and analytical reports.
- **Purpose:** Transforms data into meaningful information.
- **Output Formats**: Can generate reports in: HTML, Excel, PDF, Text, CSV, XML

Different tools and solutions available for data ingestion

Features of Pentaho

Report Designer	• Used for creating pixel-perfect reports.
Metadata Editor	• Adds user-friendly metadata domains to a data source.
Report Designer and Design Studio	• Used for fine-tuning reports and ad-hoc reporting.
Pentaho User Console Web Interface	• Facilitates easy management of reports and analysis views.
Ad-Hoc Reporting Interface	• Provides a step-by-step wizard for designing simple reports with output formats such as PDF, RTF, HTML, and XLS.
Complex Scheduling Sub-System	• Allows users to schedule and execute reports at specified intervals.
Mailing	• Enables users to email published reports to others.
Connectivity	• Ensures seamless connectivity between reporting tools and the BI server, allowing direct publishing of content to the BI server.

Different tools and solutions available for data ingestion

Install of Pentaho

Step 1: Download Pentaho

1. Visit the Pentaho Website: Go to the official Pentaho website.
2. Choose the Version: Select the version of Pentaho that suits your needs.
3. Download: Download the installer for your operating system (Windows, macOS, Linux).

Different tools and solutions available for data ingestion

Install of Pentaho

Step 2: Install Java

1. Check Java Installation:
    1. Open a terminal or command prompt.
    2. Type java -version to check if Java is installed.
2. Install Java:
    1. If Java is not installed, download and install the latest version of JDK from Oracle's website.

Different tools and solutions available for data ingestion

Install of Pentaho

Step 3: Extract Pentaho Files

1. Locate the Downloaded File:
    1. Find the downloaded Pentaho zip file.
2. Extract the Files:
    1. Right-click the zip file and select "Extract Here" (or use a similar option depending on your OS).

Different tools and solutions available for data ingestion

Install of Pentaho

Step 4: Set Up Environment Variables (Optional but Recommended)

1. Configure JAVA_HOME:
   1. Set the JAVA_HOME environment variable to point to your Java installation directory.
2. Configure PENTAHO_HOME:
   1. Set the PENTAHO_HOME environment variable to point to your extracted Pentaho directory.

Different tools and solutions available for data ingestion

Install of Pentaho

Step 5: Start Pentaho Server

1. Navigate to the Pentaho Directory:
    1. Open a terminal or command prompt and navigate to the biserver-ce directory within the extracted Pentaho files.
2. Start the Server:Run the startup script
    1. :On Windows: start-pentaho.batOn macOS/Linux: ./start-pentaho.sh
3. Wait for Startup:
    1. Wait for the server to start. This may take a few minutes.

Different tools and solutions available for data ingestion

Install of Pentaho

Step 6: Access Pentaho User Console

1. Open a Web Browser: Open your preferred web browser.
2. Enter the URL:Type http://localhost:8080 and press Enter.
3. Login: Use the default credentials to log in (usually admin / password).

Different tools and solutions available for data ingestion

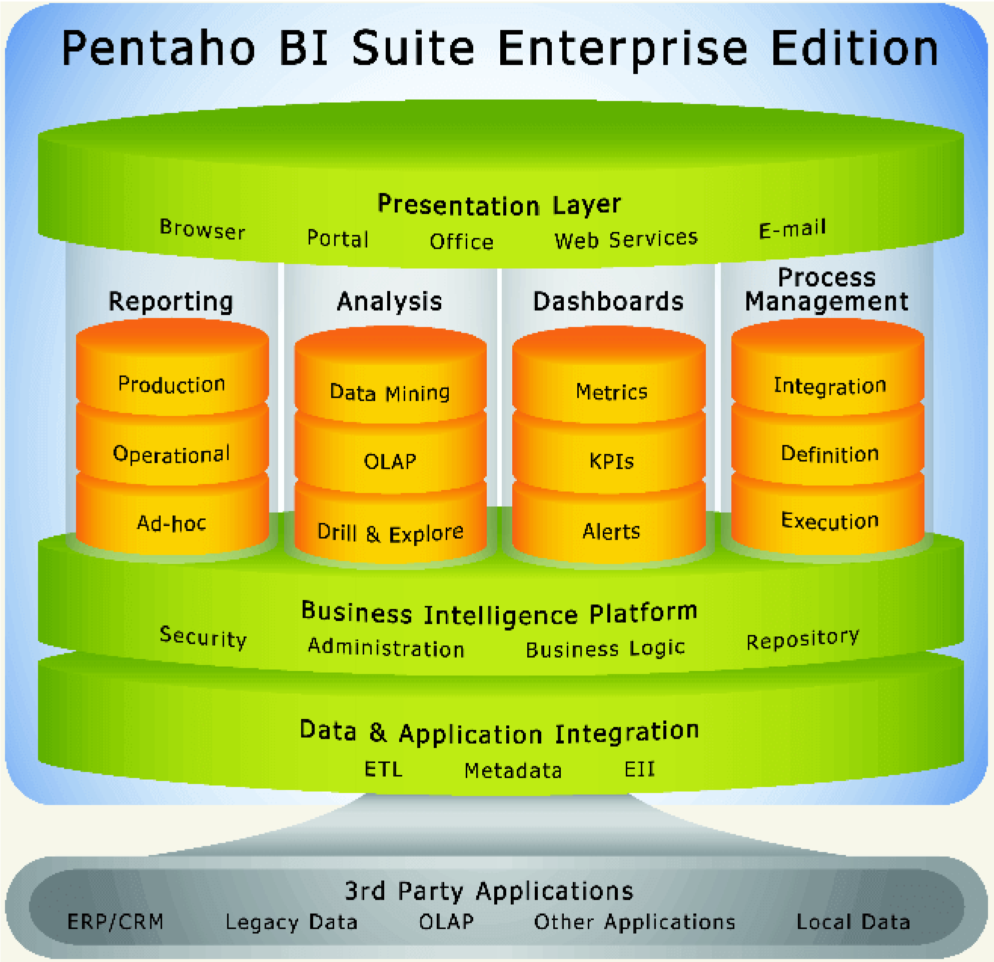
Install of Pentaho

Step 7: Verify Installation

1. Check the Dashboard: Ensure the dashboard and other features are accessible.
2. Run a Sample Report: Test the installation by running a sample report.

Different tools and solutions available for data ingestion

Pentaho BI suite



Different tools and solutions available for data ingestion

Pentaho BI suite Components

- **Pentaho Reporting**
- Based on: JFreeReport project.
- Purpose: Meets business reporting needs.
- Features:
    - Scheduled and on-demand report publishing.
    - Supports formats: XLS, PDF, TXT, and HTML.

Different tools and solutions available for data ingestion

Pentaho BI suite Components

Pentaho Reporting

- Based on: JFreeReport project.
- Purpose: Meets business reporting needs.
- Features:
  - Scheduled and on-demand report publishing.
  - Supports formats: XLS, PDF, TXT, and HTML.

Different tools and solutions available for data ingestion

Pentaho BI suite Components

Analysis

- Features:
  - Pivot table view.
  - Enhanced GUI with Flash or SVG.
  - Integrated dashboard widgets.
  - Portal and workflow integration.
- Pentaho Spreadsheet Services: Allows browsing, pivoting, and charting within MS Excel.

Different tools and solutions available for data ingestion

Pentaho BI suite Components

Dashboards

- Content: Includes Reporting and Analysis.
- Self-Service Dashboard Designer:
  - Built-in dashboard templates and layouts.
  - Enables business users to create personalized dashboards with minimal training..

Different tools and solutions available for data ingestion

Pentaho BI suite Components

Data Mining

- **Purpose:** Discovers hidden patterns and future performance indicators.
- **Algorithms:** Comprehensive set from the Weka project, including clustering, decision trees, random forests, principal component analysis, neural networks.
- **Features:** Graphical data visualization. Programmatic interaction. Multiple data sources for reports, analysis, and processes.

Different tools and solutions available for data ingestion

Pentaho BI suite Components

Data Mining

- **Purpose:** Discovers hidden patterns and future performance indicators.
- **Algorithms:** Comprehensive set from the Weka project, including clustering, decision trees, random forests, principal component analysis, neural networks.
- **Features:** Graphical data visualization. Programmatic interaction. Multiple data sources for reports, analysis, and processes.

Different tools and solutions available for data ingestion

Advantages of Pentaho

- Intuitive and User-Friendly: Easy to use with basic concepts, making it accessible for new users.
- Comprehensive BI Capabilities: Includes reporting, dashboards, interactive analysis, data integration, and data mining.
- Versatile Data Retrieval: Provides tools to retrieve data from multiple sources in a single package.
- User Interface: Features a user-friendly interface that enhances the user experience.
- Editions: Available in both Community (with many contributors) and Enterprise editions.
- Hadoop Compatibility: Can run on a Hadoop cluster, supporting big data analytics.Reusable JavaScript Code: JavaScript code written in step components can be reused across other components, enhancing efficiency.

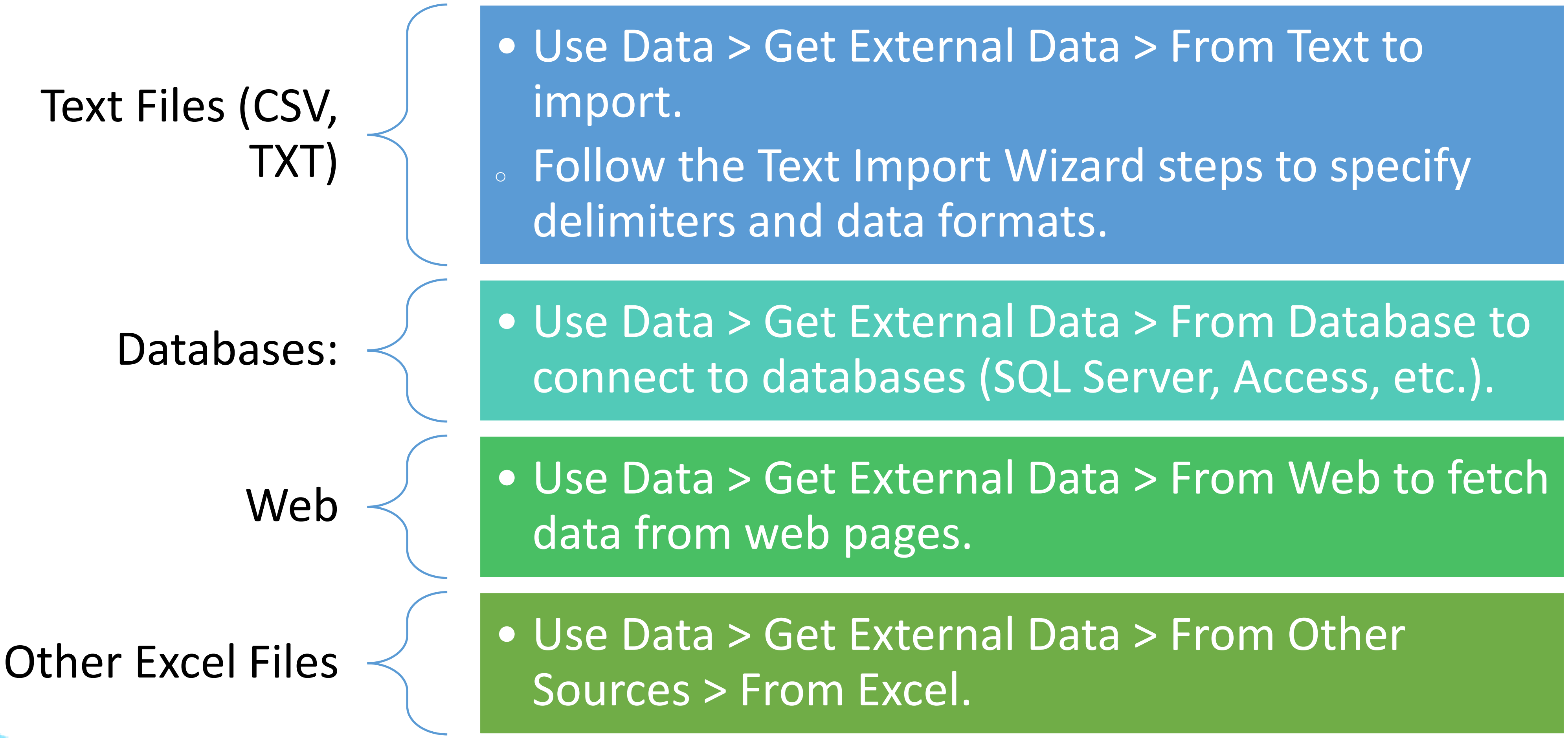
Different tools and solutions available for data ingestion

Disadvantages of Pentaho

- Interface Design: The interface design can be weak, and there is no unified interface for all components.
- Slow Tool Evolution: The tool evolves much slower compared to other BI tools.
- Limited Components: Pentaho Business Analytics offers a limited number of components.
- Poor Community Support: Limited community support means that if a component isn't working, you may need to wait for the next version for a fix.

Different tools and solutions available for data ingestion

Retrieving Data : Data can be imported into Excel from various sources, including:



Text Files (CSV, TXT)

- Use Data > Get External Data > From Text to import.
  - Follow the Text Import Wizard steps to specify delimiters and data formats.

Databases:

- Use Data > Get External Data > From Database to connect to databases (SQL Server, Access, etc.).

Web

- Use Data > Get External Data > From Web to fetch data from web pages.

Other Excel Files

- Use Data > Get External Data > From Other Sources > From Excel.

Different tools and solutions available for data ingestion

Cleaning Data

- **Removing Duplicates :**
  - Select the data range.
  - Go to Data > Remove Duplicates.
  - Select columns to check for duplicates and confirm.
- **Handling Missing Data :**
  - Removing Missing Data:
    - Use Filter to identify and delete rows with missing values.
  - Filling Missing Data:
    - Use =IF(ISBLANK(A1), "Value", A1) to fill in blanks with a specific value.
  - Imputing Missing Data:
    - Use statistical methods (mean, median, mode) to fill gaps. Example: =IF(ISBLANK(A1), AVERAGE(A:A), A1).

Different tools and solutions available for data ingestion

Cleaning Data

- **Managing Irrelevant Data :**
  - Filtering:
    - Use Data > Filter to display only relevant rows..
  - Deleting Columns/Rows:
    - Select columns/rows and right-click to delete irrelevant data.

Summary

Data Ingestion Techniques

Data Ingestion

The process of importing, transferring, loading, and processing data for later use or storage in a database.

Data Integration

Combining data from different sources to provide a unified view, ensuring consistency and accessibility.

Data Ingestion Challenges

Include handling diverse data formats, ensuring data quality, dealing with high data velocity, and managing large volumes.

Types of Data Ingestion Tools

Tools include batch processing tools (e.g., Apache Nifi) and real-time processing tools (e.g., Apache Kafka).

Data Ingestion Techniques

Benefits of Data Ingestion

Facilitates real-time analytics, improves data accessibility, enhances decision-making, and supports seamless integration with data warehouses and lakes.

Data Ingestion Framework

A set of tools and methodologies designed to streamline the process of ingesting data from various sources into a central repository.

Batch Data Ingestion

Involves collecting and processing data in chunks or batches at scheduled intervals, suitable for non-time-sensitive data.

Real-Time Data Ingestion

Captures and processes data immediately as it is created or received, enabling timely analytics and responses.

Data Quality Management

Ensures that ingested data is accurate, consistent, and reliable, which is crucial for effective data analysis and decision-making.

Knowledge Check

Data Ingestion Techniques

Data Ingestion: Data Integration

Q1 : Which of the following best describes data ingestion in the context of data integration?

a) The process of cleaning and transforming data for analysis

b) The process of combining data from different sources into a single, unified view

c) The process of capturing and importing data for immediate use or storage in a database

d) The process of analyzing data to extract meaningful insights

Data Ingestion Techniques

Data Ingestion: Data Integration

Q2 : What is the primary objective of data integration?

a) To improve data quality by cleaning and preprocessing raw data

b) To store large volumes of data in a centralized location

c) To combine data from multiple sources into a cohesive and unified view

d) To ensure the security and privacy of data

Data Ingestion Techniques

Data Ingestion Challenges

Q3 : Which of the following is a common challenge in the data ingestion process?

a) Ensuring data is stored in a centralized location

b) Maintaining data quality and consistency across different sources

c) To combine data from multiple sources into a cohesive and unified view

d) To ensure the security and privacy of data

Data Ingestion Techniques

Data Ingestion Challenges

Q4 : Which of the following is a technical challenge associated with real-time data ingestion?

a) Ensuring data privacy and compliance with regulations

b) Handling high velocity and volume of data streams

c) Merging data from relational databases

d) Creating data visualizations for business users

Data Ingestion Techniques

Types of Data Ingestion Tools

Q5 :Which data ingestion tool is best suited for real-time data ingestion?

a) Apache Spark

b) Apache Flume

c) Talend

d) IBM InfoSphere DataStage

Data Ingestion Techniques

Benefits of Data Ingestion

Q6 : How does real-time data ingestion benefit organizations?

a) It delays the processing of data to ensure thorough analysis.

b) It allows immediate processing and analysis of data as it is generated.

c) It reduces the overall volume of data collected.

d) It makes data less accessible to stakeholders.

Data Ingestion Techniques

Benefits of Data Ingestion

Q7 : Which of the following is a primary benefit of data ingestion in an organization?

a) It reduces the need for data storage solutions.

b) It ensures data from various sources is available in a centralized repository for analysis.

c) It eliminates the need for data validation and cleansing.

d) It replaces the need for real-time data processing.

Thank You !!!