

M.Tech Program

Advanced Industry Integrated Programs

Jointly offered by University and LTIMindTree

Data Engineering

Knowledge partner



Implementation partner



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 to cover....

1. Data Types & Formats
2. Data Ingestion techniques
3. Data Profiling & Visual Representation via various tools (Pandas)
4. Storage and Retrieval Methods
5. Data Lineage Analysis

Data Lineage Analysis

Data Lineage Analysis - *Learning Outcomes..*

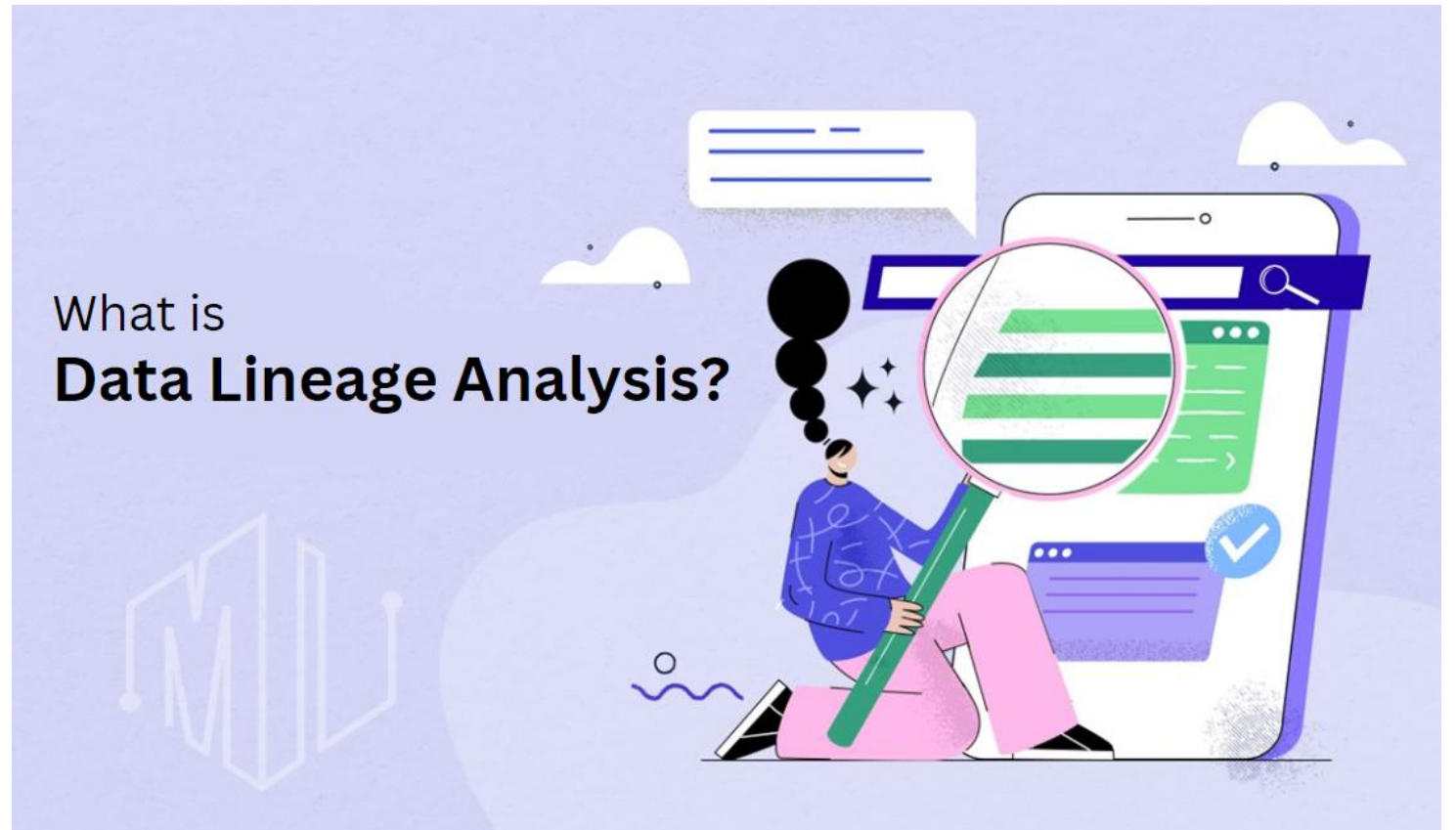
- Define the Importance of data lineage analysis in data engineering
- Definition and Importance of building a data flow in data engineering Overview of the steps involved in building a data flow
- Definition and Importance of ETL process in data engineering Overview of the steps involved in ETL process
- Overview of the benefits of edge intelligence
- Definition and Importance of data lineage
- Provide comprehensive visualization, tracking, and management of data flow, ensuring data quality, compliance, and efficient impact analysis. Importance of data lineage tools in data engineering

How to build a data flow - ETL

How to build a data flow – ETL

Introduction to Data Lineage Analysis

What is
Data Lineage Analysis?



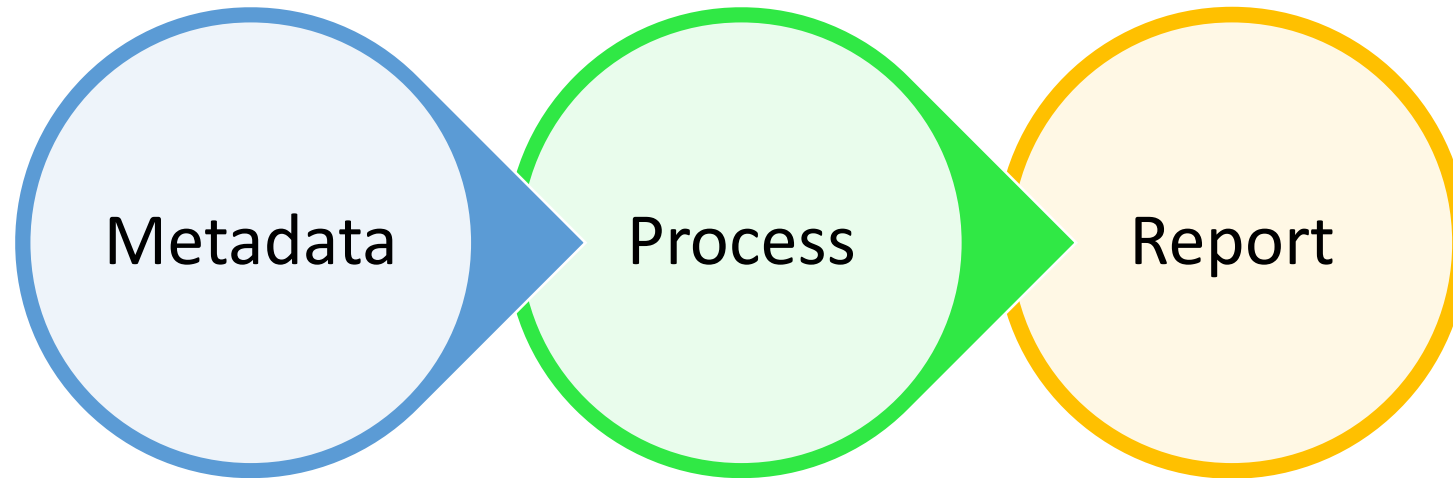
How to build a data flow – ETL

Introduction to Data Lineage Analysis

- Data lineage is the process of tracking the flow of data over time, providing a clear understanding of where the data originated, how it has changed, and its ultimate destination within the data pipeline.
- It provides a record of data throughout its lifecycle, including source information.

How to build a data flow – ETL

Introduction to Data Lineage Analysis



How to build a data flow – ETL

Introduction to Data Lineage Analysis

METADATA IN DATA LINEAGE :

- Metadata plays a crucial role in data lineage as it provides context and details about the data.
- It helps in understanding the origin, transformation, and flow of data throughout its lifecycle.

How to build a data flow – ETL

Introduction to Data Lineage Analysis

METADATA IN DATA LINEAGE :

Key aspects of metadata in data lineage include:

- Source Metadata
- Destination Metadata
- Lineage Metadata
- Transformation Metadata

How to build a data flow – ETL

Introduction to Data Lineage Analysis

Process of Data Lineage

The process of data lineage involves several steps to track and manage the flow and transformation of data. It includes :

Data Discovery

- Identifying and cataloging data sources, including databases, files, applications, and external sources.
- This step involves capturing metadata about the data sources and their structures.

How to build a data flow – ETL

Introduction to Data Lineage Analysis

Process of Data Lineage :

Data Mapping

- Defining how data moves from source to destination.
- This includes mapping data fields, understanding transformation rules, and documenting the flow of data between systems.

How to build a data flow – ETL

Introduction to Data Lineage Analysis

PROCESS OF DATA LINEAGE :

Data Transformation Tracking

- Monitoring and documenting the changes applied to data as it moves through various stages.
- This includes capturing information about data cleansing, aggregation, enrichment, and any other transformations.

How to build a data flow – ETL

Introduction to Data Lineage Analysis

PROCESS OF DATA LINEAGE :

Data Integration:

- Combining data from different sources and ensuring that lineage information is preserved throughout the integration process.
- This involves maintaining a coherent view of data flow and transformations.

How to build a data flow – ETL

Introduction to Data Lineage Analysis

PROCESS OF DATA LINEAGE :

Data Visualization

- Creating visual representations of data lineage to help users understand the flow and transformation of data.
- This can include graphical tools and dashboards that illustrate data origins, transformations, and destinations.

How to build a data flow – ETL

Introduction to Data Lineage Analysis

PROCESS OF DATA LINEAGE :

Data Lineage Documentation:

- Recording detailed information about data processes, transformations, and dependencies.
- This documentation serves as a reference for understanding data flow and ensuring compliance with data governance policies.

How to build a data flow – ETL

Introduction to Data Lineage Analysis

PROCESS OF DATA LINEAGE :

Validation and Verification:

- Regularly validating data lineage information to ensure accuracy and completeness.
- This involves checking that data transformations and flows are correctly documented and match the actual data processes.

How to build a data flow – ETL

Introduction to Data Lineage Analysis

REPORTING IN DATA LINEAGE:

Reporting in data lineage involves generating insights and summaries based on the captured metadata and data lineage information. It includes :

Data Lineage Reports:

- Detailed reports that provide insights into the origin, flow, and transformation of data.
- These reports can include graphical representations of data flows, transformation steps, and dependencies.

How to build a data flow – ETL

Introduction to Data Lineage Analysis

REPORTING IN DATA LINEAGE:

Impact Analysis Reports:

- Assessing the impact of changes to data sources, transformations, or processes on the overall data lineage.
- This helps in understanding how modifications affect data quality and integrity.

How to build a data flow – ETL

Introduction to Data Lineage Analysis

REPORTING IN DATA LINEAGE:

Compliance and Audit Reports:

- Reports that demonstrate compliance with data governance and regulatory requirements.
- These reports typically include information on data lineage to support audit trails and data stewardship.

How to build a data flow – ETL

Introduction to Data Lineage Analysis

REPORTING IN DATA LINEAGE:

Data Quality Reports:

- Summarizing data quality metrics and issues related to data lineage.
- These reports help in identifying data quality problems and ensuring that data transformations are performed correctly.

How to build a data flow – ETL

Building a Data Flow

- ETL (Extract, Transform, Load) is an automated process which takes raw data, extracts the information required for analysis, transforms it into a format that can serve business needs, and loads it to a data warehouse.
- ETL typically summarizes data to reduce its size and improve performance for specific types of analysis.

How to build a data flow – ETL

Building a Data Flow

Building a Data Flow can be done in two ways :

**Batch
Processing**

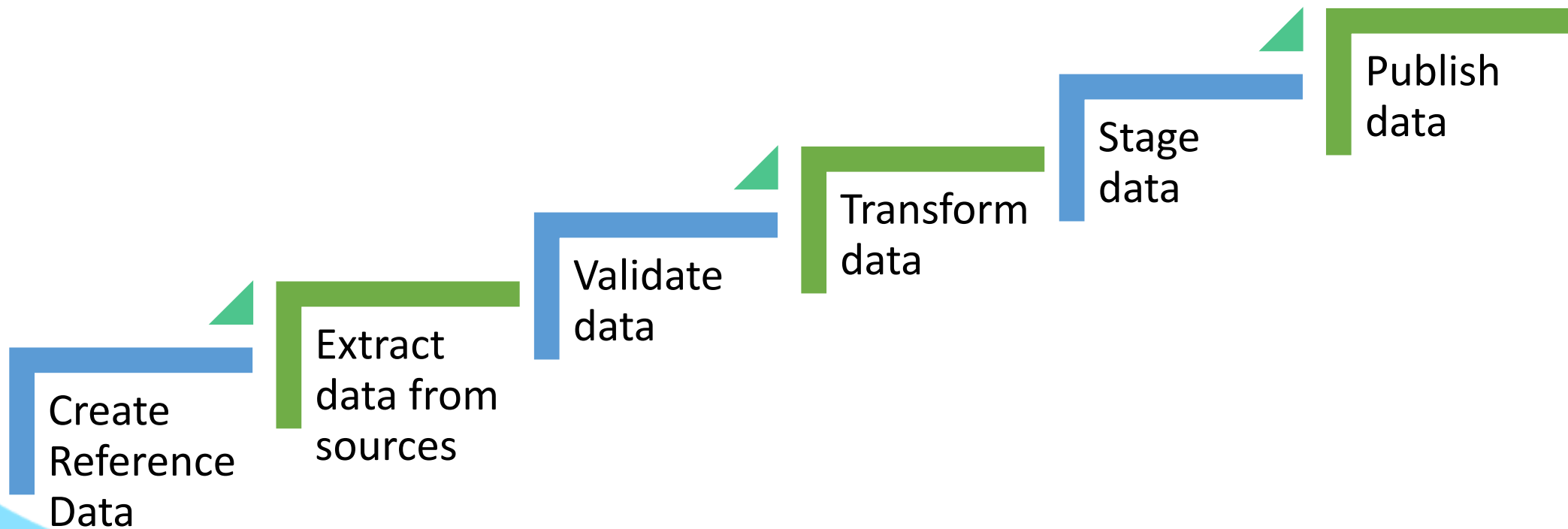
**Stream
Processing**

How to build a data flow – ETL

Building a Data Flow

Batch Processing

In a traditional ETL pipeline, data is processed in batches from source databases to a data warehouse.



How to build a data flow – ETL

Building a Data Flow

Batch Processing

- Batch processing involves collecting large volumes of data from various sources and storing it in a staging area before processing.
- The data is then processed in batches using frameworks.
- This processing includes tasks such as cleaning, aggregating, and transforming the data according to specific business logic.

How to build a data flow – ETL

Building a Data Flow

Batch Processing

- The results are stored in a data warehouse, database, or file system for further analysis.
- The processed data is then used for generating reports, feeding into dashboards, or serving machine learning models, with results being delivered on a scheduled basis

How to build a data flow – ETL

Building a Data Flow

Stream Processing

- Modern data processes often include real-time data, such as web analytics data from a large e-commerce website.
- In these cases, you cannot extract and transform data in large batches but instead, need to perform ETL on data streams.
- Thus, as client applications write data to the data source, you need to clean and transform it while it's in transit to the target data store.

How to build a data flow – ETL

Building a Data Flow

Stream Processing

- For example, ETL based on Kafka involves :



How to build a data flow – ETL

Building a Data Flow

Stream Processing

- Stream processing involves the continuous ingestion of data from various sources, processing it in real-time as it arrives.
- Unlike batch processing, data is not stored before processing; instead, it is immediately passed through a stream processing engine like Apache Kafka, Apache Flink, or Apache Storm.

How to build a data flow – ETL

Building a Data Flow

Stream Processing

- The engine applies transformations, filtering, and aggregations to the data on the fly, ensuring near-instantaneous processing and analysis.
- Processed data is then pushed to real-time dashboards, triggering alerts, or feeding into other systems like databases or machine learning models.
- The entire flow is monitored for latency, throughput, and potential errors to maintain the required real-time performance and reliability.

How to build a data flow – ETL

ETL (Extract, Transform, Load) Process

- ETL—meaning extract, transform, load—is a data integration process that combines, cleans and organizes data from multiple sources into a single, consistent data set for storage in a data warehouse, data lake or other target system.
- ETL data pipelines provide the foundation for data analytics and machine learning workstreams

How to build a data flow – ETL

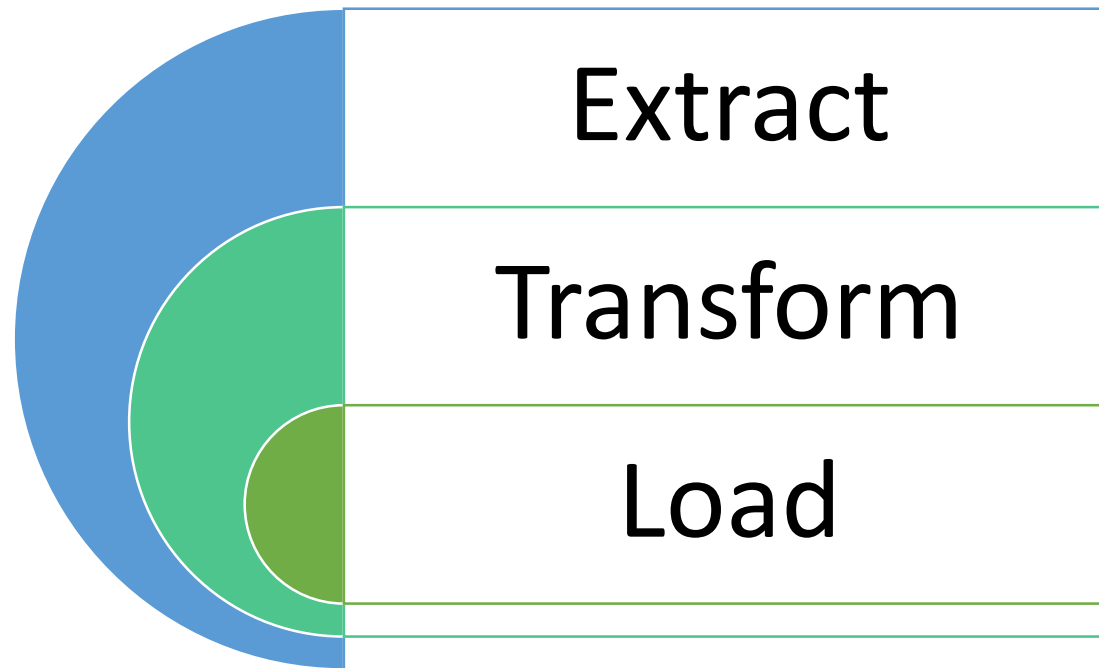
ETL (Extract, Transform, Load) Process

- The ETL process is an iterative process that is repeated as new data is added to the warehouse.
- The process is important because it ensures that the data in the data warehouse is accurate, complete, and up-to-date.
- It also helps to ensure that the data is in the format required for data mining and reporting.

How to build a data flow – ETL

ETL (Extract, Transform, Load) Process

- The ETL process is an iterative process that is repeated as new data is added to the warehouse.



How to build a data flow – ETL

ETL (Extract, Transform, Load) Process

Extract :

- The first stage in the ETL process is to extract data from various sources such as transactional systems, spreadsheets, and flat files.
- This step involves reading data from the source systems and storing it in a staging area.
- The extract process creates an extract file that contains the selected set of rows from one or more tables.

How to build a data flow – ETL

ETL (Extract, Transform, Load) Process

Transform:

- In this stage, the extracted data is transformed into a format that is suitable for loading into the data warehouse.
- This may involve cleaning and validating the data, converting data types, combining data from multiple sources, and creating new data fields.
- The goal is to transform the source data into a format that's suitable for the target destination.

How to build a data flow – ETL

ETL (Extract, Transform, Load) Process

Load:

- The load process in ETL (Extract, Transform, Load) is the final step in the process of preparing raw data for analysis and reporting.
- In this step, transformed data is written from a staging area to a target database, data warehouse, data lake, or other storage.
- The destination depends on the data's size and complexity, as well as the organization's requirements.

When to use data warehouse and when not to

When to use data warehouse and when not to

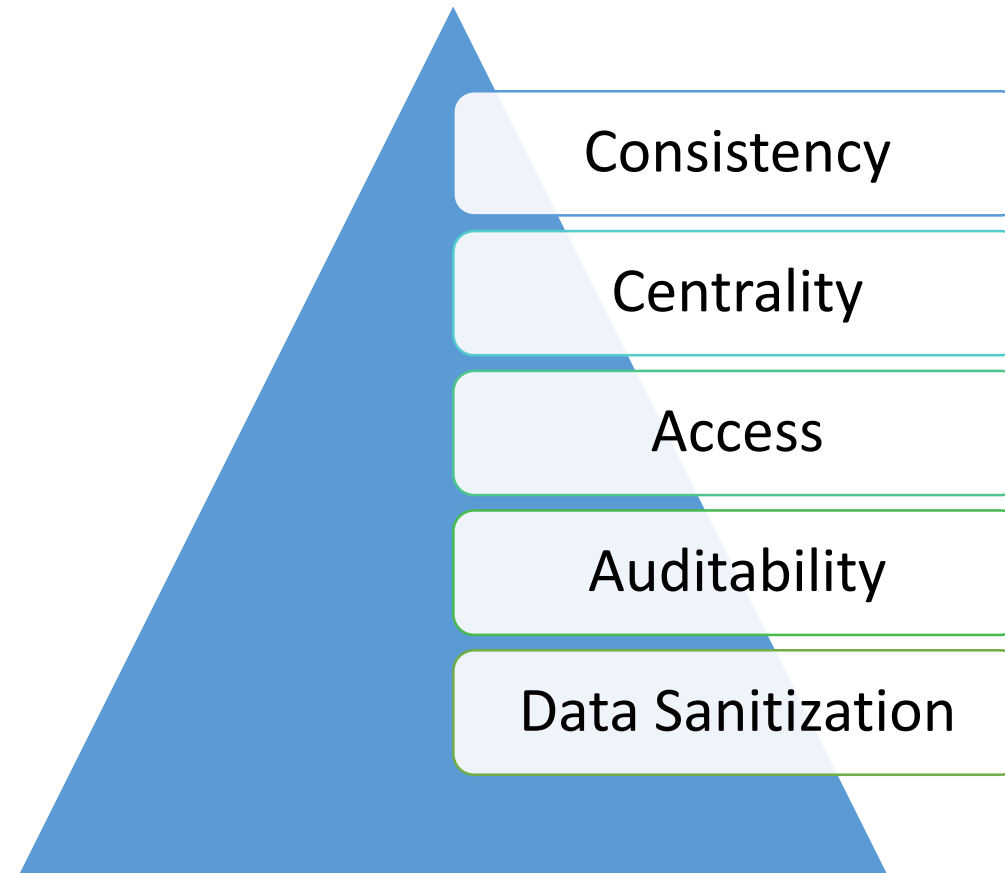
Usage of Data Warehouse

- Data warehouses store and process large amounts of data from various sources within a business.
- Data warehouses help companies make better, more informed decisions by applying data analytics to large volumes of information.
- Data warehouses are relational environments that are used for data analysis, particularly of historical data.

When to use data warehouse and when not to

Usage of Data Warehouse

Benefits of Data Warehouse



When to use data warehouse and when not to

Usage of Data Warehouse

Benefits of Data Warehouse

Consistency:

- Warehousing ensures that all data is uniformly stored and managed, providing a single source of truth for accurate reporting and analysis.
- This consistency reduces discrepancies and errors across different data sources.

Centrality:

- A centralized data warehouse consolidates information from various sources, making it easier to manage, access, and analyze data in one location.

When to use data warehouse and when not to

Usage of Data Warehouse

Benefits of Data Warehouse

Centrality:

- This centralization simplifies data governance and streamlines decision-making processes.

Access:

- Warehousing provides easy and controlled access to data for authorized users, enabling efficient data retrieval for analysis and reporting.

When to use data warehouse and when not to

Usage of Data Warehouse

Benefits of Data Warehouse

Access:

- It supports different user roles and permissions, ensuring that data is accessible to those who need it.

Auditability:

- Data warehousing enhances auditability by maintaining detailed logs and histories of data changes, allowing for thorough tracking and review.

When to use data warehouse and when not to

Usage of Data Warehouse

Benefits of Data Warehouse

Auditability:

- This ensures compliance with regulatory requirements and supports transparency in data management.

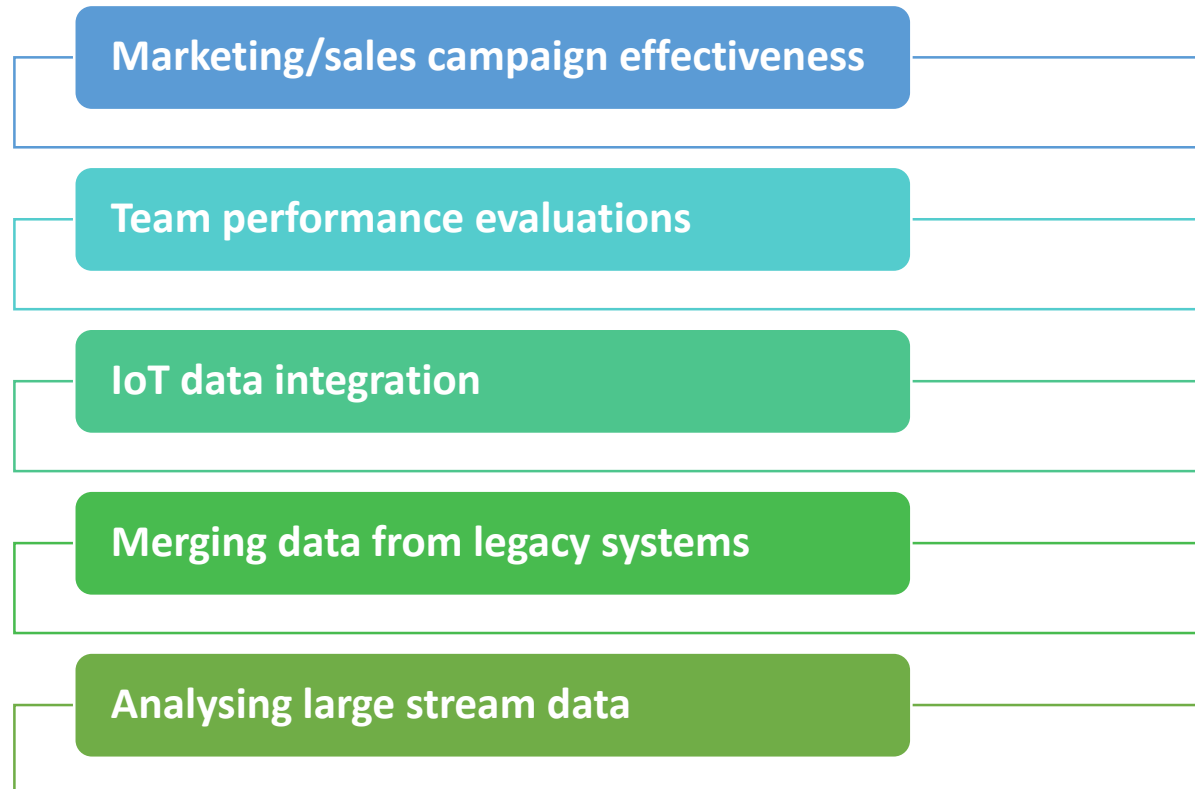
Data Sanitization:

- Warehousing includes processes for data cleansing and sanitization, ensuring that stored data is accurate, consistent, and free of errors.
- This improves the quality of data, leading to more reliable insights and decisions.

When to use data warehouse and when not to

Usage of Data Warehouse

Use cases of Data Warehouse



When to use data warehouse and when not to

Usage of Data Warehouse

Use cases of Data Warehouse

Marketing/Sales Campaign Effectiveness:

- Warehousing aggregates data from various marketing and sales channels, allowing for comprehensive analysis of campaign performance and customer behavior.
- By storing historical data on customer interactions, preferences, and responses, warehousing also facilitates longitudinal analysis, enabling the identification of trends and patterns for improved targeting and personalization in future campaigns.

When to use data warehouse and when not to

Usage of Data Warehouse

Use cases of Data Warehouse

Team Performance Evaluations:

- Data warehousing consolidates performance metrics and feedback from multiple sources, enabling detailed analysis of team performance and individual contributions. Additionally, by integrating data from various HR systems, project management tools, and employee feedback platforms, warehousing provides a holistic view of team dynamics and productivity, helping to identify areas for professional development and improvement.

When to use data warehouse and when not to

Usage of Data Warehouse

Use cases of Data Warehouse

IoT Data Integration:

- Warehousing facilitates the integration and analysis of vast amounts of data generated by IoT devices, providing a unified view of sensor data.
- This enables real-time insights and long-term trend analysis, enhancing operational efficiency and decision-making.

When to use data warehouse and when not to

Usage of Data Warehouse

Use cases of Data Warehouse

Merging Data from Legacy Systems:

- Warehousing allows for the seamless integration of data from outdated or disparate legacy systems into a modern, centralized repository.
- This enables comprehensive data analysis and reporting while preserving historical data.

When to use data warehouse and when not to

Usage of Data Warehouse

Use cases of Data Warehouse

Analyzing Large Stream Data:

- By consolidating and processing large volumes of streaming data, warehousing supports real-time analytics and trend identification.
- This capability enhances the ability to make timely decisions based on up-to-date information.

Edge Intelligence in Data Flow

Edge Intelligence in Data Flow

What is Edge computing

- Edge computing is a distributed computing model that involves processing, analyzing, and storing data near where it is generated. This allows for faster, near real-time analysis and response.
- It is a distributed computing framework that brings applications closer to data sources such as IoT devices, local end devices, or edge servers.

Edge Intelligence in Data Flow

Key capabilities of Edge Computing

Manage the distribution of software at massive scale

Leverage open-source technology

Address security concerns

Engage a trusted partner with deep industry expertise

Edge Intelligence in Data Flow

Key capabilities of Edge Computing

Manage distribution of software at massive scale:

- This capability allows organizations to deploy updates and patches across thousands of edge devices simultaneously, minimizing downtime and maintaining consistent software performance.
- It also supports the management of diverse hardware environments, ensuring compatibility and operational efficiency across the entire edge network.

Edge Intelligence in Data Flow

Key capabilities of Edge Computing

Leverage open source technology

- Open source tools in edge computing provide a robust foundation for building and scaling edge solutions, promoting community-driven innovation and quick adaptation to evolving technologies.
- These technologies often come with extensive community support and resources, making it easier to integrate and maintain edge systems.

Edge Intelligence in Data Flow

Key capabilities of Edge Computing

Address security concerns:

- By processing data at the edge, sensitive information is kept closer to its source, reducing exposure to potential threats during data transmission.
- Additionally, edge computing
- can incorporate advanced security measures like encryption and real-time threat detection, enhancing the overall security posture.

Edge Intelligence in Data Flow

Key capabilities of Edge Computing

Engage a trusted partner with deep industry expertise:

- A knowledgeable partner can offer tailored solutions that meet the specific regulatory, technical, and operational requirements of your industry, ensuring compliance and optimal performance.
- Their expertise also helps in navigating the complexities of edge computing, from deployment to ongoing management, reducing the risk of costly errors and delays.

Edge Intelligence in Data Flow

Major components of Edge Computing

Micro controllers

Micro processors

AI accelerator

Operating systems

AI frameworks

Libraries

Edge Intelligence in Data Flow

Major components of Edge Computing

Microcontrollers:

- Microcontrollers are compact, low-power devices designed to execute simple control tasks at the edge, often used in IoT devices for real-time data processing and sensor management.

Microprocessors:

- Microprocessors are more powerful processors that handle complex computations at the edge, enabling advanced data processing, multitasking, and communication with other systems.

Edge Intelligence in Data Flow

Major components of Edge Computing

AI Accelerators:

- AI accelerators are specialized hardware components optimized for running AI and machine learning algorithms at the edge, providing high-performance inference with low latency.

Operating Systems:

- Operating systems for edge devices manage hardware resources and provide a platform for running applications, ensuring efficient execution of tasks and seamless interaction with other components.

Edge Intelligence in Data Flow

Major components of Edge Computing

AI Frameworks:

- AI frameworks provide the necessary tools and libraries for developing and deploying AI models at the edge, facilitating rapid prototyping, training, and inference on local devices.

Libraries:

- Libraries in edge computing offer pre-built code modules that simplify the development process by providing reusable functions and protocols for data processing, communication, and device management.

Edge Intelligence in Data Flow

Use cases of Edge Computing

Autonomous vehicles

Remote monitoring of assets in the oil and gas industry

Smart grid

Predictive maintenance

In-hospital patient monitoring

Virtualized radio networks and 5G (vRAN)

Cloud gaming

Content delivery

Edge Intelligence in Data Flow

Use cases of Edge Computing

Autonomous Vehicles:

- Edge computing enables real-time data processing in autonomous vehicles, allowing for rapid decision-making and response to changing road conditions, enhancing safety and performance.

Remote Monitoring of Oil and Gas Assets:

- Edge computing facilitates real-time monitoring and analysis of oil and gas assets in remote locations, enabling immediate detection of anomalies and reducing the need for manual inspections.

Edge Intelligence in Data Flow

Use cases of Edge Computing

Smart Grid:

- In smart grids, edge computing processes data from sensors and meters locally, optimizing energy distribution and load balancing while minimizing latency.
- This helps in real-time demand response, improving the efficiency and reliability of the power grid.

Edge Intelligence in Data Flow

Use cases of Edge Computing

Predictive Maintenance:

- Edge computing enables the analysis of machine data at the source, allowing for early detection of potential failures and timely maintenance, reducing downtime and operational costs.
- This real-time monitoring also extends the lifespan of equipment by addressing issues before they become critical.

Edge Intelligence in Data Flow

Use cases of Edge Computing

In-Hospital Patient Monitoring:

- Edge computing processes patient data at the bedside, providing immediate insights to healthcare providers, improving response times, and enhancing patient care.
- It supports continuous monitoring and alerts for critical conditions, reducing the reliance on centralized systems.

Edge Intelligence in Data Flow

Use cases of Edge Computing

Virtualized Radio Networks and 5G (vRAN):

- Edge computing supports the deployment of virtualized radio networks, enhancing the flexibility and scalability of 5G networks.
- It enables low-latency services by processing data closer to the user, improving the performance of mobile and IoT applications.

Edge Intelligence in Data Flow

Use cases of Edge Computing

Cloud Gaming:

- Edge computing reduces latency in cloud gaming by processing game data closer to the player, ensuring smoother gameplay and faster response times.
- It also supports the delivery of high-quality graphics and real-time interactions, enhancing the gaming experience.

Edge Intelligence in Data Flow

Use cases of Edge Computing

Content Delivery:

- Edge computing enhances content delivery by caching and processing data at edge locations, reducing latency and improving the user experience for streaming and downloads.
- It also optimizes bandwidth usage, ensuring faster access to content for end-users.

Edge Intelligence in Data Flow

Benefits of Edge Computing

Faster
response
times

Better
security

Reduced
network
congestion

Higher
reliability

Lower costs

Edge Intelligence in Data Flow

Benefits of Edge Computing

Faster Response Times:

- Edge computing processes data closer to its source, significantly reducing latency and providing quicker responses for real-time applications and services.

Better Security:

- By keeping data local and minimizing the need for transmission over networks,
- edge computing reduces exposure to potential breaches and enhances overall data security.

Edge Intelligence in Data Flow

Benefits of Edge Computing

Reduced Network Congestion:

- Edge computing offloads data processing from centralized servers to
- local devices, decreasing the volume of data transmitted over networks and alleviating congestion.

Edge Intelligence in Data Flow

Benefits of Edge Computing

Higher Reliability:

- By distributing data processing across multiple edge devices, edge computing minimizes single points of failure, ensuring continuous operation even if individual devices encounter issues.
- This distributed architecture also allows for localized redundancy and fault tolerance, further enhancing system reliability.

Edge Intelligence in Data Flow

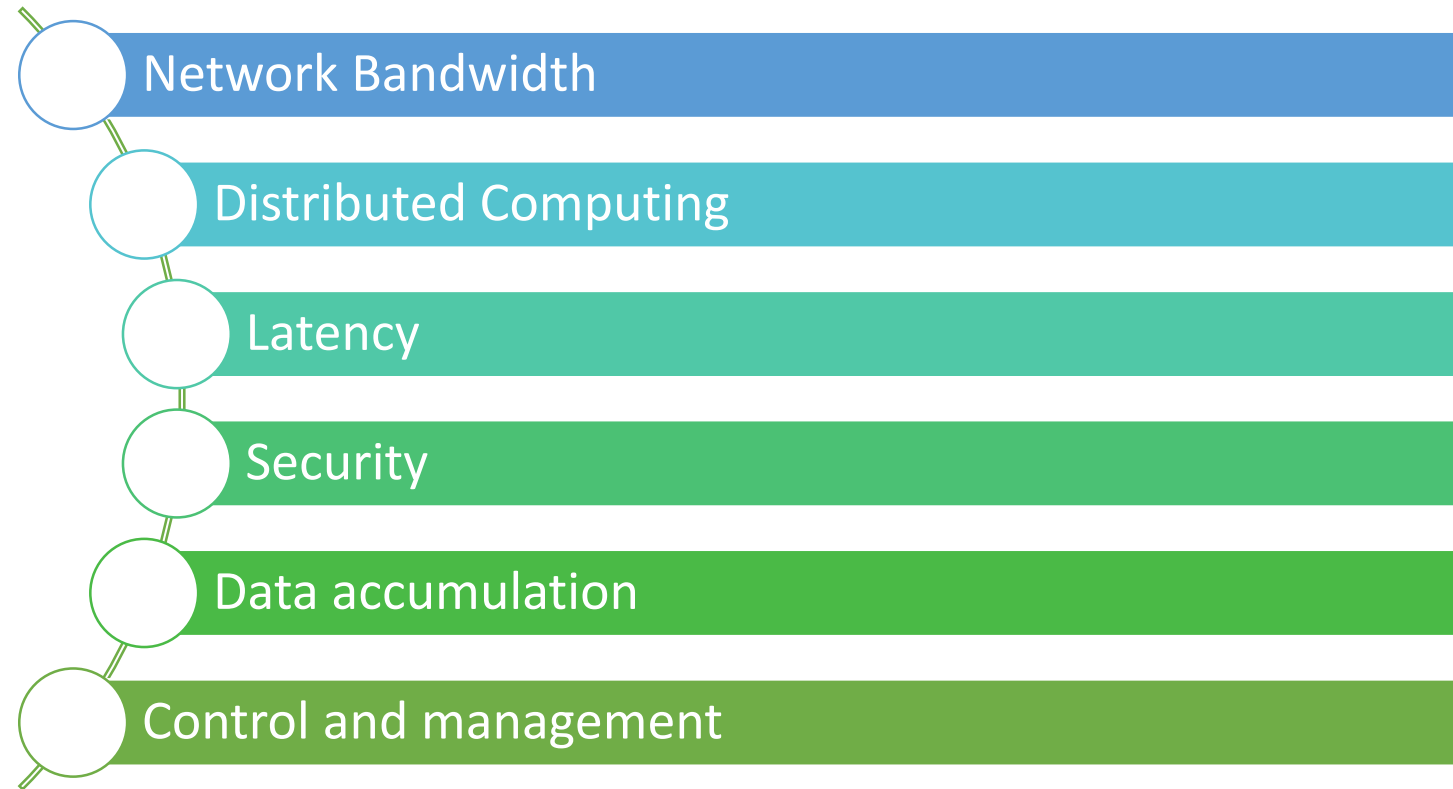
Benefits of Edge Computing

Lower Costs:

- Edge computing minimizes the need for data center resources and reduces the volume of data transmitted to and from central servers, leading to lower operational and bandwidth costs.
- Additionally, local processing decreases the strain on centralized infrastructure, potentially extending the lifespan of existing resources and reducing overall expenses.

Edge Intelligence in Data Flow

Challenges of Edge Computing



Edge Intelligence in Data Flow

Challenges of Edge Computing

Network Bandwidth:

- Edge devices often operate in bandwidth-constrained environments, limiting data transfer capabilities and impacting real-time processing.
- Optimizing bandwidth usage while maintaining data integrity is a significant challenge.

Distributed Computing:

- Managing and coordinating tasks across numerous distributed edge devices can be complex, requiring efficient algorithms and robust infrastructure to ensure consistent performance.

Edge Intelligence in Data Flow

Challenges of Edge Computing

Latency:

- Edge computing aims to minimize latency, but achieving low-latency processing is challenging due to network variability, especially in remote or high-demand scenarios.

Security:

- With data processing occurring closer to the source, edge devices are more vulnerable to physical and cyber threats, making security a critical concern in protecting sensitive information.

Edge Intelligence in Data Flow

Challenges of Edge Computing

Data Accumulation:

- Edge devices generate vast amounts of data, necessitating efficient data aggregation, filtering, and storage mechanisms to prevent overload and ensure actionable insights.

Control and Management:

- Maintaining centralized control and managing updates, configurations, and monitoring across a widely distributed edge network can be difficult, requiring sophisticated tools and strategies.

What is Data Lineage

What is Data Lineage

Understanding Data Lineage

- Data lineage is a metadata map that traces the relationship between upstream and downstream dependencies in your data pipelines to illustrate where your data comes from, how it changes, and where it's surfaced to end users.
- The data lineage life cycle shows at-a-glance how data is collected and activated to help data teams understand and monitor upstream and downstream dependencies.

What is Data Lineage

Understanding Data Lineage

Data Lineage helps to

- Track errors in data processes
- Implement process changes with lower risk
- Perform system migrations with confidence
- Combine data discovery with a comprehensive view of metadata, to create a data mapping framework

What is Data Lineage

Impact of Data Lineage

Strategic reliance on data :

- Detailed information provided through data lineage helps better understand the meaning and validity of this data.

Data in flux :

- Data lineage provides tracking capabilities that make it possible to reconcile and make the best use of old and new datasets.

What is Data Lineage

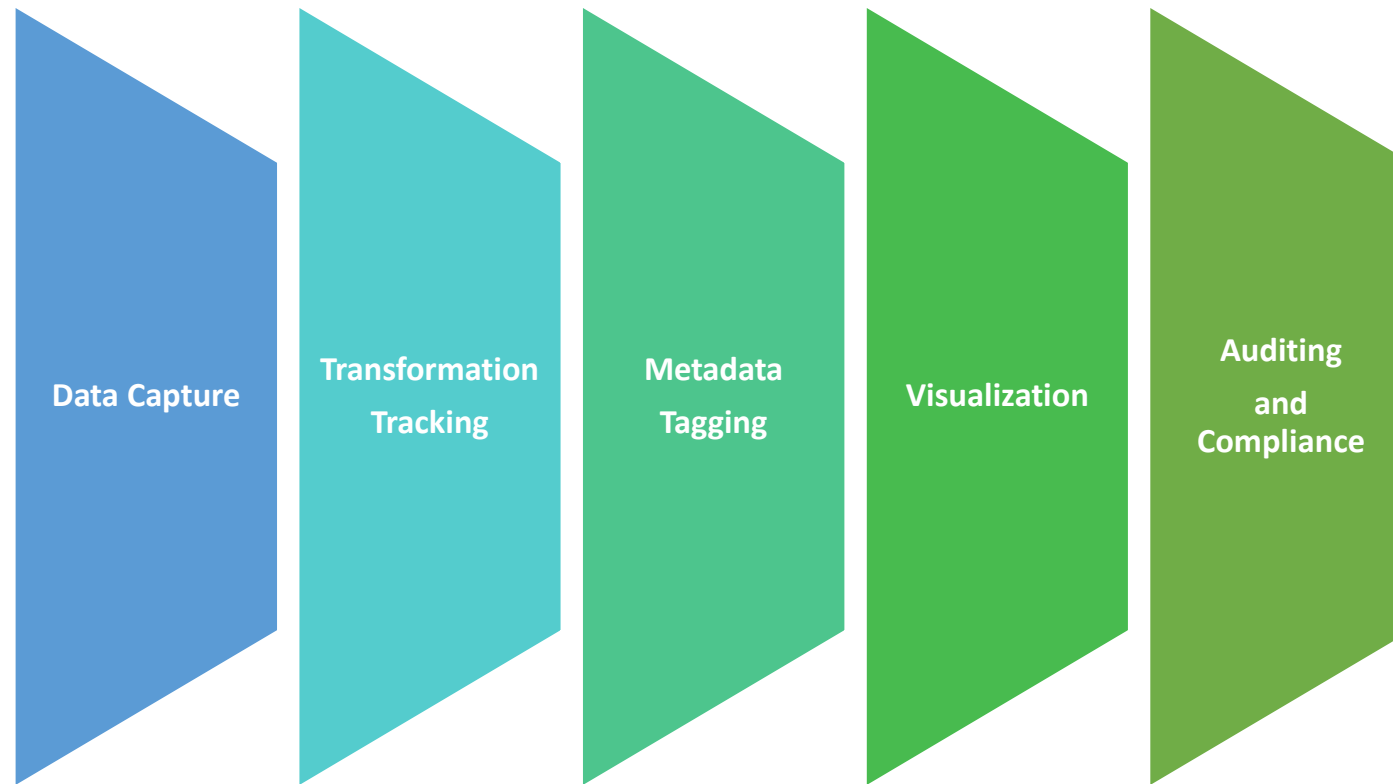
Impact of Data Lineage

Data Governance :

- The details tracked in data lineage are a good way to provide compliance auditing, improve risk management, and ensure data is stored and processed in line with organizational policies and regulatory standards.

What is Data Lineage

How Data Lineage works



What is Data Lineage

How Data Lineage works

Data lineage tracks the movement and transformation of data from its origin through various stages to its final destination. It works by capturing and documenting each step in the data lifecycle, including data sources, transformations, and destinations.

Data Capture:

- Initial data is collected from various sources, such as databases, applications, or external feeds.

What is Data Lineage

How Data Lineage works

Transformation Tracking:

- As data is processed, transformed, or aggregated, each step is recorded to capture how data changes and flows through different systems.

Metadata Tagging:

- Metadata tags or annotations are attached to data elements to provide context and traceability for each transformation.

What is Data Lineage

How Data Lineage works

Visualization:

- Data lineage tools generate visual representations of data flow, showing how data moves and changes through systems.

Auditing and Compliance:

- By documenting all transformations and data movements, lineage supports auditing and regulatory compliance, ensuring transparency and accountability.

What is Data Lineage

Benefits of Data Lineage

Compliance

Trust

Impact
analysis

Technical debt
reduction

Data quality

What is Data Lineage

Benefits of Data Lineage

Compliance:

- Data lineage ensures adherence to regulatory requirements by documenting data transformations and flows, providing a clear audit trail for compliance and governance.

Trust:

- By offering transparency into data origins and transformations, data lineage builds trust among stakeholders by verifying the accuracy and reliability of data used for decision-making.

What is Data Lineage

Benefits of Data Lineage

Impact Analysis:

- Data lineage enables assessment of the effects of changes to data sources or processes, helping to predict and manage potential impacts on downstream systems and reports.

Technical Debt Reduction:

- By clearly documenting data processes and dependencies, data lineage helps identify and address technical debt, simplifying maintenance and reducing complexity in data systems.

What is Data Lineage

Benefits of Data Lineage

Data Quality:

- Data lineage improves data quality by tracking and validating data transformations, helping to identify and rectify errors and inconsistencies throughout the data lifecycle.

What is Data Lineage

Features of Data lineage tools

A modern data lineage tool must support :

- Robust data import capabilities
- Column- and field-level lineage
- Compatibility with upstream producers and downstream consumers
- Data lineage usability and User Experience (UX)
- Collaboration and open API support
- Active metadata support

What is Data Lineage

Features of Data lineage tools

Robust Data Import Capabilities:

- Robust data import capabilities enable seamless integration of diverse data sources into a centralized system, ensuring comprehensive and accurate data ingestion.
- This capability supports various formats and protocols, facilitating smooth and efficient data consolidation.

What is Data Lineage

Features of Data lineage tools

Column- and Field-Level Lineage:

- Column- and field-level lineage tracks data transformations at the most granular level, providing detailed insights into how individual data elements are processed and modified.
- This precision enhances transparency and accuracy in data tracking.

What is Data Lineage

Features of Data lineage tools

Compatibility with Upstream Producers and Downstream Consumers:

- Compatibility ensures that data lineage tools can effectively interact with both upstream data sources and downstream applications, enabling smooth data flow and integration across the entire data ecosystem.

What is Data Lineage

Features of Data lineage tools

Data Lineage Usability and User Experience (UX):

- Usability and UX focus on creating intuitive and user-friendly interfaces for exploring and interacting with data lineage information, making it easier for users to understand data flow and transformations without technical expertise.

What is Data Lineage

Features of Data lineage tools

Collaboration and Open API Support:

- Collaboration and open API support facilitate integration with other tools and systems, allowing teams to share lineage information and enhance data management practices through customizable and interoperable solutions.

What is Data Lineage

Features of Data lineage tools

Active Metadata Support:

- Active metadata support involves dynamically updating and managing metadata associated with data lineage, ensuring that lineage information remains current and accurately reflects ongoing data changes and transformations.

What is Data Lineage

Techniques in Data Lineage

Pattern based Lineage

Lineage by Data tagging

Self contained Lineage

Lineage by parsing

What is Data Lineage

Techniques in Data Lineage

Pattern based Lineage

- Pattern-Based Lineage traces data movement by identifying and analyzing recurring processing patterns throughout the data lifecycle.
- This approach provides clear visibility into how data transforms from its origin to its final destination.
- Key benefits include improved traceability, enhanced data quality by spotting inconsistencies, and simplified data management.

What is Data Lineage

Techniques in Data Lineage

Lineage by Data tagging

- Lineage by Data Tagging tracks the flow of data by attaching metadata tags to data elements as they move through the system.
- This method provides detailed visibility into data transformations and dependencies by following these tags from source to destination.
- Key benefits include enhanced traceability, as each data element's path is documented through its tags, and improved data quality by identifying data issues and inconsistencies.

What is Data Lineage

Techniques in Data Lineage

Self contained Lineage

- Self-Contained Lineage tracks data movement and transformation within isolated data environments or systems.
- It provides a complete view of data processing by maintaining all necessary lineage information within the system itself, without relying on external sources.
- Key benefits include enhanced visibility into data workflows, improved accuracy in tracking data changes, and simplified data management through built-in lineage documentation.

What is Data Lineage

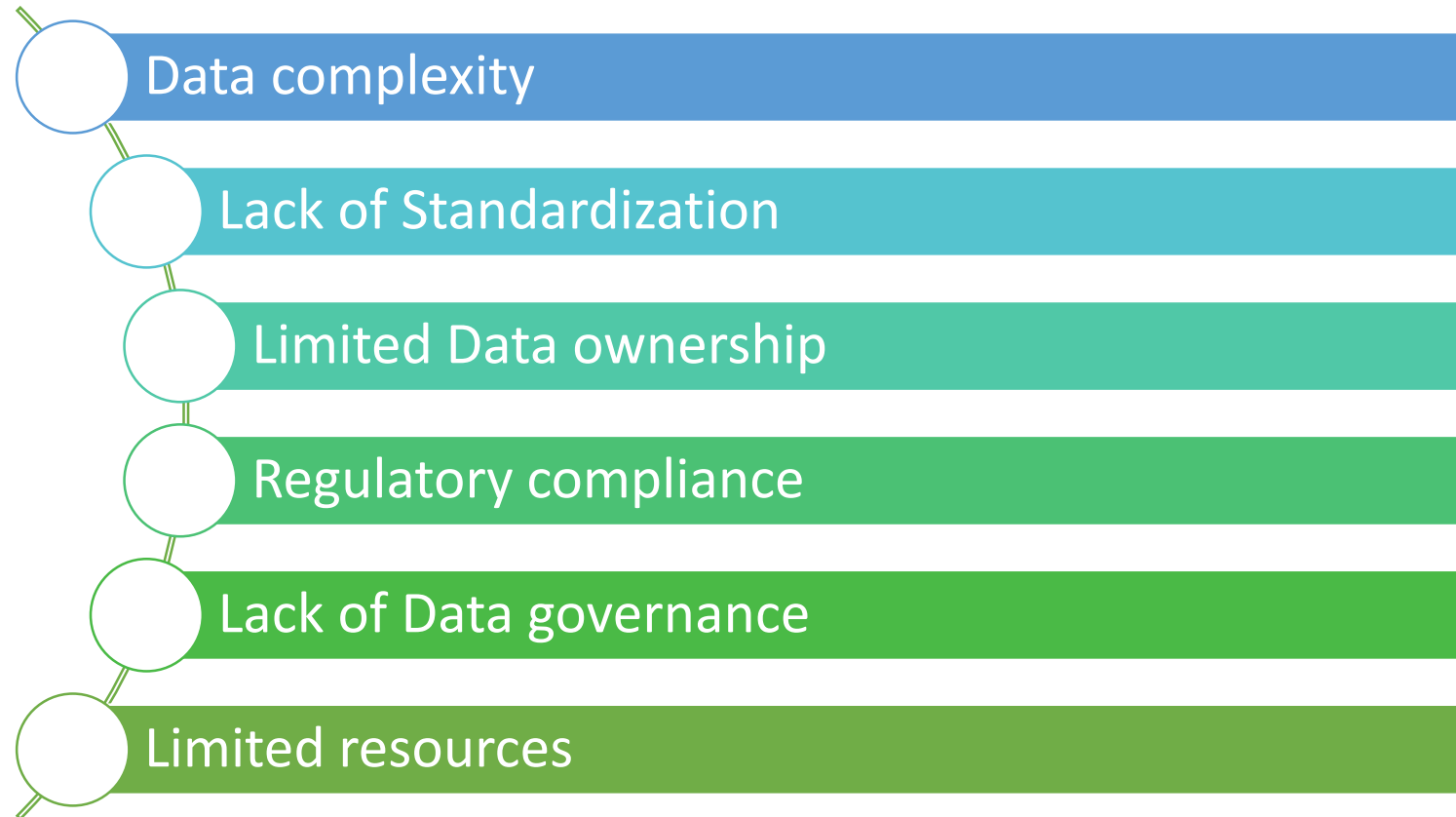
Techniques in Data Lineage

Lineage by Parsing

- Lineage by Parsing involves analyzing and interpreting data processing scripts, logs, or configuration files to track data transformations and flow.
- This method provides insights into how data is manipulated by examining the structure and content of processing instructions.
- Key benefits include detailed visibility into data workflows by decoding processing logic, improved traceability by linking transformations to specific scripts or configurations, and easier troubleshooting through clear understanding of data handling procedures.

What is Data Lineage

Challenges in Data lineage



What is Data Lineage

Challenges in Data lineage

Data Complexity:

- The complexity of modern data environments, with numerous sources, transformations, and integrations, makes it challenging to capture and maintain accurate lineage.
- Managing this complexity requires sophisticated tools and methodologies to ensure comprehensive tracking and visibility.

What is Data Lineage

Challenges in Data lineage

Lack of Standardization:

- The absence of standardized practices and formats for documenting data lineage can lead to inconsistencies and difficulties in integrating lineage information across different systems and platforms.
- This lack of standardization hampers effective data management and analysis.

What is Data Lineage

Challenges in Data lineage

Limited Data Ownership:

- In organizations with fragmented data ownership, responsibilities for data management and lineage can be unclear or poorly defined.
- This limitation complicates efforts to track and maintain accurate data lineage, leading to potential gaps and inaccuracies.

What is Data Lineage

Challenges in Data lineage

Regulatory Compliance:

- Meeting regulatory requirements for data lineage can be challenging due to the need for detailed documentation and transparency across various data processes.
- Ensuring compliance requires robust lineage tracking and reporting mechanisms to adhere to regulatory standards.

What is Data Lineage

Challenges in Data lineage

Lack of Data Governance:

- Inadequate data governance practices can result in insufficient oversight and control over data lineage processes.
- Without effective governance, maintaining accurate and consistent lineage information becomes difficult, impacting data quality and reliability.

What is Data Lineage

Challenges in Data lineage

Limited Resources:

- Resource constraints, including budget, personnel, and technical capabilities, can impede the implementation and maintenance of comprehensive data lineage solutions.
- Limited resources may restrict the ability to fully address lineage challenges and invest in necessary tools and technologies.

Summary

Data Lineage Analysis

Data Lineage Analysis

Data Lineage analysis tracks and visualizes the flow of data through systems, documenting its origins, transformations and destinations.

Building a Data flow

Building a data flow involves data sources, designing transformation processes and mapping data pathways.

ETL process

The ETL (Extract, Transform, Load) process involves extracting data from various sources, transforming it to fit operational needs, and loading it into a target system for analysis.

Data Lineage Analysis

Metadata

Metadata in data lineage provides essential context about data sources, transformations, and destinations.

Process

The process of data lineage involves discovering and cataloging data sources, mapping data flows and transformations, and tracking how data changes throughout its lifecycle.

Report

Reporting in data lineage involves generating detailed summaries and visual representations of data origins, transformations, and destinations.

Data Lineage Analysis

Data Warehouse

A data warehouse is a centralized repository that stores large volumes of structured data from multiple sources.

Edge computing

Edge computing refers to processing data closer to its source, rather than relying on centralized data centers.

Benefits of Edge computing

Key benefits include faster response times, improved security, reduced network congestion, and higher reliability by distributing computing tasks.

Data Lineage Analysis

Challenges of Edge computing

Edge computing faces challenges such as managing scalability across numerous devices and ensuring data consistency in distributed environments.

Understanding of Data Lineage

Data lineage tracks the journey of data from its origin through various transformations to its final destination. It provides visibility into data processes, enhancing traceability, data quality, and compliance.

How data lineage works

Data lineage works by documenting and visualizing the flow of data from its source through transformations to its final destination.

Data Lineage Analysis

Benefits of Data lineage

Data lineage enhances traceability by providing a clear view of data's journey and transformations. It improves data quality and supports compliance by documenting data processes and changes.

Data lineage tool features

Data lineage tools offer visual mapping of data flows, capturing metadata for detailed tracking of transformations and sources.

Techniques in data lineage

Techniques in data lineage include pattern-based lineage, which identifies recurring data processing patterns; tagging, which attaches metadata to track transformations; and parsing, which analyzes scripts and logs to trace data flow

Data Lineage Analysis

- The module consisted of Data lineage which involves tracking and visualizing data from its origin through transformations to its final destination, enhancing traceability and quality.
- Building effective data flows and implementing the ETL process ensure accurate data management and reporting in data warehouses.
- Edge computing processes data closer to its source, offering benefits like reduced latency and improved security, though it faces challenges such as scalability and resource constraints.
- Understanding and utilizing data lineage tools and techniques, including pattern-based lineage and tagging, address complexities and support compliance.
- Overall, these concepts are crucial for efficient data management, analysis, and operational success.

Knowledge Check

Data Lineage Analysis

Q1 : What is the primary purpose of data lineage?

- A) Data storage optimization
- B) Tracking data flow and transformations**
- C) Data encryption
- D) Data visualization

Data Lineage Analysis

Q2 : Which process involves extracting data from sources, transforming it, and then loading it into a data warehouse?

- A) Data Migration
- B) ETL (Extract, Transform, Load)**
- C) Data Aggregation
- D) Data Visualization

Data Lineage Analysis

Q3 : What is a key benefit of edge computing?

- A) Increased central server load
- B) Reduced data processing speed
- C) **Faster response times**
- D) Higher data transmission costs

Data Lineage Analysis

Q4 : Which technique in data lineage involves identifying and analyzing recurring processing patterns?

- A) Tagging
- B) Parsing
- C) **Pattern-Based Lineage**
- D) Data Mapping

Data Lineage Analysis

Q5 : What is one of the main challenges of edge computing?

- A) Centralized data processing
- B) Reduced network congestion
- C) **Managing scalability across numerous devices**
- D) Simplified security measures

Data Lineage Analysis

Q6 : How does a data warehouse support business intelligence?

- A) By storing data in real-time
- B) By integrating and consolidating data for complex queries**
- C) By encrypting data at rest
- D) By providing data visualization tools

Data Lineage Analysis

Q7 : Which of the following is NOT a typical feature of data lineage tools?

- A) Visual Mapping
- B) Metadata Management
- C) **Real-Time Data Processing**
- D) Impact Analysis

Data Lineage Analysis

Q8 : What does the 'Transform' step in the ETL process involve?

- A) Extracting data from sources
- B) Loading data into a target system
- C) **Cleaning and modifying data**
- D) Encrypting data

Data Lineage Analysis

Q9 : Which use case is a good example of edge computing application?

- A) Large-scale data warehousing
- B) **Autonomous Vehicles**
- C) Centralized data management
- D) Long-term data storage

Data Lineage Analysis

Q10 : What is a common challenge associated with data lineage?

- A) Data visualization
- B) Managing data complexity**
- C) Data encryption
- D) Data aggregation



Thank You !!!