DATA ENGINEERING

60+

DATA ENGINEERING 101

ETL - TERMINOLOGY

Everything you need to begin the journey

Swipe Left

ETL System

Extract, Transform, Load (ETL) system is responsible for extracting data from source systems, transforming it to fit business needs, and loading it into a data warehouse.

A retail company extracts sales data from its point-of-sale system, transforms it to analyze seasonal trends, and loads it into a data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Data Warehouse

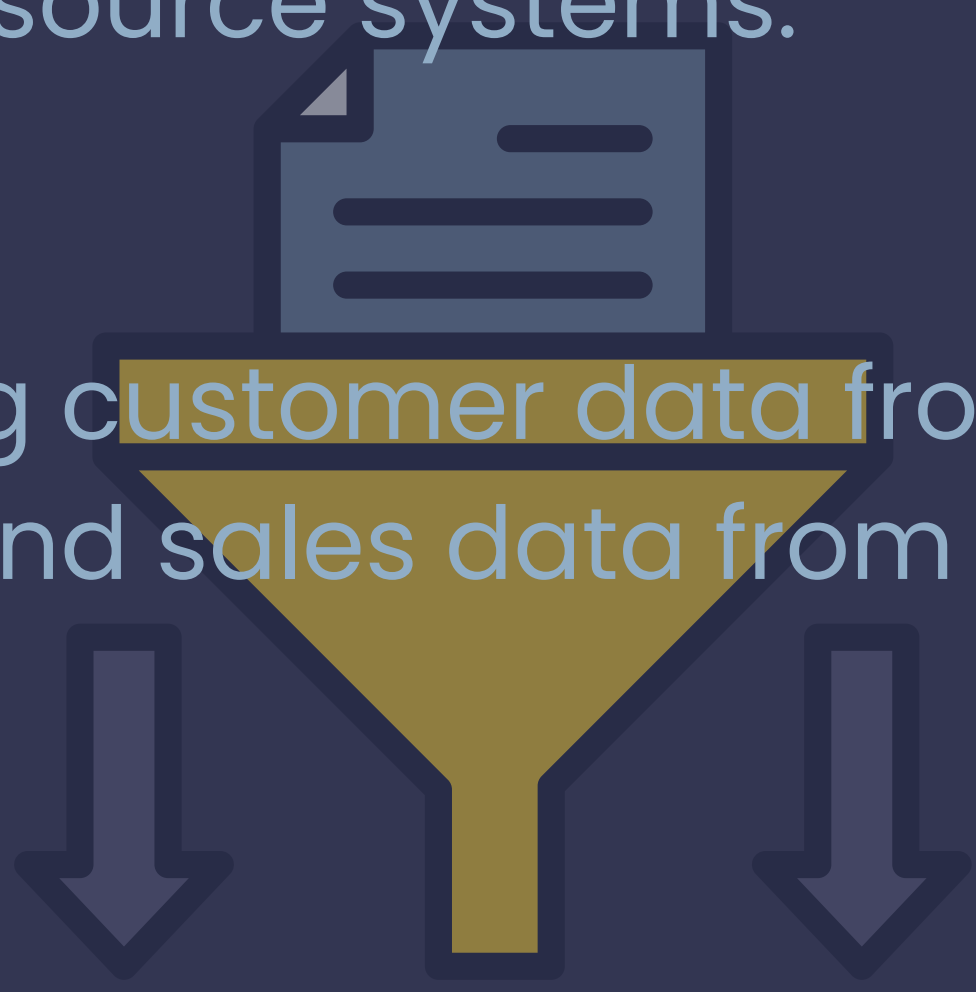
A central repository of integrated data from multiple sources, designed to support decision making and business intelligence activities.

A healthcare organization uses a data warehouse to integrate patient data from various departments for comprehensive reporting and analysis.

Shwetank Singh
GritSetGrow - GSGLearn.com

Extract

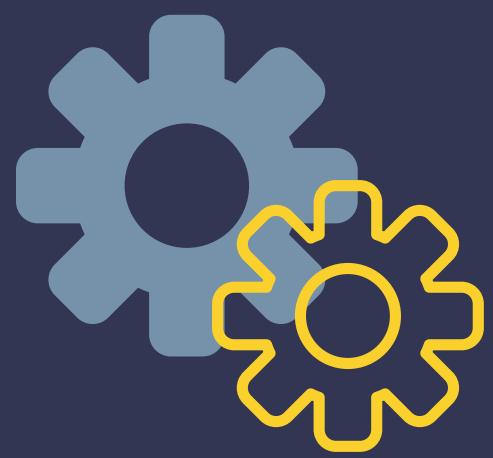
The process of retrieving data from different source systems.



Extracting customer data from a CRM system and sales data from an ERP system.

Shwetank Singh
GritSetGrow - GSGLearn.com

Transform



The process of converting extracted data into a format that can be loaded into the data warehouse, including cleaning, conforming, and integrating data from multiple sources.

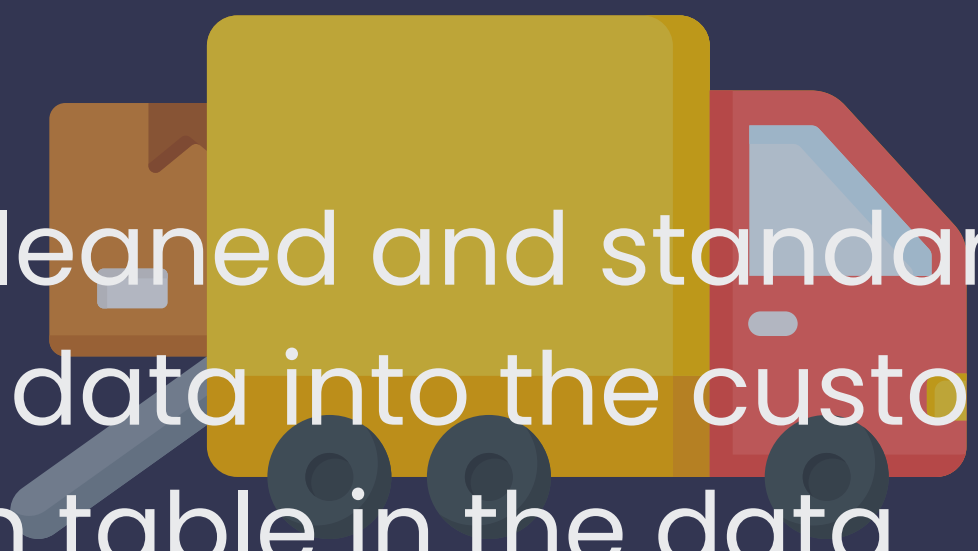
Standardizing date formats and removing duplicates from customer records before loading them into the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Load

The process of loading transformed data into the target data warehouse.



Loading cleaned and standardized customer data into the customer dimension table in the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Staging Area

An intermediate storage area used for data processing during the ETL process.

Temporarily storing raw sales data extracted from an ERP system before transformation.

Shwetank Singh
GritSetGrow - GSGLearn.com

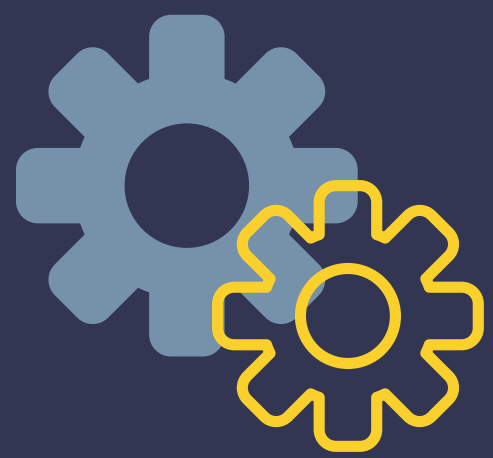
Data Profiling

Analyzing source data to assess its quality, completeness, and fitness for purpose.

Using a data profiling tool to find missing values, outliers, and inconsistencies in the customer data from the CRM system.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Cleaning

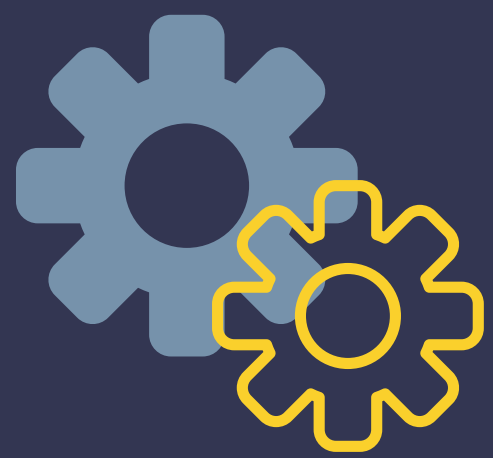


The process of identifying and fixing errors and omissions in the data.

Standardizing addresses in customer data by correcting misspellings and formatting issues.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Conforming



Resolving labeling conflicts between potentially incompatible data sources so that they can be used together in the data warehouse.

Conforming product category names from different source systems to a standard set of categories used in the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Change Data Capture

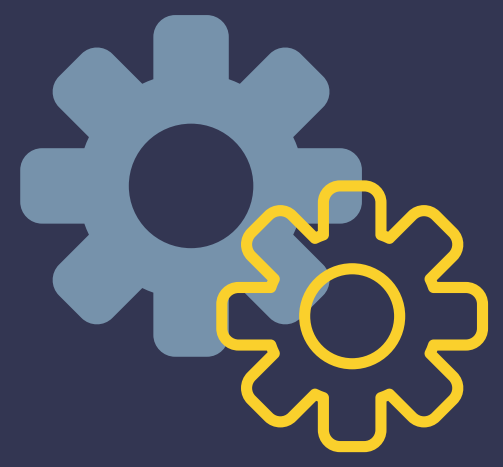
Techniques for capturing changes in the source data to keep the data warehouse up-to-date.

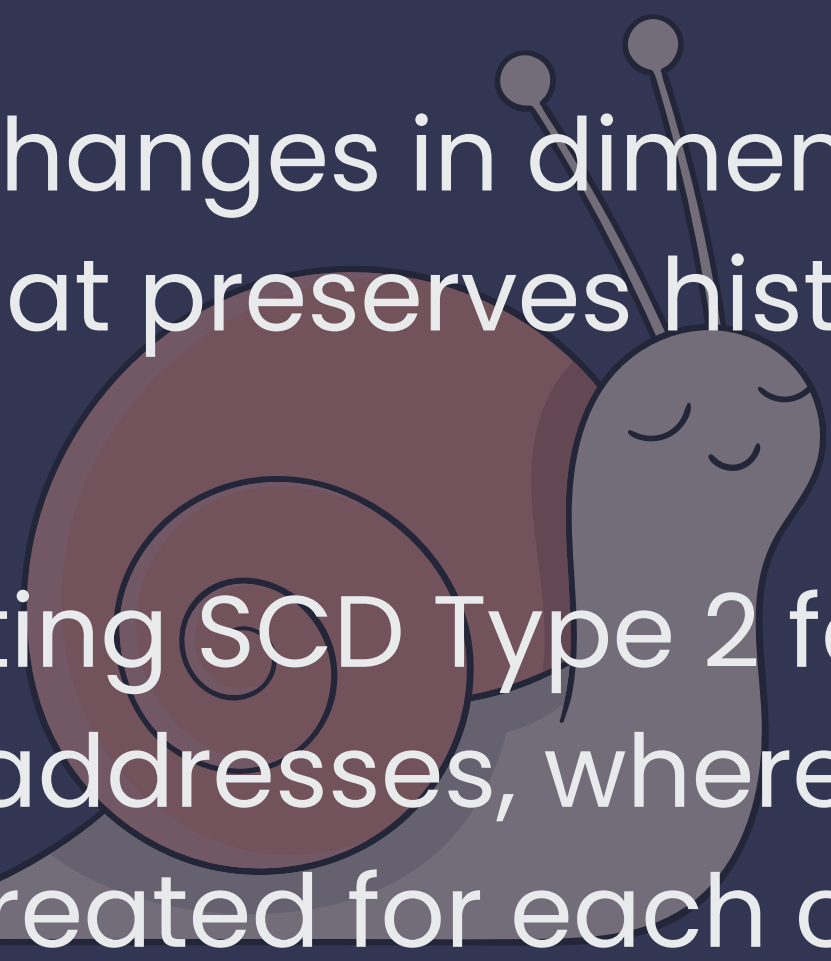
Implementing triggers in the CRM system to capture inserts, updates, and deletes and storing these changes in a staging table.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Slowly Changing Dimensions (SCD)



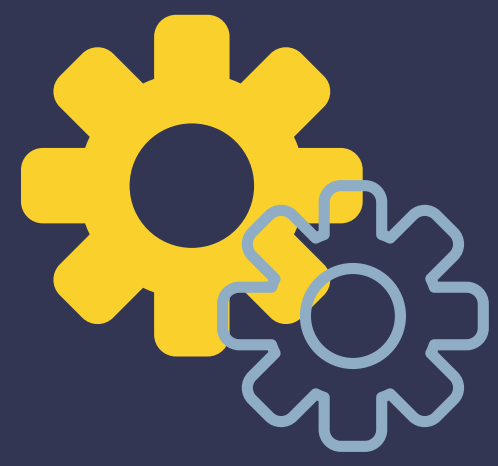


Handling changes in dimension data in a way that preserves historical data.

Implementing SCD Type 2 for customer addresses, where a new record is created for each change, preserving the history of address changes.

Shwetank Singh
GritSetGrow - GSGLearn.com

Surrogate Key Assignment



Assigning unique keys to records in the dimension tables to avoid using business keys from the source systems.

Generating surrogate keys for customer records in the data warehouse to replace the customer IDs from the CRM system.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

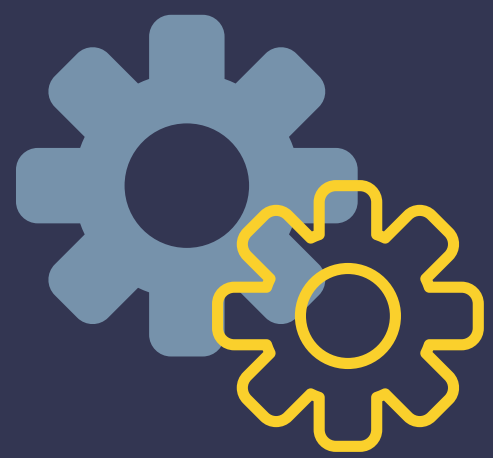
Fact Table Loading

The process of loading transactional data into fact tables, often involving complex transformations and calculations.

Loading sales transactions into the sales fact table, including calculations for total sales amount and discount applied.

Shwetank Singh
GritSetGrow - GSGLearn.com

ETL Process Automation



Automating the ETL process to run at scheduled intervals or in response to specific events.

Setting up a workflow in an ETL tool to automatically extract, transform, and load sales data from the source systems to the data warehouse every night.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Logical Data Map

A document that ties the beginning of the ETL system to the end, describing the relationship between source fields and destination fields.

A logical data map for a sales data warehouse showing how sales records from different source systems map to the unified sales table in the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Discovery

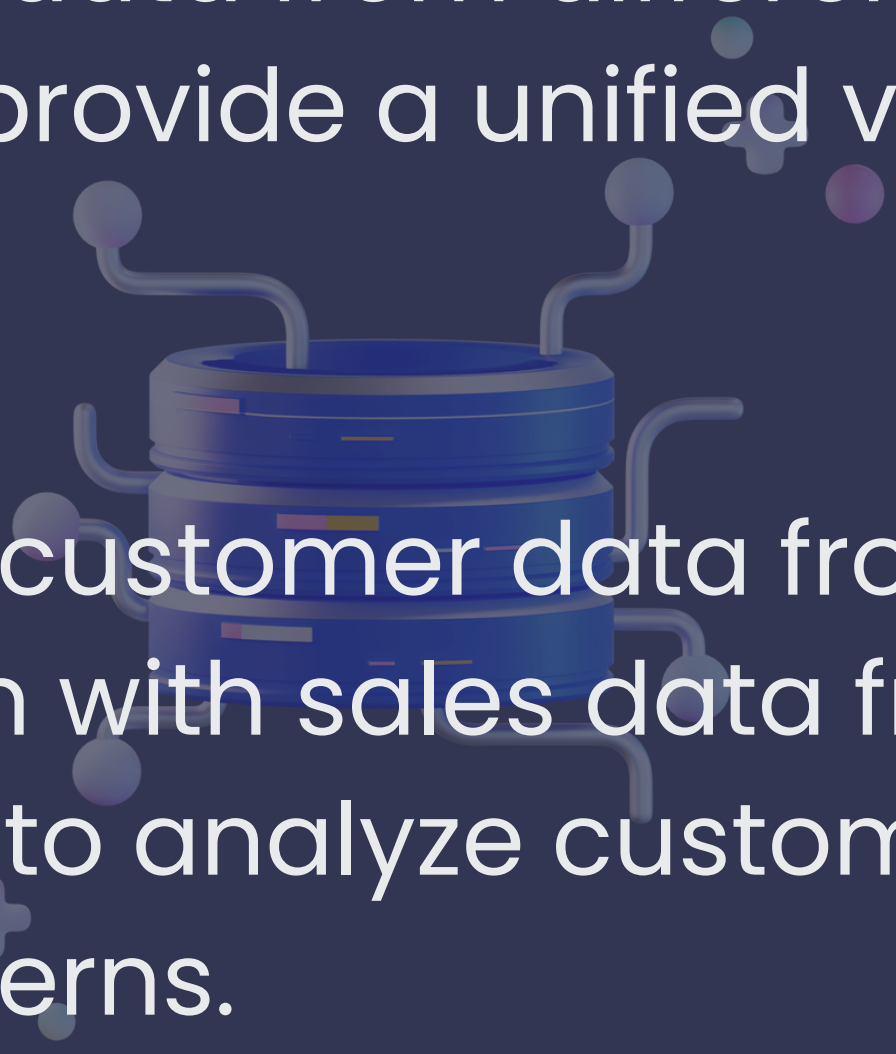
The phase where source systems are identified and analyzed to determine the required source data for the data warehouse.

Identifying the CRM system as the primary source for customer data and analyzing its data structure.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Integration

Combining data from different sources to provide a unified view.

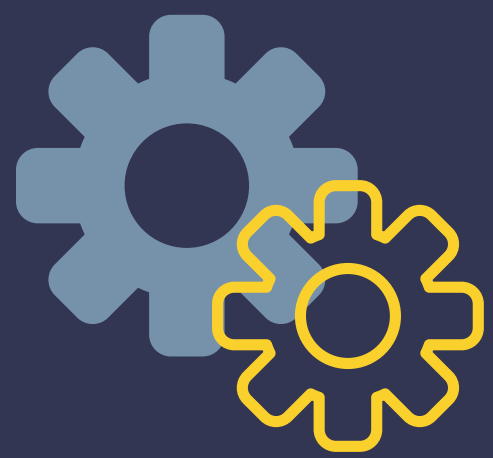


Integrating customer data from the CRM system with sales data from the ERP system to analyze customer buying patterns.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Metadata Management



Managing data about data, providing information about the source, transformation, and destination of data.

Documenting the source, transformation rules, and target tables for sales data in the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Lineage

Tracking the flow of data from source to destination, showing how data has been transformed.

Tracing a sales figure in a report back to the original transaction in the point-of-sale system.

Shwetank Singh
GritSetGrow - GSGLearn.com

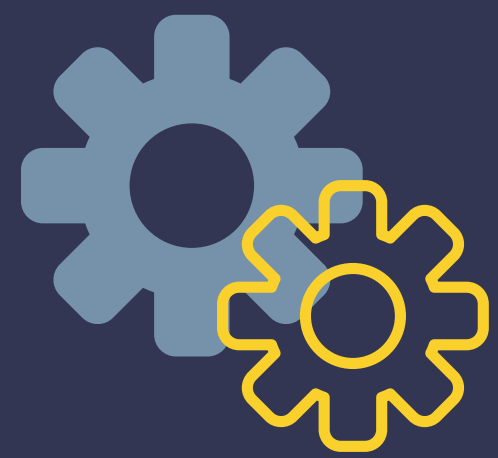
Data Quality

Ensuring that the data in the data warehouse is accurate, complete, and reliable.

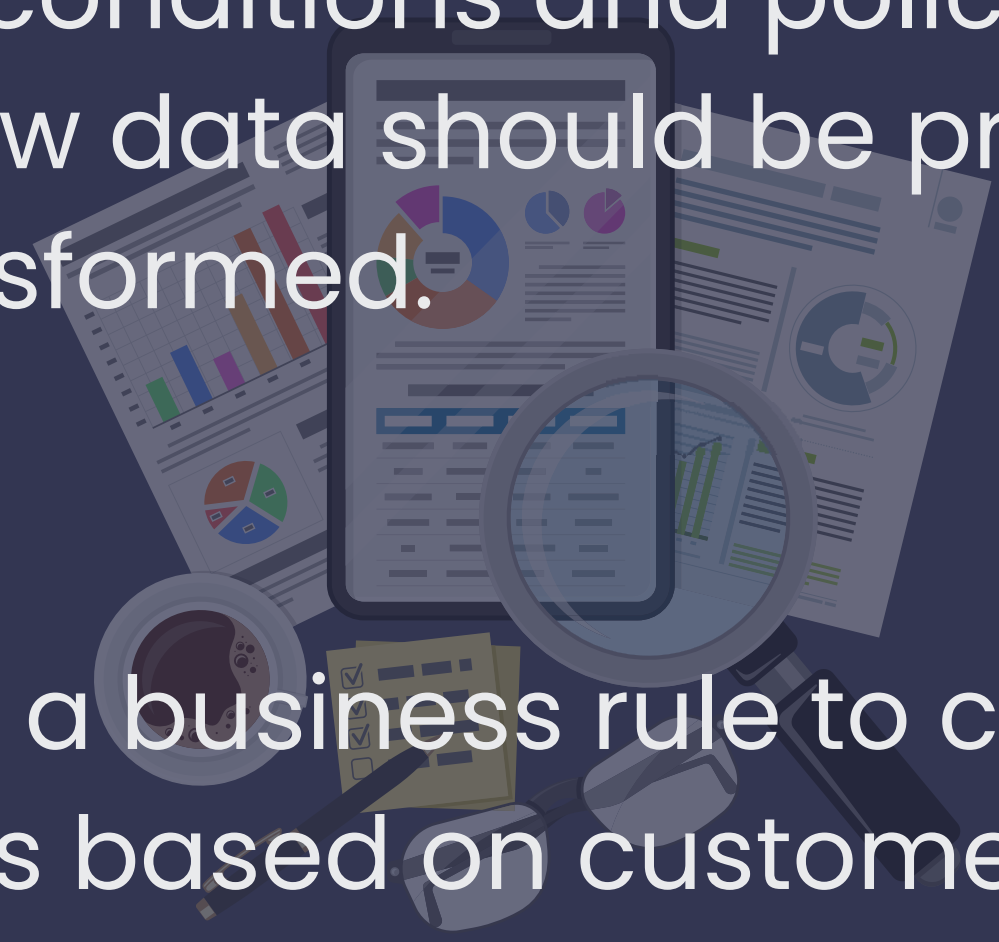
Implementing data validation rules to check for missing values and incorrect data types in customer records.

Shwetank Singh
GritSetGrow - GSGLearn.com

Business Rules



Specific conditions and policies that guide how data should be processed and transformed.



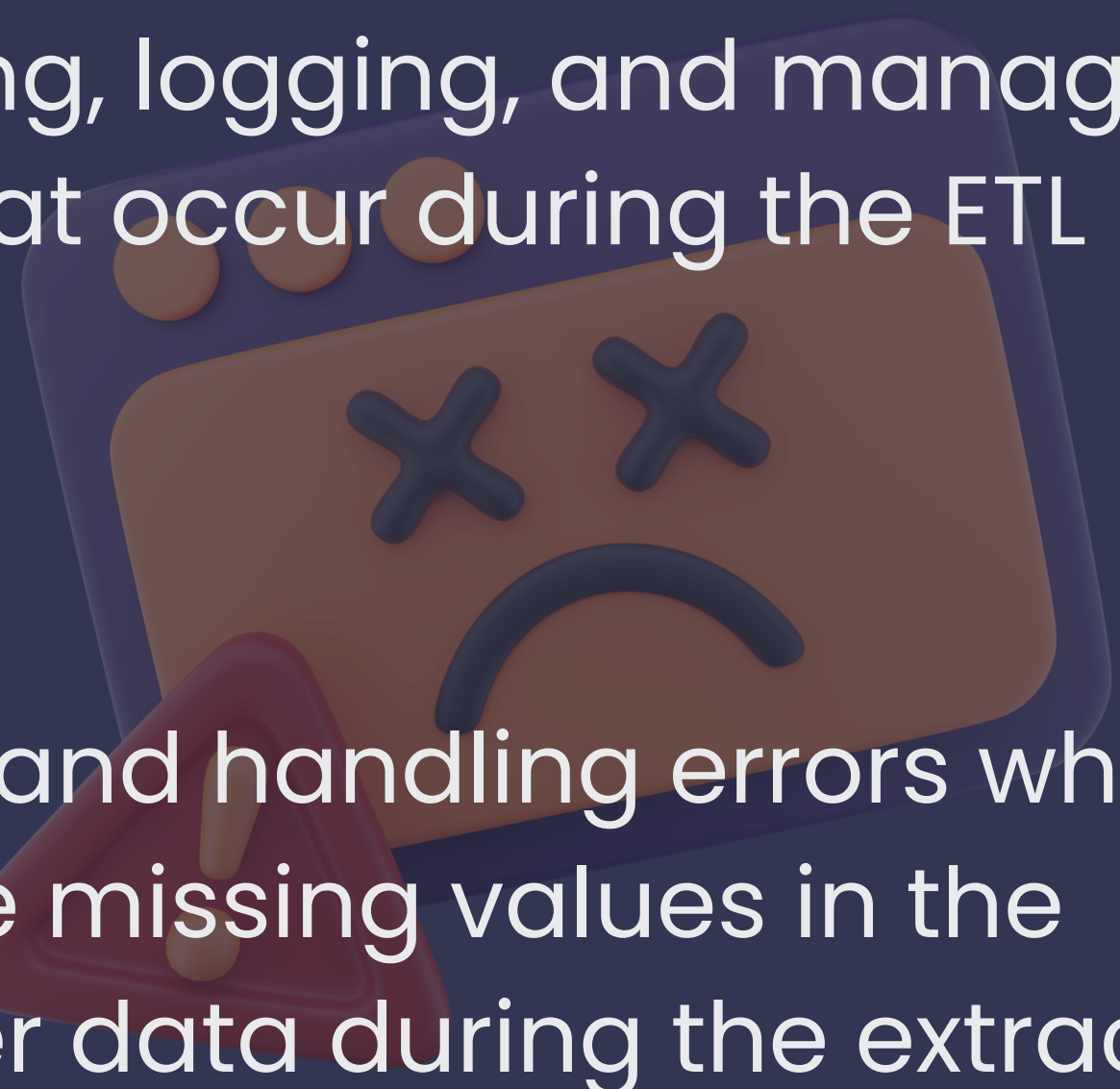
Applying a business rule to calculate discounts based on customer loyalty status during the ETL process.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Error Handling

Identifying, logging, and managing errors that occur during the ETL process.



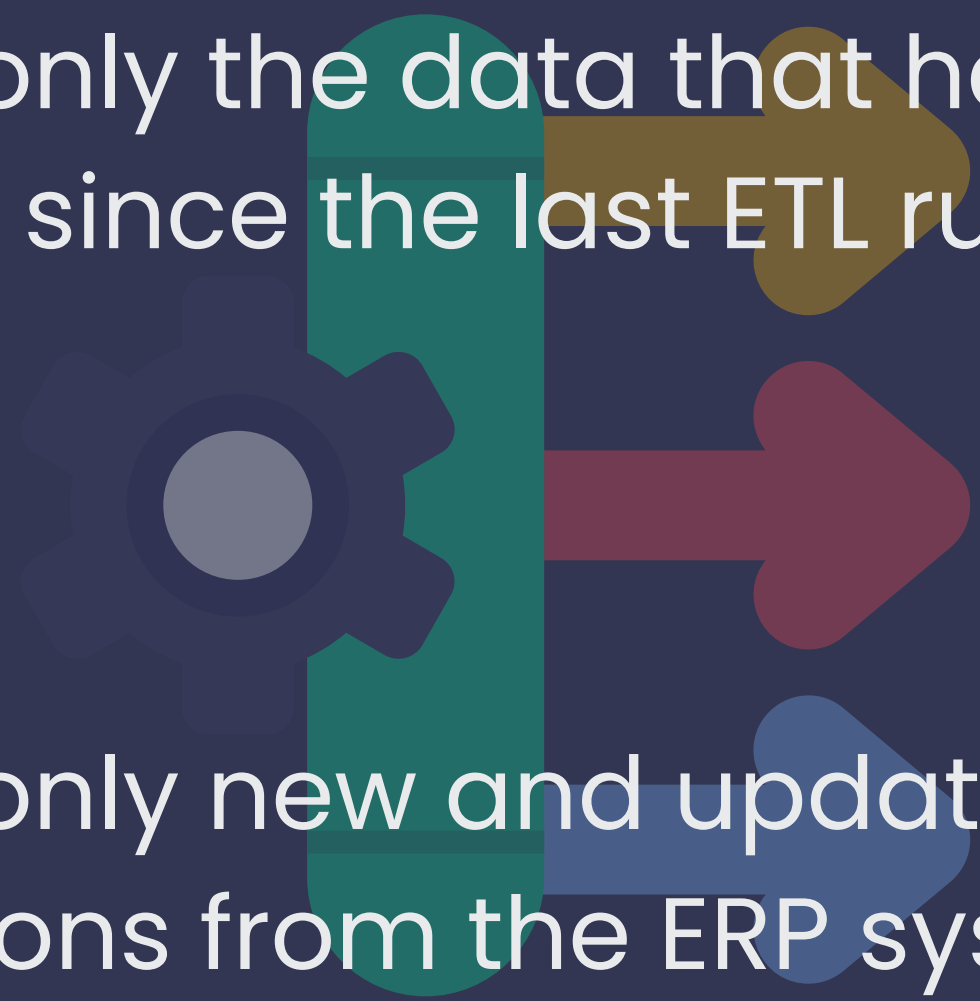
Logging and handling errors when there are missing values in the customer data during the extraction process.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Incremental Load

Loading only the data that has changed since the last ETL run.

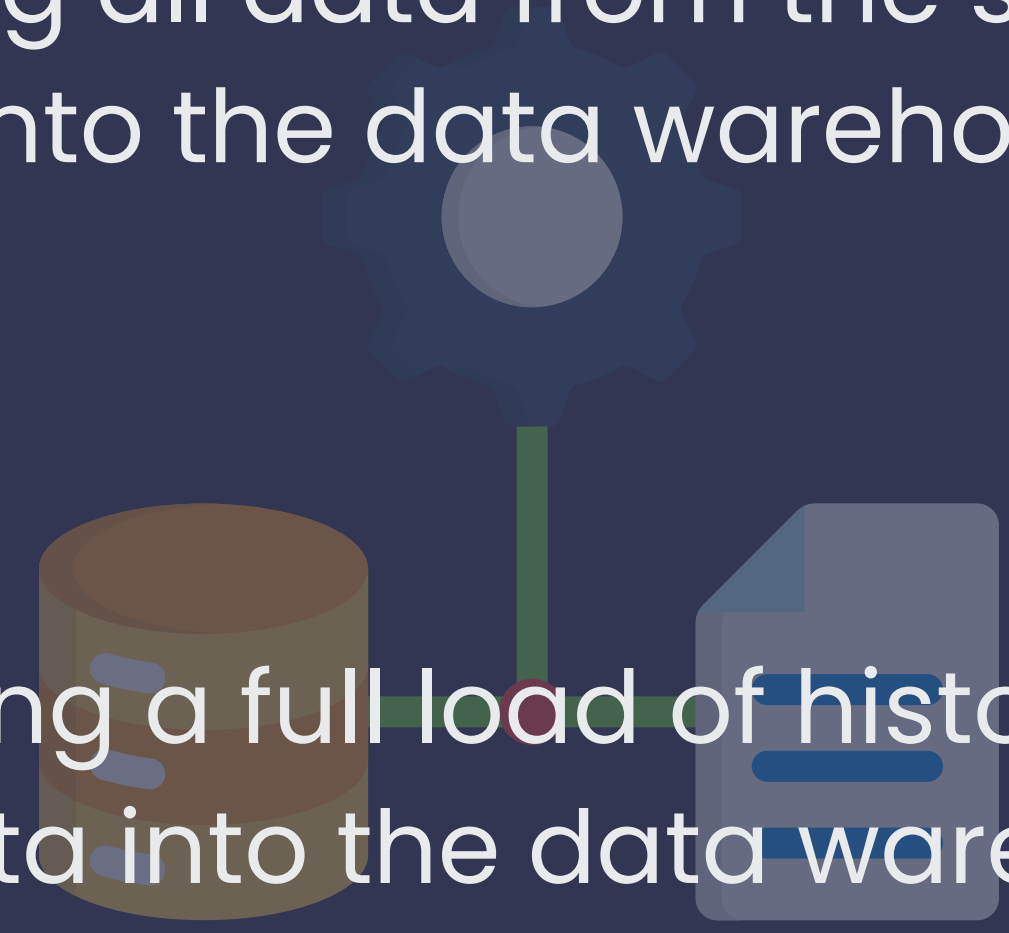


Loading only new and updated sales transactions from the ERP system into the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Full Load

Reloading all data from the source system into the data warehouse.



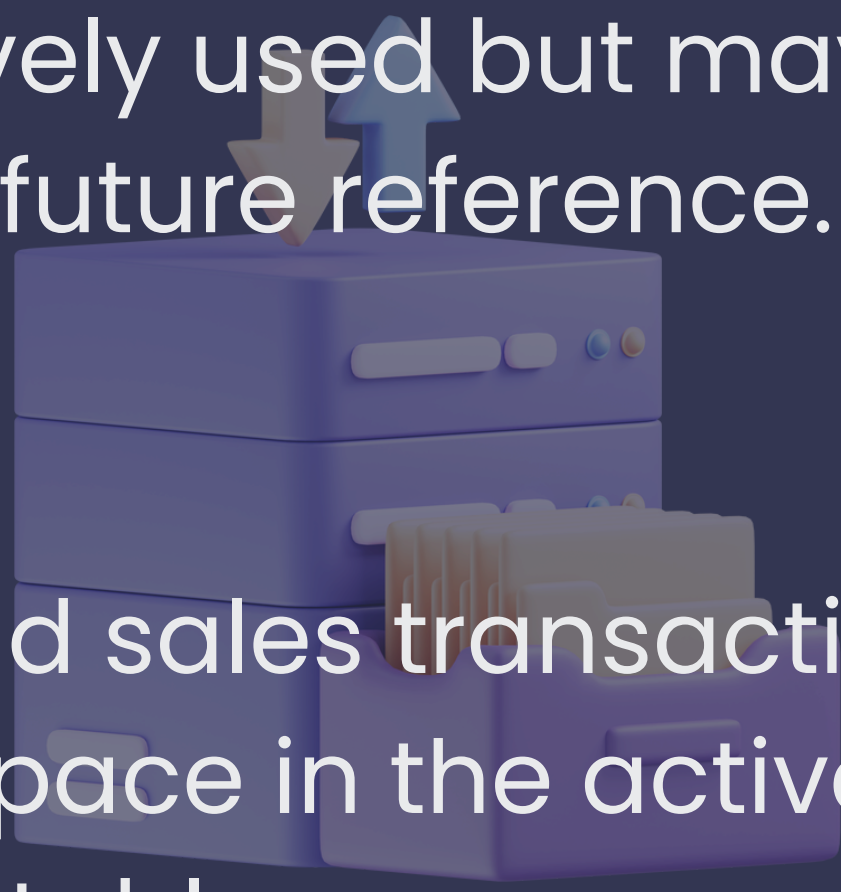
Performing a full load of historical sales data into the data warehouse during the initial ETL process.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Data Archiving

Storing historical data that is no longer actively used but may be needed for future reference.



Archiving old sales transaction data to free up space in the active data warehouse tables.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Data Purging

Removing data that is no longer needed from the data warehouse.

Deleting customer records that have been inactive for more than ten years.

Shwetank Singh
GritSetGrow - GSGLearn.com

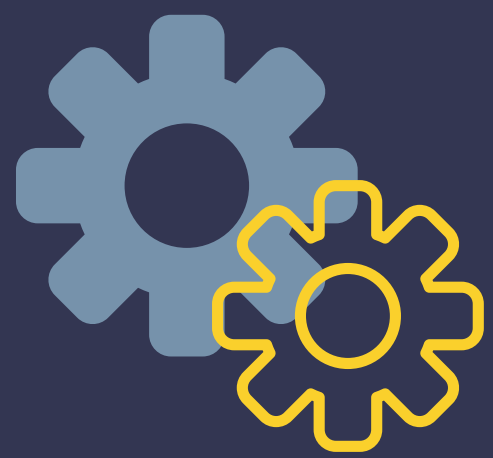
Data Backup

Creating copies of data to protect against data loss.

Regularly backing up the data warehouse to an off-site storage location.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Recovery



Restoring data from backups in case of data loss or corruption.

Restoring the data warehouse from a backup after a hardware failure.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Data Anonymization

Removing or masking personally identifiable information (PII) in the data warehouse.

Anonymizing customer names and addresses in a dataset used for public analysis.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Encryption

Protecting data by converting it into a secure format that can only be read with the proper decryption key.

Encrypting sensitive customer data before loading it into the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Masking

Hiding sensitive data by replacing it with fictional but realistic data.

Masking credit card numbers in the data warehouse to protect against unauthorized access.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

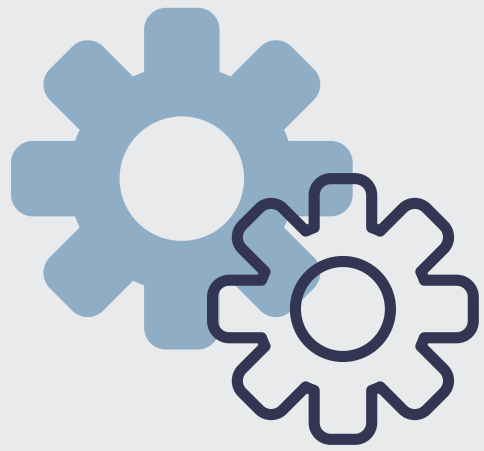
Data Transformation

Converting data from one format or structure to another to meet the requirements of the data warehouse.

Transforming date formats from MM/DD/YYYY to YYYY-MM-DD during the ETL process.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Validation



Checking data for accuracy and consistency before loading it into the data warehouse.

Validating email addresses in customer records to ensure they are in the correct format.

Shwetank Singh
GritSetGrow - GSGLearn.com

ETL Logging

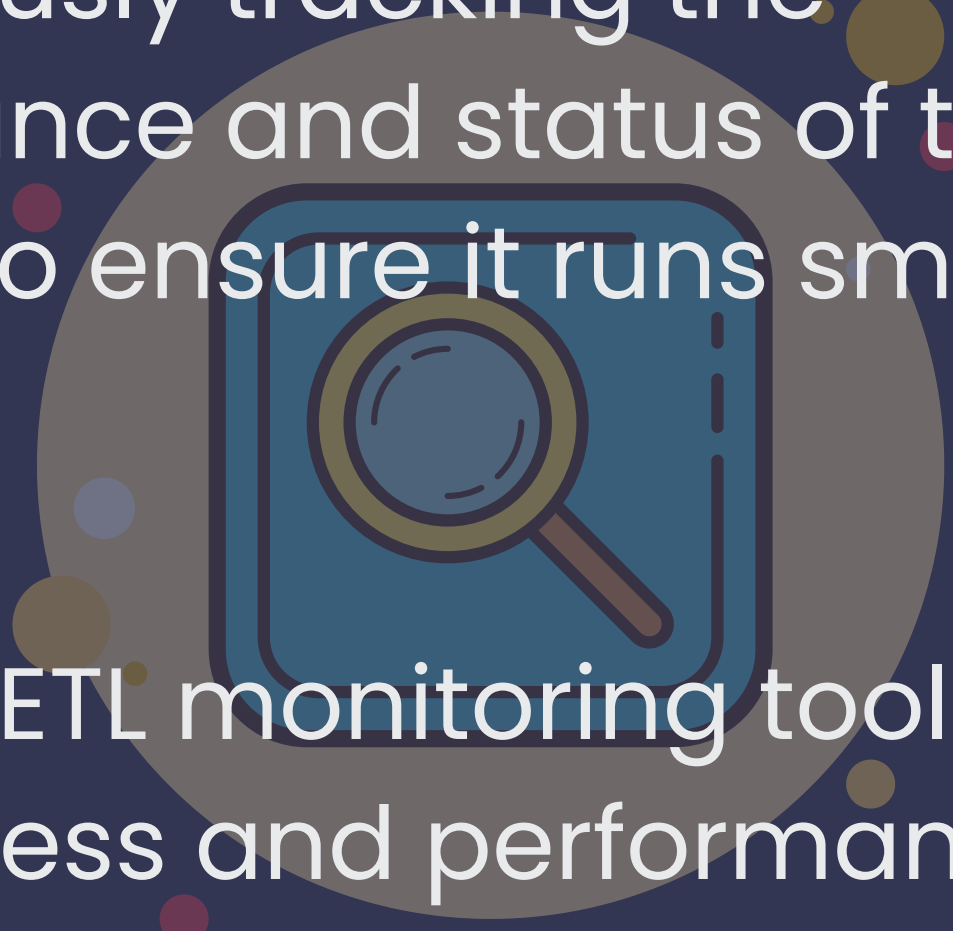
Recording the steps and events that occur during the ETL process for monitoring and debugging purposes.

Keeping a log of all extraction, transformation, and load activities, including any errors encountered.

Shwetank Singh
GritSetGrow - GSGLearn.com

ETL Monitoring

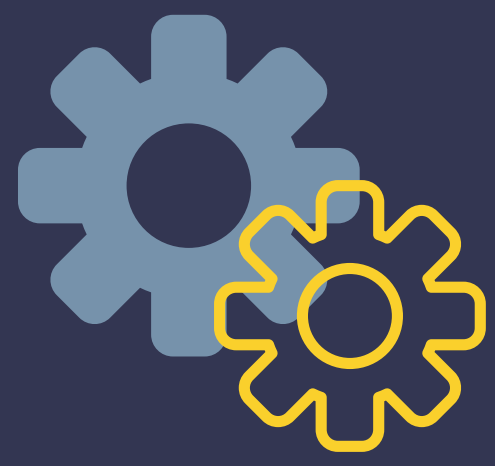
Continuously tracking the performance and status of the ETL process to ensure it runs smoothly.



Using an ETL monitoring tool to track the progress and performance of daily data loads.

Shwetank Singh
GritSetGrow - GSGLearn.com

ETL Performance Tuning



Optimizing the ETL process to improve speed and efficiency.

Adjusting the ETL process to reduce the time it takes to load sales data by optimizing SQL queries and parallelizing tasks.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Data Mart

A subset of the data warehouse that focuses on a specific area or department within the organization.

Creating a sales data mart that includes only sales-related data for the sales department.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

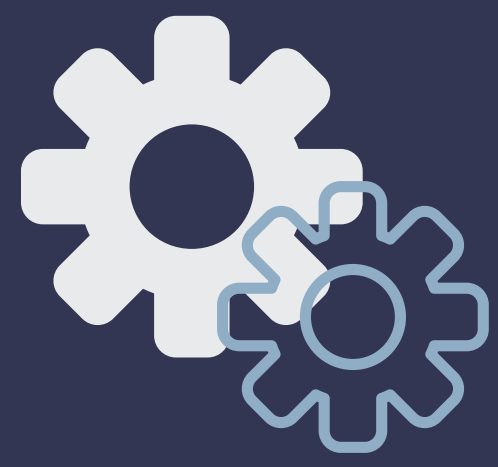
Data Lake

A centralized repository that allows you to store all your structured and unstructured data at any scale.

Storing raw and unprocessed data from various sources in a data lake for later analysis.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Integration Tools



Software tools that help to combine data from different sources and provide a unified view.

Using Informatica or Talend for integrating data from multiple source systems into the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

ETL Tool

Software tools that facilitate the extraction, transformation, and loading of data into the data warehouse.

Using tools like Apache Nifi or Microsoft SSIS to automate and manage the ETL process.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Data Governance

The management of data availability, usability, integrity, and security in the data warehouse.

Implementing data governance policies to ensure data quality and compliance in the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Stewardship

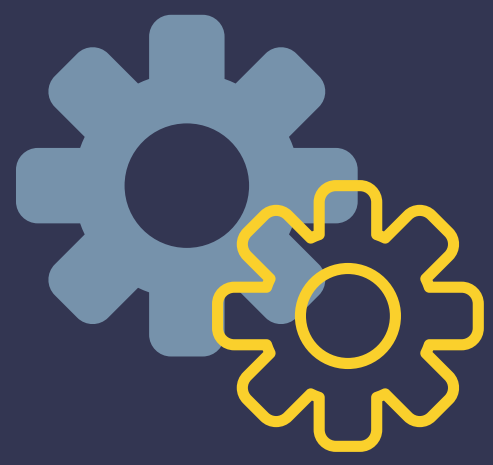
The management and oversight of an organization's data assets to help provide business users with high-quality data that is easily accessible.

Assigning data stewards to oversee data quality and data management practices in the organization.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Data Warehouse Architecture



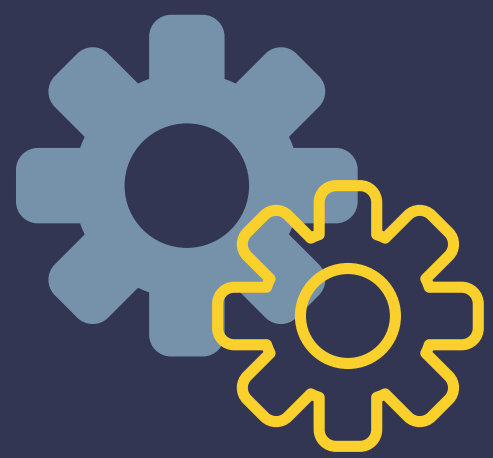
The design and structure of a data warehouse, including its components and their relationships.

Designing a data warehouse architecture that includes a staging area, ETL process, and presentation layer.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Dimensional Modeling



A data modeling technique optimized for data warehouse and OLAP cube implementations.

Designing a star schema or snowflake schema for the sales data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Star Schema

A type of database schema that is composed of a single fact table and multiple dimension tables.

Designing a star schema with a central sales fact table connected to dimension tables for products, customers, and time.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Snowflake Schema

A type of database schema that is composed of a fact table and multiple dimension tables, where the dimension tables are normalized.

Designing a snowflake schema where the product dimension table is normalized into separate tables for product categories and subcategories.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Fact Table

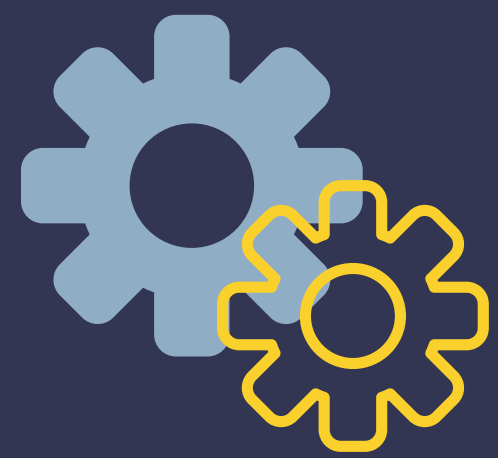
A central table in a star or snowflake schema that contains the numerical measures of a business process.

Creating a sales fact table that records sales transactions, including quantities sold and sales amounts.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Dimension Table



A table in a star or snowflake schema that contains descriptive attributes related to the dimensions of the business process.

Creating a customer dimension table that includes attributes like customer name, address, and contact information.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Aggregation

The process of summarizing detailed data for analysis and reporting.

Aggregating daily sales data into monthly sales summaries for trend analysis.

Shwetank Singh
GritSetGrow - GSGLearn.com

Granularity

The level of detail represented by the data in the data warehouse.

Deciding on the granularity of the sales fact table, such as recording sales at the transaction level versus daily summaries.

Shwetank Singh
GritSetGrow - GSGLearn.com

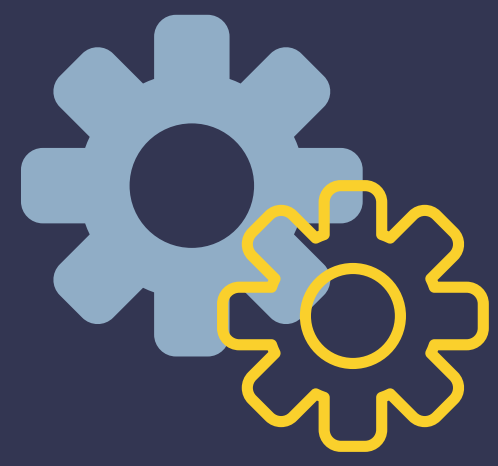
Historical Data

Data that represents past events and is stored in the data warehouse for analysis.

Storing historical sales data for the past ten years to analyze long-term trends.

Shwetank Singh
GritSetGrow - GSGLearn.com

Real-Time Data Warehousing



The process of updating the data warehouse in real-time as new data becomes available.

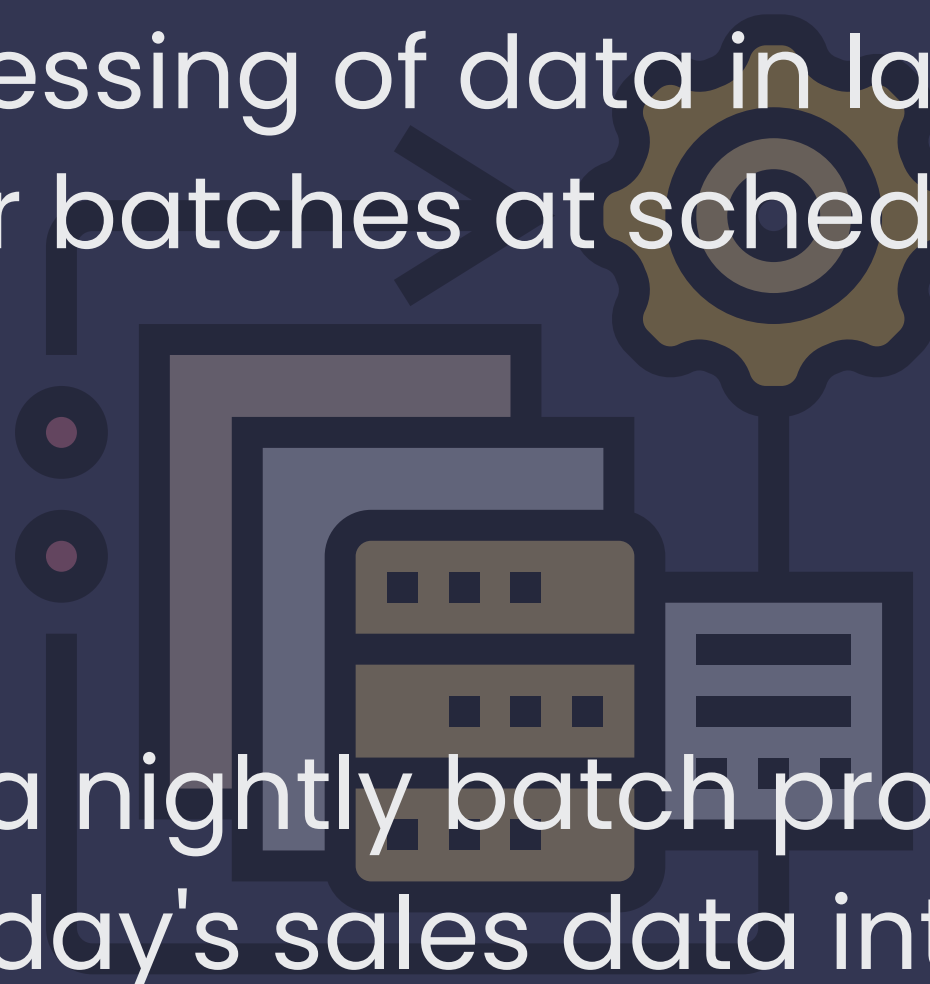
Implementing a real-time ETL process to load streaming data from IoT devices into the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Batch Processing

The processing of data in large groups or batches at scheduled intervals.



Running a nightly batch process to load the day's sales data into the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Parallel Processing

The simultaneous processing of multiple tasks to speed up the ETL process.

Using parallel processing to transform and load multiple data sources concurrently.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Data Lakehouse

An architecture that combines the benefits of data lakes and data warehouses.

Implementing a data lakehouse to store both structured and unstructured data for comprehensive analysis.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Data Vault Modeling

A database modeling method designed to provide long-term historical storage of data from multiple operational systems.

Using data vault modeling to capture and store all historical changes in the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

Data Quality Dimensions

Attributes used to measure data quality, including accuracy, completeness, consistency, timeliness, and validity.

Assessing the accuracy and completeness of customer data in the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

ETL Metadata

Data about the ETL process, including source-to-target mappings, data lineage, and transformation rules.

Documenting the transformation rules and source-to-target mappings in the ETL metadata repository.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

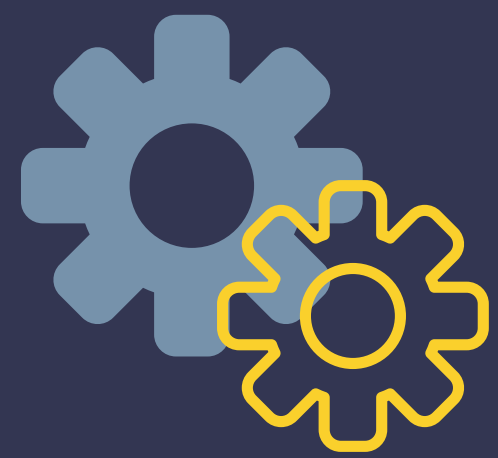
ETL Orchestration

Coordinating and managing the execution of multiple ETL processes.

Using an ETL orchestration tool like Apache Airflow to manage the workflow of data extraction, transformation, and loading tasks.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Source



The system or location from which data is extracted for the ETL process.

Extracting customer data from a CRM system and sales data from an ERP system.

Shwetank Singh
GritSetGrow - GSGLearn.com

Target System

The system or location where transformed data is loaded during the ETL process.

Loading cleaned and transformed data into the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

ETL Scheduler

A tool that manages the timing and execution of ETL processes.

Using an ETL scheduler to run data extraction and loading jobs at specified times.

Shwetank Singh
GritSetGrow - GSGLearn.com

ETL Job

A single task or unit of work in the ETL process.

Creating an ETL job to extract customer data from the CRM system.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Transformation Rules

Specific rules and logic applied to data during the transformation process.

Defining transformation rules to convert all product prices to a common currency.

Shwetank Singh
GritSetGrow - GSGLearn.com

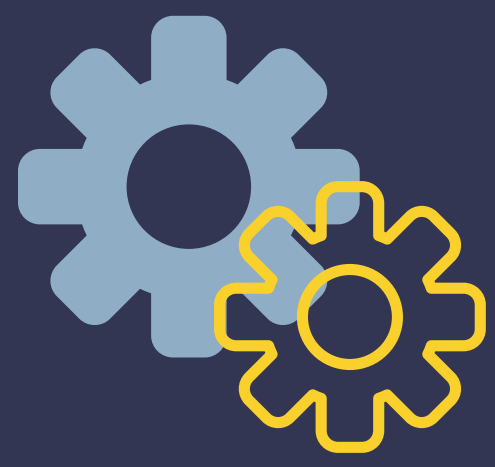
ETL Testing

The process of verifying that the ETL process is working correctly and that the data is accurate and complete.

Conducting ETL testing to ensure that all data transformations have been applied correctly and that the loaded data matches the source data.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Validation Rules



Specific criteria and checks to ensure data validity during the ETL process.

Implementing validation rules to ensure that email addresses are in the correct format and that date fields contain valid dates.

Shwetank Singh
GritSetGrow - GSGLearn.com

ETL Resiliency

The ability of the ETL process to recover from failures and continue processing.

Implementing checkpointing in ETL processes to allow for restarting from the last successful step after a failure.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Engineering 101 - ETL Terminology

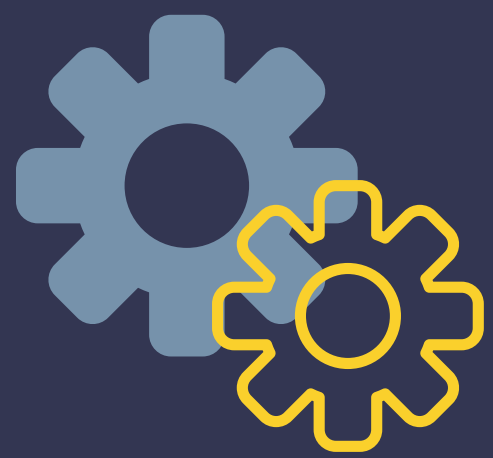
ETL Load Balancing

Distributing ETL workloads across multiple servers or processes to optimize performance.

Implementing load balancing to distribute data transformation tasks across a cluster of ETL servers.

Shwetank Singh
GritSetGrow - GSGLearn.com

ETL Audit Trail



A record of all changes and transformations applied to data during the ETL process.

Maintaining an audit trail to track all data transformations and changes for compliance and debugging purposes.

Shwetank Singh
GritSetGrow - GSGLearn.com

THANK YOU