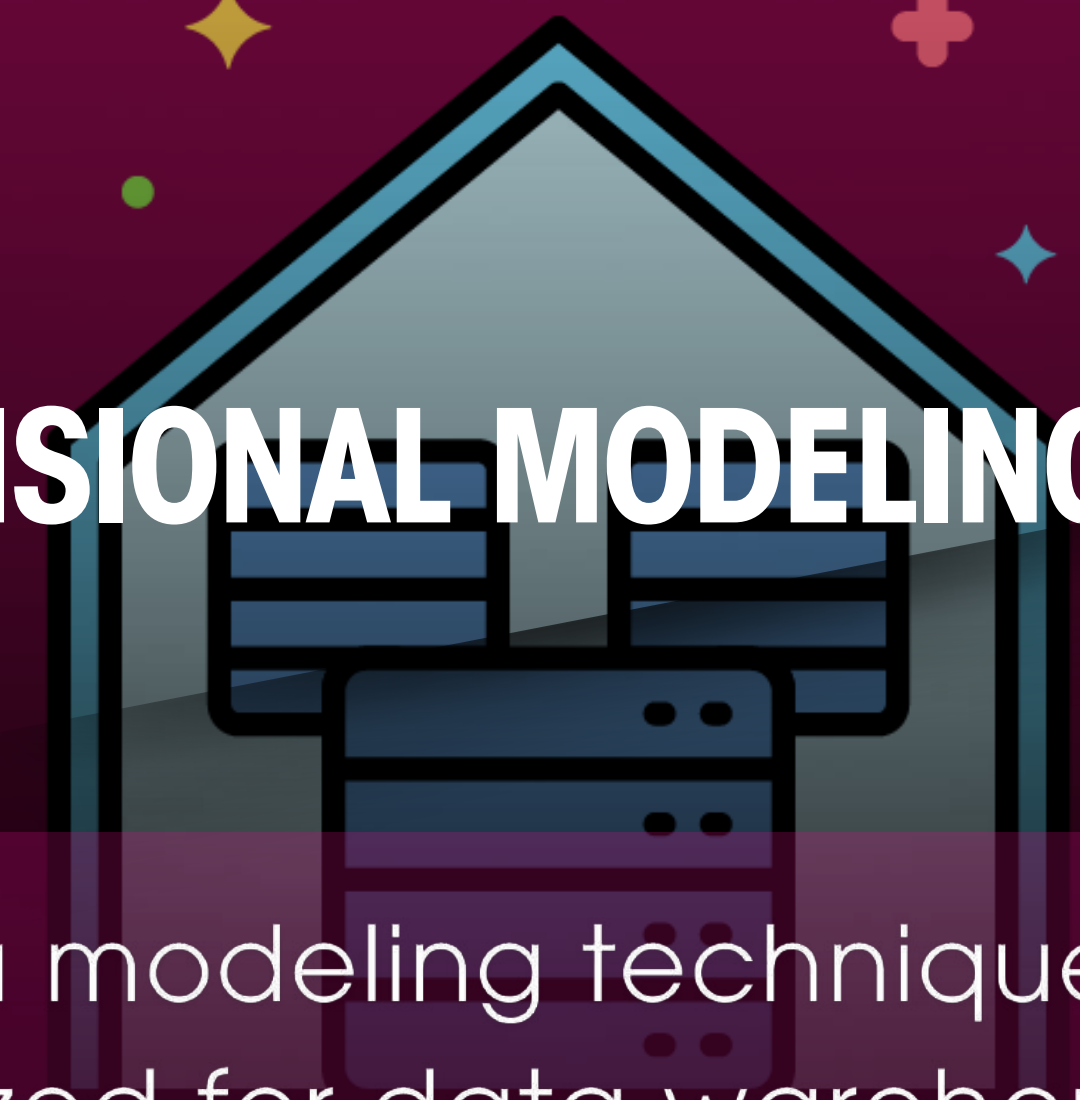
DATA WAREHOUSING

DIMENSIONAL MODELLING

Shwetank Singh
GritSetGrow - GSGLearn.com

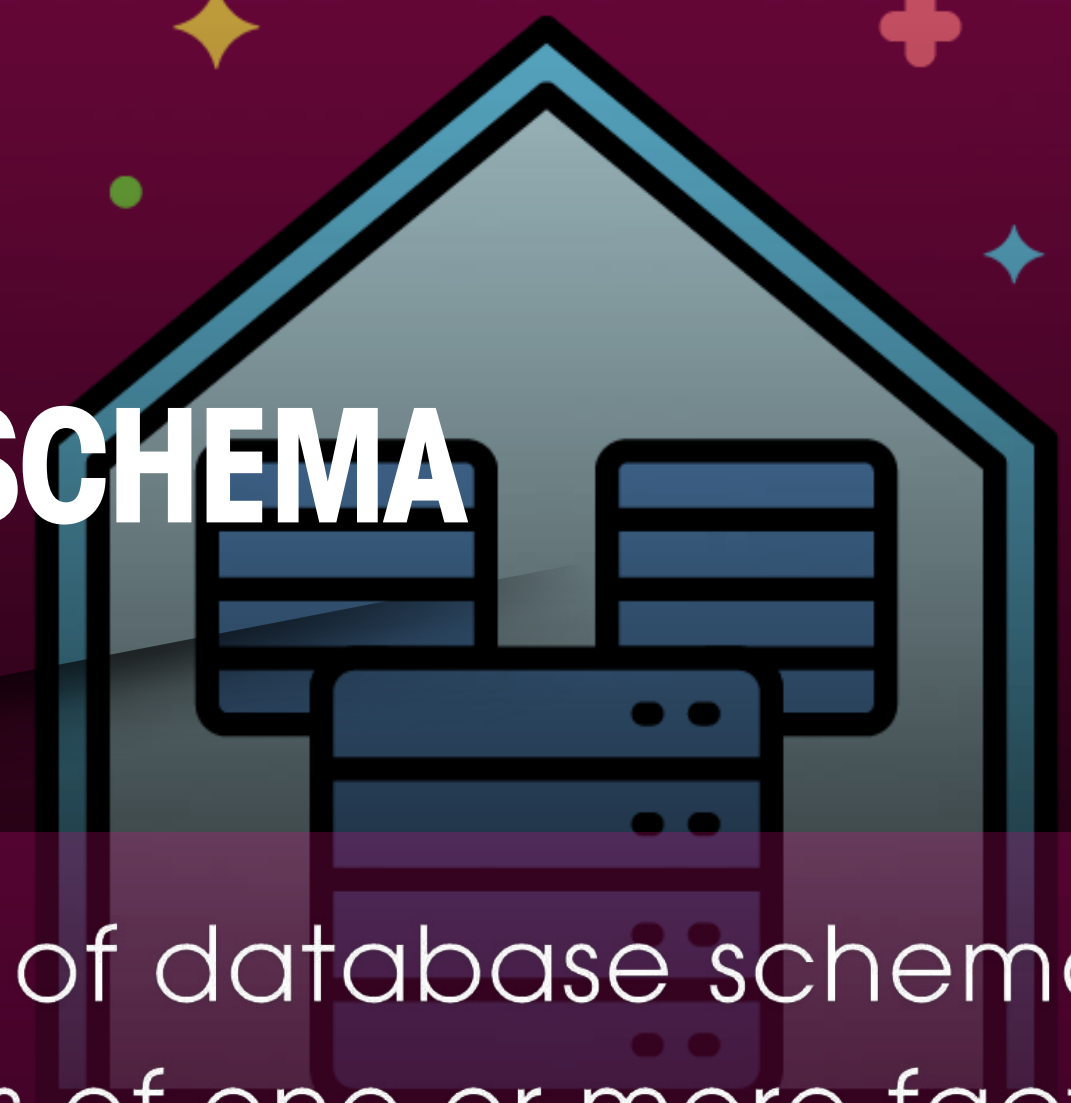


DIMENSIONAL MODELING

A data modeling technique optimized for data warehousing and decision support systems, focusing on ease of querying.

Using star and snowflake schemas to organize data in a data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com



STAR SCHEMA

A type of database schema that consists of one or more fact tables referencing any number of dimension tables.

A sales database where a fact table contains sales data, and dimension tables include information about products, customers, and time.

Shwetank Singh
GritSetGrow - GSGLearn.com

SNOWFLAKE SCHEMA

A more complex version of the star schema where dimension tables are normalized into multiple related tables.

A sales database where dimension tables such as products and customers are further split into related tables like product categories and customer demographics.

Shwetank Singh
GritSetGrow - GSGLearn.com

FACT TABLE

Central table in a star schema that contains quantitative data for analysis.

A sales fact table that includes measures like sales revenue, quantity sold, and discount applied.

Shwetank Singh
GritSetGrow - GSGLearn.com

DIMENSION TABLE

Tables that contain descriptive attributes (dimensions) related to the facts.

A product dimension table that includes product names, categories, and descriptions.

Shwetank Singh
GritSetGrow - GSGLearn.com

| GRAIN

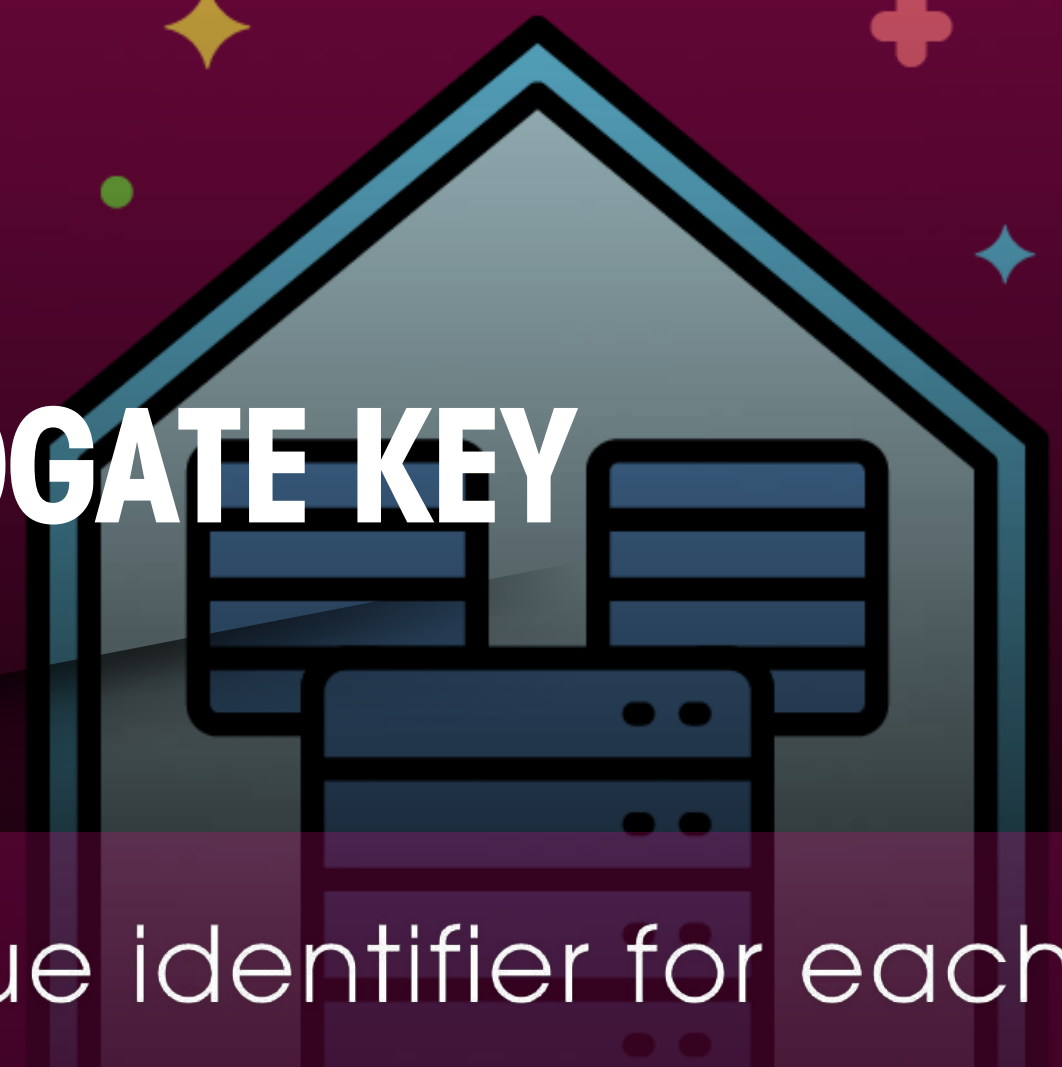
The level of detail or granularity of the data stored in a fact table.

Daily sales data versus monthly sales data in a fact table.

Shwetank Singh

GritSetGrow - GSGLearn.com

SURROGATE KEY



A unique identifier for each row in a dimension table, not derived from application data.

An auto-incremented integer used as the primary key in a customer dimension table.

Shwetank Singh
GritSetGrow - GSGLearn.com

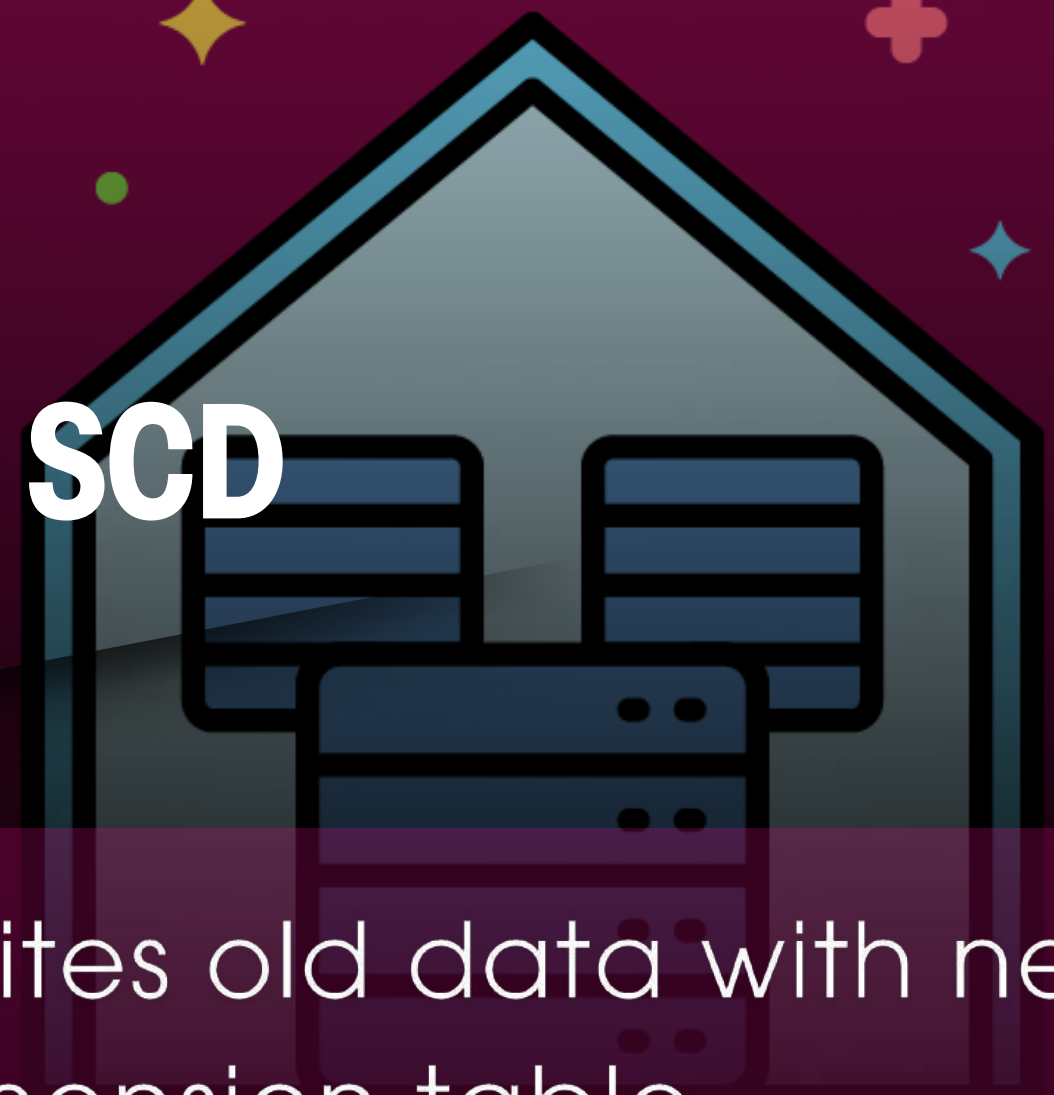
SLOWLY CHANGING DIMENSIONS (SCD)

Techniques for managing and tracking changes in dimension table attributes over time.

Type 1 SCD updates data directly, Type 2 SCD adds new rows, and Type 3 SCD adds new columns to track changes.

Shwetank Singh
GritSetGrow - GSGLearn.com

TYPE 1 SCD



Overwrites old data with new data in a dimension table.

Updating a customer's address directly in the customer dimension table.

Shwetank Singh

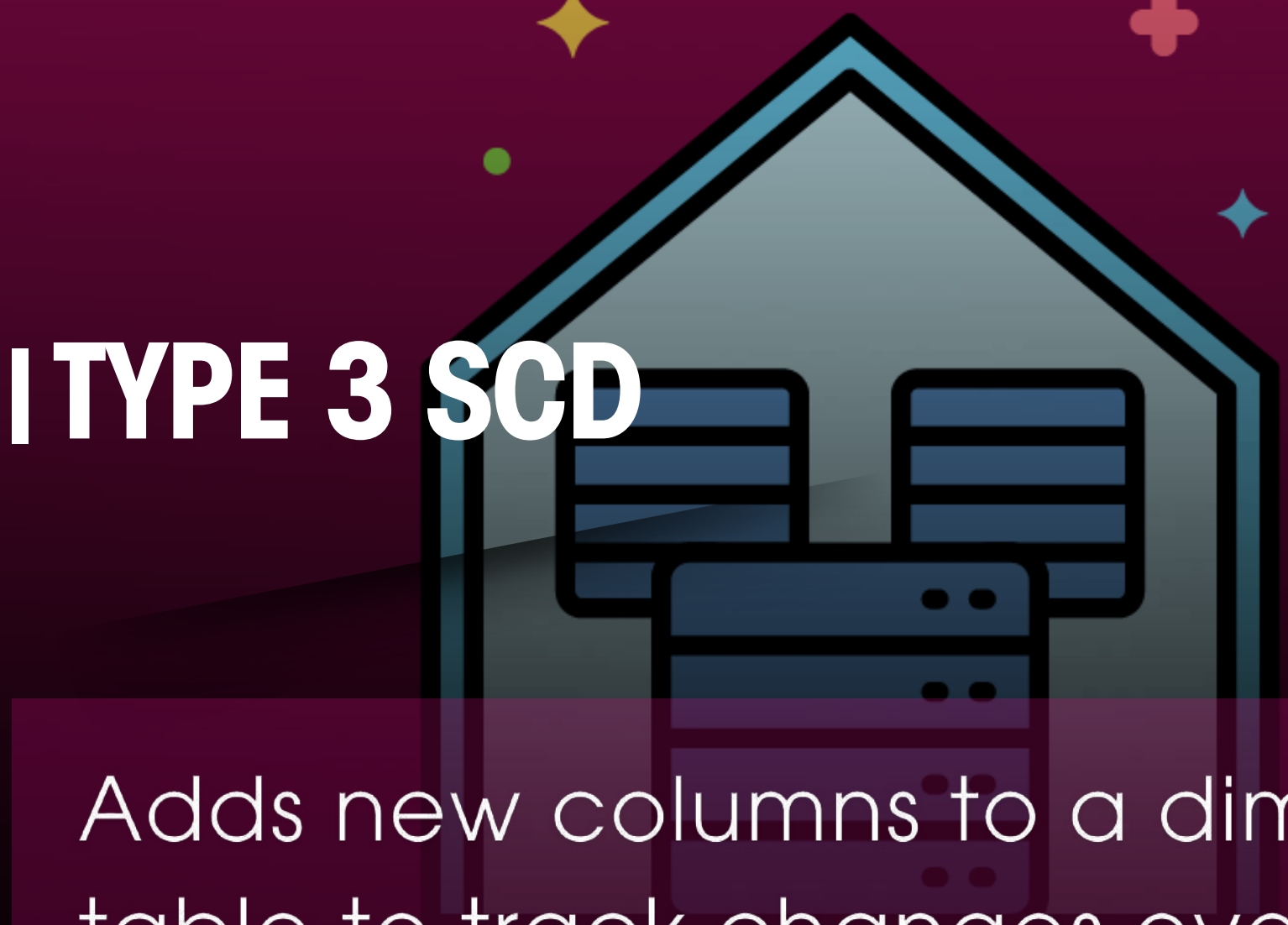
GritSetGrow - GSGLearn.com

TYPE 2 SCD

Creates a new record with a new surrogate key when a change occurs in the dimension data.

Adding a new row for a customer who has moved to a new address, retaining the history of the old address.

Shwetank Singh
GritSetGrow - GSGLearn.com



TYPE 3 SCD

Adds new columns to a dimension table to track changes over time.

Adding columns for "previous address" and "current address" in a customer dimension table.

Shwetank Singh
GritSetGrow - GSGLearn.com

CONFORMED DIMENSIONS

Dimensions that are shared across multiple fact tables and/or data marts.

A date dimension table used across sales, inventory, and finance data marts.

Shwetank Singh
GritSetGrow - GSGLearn.com

JUNK DIMENSION

Combines low-cardinality flags and indicators into a single dimension table.

Combining boolean flags like "is_promotional" and "is_returned" into a single junk dimension table.

Shwetank Singh
GritSetGrow - GSGLearn.com

ROLE-PLAYING DIMENSIONS

A single physical dimension table used in different contexts within the schema.

A date dimension table used for "order date," "ship date," and "delivery date" in a sales schema.

Shwetank Singh
GritSetGrow - GSGLearn.com

FACTLESS FACT TABLE

A fact table that captures events or conditions with no associated numeric facts.

A table capturing student attendance records without any numerical measures.

Shwetank Singh
GritSetGrow - GSGLearn.com

AGGREGATE FACT TABLE

A summarized fact table that improves query performance by reducing the amount of data processed.

A monthly sales summary table that aggregates daily sales data.

Shwetank Singh
GritSetGrow - GSGLearn.com

| BRIDGE TABLE

A table used to handle many-to-many relationships between fact and dimension tables.

A table linking customers to multiple sales regions they belong to.

Shwetank Singh
GritSetGrow - GSGLearn.com

DEGENERATE DIMENSION

A dimension attribute stored in the fact table itself.

An order number stored directly in the sales fact table without a separate dimension table.

Shwetank Singh
GritSetGrow - GSGLearn.com

ETL (EXTRACT, TRANSFORM, LOAD)

The process of extracting data from source systems, transforming it, and loading it into a data warehouse.

Extracting sales data from transactional databases, transforming it to match the warehouse schema, and loading it into the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

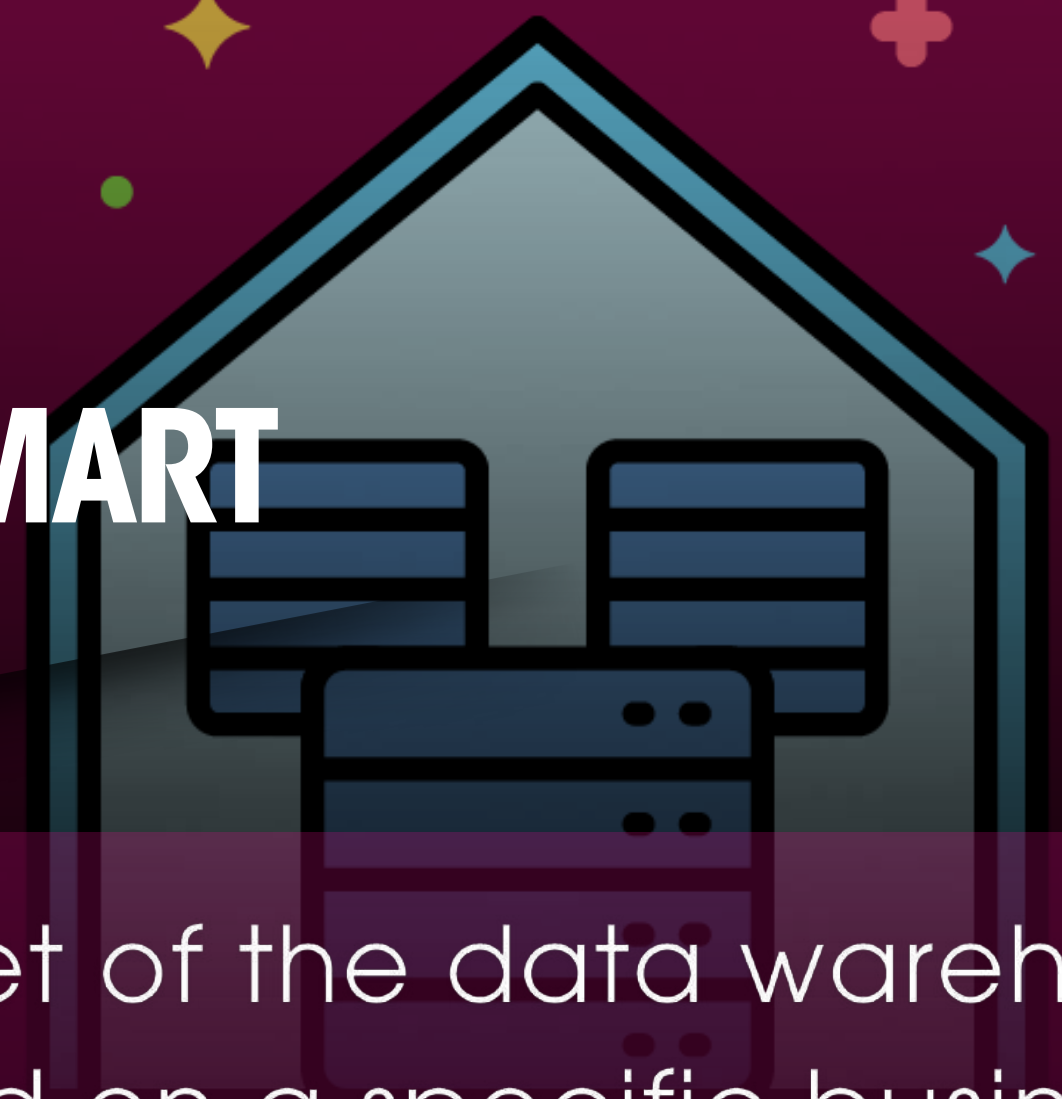
BUS ARCHITECTURE

A framework for organizing data marts and warehouses using conformed dimensions and facts.

Ensuring that all data marts use the same date dimension table to provide a consistent view of time across the organization.

Shwetank Singh
GritSetGrow - GSGLearn.com

DATA MART



A subset of the data warehouse focused on a specific business area or department.

A sales data mart that contains sales-related data for analysis by the sales department.

Shwetank Singh
GritSetGrow - GSGLearn.com

GRAIN DECLARATION

The process of defining the granularity of the fact table during design.

Deciding that the grain of the sales fact table will be at the daily transaction level.

Shwetank Singh
GritSetGrow - GSGLearn.com

DIMENSIONAL HIERARCHY

A structure that organizes dimension attributes into levels of granularity.

A time hierarchy with levels for year, quarter, month, and day.

Shwetank Singh
GritSetGrow - GSGLearn.com

MULTIVALUED DIMENSIONS

Dimensions where attributes can have multiple values for a single entity.

A customer dimension where a customer can have multiple phone numbers or email addresses.

Shwetank Singh
GritSetGrow - GSGLearn.com

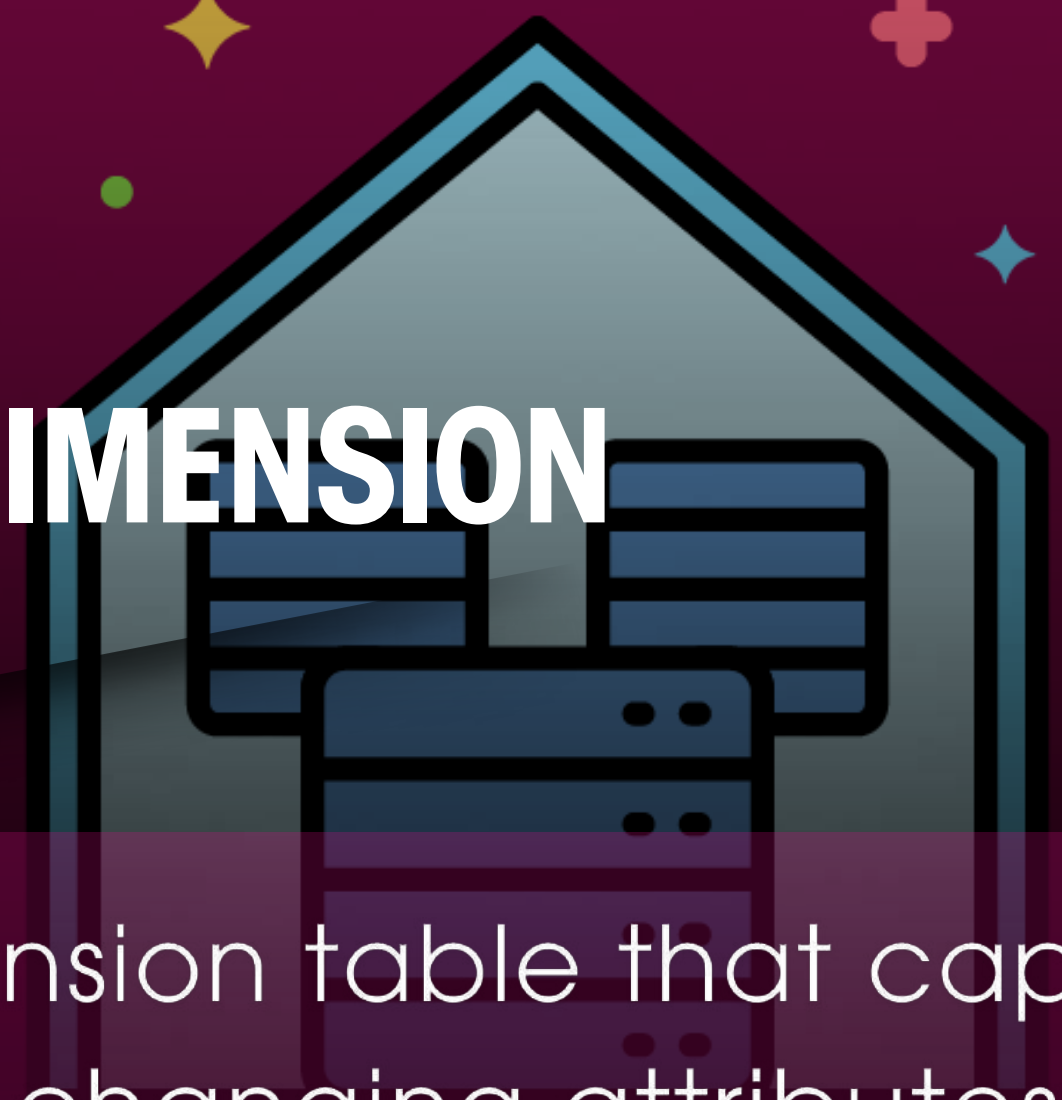
OUTRIGGER DIMENSION

A dimension table that is linked to another dimension table rather than directly to a fact table.

A product subcategory table linked to a product category table, which in turn is linked to the fact table.

Shwetank Singh
GritSetGrow - GSGLearn.com

| MINI-DIMENSION



A dimension table that captures rapidly changing attributes, separated from the main dimension table.

A mini-dimension for tracking changes in customer preferences separate from the main customer dimension table.

Shwetank Singh
GritSetGrow - GSGLearn.com

DATA STEWARDSHIP

The management and oversight of an organization's data assets to ensure data quality and governance.

Implementing policies and procedures for data entry, validation, and maintenance to ensure accurate and reliable data in the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

SURROGATE KEY PIPELINE

The process of assigning surrogate keys to records as they are loaded into the data warehouse.

Generating unique integer surrogate keys for customer records during the ETL process.

Shwetank Singh
GritSetGrow - GSGLearn.com

SLOWLY CHANGING MEASURE

Measures in a fact table that change over time and need to be tracked historically.

Tracking historical changes in product prices in the sales fact table.

Shwetank Singh
GritSetGrow - GSGLearn.com

FACT TABLE GRAIN

The level of detail or granularity of the data stored in a fact table.

Storing sales data at the transaction level versus the daily summary level.

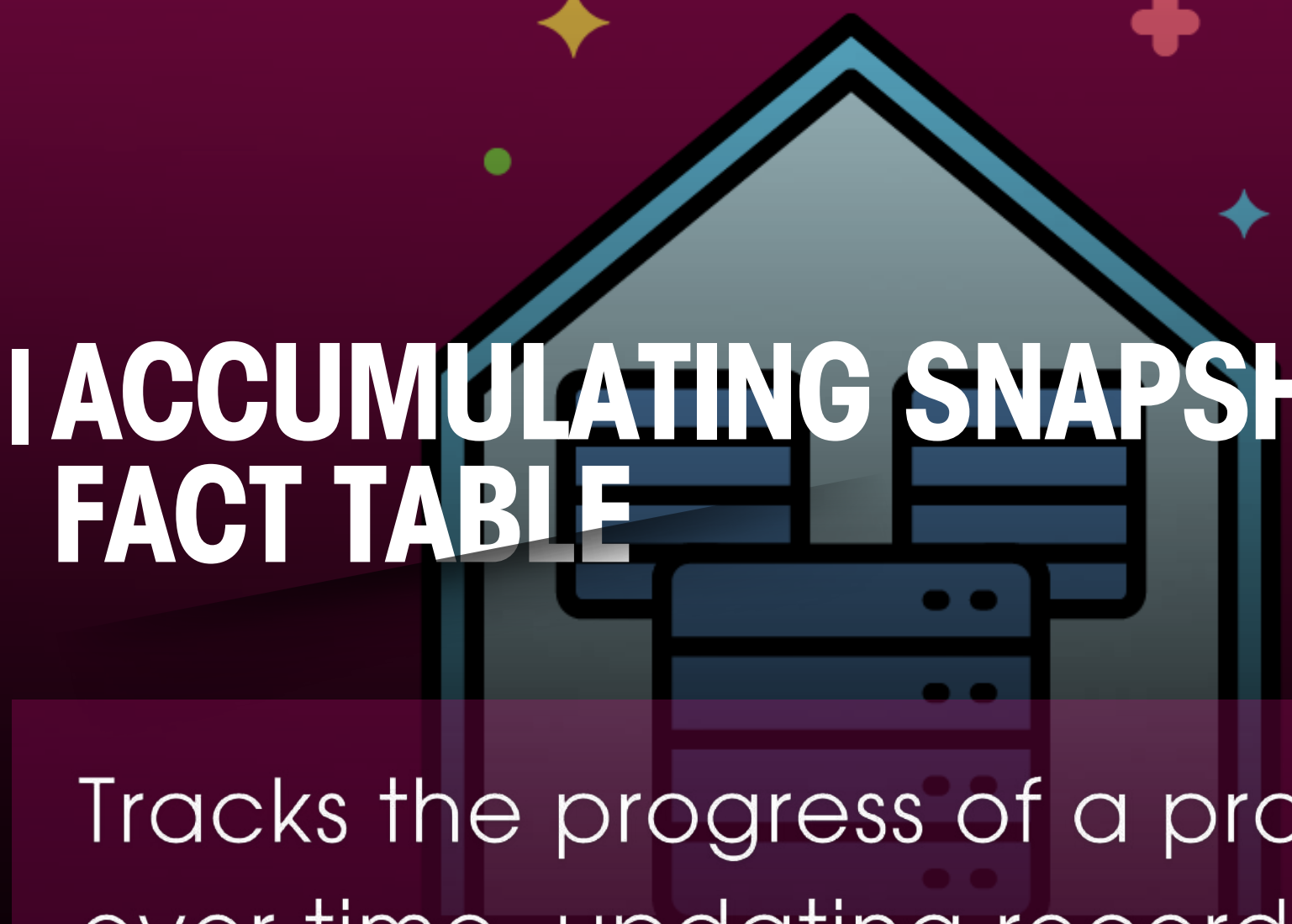
Shwetank Singh
GritSetGrow - GSGLearn.com

SNAPSHOT FACT TABLE

Captures the state of a process at a specific point in time.

A table capturing the month-end inventory levels for each product.

Shwetank Singh
GritSetGrow - GSGLearn.com



ACCUMULATING SNAPSHOT FACT TABLE

Tracks the progress of a process over time, updating records as milestones are reached.

A table tracking the stages of an order from placement to shipment and delivery, with updates to the same record as the order progresses.

Shwetank Singh
GritSetGrow - GSGLearn.com

SEMI-ADDITIVE MEASURES

Measures that can be summed across some dimensions but not others.

Bank account balances that can be summed across time but not across accounts.

Shwetank Singh
GritSetGrow - GSGLearn.com

| NON-ADDITIVE MEASURES

Measures that cannot be summed across any dimension.

Ratios or percentages, such as profit margins, that cannot be summed meaningfully.

Shwetank Singh
GritSetGrow - GSGLearn.com

ETL STAGING AREA

A temporary storage area used during the ETL process to hold data before it is transformed and loaded.

Using a staging database to store raw sales data extracted from transactional systems before transforming and loading it into the data warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

DATA QUALITY

Ensuring the accuracy, completeness, and reliability of data in the data warehouse.

Implementing data validation checks during the ETL process to ensure accurate and consistent data.

Shwetank Singh
GritSetGrow - GSGLearn.com

DATA LINEAGE

Tracking the origin and transformations of data as it moves through the data warehouse.

Maintaining metadata that documents the source systems, transformations, and loading processes for each data element in the warehouse.

Shwetank Singh
GritSetGrow - GSGLearn.com

BUSINESS PROCESS

A series of activities or tasks that produce a specific outcome, often used as the basis for defining fact tables.

The order fulfillment process, which includes order placement, processing, shipment, and delivery.

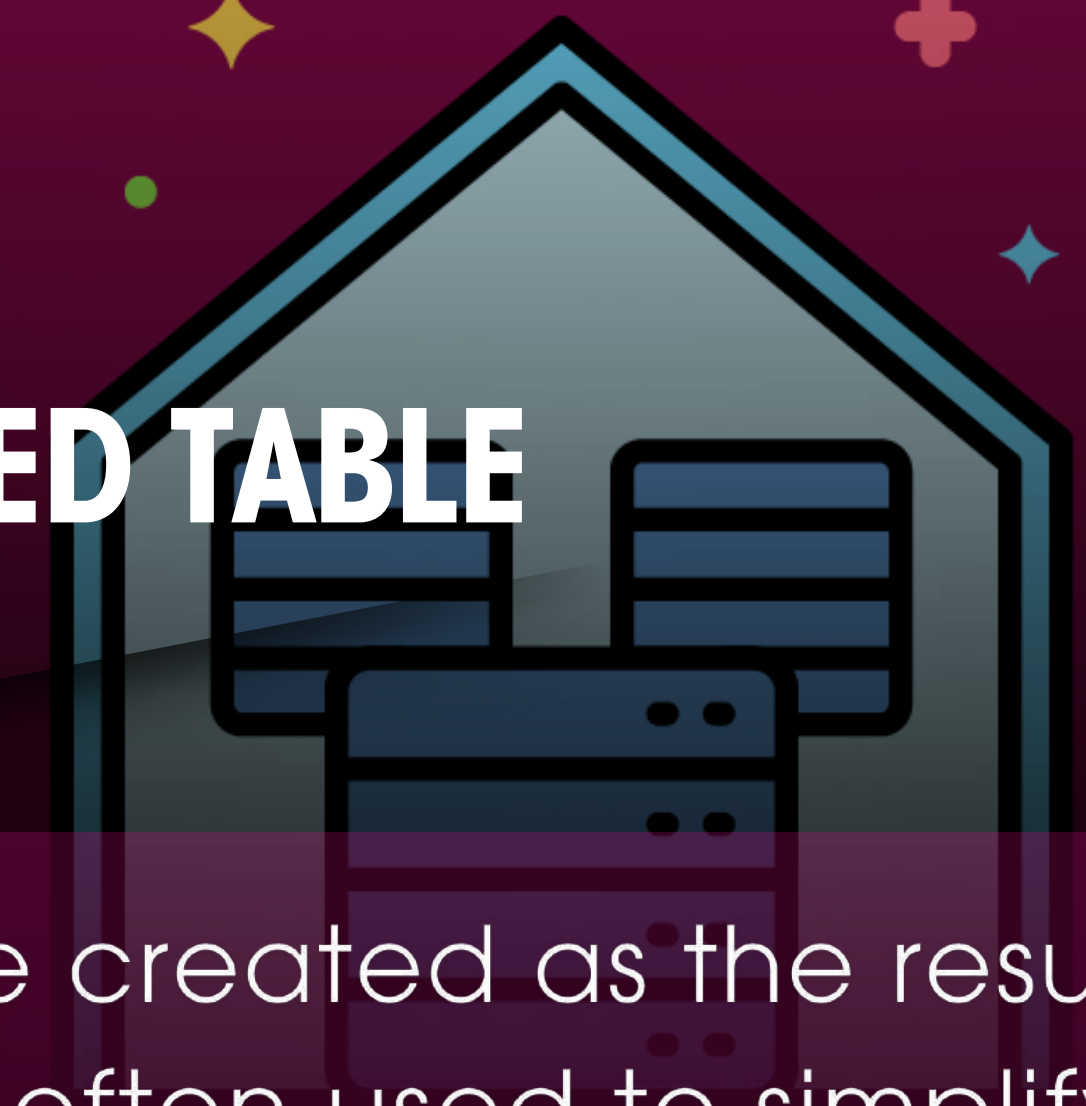
Shwetank Singh
GritSetGrow - GSGLearn.com

DRILL-DOWN ANALYSIS

The ability to navigate from summarized data to more detailed data.

Analyzing sales data by drilling down from monthly sales totals to daily sales transactions.

Shwetank Singh
GritSetGrow - GSGLearn.com



| DERIVED TABLE

A table created as the result of a query, often used to simplify complex joins and calculations.

Creating a derived table to calculate the average order value for each customer segment.

Shwetank Singh
GritSetGrow - GSGLearn.com

FACT TABLE AGGREGATION

The process of summarizing detailed data in a fact table to improve query performance.

Aggregating daily sales data into monthly sales summaries to speed up reporting queries.

Shwetank Singh
GritSetGrow - GSGLearn.com

PERFORMANCE TUNING

Techniques used to optimize the performance of the data warehouse and its queries.

Indexing key columns in fact and dimension tables to improve query performance.

Shwetank Singh
GritSetGrow - GSGLearn.com

DATA WAREHOUSE ARCHITECTURE

The overall structure and organization of the data warehouse, including schemas, ETL processes, and data storage.

Implementing a three-tier architecture with staging, integration, and presentation layers.

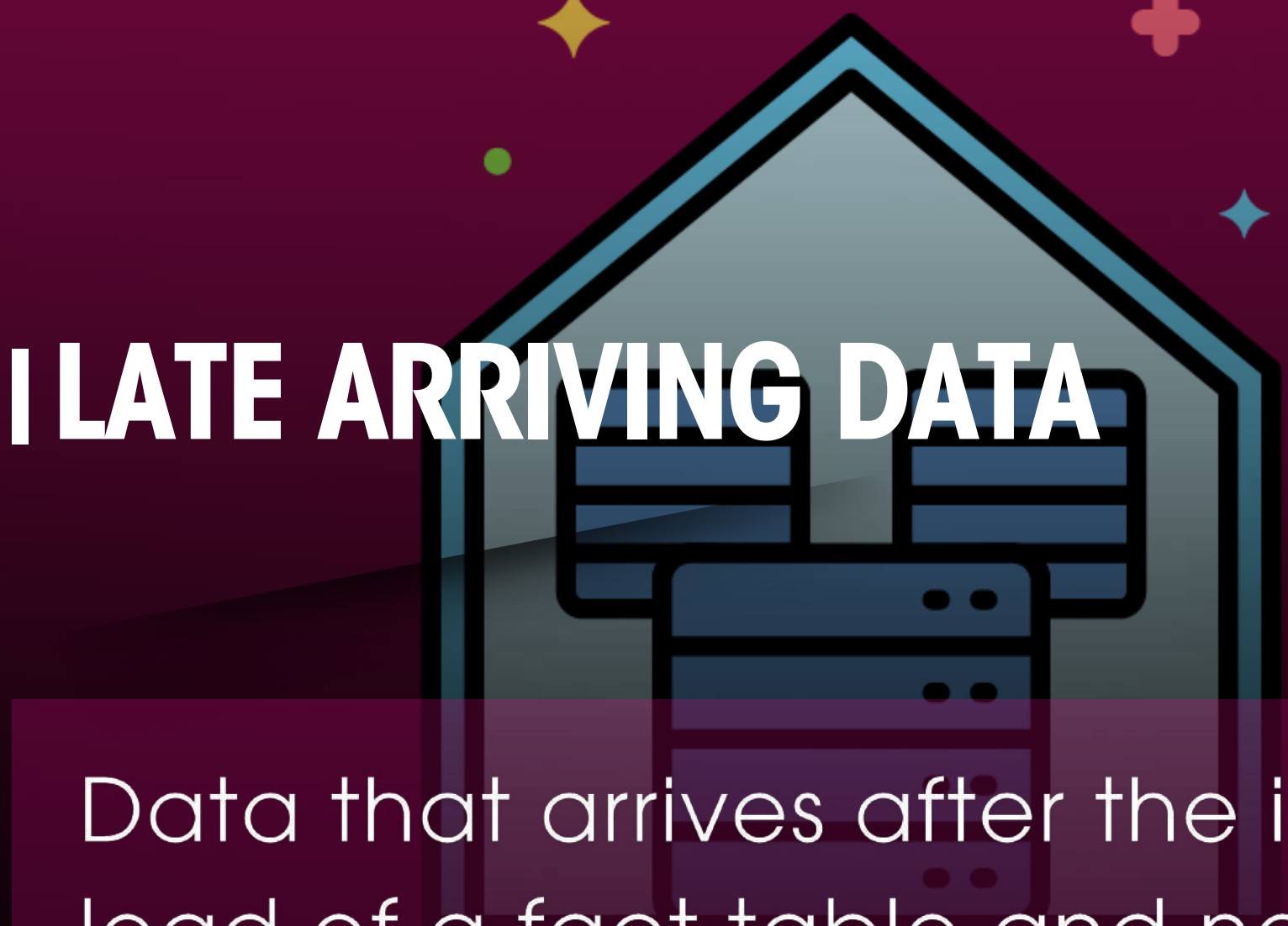
Shwetank Singh
GritSetGrow - GSGLearn.com

DIMENSIONAL INTEGRITY

Ensuring consistency and accuracy of dimensions across the data warehouse.

Implementing referential integrity constraints to ensure dimension keys in fact tables match primary keys in dimension tables.

Shwetank Singh
GritSetGrow - GSGLearn.com



LATE ARRIVING DATA

Data that arrives after the initial load of a fact table and needs to be integrated into the existing data.

Handling late arriving sales transactions that need to be added to a fact table after the end of the reporting period.

Shwetank Singh
GritSetGrow - GSGLearn.com

DATA WAREHOUSE LIFECYCLE

The stages of development and maintenance of a data warehouse, from initial planning to ongoing management.

Following a lifecycle approach that includes requirements gathering, design, implementation, testing, and maintenance phases.

Shwetank Singh
GritSetGrow - GSGLearn.com

OLAP (ONLINE ANALYTICAL PROCESSING)

A category of software tools that provide analysis of data stored in a database, often used in data warehousing.

Using OLAP tools to perform multidimensional analysis of sales data, such as slicing, dicing, and pivoting.

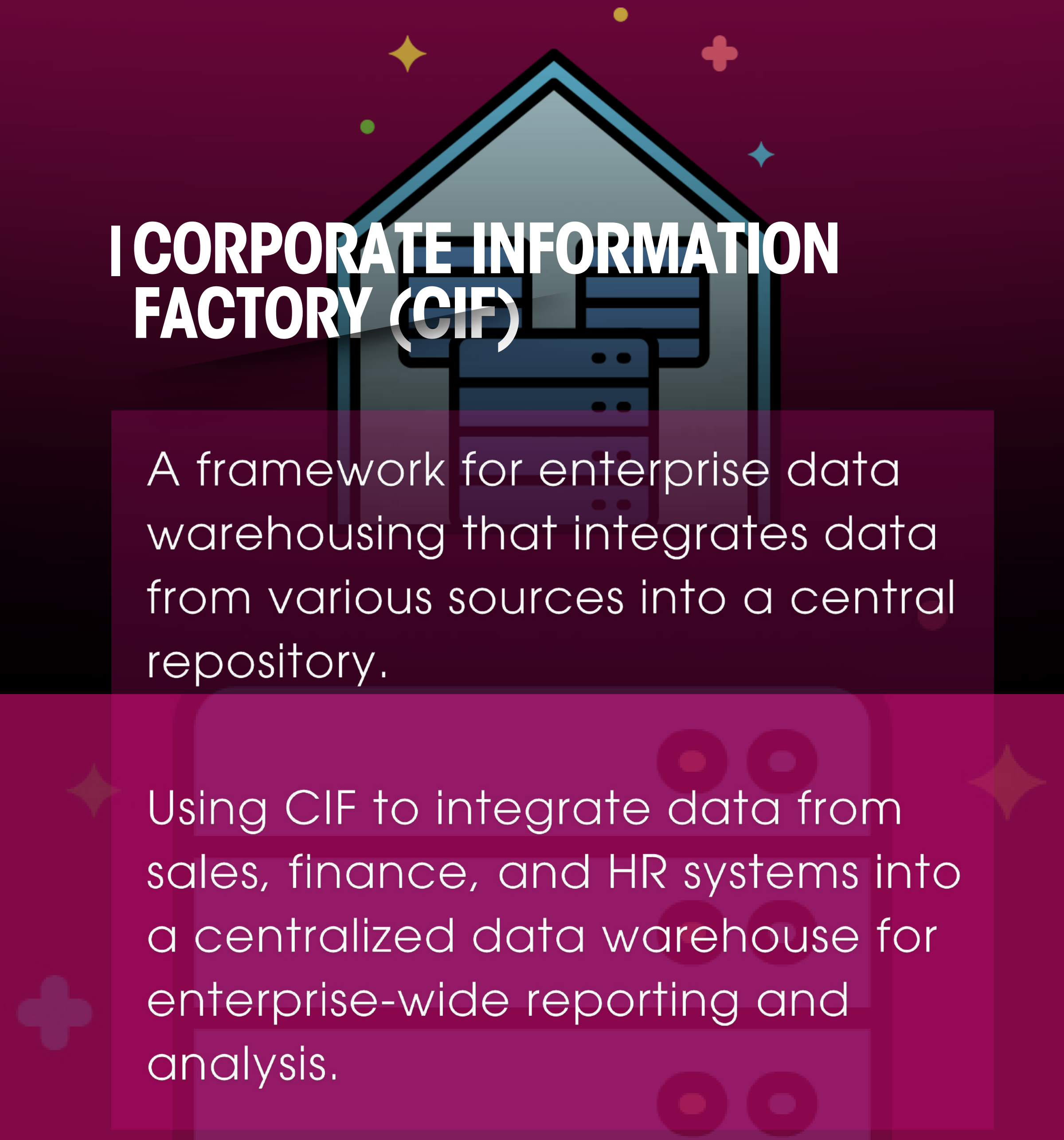
Shwetank Singh
GritSetGrow - GSGLearn.com

KIMBALL METHODOLOGY

An approach to data warehousing that emphasizes the use of dimensional modeling and incremental development.

Implementing a data warehouse using the Kimball Methodology, starting with a data mart and gradually integrating additional data sources.

Shwetank Singh
GritSetGrow - GSGLearn.com

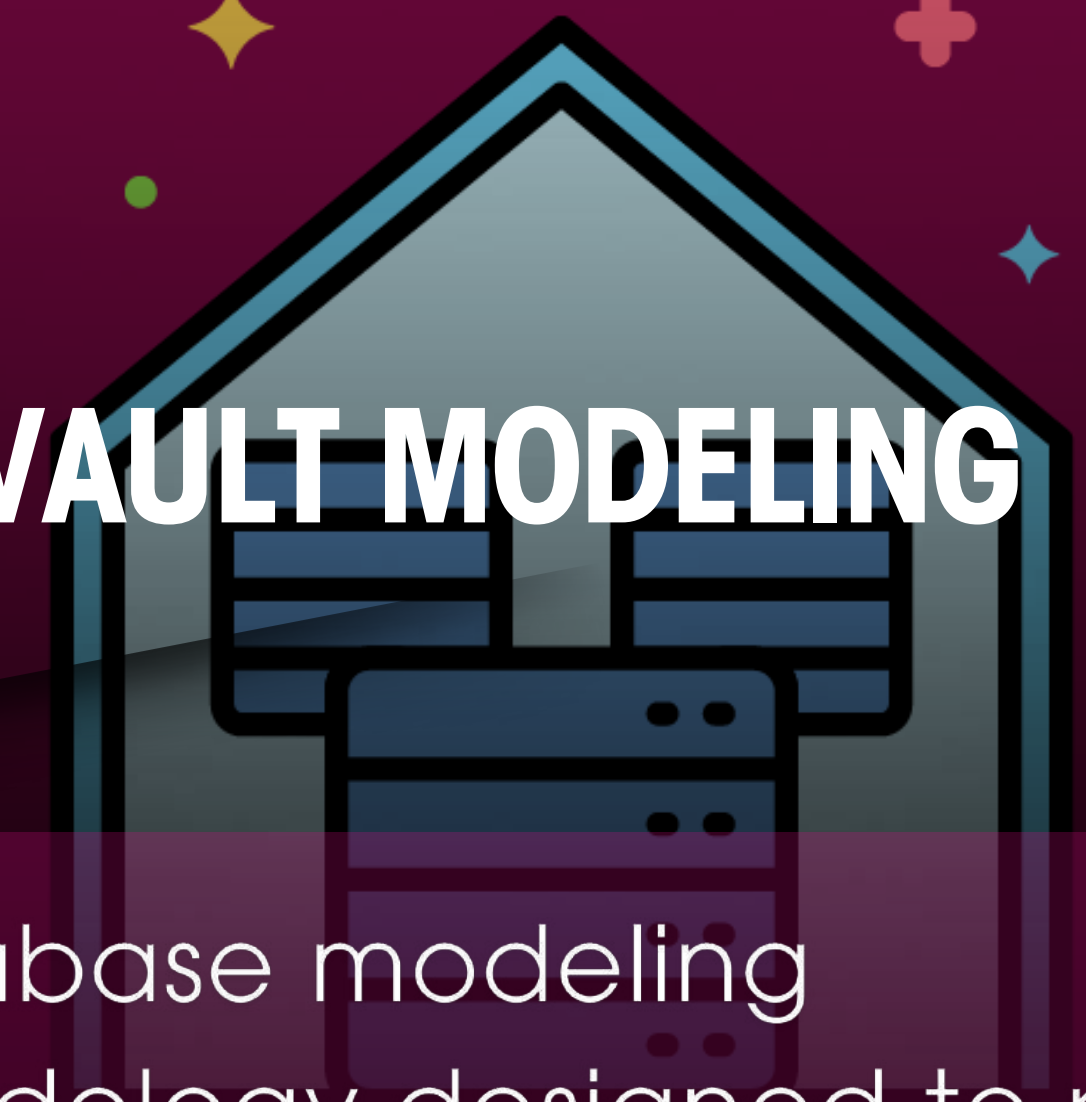


CORPORATE INFORMATION FACTORY (CIF)

A framework for enterprise data warehousing that integrates data from various sources into a central repository.

Using CIF to integrate data from sales, finance, and HR systems into a centralized data warehouse for enterprise-wide reporting and analysis.

Shwetank Singh
GritSetGrow - GSGLearn.com



DATA VAULT MODELING

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

Implementing a data vault model to store historical sales data from multiple transactional systems in a central repository.

Shwetank Singh
GritSetGrow - GSGLearn.com

Thank you