DW is a System designed to support BI, particularly analytics. It focuses on querying large amounts of historical data, helping businesses make data driven decisions. Diff b/w OLTP and DW -> OLTP systems are primarily operational systems and analytical queries can cause performance issues. OLTP systems are normalized, which makes them highly unfriendly for ad-hoc queries and makes it difficult for business users to interact with data. OLTP systems doesn’t store data for long period of time, and only keep what is necessary. Success Criteria – Accessibility, Consistency, Adaptability, Timeliness, Security, Trustworthiness, Acceptance. OLAP cube -> The application of dimensional modeling to create a multidimensional cube for data analysis. Kimball – Denormalized, star schema, facts and dimensions. Inmon – Normalized, Enterprise wide, high firsthand cost.

A computer screen shot of a diagram

Description automatically generated A screenshot of a computer

Description automatically generated

Database is an organized collection of structured information, or data typically stored electronically in a computer system. ER Model – The standard modeling method for relational databases, leverages the concept of relationships b/w tables. Normalization – Process of splitting data into multiple tables to reduce redundancy and increase consistency., DDL – Used to define and modify the schema and core database elements, ex Create. DCL – Used to define rights and permission to users, ex Grant. DML – used to manipulate data in db, ex Update. TCL – Manages the changes made by DML stmnts, ex Commit. DQL – Used to query db, ex Select. INNER JOIN: Combines rows from two or more tables based on a related column (primary and foreign key relationship) where a matching value exists in both tables LEFT (outer) JOIN: Returns all records from the “left” table, and any records from the right table that match the left RIGHT (outer) JOIN): Returns all records from the “right” table, and any records from the left table that match the right FULL (outer) JOIN: Returns all records when there is a match in either table. DM is a methodology for designing DW that seeks to deliver data to business users in an understandable way, while providing for fast query performance. Facts – measurable outcomes of a business process. Dimensions – Provide context to facts. STAR vs OLAP – STAR - Simplified db schema designed to query with facts and connected dimensions. OLAP – multi dimensional db designed for quick data slicing and analysis. Limitations of cubes – Specially architected by IT to support new business cases, slices of data, drill down and drill up. Must recomputed during any change.

The data warehouse lifecycle - Select the business process, Declare the granularity, Identify the dimensions, Identify the facts. Business Processes: A business process represents the operational activities conducted by an organization and the steps performed by people/technology to achieve a goal. Essential for identifying what needs to be measured. Granularity: What level is needed, what are business reporting wants/needs, what granular data we have. Atomic grain: Most detailed level of grain. Rolled-up Grain: Aggregation of the grains. Identifying Dimensions: provide context to a business process. Who, What, When, Why, How, Where. Enterprise Bus Matrix is a dimensional modelling technique designed to assist in the mapping of processes to potential common dimensions. These will form the foundations for your Star schema relationships when conducting the logical design. The bus matrix columns (dimensions) and rows (facts) will directly depend on the processes and descriptors relevant to the industry and specific business. Facts are measurements of a business process. Source of data: Understanding the available data helps to Provide context to business process, balance process requirements with reality, prevent you from overcommitting.

Dimension Models in dimensional modeling provide the context facts. Dimension table typically contain descriptive attributes like customer name or product category. Surrogate Keys: PK for a record but no business meaning. Natural Key/ Business Key: A unique value that can be used as a primary key but has a business meaning eg SSN. Dimension tables should almost always use a surrogate key as a primary key. Degenerate Dimension: Dimensions with no actual data, often placed in fact tables, eg Invoice Number. Hierarchies: Defines aggregation paths for data. Relationships b/w different levels of data, commonly used for drill down reporting. Fixed Hierarchy: Each attribute as a one-to-many relationship with the next attribute and always has the same number of attribute levels. If hierarchy is fixed, each attribute will have a subcategory, category and department (Retail case). Variable Depth hierarchy: Do not have fixed number of levels. 1. Slightly Ragged hierarchy: Minor variations in hierarchy, consider an example of Address- Simple location, medium location and Complex location. 2. Ragged Hierarchy: Higher variations in levels and sometimes no clear pathway upward in the hierarchy. Few techniques to address to flatten ragged hierarchy – 1. Recursion: A method of representing a parent child relationship by having the parent surrogate key as an attribute of the child record. 2. Hierarchical Bridge table: An immediately table that connects a ragged dimension with a fact table and provides information about all the rollup possibilities for the ragged hierarchy. Conformed Dimensions: Is a dimension that can be tied to more than one fact table. They create flexibility and scalability, as the data in the dimension is all in the “same language” and is fit for multiple contexts. For ex-Date dimension should contain all the attributes needed by the organization in a consistent format – year, quarter, season, month, day of month, day of week, holiday, etc. Shrunken Dimension: These represent a subset of conformed dimension and are used when the fact table doesn’t require full granularity. For ex – shrunken date dimension may relate to a fact table that only captures sales by quarter and year, but not at a lower granularity – (year, quarter). Slowly Changin dimensions: Dimension attributes which could be changed over time or require an update. They are categorized into 3 different types. Type 0: Unchanging dimension attributes, Type 1: Change attributes, only requires attributes to be overwritten. Overwrites can be destructive, as all history of previous value is lost. Overwrites can also have serious regulatory and legal consequences, depending on the organization and its reporting requirements. Type 2: Add anew row for each change, allowing you to track historical data while maintain current data. Type 3: Add a new column to the dimension table both old and new values. This method is not preferred as it creates conflicting elements in a single record and is less scalable over time. Common Attributes: Date/Time dimension, Product, Customer.

Fact Tables: Store quantifiable data generated by business processes. PK: Unique (Natural key, Surrogate key, composite key) FK: Link to dimension tables. Facts: Numeric values representing key measurements. Types of fact tables: Fact tables are designed according to the nature of the data and its time related behavior. Transactional fact table: Capture individual events at their atomic level. Event driven and capture facts at a specific time. Periodic Snapshot fact tables: Store aggregated facts for a specific time such as daily or monthly totals. Accumulating snapshot fact tables: Track a business process with multiple milestones over time (eg order placement to shipment). These accumulate data over each stage of the business. Types of facts: 1. Additive facts: Summed across dimensions Eg total sales across time, customers or locations. 2. Semi additive facts: Summed across some dimensions but not others, ex acc balance can be summed across customers but not time. 3. Non-additive facts: These cannot be summed at all. Common example is percentage discounts. Handling Null values: Can be present in fact table, but not allowed in foreign keys. TO handle this case, a wild card row in dimension table is added. Conformed Facts: Facts that are consistent across multiple fact tables, computed in the same and used across multiple business processes. Factless fact table: They do not contain measurable data but instead track events or occurrences (e.g. student attendance.), without numeric value. Aggregate fact tables: These store pre-aggregated data, such as sales totals to improve query performance. DataMart is a subset of DW, focusing on a particular business line. OLAP Cubes are multi-dimensional databases that store data in a format that is designed for rapid analysis instead of a relational format. Minimizes real time processing by precomputing combinations of values for instant reporting. Consolidated facts tables: Sometimes it may make sense to combine facts from different business processes in one table. Can make loading data more complex but drastically simplify business intelligence. Common fact tables: Sales, Sales forecast, Accounts receivable, accounts payable, shipping, customer support, health insurance claim.

A screenshot of a computer

Description automatically generated A computer screen shot of a computer screen

Description automatically generated