Data Warehouse Design – Exercise

Learning goal: carry out some research into data warehouse design and gain an understanding of some of the key concepts of data warehousing design.

Read and provide descriptions as required.

Two main methodologies for developing an enterprise data warehouse (EDW)

1. Kimball's Business Dimensional Lifecyle or Kimball's Life Cycle (Kimball, 2008)
2. Inmon's Corporate Information Factory (CIF) methodology (Inmon, 2001)

Inmon's architecture is called an enterprise data warehouse. Kimball's architecture is known as the dimensional data warehouse

Describe Inmon's Corporate Information Factory (CIF) methodology.

Kimball uses the dimensional model such as star schemas or snowflakes to organize the data in dimensional data warehouse while Inmon uses ER model in enterprise data warehouse. Inmon only uses dimensional model for data marts only while Kimball uses it for all data.

The Kimball lifecycle is a methodology for developing data warehouses. Describe its key phases.

Describe the star schema.

Dimensional Modelling is logical design technique that aims to present data in a standard, intuitive form that allows for high performance access.

Every Dimensional model (DM) is composed of one table with a composite primary key, called the fact table, and a set of smaller tables called dimension tables.

Each dimension table has a simple primary key that corresponds exactly to one of the components of the composite key in the

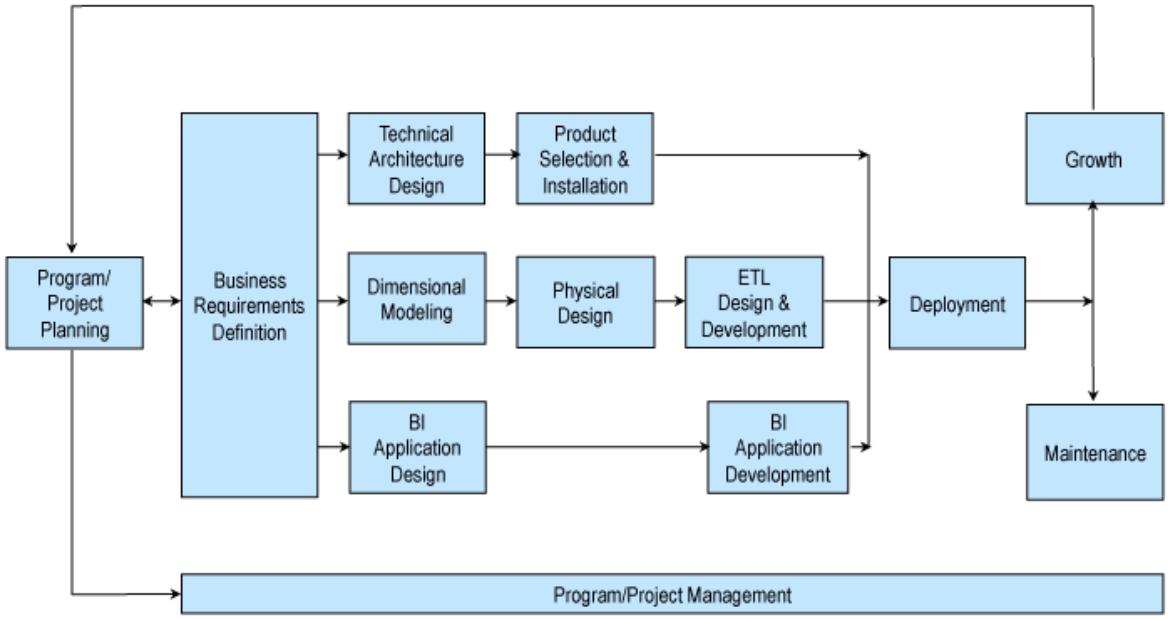
Describe the snowflake schema.

1 - Describe Inmon’s Corporate Information Factory (CIF) methodology.

Inmon's Corporate Information Factory (CIF) methodology is a data warehousing approach that focuses on creating an integrated and standardized information architecture within an organization. It was developed by Bill Inmon, a prominent figure in the field of data warehousing. The CIF methodology is designed to provide a framework for organizing and managing an enterprise's data assets in a way that facilitates effective business intelligence and decision-making.

2 - The Kimball lifecycle is a methodology for developing data warehouses. Describe its key phases.

The Kimball Lifecycle is a data warehousing methodology developed by Ralph Kimball, another pioneer in the field of data warehousing. This methodology is often contrasted with Bill Inmon's approach, with Kimball's focus being more on delivering quickly and incrementally.



It consists of the following key phases:

- Project Planning: In this initial phase, the overall scope and goals of the data warehousing project are defined. This includes project initiation, forming a project team, defining business requirements, and establishing a high-level project plan.

- Business Requirements Definition: The focus in this phase is on defining and documenting the business requirements for the data warehouse. The project team works closely with business stakeholders to understand their information needs, reporting requirements, and analytical goals.

- Data Warehouse Design: In the Data Warehouse Design phase, the technical architecture of the data warehouse is defined. This includes decisions on data modelling, schema design, and the selection of tools and technologies for extraction, transformation, and loading (ETL) processes.

- ETL Design and Development: The Extract, Transform, Load (ETL) processes are important for moving and transforming data from source systems to the data warehouse. In this phase, the ETL processes are designed and developed based on the requirements identified in earlier phases.

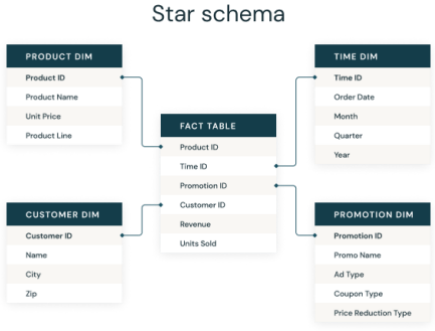
- Deployment: The data warehouse is deployed for use by end-users.

- Maintenance and Growth: The final phase involves the ongoing maintenance and evolution of the data warehouse.

3- Describe the star schema.

An easy-to-understand and analyse multi-dimensional data model for organising data in a database. Databases, data marts, data warehouses, and other technologies can all employ star schemas. Large data sets can be queried more efficiently with the star schema approach.

Star schemas were first introduced by Ralph Kimball in the 1990s. They are effective for storing data, preserving history, and updating data by minimising the need for duplicate business definitions. This allows for quick data aggregation and filtering in the data warehouse.



4 - Describe the snowflake schema.

A multi-dimensional data model called a snowflake schema is an expansion of a star schema in which dimension tables are divided into smaller subdimensions. Snowflake schemas are frequently utilised in relational databases, OLAP data warehouses, and data marts for reporting and business intelligence purposes.

Engineers divide up individual dimension tables into logical subdimensions in a snowflake structure. Although this increases the complexity of the data model, analysts may find it easier to work with—especially for some types of data.

The reason it's named a snowflake schema is that its entity-relationship diagram (ERD) resembles one.

References:

https://www.astera.com/type/blog/data-warehouse-concepts/

https://en.wikipedia.org/wiki/Bill_Inmon

https://en.wikipedia.org/wiki/Ralph_Kimball

https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dw-bi-lifecycle-method/

http://www.kimballgroup.com/wp-content/uploads/2012/05/DT115KimballLifecycleNutshell.pdf

https://www.databricks.com/glossary/star-schema

https://en.wikipedia.org/wiki/Snowflake_schema

https://learn.microsoft.com/en-us/power-bi/guidance/star-schema

https://www.geeksforgeeks.org/snowflake-schema-in-data-warehouse-model/#

https://www.databricks.com/glossary/snowflake-schema