Topic: Decision Support Systems (DSS); Introduction to Data Warehousing
Student's Full Name: Sourav Mangla
Course Number and Title: INFO 531: Data Warehousing and Analytics in the Cloud
Term name and year: Spring 2023 Submission
Week: Week 1 Assignment
Instructor's Name: Nayem Rahman
Date of Submission: 15 Jan 2023

Summary:
Data warehousing is a group of decision-supporting technology designed to help professionals to make choices more quickly and effectively. Data warehouses are often orders of magnitude bigger than operational databases because they combine data from several operational databases over potentially long periods of time; enterprise data warehouses are anticipated to be hundreds of gigabytes to terabytes in size. Attempting to run complicated OLAP queries against operational databases would result in unsatisfactory performance since operational databases are carefully optimized to accommodate known OLTP workloads. One or more warehouse servers store and manage data in the warehouse and data marts and provide multidimensional representations of the data to various front-end tools, including query tools, report writers, analysis tools, and data mining tools.
A data warehouse is a computer system that stores and processes large amounts of data in a variety of formats. It may include tools for loading, cleaning, transforming, and loading data from multiple sources into the data warehouse; and for periodically refreshing the warehouse. In addition to the main warehouse, there may be several departmental marts, each with its own dedicated servers and management tools.
Data warehousing systems use a variety of data extraction and cleaning tools and load and refresh utilities for populating warehouses. Data extraction from "foreign" sources is usually implemented via gateways and standard interfaces. Detecting anomalies, inconsistent field lengths, conflicting descriptions, inconsistent value assignments, missing entries, and integrity constraint breaches are the fundamental cleaning stages. For cleaning Data migration tools, Data scrubbing tools, and Data auditing tools are also used. Data has to be put into the warehouse once it has been extracted, cleaned, and transformed. For this, batch load utilities are typically employed. Compared to operational databases, the load utilities for data warehouses must handle substantially bigger data quantities. In order to update the base data and derived data stored in a warehouse, updates on the source data must be propagated to the warehouse. The warehouse is often updated on occasion like daily or weekly.
Data warehouses utilize a multidimensional representation of the data. A collection of numerical measurements make up a multidimensional data model's analytical targets. It is thought that the dimensions taken together will always produce the measure. A collection of characteristics are used to characterize each dimension. Most data warehouses use a star schema or Snowflake schemas or Fact constellations structure to represent multidimensional data models. We need highly efficient access methods and query processing techniques to answer queries efficiently. Choosing which indices to build and which views to materialize is an important physical design problem. A number of query-processing techniques that exploit

indices are useful. For instance, the selectivities of multiple conditions can be exploited through index intersection.
A lot of data warehouse queries request summary data, which necessitates the usage of aggregates. It should be made clear that the aggregating functions' algebraic characteristics are what allow for roll-up from a partially aggregated result. Processing large databases involve a lot of parallelisms. Some of the important technologies were invented by Teradata. Data partitioning and parallel query processing technology are currently offered by all significant database management system providers. The server architectures for query processing include specialized SQL Servers, ROLAP Servers, and MOLAP Servers. In extended relational servers, a number of SQL extensions that simplify the phrasing and execution of OLAP queries have been suggested or implemented.
Metadata management is a crucial component of warehousing architecture since a data warehouse represents the business model of an organization. Metadata comes in a variety of forms that must be maintained. The history of migrated and converted data, the current status of the data in the warehouse, and monitoring data like use statistics, error reports, and audit trails are all included in operational metadata. The well-known issues of index selection, data partitioning, and materialized view selection should garner further interest in view of the difficulty of the physical architecture of data warehouses. Workflow technology adoption and application might be beneficial, but further research is required.

Data warehouses contain consolidated data from several operational databases, over potentially long periods of time, they tend to be orders of magnitude larger than operational databases.
The warehouse is distributed for load balancing, scalability, and higher availability.
Data in the warehouse and data marts is stored and managed by one or more warehouse servers, which present multidimensional views of data to a variety of front-end tools: query tools, report writers, analysis tools, and data mining tools.
Architecture and backend technologies of the data warehouse to perform different operations like Data Cleaning, data refreshing, and data loading.
Database Design Methodologies to design data warehouse in star, snowflake, or fact constellation models and their pros and cons.