Assignment / Explore Data Warehoues

Connor Clancy - clancy.co@northeastern.edu

Spring 2023

Question 1

Data warehouses are often constructed using relational databases. Explain the use of fact tables and star schemas to construct a data warehouse in a relational database. Also comment on whether a transactional database can and should be used to OLAP.

Fact tables and Star Schemas can be used in an *Online Analytical Processing (OLAP)* environment to contruct a relational database that is optimized for analytical querying rather than the traditional *Online Transaction Processing (OLTP)* which is optimized for transactional querying. Another way to think about this is that OLAP databases are optimized for read queries where as OLTP databases are optimized for write queries. Both environments can be created in traditional RDBS programs, the main difference is how the tables and schemas are designed.

In OLAP, fact tables usually consist of pre-calculated and aggregated data points in a non-normalized schema. This helps to reduce query time and complexity for the analytical user and allows them to answer questions about the data in a minimal amount of time. Dimension or 'dim' tables are then set up around the fact table to layer in additional data not included in the fact table. Star databases usually only require one-layer of joins from the fact table to get to a dimension table. If additional layers are added, this is called a snowflake design and can sometimes be appropriate, but can also slow down processing time for queries the more complicated these schemas get.

Traditional OLTP databases can be used for analyzing data, but it is likely that this would be painful for an analytical users who need to use this databases as their queries would take a long time to develop (due to the complex nature of the schema they are working work) and the query would take a long time to execute as compared to the same data in an OLAP Star Schema database.

Question 2

Explain the difference between a data warehouse, a data mart, and a data lake. Provide at least one example of their use from your experience or how you believe they might be used in practice. Find at least one video, article, or tutorial online that explains the differences and embed that into your notebook.

Data lakes are the least curated and least structured analytical data store; they tend to be raw collections of data that are being centralized from their original sources. My company uses S3 storage buckets as a data lake to act as a data centralization point entering our cloud environment. Here you can find JSONs, CSVs, Parquete, and database table files.

Data warehouses exist after an Extract, Transform, & Load process (ETL). This process involves *extracting* data from its source (either the data lake, or possible another source), *transforming* the data to match the schema of the data in the data warehouse, and then *loading* the data into the dataw arehouse tables. My company uses Hadoop to do the ETL and datawarehousing processes. Once the data ETL'd it can be accesses using Hive SQL commands by our data engineering and data management teams.

Data marts are the most curated data stores we are covering in this assignment. They are subsets of a data warehouse used by a specific domain (marketing, finance, security, etc.). My company uses Snowflake to house our data marts; our data marts serve two purposee. The first is to simplify the domain knowledge requirements for analysts, and the second is to ensure that we are only exposing analysts to the data necessary to do their jobs as a security and privacy control.

Question 3

After the general explanation of fact tables and star schemas, design an appropriate fact table for Practicum I's bird strike database. Of course, there are many fact tables one could build, so pick some analytics problem and design a fact table for that. Be sure to explain your approach and design reasons.

Let's say we wanted to run a monthly analysis to trend the number of bird strikes at each airport to determine if the number of incidents is getting better or worse by location. We could pre-compute aggregations of strikes by airport and month into a fact table with dimension tables for the airport metadata. This would allow us to easily analyze a trend and build visualizations to easily see the data over time.

Data Warehouse UML