**ETL Process Implementation**

The second major part of the project involved the implementation of the extract, transform, load process (ETL). In this process the data was being moved from the source system to the data warehouse, and cleaned, transformed and then loaded into the right tables. For the project to succeed, having clean and consistent data in the fraud detection system was critical.

First step of ETL was to extract data from the source system. Those who did this accessed the operational database, retrieved raw transactional data (which may or may most’ve have been messy, incomplete, unformatted). Supply was extracted through SQL queries that selected appropriate fields like i.e. transaction IDs, product names, customer IDs etc., for sales as well as refunds. The biggest challenge in this step was that the extracted data have to be full – no information should have been omitted in the extraction process, and that the extraction process itself won’t affect the source database’s performance.

The next step was to extract the data and then transform that data to a format that is more pungent to the data warehouse. Data transformation process included cleansing and mapping of the data along with reformatting of it. For example, the transaction date was standardized, matching product name to the correct category, and handling of missing values was done according to pre-defined rules. Data deduplication was one of the most critical pieces to transformation and the reason being that fraud transactions could consist of duplicate records. To ensure that the data loaded into the warehouse was reliable and accurate a series of transformation rules were written to identify and remove all duplicate rows.

The last step in the process to ETL was to transform the data to the data warehouse. According to the pre-defined schema, data was loaded in fact and dimension tables. It involved delicate handling of relationships between the fact table and the numerous dimension tables, but the maps of data onto the appropriate attributes in the warehouse had to be right. Furthermore, in order provide efficiency, the loading process required to process large amounts of data needed to be efficient. Batch processing of the data was exploited to load data in batches to avoid putting strain on system resources.

There was a major challenge when it came to data integrity in the ETL process. This was especially important since fraud detection is based on accurate and consistent data. Erroneous conclusion is possible since any discrepancies in the data. A few checks and validations were added to ensure the data was right before loading it to the warehouse. Validation ensured data types were the same, and no fields were blank, and cross-checking extraction data with third party sources when required.

Using SQL Server Integration Services (SSIS) packages the team automated the ETL process giving the ability to schedule and run the ETL jobs at regular intervals. It was essential that the data warehouse was automatically updated with fresh data without any intervention from humans. It also served to make sure that the data was always current (for spotting a fraud in real time).