**ETL Project**

**Summary – Group 7**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mohamed Jouhari (DE) | Vincent Pinneau (DA) | Christine Ismail (DE) | Abdifatah Farah (DE) |
| STA Phase | * (technical) Resolved the encoding problem that prevented the order\_reviews\_dataset to be loaded in the STA phase |  | * did a table in the database for each dataset and used the foreach loop for orders * used utf-8 for all datasets | -In the Staging phase, I set up SQL tables to match our raw data files (orders, customers, etc.) and used SSIS to automate data imports, especially with multiple files for orders. We hit a technical issue with special characters in the order\_reviews\_dataset, so I adjusted the encoding to UTF-8, ensuring all data loaded smoothly. This gave us a clean starting point for transformations ; |
| ODS Phase |  | * In his ETL, he preferred to create the different tables of the final model with look ups at the ODS stage (before the column resizing). * (technical) In the original final template of the project, the ADM Rejects table was truncated at the beginning of each ODS. In Vincent’s ETL he didn’t include an ADM Reject truncate because he preferred to keep a ’tracking’ of the rejects, and filter by date in SQL when he needed the more recent rejects. He said that we could keep the truncating in the final version but that the truncation had to be only done once, because the different steps of the ODS all shared the same Rejects table. Also, initially, in the sequence container, the ODS phases were not linked together, so he asked to link them and put the ODS that had the ADM reject at the top of the ODS pipeline. | * Created the tables in the ods database exactly like the sta without removing certain columns * Only put added the date and col of the reject without specifying further description |  |
| DWH Phase | * Wanted to have all the geolocation information of customers and sellers regrouped in a geolocation dimension table to reduce the size of the data but also to make the model more logical. | * Had the same idea as Mohamed to move the city/state information from the customer/seller tables to the geolocation dimension to avoid duplicates. Also from a data analyst point of view, having all the gps geolocations already in one table can allow us to do a map visualization comparing customer et seller location more easily. The only thing that would need to be added is a category column in the geolocation table differentiating customers and sellers to filter or do a legend on the map. * Was in favor of putting most information regarding orders (and not orderitems) in a separate dimension table to prevent duplicating information in the orderitems fact table since some orders appear in several rows of the fact table (multiple items). This logic was partially kept in the final ETL by putting the reviews in a separate table dimension table. | * At first did not follow the snowflake scheme but after the meetings agreed to the fact that this scheme will help us retrieve data in a better way so followed the team suggestion and linked the DimGeo table with DimCustomer and DimSeller and the DimDate with DimRev * Did not add a lot of details such as Day ,Day of Year,Suffix…to the DimDate Table I preferred keeping it simple | -I shared the team’s vision for the Data Warehouse schema, including ideas like organizing geolocation details into a single dimension, but I didn’t have the time to work on the implementation directly. |
| Data analysis | * SQL code * Report comments | * Report comments | * SQL | -Throughout the project, I used SQL for organizing, cleaning, and analyzing data. In the ODS phase, I created a **Rejects table** to track data errors, which made troubleshooting easier. In the Data Warehouse phase, I wrote SQL queries to pull insights, like identifying top product categories and analyzing payment methods’ impact on order times. SQL was key to making our data accurate and valuable for business use. |