**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 |  |  |  |
| 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. |  |  |
| 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. |  |  |
| Data analysis | * SQL code * Report comments | * Report comments |  |  |