**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 |  | * **(technical)** 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 past rejects, and filter by date in SQL when he only 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 (ODS Orders) that had the truncation of the ADM reject table at the top of the ODS pipeline, to be sure that the rejects from all the ODS packages were taken into account when launching the entire pipeline. * **(Modelling)** In his initial model, only one date was kept and converted to date: order\_approved\_at. The reasoning behind this was that the information given by the dates ‘order\_delivered\_carrier\_date’ and ‘order\_delivered\_customer\_date’ was already partially given by the ‘order\_status’ column. He considered that the combination of ‘order\_approved\_at’ and ‘order\_status’ were enough for a database destined to be used by sales teams. The other dates, especially the estimated\_delivery\_date or the shipping\_limit\_date were more for a customers relations/logistics team that would need to build a visualization table to follow in real time the progress of the deliveries (some type of KPI-like table where rows or symbols would turn red if the delivery was made after the ‘estimated\_delivery\_date’). In the final version of the group ETL, purchase date and delivered\_date were kept to be able to answer to questions like what parameters can influence the length of a delivery. * **(Modelling)** For the payments, a similar reasoning was adopted when choosing not to keep the payment\_sequential or payment\_installments, considering that this would be more useful for an accounting team than a sales team. |  |  |
| DWH Phase | * **(Modelling)** 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. * **(Technical)** Modifications made on the GPS values * **(Technical/Modelling)** Creation of the column Total\_invoiced, payment\_value deleted, and columns price + freight\_value moved to products table | * **(Modelling)** Moving the city/state information from the customer/seller tables to the geolocation dimension to avoid duplicates. Also, from a data analyst point of view, keeping all the gps geolocations in one table and not splitting them between the customer and seller tables, can allow us to do a visualization comparing customer et seller location on a unique map 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. * **(Modelling)** 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 (those that have multiple items). In the final form of the project, a mixed strategy was adopted. The strategy previously stated was kept for the orders information that could take a lot of memory and processing power such as comments, so a separate DimReview table was created. On the other hand, the rest of the orders information was put in the main fact table to make it easier to query, but there was still a strategy of maintaining the size to a minimum by taking only the most relevant columns (choices made in ODS phase) |  |  |
| Data analysis | * **(technical)** SQL coding * Report comments | * Report comments |  |  |