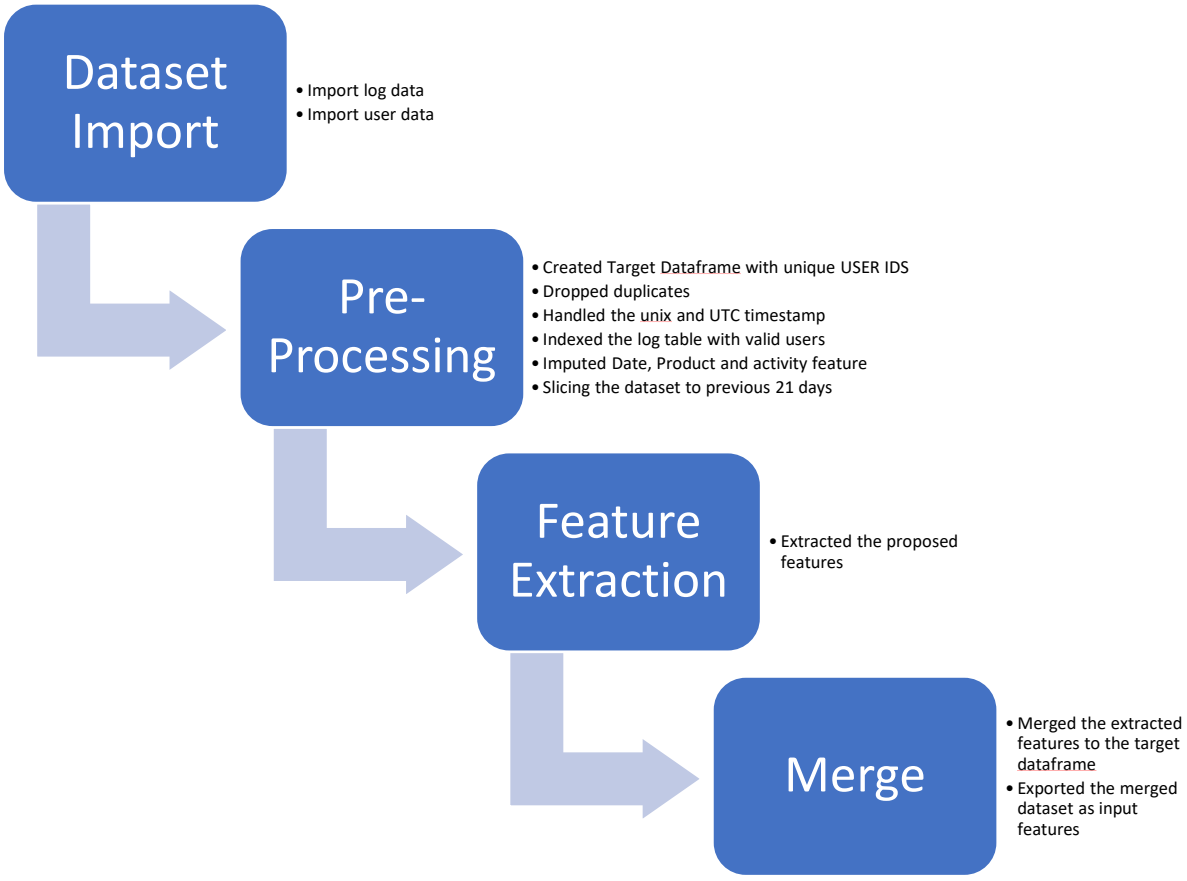
Approach Document: Feature Extraction ETL



Data Pre-Processing Tried:

1) Date Parse:
    a. Selected only date month and year – Not worked out
    b. Selected timestamp with nanoseconds – Worked out
2) Dropped duplicates
3) Datetime Imputation:
    a. Tried dropping null dates – Dint work out
    b. Tried with imputing null values with the Start date of the dataset(21 days prior to the current date) – Dint work out
    c. Tried imputing the missing dates with min of the product group per user – Dint work out
    d. While looking at the dataset(Sorted with user ID and Date) there seems to be a pattern with product ID grouped with median value per product group, Imputed null date values as the mid day of the user per product – Worked out
4) Activity, Product imputation:
    a. Forward filled the product ID with sorted date for every user, then grouped the user id,product to forward fill the activity feature as well

Feature Extraction Tried:

5) Vintage:
    a. Tried float values of days – Not worked out
    b. Tried days as integers rounded – Worked out
6) No of products viewed:
    a. Tried only with pageloads – Not worked out
    b. Tried with only Clicks – Not worked out
    c. Tried picking the unique product viewed by the user – Worked out

Feauture extraction was straight forward once data imputation is done.

Imputation:

1) Kept null as null from the log dataset with other values(UNIX and UTC) converted to timestamp
2) Sorted the Dataframe with timestamp per user
3) If a product inbetween two same products, imputed the product ID(Using Mask)
4) Filled the null timestamp with the median of the product group
5) Grouped the product ID to forward fill the activity column