Predictive Analytics & Dashboarding for ATD Optimization

Miguel Angel Fernandez Castresana

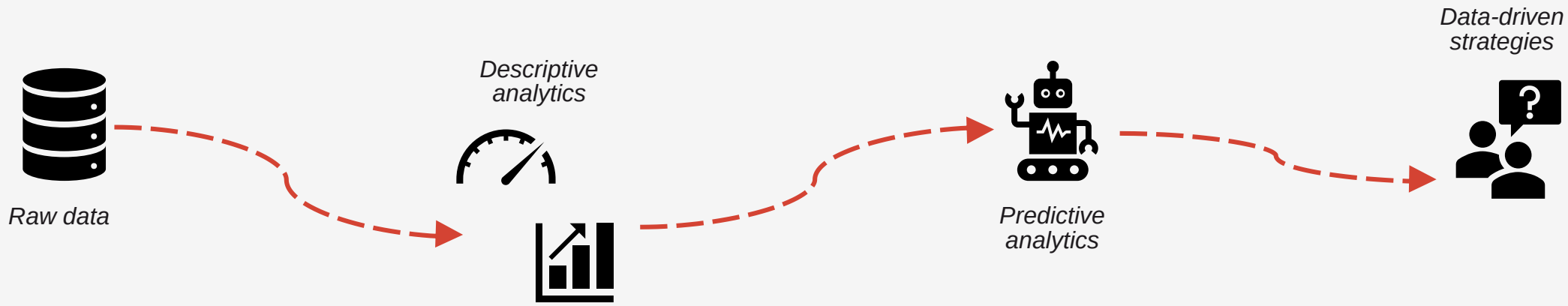
Index

Uber Eats

Business context

Context and problem statement

Data	***Expected outcome***	***Need***
Comprehensive dataset detailing deliveries across the marketplace in March and April for Mexico	Deliver high-quality descriptive and predictive insights to our customers, empowering them to formulate data-driven strategies; focused in improve the ATD	Develop an automated data extraction pipeline to directly support and enhance descriptive and predictive analytics efforts

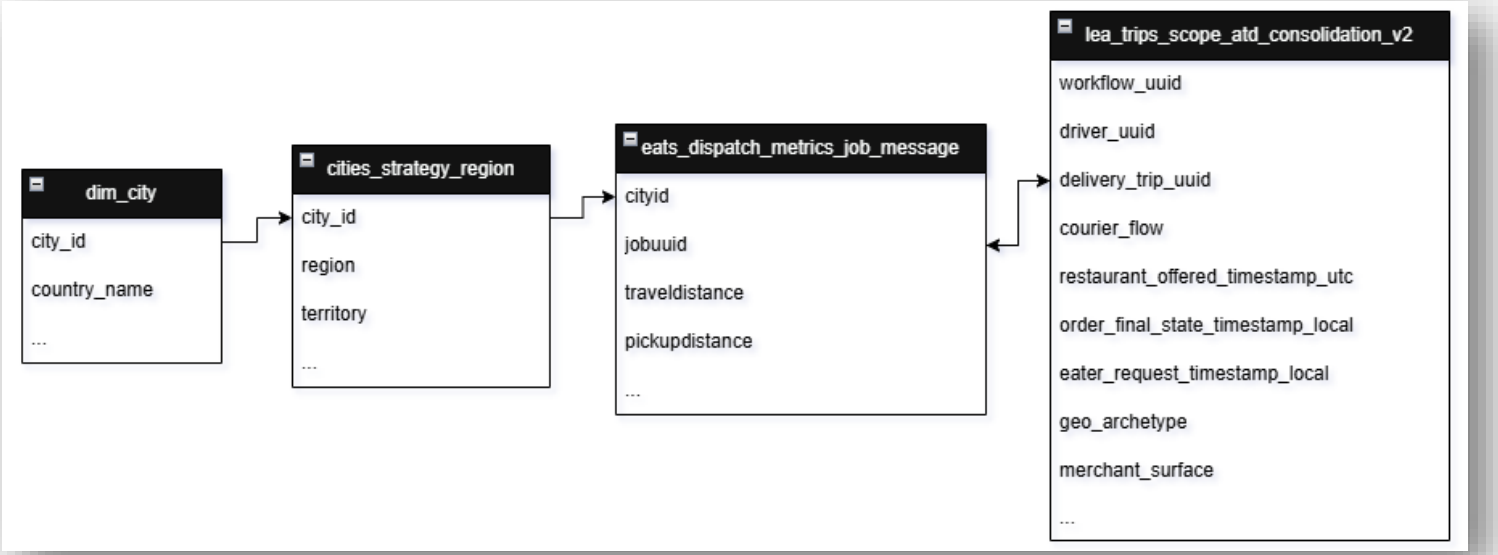


Data extraction

Data model & query composition

Along the query to extract the information, some modifications and adaptations that are important to keep in mind:

- As the date and numerical columns have different format and in some cases null values represented by the text '\N', the query is treating these columns as VARCHAR type and the transformation to the correct format is done along the query
- As in Mexico we have three different hour time zones and we don't have the specification of the state or the *currency_code*; I'm inferring the hours of difference with the *eater_request_timestamp_local* variable
- As asked, for every date provided in the variable *{ds},* we are providing the deliveries data of the past week, here the calculated week is considered to run from Monday to Sunday

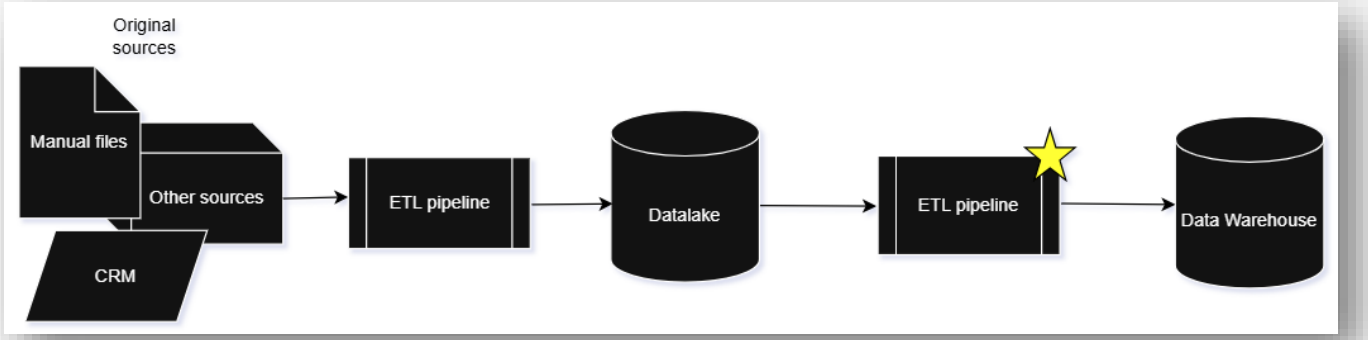


Data extraction

Proposed pipeline

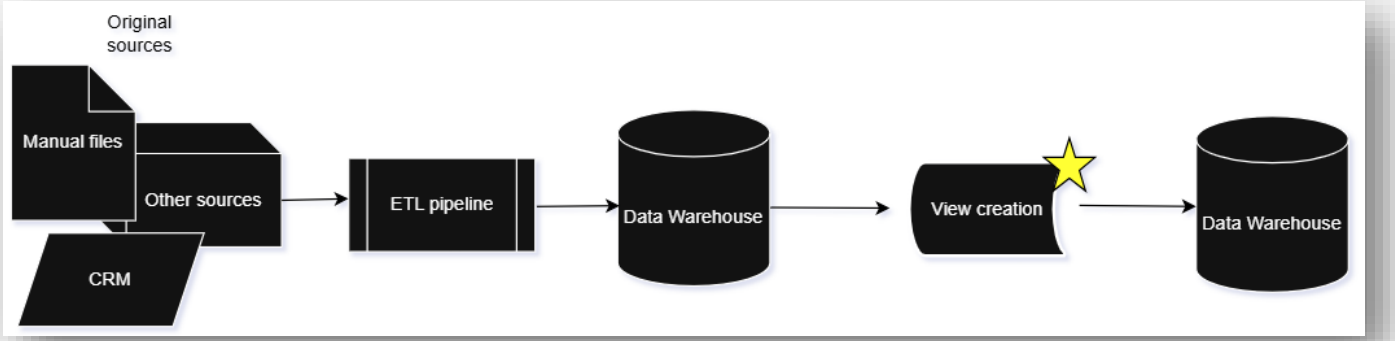
Case 1

The final table where we will storage the data (AA_tables schema) is in a different storage space than the one where the data is coming from, for example the data is coming from the Datalake (Query.sql execution highlighted with a star)



Case 2

The final table where we will storage the data (AA_tables schema) is in the same storage space than the one where the data is coming from (different schemas in the same database) (Query.sql execution highlighted with a star)



Streamlit dashboard

General composition

Composition

The dashboard is divided in two pages:

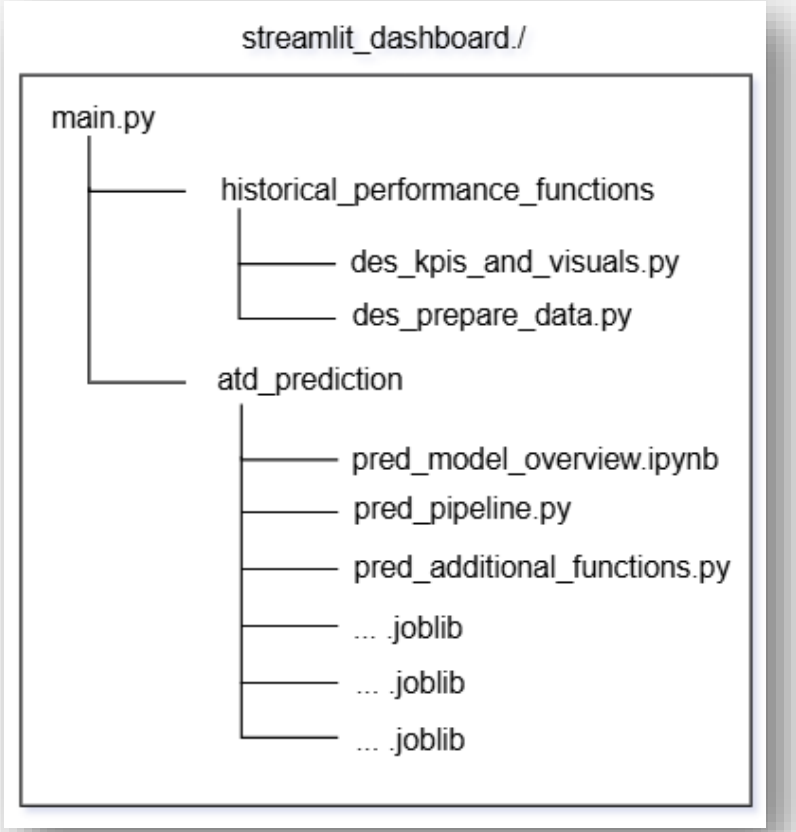
Historical & Trend analysis

- Creation of multiple descriptive visualizations to clearly understand data patterns across time, velocity, distance, location, and day of the delivery
- KPIs to track the overall performance
- Development of a time series trend graph integrating SARIMA predictions for the next fourteen days for both the ATD and total delivery volume

ATD prediction

- Analysis of XGBoost feature importance to identify the variables with the highest impact on the ATD prediction
- Implementation and comparison of Random Forest, Multilayer Neural Network (MLNN), and XGBoost models, reporting the specific error metric for each ATD prediction model

Modularization



ATD predictions

ATD variable composition

Formula

$$ATD\ (Actual\ Time\ of\ Delivery) = Seconds(order\_final\_state\_timestamp - Converted\ to\ local\ time\ (restaurant\_offered\_timestamp\_utc)) / 60$$

Minutes

$$= Seconds(order\_final\_state\_timestamp - (restaurant\_offered\_timestamp\_utc - Hours(restaurant\_offered\_timestamp\_utc - eater\_request\_timestamp\_local))) / 60$$

In the created query to extract the information, as we don't have any information about the state or the currency_code, we are inferring this conversion, calculating the difference in hours between *restaurant_offered_timestamp_utc* and *eater_request_timestamp_local*

Example

$$ATD\ (Actual\ Time\ of\ Delivery) = \text{Seconds(2025-03-29 09:34:09 – 2025-03-29 13:57:11 – Hours(2025-03-29 13:57:11 – 2025-03-29 08:57:11)) / 60}$$

$$= \text{Seconds(2025-03-29 09:34:09 – 2025-03-29 13:57:11 – 5 hours) / 60}$$

$$= 2218 / 60$$

$$= 36.97 \text{ minutes}$$

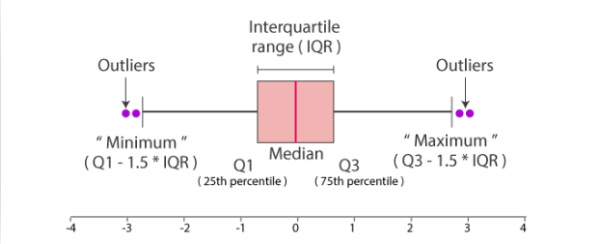
Uber Eats

ATD predictions

Missing values, Outliers & ATD transformation

Outliers

Methodology



- Outliers were removed after identification; imputation was deemed unnecessary to preserve the distribution because our sample size was large enough, and limitation could potentially bias the results
- The upper bound threshold was specifically modified to prevent 'cutting' or artificially truncating the natural distribution of the remaining data:

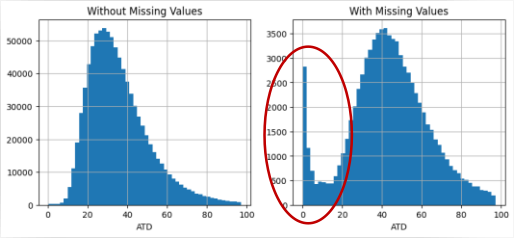
```
IQR = Q3-Q1
upper_bound = Q3 + 2.5 * IQR
lower_bound = Q1 - 1.5 * IQR
```

- The total number of deletions represented 1.5% of the overall sample size

Missing values

Exploration steps

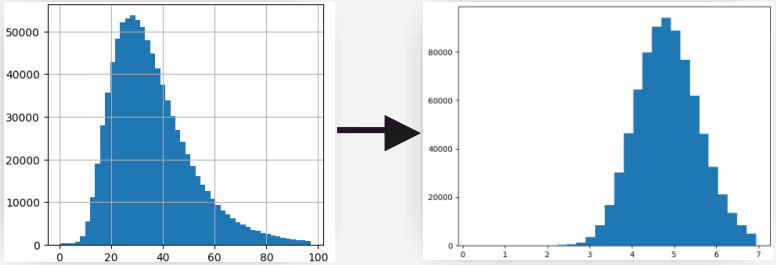
1 – Discard correlation with specific categorical variables
2 – T-test used to discard significant differences between the samples



- A unique ID column for the missing values from all the columns was created to do the analysis
- We the T-test we confirmed the significance of the missing values; as this cluster showed a significant low ATD value it is inferred that the cases with missing values are errors (cancelled deliveries, deliveries that were not addressed, …)
- The missing values were removed from the sample as them are distributed along the multiple categorical variables; imputation was deemed unnecessary to preserve the distribution because our sample size was large enough again

ATD variable

Transformation



- With the objective of achieving a Normal distribution for the Adjusted Total Delivery (ATD) variable, we are applying the Yeo-Johnson transformation (Box-Cox extension)
- With this transformed variable we will be training the models to ensure the best possible adjustment and predictive performance (when making a prediction we will use an inverse of this transformation)

ATD predictions

Feature engineering

territory | courier_flow | geo_archetype | merchant_surface | weekday | hour_of_day | distances

feature_territory_average_velocity (FunctionTransformer) | feature_map_categories_courier_flow (FunctionTransformer) → OneHotEncoder | feature_map_categories_archetype (FunctionTransformer) → OneHotEncoder | OneHotEncoder | feature_day_cyclical (FunctionTransformer) | feature_hour_cyclical (FunctionTransformer) | PowerTransformer

Instead of including the territory names as a categorical feature, we are incorporating the average velocity in Kilometers per Minute (Km/Minute)

```
territory_map = {
    'Central': 0.1187,
    'Long Tail - Region': 0.1567,
    'North': .1715,
    'South East': .1665,
    'West': .1725
}
```

This calculated feature will help the model better understand the nature of the territory's traffic and local dynamics

First, we apply a categories mapping (grouping) to combine those categories that represent a small percentage of the total data, this to reduce the overall cardinality

```
rare_classes = ['UberX',
                'Fleet',
                'SUV',
                'Onboarder']
new_name = 'UberX_Fleet_SUV_Onboarder'
```

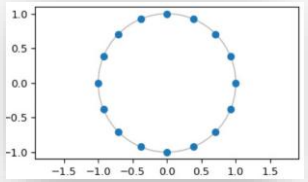
```
rare_classes = ['Build experience',
                'Unspecified',
                'Unlaunched']
new_name = 'Build experience_Unspecified_Unlaunched'
```

Then, we apply the use of a OneHotEncoder to translate the resulting categorical variables into a numerical format suitable for model training

Usage of a OneHot Encoder to translate the resulting categorical variables into a numerical format

We are applying a Cyclical Time Encoding, translating our increasing time variables into cyclical variables with the *sin* and *cos* coordinates:

- *Hour:* 0, 1, 2, … 22, 23 *(m= 24)*
- *Day:* 0, 2, … 6 *(m=12)*



$$CoordinateY = \sin\left(\frac{2\pi t}{m}\right)$$

$$CoordinateX = \cos\left(\frac{2\pi t}{m}\right)$$

With this new mapping, through the coordinates, as it is a mapping of a circumference, our first and last records will be side by side, ensuring that proximity is simulated

Similar to the ATD variable, we will apply a Yeo-Johnson transformation here as well

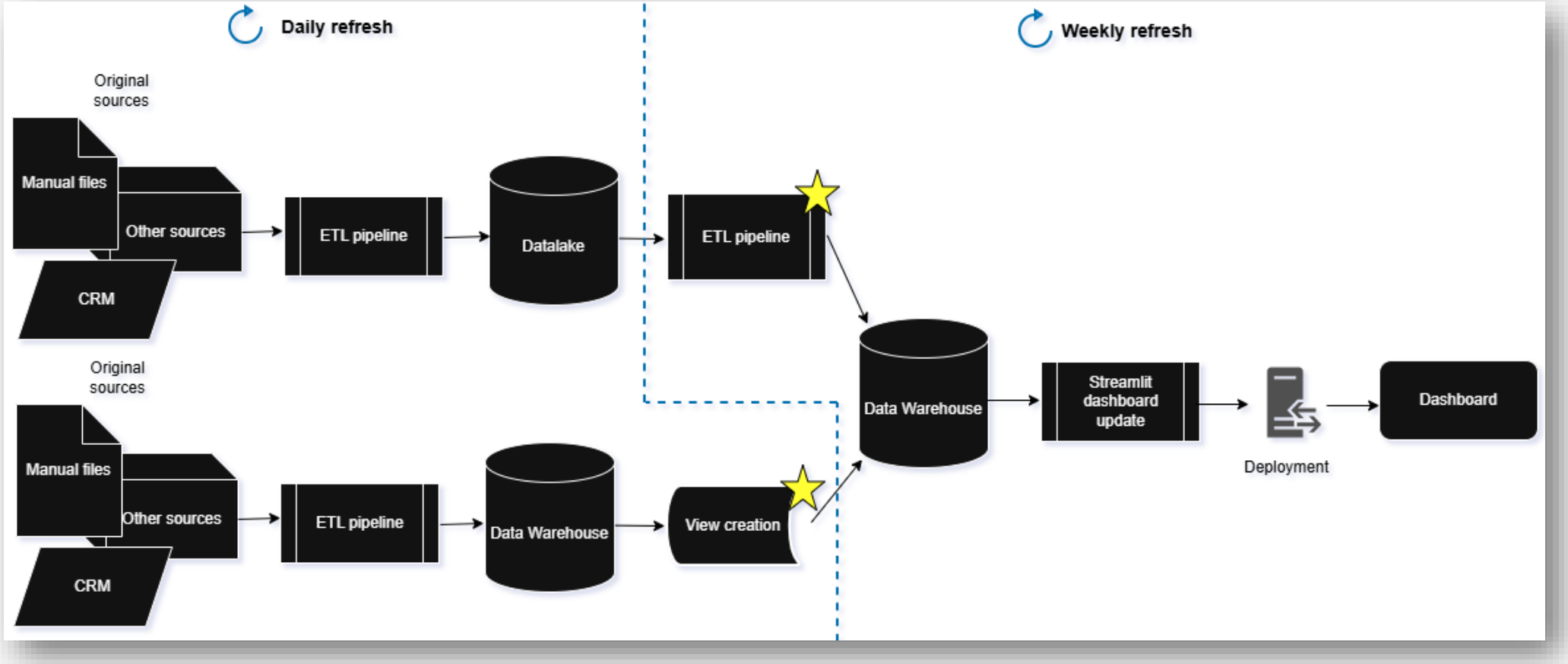
We are also applying standardization (scaling), this ensures the variable has the same magnitude as the rest of the features

Uber Eats

ATD predictions

Proposed pipeline

Independent from the case defined for the query update, the proposed pipeline to refresh the dashboard is the following (Query.sql execution highlighted with a star):



This separate pipeline ensures the dashboard's data availability remains unaffected by ongoing query development