Developing Innovative Metrics for Applaudo Studios' Data Lakehouse

Internship at Applaudo Studios

Javier Cladellas

Applaudo Studios

Applaudo is a digital solutions company that help brands streamline their IT solutions, optimize delivery costs, and speed up their digital transformation.

Solutions:

- Digital Transformation
- Web and Mobile Development
- Cloud Computing
- AI and Machine Learning

Fig. 1. Hero image of the 'our values' page of the Applaudo Studios Website. (https://applaudo.com/our-values/)

Applaudo’s Timeline

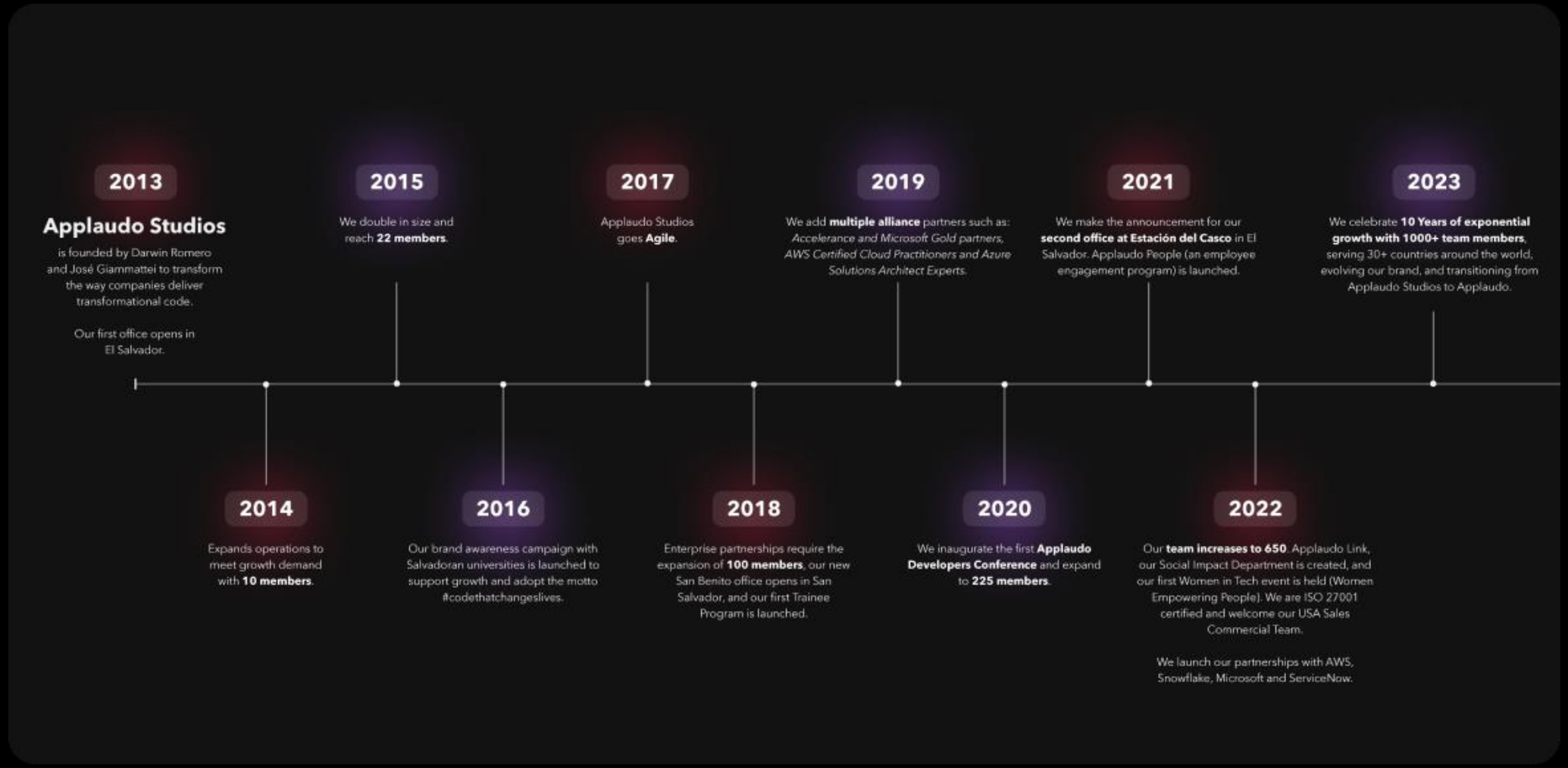

Fig. 2. Applaudo Studios’ timeline (https://applaudo.com/our-story/)

Fig. 3. Applaudo Studios' presence on the world (https://applaudo.com/our-story/)

Notable Clients

Context

- Need to accelerate company’s internal processes.
- Need to improve productivity of various departments.

Problems

- Data is fragmented
- Data extraction is manual

Solution

Develop a Data Lakehouse

Data Lakehouse

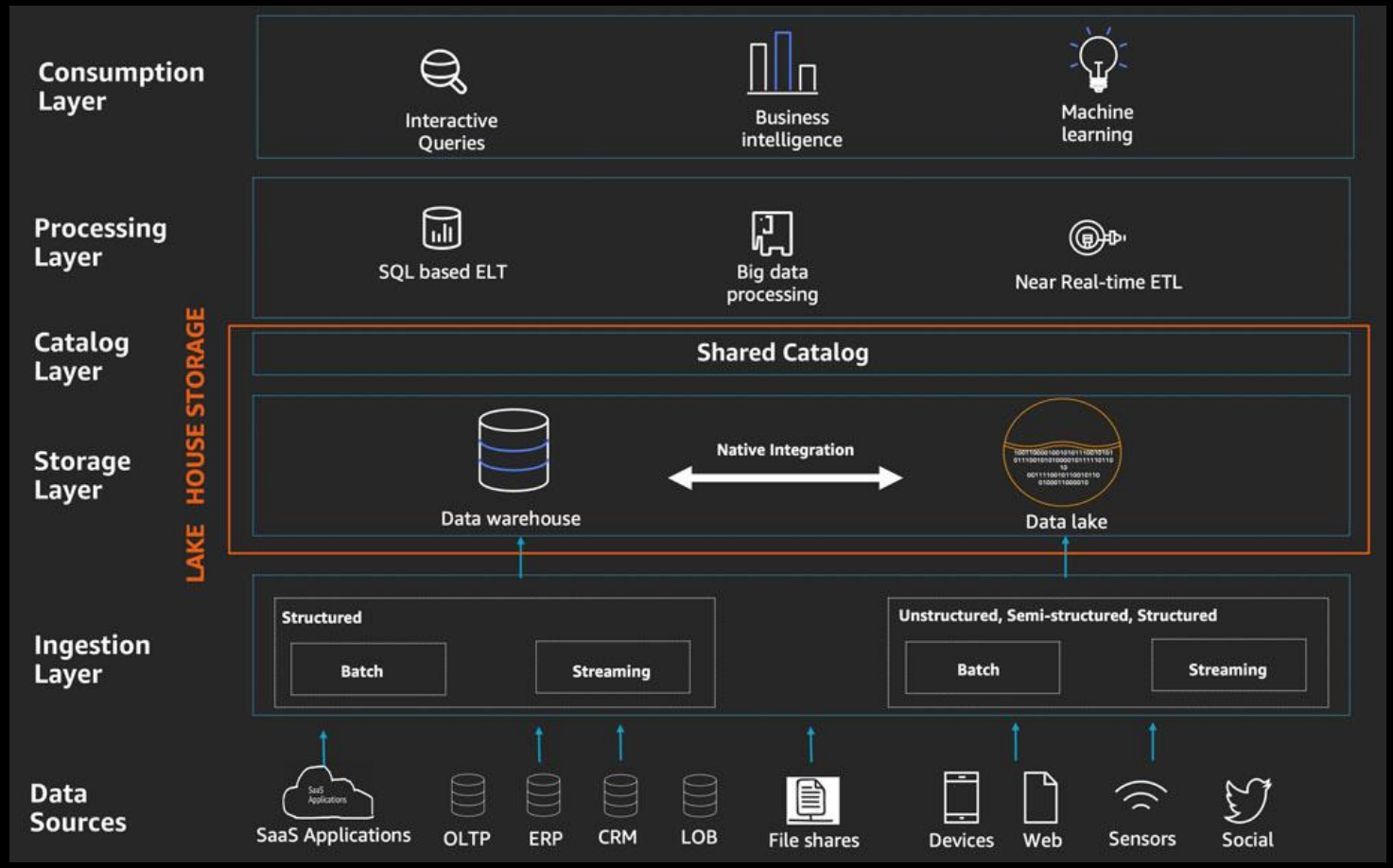

Fig. 4. Data Lakehouse architecture according to AWS. (https://aws.amazon.com/blogs/big-data/build-a-lake-house-architecture-on-aws/)

Internship’s Objectives

Propose, design and develop new metrics for the data lakehouse

- Propose new informative metrics
- Develop the metrics with a processing and transformation pipeline
- Propose a way to integrate the results into the data lakehouse
- Optimize the code and scripts to reduce computing and storage costs
- Document the process

Tools

Python

- Pandas
- Scipy
- Scikit-learn
- Tensorflow
- Statsmodels
- Plotly

Google Cloud Platform

- Google Cloud Storage
- BigQuery Storage
- BigQuery ML
- GoogleSQL

PowerBI

- DAX
- M language

Metrics

- Headcount forecasting
- Income forecasting
- Project Clustering
- Billable Percentage visualization
- Leaves Cost visualization

Headcount forecasting

Metric Description

Use employee information from Applaudo's roster to predict the future number of employees of the company, filtered by departament or technology.

Possible Uses

- Workforce planning
- Budgeting and ressource allocation
- Talent acquisition

Data sources

- Zoho Roster

SARIMA models

Seasonal AutoRegressive Integrated Moving Average

Let Y_t be the value of a time series at time t.

Let Y_t^d be the d-th difference of Y_t, given by the $I(d)$component.

$$\hat{Y}_t^d = c + \underbrace{\sum_{i=1}^{p} \phi_i Y_{t-i}^d}_{AR(p)} + \underbrace{\sum_{i=1}^{q} \theta_i \varepsilon_{t-i}}_{MA(q)} + \underbrace{\sum_{i=1}^{P} \Phi_i Y_{t-iS}^{D+d}}_{SAR(P,S)} + \underbrace{\sum_{i=1}^{Q} \Theta_i \varepsilon_{t-iS}}_{SMA(Q,S)} + \varepsilon_t$$

Forecasting

- Aggregate the data into a time series.
- Find the parameters of the SARIMA model:

Terms	Non-seasonal	Seasonal
AutoRegressive	p=1	P=1
Moving Average	q=1	Q=1
Differencing	d=1	D=0

- Test the results
- Predict future data

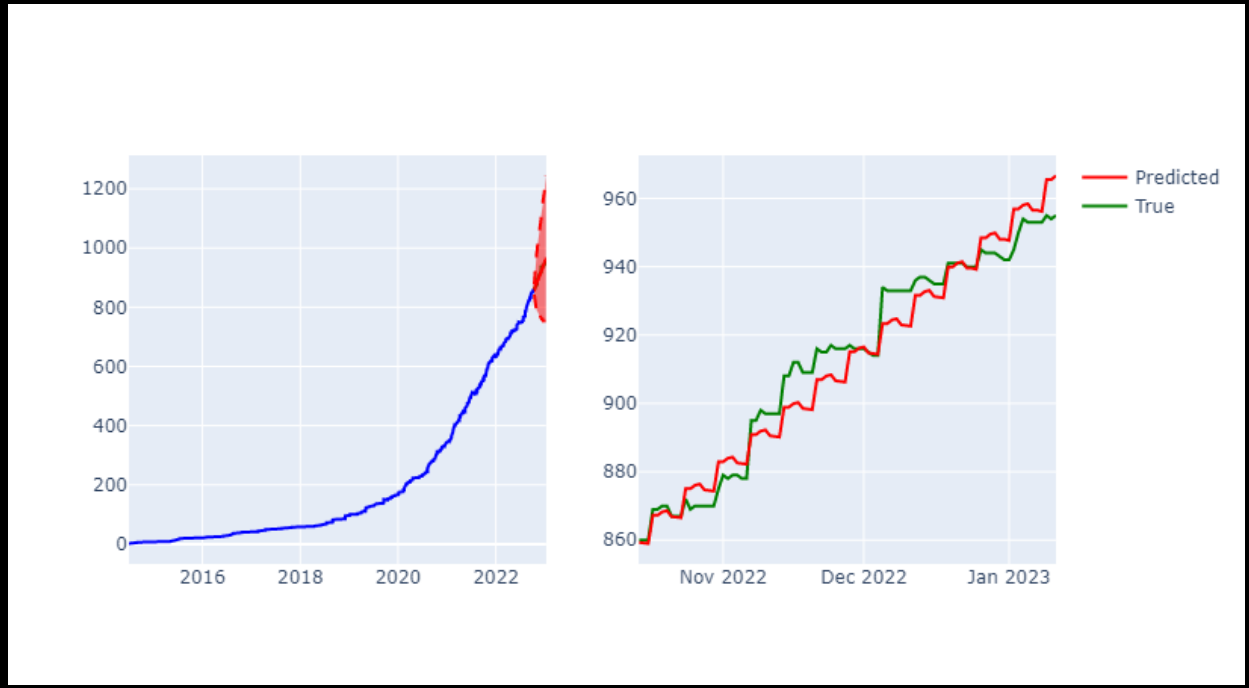

Fig. 5. Predicted and true total headcount using the SARIMA(1,1,1,1,0,1,7) model.

Income forecasting

Metric Description

Use historical information on the company's time report to predict the total billable amount, the worked hours and the billable rate, over a long term period.

Possible Uses

- Financial planning
- Performance evaluation
- Marketing and sales

Data sources

- Harvest time report

LSTM

Long Short-Term Memory

- Useful for long sequential data
- Remembers and forgets information over time

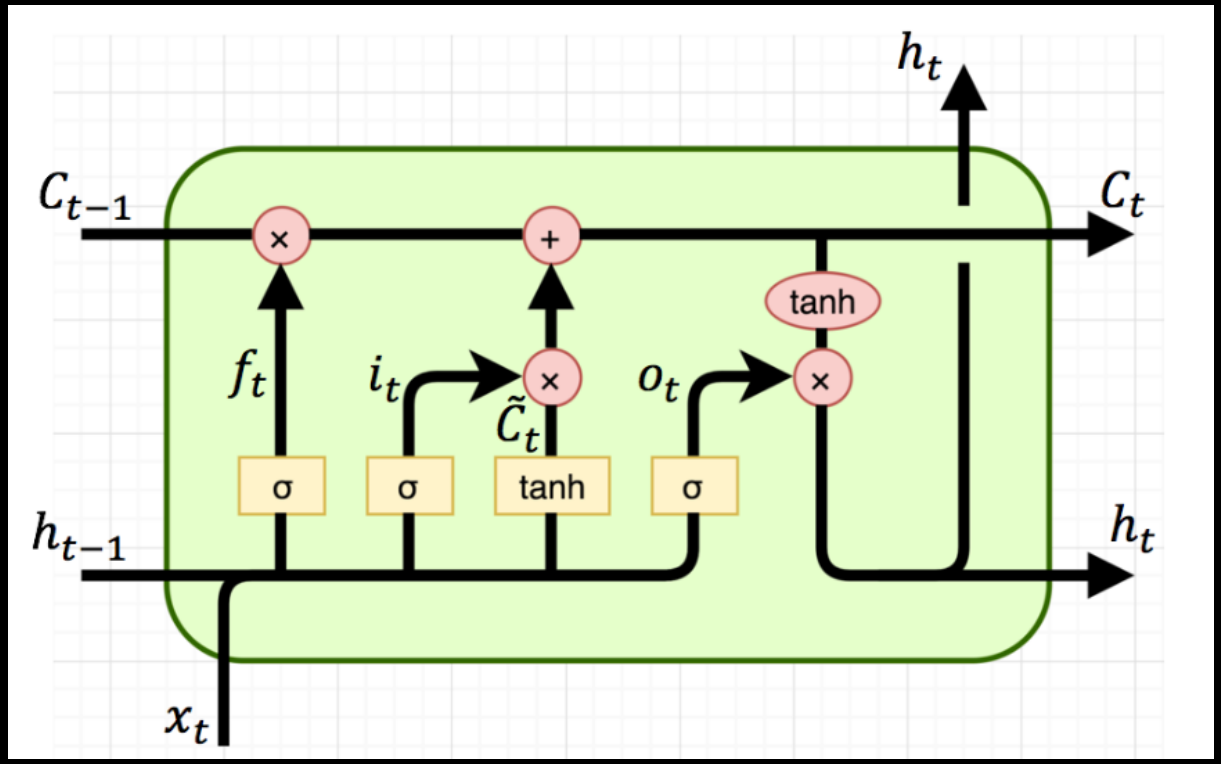

Fig. 6. Schema of an LSTM cell.
(https://www.researchgate.net/figure/Representation-visuelle-dune-cellule-LSTM_fig4_365072370)

Pre-processing

- Aggregate and re-index by a daily frequency.
- Remove outliers.
- Reindex by week to reduce noise.

Fig. 7. Visualization of daily income, cost and hours just after aggregating and re-indexing.

Pre-processing

- Aggregate and re-index by a daily frequency.
- Remove outliers.
- Reindex by week to reduce noise.

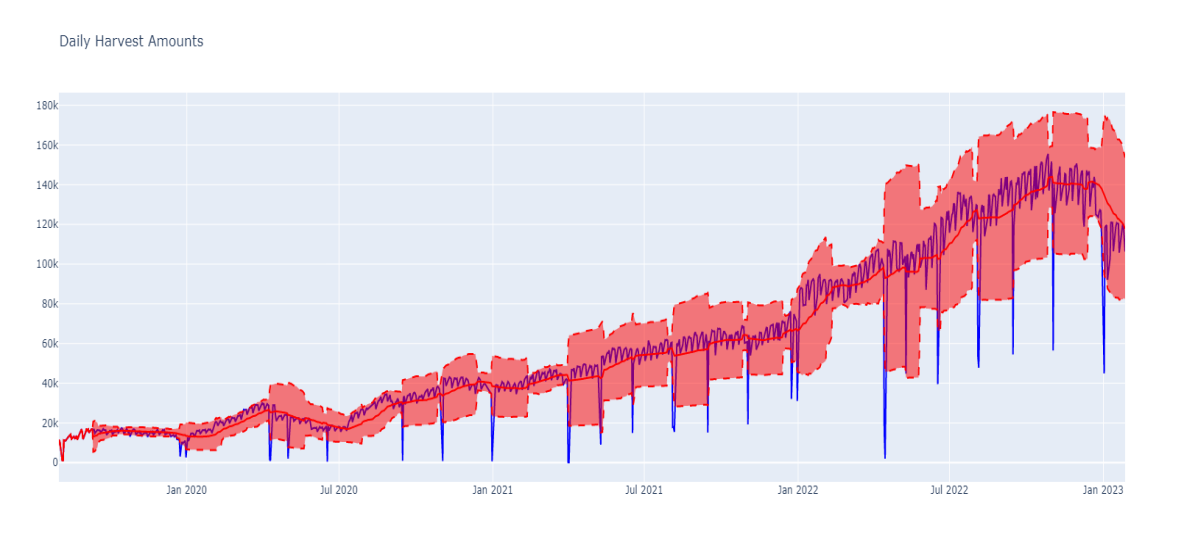

Fig. 8. Representation of an outlier detection method.

Pre-processing

- Aggregate and re-index by a daily frequency.
- Remove outliers.
- Re-index by a weekly frequency to reduce noise.

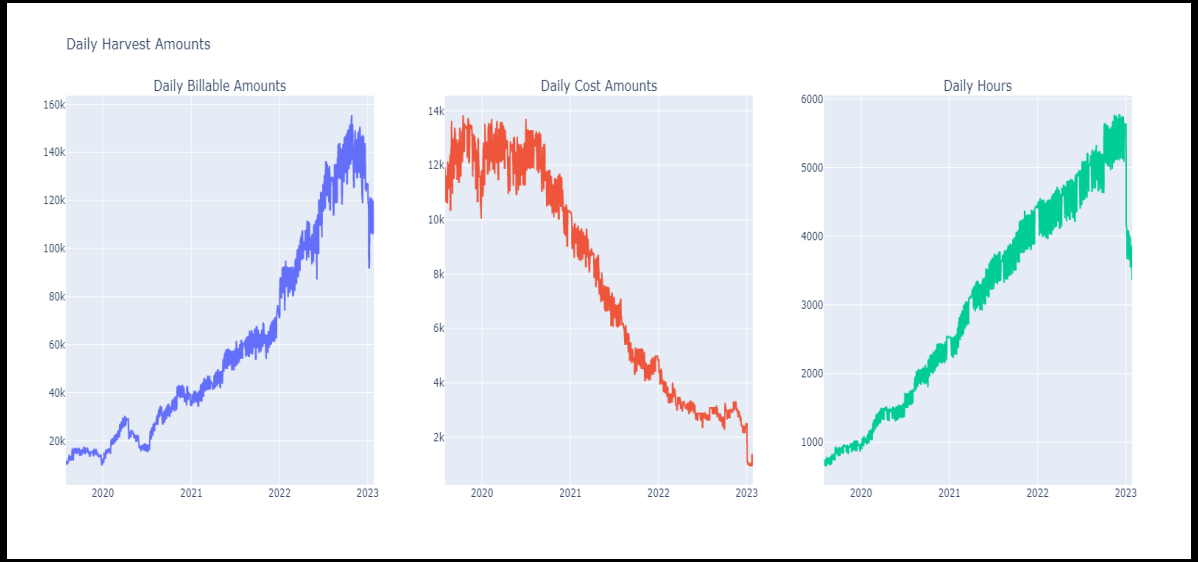

Fig. 9. Visualization of daily income, cost and hours after removing outliers.

Forecasting

- Data is reshaped into windows
- Hyperparameter tuning using a grid search
- On GCP:
  - Preprocessing–training–saving (BigQuery/python)
  - Processing-predicting (BigQueryML)

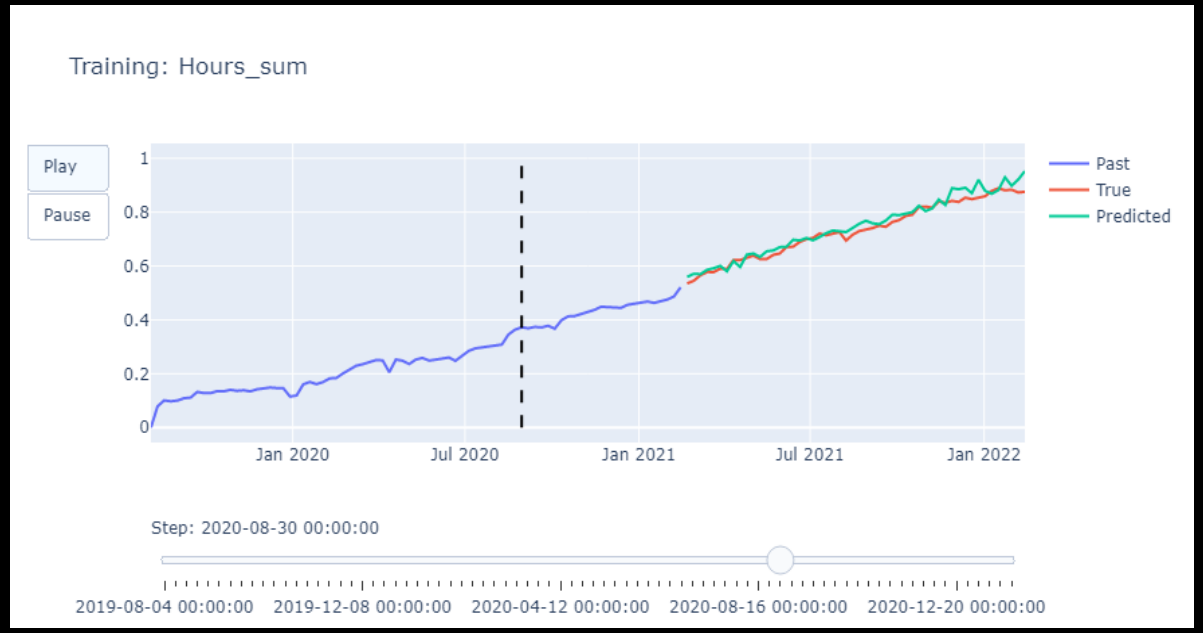

Fig. 10. Worked hours on a training window of the LSTM model.
The model shows accurate predictions of the window without overfitting.

Project Clustering

Metric Description

Group similar projects depending on their progress, based on multiple features.

Possible Uses

- Ressource allocation
- Performance comparison
- Project prioritization

Data sources

- Harvest time report
- Forecast

K-means algorithm

Iterative method :

- Define number of clusters (Elbow method)
- Randomly initialize the centroids
- Assign data points to nearest centroid
- Recalculate centroid position as the mean of the current cluster points

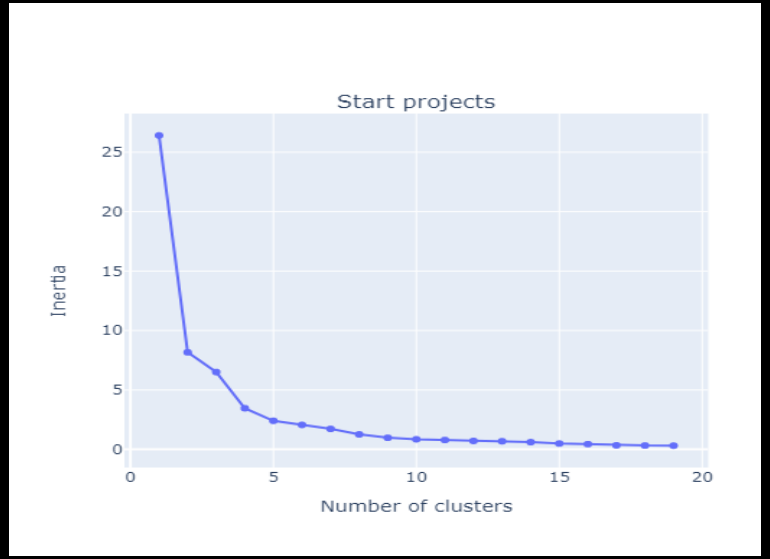

Fig. 11. Representation of the elbow method used to detect the optimal number of clusters for the projects with the 'start' label.

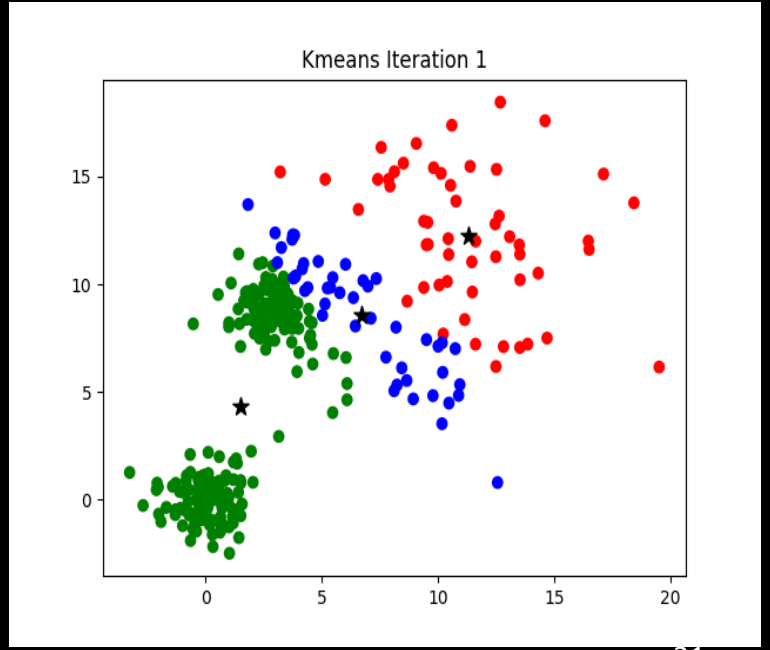

Fig. 12. Iterations of the K-means algorithm. https://towardsdatascience.com/k-means-clustering-from-a-to-z-f6242a314e9a

Project Clustering

- Compute the project's progress and split (using forecasted hours and time reports)
- Normalize (min-max)
- Determine optimal number of clusters for each project stage
- Apply K-means algorithm
- Plot the groups on 2D space using PCA
- For a requested project:
  - Compute the distance of each cluster point to the requested one
  - Return a sorted list of similar projects

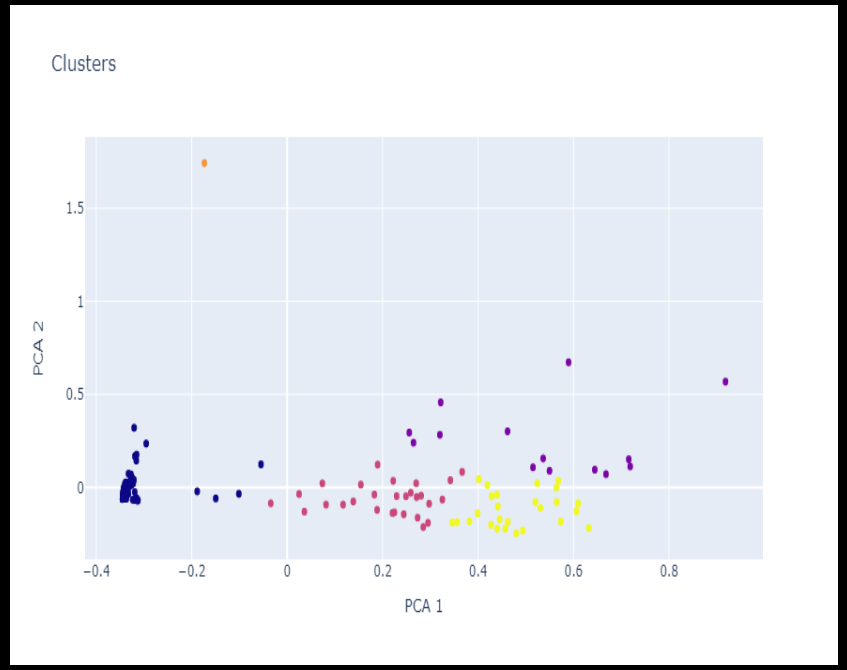

Fig. 13. Clusters for the projects with the 'start' label.

Billable Percentage Visualization

Metric Description

Generate descriptive figures regarding the billable percentage of the company's clients and projects.

Possible Uses

- Financial Performance Analyis
- Performance Monitoring
- Trend identification

Data sources

- Harvest time report
- Zoho Roster

$$BP = 100 * \frac{Worked\ billable\ hours}{Available\ hours}$$

Notable visualizations

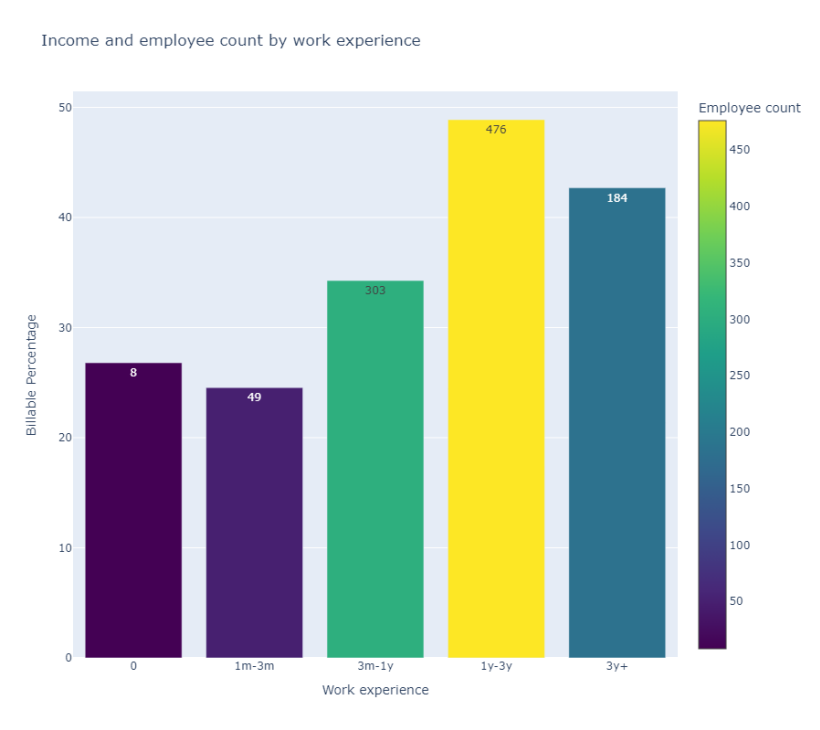

Fig. 14. Billable percentage and headcount by work experience.

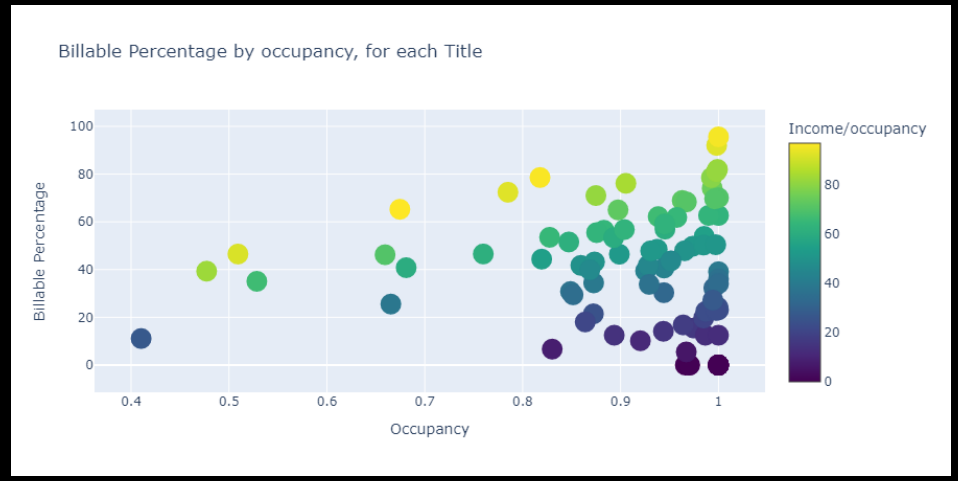

Fig. 15. Job titles depending on their average billable percentage and occupancy.

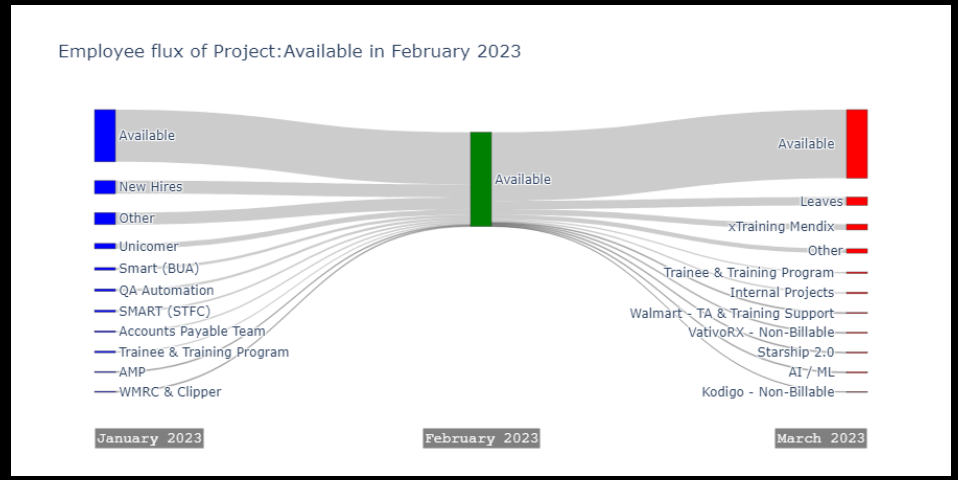

Fig. 16. Employee flux of employees in the 'available' project for February 2023.

Leaves Cost Visualization

Metric Description

Show the impact of the leave reasons on the cost and worked hours of the company.

Possible Uses

- Ressource allocation
- Policy adjustments
- Health improvements
- Performance evaluation

Data sources

- Harvest time report

Notable visualizations

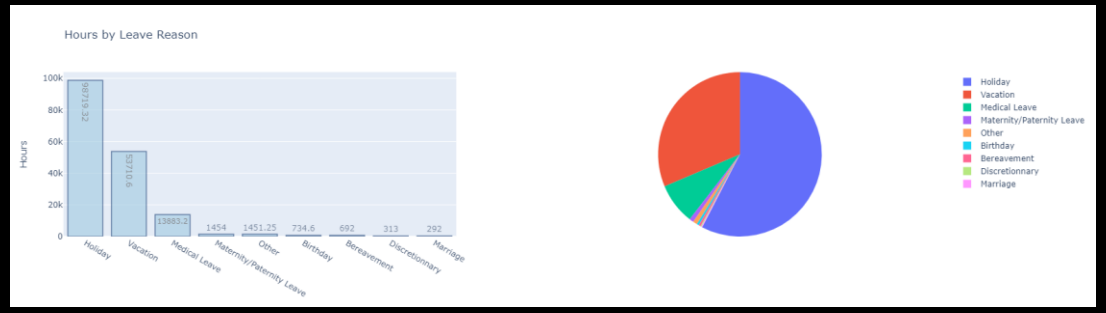

Fig. 17. Total worked hours by leave reason.

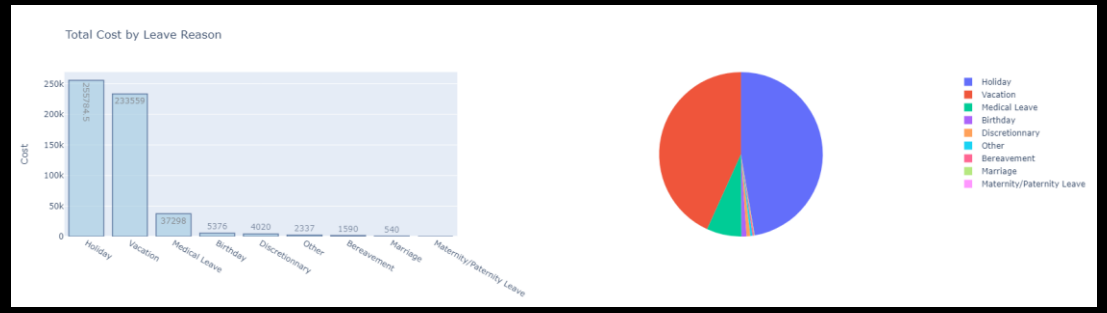

Fig. 19. Total cost by leave reason.

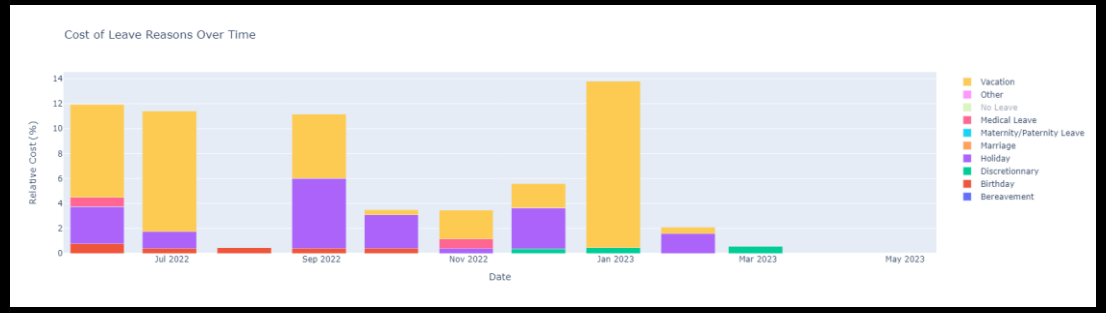

Fig. 17. Stacked bar chart showing the monthly relative cost of each leave reason.

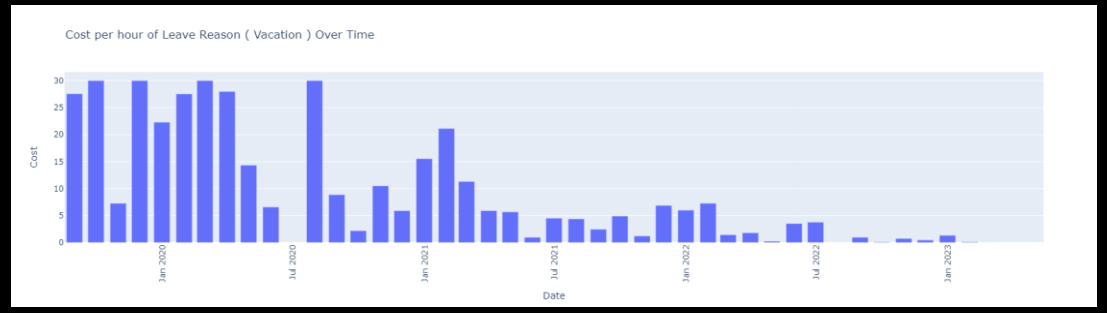

Fig. 20. Bar chart showing the monthly cost of the 'vacation' leave reason.

Conclusions

- Metrics were developed locally, using the available manually downloaded files from the data sources.
- A GCP implementation was proposed for each metric.
- When the data lakehouse is deployed, the metrics should be adapted to the final structure and column names.
- For each view, a cost and optimization analysis should be done to choose whether to calculate it in Power BI or store it in BigQuery Storage.