Enefit: Energy Behavior Analysis of Prosumers in Estonia

Forecast Prosumer Energy Patterns and Minimize Imbalance Costs

---

CSCI E-103 Final Project

DataGeeks (Group 4)
Michael Leung
Manos Sinha
Melahat Tayli
Alex Coward
Atul Kant
Nihal Pai

Business Use Case

Motivation

- Electricity consumption & generation vary every minute, causing energy imbalance.
- Solving energy imbalance is inherently difficult and requires careful planning.
- Undersupply results in inconvenience to consumers, and economic losses to factories and businesses.
- Oversupply results in price reduction and potential damages to generators, due to a higher rotational speed.
- Solar panel adds complexity to building models for forecasting as there are multiple elements to factor in

Proposed Solution

- Two t-series models to forecast on the consumption and production data on a hourly basis.
- The net of the two will be the energy imbalance (surplus / deficit) that the energy company will have to expect, and prepare for. This is where this project adds value.
- The features of our models include but are not limited to: historical data consumption, weather info (forecast and historical), gas price, electricity price among others.

Agenda

01 **Data Architecture**

ER Diagram & Data Flow Diagram Overview

02 **Data Engineering**

Ingestion Within Medallion Framework

03 **Machine Learning**

Forecasting Model Using AutoML

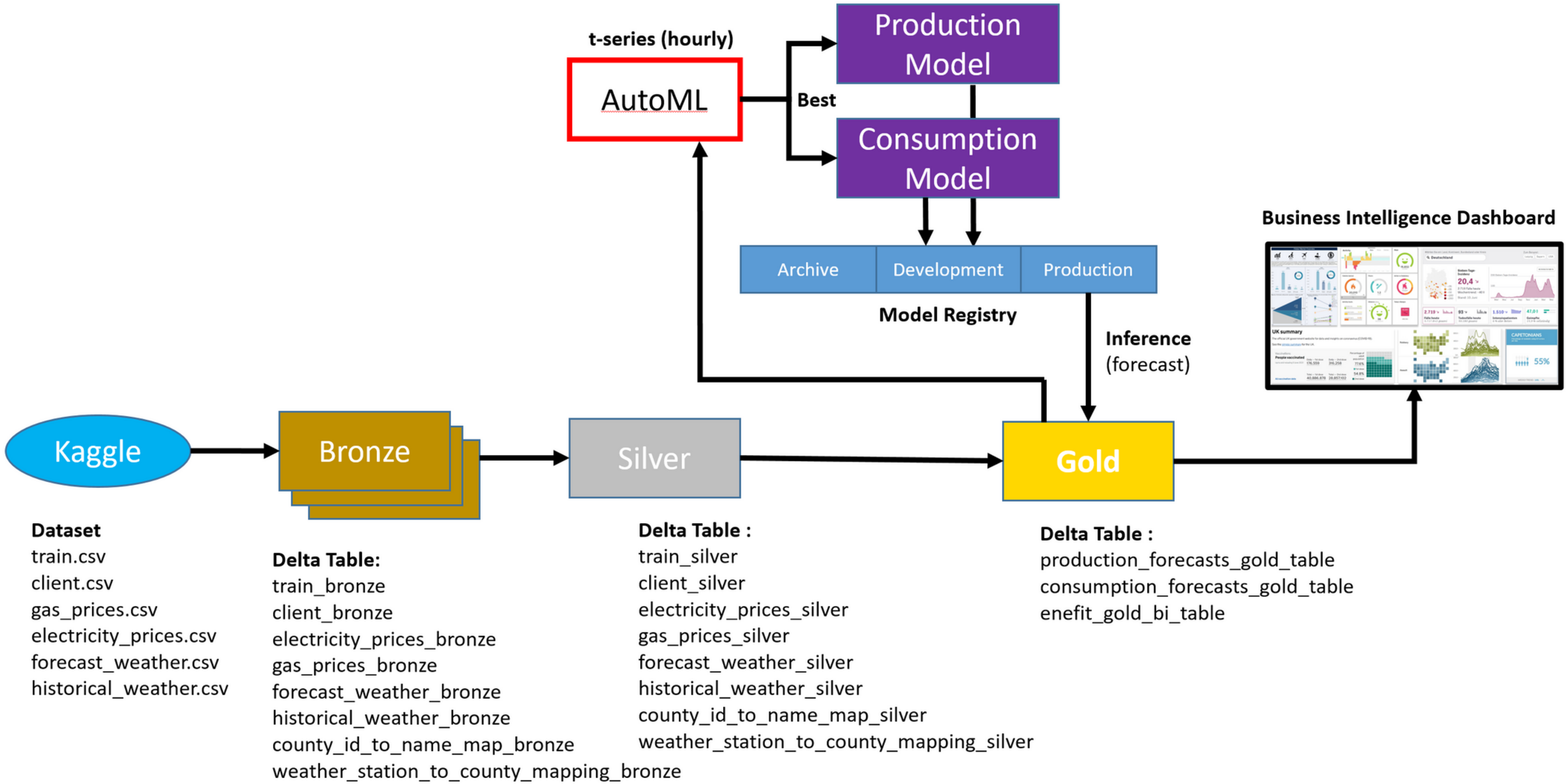
04 **BI**

Forecast Visualizations and Insights

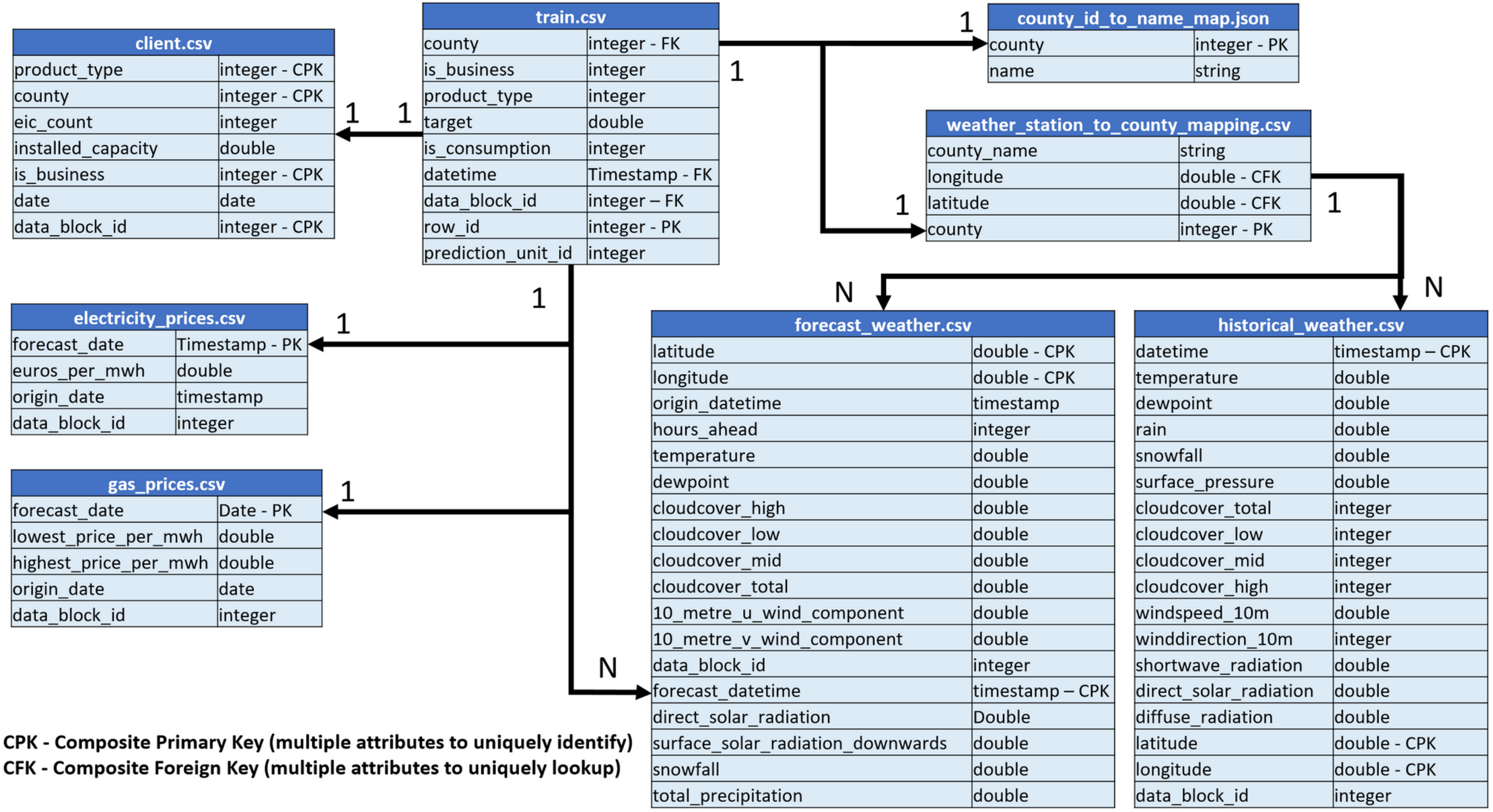
05 **Other Design Considerations**

CI/CD, DR, Partition Choices, What-if Streaming

Data Flow Diagram



ER Diagram (Raw Data)



CPK - Composite Primary Key (multiple attributes to uniquely identify)
CFK - Composite Foreign Key (multiple attributes to uniquely lookup)

Data Engineering

Step 1 : Exploratory Data Analysis (EDA)

- Explore the categorical and continuous features
- Define primary keys for each table (combination of distinct features)
- Determine null values and duplicated rows
- Relationship of features at table level for feature correlations and trends
- Foreign key relationship of tables

Categorical Features

- County
- Business/Non-business
- Product type

Dates

Gas Prices: **Daily**
Electricity Prices: **Hourly**
Weather Data: **Hourly**

Continuous Features

Installed **Capacity**
Gas Price Predictions
Electricity Price Predictions
Weather Forecasts
Historical Weather Data

Target

1. **Consumption** Amount
2. **Production** Amount

Data Engineering

Step 1 : Exploratory Data Analysis (EDA)

- Over 2M training data points
  (0.03% missing values dropped)
- 15 counties of Estonia
  (data with unknown county is dropped)
- 112 weather stations for weather forecasts
  (56% null values)

Weather stations not located in Estonian counties

Weather stations located in Estonian counties

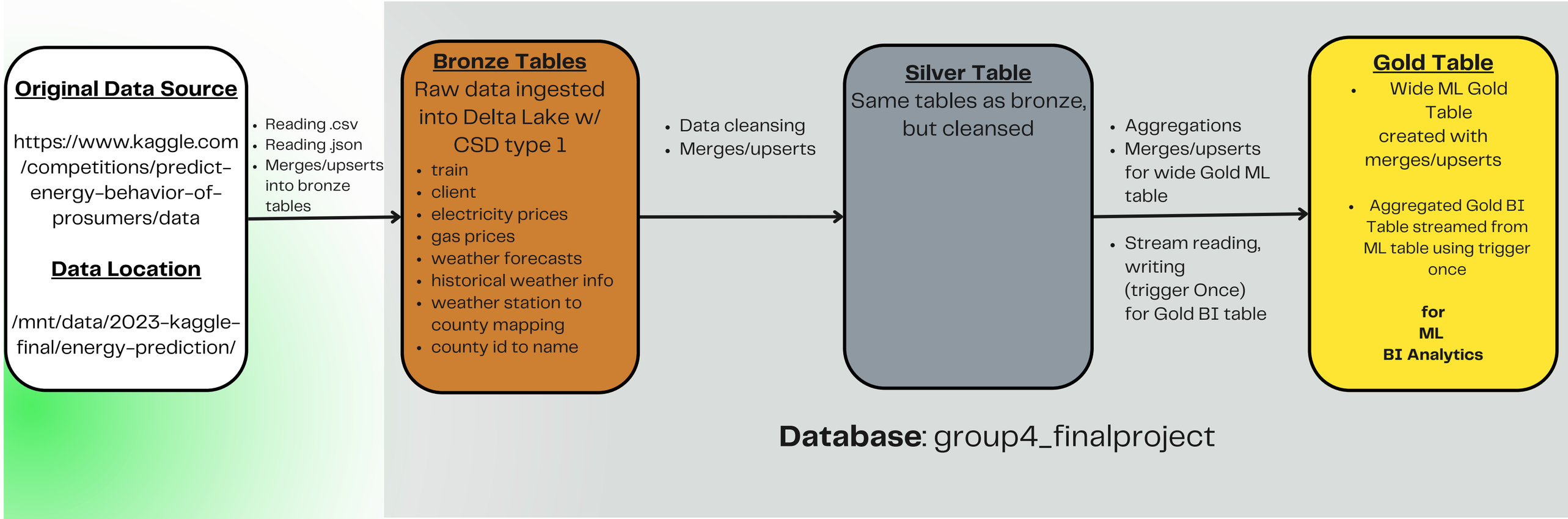
Imputation of Null Counties in weather station dataset:
weather stations **mapped** to the **closest county** in silver table
(using python geopy library)

- Daily gas and hourly electricity prices
  (available in train data with 1 day lag)
- Weather forecasts for the next 48 h
  (only 24 hours of forecasts is used, forecasts are available in train data with 1 day lag)
- Historical weather data
  (1st 11 hours of weather stats available in 1 day lag, the rest is available in 2 days lag)

Engineered Feature
Available datetime
to
Join Tables

Data Engineering

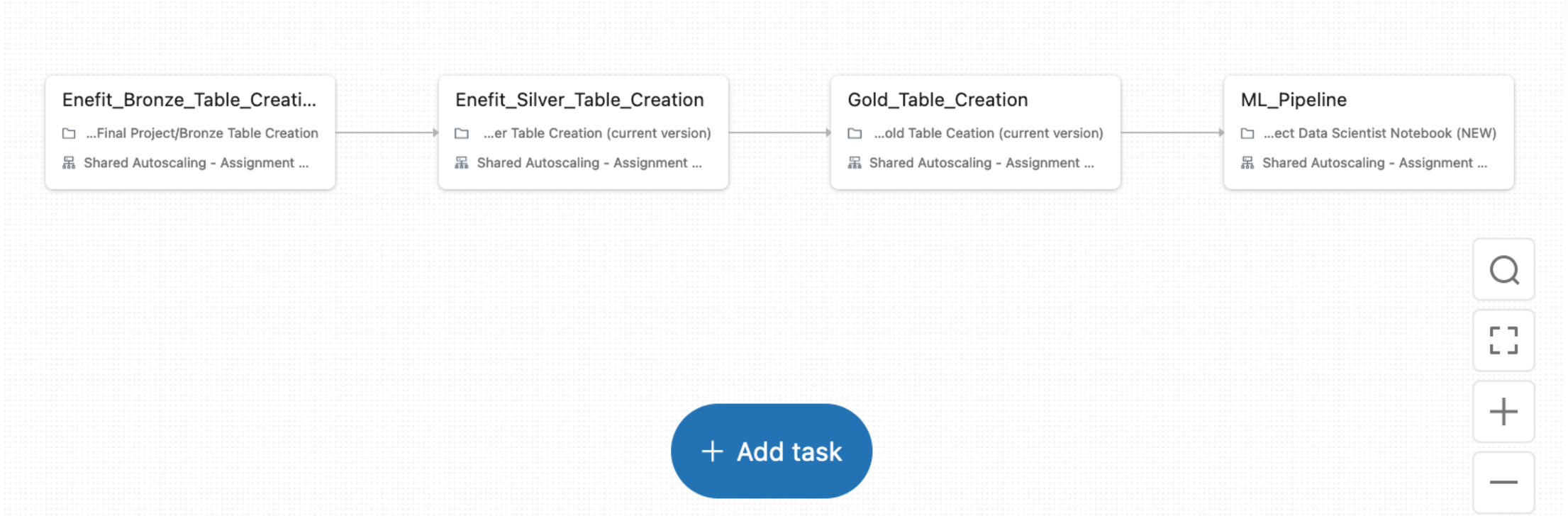
Step 2 : Building Data Pipelines With Medallion Architecture



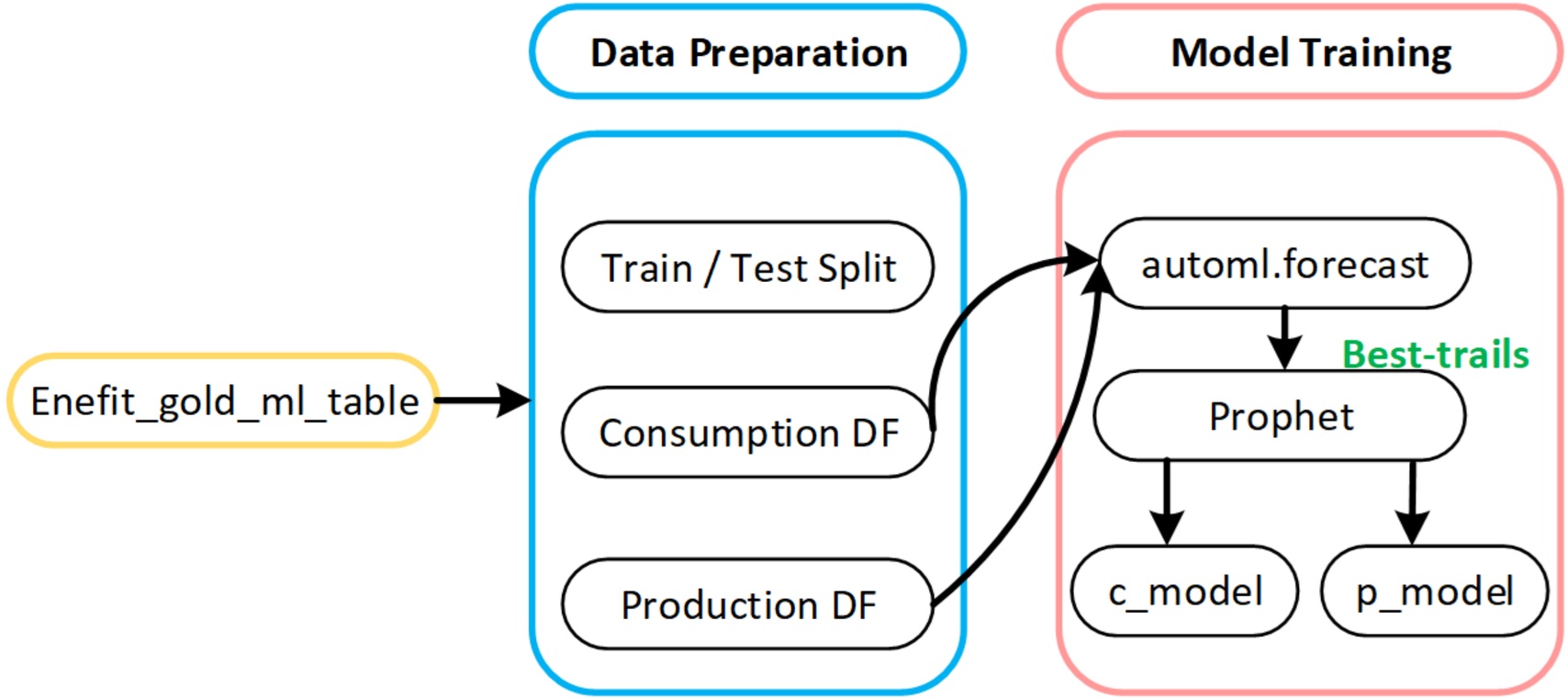
Data Engineering

Step 2 : Building Data Pipelines With Medallion Architecture

Robust Data Pipeline Using Databricks Workflows

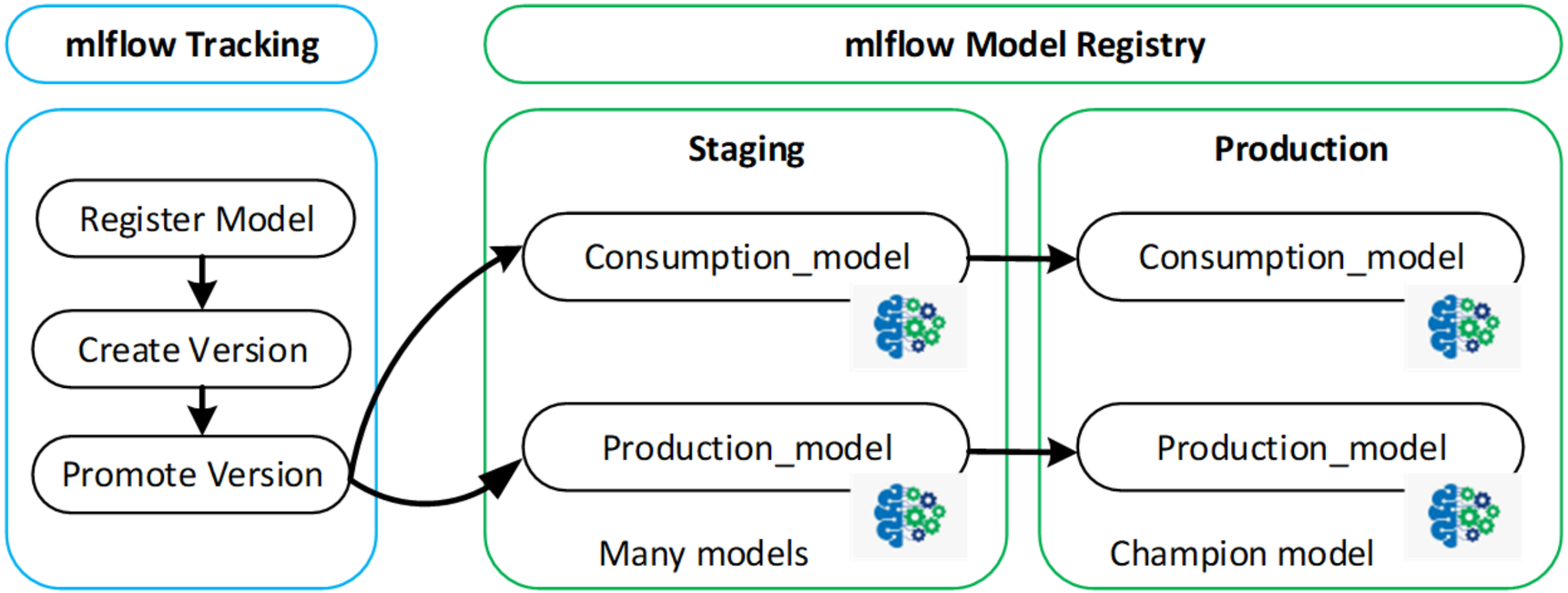


Basic Model Training



- Ingested prepared data from date pipeline
- Chronologically train/test split
- Data separation for the two use cases
- AutoML Parameters:
  - hourly frequency
  - 7–day horizon
  - mdape metric for best trials
- Two models created
  - Consumption
  - Production

Model Lifecycle Management



- Two models are registered
- New versions per model is created for new runs
- Latest version is promoted to staging
- Simulation to production
- As new data arrives in the pipeline, new models can be trained and versions can be created
- Latest version can be used to draw new inferences

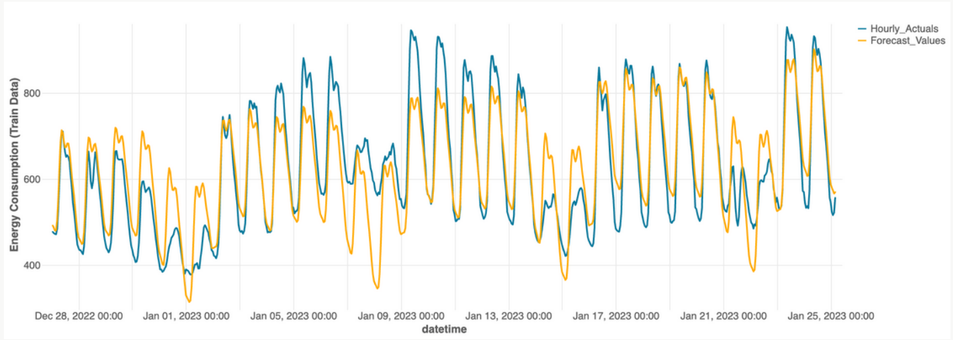
Base Model Inferencing & Performance

e.g. Consumption

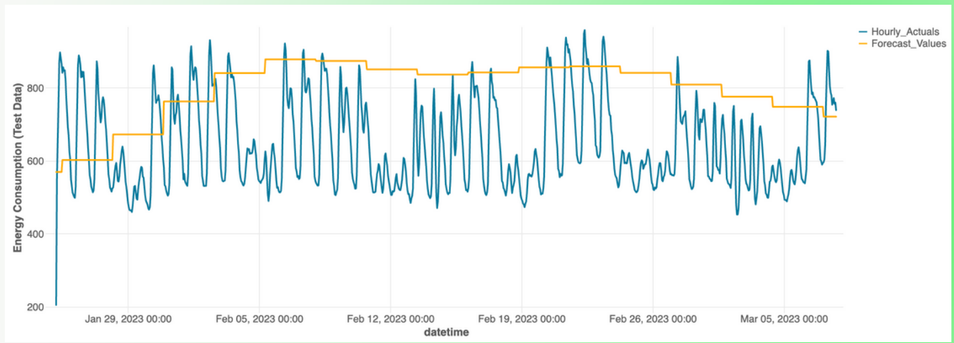
Metrics for the best trial:

	Validation
mdape	0.065
coverage	0.833
mae	50.781
smape	0.076
rmse	61.587
mape	0.079
mse	3809.261

Training Data



Testing Data

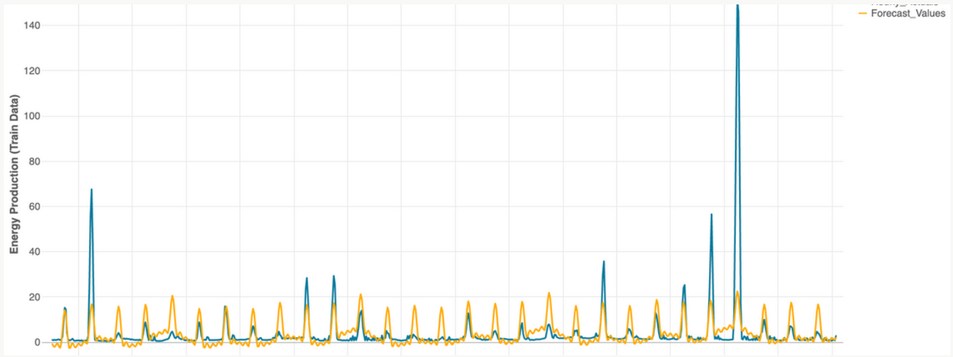


e.g. Production

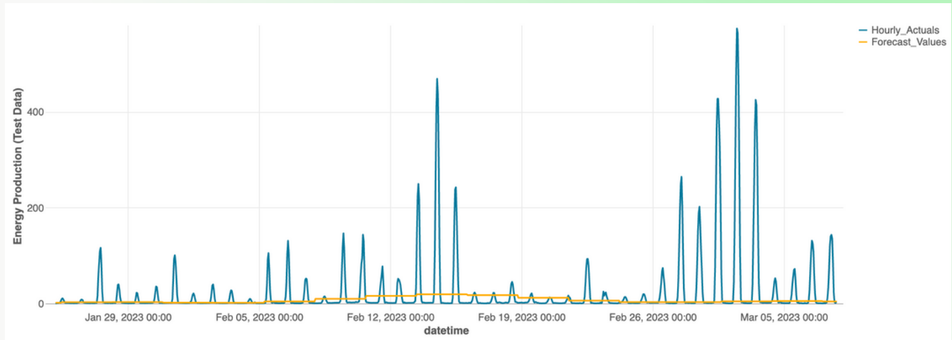
Metrics for the best trial:

	Validation
mdape	1.010
coverage	0.969
mae	2.769
smape	1.010
rmse	4.591
mape	1.261
mse	21.178

Training Data



Testing Data



Model Improvements

- Manually create base model as autoML is not highly configurable for model tuning in following ways:
  - Hyperparameter tuning via time series cross-validation
  - More robust EDA to determine whether training data is imbalanced and needs to be over/under sampled
  - Incorporation of loss function to penalize aspects of forecast according to business goals (e.g. penalize overforecasting more than underforecasting)

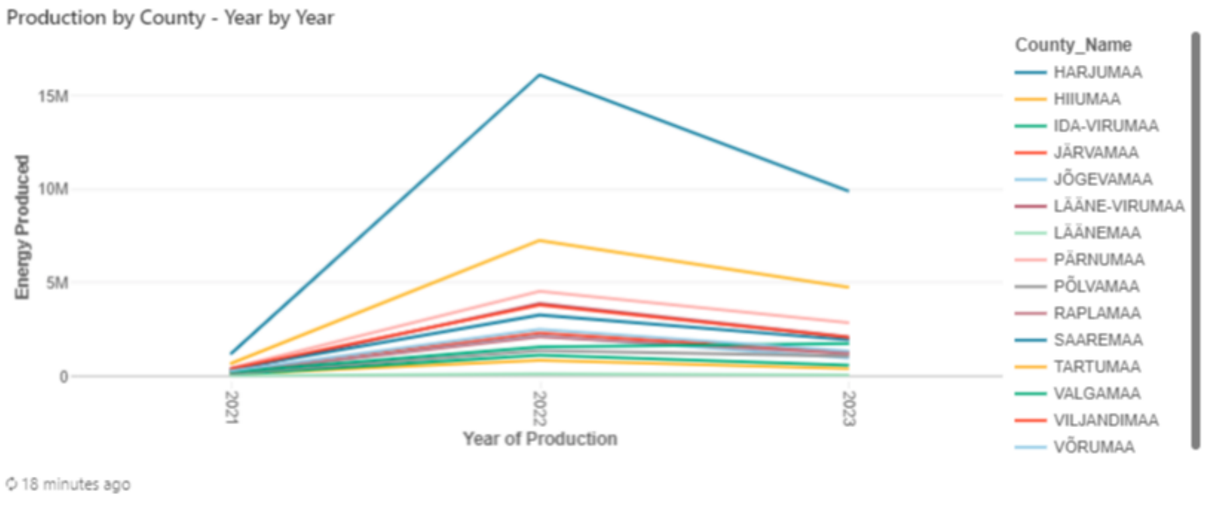
- Perform ML pipelines and feature engineering iteratively to get the best model

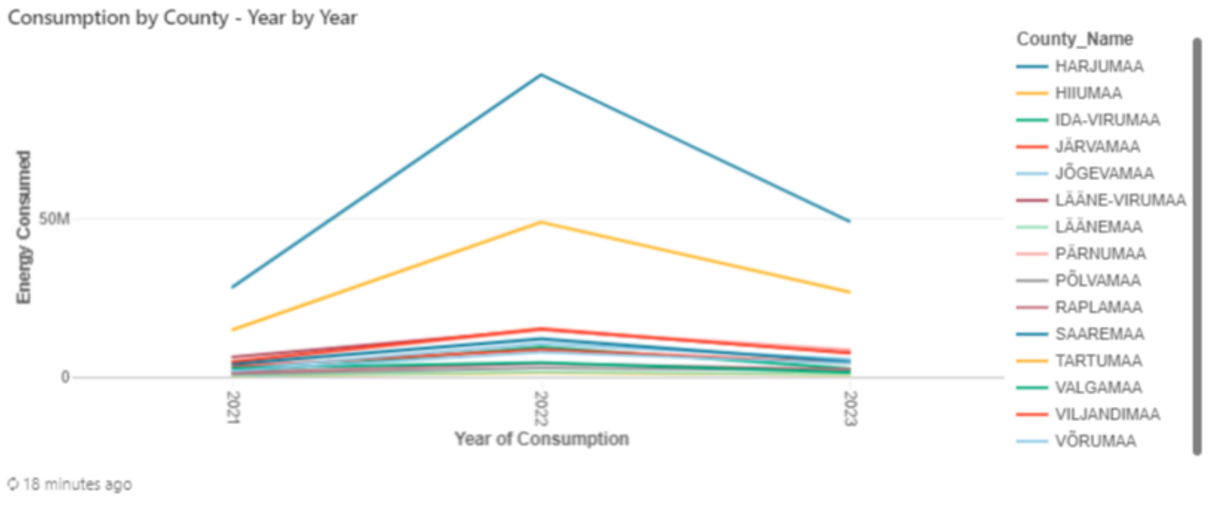
- Use ensemble methods to combine predictions from different models and produce one robust, accurate model
  - e.g. model suitable for timeseries features can be combined with model suitable for non-time series features (e.g. geographical data)

Nihal Pai’s Presentation

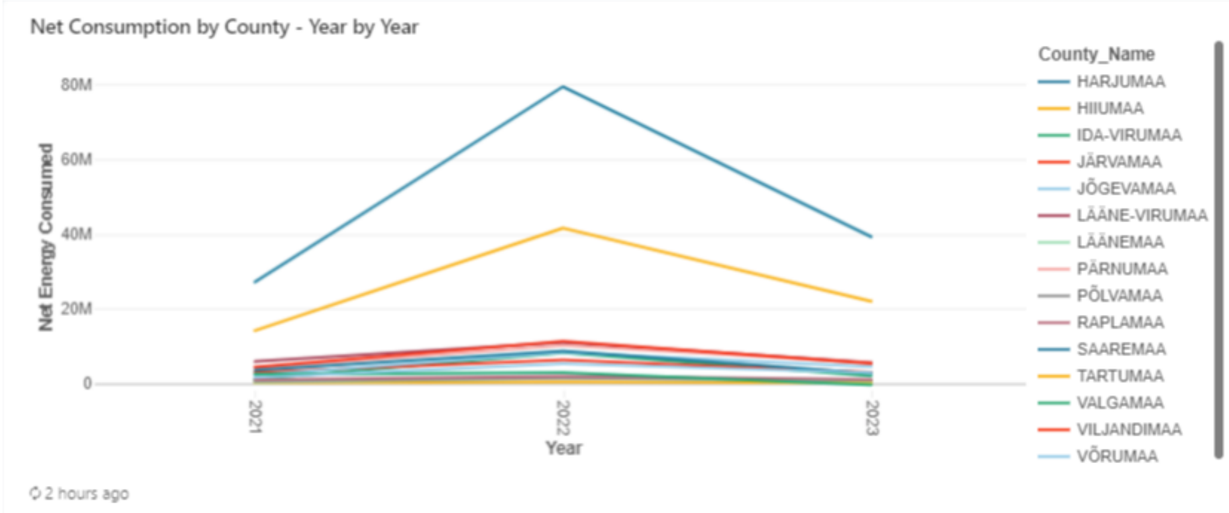
Link: https://youtu.be/tnXQlBjC-f4

BI Dashboard

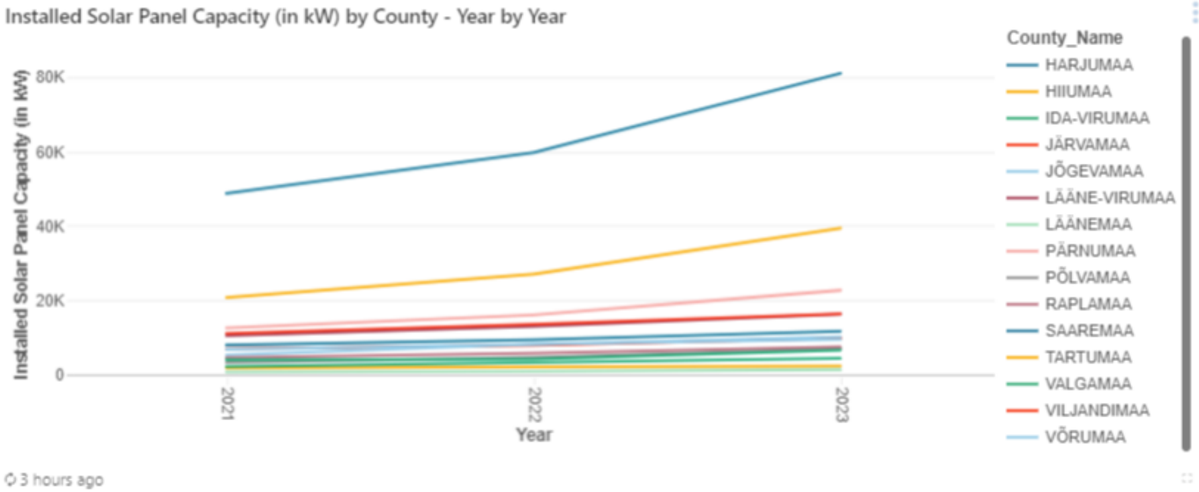


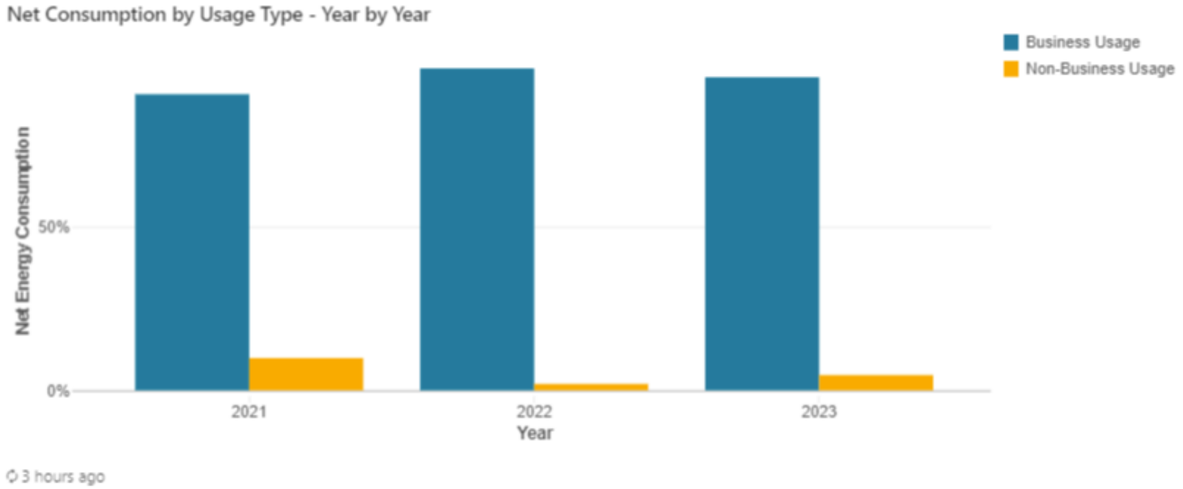


- Energy production < consumption –> Shortages of energy resulting in power cut
- Some counties are consuming/producing more than others: Harjumaa and Tartumaa
- Geopolitical terrain in Europe/Baltic states could be further aggravating the energy situation –> Decoupling from non-local power grids as early as 2025 (Reuters, 2023)



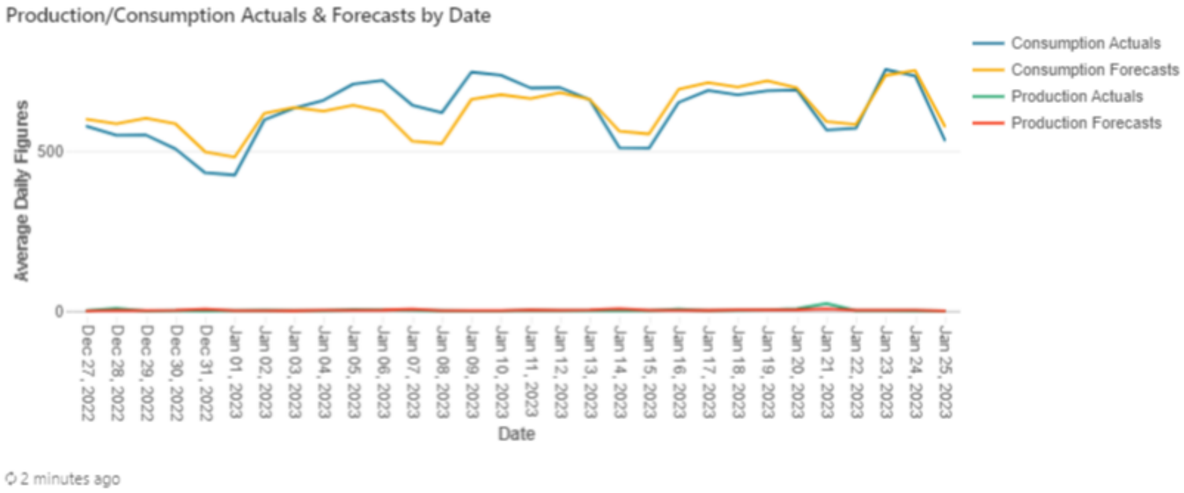
BI Dashboard





- Estonia is reliant on the burning of fossil fuels (peat) to generate power –> Moving to a renewable energy target by 2030 to eliminate shortages (ESG News, 2023)
- Although small in number, there is a spike in installed solar panel capacity as early as 2023, with Harjumaa and Tartumaa counties leading the way

- Energy consumption by business is higher than non-business –> Reform suggested by 2023 EU Country Report to Estonia's tax system (European Commission, 2023)
- The ML model reduces the energy variance over the period of a month



CI/CD & DR

CI/CD

- CI/CD is all about automation.
- Crucial for complex process, such as forecasting through an AI model
- CI/CD can be divided into 2 parts (code & model).
- **For deploying the code**, we need to: version control, pipeline to compilation & testing. With Databricks:
  - create a **workflow**
  - define the dependencies
  - link workflow to **Git** for versioning
  - use **API** or **CLI** to trigger CI/CD
  - **autolog()** to monitor & capture logs

Disaster Recovery (DR)

- General Ideas
  - duplicated cloud (on-prem & commercial)
  - periodic backup with automation
- **For data:**
  - **Deep Clone** and **Shallow Clone** features
  - **Time Travel** can roll back in time.
  - Define **retention** period, then use **Vaccum** feature to reduce stale data
- **For code:**
  - IaC tools like Terraform or Azure to recreate the ENV, lib dependencies, etc.
  - external **Git** to maintain version control
  - **workflow** to automate the back up process

What-if: Streaming Scenario

- Energy prediction is an ideal candidate for streaming.
- Continuous influx of real-time data from various sources : weather station, consumer usage changes and generator supply changes among others.
- The majority of the pipeline stages will remain the same.
- One change would be in the **ingestion** stage:
  - File-based streaming could be used.
  - Event-based streaming will be too overwhelming.
- Another change would be a **new t-series model** because:
  - Time interval to accumulate data is different (every minute or every 30 min, etc.)
  - A new triggering frequency for the pipeline to execute is changed.
  - There may be more data aggregation tasks & pre-processing in the silver table. For example: New lag analysis, new stationary analysis, new feature engineering
- BI dashboard can have a new update as frequent as the streaming requirement.

Table Partition Choices

- **data_block_id**: It' a serial no. to identify observations based on the day. Any data points collected within the 24 hour window will have the same id. This granularity meets perfectly the daily update requirement of the dashboard.

- Other potential candidates for partitioning:
  - **county**: Estonia has 15 counties. We could potentially partition the data geographically, and perform t-series analysis for each county independently.
  - **any attributes with a timestamp datatype**: If interested in balancing electricity by minutes, the current choice of data_block_id will be inappropriate. Finer granularity will be derived from the timestamp.

Questions ?

References

1. EneFIT. (2023). Predict Energy Behavior of Prosumers | Kaggle. https://www.kaggle.com/competitions/predict-energy-behavior-of-prosumers/data

2. Khare, P. (2023, October 13). Understanding FB Prophet: A Time series Forecasting Algorithm. Medium. https://medium.com/illumination/understanding-fb-prophet-a-time-series-forecasting-algorithm-c998bc52ca10#:~:text=The%20core%20idea%20behind%20FBProphet,underlying%20patterns%20in%20the%20data

3. ESG News. (2023, December 11). Estonia aims for 80% GHG reductions by 2035 and 100% renewable energy by 2030. https://esgnews.com/estonia-aims-for-80-ghg-reductions-by-2035-and-100-renewable-energy-by-2030/#:~:text=For%20example%2C%20Estonia's%20parliament%20has,2035%2C%20compared%20to%201990%20levels

4. Sytas, A. (2023, July 17). Baltic states set to decouple from Russian power grid in early 2025. Reuters. https://www.reuters.com/business/energy/estonia-lithuania-says-baltic-states-decouple-russian-power-grid-early-2025-2023-07-17/

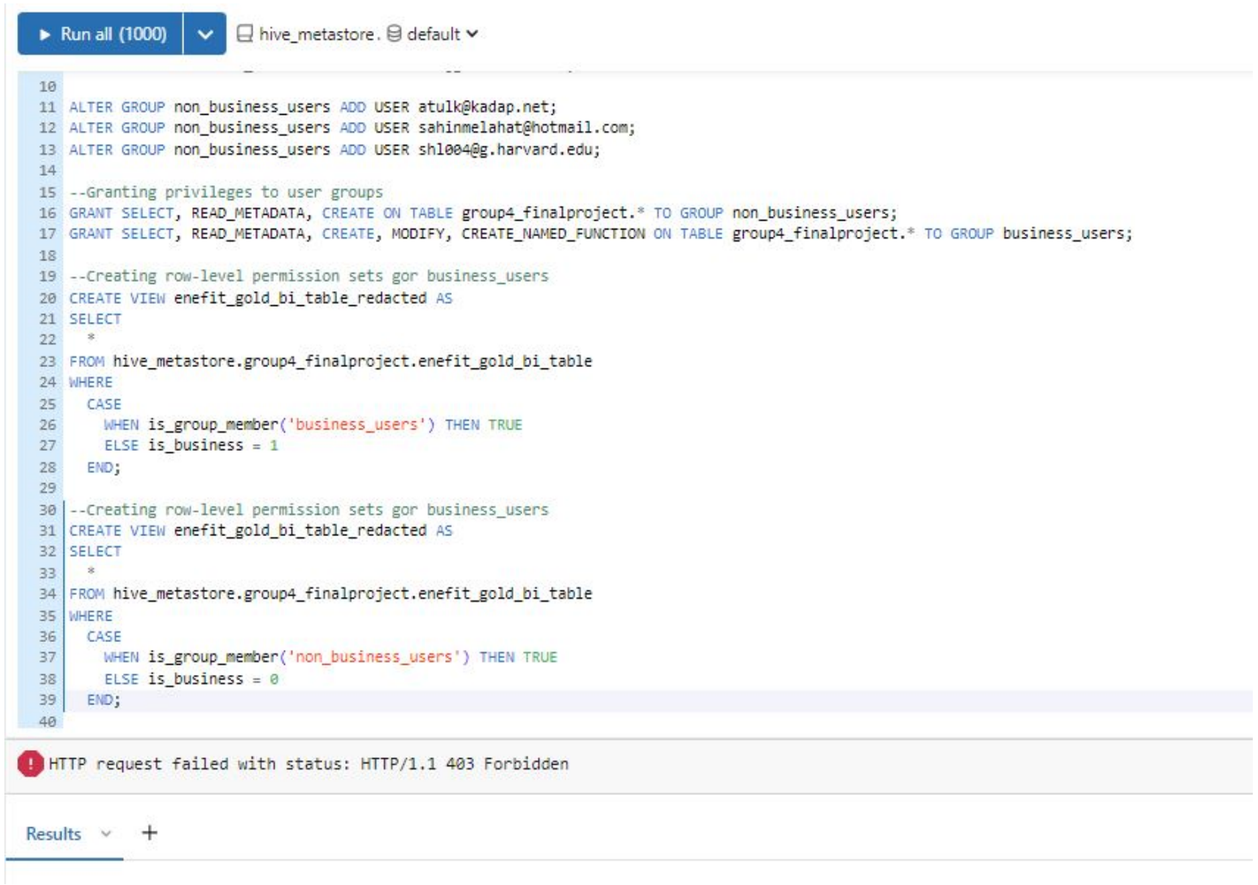
5. European Commission. (2023, June 1) 2023 Country Report Estonia. Institutional Paper 230. European Economy. https://economy-finance.ec.europa.eu/system/files/2023-06/ip230_en.pdf

Addendum

Security Model

Desired Security Model:

As we did not have admin rights to create groups, we were not able to create different permission sets and restrict users at the row-level as such:



Path to Inactive_Grants_Privileges.sql: https://harvard-e103.cloud.databricks.com/sql/editor/d2ded1ec-1901-49f7-b448-347a40f123fe?o=4720099931249 27

Security Model

In-use Security Model:

As we could not produce the desired security model, we were restricted to creating permission sets at the user level:

```
--Granting users all access to the entire schema
GRANT ALL PRIVILEGES ON SCHEMA group4_finalproject TO `mas4226@g.harvard.edu`;
GRANT ALL PRIVILEGES ON SCHEMA group4_finalproject TO `nihal.pai@gmail.com`;
GRANT ALL PRIVILEGES ON SCHEMA group4_finalproject TO `cac0888@g.harvard.edu`;
GRANT ALL PRIVILEGES ON SCHEMA group4_finalproject TO `atulk@kadap.net`;
GRANT ALL PRIVILEGES ON SCHEMA group4_finalproject TO `sahinmelahat@hotmail.com`;

--Granting user limited access to all tables in the schema
GRANT SELECT, READ_METADATA, CREATE, MODIFY ON SCHEMA group4_finalproject TO `sh1004@g.harvard.edu`;
```

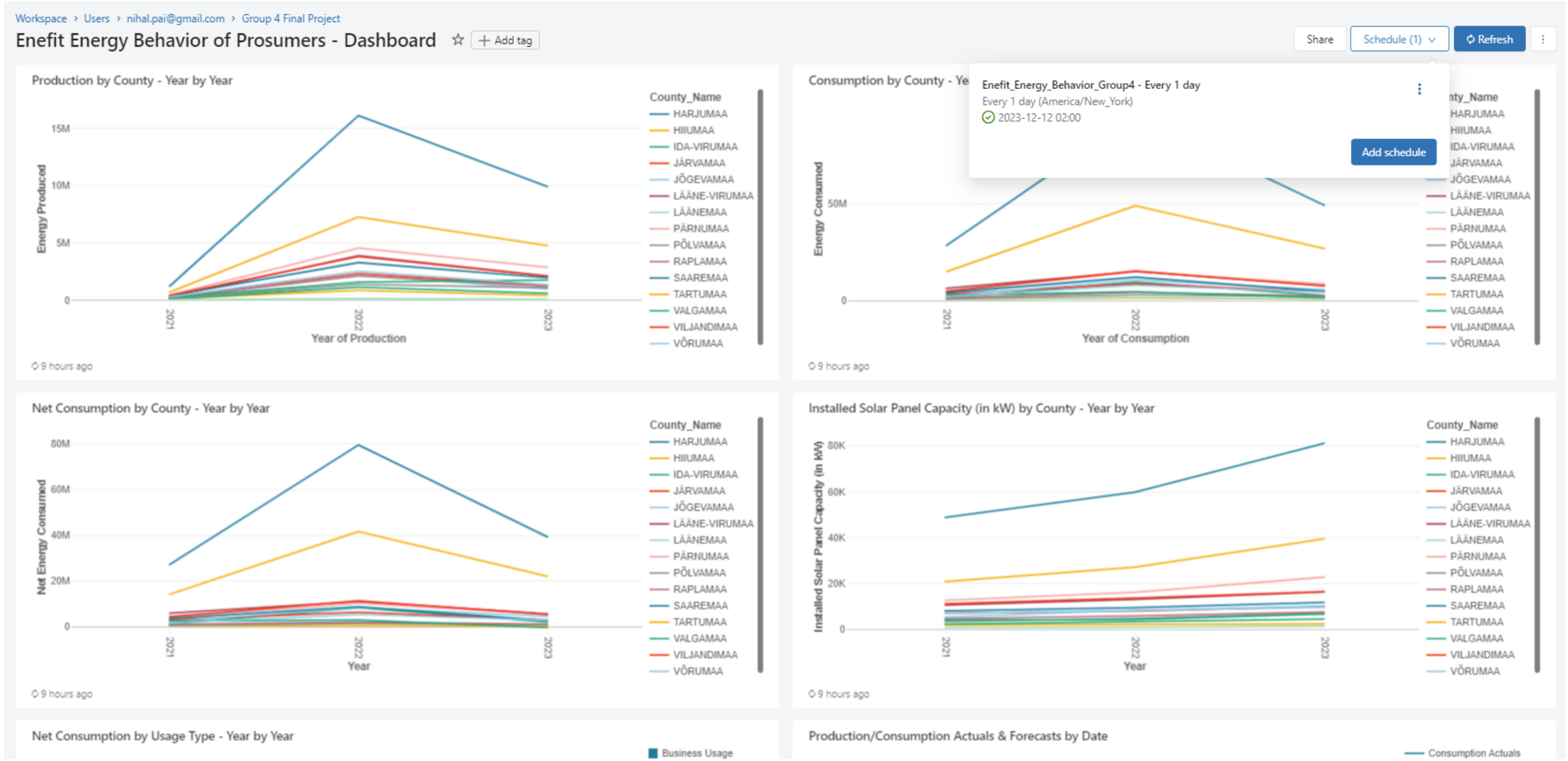
Results +

#	result
1	--Granting users all access to the entire schema GRANT ALL PRIVILEGES ON SCHEMA group4_finalproject TO `mas4226@g.harvard.ed...

Path to Active_Grants_Privileges.sql: https://harvard-e103.cloud.databricks.com/sql/editor/d2ded1ec-1901-49f7-b448-347a40f123fe?o=4720099931249 27

Dashboard

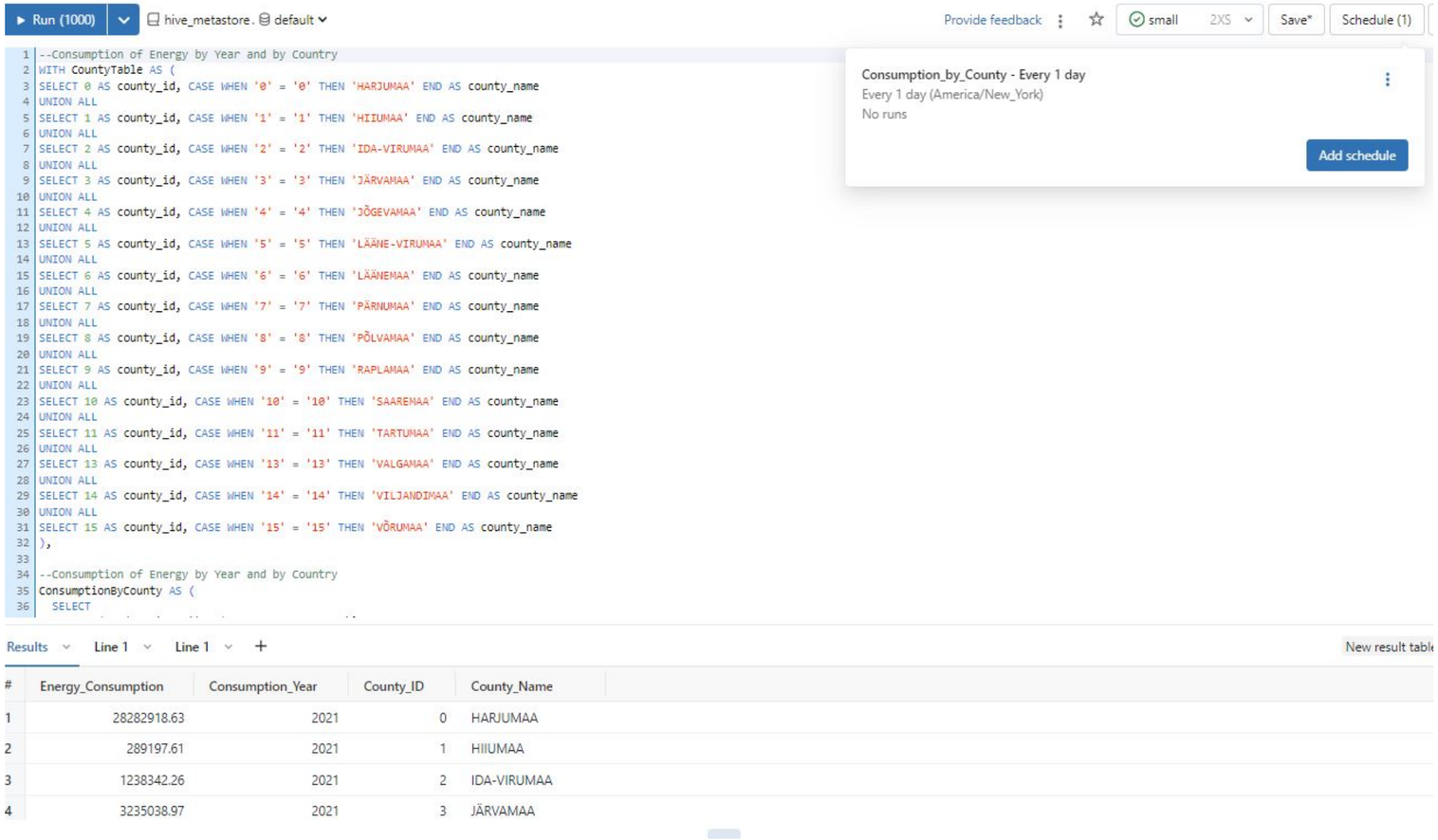
As much as we desired to include the refreshing of our dashboard with the workflow job we created, we were not able to do it as the instance we have access to did not permit us to do so. Therefore, we scheduled the dashboard to run at a certain time in the day as such:



Path to the Dashboard: https://harvard-e103.cloud.databricks.com/sql/editor/d2ded1ec-1901-49f7-b448-347a40f123fe?o=4720099931249 27

Dashboard Queries

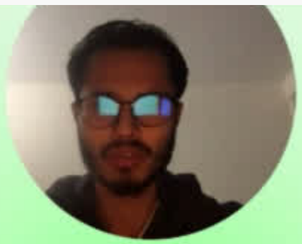
As much as we desired to include the refreshing of our dashboard queries with the workflow job we created, we were not able to do it as the instance we have access to did not permit us to do so. Therefore, we scheduled the dashboard queries to run at a certain time in the day prior to the ingestion of the data into the dashboard as such:



Path to the Dashboard Queries Folder: https://harvard-e103.cloud.databricks.com/sql/editor/d2ded1ec-1901-49f7-b448-347a40f123fe?o=4720099931249 27

Model Inference & Improvements

Base Model Inferencing & Performance

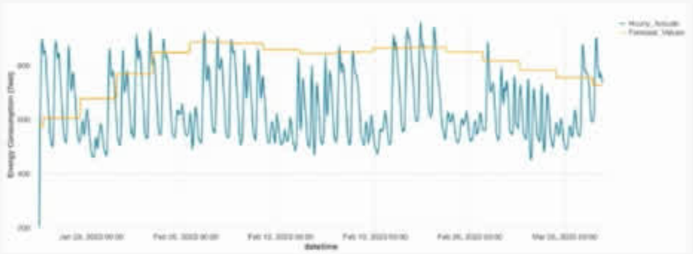


e.g. Consumption

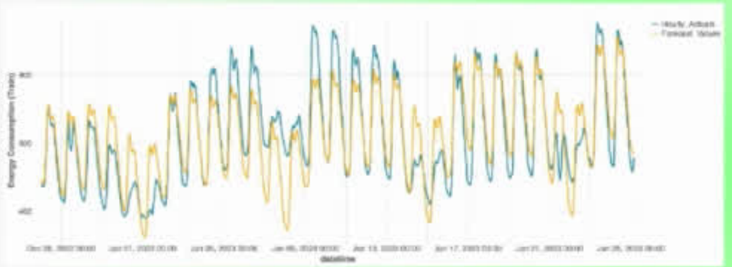
Metrics for the best trial:

	Validation
mdape	0.065
coverage	0.833
mae	50.781
smape	0.076
rmse	61.587
mape	0.079
mse	3809.261

Training Data



Testing Data

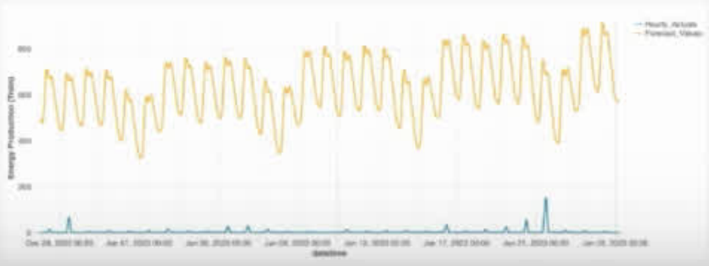


e.g. Production

Metrics for the best trial:

	Validation
mdape	1.010
coverage	0.969
mae	2.769
smape	1.010
rmse	4.591
mape	1.261
mse	21.178

Training Data



Testing Data