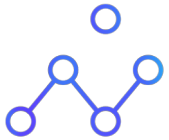
STARSCHEMA
An HCL Technologies Company

Textual summarization of time series anomalies

Balázs Zempléni and Bálint Kovács

Big Picture - Agenda

Big Picture



Anomaly detection

NLG*

*Special thanks to Szilvia Hodvonger

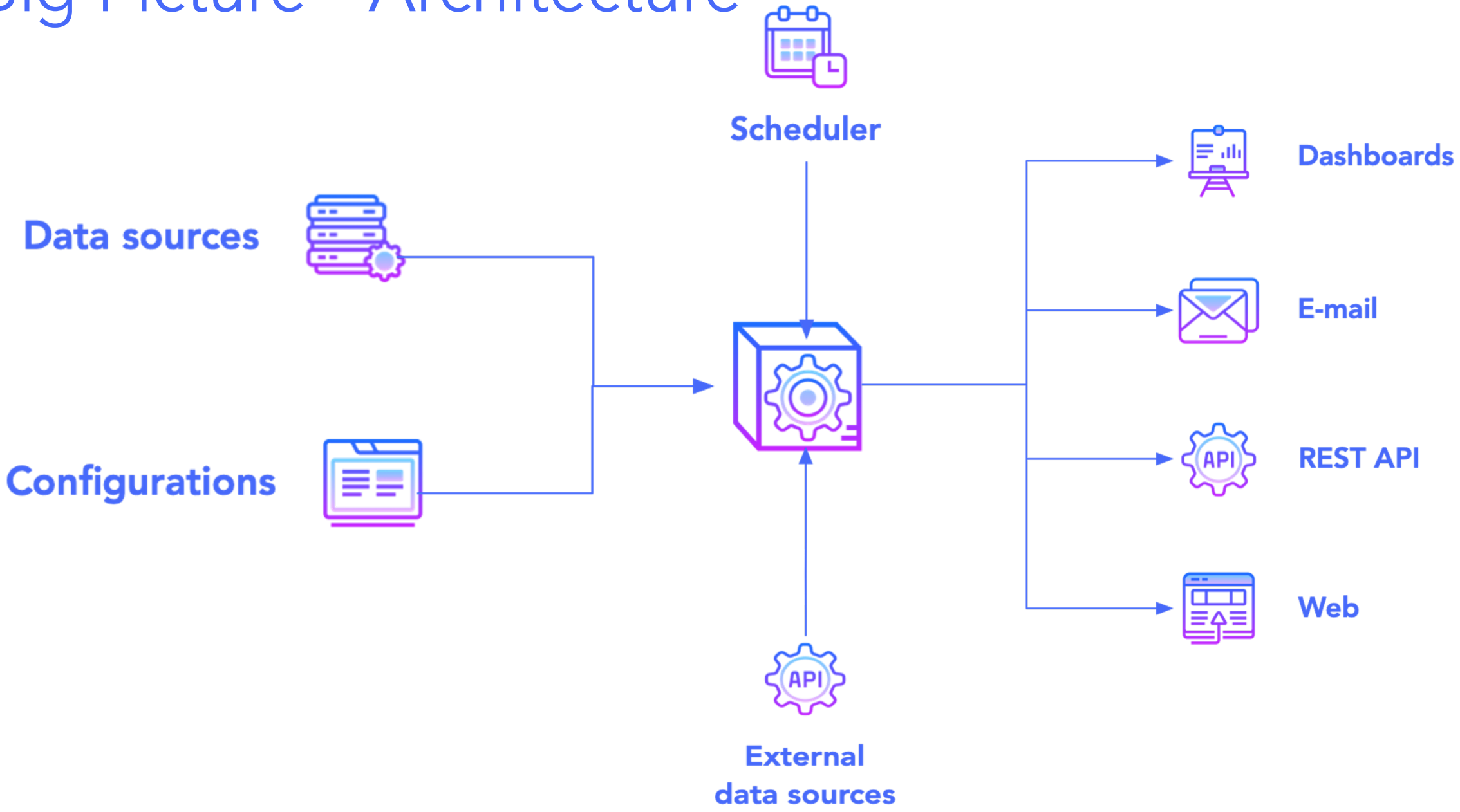
Big Picture

What is the purpose of this project?

- Multiple companies asked help to find relevant ‘news’ in their data, as the higher management wants to understand what happened in their business.

- In most cases they had to navigate in dozens of dashboard, and find the relevant KPIs and charts, to have an overall understanding of what happened on the given time.

Big Picture - Architecture



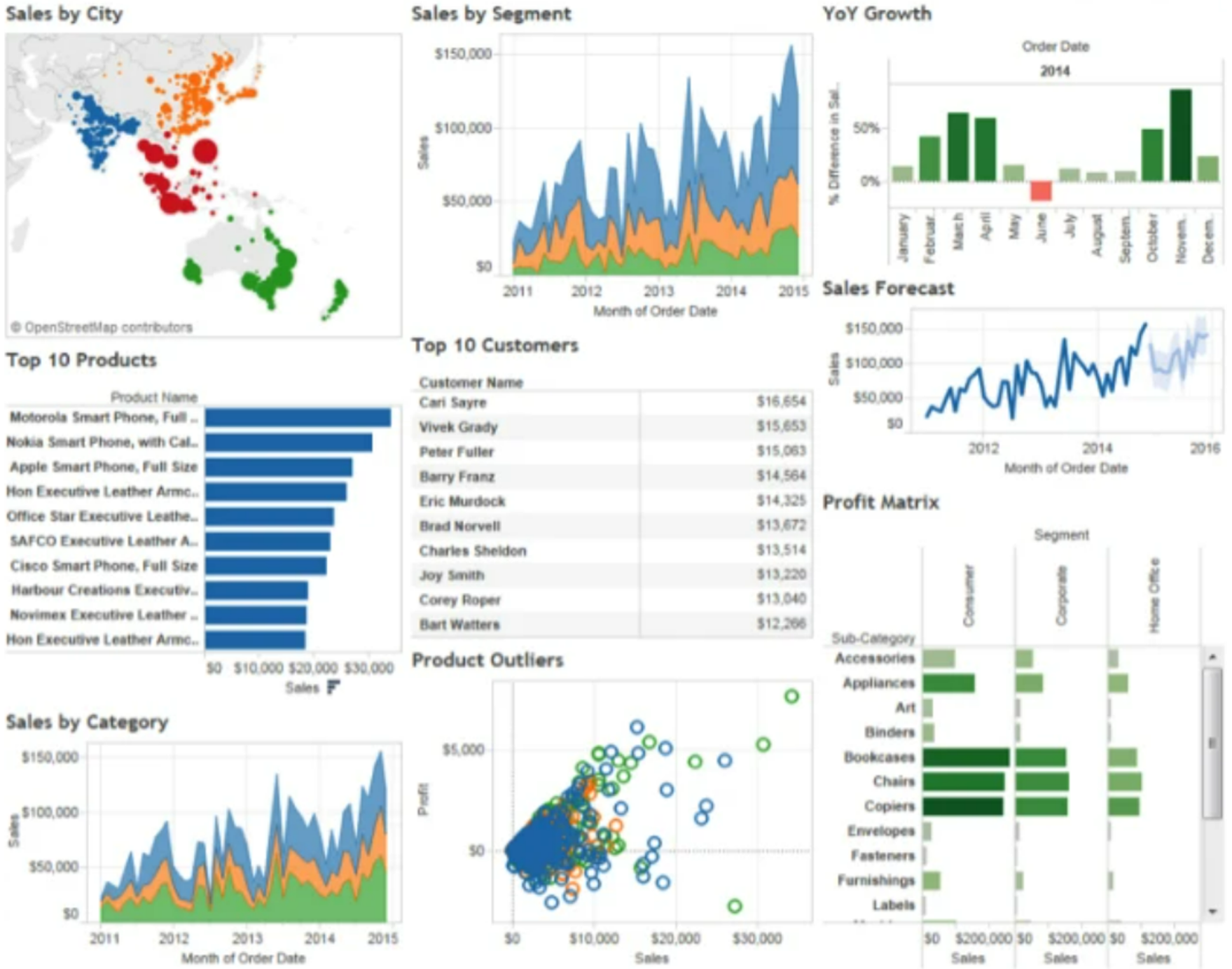
Anomaly detection

Big Picture

Anomaly detection

NLG

A complex sales dashboard

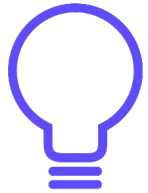


Anomaly detection in general

Problem

Dashboards are messy

- Data quality
- Manually filtering dashboards
- Many driver factors rest unrecognized



Goal

Automate the process

- Detect what is important
- Keep analyst out-of-the loop
- Summarize key events
- Reusability

Solution

ML and data engineering

- Apply time series anomaly detection
- Automatic hyperparameter configuration
- Create structured output

Challenges

Define what is interesting

- Back-and-forth talks with the customer
- Findings often had an obvious cause

Connect multiple data sources

- Many KPIs and factors
- External data sources with holidays and events

Make it reusable

- Configuration options
- Unified output structure for all type of use case
- How to generate correct textual summary

The data

Factors

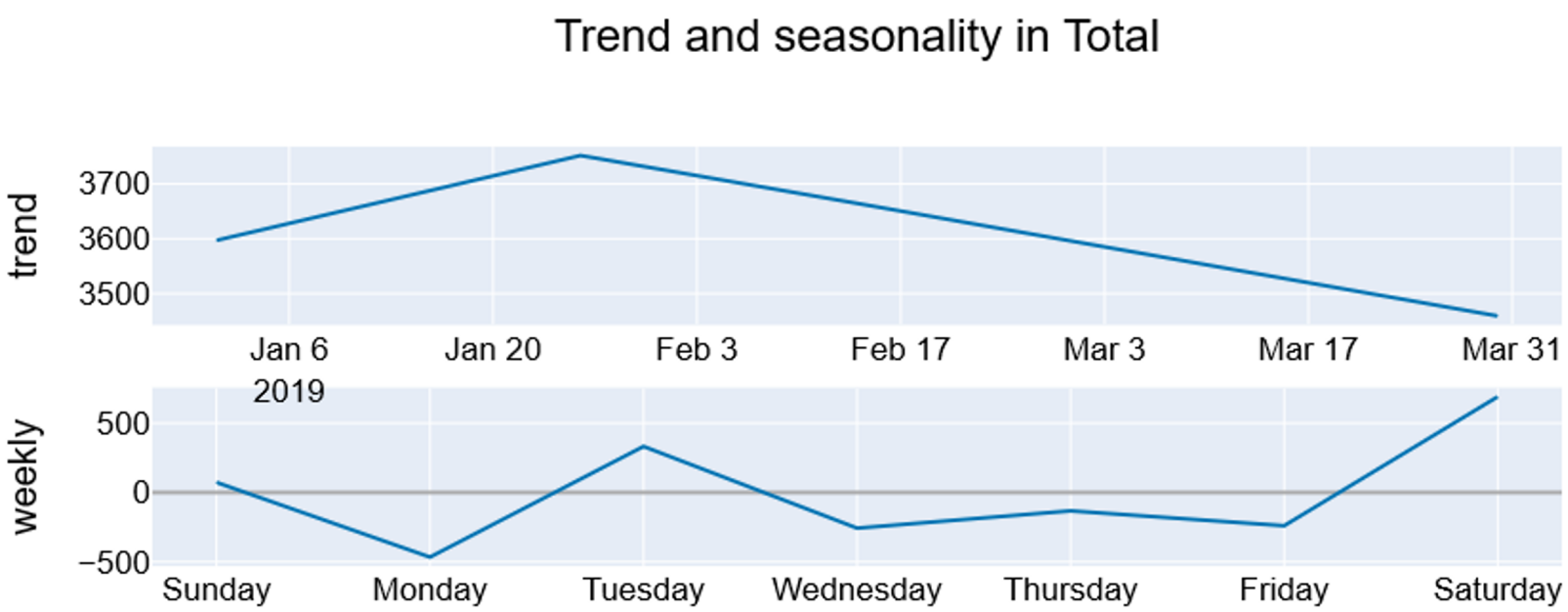
Metric

Aggregate

	City	Gender	Product line	Payment	Total	Date
0	Yangon	Female	Health and beauty	Ewallet	548.9715	2019-01-05
1	Naypyitaw	Female	Electronic accessories	Cash	80.2200	2019-03-08
2	Yangon	Male	Home and lifestyle	Credit card	340.5255	2019-03-03
3	Yangon	Male	Health and beauty	Ewallet	489.0480	2019-01-27
4	Yangon	Male	Sports and travel	Ewallet	634.3785	2019-02-08

https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales

Technical introduction



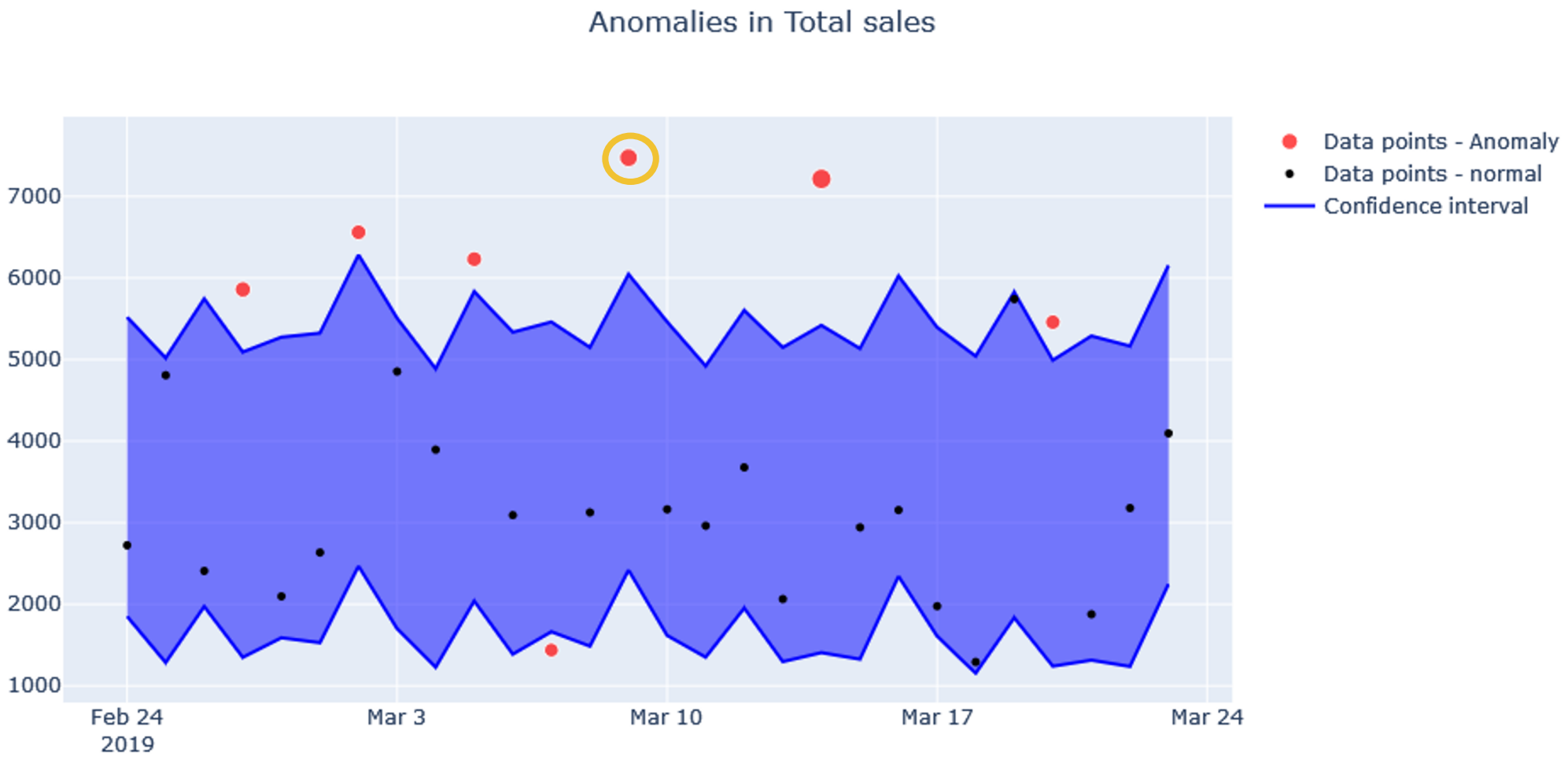
Example of time series decomposition

PROPHET

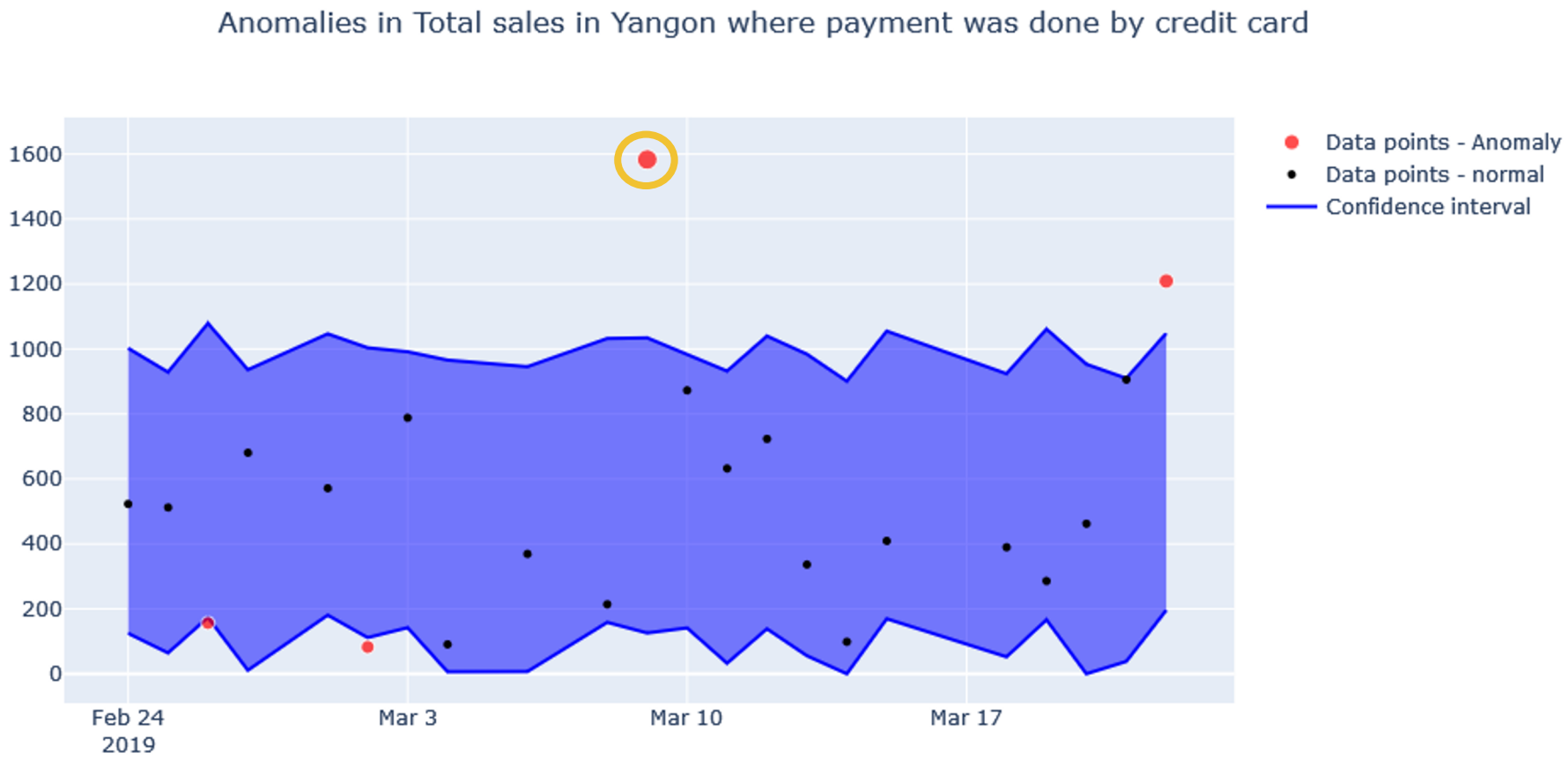
python™

SQL

Example of detected anomalies

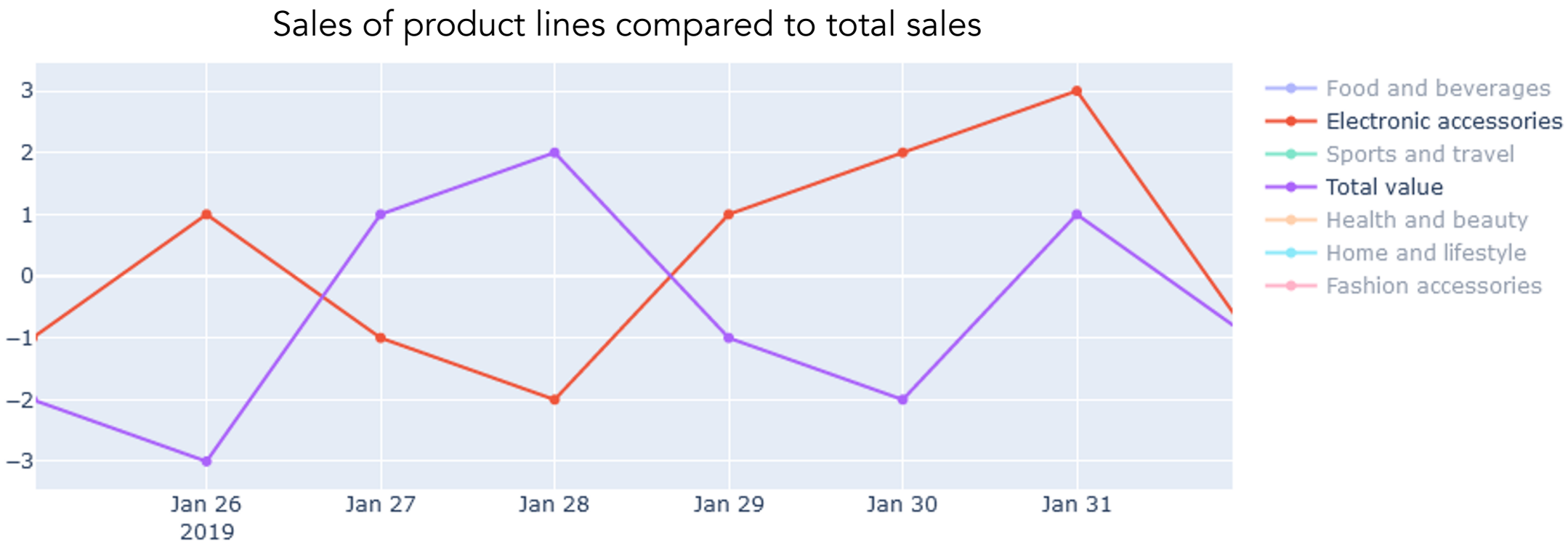


Example of factor with high importance



Date	2019-03-09
Product line	-
City	Yangon
Gender	-
Payment	Credit card
importance	0.3468
actual_prediction_difference	1031.44
actual_prediction_ratio	2.8695

Trend change



streak_end	streak_length	streak_value
2019-01-30	5	moves_different_than_total

Natural Language Generation - NLG

Big Picture

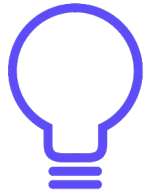
Anomaly detection

NLG

Natural Language Generation in general

Challenge

- Generate text with template become complex very fast.
- The state-of-the-art Natural Language Generation models are not open source, this would make us **transfer the sensitive financial data** over a not secure connection, which could be logged as well.
- Currently text generation models, does not take **factual data as input.**



Goal

- Generate sentences from input data, with high semantic **fluency** and high **fidelity.**

Solution

- Modify and retrain a model that can generate text.
- Transform anomaly detection data for the required form.

Natural Language Generation - NLG

Input data

date	2019-01-30
product_line	Health and Beauty
...	...

On 2019-01-03, for 2 days in a row, Health and beauty had an increasing trend, while the rest of the categories among Product lines were decreasing.

Transfer learning

We are using a transformer model + transfer learning

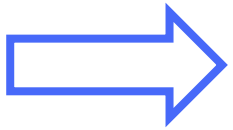
Sentence generation

Syntactically and semantically correct sentences, that contains all the information.

NLG - Input data

Challenge

- Limited data available with structural and textual input.
- Most available open-source data is mostly about restaurants, sport or airports.
- We need to generate input data for our use case, which is expensive.



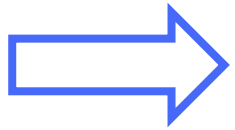
Solution

- Gather all available open-source data set, modify it to have the same format

NLG - Transfer learning

Challenge

- Out of the box none of the Transformer variants work properly for this kind of text generation



Solution

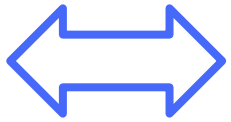
- Tested multiple transformer variants and chose the most suitable one for text generation.
- Started with a pretrained model, which was trained on multiple different open-source dataset.
- Used transfer learning and trained for our specific smaller data sets.

NLG - Sentence generation

Semantic Fidelity

Fluency

- Semantic fidelity: text that conveys the meaning accurately



- Fluency: text that sounds very fluent

Low semantic fidelity and high fluency:

In March the sales was over performing in Asia.

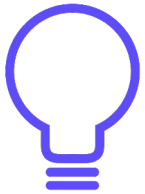
Low fluency high semantic fidelity:

Total sales were $1583.15. Customers who paid with credit card. In Yangon the Total sales were overperforming.

High semantic fidelity and fluency:

On 2019-03-09, among customers paid with credit card in Yangon, the Total sales were $1583.15, which overperformed predictions by $1031.44.

NLG - External data source



Idea

- Pull data from external data sources, and filter for the parameters of the anomaly

Solution

- This will give additional information i.e.
  - was there any stock
  - how was the weather on that day
  - was there any promotion

Anomaly description:

On 2019-03-09, among customers paid with credit card in Yangon, the Total sales were $1583.15, which overperformed predictions by $1031.44.

Additional information:

On 2019-03-09 there was a 10% discount in Yangon for those who paid with credit card.

Summary

- We developed multiple algorithms to be able to find interesting ‘news’ in data.
- Trained a transformer model which generates text from input data.
- In our use case it overperforms the current state of the art solution.

Q & A