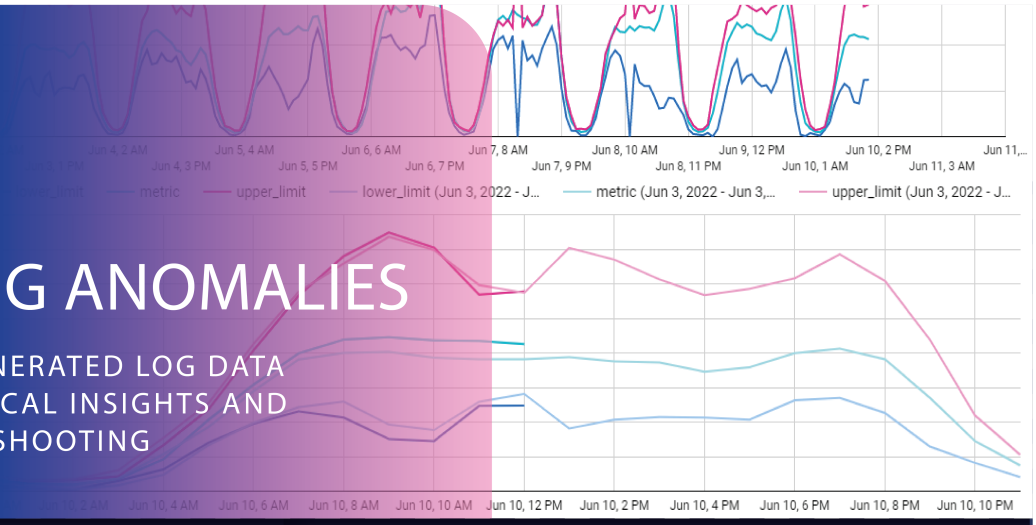

PREDICTING ANOMALIES

UTILISING USER GENERATED LOG DATA TO CREATE ANALYTICAL INSIGHTS AND ASSIST IN TROUBLESHOOTING

ABSTRACT

A lot of log data is generated from products these days but it is not always used for analytical purposes. If that log data could be analysed and used for detecting when abnormal behaviour is occurring it could provide an efficient way to notice problems in the product and also assist in troubleshooting said problems. Log data generated by users of the online shopping cart at IKEA IT were analysed to create an anomaly detection prototype on the Google Cloud Platform. This prototype alerted when user activity was noted to be outside thresholds that were calculated with standard deviation on the hourly web user's cart usage in Germany. It also present relevant data for troubleshooting in a dashboard created in Google Data Studio if abnormal behaviour was found. The resulting prototype is a proof of concept that looked at a subset of log data but still proves the possibility of determining abnormal system behaviour from user generated data. The usefulness of such monitoring for companies in many fields can be large but future development of faster and more reliable models and alerting are needed.

PROBLEMS

- What data can be extracted from the logs?
- What user generated log data are of interest for finding potential anomalies?
- How can the project be run cost-efficiently?
- How to implement the thresholds to reduce, or increase the amount of alerts?
- What data to visualise for assisting in troubleshooting?
- Which mathematical models can be used to perform relevant data analysis?

METHOD

1. Gathering information
2. Collecting data
3. Analyzing data
4. Creating a threshold model
5. Setting up alerting through GCP
6. Creating a dashboard
7. Testing the prototype

Oskar Karlsson & Michael Lord

SOLUTION

By analysing the user-generated log data collected by the systems, a general sense of what data is optimal to look at to determine abnormal behaviour is gained. Following this, a model for determining the upper and lower thresholds of what defines normal behaviour has to be researched and created. When this has been done creating an alerting system that notices when the thresholds have been crossed is to be implemented. Finally, choosing what data is relevant for troubleshooting and creating a place where this data can be shown is done.

RESULTS

- A prototype with four parts
  - •Bigquery Tables - where all the data for the prototype is stored, including threshold values and averages.
  - •Scheduled Queries - handles the summarization of new data per time-frame to the tables and calculations of thresholds.
  - •Cloud Functions - the functions that compare the usage against the boundaries set to find abnormal behaviour.
  - • Dashboard - presents relevant data for troubleshooting.
- Testing of the system
  - • By testing a downtime of 38 minutes the prototype was determined to work as intended.

DISCUSSION

The prototype created is a proof of concept that does prove that this type of detection is possible with this type of data. Furthermore, the cost of it was concluded to be very low compared to the amount of money it would save.

However, the potential for future development this prototype has is big. The current time-frame for detection is an hour, meaning there is a lot of improvement to be done. Increasing the frequency is one of the things to look into.

The model for determining thresholds are also a point of possible improvement. The current model is a standard deviation for making an upper and lower limit. More complex models could most likely be used to determine these limitations with better accuracy. The best model would most likely be a machine learning one.

Lastly the dashboard and usage areas of this detection could be expanded to basically all fields that have a product which gets some type of data that can represent user activity.