Erratic server behavior detection using machine learning on streams of monitoring data

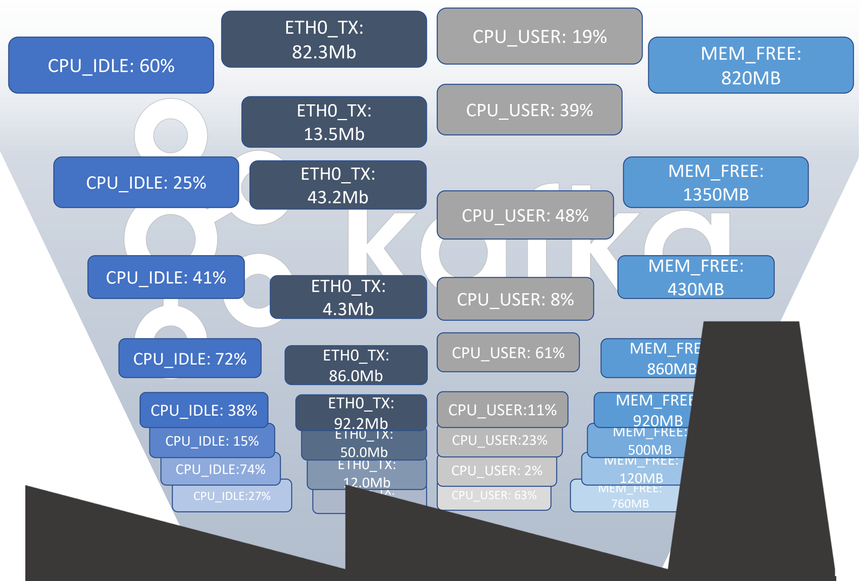
M. Adam
Academy of Sciences of the Czech Republic
L. Magnoni
CERN
D. Adamová
Academy of Sciences of the Czech Republic
M. Pilát
Charles Uni., Faculty of Mathematics and Physics

Introduction

In modern data centers the variety and complexity of services creates an overflow in monitoring data, both in terms of volume and diversity. Standard threshold-based alarms are by nature limited in coverage and can produce alert-fatigue and problem-miss.

Goals

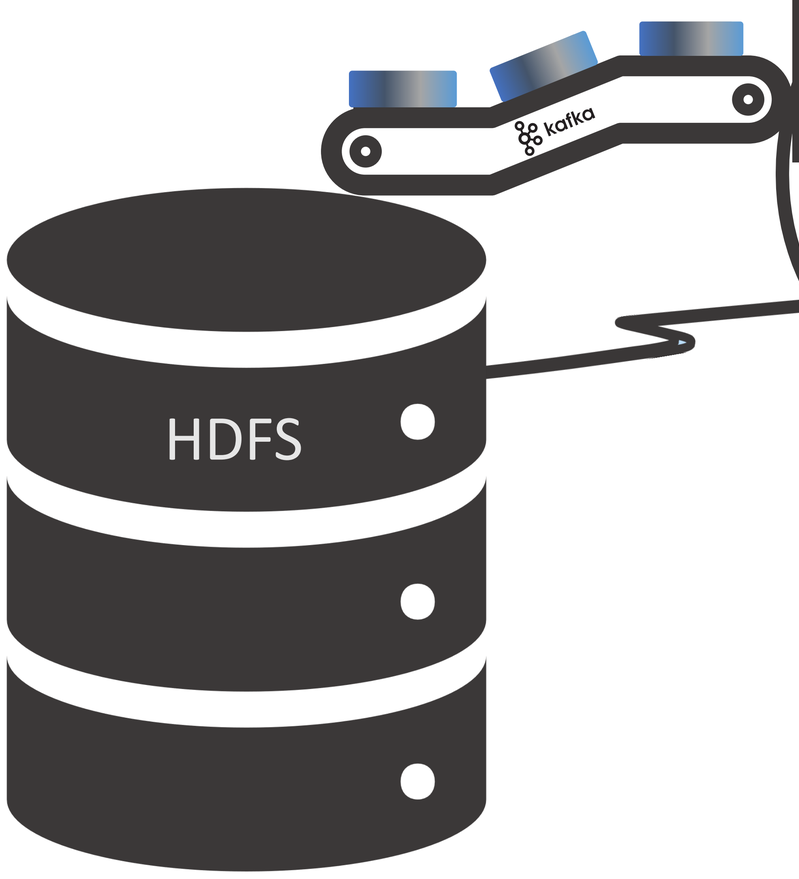
1) Gather a ML ready dataset
2) Confirm, that given a history of a **distributed application** workload and a **cluster** of servers running said application, it is possible to **detect an erratic server** by analyzing **only the os-level metric** data.

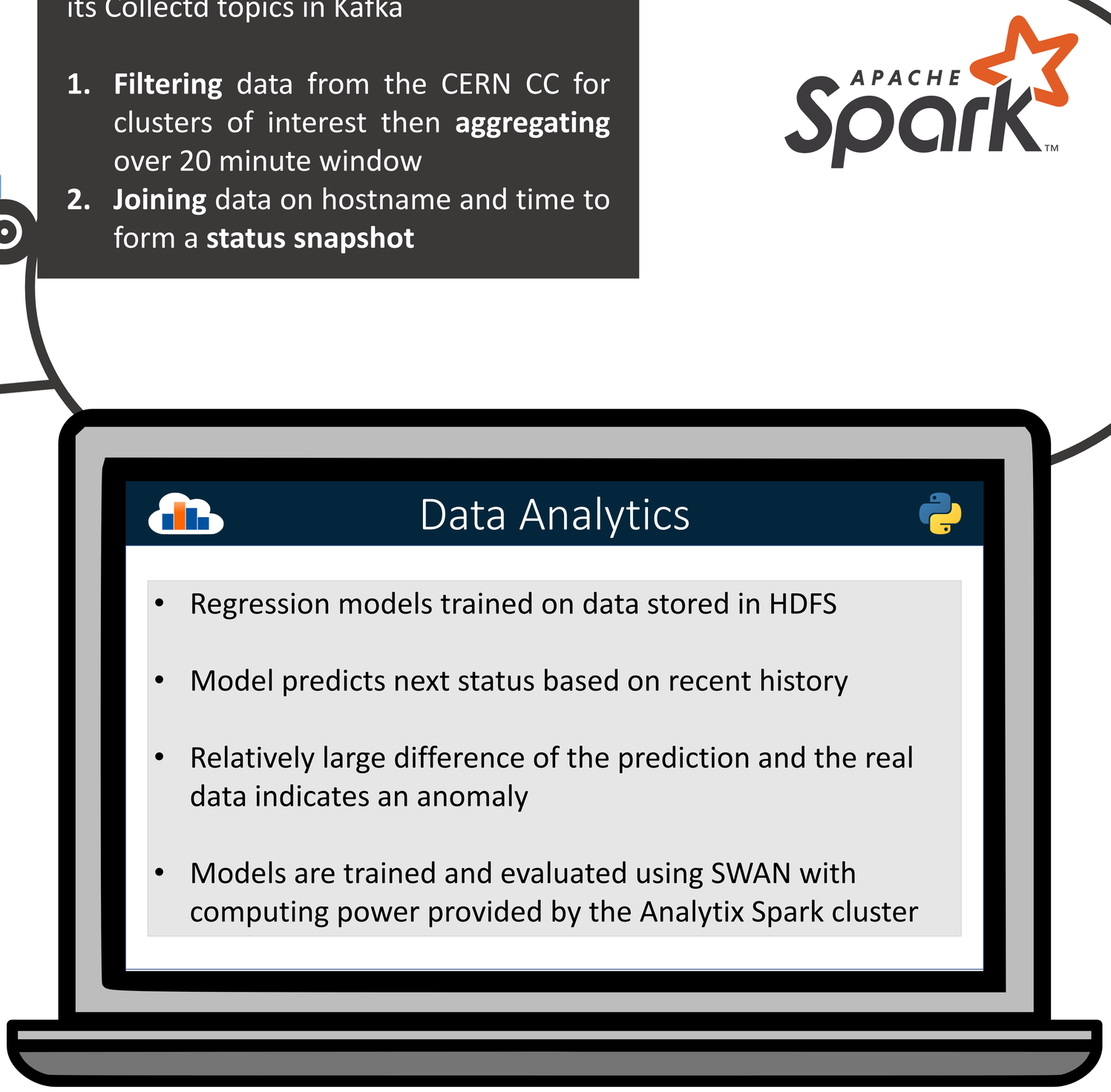


Data Gathering & Preprocessing

2 Spark streaming jobs running in the MONIT infrastructure reading data from its Collectd topics in Kafka

1. **Filtering** data from the CERN CC for clusters of interest then **aggregating** over 20 minute window
2. **Joining** data on hostname and time to form a **status snapshot**





Results & Acknowledgements

Both tested types of models (Linear regression and Random forest regressor) give satisfactory results on simple anomalies with basic preprocessing. Attempts to further pre-process the data using PCA yield mixed results.

Special thanks to the MONIT team and to the Hadoop and SWAN team from CERN.

Conclusion & Future Work

The **concept of a ML model detecting** a simple artificial **anomaly only using the os-level monitoring metrics has been confirmed**. The next step is studying a wider array of anomalies as well as model universalness. Final challenge would be converting this experimental setup to a production ready application running without user interference.