Masterthesis

Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Masterthesis

Outlier Analysis in Live Systems from Application Logs

Detecting outliers, i.e., errors or anomalies in today's applications is a difficult task because (1) such applications generate massive amounts of data, usually collected as logs, which are heterogeneous, i.e., they are composed of information of various types (numerical, categorical, images, …), and the format of such logs differs from one application to another. Also, (2) one must analyze such errors in real-time, because application downtime usually translate into high cost for the operator – One must not only detect outliers/errors, but also understand what happened.

In the automotive sector, take for example the task of managing a fleet of vehicles at Porsche: Between front-end and back-end systems, numerous "data collectors" generate logs which are subsequently stored into a data lake. In such setting, it is important to detect and classify errors (CPU issues, platform issues, etc…) whenever they occur to avoid costly downtimes of the live application. To design reliable and adaptive systems, it is not only necessary to detect errors, but also to know which kinds of error occur, to ensure a quick recovery.

Thus, the focus of this thesis is the development of a generic framework to detect and analyze outliers from streams of application logs. This results in the following tasks:

- In-depth literature review of outlier detection and analysis in streams.
- Implementation of an approach for the detection of outlier, paired with an automated clustering of the outliers, to ensure better explanation of the error occurring. The approach should be generic, i.e., it should deal with data from various projects with different log structures.
- Evaluation of the developed approach against baselines and state-of-the-art algorithms. Hereby, the student will test against (1) real-world Porsche data and (2) public data (e.g., from Kaggle, UCI ML).

We offer:

- Thorough mentoring, highly motivated and fun team, teamwork with Porsche Ludwigsburg.
- Access to state of the art big data and automotive infrastructure.
- Help to "go for the extra mile" and publish parts of your results in scientific conferences.

We expect:

- Good knowledge in Python/Scala programming and Data Mining, knowledge about Big Data technologies (Spark, Hadoop, AWS or Cloudera Data Platform) is a plus.
- Ability to plan and work independently.
- Very good knowledge of English (German is a plus).
- High level of motivation, enthusiasm and curiosity.
- Interest in bringing the work up to the standard of a scientific publication.

Throughout this work, the student will acquire knowledge and practical experience in the domain of data stream and related fields. The thesis takes place as a joint project between KIT and Porsche; The student will benefit from supervision by both institutions. Optionally, the student can start with a 2-3 months internship before the thesis to get familiarity with Porsche big data infrastructure. Please send your application to Rosina Kazakova (rosina-teodora.kazakova@porsche.de) and Edouard Fouché (edouard.fouche@kit.edu), containing at least a transcript of records and a curriculum vitae.

Ansprechpartner

Edouard Fouché, M. Sc. edouard.fouche@kit.edu +49 721 608-47337 Raum: 342

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34