Technische Universität München

Fakültät für Informatik
Lehrstuhl 22
Software Engineering
Prof. Dr. Alexander Pretschner

Boltzmannstraße 3 85748
Garching bei München

Tel: +49 89 28917885

Web: http://www22.in.tum.de

Anomaly detection and prediction in distributed software systems using machine learning

Bachelor thesis

Fabian Huch
Supervisors: Prof. Dr. Alexander Pretschner, Ana Petrovska, Mojdeh Golagha

Context

Anomalies are an inevitable part of operating a software system. Thus, it is desirable to design systems resilient - meaning that they can overcome anomalies by using alternative approaches or by trying to counteract malfunctions. The antifragile approach on the other hand tries to improve the system on occurring anomalies, just like the human muscle does when encountering extensive stress. [1]

But reliable anomaly detection and localization in big, highly distributed software systems is hard to accomplish by simple analysis of system properties like response times, memory usage or CPU load. This is because firstly, those properties on their own don't necessarily yield accurate information about the state of the system. Secondly, due to the distributed nature of those systems, the properties of a single micro-service do not reflect on the whole system. Traditional operations frameworks, like RedHat's OpenShift, usually use liveness requests to the different services in order to determine their health status [2]. This is an easy way to achieve resilience, however it does not provide for fine-grain, intelligent decision making that would be suitable for completely automated DevOps.

Goal

Analyzing the time series of those properties across multiple subsystems, one can recognize patterns which correspond to system health. It is hence the goal of this thesis to reliably identify system anomalies by those patterns using machine learning algorithms. While known anomalies will form the basis for the training data, the trained system should also identify yet unknown anomalies, possibly triggered by code changes.

An explicit example of such a pattern is when the relative average memory usage of a node increases to a critical point, while the request load doesn't differ from normal operations and CPU usage indicates that no process like a loader is running.

As the use-case of anomaly detection is to make automated operation decisions, real-time processing of the large amount of measurement data (potentially ten thousands of measurements on dozens of machines every few seconds) is aimed to be achieved with the help of Big Data technologies.

Workplan

1. Assessment of the available data and manual classification of some anomalies
2. Setup of the required cloud infrastructure
3. Data Pre-Processing
    I. Development of a feature abstraction
    II. Cleansing of available feature set by filtering out constant or irrelevant features
    III. Selection of features/feature combinations associated to feature abstractions
4. Implementation of the basic structure to do Machine Learning using Apache Spark MLlib for distributed ML computing, connecting to a SolR cloud as distributed data store
5. Execution of different ML approaches, using Classification, Clustering and Association Rule Mining techniques
    I. Modeling, parameterization and training

Technische Universität München

Fakültät für Informatik
Lehrstuhl 22
Software Engineering
Prof. Dr. Alexander Pretschner

Boltzmannstraße 3 85748
Garching bei München

Tel: +49 89 28917885

Web: http://www22.in.tum.de

II. Feature selection
6. Evaluation of the different approaches with the support of experts
7. Writing the thesis document

Time Table

Task Nr.	Start Date	End Date	Work Days
1	Jun 6	Jun 15	done
2	Jun 19	Jun 22	2
3.1	Jun 27	Jun 30	2
3.2	Jun 30	Jul 4	3
3.3	Jun 8	Jul 7	4
4	Jun 14	Jul 26	12
5.1	Jul 27	Aug 30	11
5.2	Aug 1	Aug 30	8
6	Aug 17	Sep 5	6
7	Aug 10	Oct 15	38

References

1. Ibryam, Bilgin, From Fragile to Antifragile Software.
2. redhat, Chapter 20. Application Health.
3. Omar, S., Ngadi, A., Jebur, H. H., Machine Learning Techniques for Anomaly Detection: An Overview, *International Journal of Computer Applications*, October 2013
4. Zamani, M., Movahedi, M., Machine Learning Techniques for Intrusion Detection, May 2015
5. Patcha, A., Park, J., An overview of anomaly detection techniques: Existing solutions and latest technological trends, *Elsevier B.V.*, 2007