

A big data infrastructure for predictive maintenance of large-scale data centers

Friday, 28 August 2020 09:00 (30 minutes)

Predictive maintenance is emerging as a new trend in research, due to its advantages compared to the alternative methodologies of corrective and preventive maintenance. The ability to predict faults and intervene before they occurs, allows saving money in a wide set of application domains, among which management of data centers. Savings are usually directly proportional to the size of the involved entities. Due to the novelty of this approach, and to the variety and heterogeneity of the involved data sources, identifying, extracting and processing valuable information in an efficient and effective way is a challenging task. In fact, in such scenario data sources may include log files produced by each computing node, as well as infrastructure monitoring data (e.g. cabinets and rack sensors) and environmental data produced by sensors (e.g., temperature/humidity, fire, flooding etc.) installed in the data center.

We hereby present a layered scalable big data infrastructure aimed at predictive maintenance for large data centers. Despite leveraging open source Apache technologies, the proposed infrastructure (supporting both batch and real-time analysis) follows a general approach ensuring its portability to different frameworks. On the bottom level, Apache Flume performs data ingestion dealing with different data sources (e.g. syslog, time series databases). Timely distribution to processing nodes is carried out at a higher level by the topic-based publish-subscribe engine Apache Kafka. Processing is performed through Apache Spark and Spark Streaming instances. Data persistence is achieved through Apache's distributed filesystem HDFS, that allows saving both the original log files and the results of the analysis. The topmost layer constitutes the presentation layer, including the visualization tools through which the final users (i.e., sysadmins) may monitor the status of the system be notified about foreseen faults. This work is framed in the context of the INFN Tier-1 data center, involving data from approximately 1200 nodes. DODAS (Dynamic On Demand Analysis Service) has been adopted to deploy and easily replicate the deployment of the analysis cluster. DODAS is a Platform as a Service (PaaS) tool allowing local and remote deployment with a minimal effort, based on the specifications included in TOSCA templates. It supports any cloud provider, only requiring the access credentials. Within DODAS, the Infrastructure Manager (IM) provides an abstraction over the underlying architecture.

Primary author: Dr VIOLA, Fabio (INFN-CNAF)

Co-authors: Dr FALABELLA, Antonio (INFN); Dr MARTELLI, Barbara (INFN - CNAF); Prof. BONACORSI, Daniele (University of Bologna); Ms DECKER DE SOUSA, Leticia (Università di Bologna (UNIBO) and Italian Institute of Nuclear Physics (INFN)); Dr SPIGA, daniele (INFN-PG); Mr ROSSI TISBENI, simone (INFN - CNAF)

Presenter: Dr VIOLA, Fabio (INFN-CNAF)

Session Classification: Infrastructure Clouds and Virtualisation Session

Track Classification: Infrastructure Clouds and Virtualisation