Big Data Meets HPC Log Analytics: Scalable Approach to Understanding Systems at Extreme Scale

Achievement: Designed and implemented HPC log data analytics framework

Significance and Impact: Accomplished Big Data approach to HPC log analytics framework for a wide range of researchers. The framework tightly couples scalable data (Cassandra) and compute (Spark) models to enable various analytic capabilities. As exascale systems coming available, the framework will play important role in understanding system health, status, and utilization of the system.

Research Details:

- Designed both offline and online analytics framework based on Big Data and NoSQL technologies.
- Coupling of data model and compute model to maximize data locality
- Integration of event streams from OLCF with Spark Streaming capabilities for realtime event monitoring and analytics

Sponsor/Facility: Office of Advanced Scientific Computing Research (program manager Lucy Nowell). Resources of OLCF and CADES are used.

PI and affiliation: Christian Engelmann – Oak Ridge National Laboratory

Team: Byung-Hoon (Hoony) Park, Saurabh Hukerikar, Ryan Adamson, and Christian Engelmann

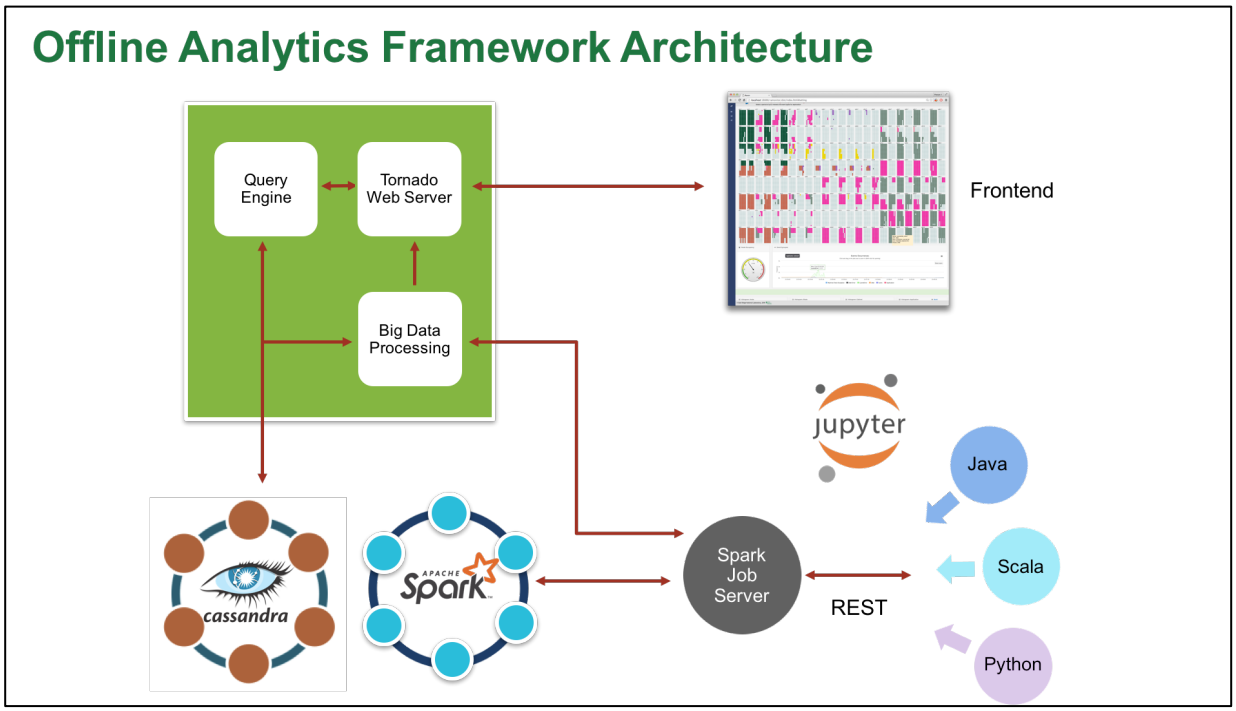

Figure 1 Offline Monitoring & Analytics Framework. The framework offers not only simple data queries but also compute intensive queries through the default analytic frontend. The framework also supports user customized analytic codes written in various languages through Spark job server.

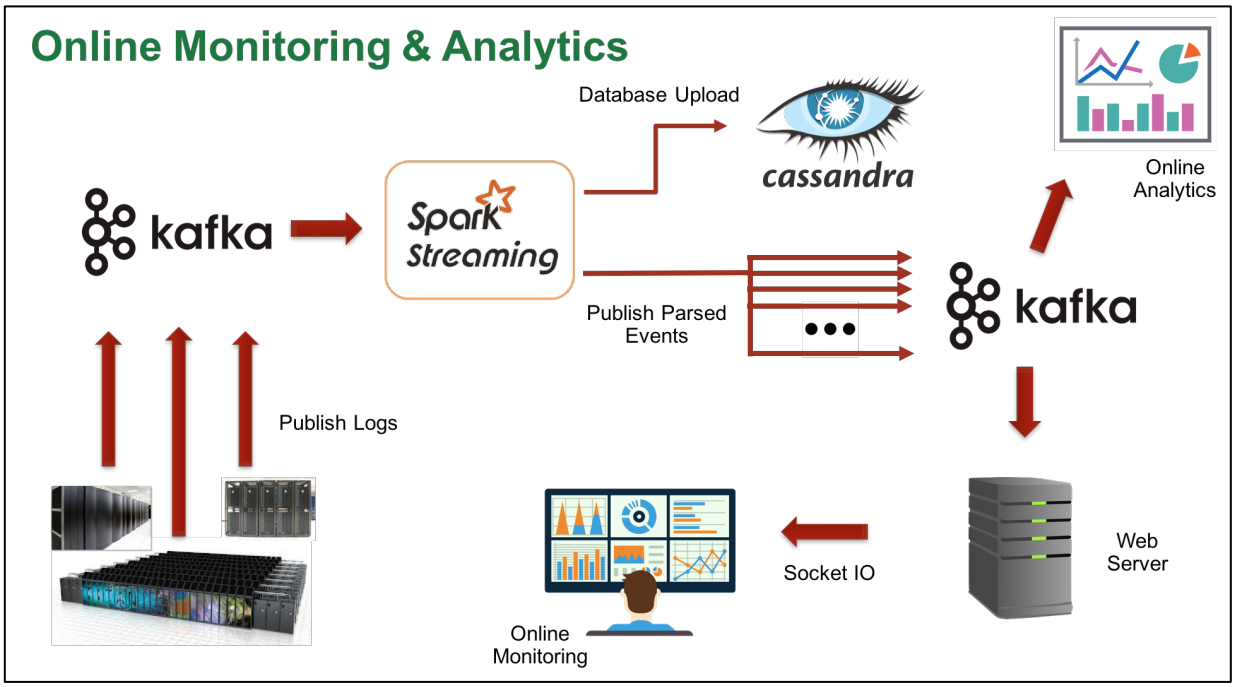

Figure 2 Online Monitoring & Analytics Framework. The online framework ingests event streams from OLCF and processes them using Spark Streaming capability. The results are forwarded to real-time monitoring and analytics components.

Publication: Byung H. Park, Saurabh Hukerikar, Ryam Adamson, and Christian Engelmann, ***Big Data Meets HPC Log Analytics: Scalable Approach to Understanding Systems at Extreme Scale***, Workshop on Monitoring and Analysis for High Performance Computing Systems Plus Applications (HPCMASPA) in IEEE Cluster 2017

Overview:

Today's high-performance computing (HPC) systems are heavily instrumented, generating logs containing in- formation about abnormal events, such as critical conditions, faults, errors and failures, system resource utilization, and about the resource usage of user applications. These logs, once fully analyzed and correlated, can produce detailed information about the system health, root causes of failures, and analyze an application's interactions with the system, providing valuable insights to domain scientists and system administrators. However, processing HPC logs requires a deep understanding of hardware and software components at multiple layers of the system stack. Moreover, most log data is unstructured and voluminous, making it more difficult for system users and administrators to manually inspect the data. With rapid increases in the scale and complexity of HPC systems, log data processing is becoming a big data challenge. This work introduces a HPC log data analytics frame- work that is based on a distributed NoSQL database technology, which provides scalability and high availability, and the Apache Spark framework for rapid in-memory processing of the log data. The analytics framework enables the extraction of a range of information about the system so that system administrators and end users alike can obtain necessary insights for their specific needs. We describe our experience with using this framework to glean insights from the log data about system behavior from the Titan supercomputer at the Oak Ridge National Laboratory.