conferenceseries.com SciTechnol

J Comput Eng Inf Technol, 6:5
DOI: 10.4172/2324-9307-C1-015

4th International Conference on

BIG DATA ANALYSIS AND DATA MINING

September 07-08, 2017 | Paris, France

Memory management in modern data analytics platforms

Mayuresh Kunjir
Duke University, USA

The field of data processing is seeing an explosion of computation platforms such as Hadoop, Spark, HBase, Cassandra, Flink, Storm, and others. Some key observations of these platforms are: They are JVM-based and therefore developer friendly, stable, and secure; they run in multi-tenant environments where resources are isolated using containers, e.g. Yarn, Mesos, and Docker and; they use memory extensively both for data storage and computations. The central premise of this work is that data analytics will increasingly be done in memory on shared-nothing clusters running JVM inside containers. Given this, it is high time for a better understanding of memory management in these modern data analytics platforms. This work uses Thoth, a data-driven platform for multi-system cluster management developed at Duke University, to build a deep understanding of different interplays of memory-management decisions made at multiple levels such as the resource manager (e.g. Yarn), the application framework (e.g. Spark), and the JVM. An empirical study is carried out over a suite of industry-standard analytical workloads to show the impact of various memory management options on response time, efficiency, and reliability. A set of guidelines follows to ensure a safe and efficient execution of the workloads.

mayuresh.kunjir@gmail.com