Abstract

The increasing trends of data growth with the Internet and Internet of Things (IoT), the big data topic is becoming not only important but also very challenging for Data Centers. Apache Hadoop is a framework that allows for the distributed processing of huge amount of datasets across clusters of computers. Big Data Analytics applications have already started to move beyond the classic Hadoop architecture towards very close to real-time architectures such as Spark etc. In this sense, a fundamental understanding of a Hadoop and MapReduce principles and services (e.g. Hive, HBase etc.,) where operates on top of the Hadoop core, can be considered a very good starting point to have a good view of the Big Data World. This manuscript presents not only the design and deployment, but also a performance evaluation of benchmarks and stress testing of a Hadoop cluster. Given the fact that the raspberry pi is an affordable single board computer (SBC) gives the chance to everyone to enhance its knowledge and contribute, in a reasonable degree to the academic community, based on Raspberry Pi 2 abilities as an integrated computer. The current model is comprised of 15 low cost Raspberry Pi 2 model B computers with CPU 900 MHz, 32-bit quad-core ARM Cortex-A7 CPU processors.
and RAM 1GHz each node. The most common benchmarking and testing tools that are included in the Apache Hadoop distribution, are the TestDFSIO, TeraSort, NNBench and MRbench tools. Broadly speaking, the above mentioned tools are very popular choices to benchmark and stress test a Hadoop cluster to measure the performance, to compare the results and to share the outcome with other people who are interested in the topic. In this project the TestDFSIO tool is used to stress test the Hadoop cluster.

References


Index Terms

Computer Science Information Sciences
Keywords

Raspberry Pi Hadoop cluster, Cloud Computing, Hadoop, Big Data, Big Data Analytics, Parallel Computing, MapReduce, Hadoop cluster benchmark.