Hadoop and Map Reduce Performance Analysis

Team Member:

1. Deep Desai (Student ID : 1059666)
2. Tapan Parikh (Student ID : 1101133)
3. Yash Modi (Student ID :)

- The objective of your project:

Mainly Hadoop is for process very large data without any worry about the structure of data. Large data is not like 1 or 10 GB data. It's around almost 10-100 Gigabytes Data.
The main objective of the project is to measure the performance of the cluster system using Hadoop and MapReduce Algorithm. We will measure the performance measurement of Single node cluster and multiple node clusters and evaluate the result of the every node.

- Proposed deliverables

In first phase, we will configure the Single node environment on Linux system with Hadoop 2.7 and Java 7. After that we will perform the Hadoop distributed file system (HDFS) of the data-set. See the figure for the HDFS architecture [7]. Once data-set gets distributed we will apply MapReduce algorithm to process vast amount of data in parallel. Test preparation to compute cluster performance. Now we will apply different test cases by changing the cluster parameter such as **data-set size**, **data replication value** and **data block size** to analyze the cluster performance in different environments.

On successfully completion of first phase. In second phase, we will configure the multiple node environments and will follow the same procedure as single node environment for performance analysis.

After completion of both the phases we will analyze the performance of Single node environment over Multiple node environments or vice-versa.

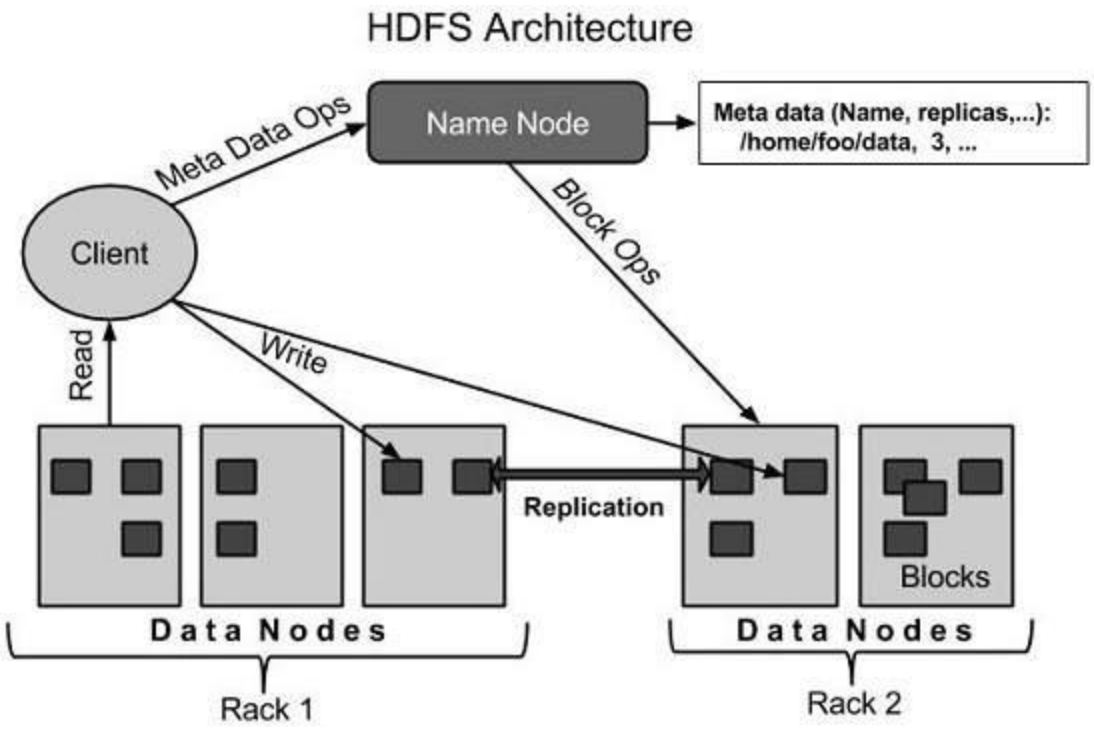
- Proposed project Timeline

We are planning to complete first phase of the project after in October and then will plan to learn clustering and try to complete it before December. In December we could focus on testing and project documentation.

- Team collaboration (task assigned to each member)

Mainly we are very new to this technology so first of all we focused on learning and sharing information on Prerequisite, Some Research Paper and also get information regarding big data concepts and Hadoop framework and its applications. Along with this we look into some the big-data data-set on which we can carry out our plans.

After this stuff, we plan to setting up the environment in individual systems. On completion we will start learning the MapReduce algorithm in detail simultaneously we start implementing WordCount example on individual systems to get clear understanding of Hadoop and MapReduce Algorithm. Then 2 members will start working MapReduce Algorithm for actual system and 1 member will start looking into the establishment of the cluster environment for the second phase.



Source: *Tutorials Point [7]*

References

1. http://www1.cse.wustl.edu/~jain/cse570-13/ftp/bigdatap/index.html
2. https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
3. https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html
4. https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/ClusterSetup.html
5. http://www.michael-noll.com/blog/2011/04/09/benchmarking-and-stress-testing-an-hadoop-cluster-with-terasort-testdfsio-nnbench-mrbench/
6. http://www.tutorialspoint.com//hadoop/index.htm
7. http://www.tutorialspoint.com/hadoop/hadoop_hdfs_overview.htm

Data-set Reference

1. https://snap.stanford.edu/data/amazon-meta.html
2. https://www.kaggle.com/competitions