16:137:602:900C
Introduction to Cloud and Big Data Systems (Spring 2017)

Assignment 2

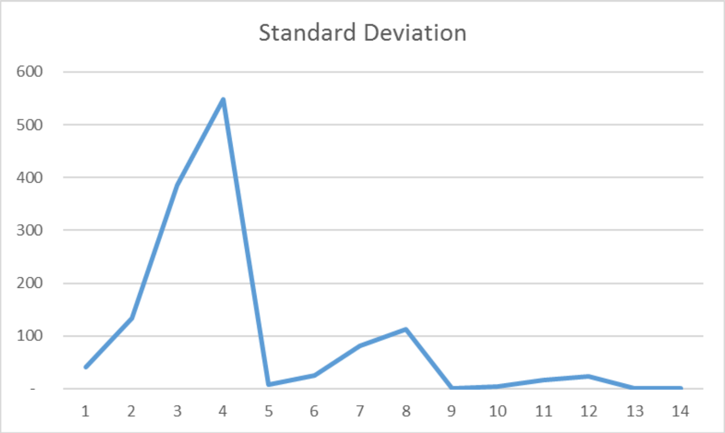
Maruthi Ayyappan – Aishwarya Gunde – Beethoven Plaisir

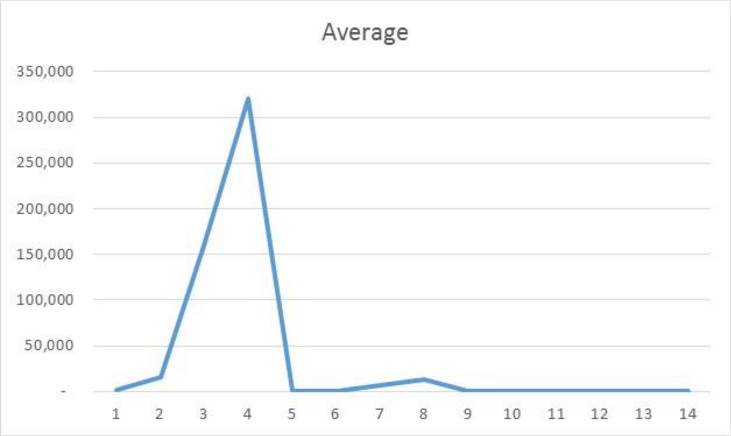
1. **Provide the code that you have developed to run the proposed problem in Hadoop/HDFS.**

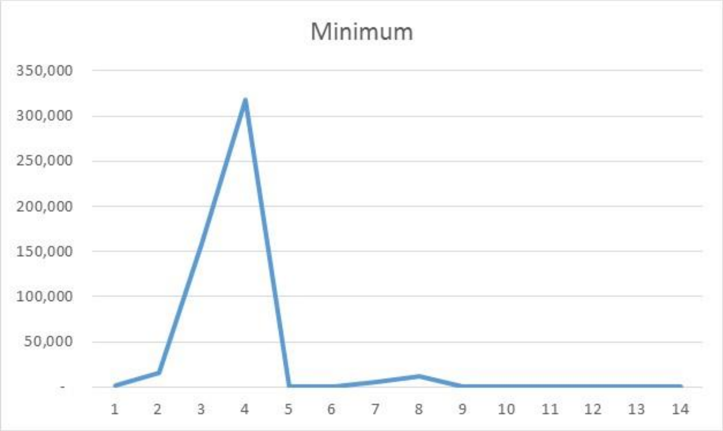
Ans. Code included in the CBD A2 Code.txt

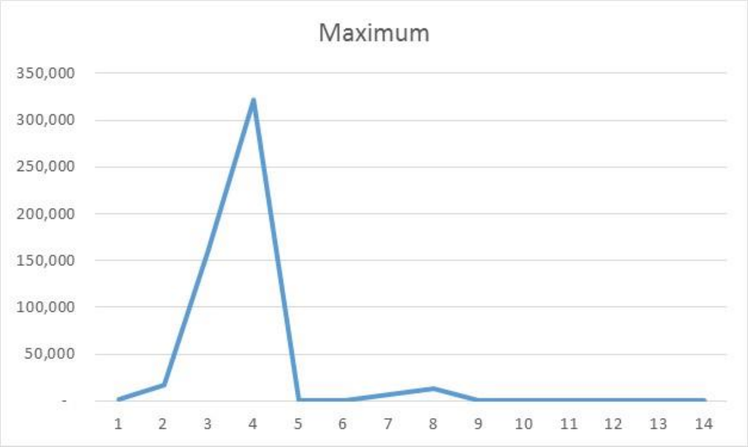
2. **Provide the results obtained (statistics e.g., in a plot).**

Ans.









3. **Provide your observations and conclusions based on your experiments.**

Ans. **MapReduce** is a programming model for distributed computing containing two major tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are split down into key-value pairs. Reduce task takes the output from a map as an input and combines those data pairs into a smaller set of tuples. The reduce task is always performed after the map job.

1. Mapping: The map or mapper's job is to process the input data and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of key-value pairs of data.

2. Reducing: The Reducer's job is to process the data taken as an input from the mapper. It produces a new set of output, integrating the tuples, and store it in the HDFS.

3. Hadoop then sends the Map and Reduce tasks to the appropriate servers in the cluster. HDFS manages issuing tasks, verifying task completion, and copying data around the cluster between the nodes.

4. The cluster collects and reduces the data to give required output, and sends it back to the Hadoop server after completion of the given tasks.

4. Specific question: describe the order in which map and reduce tasks are executed in Hadoop for different configurations.

Ans.

- It can be observed that mapping and reducing takes place simultaneously, i.e. reducer begins approximately after 50% of mapping, for words with less number of characters. However, for the words with more number of characters, reducer begins implementation once mapping is done 100%.
- The total megabyte-milliseconds taken by all map tasks ranges from minimum 3803136 for small length words to maximum 399156224 for higher length words.
- The total megabyte-milliseconds taken by all reduce tasks ranges from minimum 1972224 for small length words to maximum 189047808 for higher length words.