Project Paper Draft (CSP 554 Big Data Technologies)

Project Topic: Factorial of a Number using Big Data Technologies.

Introduction:

In this project, we use Big Data Technologies to find out the factorial of a number and benchmark the results with traditional methods.

Description:

The goal of the project is to use Big Data Technologies like map-reduce, spark and benchmark the timing results with traditional methods like recursion and loop.

Big Data Tools:

HDFS:

The Hadoop Distributed File system (HDFS) is a distributed file system designed to run on commodity software. It is highly fault-tolerant and designed to be deployed on low-cost hardware. HDFS provides high throughput to application data and is highly useful for applications that have large datasets. It is reliable as the data is duplicated on multiple nodes in a cluster. HDFS provides an interface for applications to move themselves closer to where the data is located. HDFS has been designed to be easily portable from one platform to another. Moreover, it supports a very traditional hierarchical file organization

Map-Reduce:

MapReduce is Hadoop's primary framework for processing big data on a shared cluster. It allows for the distributed processing of the map and reduction operations. Maps can be performed in parallel, provided that each mapping operation is independent of the others; in practice, this is limited by the number of independent data sources and/or the number of CPUs near each source. Similarly, a set of 'reducers' can perform the reduction phase, provided that all outputs of the map operation that share the same key are presented to the same reducer at the same time, or that the reduction function is associative.

Spark:

Spark is a distributed and open-source processing system. It is used for the workloads of 'Big data'. Spark utilizes optimized query execution and in-memory caching for rapid queries across any size of data. It is simply a general and fast engine for much large-scale processing of data.

References:

https://www.javatpoint.com/spark-big-data
http://datascienceguide.github.io/map-reduce
https://blog.matthewrathbone.com/2013/11/17/python-map-reduce-on-hadoop-a-beginners-tutorial.html
http://techsquids.com/bd/matrices-sum-map-reducer/
https://medium.com/@aw.shubh/matrix-multiplication-through-map-reduce-c72be2f4f90