

**CSE488: Big Data Analytics**

**[SPRING 2023]**

**Lab-4 Offline Task**

**Computing Average and Performance Comparison**

**Submitted by:**

Student ID: **2019-3-60-046**

Student Name: Mohsenul Kabir Mithun

# Comparison between Hadoop MapReduce program with HDFS and C++ Program in LFS

Execution timetable is given bellow:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Dataset Size | Map Time (M) (millisecond) | Reduce Time (R) (millisecond) | Total (M+P) Time (millisecond) | C++ Time (millisecond) |
| 10000 | 7485 | 10857 | 18342 | 163 |
| 30000 | 8499 | 8155 | 16654 | 454 |
| 60000 | 7690 | 8104 | 15794 | 903 |

# Scatter Plotting:

# Bar Chart:

**Basic knowledge:**

Hadoop is a distributed computing framework that is designed to handle large-scale data processing tasks, particularly those involving Big Data. Hadoop is optimized for handling large data sets and for distributing the processing of those data sets across multiple machines in a cluster. This distributed processing capability makes Hadoop well-suited for handling large-scale data processing tasks that require high scalability and fault tolerance.

On the other hand, C++ is a programming language that is well-suited for building high-performance, low-latency applications. C++ can be used to implement algorithms that are optimized for specific hardware architectures, which can result in very fast execution times for certain tasks.

**Observation:**

If you are performing a simple average calculation operation on a small data set, then C++ become faster than Hadoop, since the overhead of setting up a Hadoop cluster and distributing the processing across multiple machines may outweigh the benefits of parallel processing. However, if you are working with a large data set or if you need to perform complex data processing tasks, then Hadoop becomes faster than C++, since it can distribute the processing across multiple machines and scale up as needed.

Based on the information presented in the **scatter plot and bar chart** diagrams, it appears that as the size of the dataset increases, the performance of C++ worsens, while the performance of Hadoop improves with an increasing number of datasets. This observation is consistent with the general characteristics of C++ and Hadoop, where C++ may be limited by the resources of a single machine, whereas Hadoop's distributed processing capability allows for better scalability and potentially faster processing times.