Literature Survey – Map Reduce

# Student Details:

Shyam Kantesariya (40042715) Reza Behbahani (40039400)

# Introduction:

Considering the growth of data day-by-day, a sequential system would not be able to process this large volume of data within operational time. We need to have a parallel processing approach to extract value from this data. Google has published an approach to analyze this Big data using MapReduce framework which provided a base for Hadoop distributed computing platform, a top level Apache open source project which is being implemented by all the major organizations including Google, Facebook, Yahoo to name a few.

# Topics to be covered in this survey

In spite of complete abstraction of decomposition and fault tolerant capabilities, MapReduce programming can’t be the solution for all parallel computing use cases as there is no one-fits-all real world solution.

Considering that we would be discussing following topics as part of this literature survey:

* Motivation behind MapReduce framework and how it is different from MPI programming
* Components of MapReduce framework
  + Map
  + Sort
  + Partition
  + Shuffle
  + Merge
  + Reduce
* Performance issues and area of improvements
  + Absence of execution plan
  + Management of intermediate results between Map and Reduce phases
  + I/O optimization
* MapReduce vs Parallel Databases
* Industry use case of MapReduce
* Recent research in MapReduce

To support our content, we would be referring papers as attached:

 