MapReduce: Simplified Data Processing on Large Clusters

The paper describes about MapReduce which is a model that helps in processing and generating big data sets. The model is used by Google which allows the programmers to manage large amount of data. This process is followed by the sections which are, Programming model, Implementation, Refinements, and Performance. Map function takes an info pair and delivers a bunch of intermediate key-value pairs and the Reduce function is additionally composed by the client and acknowledges a intermediate key and a bunch of values for that key, then, at that point consolidates these values to frame a perhaps more modest set of values. The execution for map is portrayed as dividing the input data which turns out in M parts and for lessen is depicted by parceling the key in R pieces. The master assists with putting away data structures which has the state and identity for the machine inside it. The undertaking ought to be partitioned in appropriate extent as to get legitimate output file. Estimations of the assignments differ dependent on the implementation. For the arranging task, typical execution time is generally not exactly the no-reinforcement execution time. At the point when a few pieces of the program were deliberately eliminated to test the machine disappointments, the program naturally re-executed the connected cycle and it requires some time as the ordinary execution time. The usefulness of Map/Reduce is sufficiently adequate, however it very well may be improved by partitioning , combiner functions, avoiding the terrible records and executing a nearby framework to assist with making the troubleshooting system simpler.The Map/Reduce model is very easy to use. Hence, it is used various domains such as Google News, Google search engine, Google Zeitgeist. So, from the all side of view, we can conclude that it can be applied to machine learning or clustering problems. Also, when the programming models are limited it becomes reliable to discover the issue, network transmission capacity is likewise saved and the superfluous execution helps in disappointment and information loss in machines.