4. Internals of MapReduce

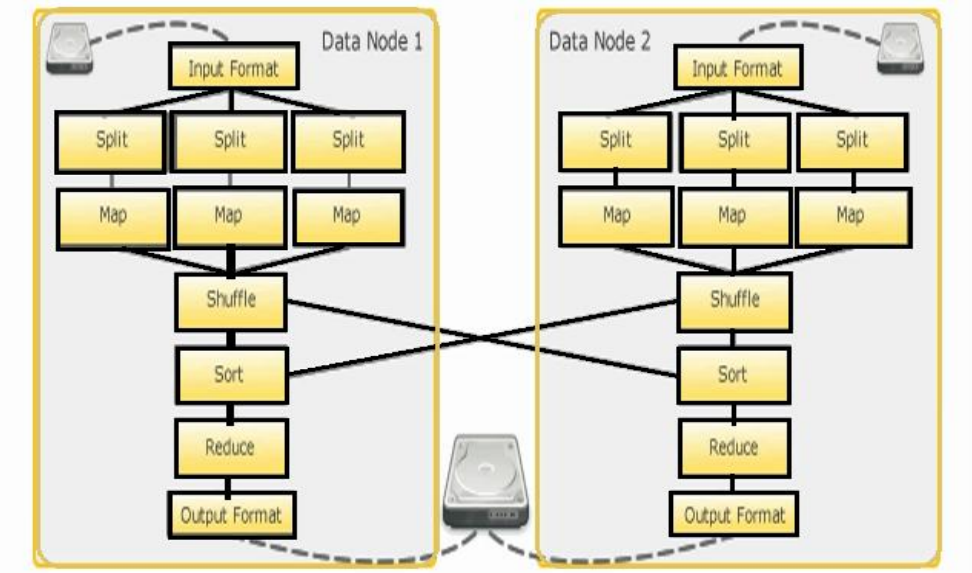

Figure 2: Internals of the MapReduce

In Figure 2, we can see the internals of the MapReduce. These are the phases of the MapReduce. The flows through these phases and gives us output

4.1 Spilt Phase

The split phase uses the input format to bring data off the disk out of HDFS and split it up. The default input format is the text input format which breaks up the data line by line. Each line is split and send to the mapper. So, if we have large data we could have thousands and thousands of mappers running (depending on the cluster setup) simultaneously against this data .There are variety of input format depending on the kind of the data which we storing the HDFS. If we storing the image data there is binary input format. If we storing database there is database input format .

4.2 Map

The mapper just goes and gets the data we want. It runs on the key that we pairs. It transforms the input split into the pairs based on user-defined code. Mapper is based on keys get the values we want.

4.3 Shuffle & Sort

It gonna takes all the data nodes that are part of this job, shuffle which is portioning and grouping the data and sorting it and send it to the reducers.

4.4 Reducers

Reducers work on the sorted data and aggregated all the results.

4.5 Output Format

Output Format sends the data to the HDFS.

5. Programming Model

To Use MapReduce, the programmer expresses their desired computation as a series of job. The input to a job is an input specification that will yield key-value pairs. Each job consists of two stages : first a user defined map function is applied to each input records to produce a list of intermediate key-value pairs .Second a user-defined reduce function is called once for each distinct key in map output and passed the list of intermediate values associated with key. The MapReduce framework automatically parallelizes the execution of these functions and ensures fault tolerance.

Optionally, the user can supply a combiner function. Combiners are similar to reduce functions, except that they are not passed all the values for a given key: instead, a combiner emits an output value that summarizes the input value is passed.

```
public interface Mapper<K1, V1, K2, V2> {
void map(K1 key, V1 value,
OutputCollector<K2, V2> output);
void close();
}
```

6. Execution in MapReduce

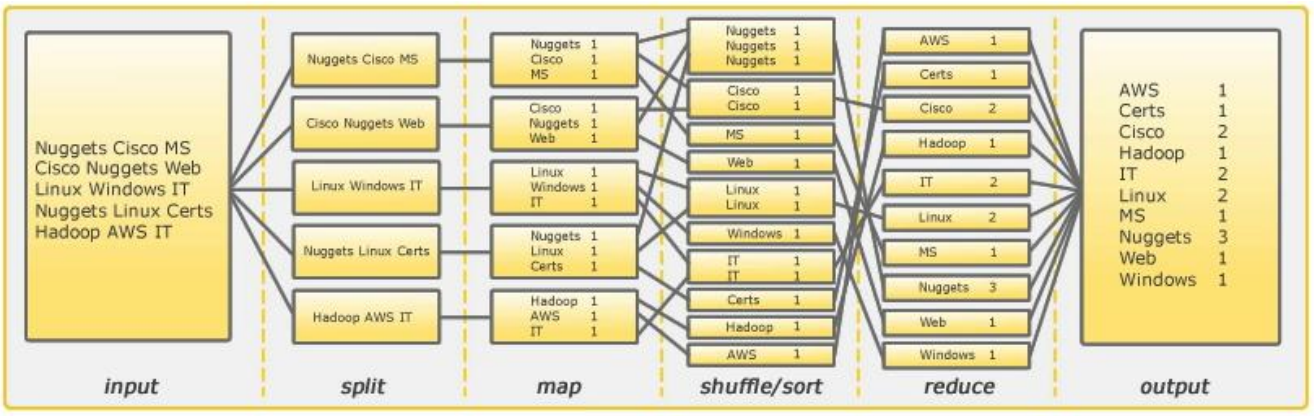

Figure 3: Data Through the MapReduce Internals

In figure 3. We have a file in input containing 5 lines .We split our input. We split our file line by line and then we will send it to the mapper. In mapper the line is broken down into words with count next to them. Then shuffle and sort phase takes all the data to the mapper, brings all the data together and shuffle it or sort it and sends it to the reduce .The reducer have simple job of aggregating all the results and give us final output

7. Conclusion

The MapReduce programming model has been successfully used at Google and many companies for different purposes .We assign this success to several reasons. First the model is easy to use. Without programmer's parallel and distributed system, it hides the details of parallelization, fault tolerance, locality optimization and load balancing. Second, very large problems are easily expressible using MapReduce as MapReduce computations. Third, the MaprReduce implementation makes efficient use of machine resources and hence suitable for many large companies like Google.

We have learnt several things from this work. First, what is really MapReduce is and application and key features of the MapReduce. Second, we get the internals of the MapReduce which give us deep knowledge of MapReduce. Third, we get how data is implemented in the MapReduce with the help of its internals and also learnt about its components and their workings.

8. Acknowledgement

The author is grateful to the participants who contributed to research and our guide Dr. Sanjay Srivastava.

References

[1] Shipa, Manjit kaur, "*Big Data and Methodology",* 10 Oct, 2013
[2] Pareedpa**,** A.; Dr.Antony Selvadoss, "*Significant Trends of Big Data*", 8 Aug, 2013
[3] Gurpeet Singh Bedi, Ashima, "*Big Data Analysis with Dataset Scaling in Yet another Resource Negotiator (YARN)",* 5April, 2013
[4] *Hadoop-The Definitive Guide*, Tom White, Edition-3, 27Jan, 2012
[5] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta,Kumar N, "*Analysis of Big data using Apache Hadoop and MapReduce",* Volume 4, May 2014
[6] IBM 2012, What is big data: Bring big data to the enterprise,http://www.01.ibm.com/software/data/bigdata/, IBM
[7] Sam Madden, *"From Databases to Big Data",* IEEE computer society,2012
[8] *"Data Mining with BigData"* ,Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding , 1041-4347/13/$31.00 © 2013 IEEE
[9] Russom, *"Big Data Analytics"* , TDWI Research,2011
[10] An Oracle White Paper, *"Hadoop and NoSQL Technologies and the Oracle DataBase",* February 2012
[11] "*MapReduce Online*", Tyson Condie ,Neil Conway ,Peter Alvaro ,Joseph Hellerstein