Software Tools

LatexAssignment

Hadoop- MapReduce

Name: Rohan Baghel
Student Number: 2021PCS1025

1 Introduction

MapReduce is a Hadoop framework used for writing applications that can process vast amounts of data on large clusters. It can also be called a programming model in which we can process large datasets across computer clusters. This application allows data to be stored in a distributed form. It simplifies enormous volumes of data and large scale computing.

There are two primary tasks in MapReduce: map and reduce. We perform the former task before the latter. In the map job, we split the input dataset into chunks. Map task processes these chunks in parallell. The map we use outputs as inputs for the reduce tasks. Reducers process the intermediate data from the maps into smaller tuples, that reduces the tasks, leading to the final output of the framework.

2 How MapReduce in Hadoop works

MapReduce Architecture and **MapReduce's phases** will help to understand working of MapReduce in Hadoop

2.1 MapReduce architecture

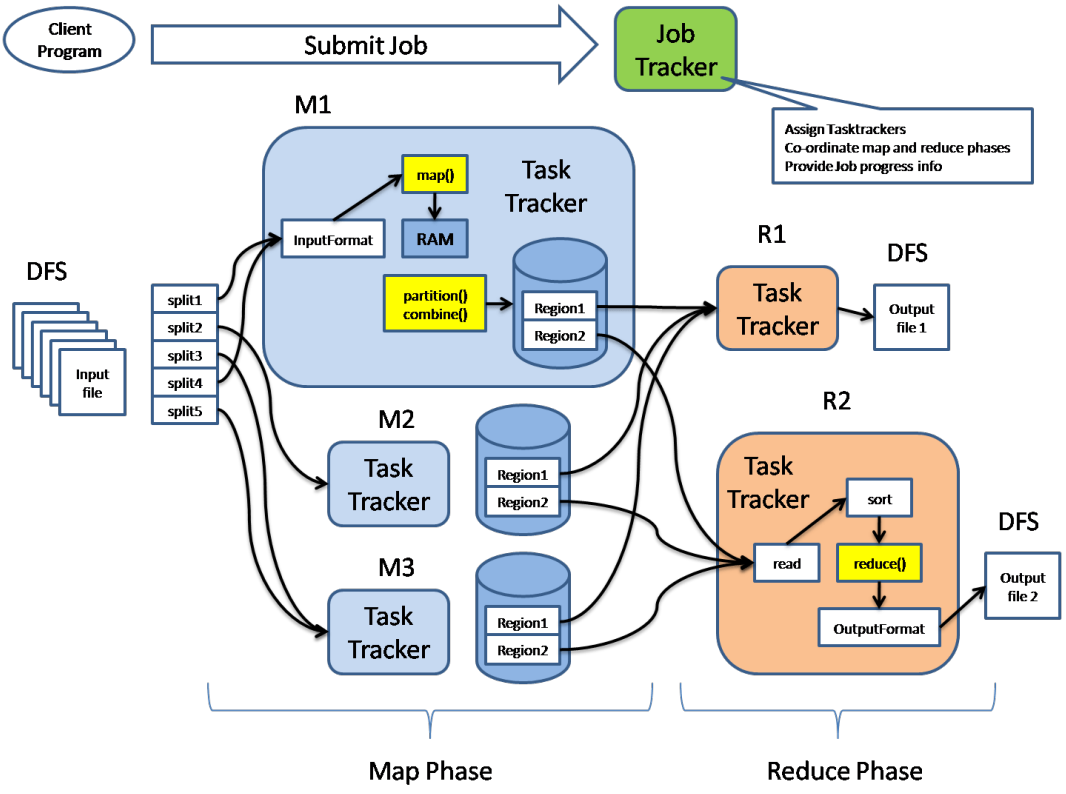

Figure 1: Map Reduce Architecture

MapReduce architecture consists of various components. Components are.

- **job:** This is the actual work that needs to be executed or processed
- **Task:** This is a piece of the actual work that needs to be executed or processed. A MapReduce job comprises many small tasks that need to be executed.
- **Job Tracker**: This tracker plays the role of scheduling jobs and tracking all jobs assigned to the task tracker.
- **Task Tracker:** This tracker plays the role of tracking tasks and reporting the status of tasks to the job tracker.
- **Input data:** This is the data used to process in the mapping phase.
- **Output data:** This is the result of mapping and reducing.
- **Client:** This is a program or Application Programming Interface (API) that submits jobs to the MapReduce. MapReduce can accept jobs from many clients.
- **Hadoop MapReduce Master:** This plays the role of dividing jobs into job-parts.
- **Job-parts:** These are sub-jobs that result from the division of the main job.

In the MapReduce architecture, clients submit jobs to the MapReduce Master. This master will then sub-divide the job into equal sub-parts. The job-parts will be used for the two main tasks in MapReduce: mapping and reducing.

The developer will write logic that satisfies the requirements of the organization or company. The input data will be split and mapped.

The intermediate data will then be sorted and merged. The reducer that will generate a final output stored in the HDFS will process the resulting output.

Diagram: Flow Diagram for the map reduce program

Figure 2: Flow diagram

2.1.1 How job trackers and task trackers work

Every job consists of two key components: mapping task and reducing task. The map task plays the role of splitting jobs into job-parts and mapping intermediate data. The reduce task plays the role of shuffling and reducing intermediate data into smaller units.

The job tracker acts as a master. It ensures that we execute all jobs. The job tracker schedules jobs that have been submitted by clients. It will assign jobs to task trackers. Each task tracker consists of a map task and reduces the task. Task trackers report the status of each assigned job to the job tracker. The following diagram summarizes how job trackers and task trackers work.

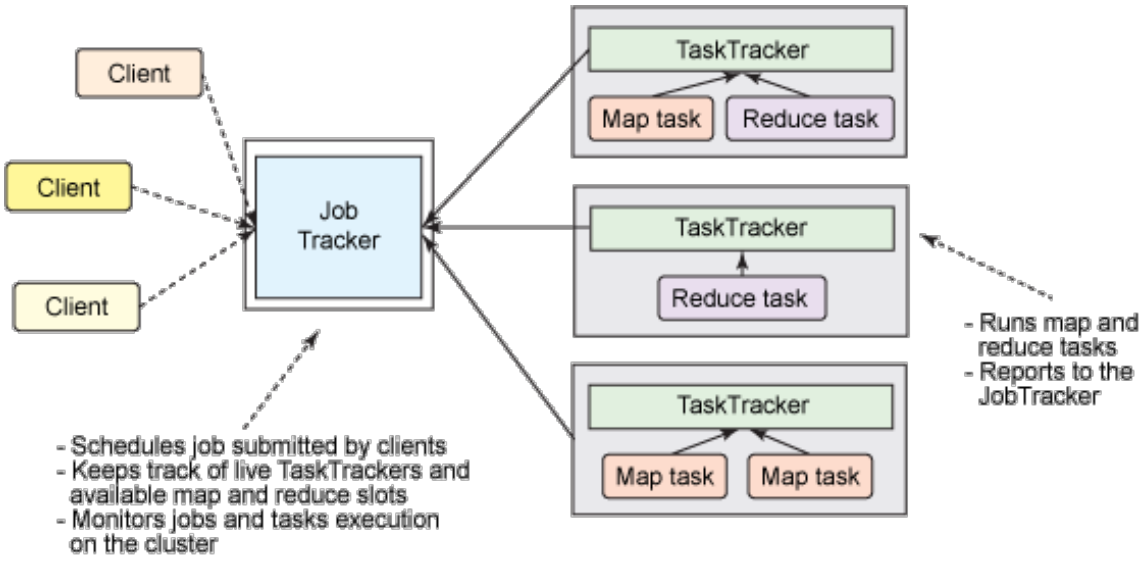

Figure 3: Flow diagram

2.2 Phase of MapReduce

The MapReduce program is executed in three main phases:

1. Mapping
2. Shuffling
3. Reducing

1 Mapping Phase

This is the first phase of the program. There are two steps in this phase: splitting and mapping. A dataset is split into equal units called chunks (input splits) in the splitting step. Hadoop consists of a RecordReader that uses TextInputFormat to transform input splits into key-value pairs.

The key-value pairs are then used as inputs in the mapping step. This is the only data format that a mapper can read or understand. The mapping step contains a coding logic that is applied to these data blocks. In this step, the mapper processes the key-value pairs and produces an output of the same form (key-value pairs).

2 Shuffling Phase

This is the second phase that takes place after the completion of the Mapping phase. It consists of two main steps: sorting and merging. In the sorting step, the key-value pairs are sorted using the keys. Merging ensures that key-value pairs are combined.

The shuffling phase facilitates the removal of duplicate values and the grouping of values. Different values with similar keys are grouped. The output of this phase will be keys and values, just like in the Mapping phase.

3 Reducing

In the reducer phase, the output of the shuffling phase is used as the input. The reducer processes this input further to reduce the intermediate values into smaller values. It provides a summary of the entire dataset. The output from this phase is stored in the HDFS.

Example

Example of MapReduce with the three main phases.

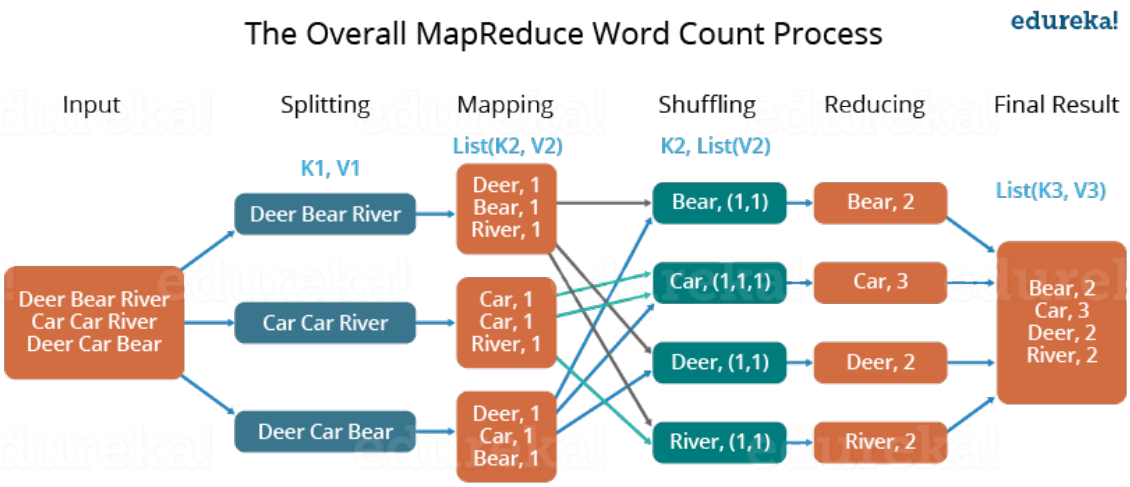

Figure 4: Example

4 Combiner Phase

This is an optional phase that's used for optimizing the MapReduce process. It's used for reducing the pap outputs at the node level. In this phase, duplicate outputs from the map outputs can be combined into a single output. The combiner phase increases speed in the Shuffling phase by improving the performance of Jobs.

Diagram Shows how all the four phases of MapReduce have been applied.

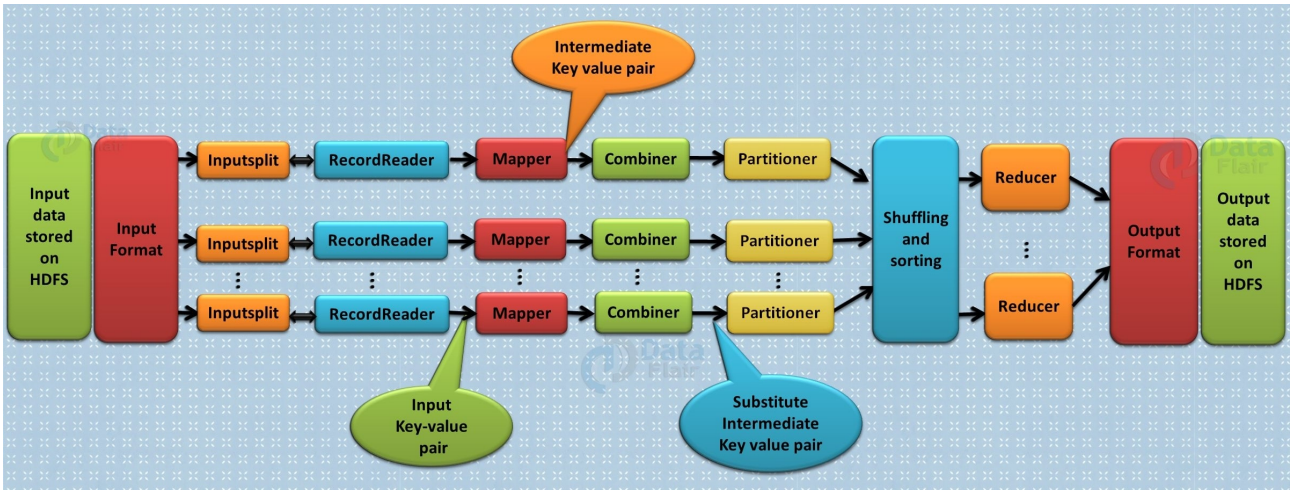

Figure 5: Combiner Phase

3 Benefits of Hadoop MapReduce

- **Speed:** MapReduce can process huge unstructured data in a short time.
- **Fault-tolerance:** The MapReduce framework can handle failures.
- **Cost-effective:** Hadoop has a scale-out feature that enables users to process or store data in a cost-effective manner.
- **Scalability:** Hadoop provides a highly scalable framework. MapReduce allows users to run applications from many nodes.
- **Data availability:** Replicas of data are sent to various nodes within the network. This ensures copies of the data are available in the event of failure.
- **Parallel Processing:** In MapReduce, multiple job-parts of the same dataset can be processed in a parallel manner. This reduces the time taken to complete a task.

4 Applications of Hadoop MapReduce

some of the practical applications of the MapReduce program

- **E-commerce**
  E-commerce companies such as Walmart, E-Bay, and Amazon use MapReduce to analyze buying behavior. MapReduce provides meaningful information that is used as the basis for developing product recommendations. Some of the information used include site records, e-commerce catalogs, purchase history, and interaction logs.
- **Social networks**
  The MapReduce programming tool can evaluate certain information on social media platforms such as Facebook, Twitter, and LinkedIn. It can evaluate important information such as who liked your status and who viewed your profile.
- **Entertainment**
  Netflix uses MapReduce to analyze the clicks and logs of online customers. This information helps the company suggest movies based on customers' interests and behavior.

5 Conclusion

MapReduce is a crucial processing component of the Hadoop framework. It's a quick, scalable, and cost-effective program that can help data analysts and developers process huge data.

This programming model is a suitable tool for analyzing usage patterns on websites and e-commerce platforms. Companies providing online services can utilize this framework to improve their marketing strategies.