hadoop Map Reduce

Prepared by: Sunil Kumar

What is Map Reduce

- **Map Reduce** is a software framework and programming model used for processing huge amounts of data.
- **MapReduce** is the processing layer of **Hadoop.**
- MapReduce programming model is designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks
- It was developed in 2004, on the basis of paper titled as **"MapReduce: Simplified Data Processing on Large Clusters,"** published by Google.
- Hadoop divides the job into tasks. **MapReduce** program work in two phases:

  1. **Map tasks** (Splits & Mapping)
  2. **Reduce tasks** (Shuffling, Reducing)
- Hadoop is capable of running Map Reduce programs written in various languages:

  1. Java 2. Ruby 3.Python 4.C++.

- The whole process of MapReduce goes through the four phases of execution –
  1. Splitting 2. Mapping 3. Shuffling 4. Reducing.

Task – A task in MapReduce is an execution of a Mapper or a Reducer on a slice of data.

- It is also called Task-In-Progress (TIP).
- It indicate that processing of data is in progress either on mapper or reducer.

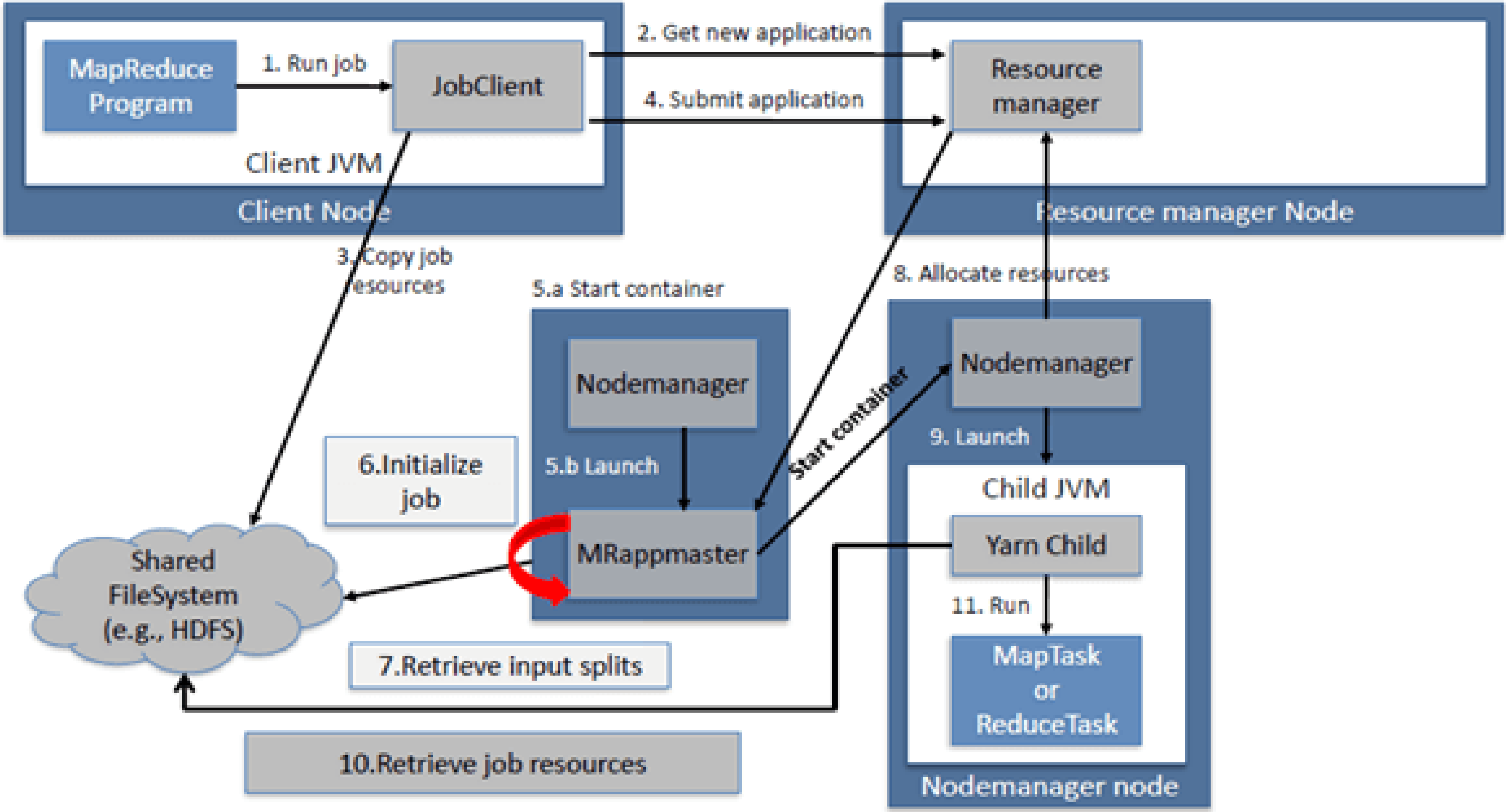
Task Attempt-

Task Attempt is a particular instance of an attempt to execute a task on a node. There is a possibility that anytime any machine can go down.

Example- While processing data if any node goes down, framework reschedules the task to some other node. This rescheduling of the task cannot be infinite. There is an upper limit for that as well. The default value of task attempt is 4.

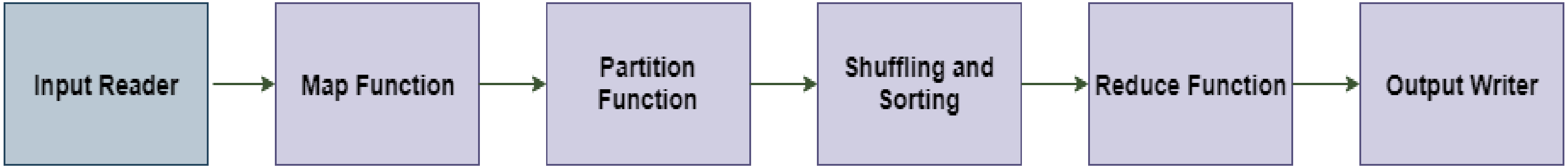
If a task (Mapper or reducer) fails 4 times, then the job is considered as a failed job. For high priority job or huge job, the value of this task attempt can also be increased.

Steps used in MapReduce



Data Flow In MapReduce

- MapReduce is used to compute the huge amount of data . To handle the upcoming data in a parallel and distributed form, the data has to flow from various phases.



Input reader - The input reader reads the upcoming data and splits it into the data blocks of the appropriate size (64 MB to 128 MB). Each data block is associated with a Map function.

Map function - The map function process the upcoming key-value pairs and generated the corresponding output key-value pairs. The map input and output type may be different from each other.

Partition function -

The partition function assigns the output of each Map function to the appropriate reducer. The available key and value provide this function. It returns the index of reducers.

Shuffling and Sorting -

The data are shuffled between/within nodes so that it moves out from the map and get ready to process for reduce function.

- The sorting operation is performed on input data for Reduce function. Here, the data is compared using comparison function and arranged in a sorted form.

Reduce function - The Reduce function is assigned to each unique key. These keys are already arranged in sorted order. The values associated with the keys can iterate the Reduce and generates the corresponding output.

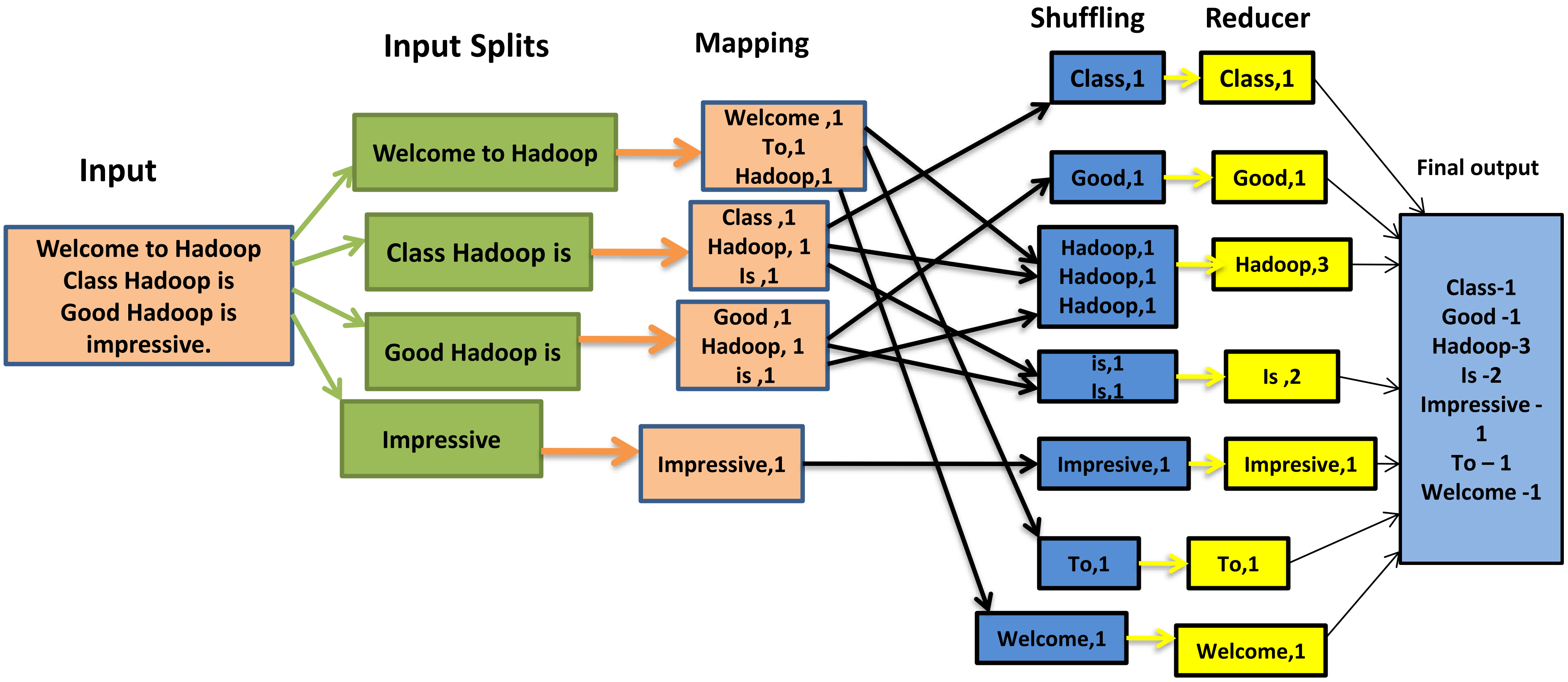
Output writer

- Once the data flow from all the above phases, Output writer executes. The role of Output writer is to write the Reduce output to the stable storage.

Input Split

- **Input Split** in Hadoop **MapReduce** is the logical representation of data.
- by default, breaks a file into **128MB** chunks (same as blocks in HDFS). Input Split in Hadoop is user defined. User can control split size according to the size of data in MapReduce program.
- The client (running the job) can calculate the splits for a job by calling **getSplit().**

Working Example of Map Reduce



Features of Map Reduce

- **Scalability.** Apache Hadoop is a highly scalable framework. .
- **Flexibility.** Map Reduce programming enables companies to access new sources of data. ...
- **Security and Authentication**. ...
- **Cost-effective solution**. ...
- **Fast**. ...
- **Simple model of programming**. ...
- **Parallel Programming**. ...
- **Availability and resilient nature**.

Advantage of MapReduce

Fault tolerance: It can handle failures without downtime.
Speed: It splits, shuffles, and reduces the unstructured data in a short time.
Cost-effective: Hadoop MapReduce has a scale-out feature that enables users to process or store the data in a cost-effective manner.
Scalability: It provides a highly scalable framework. MapReduce allows users to run applications from many nodes.
Parallel Processing: Here multiple job-parts of the same dataset can be processed in a parallel manner. This can reduce the task that can be taken to complete a task.

Limitations Of MapReduce

- MapReduce cannot cache the intermediate data in memory for a further requirement which diminishes the performance of Hadoop.
- It is only suitable for Batch Processing of a Huge amounts of Data.

Application Of MapReduce

- **Entertainment:** To discover the most popular movies, based on what you like and what you watched in this case Hadoop MapReduce help you out. It mainly focuses on their logs and clicks.
- **E-commerce:** Numerous E-commerce suppliers, like Amazon, Walmart, and eBay, utilize the MapReduce programming model to distinguish most loved items dependent on clients’ inclinations or purchasing behavior.
- It incorporates making item proposal Mechanisms for E-commerce inventories, examining website records, buy history, user interaction logs, etc.
- **Data Warehouse:** We can utilize MapReduce to analyze large data volumes in data warehouses while implementing specific business logic for data insights.
- **Fraud Detection:** Hadoop and MapReduce are utilized in monetary enterprises, including organizations like banks, insurance providers, installment areas for misrepresentation recognition, pattern distinguishing proof, or business metrics through transaction analysis.

What is alternative to MapReduce?

Best alternate for MapReduce is **Spark**, because its 10 to 100 times faster than the MapReduce. And also very easy to maintain, less coding high performance.

Thank You