Software Tools

Hadoop MapReduce

Rohan Baghel
2021PCS1025
Computer Science and Engineering
India Institute of Technology Jammu

July 21, 2025

Overview

Introduction

- MapReduce a powerful paradigm for Parallel computing
- Hadoop uses MApReduce to execute jobs on files in HDFS
- Hadoop will intelligently distribute computation over cluster
- Take computation to data

Data Flow

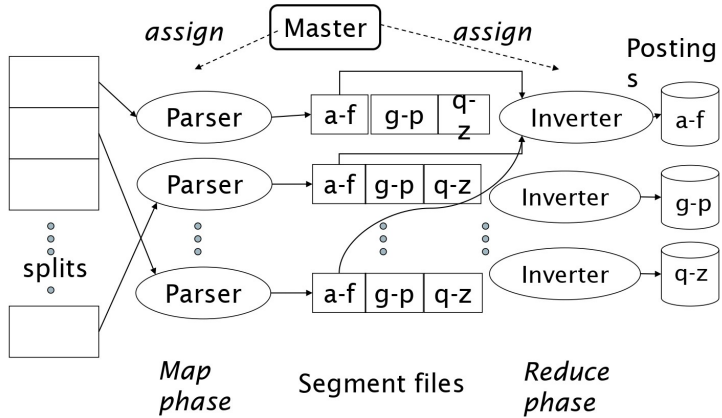

Figure: Map Reduce

Map Reduce Architecture

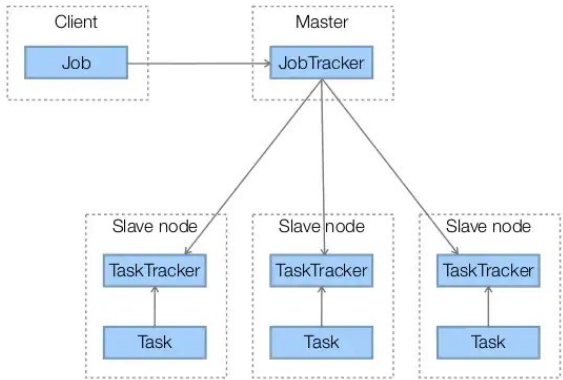

Figure: Map Reduce Architecture

Map Reduce Architecture

- Each node is part of Hadoop file system(HDFS) cluster.
- Input Data is stored in HDFS spread across nodes and replicated.
- Programmer submits job (Mapper, Reduce, input) to job tracker.
- Job Tracker -Master
  * splits input data
  * Schedules and monitor various map and reduce tasks
- Task Tracker- Slaves
  * Execute map and reduce asks
- Execute map and reduce asks

Stages of Map Reduce

3 Stages

1. Map Stage
2. Shuffle Stage
3. Reduce Stage

1. Map Stage

- The map or mapper's job is to process the input data.
- Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS).
- The input file is passed to the mapper function line by line.
- The mapper processes the data and creates several small chunks of data.

Map Stage

- Records are input as key/value pair
- Mapper output one or More intermediate key/value pair for each input.
- **map(K1 key, V1 Val, OutputCollector¡K2, V2¿output, Reporter reporter)**

Map Stage Example

- **Input:** ¡key:, offset, value:lien of a document¿
- **Output:** for each word w in input line output¡key:w value:1¿
  - **Input:**(2133, Student of Indian Institute of Technology JAMMU.)
  - **Output:**(student,1),(of,1),(indian,1),.....,(technology,1),(jammu,1)

2. Shuffle and sort Phase

- Map task output in partitioned by hashing the output key
- Number of partitions is equal to number of reducers
- Partitioning ensures all key/value pairs sharing same key belong to same partition
- The map output partition is sorted by key to group all values of the same key

3. Reduce Stage

- This stage is the combination of the
  1. Shuffle
  2. Reduce
- The Reducer's job is to process the data that comes from the mapper.
- After processing, it produces a new set of output, which will be stored in the HDFS.

Reducer Stage

- After the map phase, all the intermediate values for a given output key are combined together into a list
- reducer combines those intermediate values into one or more final key/value pair
- **reduce(K2 key, Iterator¡V2¿values, OutputCollector¡K3, V3¿ output, Reporter reporter)**

Reducer stage Example

- **Input:** ¡key: word, value:list¡integer¿¿
- **Output:** sum all values from input for the given key input list of values and output ¡Key:word value:count¿
  - **Input:** (the,[1,1,1,1],(fox,[1,1,1])…..
  - **Output:** (the, 5) (fox,3)……

Conclusion

- Map Reduce greatly simplifies writing large scale distributed application
- Used for Building search index at Google, Amazone
- Widely used for analyzing user logs, data warehousing and analytics
- Also used for large scale machine learning and data mining application.

THANK YOU