SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Artificial Intelligence and Data Science

Course Name – Big Data Analytics
III Year / V Semester

Unit 3 – DATA ANALYTICAL FRAMEWORKS

Topic - MapReduce

MapReduce

- MapReduce is a programming model for writing applications that can process Big Data in parallel on multiple nodes.
- MapReduce provides analytical capabilities for analysing huge volumes of complex data.
- MapReduce is a processing technique and a program model for distributed computing based on java.
- The MapReduce algorithm contains two important tasks, namely Map and Reduce.
- The Map task takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pairs).
- The Reduce task takes the output from the Map as an input and combines those data tuples (key-value pairs) into a smaller set of tuples.
- The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers.

MapReduce

The Algorithm

- MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

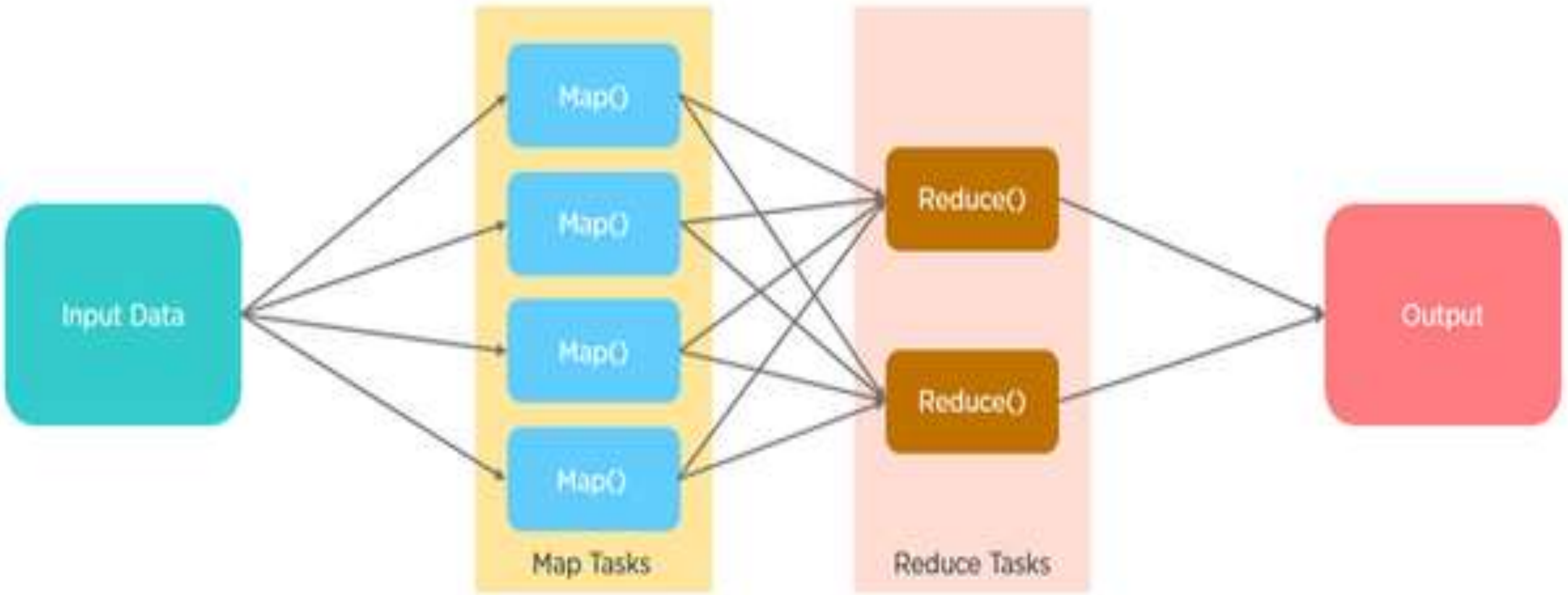
Map stage :

- 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.

Reduce stage:

- This stage is the combination of the Shuffle stage and the Reduce stage.
- 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.

MapReduce


Input Data
Map()
Map()
Map()
Map()
Map Tasks
Reduce()
Reduce()
Reduce Tasks
Output

MapReduce

Phases

- Input Phase − Here we have a Record Reader that translates each record in an input file and sends the parsed data to the mapper in the form of key-value pairs.

- Map − Map is a user-defined function, which takes a series of key-value pairs and processes each one of them to generate zero or more key-value pairs.

- Intermediate Keys − They key-value pairs generated by the mapper are known as intermediate keys.

- Combiner − It takes the intermediate keys from the mapper as input and applies a user-defined code to aggregate the values in a small scope of one mapper.

- Shuffle and Sort − It downloads the grouped key-value pairs onto the local machine, where the Reducer is running. The individual key-value pairs are sorted by key into a larger data list. The data list groups the equivalent keys together so that their values can be iterated easily in the Reducer task.

MapReduce

- Reducer – The Reducer takes the grouped key-value paired data as input and runs a Reducer function on each one of them. Here, the data can be aggregated, filtered, and combined in a number of ways, and it requires a wide range of processing. Once the execution is over, it gives zero or more key-value pairs to the final step.

- Output Phase – In the output phase, we have an output formatter that translates the final key-value pairs from the Reducer function and writes them onto a file using a record writer.

THANK YOU