DSBDAL

Assignment No:- 11

* Aim:- Write a code in Java for a single simple Word count the number of occurence of each word in given input set using the hadoop, map reduce framework on local stand alone setup

* Theory :- map reduce word count is a framework which splits the chunk of data, slots the map reduce tasks. A file-system stores the o/p & i/p of jobs. Re-execution of failed tas of the framework and monitoring them is the task of the framework.

* Map function -
Create and process the import data take in data converts it into a set of other data where the breakdown of individuals element into these tuples is done.
No API contract requiring a certain number of outputs.

* Reduce function:-
Mappers output is posted into the reduct processes the data into something usable every single mapper is passed into the reduced function.

The new o/p values are saved in HDFS.
A concept called streaming is used in a writing a code for Word count.
In python using mapReduce, let's look at the reducer code and How to execute that using a streaming a Jar file.

* How MapReduce Works?

The MapReduce algorithm contains two important task, namely map and Reduce.
The map task takes a set of data & converts it into another set of data, where individual elements are broken down into tuples (key - value pairs).

MapReduce task takes the output from the map as an input and combines those data tuples into a smaller set of tuples.
The reduce task is always performed after the map job.

1) Input phase -
Here we have a record reader that translates each record that in on input file and sends that 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 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:-
A combiner is a type of local reducer that groups similar data from the map phase into identifiable sets.

shuffle and sort - The reducer task starts with the shuffle and sort step.

* Reducer-
The Reducer takes the grouped key-value paired data as input and runs a reducer function on each of them.

* Conclusion:- Hence we studied about mapreduce.

Q.1) Explain Apache Hadoop :-
→ -Apache Hadoop project develops open-

source s/w for reliable, scalable & distributed computing.
- Apache Hadoop s/w library is a framework.
- It allows distributed processing of large data sets across clusters of computers using simple programming models.

* modules in Apache Hadoop:-
① mapReduce -
- This is a system for parallel processing of large sets.

2) Hadoop YARN -
This is a framework for job scheduling and clusters resource management.

3) HDFS - (Hadoop Distributed file system).
- HDFS is a unique design that provides storage for extremely large files with streaming data.

4) Hadoop common:-
- These are common utilities that support the other Hadoop modules.

Q.2) Explain mapReduce.
→ mapReduce is a programming model for processing large datasets using parallel, distributed and clustered compute nodes.