truelines
DATE : _/_/___

DS & BDA Group A (Exp. 2)

Title: Design a distributed application using mapReduce

Objective
1.) To explore different Big Data processing techniques with use cases.
2) To study detailed concept of mapReduce.

Software Requirements
1) Ubuntu 14.04 / 14.10
2) GNN C Compiler
3) Hadoop
4) JAVA

Problem Statement: Design a distributed applications using MAP Reduce (using JAVA) which processes a log file of a system. List out the users who have logged for maximum period on the system. Use simple log file from the internet & process it using a pseudo distribution mode on Hadoop platform.

Theory:
Introduction:
MapReduce is a framework using which we can write app^n to process huge amounts of data in parallel on large clusters of commodity hardware in a reliable manner.
^y

truelines

DATE : _/_/_

MapReduce is a processing techniques & a program model. For distributed computing based on JAVA

The MapReduce Algorithm contains two important tasks namely Map & Reduce. Map takes a set of data & converts it onto another set of data where individual elements are broken down to tuples

Secondly, reduce tasks which takes the output from map as an input & combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

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 & reducers. ~~Dem~~ Decomposing a data processing application into mapper & reducers is sometimes nontrivial. But once we write an application in map-Reduce form scaling the applicalon to run over hunderds, thousand or even ten of thosands of machines in cluster is merely a configuration change.

ALGORITHM: MapReduce program executes in 3 stages namely map stage, shuffle stage & reduce stage

truelines
DATE : _/_/_

- Map stage: The map or mappers job is to process the input data. Generally the input data is in form of file or directory & is stored in Hadoop file

- Shuffle stage: This stage combines all value associated to an identical key. For eg. (Are, 1) is there three times in input file so after the shuffling phase, the output will be like (Are, [1, 1, 1])

- Reduce stage: This stage is combination of shuffle stage & reduce stage. The reducers 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 HDFS.

* Inserting data into HDFS

Map reduce framework operates on <key, value pairs that is the framework views the input to the job as set of <key, value> pairs & produces set of <key value> pairs as the output of the job, conceivably of different types.

The key & the value classes should be serialized manner by framework & hence need to implement the Writables interface. Additionally the key classes have to implement the writable comparable interface to facilitate sorting by framework.

truelines

DATE : _/_/_

Input & output types of MapReduce job:
(Input) $<K_1 V_1>$ → map → $<K_2 V_2>$ → reduce → $<K_3 V_3>$ (output)

Conclusion: Thus we have learnt how to design a distributed application using MapReduce & process a log file of system.