DISCUSS HADOOP AS A DISTRIBUTED COMPUTING SYSTEM

HASHIM ATHMAN ABDALLA
P15/81781/2017
28/09/2019

What is Hadoop?

Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving large amounts of data and computation e.g. in big data analysis.[1]

Development of Hadoop was inspired by the realization that it was quickly becoming useful for people to store and analyze data-sets far larger than can be stored and accessed on one physical machine. Hence the use of distributed system was necessary

What makes Hadoop a distributed system?

Hadoop is a distributed system and it is made up of four modules which work together. The modules includes: -

1. **Hadoop Distributed File-System**
2. **MapReduce**
3. **Hadoop Common**
4. **Yarn.**

The Hadoop Distributed File-System (HDFS) module is the primary data storage system used by Hadoop applications and it is the one which makes the Hadoop system a distributed system[2]. The HDFS employs a *NameNode* and *DataNode* architecture to implement a distributed file system which sits on top of the file system define by the operating system of a physical machine.

When HDFS takes in data it breaks its down into separate blocks and distributes them to different nodes in the cluster. It then transfers packaged code into nodes to process the data in parallel hence achieving the main goal of distributed system to distribute resource and hide the distribution to the user since the user will be accessing the resources as a single system (***transparency***).

As a distributed system the Hadoop Distributed File-System is highly fault-tolerant. The file system replicates, or copies each piece of data multiple times and distributes the copies to individual nodes, placing at least one copy on a different machine[3]. If one node crashes the data is recovered from the other nodes.

Hadoop as a distributed system is able to achieve ***concurrency*** as operations on the chunks of data stored in different nodes are carried parallel to each other. This make the operations faster compared to doing it on a single physical machine.

Hadoop as a distributed system achieve ***relocation*** of resource by the fact that it supports rapid transfer of data between computing nodes in a given cluster of computers.

In conclusion Hadoop system through its module Hadoop Distributed File-System is able to employ all the functionalities of a distributed system.

1 https://en.wikipedia.org/wiki/Apache_Hadoop
2 https://searchdatamanagement.techtarget.com/definition/Hadoop-Distributed-File-System-HDFS
3 https://searchdatamanagement.techtarget.com/definition/Hadoop-Distributed-File-System-HDFS