Article

Why We Need a Distributed Computing System And Hadoop Ecosystem

By **Abdul Rasheed Feroz Khan** on **Dec 07 2016**

Introduction

This article is a continuation of Hadoop – Distributed Computing Environment. We will be developing knowledge about why we need Hadoop and the ecosystem of Hadoop here.

Link

Click here to get details about distributed computing environments.

Working with distributed systems needs software which can coordinate and manage the processors and machines within the distributed environment. As the scaling of giant corporations like Google keeps on increasing, they started to build new software that can run on all the distributed systems.

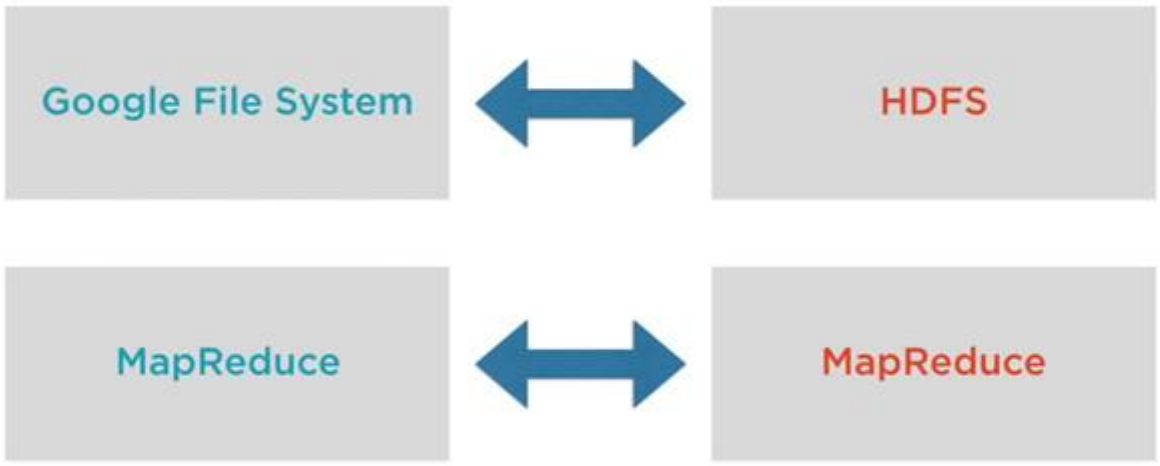
Objective of the software that was developed for distributed systems will be, as shown below.

1. Store millions of records (raw data) on multiple machines, so keeping records on what record exists on which node within the data center.
2. How do we run the processes on all these machines to simplify the data?

Thus, Google worked on these two concepts and they designed the software for this purpose.

- *Google File System* was used to store the data in a distributed manner and to solve the distributed storage.
- *Map Reduce* was used to solve the distributed computing running across multiple machines and to bring in data running across multiple machines to hop in something useful.

Both of these combine together to work in Hadoop.



Google File System works namely as Hadoop Distributed File System and Map Reduce is the Map Reduce algorithm that we have in Hadoop. Hence, HDFS and MapReduce join together with Hadoop for us.

HDFS is a file system that is used to manage the storage of the data across machines in a cluster. Perhaps MapReduce is a framework to process the data across the multiple Servers. Hadoop is distributed by Apache Software foundation whereas it's an open source.

In 2013, MapReduce into Hadoop was broken into two logics, as shown below.

1. MapReduce
2. YARN

Now, MapReduce framework is to just define the data processing task. It was focused on what logic that the raw data has to be focused on.

YARN is a framework again, which will be running the data processing task across the multiple machines, managing memory, managing processing etc.

Work allocation of Hadoop

MapReduce	Here, the user defines map and reduces tasks, using the MapReduce API. It can help us to work with Java and other defined languages. Map defineIt id program is packed into jobs which are carried out by the cluster in the Hadoop.
YARN	A job is triggered into the cluster, using YARN. It checks whether the node has the resources to run this job or not.
HDFS	YARN should sketch how and where to run this job in addition to where to store the results/data in HDFS.

Hadoop Ecosystem

Hadoop Ecosystem holds the following blocks.

Hive	It seems to be like a SQL query interface to data stored in the Big Data system.
HBase	It is a different kind of the database.
Pig	It allows us to transform unstructured data into a structured data format.
Oozie	It is a workflow management system.
Flume/Sqoop	It allows us to add data into Hadoop and get the data from Hadoop.
Spark	It allows us to perform computations in a functional manner at Big Data.

Thank you for using C# Corner