Task 1 Explain the core changes made in Hadoop 2.x

-> The progress from a Hadoop 1's more restricted processing model of batch oriented Map Reduce jobs, to more interactive and specialized processing models of Hadoop 2 will only further position the Hadoop ecosystem as the dominant big data analysis platform.

-> Map Reduce, as implemented in Hadoop 1, can be I/O intensive, not suitable for interactive analysis, and constrained in support for graph, machine learning and on other memory intensive algorithms. Hadoop developers rewrote major components of the file system to produce Hadoop 2.

-> Two of the most important advances in Hadoop 2 are the introduction of HDFS federation and the resource manager YARN.

-> HDFS

1) HDFS comprises two major components: namespaces and blocks storage service.

2) Operations on files and directories are managed by the name spac eservice manager, such as creating and modifying files and directories. The block storage service implements data node cluster management, block operations and replication.

3) In Hadoop 1, a single Name node manages the entire namespace for a Hadoop cluster.

4) HDFS federation, multiple Name node servers manage namespaces and this allows for horizontal scaling, performance improvements, and multiple namespaces.

5) The implementation of HDFS federation allows existing Name node configurations to run without changes. For Hadoop administrators, moving to HDFS federation requires formatting Name nodes, updating to use the latest Hadoop cluster software, and adding additional Name nodes to the cluster.

-> YARN

1) YARN, the other major advance in Hadoop 2, brought major performance improvements for some applications, supports additional processing models and also implements a more flexible execution engine.

2) YARN is a resource manager and it was created by separating the processing engine and resource management capabilities of Map Reduce as it was implemented in Hadoop 1.

3) YARN is often called the operating system of Hadoop because it is responsible for managing and monitoring workloads, maintaining a multi-tenant environment, implementing security controls, and managing high availability features of Hadoop.

4) YARN supports multiple processing models in addition to Map Reduce. One of the most significant benefits of this is that we are no longer limited to working the often I/O intensive, high latency Map Reduce framework.

Task 2 Explain the difference between Map Reduce 1 and Map Reduce 2 / Yarn

Map Reduce 1.0

In Hadoop cluster, racks are interconnected via core switches. Core switches should connect to top-of-rack switches Enterprises using Hadoop should consider using 10GbE, bonded Ethernet and redundant top-of-rack switches to mitigate risk in the event of failure. A file is broken

into 64MB chunks by default and distributed across Data Nodes. Each chunk has a default replication factor of 3, meaning there will be 3 copies of the data at any given time. Hadoop is “Rack Aware” and HDFS has replicated chunks on nodes on different racks. Job Tracker assign tasks to nodes closest to the data depending on the location of nodes and helps the Name Node determine the ‘closest’ chunk to a client during reads. The administrator supplies a script which tells Hadoop which rack the node is in, for example: /enterprise datacenter/rack2.

Limitations of Map Reduce 1.0 – Hadoop can scale up to 4,000 nodes. When it exceeds that limit, it raises unpredictable behavior such as cascading failures and serious deterioration of overall cluster. Another issue being multi-tenancy – it is impossible to run other frameworks than Map Reduce 1.0 on a Hadoop cluster.

Map Reduce 2.0

Map Reduce 2.0 has two components – YARN that has cluster resource management capabilities and Map Reduce.

In Map Reduce 2.0, the Job Tracker is divided into three services:

Resource Manager, a persistent YARN service that receives and runs applications on the cluster. A Map Reduce job is an application. Job History Server, to provide information about completed jobs Application Master, to manage each Map Reduce job and is terminated when the job completes. Also, the Task Tracker has been replaced with the Node Manager, a YARN service that manages resources and deployment on a node. Node Manager is responsible for launching containers that could either be a map or reduce task.

Job Tracker model is broken by this new architecture by allowing a new Resource Manager to manage resource usage across applications, with Application Masters taking the responsibility of managing the execution of jobs. This change removes a bottleneck and lets Hadoop clusters scale up to larger configurations than 4000 nodes. This architecture also allows simultaneous execution of a variety of programming models such as graph processing, iterative processing, machine learning, and general cluster computing, including the traditional Map Reduce.