NameNode and DataNode

Published by Hadoop Team at July 12, 2015

NameNode and DataNode

In this post let's talk about the 2 important types of nodes and it's functions in your Hadoop cluster – NameNode and DataNode.

What is HDFS?

We covered a great deal of information about HDFS in "HDFS – Why Another Filesystem?" chapter in the Hadoop Starter Kit course. If you are new to Hadoop, we suggest to take the free course.

Namenode

1. NameNode is the centerpiece of HDFS.
2. NameNode is also known as the Master
3. NameNode only stores the metadata of HDFS – the directory tree of all files in the file system, and tracks the files across the cluster.
4. NameNode does not store the actual data or the dataset. The data itself is actually stored in the DataNodes.
5. NameNode knows the list of the blocks and its location for any given file in HDFS. With this information NameNode knows how to construct the file from blocks.
6. NameNode is so critical to HDFS and when the NameNode is down, HDFS/Hadoop cluster is inaccessible and considered down.
7. NameNode is a single point of failure in Hadoop cluster.

is usually configured with a lot of memory (RAM). Because the block locations are help in main memory.

ode

1. DataNode is responsible for storing the actual data in HDFS.

2. DataNode is also known as the Slave
3. NameNode and DataNode are in constant communication.
4. When a DataNode starts up it announce itself to the NameNode along with the list of blocks it is responsible for.
5. When a DataNode is down, it does not affect the availability of data or the cluster. NameNode will arrange for replication for the blocks managed by the DataNode that is not available.
6. DataNode is usually configured with a lot of hard disk space. Because the actual data is stored in the DataNode.

Hardware Configuration

Hardware configuration of nodes varies from cluster to cluster and it depends on the usage of the cluster. In Some Hadoop clusters the velocity of data growth is high, in that instance more importance is given to the storage capacity. If the SLAs for the job executions are important and can not be missed then more importance is give to the processing power of nodes.

Often the term “Commodity Computers” is misunderstood. Commodity Computers or Nodes does not mean cheap or less powerful hardware, it just means in-expensive computer and deemphasize the need for specialized hardware.

Here is a sample configuration for NameNode and DataNode hardware configuration.

Name Node Configuration

Processors: 2 Quad Core CPUs running @ 2 GHz
RAM: 128 GB
Disk: 6 x 1TB SATA
Network: 10 Gigabit Ethernet

Data Node Configuration

Processors: 2 Quad Core CPUs running @ 2 GHz
RAM: 64 GB
Disk: 12-24 x 1TB SATA
Network: 10 Gigabit Ethernet

Hadoop Team

We are a group of Senior Hadoop Consultants who are passionate about Hadoop and Big Data technologies. Our collective experience ranges from finance, retail, social media and gaming. We have worked with Hadoop clusters ranging from 50 all the way to 800 nodes.

Related posts

March 2, 2017 February 27, 2017 February 23, 2017

Dissecting MapReduce Program (Part 2)

Read more

Dissecting MapReduce Program (Part 1)

Read more

Dissecting MapReduce Components

Read more

ıt