IoT & Applications

Hadoop Cluster Setup

Module-5: Class-3

Silicon ...beyond teaching
Silicon Institute of Technology
Bhubaneswar

Introduction

- In general, a computer cluster is a collection of various computers that work collectively as a single system.
- "A hadoop cluster is a collection of independent components connected through a dedicated network to work as a single centralized data processing resource".
- "A hadoop cluster can be referred to as a computational computer cluster for storing and analyzing big data (structured, semi-structured and unstructured) in a distributed environment".
- "A computational computer cluster that distributes data analysis workload across various cluster nodes that work collectively to process the data in parallel".

Hadoop Cluster Architecture

- A hadoop cluster architecture consists of a data centre, rack and the node that actually executes the jobs.
- **Data centre consists of the racks and racks consists of nodes.**

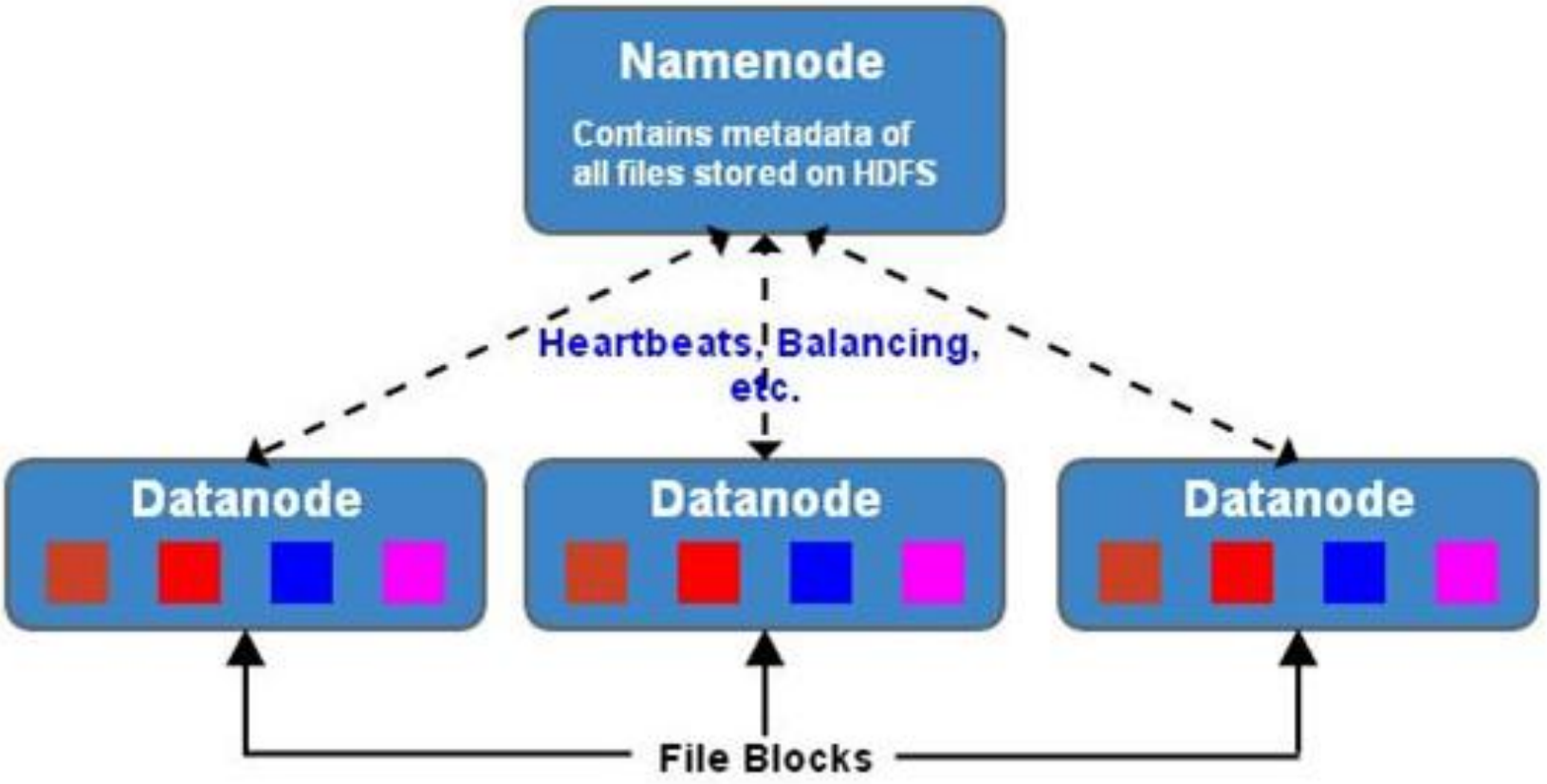
- A medium to large cluster consists of a two or three level hadoop cluster architecture that is built with rack mounted servers.
- Every rack of servers is interconnected through 1 gigabyte of Ethernet (1 GigE).
- Each rack level switch in a hadoop cluster is connected to a cluster level switch which are in turn connected to other cluster level switches or they uplink to other switching infrastructure.

Components of a Hadoop Cluster

Hadoop cluster consists of three components -

- **Master Node –** Master node in a hadoop cluster is responsible for storing data in HDFS and executing parallel computation the stored data using MapReduce. Master Node has 3 nodes – NameNode, Secondary NameNode and JobTracker. JobTracker monitors the parallel processing of data using MapReduce while the NameNode handles the data storage function with HDFS. NameNode keeps a track of all the information on files (i.e. the metadata on files) such as the access time of the file, which user is accessing a file on current time and which file is saved in which hadoop cluster. The secondary NameNode keeps a backup of the NameNode data.
- **Slave/Worker Node-** This component in a hadoop cluster is responsible for storing the data and performing computations. Every slave/worker node runs both a TaskTracker and a DataNode service to communicate with the Master node in the cluster. The DataNode service is secondary to the NameNode and the TaskTracker service is secondary to the JobTracker.
- **Client Nodes –** Client node has hadoop installed with all the required cluster configuration settings and is responsible for loading all the data into the hadoop cluster. Client node submits mapreduce jobs describing on how data needs to be processed and then the output is retrieved by the client node once the job processing is completed.

Components of a Hadoop Cluster



What is cluster size in Hadoop?

A Hadoop cluster size is a set of metrics that defines storage and compute capabilities to run Hadoop workloads, namely :

- **Number of nodes**: number of Master nodes, number of Edge Nodes, number of Worker Nodes.
- **Configuration of each type node**: number of cores per node, RAM and Disk Volume.

Advantages of a Hadoop Cluster Setup

- As big data grows exponentially, parallel processing capabilities of a Hadoop cluster help in increasing the speed of analysis process. However, the processing power of a hadoop cluster might become inadequate with increasing volume of data. In such scenarios, hadoop clusters can scaled out easily to keep up with speed of analysis by adding extra cluster nodes without having to make modifications to the application logic.
- Hadoop cluster setup is inexpensive as they are held down by cheap commodity hardware. Any organization can setup a powerful hadoop cluster without having to spend on expensive server hardware.
- Hadoop clusters are resilient to failure, meaning whenever data is sent to a particular node for analysis, it is also replicated to other nodes on the hadoop cluster. If the node fails then the replicated copy of the data present on the other node in the cluster can be used for analysis.

Challenges of a Hadoop Cluster

- Issue with small files - Hadoop struggles with large volumes of small files - smaller than the Hadoop block size of 128MB or 256MB by default. It wasn't designed to support big data in a scalable way. Instead, Hadoop works well when there are a small number of large files. Ultimately when you increase the volume of small files, it overloads the Namenode as it stores namespace for the system.
- High processing overhead - reading and writing operations in Hadoop can get very expensive quickly especially when processing large amounts of data. This all comes down to Hadoop's inability to do in-memory processing and instead data is read and written from and to the disk.
- Only batch processing is supported - Hadoop is built for small volumes of large files in batches. This goes back to the way data is collected and stored which all has to be done before processing starts. What this ultimately means is that streaming data is not supported and it cannot do real-time processing with low latency.
- Iterative Processing - Hadoop has a data flow structure is set-up in sequential stages which makes it impossible to do iterative processing or use for ML.

Thanks.