Cluster

Krishno Dey

Lecturer

Department of Computer Science and Engineering

Agenda

- What is Hadoop Cluster
- Hadoop Cluster Architecture
- Size of Hadoop Architecture
- Single Node and Muti-Node Cluster
- Communication Protocol in Hadoop Cluster
- Benefits of Hadoop Cluster
- Challenges of Hadoop Cluster

What is Cluster

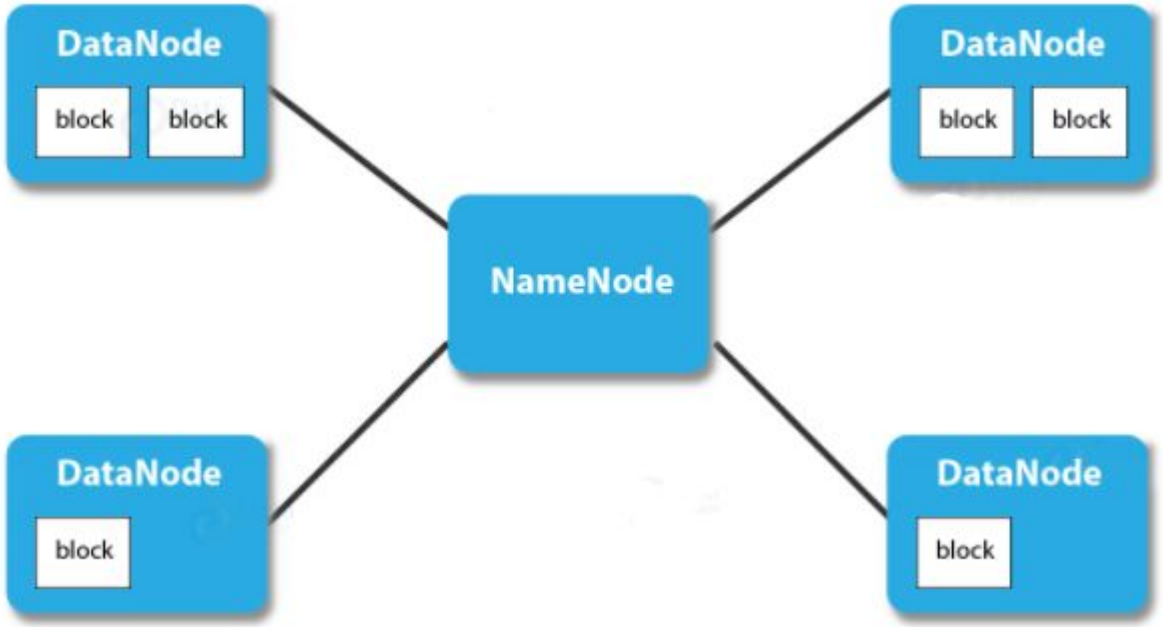
- A Hadoop cluster is a collection of computers, known as nodes, that are networked together to perform parallel computations on big data sets.

- Hadoop clusters consist of a network of connected master and slave nodes that utilize high availability, low-cost commodity hardware.

Hadoop Cluster Architecture

Hadoop cluster has master-slave architecture.

- Master in Hadoop Cluster
- Slaves in Hadoop Cluster



Hadoop Cluster Architecture

Master in Hadoop Cluster Consists of

- NameNode
- Resource Manager

Hadoop Cluster Architecture

Slave in the Hadoop Cluster Consists of

- DataNode
- NodeManager
- Client Node

Single Node Cluster

- Consist of only one node.
- All the NameNode, DataNode, ResourceManager, and NodeManager are on a single machine.
- Only one DataNode is running.
- Mainly used for studying and testing purposes

Multi-Node Cluster

- Consist of multiple node.
- Daemons run on separate host or machine.
- NameNode daemon run on the master machine
- DataNode daemon runs on the slave machines.
- Used in organizations for analyzing Big Data.

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.

Calculating Hadoop Cluster Capacity

For example: Let's say that you need 500TB of space. If you have a JBOD of 12 disks, and each disk can store 6TB of data, then the data node capacity, or the maximum amount of data that each node can store, will be 72 TB. Data nodes can be added as the data grows, so to start with its better to select the lowest number of data nodes required.

In this case, the number of data nodes required to store 500TB of data equals 500/72, or approximately 7.

Communication Protocols Used in Hadoop Clusters

- The HDFS communication protocol works on the top of TCP/IP protocol.
- The client establishes a connection with NameNode using configurable TCP port.
- Hadoop cluster establishes the connection to the client using client protocol.
- DataNode talks to NameNode using the DataNode Protocol.

Benefits of Hadoop Clusters

- Robustness
- Data disks failures, heartbeats and re-replication
- Cluster Rebalancing
- Data integrity
- Metadata disk failure
- Snapshot

What are the challenges of a Hadoop Cluster?

- Issue with small files
- High processing overhead
- Only batch processing is supported
- Iterative Processing

References

- Data-Flair
- Data-Bricks