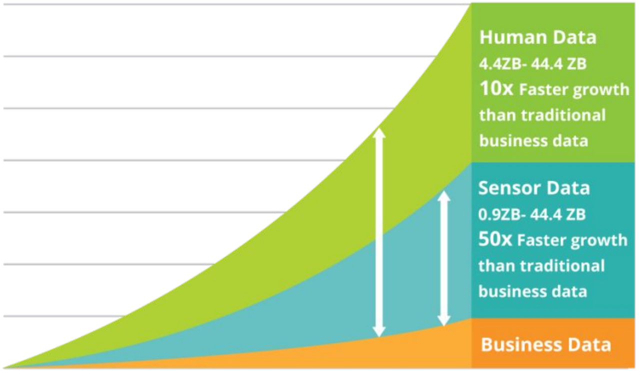
Doing Data Analytics I

15 July 2023

Introduction to Big Data

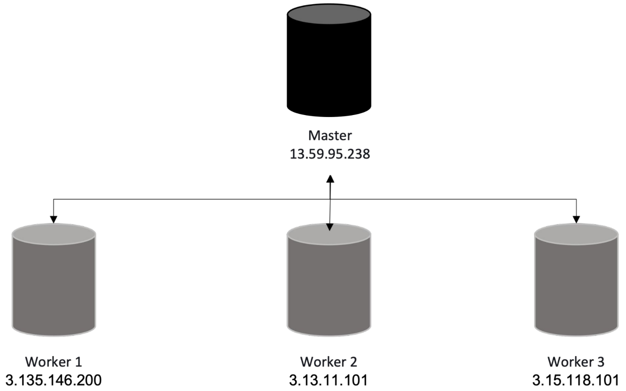
- With the development of e-commerce, businesses could capture not only the sales transactions but also the customer behaviour data.
- Customer behaviour data includes products viewed and added to the cart with product attributes such as brand, price and category etc.
- Since this capture data points at multiple touch points, the volume of data collected can be significantly larger than the sales transactions.
- For example, a multi category e-commerce website could generate sales transactions data of GB level but customer behaviour data of TB or even PB level.
- We'll be using customer behaviour data of an e-commerce website in the month of October in 2019.
- The size of this dataset is ~5GB which can't be handled using our R Studio or Co-Lab IDEs.



Source: Inside big data

Introduction to Distributed Computing

- When you have a data set which can not be processed or analysed using a single machine, you can consider it as big data.
- At this point, you need big data tools and technologies to process and analyse data.
- Today the main technology behind the big data tools is Distributed Computing.
- Distributed computing is the method of making multiple computers work together to solve a common problem.
- It makes a computer network appear as a powerful single computer that provides large-scale resources to deal with complex challenges.



Distributed Big Data Technologies

Data companies such as Google, Yahoo, Facebook and Amazon adopted distributed computing technology to deal with big data.

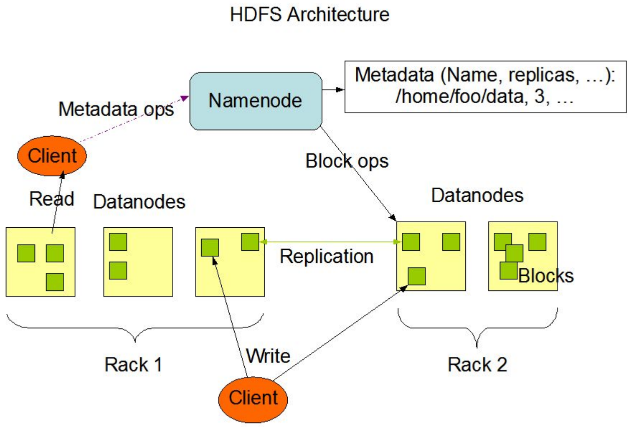
Currently, there are several distributed big data tools used for big data use cases.

This session discusses three distributed big data platforms developed by Apache (www.apache.org)

1. **Hadoop MapReduce**
2. **Hive**
3. **Spark**

Hadoop Distributed File System

- Hadoop Distributed File System
- The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware.



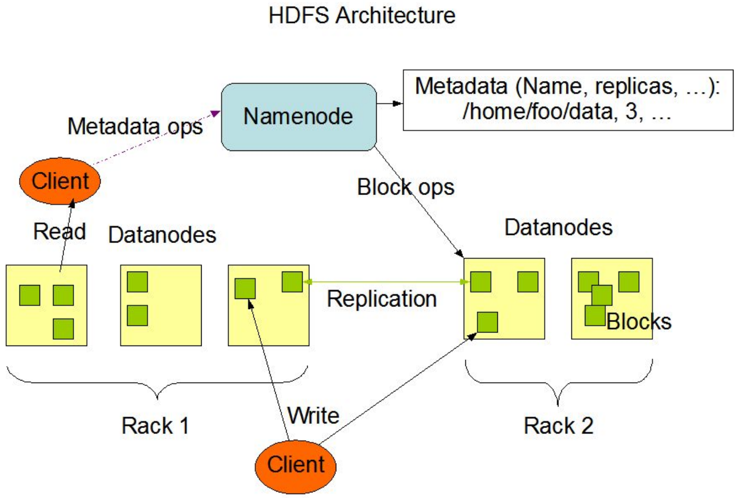
Source: HDFS Architecture Guide

HDFS Benefits

- **Fault tolerance**: HDFS creates a replica of data on other available machines in the cluster if suddenly one machine fails.
- **Failure recovery**: If a node fails in the cluster, HDFS has the ability to detect it and recover quickly and automatically.
- **Support large files**: A typical file in HDFS is gigabytes to terabytes in size. Thus, HDFS is tuned to support large files.
- **High throughput data access**: HDFS is a write-once-read-many access model for files. A file once created, written, and closed need not be changed.
- **Portability across heterogeneous hardware and software**: HDFS is written in JAVA. Usage of the highly portable Java language means that HDFS can be deployed on a wide range of machines.

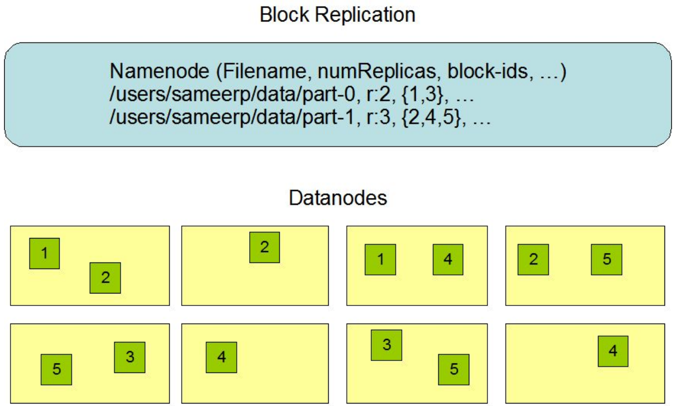
NameNode and DataNodes

- HDFS has a master/worker architecture
- NameNode (master) manages the file system namespace and regulates access to files.
- There are number of DataNodes (workers) which manage storage
- A file is split into fixed size blocks (64MB, 128MB) and replicated among the DataNodes.
- DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode.



Data Replication

- The NameNode makes all decisions regarding replication of blocks.
- It periodically receives a Heartbeat and a Blockreport from each of the DataNodes in the cluster.
- It puts one replica on one node in the local rack, another on a node in a different (remote) rack, and the last on a different node in the same remote rack.
- The NameNode keeps an image of the entire file system namespace and file Blockmap in memory.



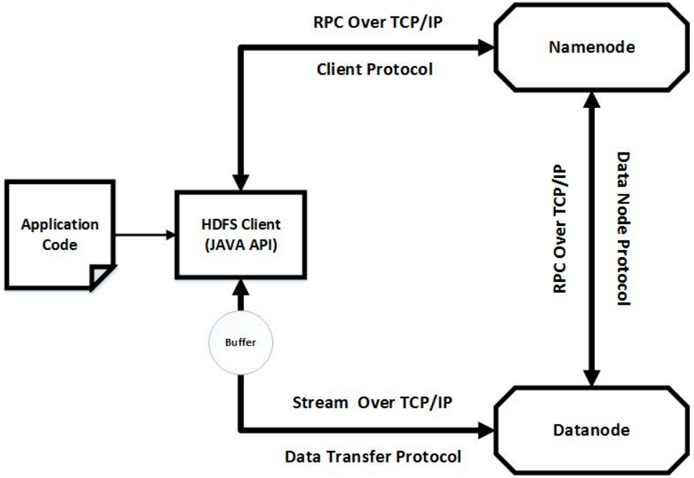
Block Replication

Namenode (Filename, numReplicas, block-ids, …)
/users/sameerp/data/part-0, r:2, {1,3}, …
/users/sameerp/data/part-1, r:3, {2,4,5}, …

Datanodes

The Communication Protocol

- All HDFS communication protocols are layered on top of the TCP/IP protocol.
- Each DataNode sends a Heartbeat message to the NameNode periodically.
- The NameNode marks DataNodes without recent Heartbeats as dead and does not forward any new IO requests to them.
- The NameNode constantly tracks which blocks need to be replicated and initiates replication whenever necessary.
- Remote Procedure Call (RPC) protocol is used over TCP/IP for all the communications.



HDFS Shell Commands

HDFS provides a command line interface called FS shell that lets a user interact with the data in HDFS. Some of the frequently used commands are as follows.

Command	Operation
hadoop fs -ls /temp	List HDFS files and directories inside temp
hadoop fs -mkdir /temp	Create temp directory in HDFS
hadoop fs -rmr /temp	Remove directory temp in HDFS
hadoop fs -copyFromLocal sample.txt /temp/sample.txt	Copy local file sample.txt to HDFS location.
hadoop fs -copyToLocal /temp/sample.txt sample.txt	Copy a HDFS file to local file system

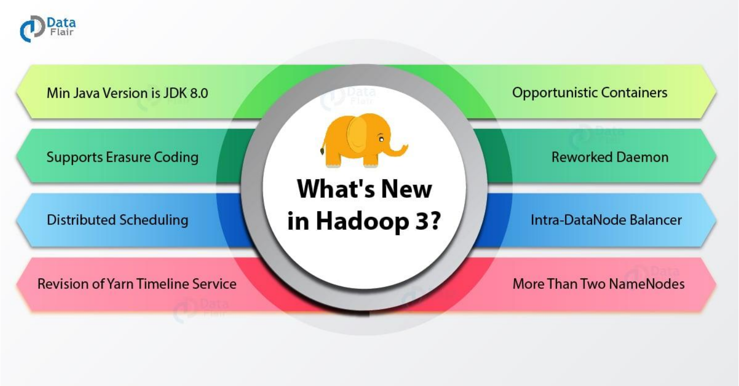
Source: File System Shell Guide

Hadoop 3.x

- **High Availability:** The loss of NameNode can crash the cluster. high-availability was introduced to help recover from NameNode failure. In Hadoop 3.x we can have two passive NameNodes along with the active node, as well as five JournalNodes.

- **Intra-DataNode Balancer:** Hadoop 3.x introduces intra-DataNode balancer to balance the physical disk inside each DataNode to reduce the skew of the data.

- **Erasure Coding (EC):** Typical HDFS installation has a replication factor of 3 which requires large storage capacity in the cluster. EC is a method of data protection in which data is broken into fragments, expanded, encoded with redundant data pieces and stored across a set of different locations or storage. This can brings down the replication factor from 3 to about 1.4.



Source:Data Flair

Questions & Answers

Use your Google Classroom stream to post any questions or start discussions.
[*https://classroom.google.com/c/NDk3MzI3NzgxNDM5?cjc=ltzuypk*](https://classroom.google.com/c/NDk3MzI3NzgxNDM5?cjc=ltzuypk)