Master Executive di II Livello

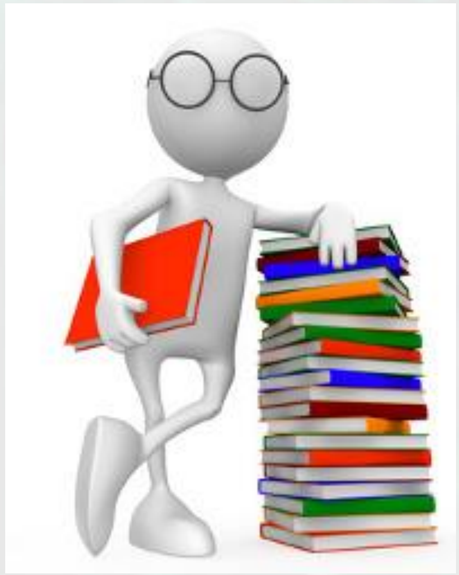
BIG DATA ANALYSIS AND BUSINESS INTELLIGENCE

Vamsi Krishna Varma Gunturi
Data science intern at ISTAT
vamsivarmagunturi@gmail.com

HDFS wrap-up

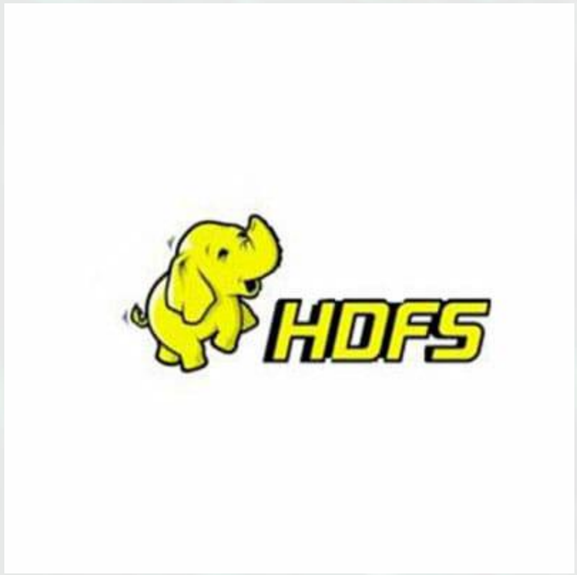
fondazione INUIT TORVERGATA

Topics



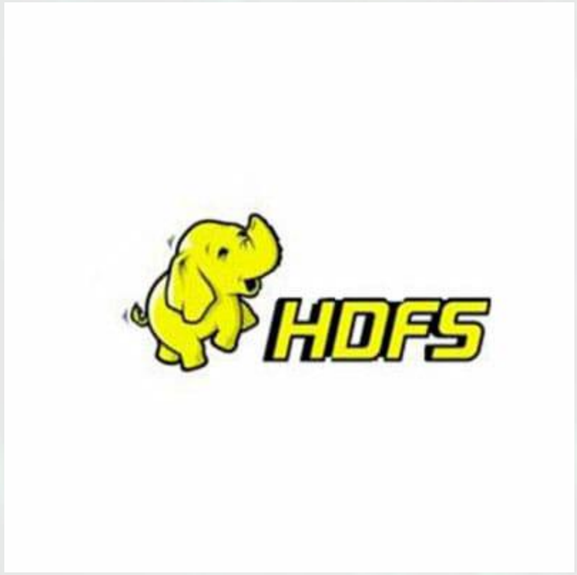
- HDFS Recap
- Overview of HDFS
- Configuration files
- Listing files in HDFS
- Creating directories in HDFS
- File and directory permissions overview
- Previewing text files in HDFS
- Rack awareness
- Overview of block size and replication factor
- Getting metadata of files using “hdfs fsck”
- Hadoop cluster architecture and block placement
- Basic administration commands

HDFS recap



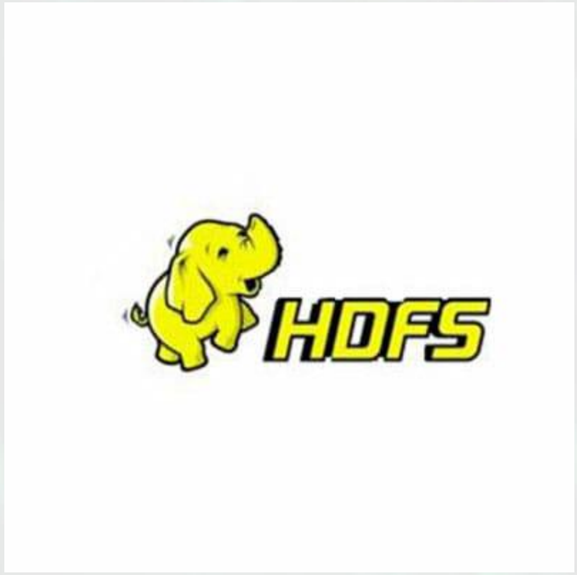
- HDFS is for storing and managing data across cluster of computers.
- HDFS daemons:
  - Name node
  - Data node
- Data replication
- Reading and writing files in HDFS

Overview of HDFS



- Hadoop distributed file system which handles big files by breaking them in to small blocks stored across several commodity computers
- HDFS allows your big data to be stored across an entire cluster in a distributed manner in a reliable manner and allows your applications to analyse that data to access that data quickly and reliably.
- Each block is about 128 MB’s large so it can accommodate pretty large files
- In order to handle failure. it will actually store more than one copy of each block and so that way if one of these individual computers goes down HDFS can deal with that and actually start retrieving information from a different computer that had a backup copy of that block.

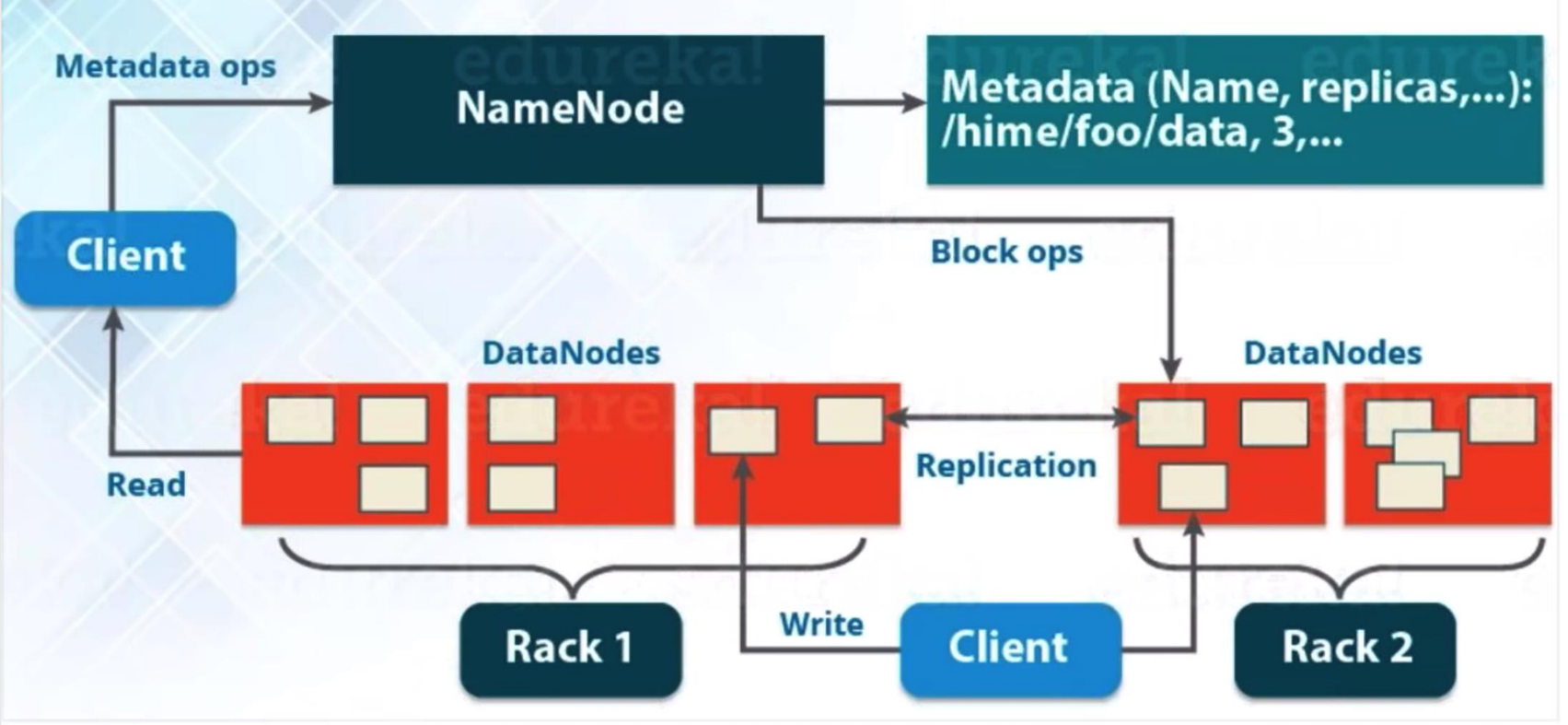
Configuration files



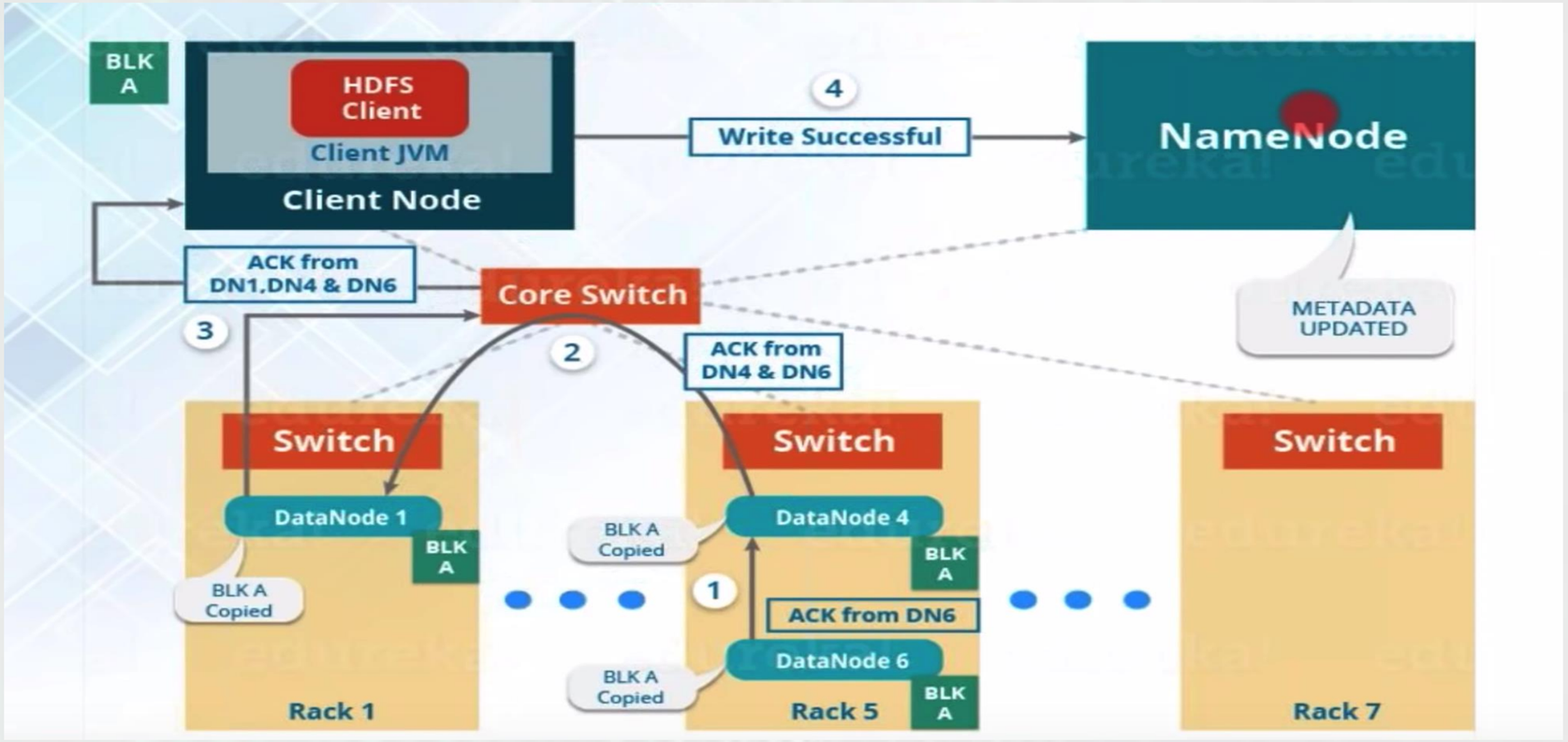
Primarily there are 4 configuration files in Hadoop (inside /etc/Hadoop/conf folder) namely,

- **core-site.xml** – contains the locations of Hadoop processes such as JobHistoryServer, Resource manager, Node Manager, Name node and DataNode
- **hdfs-site.xml** - Configuration options for HDFS such as block size and replication factor etc..,
- **yarn-site.xml** - Configuration options for resource negotiator
- **mapred-site.xml** - map reduce configuration file

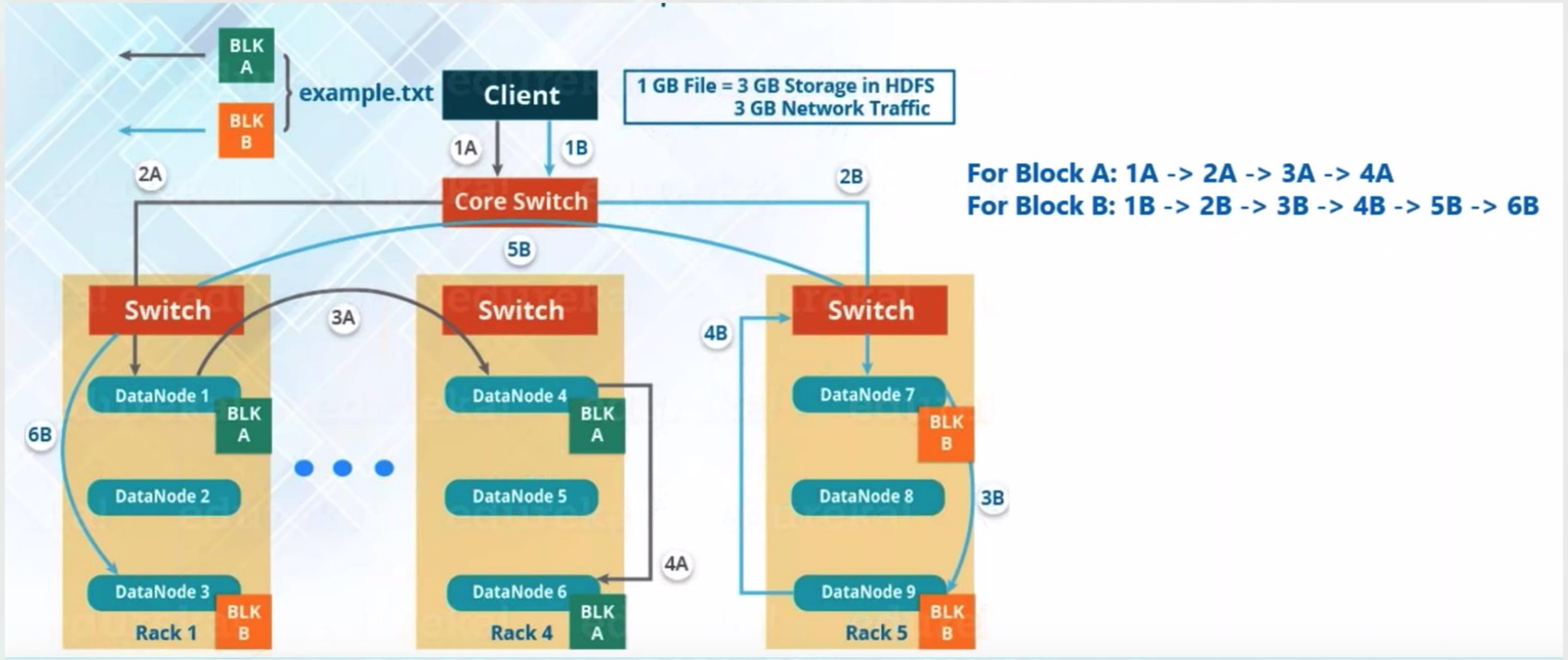
HDFS Architecture



HDFS write mechanism

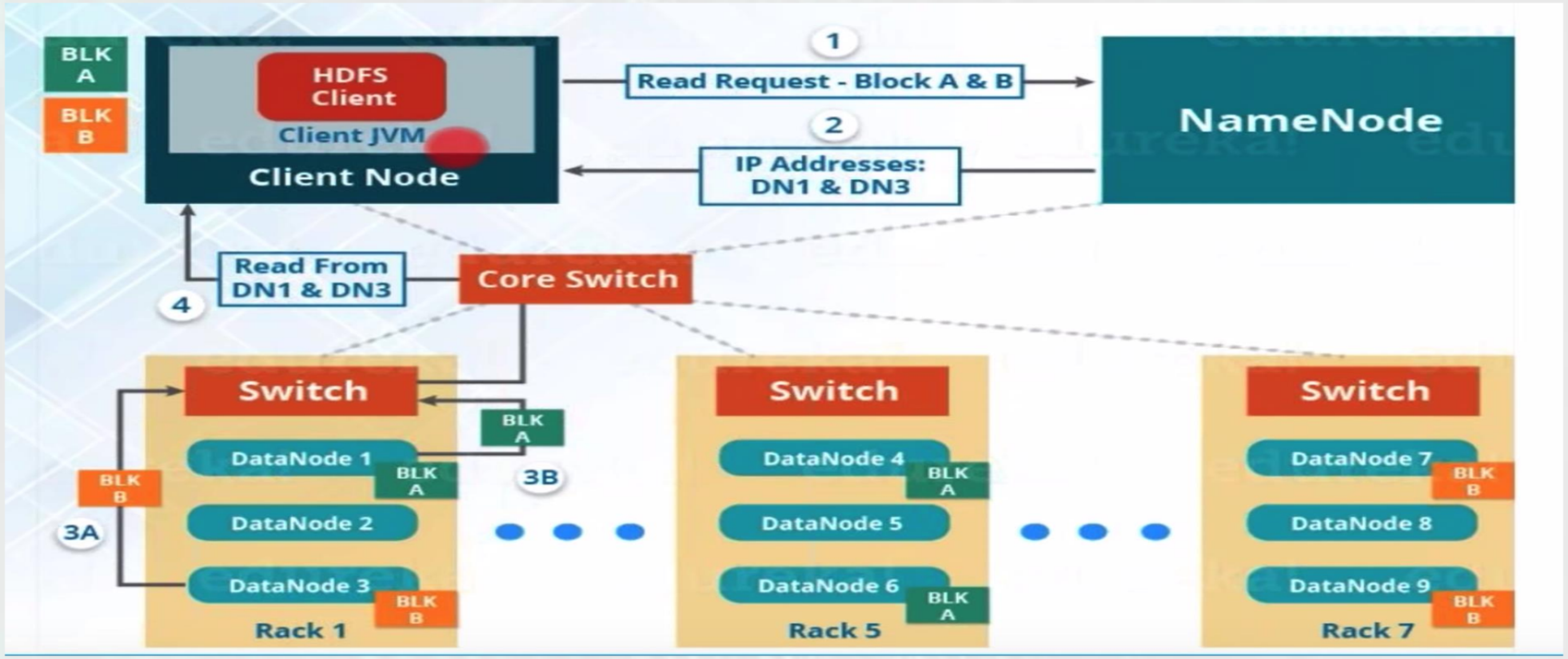


HDFS multi-write pipeline

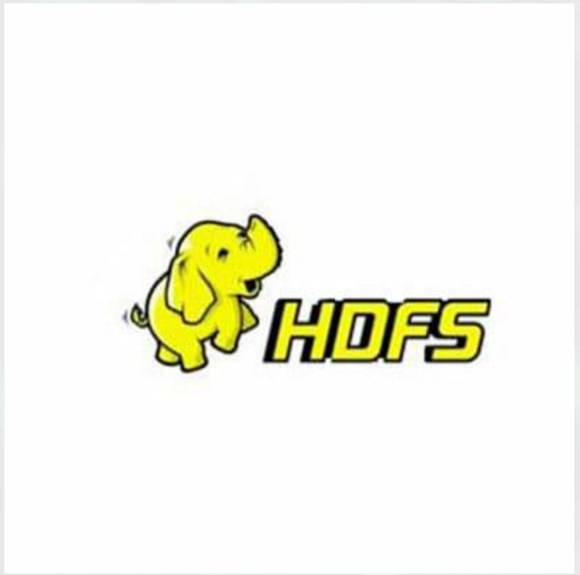


Vamsi Krishna Varma Gunturi

HDFS read pipeline



Listing files in HDFS

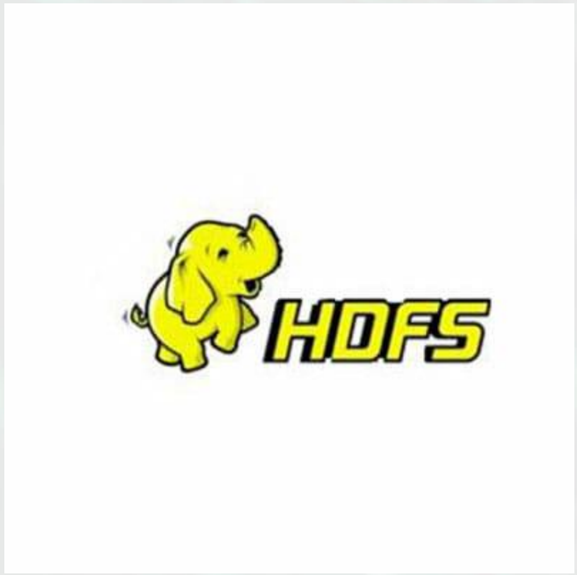


- To access HDFS manual for a particular command syntax we can use –help option
  Usage: hadoop fs –help ls

- To get the file sizes in readable format (in MB's, GB's instead of in bytes) we use –S –h option in ls command
  Usage: hadoop fs –ls –S –h /public/hdfs_dataset

- To recursively search for directories with a particular name inside a directory we use –R with grep command
  Usage: hadoop fs –ls –R /public | grep order (to find folders with sub-string order)

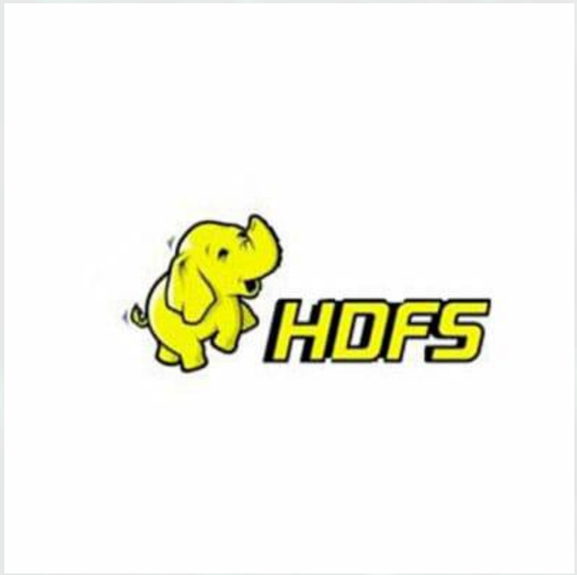
Creating directories



Commands for creating and removing directories –

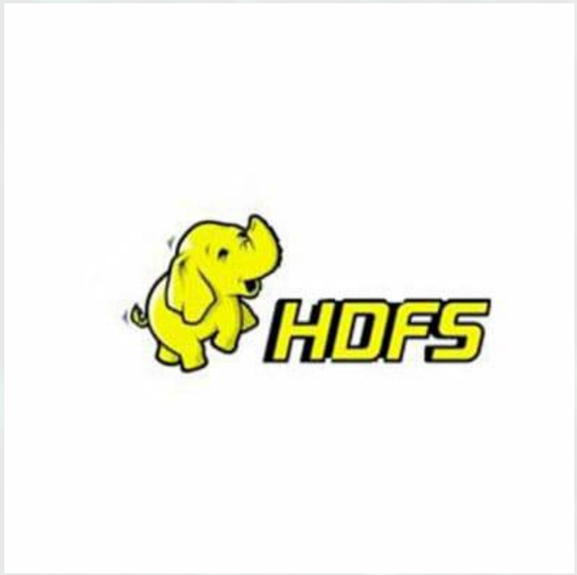
- hadoop fs –mkdir /user/hadoop/hdfs_new - creates a new directory with name hdfs_new in the metioned path
- hadoop fs –rm -R /user/hadoop/hdfs_new - Recursively deletes the hdfs_new folder and its contents
- hadoop fs –rmdir –-ignore-fail-on-non-empty /user/hadoop/hdfs_new – rmdir is used to remove empty directories. So if the directory is non-empty it will throw an error and terminate

File permissions



- Similar to LINUX/UNIX file/directory permissions
- The administrators are responsible for changing the permissions or ownership of files/directories
- Has 3 categories – owner, group and others

Previewing files



We can preview text files on HDFS using cat and tail commands. So we need not download a large file in HDFS to view its contents and structure. Below is the usage of these commands

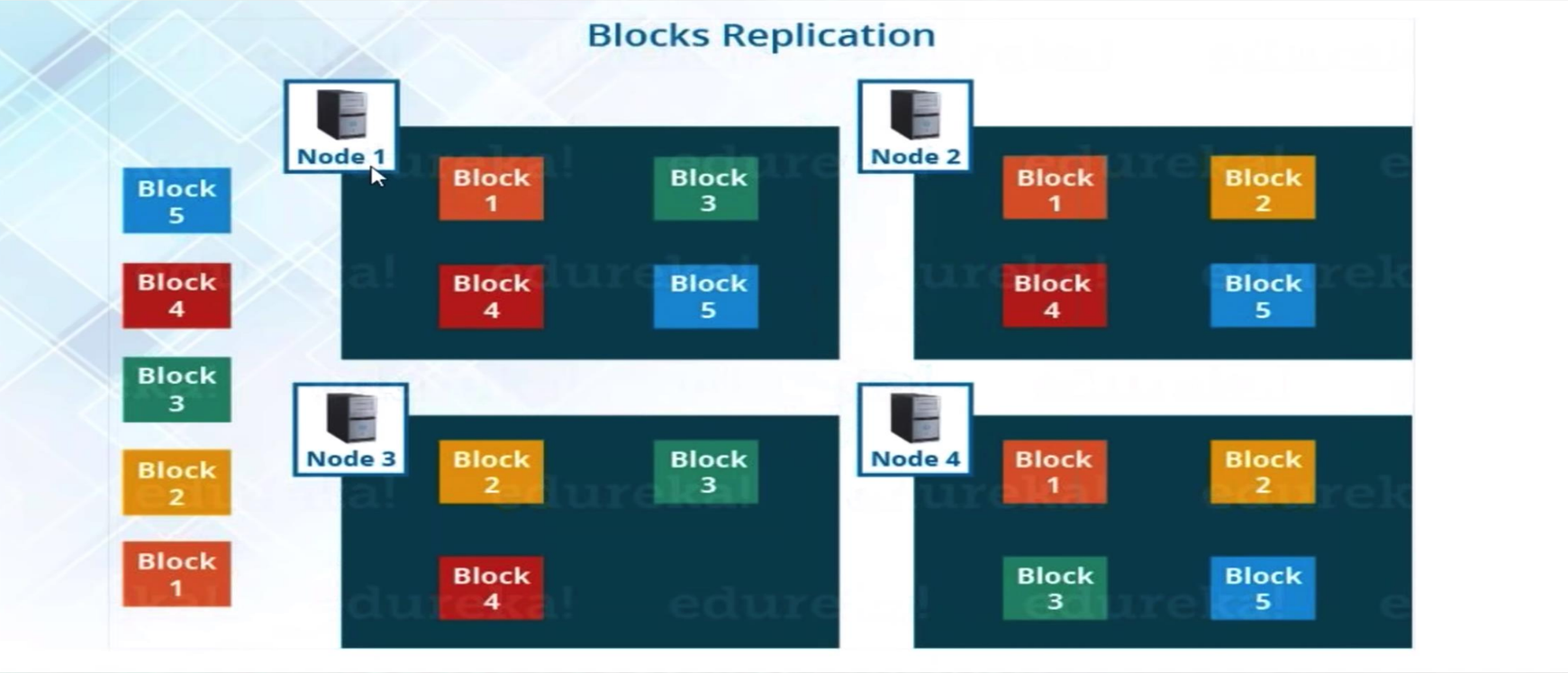
- **cat**: hadoop fs -help cat – displays all the file contents. If the file is large this is not a suggested way. Instead use tail command. We can apply patterns and regular expressions using cat command

Usage: Hadoop fs –cat /public/txt_files/*|more – to view first few lines of a large files.

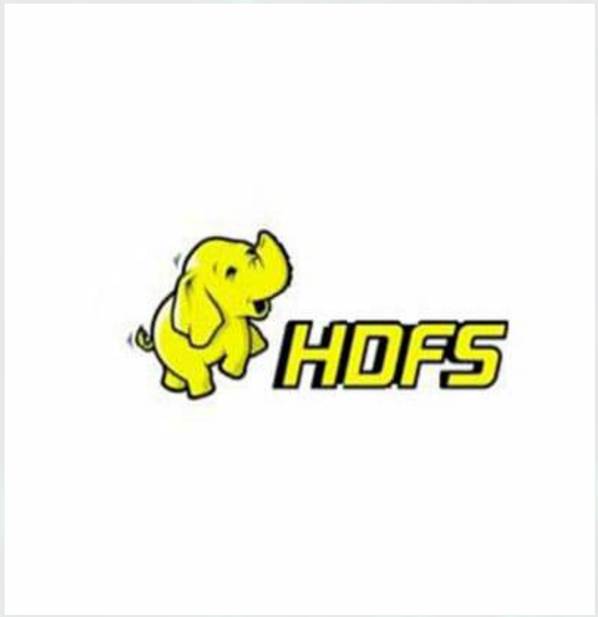
- **tail**: hadoop fs –help tail - displays the last 1 KB of the file. We cannot apply patterns using tail command.

Usage: hadoop fs –tail /public/txt_files/large_file.txt

Block size and replication



Getting metadata



- To get meta data we can use **fsck** command which stands for file system check. So we can get information on what type of files and the location of different blocks of a particular HDFS directory using this command so that we know where our blocks are stored and what is the health of our directory in general and other useful metadata using this command fsck. All this metadata is stored in in-memory component of the data node

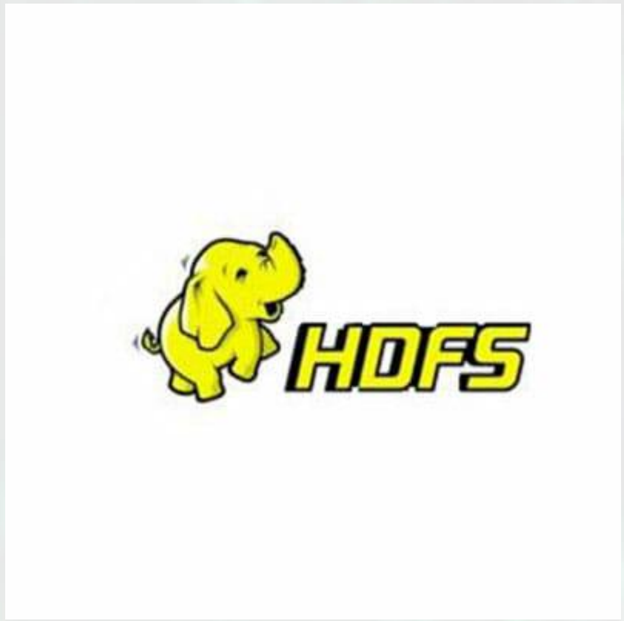
Usage:

> hdfs fsck /user/data/hdfs_data –files – Gives the details wrto files in this location but the details of blocks used to store those files.

> hdfs fsck /user/data/hdfs_data –files -blocks – Gives the metadata related to blocks. Every block contains a unique block id and block index across all the files in the HDFS.

> hdfs fsck /user/data/hdfs_data –files -blocks -locations – To get the locations of the blocks i.e.., IP addresses of the data nodes

Useful HDFS Commands

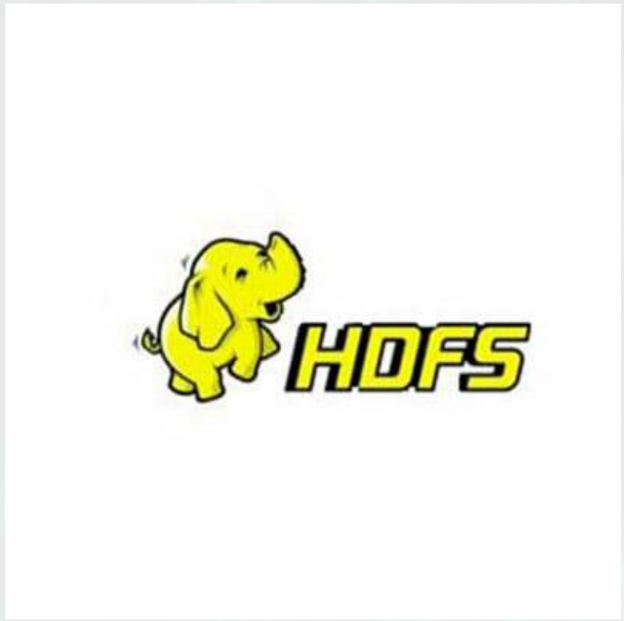


- **> hdfs dfs - Check all the HDFS commands. This is similar to Hadoop fs command.**
- **> sbin/start-all.sh** : To start all the Hadoop services
- **> jps** : To check list of active services and their port numbers
- > **hdfs dfs -ls /** : Prints all the directories present in HDFS
- > **dfs -mkdir <folder name>** : Create a new directory in HDFS
- > **hdfs dfs -touchz <file_path>** : Creates an empty file
- > **hdfs dfs -copyFromLocal <local file path> <dest(present on hdfs)>** : To copy the files/folders from local file system to HDFS store. We can also use –put instead of –copyFromLocal command.
- > **hdfs dfs -cat <path>** : Prints the file contents

Useful HDFS Commands

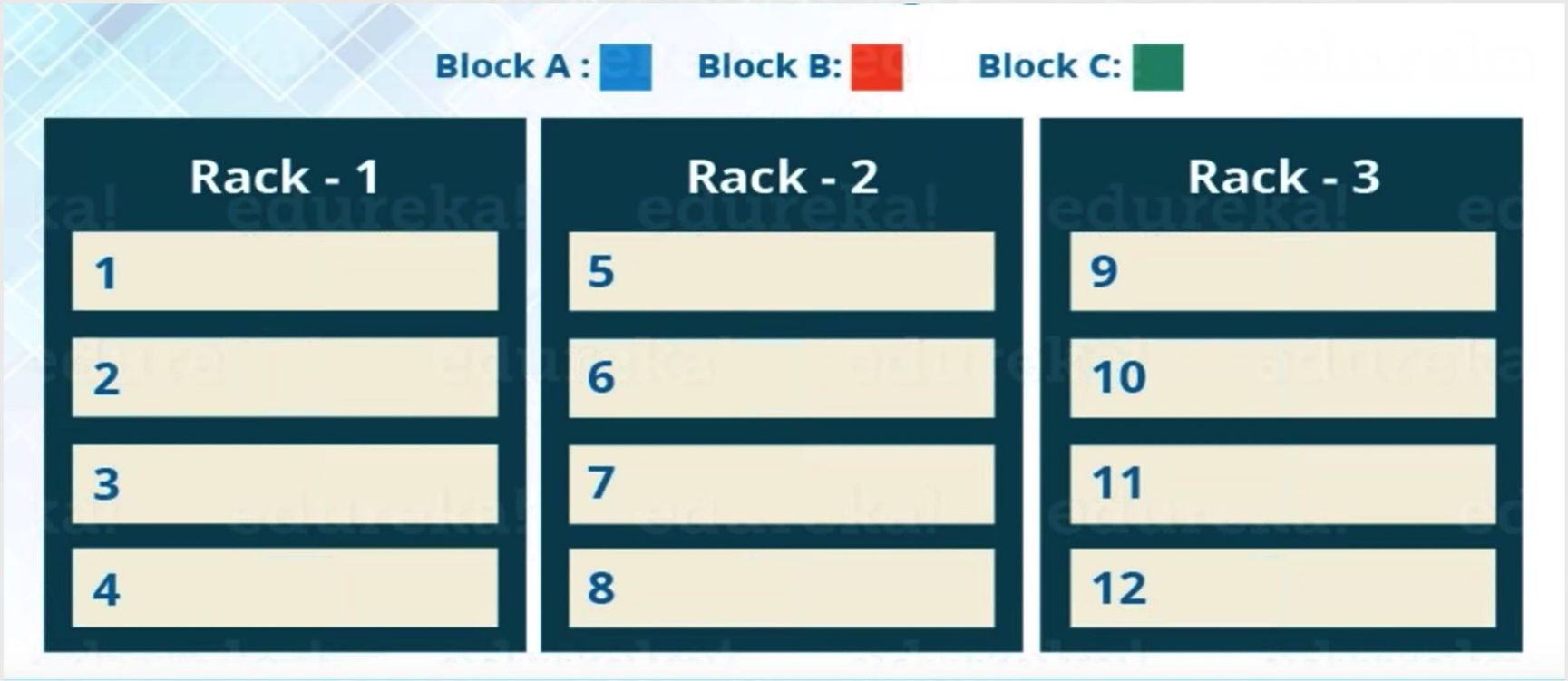
- **> hdfs dfs -copyToLocal <<srcfile(on hdfs)> <local file dest>** : To copy files/folders from HDFS store to the local file system. We can also use –get instead of -copyToLocal
- **> hdfs dfs -cp <src(on hdfs)> <dest(on hdfs)>** : Copy files with in HDFS
- > **hdfs dfs -rmr <filename/directoryName>** : Deletes a file from HDFS recursively
- > **hdfs dfs -du <dirName>** : Gives the size of each file in the directory
- > **hdfs dfs -setrep -R -w 6 <filename/directoryName>** : Change the replication factor of file/folder inside HDFS. By default it is 3.

Useful HDFS Commands

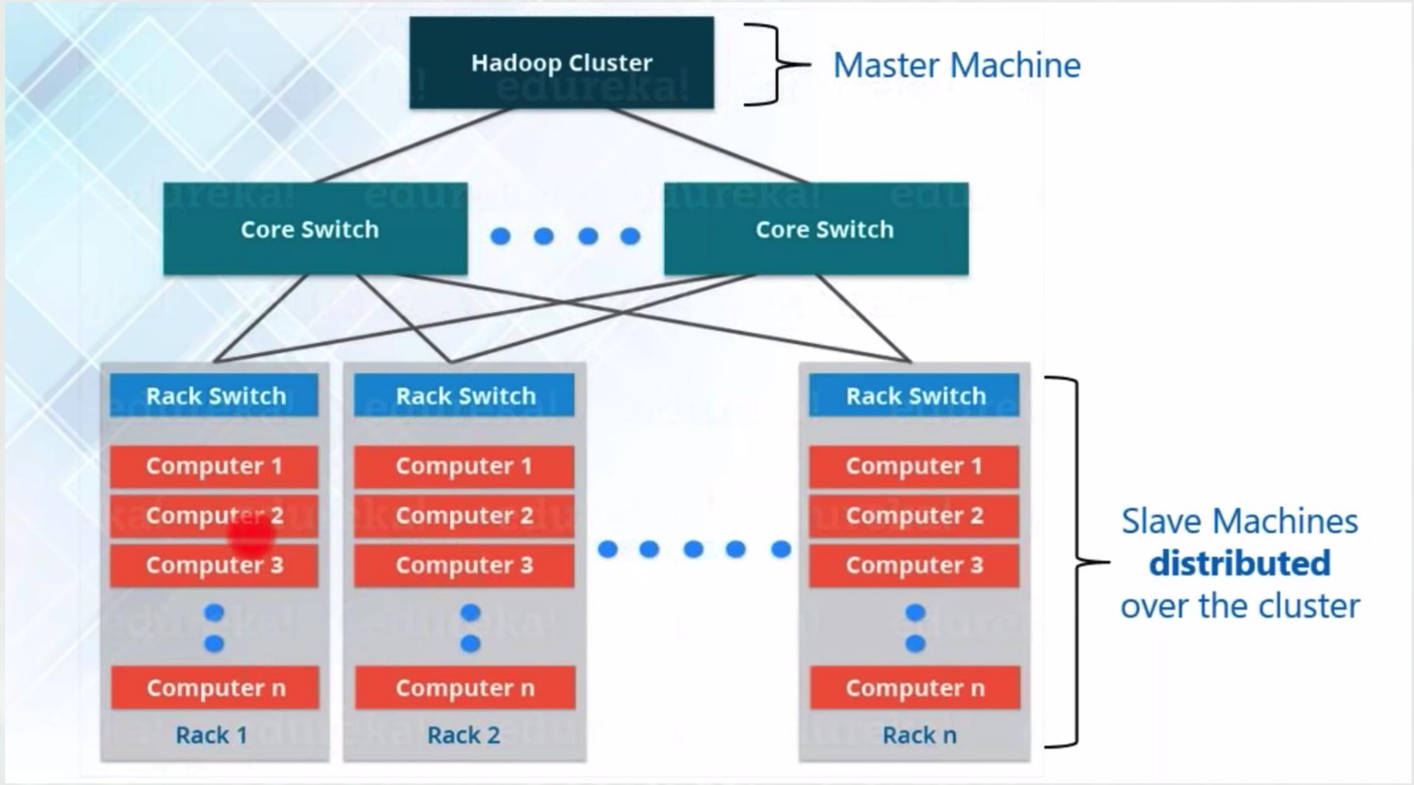


- **> hdfs dfs –df –h /user/data/hdfs_data –** displays the size of hdfs_data folder in human readable format. It gives information such as Filesystem, Size, Used, Available, Use%.
- **> hdfs dfs –du –h /user/data/hdfs_data –** displays the disk usage of a particular folder on HDFS in human readable format. So if the directory has many sub-directories then the size is computed at each and every directory level. This command gives the original data size and it doesn't factor the replication factor.
- **> hdfs dfs –du -s –h /user/data/hdfs_data –** summarize the disk usage of entire directory with out the sub-directories information.
- **> hdfs dfs –Ddfs.replication=3 –put /user/data/hdfs_data/large_file.txt /user/hdfs_main –** to change the replication factor to 3 on run time.

Hadoop Rack awareness



Hadoop Cluster Architecture



Master Executive di II Livello

BIG DATA ANALYSIS AND BUSINESS INTELLIGENCE

Vamsi Krishna Varma Gunturi
Data science intern at ISTAT
vamsivarmagunturi@gmail.com

Grazie

fondazione INUIT TORVERGATA