Big Data Analytics

VIT Vidyalankar Institute of Technology
Accredited A+ by NAAC

Academic Year 2022-23

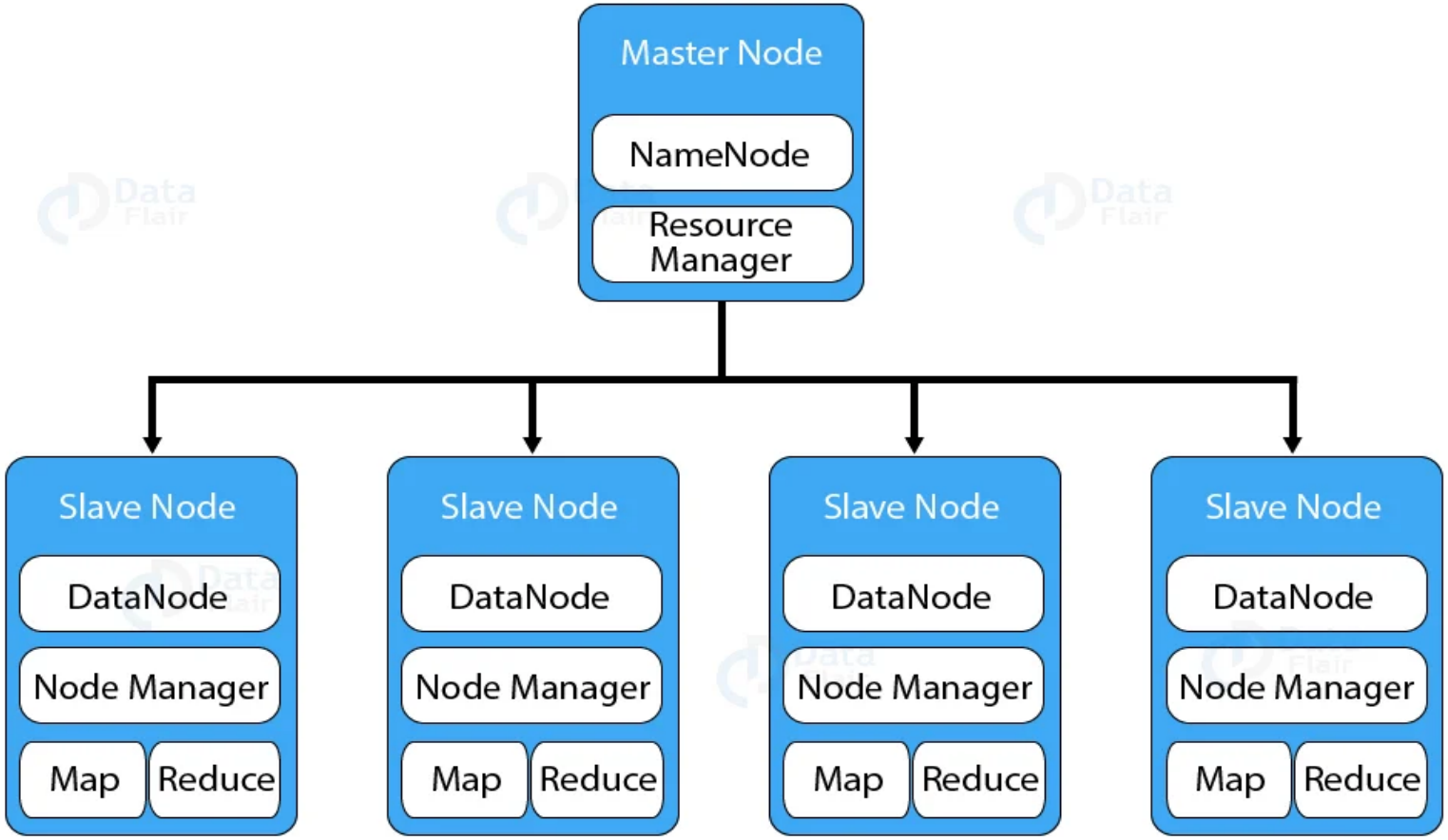
Index

1. HDFS
2. HDFS Architecture
3. HDFS Read / Write Mechanism

Subject Teacher: **Prof. Prakash Parmar, Assistant Professor, CMPN**

Hadoop Architecture

Hadoop have a master-slave architecture. In this, we have one master node and multiple slave nodes.



HDFS

- HDFS is a file system designed for storing very large files with streaming data access pattens, running on clusters of commodity hardware.

- Hadoop is designed to run on clusters of commodity hardware, where the chance of node failure is high. HDFS is designed to carry on working without a noticeable interruption in the face of such a failure.

Component of HDFS

Major component of HDFS

NameNode

Master Node contains metadata in ram and disk

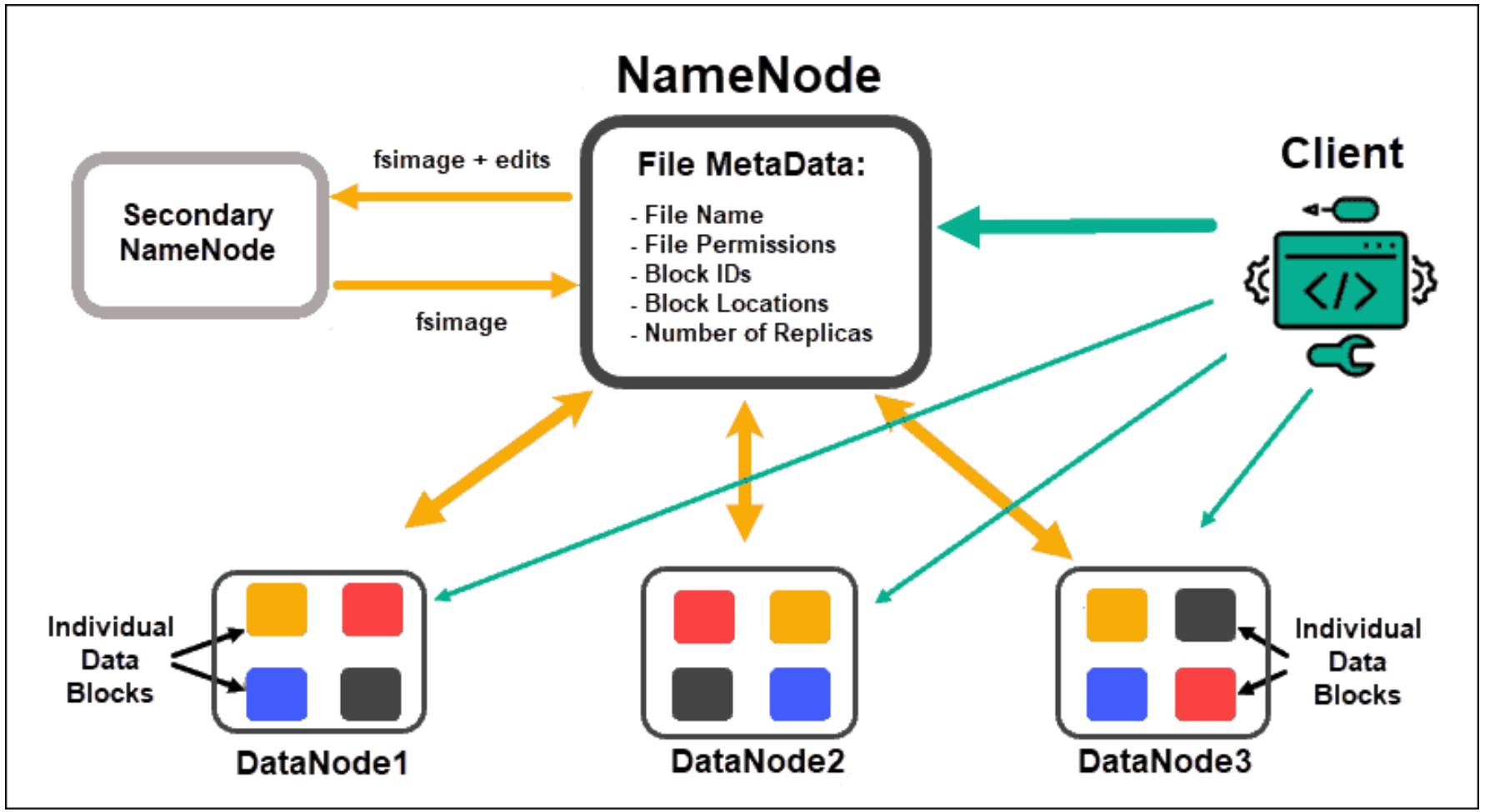
Secondary NameNode

Has a copy of Name node's metadata in disk

DataNode

Contains the actual data in the form of blocks

HDFS Architecture



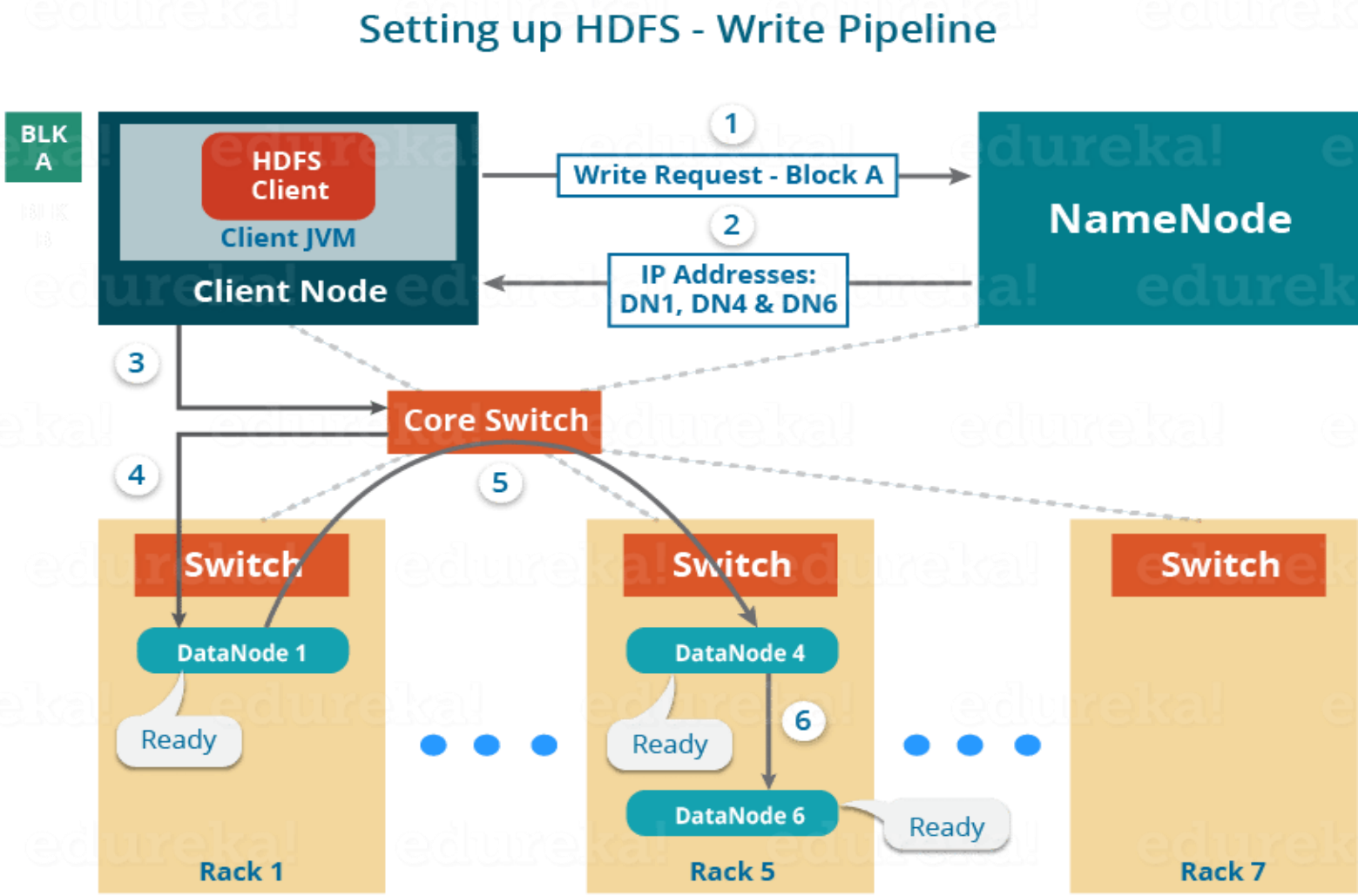
HDFS Data Write Mechanism

- HDFS follow Write once Read many models, so we cannot edit files already stored in HDFS, but we can append data by reopening the file.
- Suppose a situation where an HDFS client, wants to write a file named “example.txt” of size 248 MB.
- Assume that the system block size is configured for 128 MB (default). So, the client will be dividing the file “example.txt” into 2 blocks – one of 128 MB (Block A) and the other of 120 MB (block B).

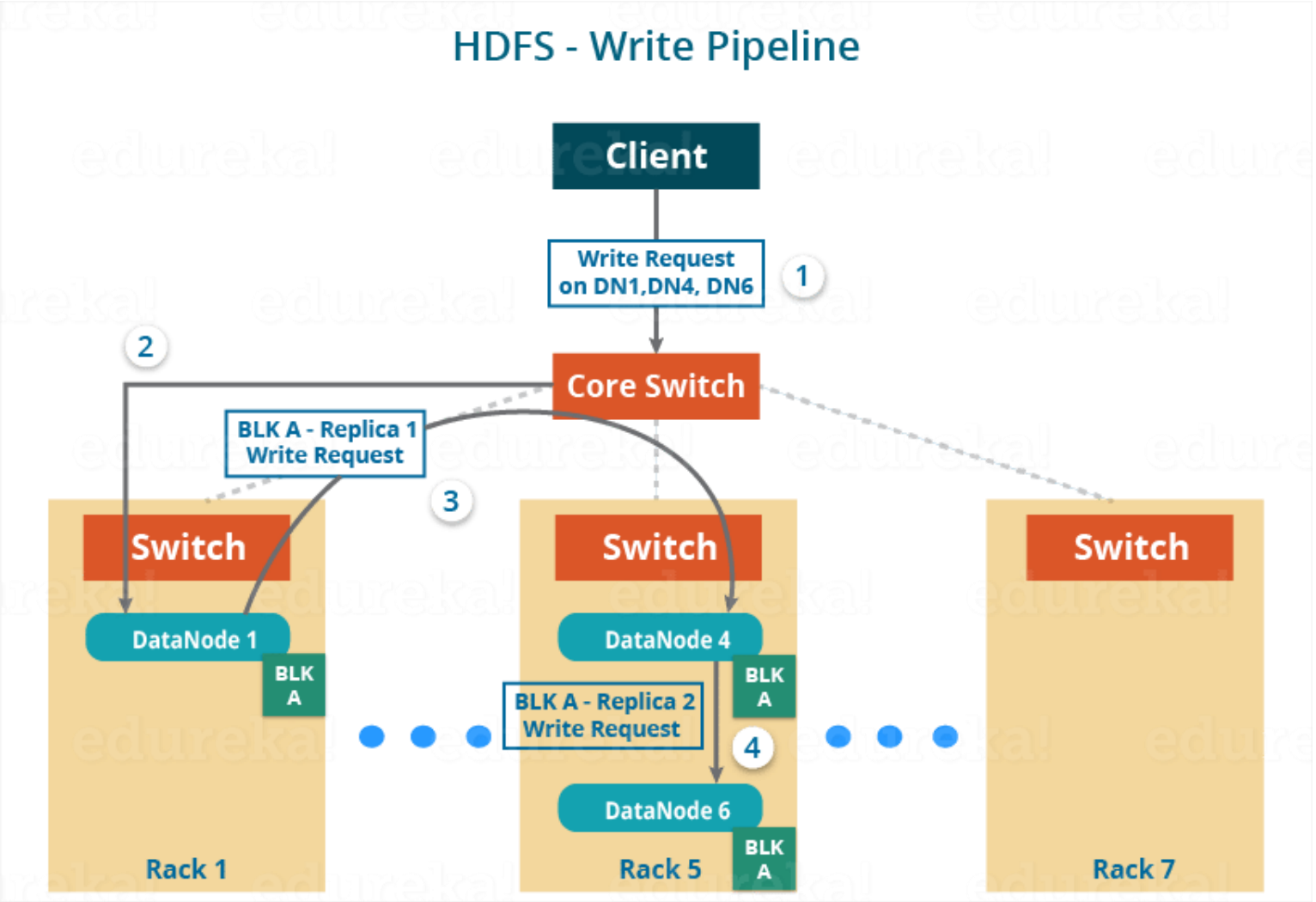
HDFS Data Write Mechanism

1. Set up of Pipeline
2. Data streaming and replication
3. Shutdown of Pipeline (Acknowledgement stage)

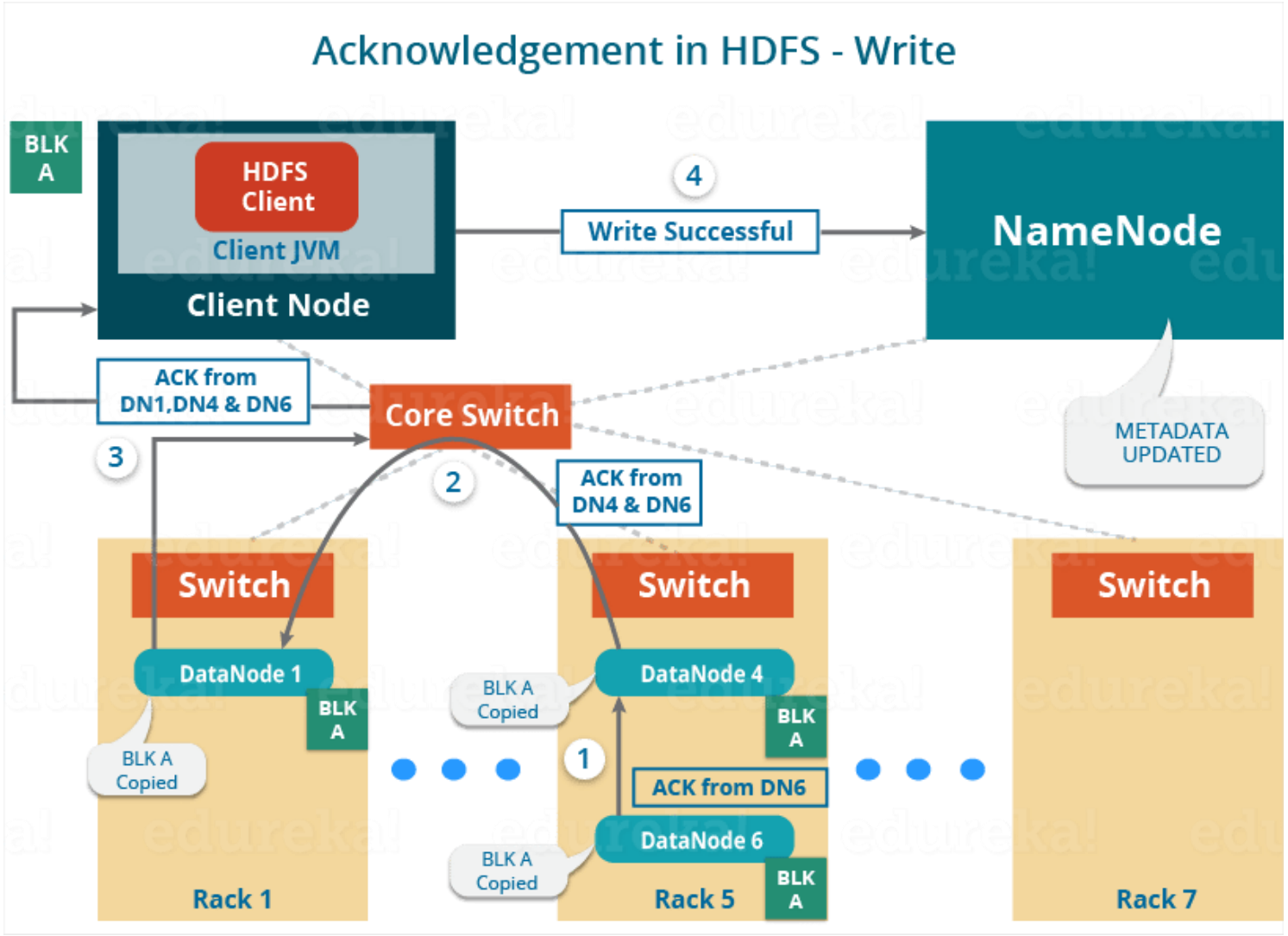
HDFS Write: Set up of Pipeline



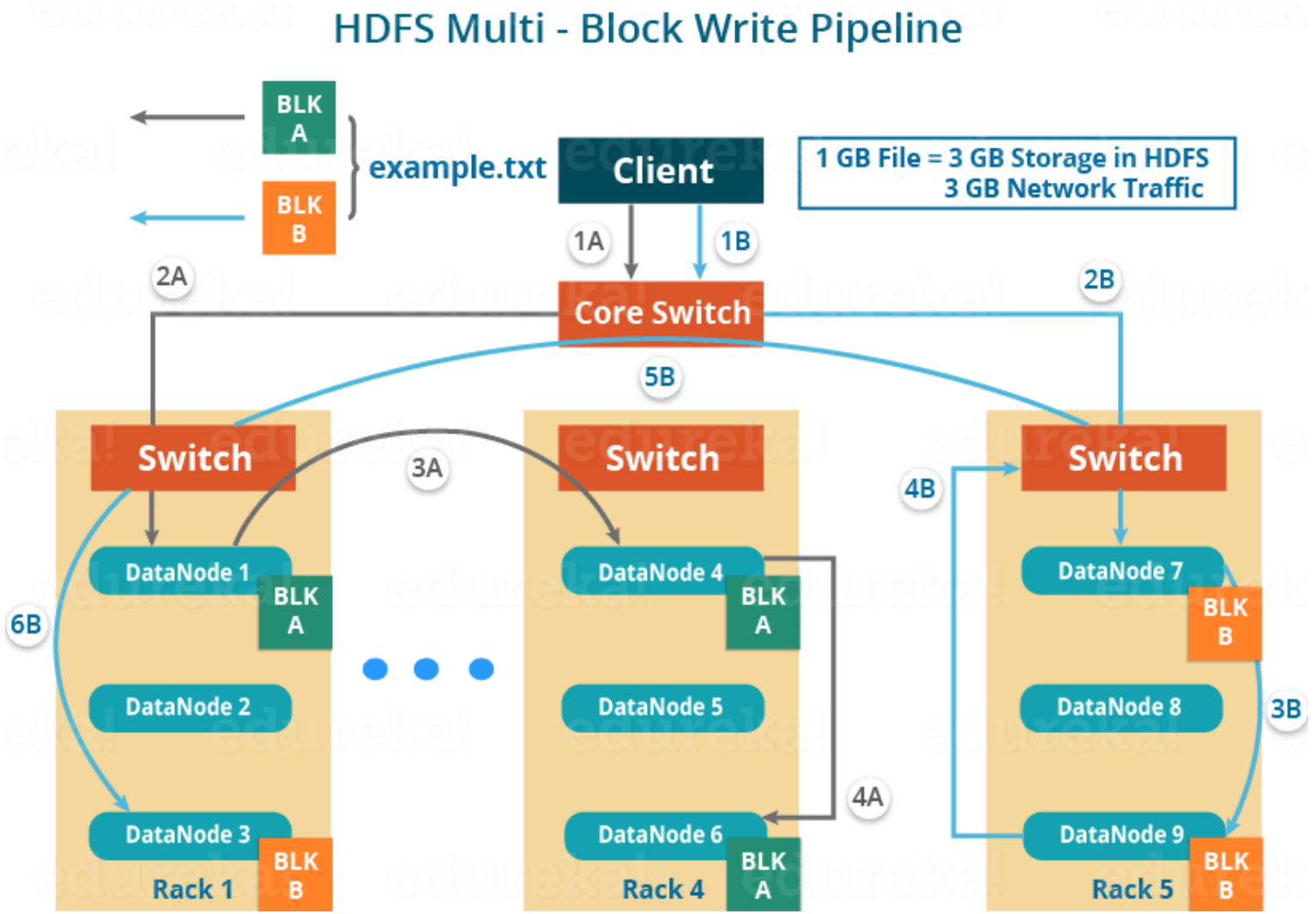
HDFS Write: Data streaming and replication



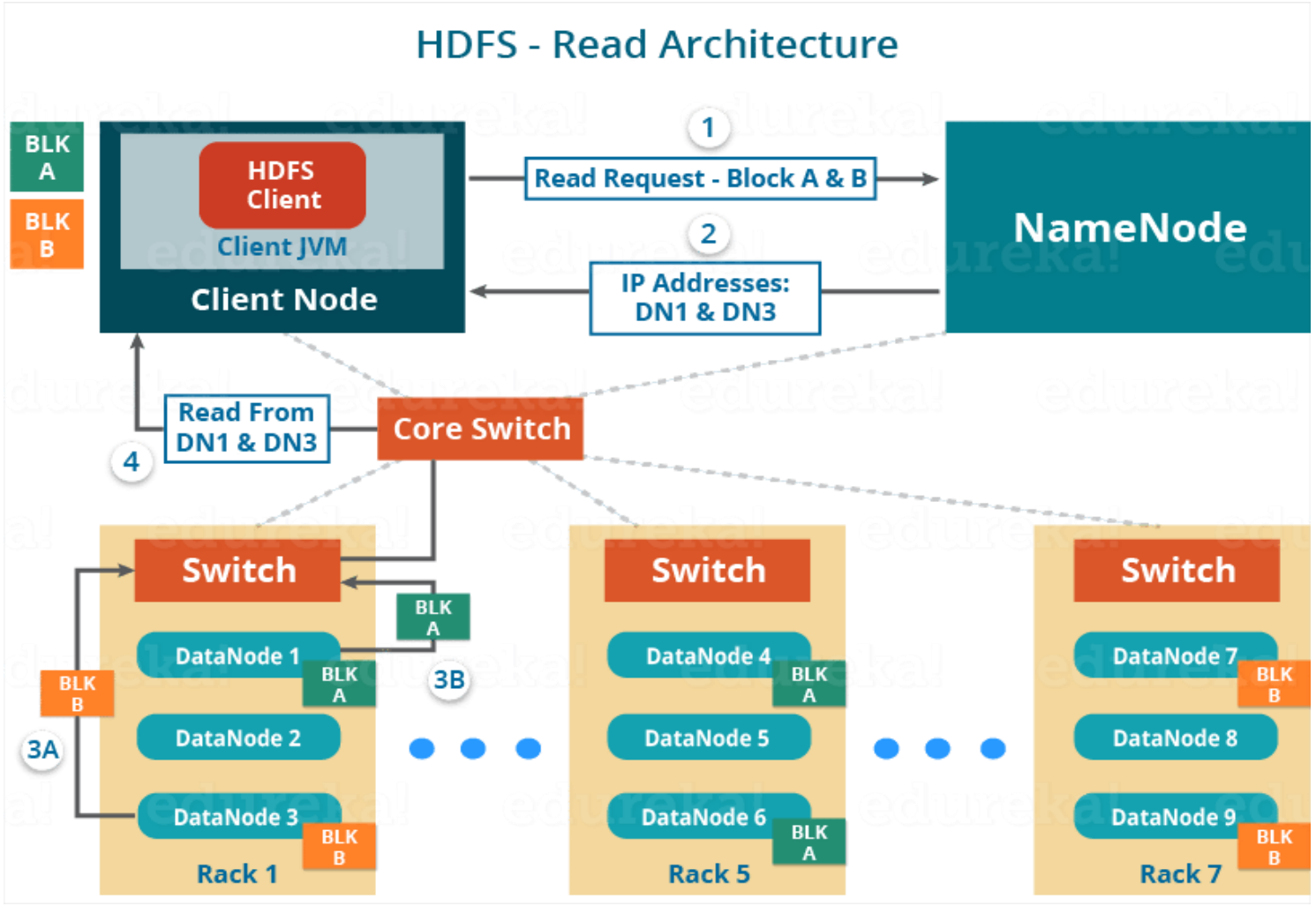
HDFS Write: Acknowledgement



HDFS Data Write Architecture



HDFS Data Read Architecture



Features of HDFS

- **HDFS has high fault-tolerance**
- HDFS may consist of thousands of server machines. Each machine stores a part of the file system data. HDFS detects faults that can occur on any of the machines and recovers it quickly and automatically.
- **HDFS has high throughput**
- HDFS is designed to store and scan millions of rows of data and to count or add some subsets of the data. The time required in this process is dependent on the complexities involved.
- It has been designed to support large datasets in batch-style jobs. However, the emphasis is on high throughput of data access rather than low latency.
- **HDFS is economical**
- HDFS is designed in such a way that it can be built on commodity hardware and heterogeneous platforms, which is low-priced and easily available.

Learn Fundamentals & Enjoy Engineering

Prof. Prakash Parmar
Assistant Professor
Computer Engineering Department
Vidyalankar Classes CSE GATE Faculty