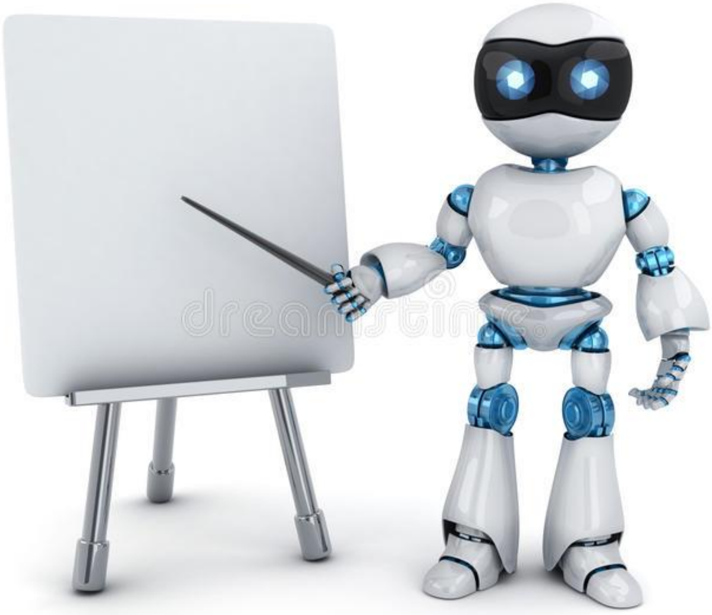


Hadoop

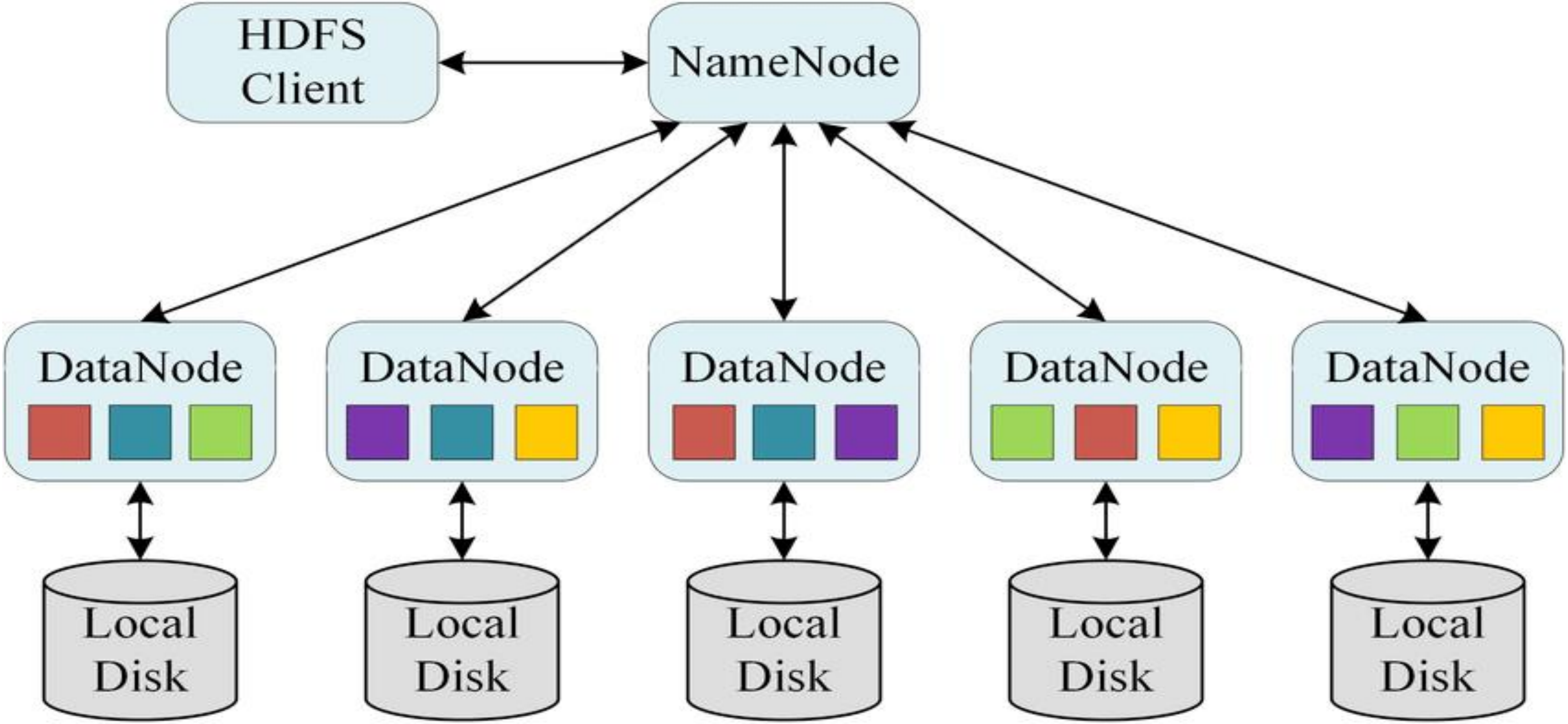
HDFS

What is HDFS?

HDFS stands for Hadoop Distributed File System. It is a distributed file system that is designed to store large amounts of data across a cluster of commodity hardware. HDFS is a core component of the Apache Hadoop ecosystem and is used to store and manage big data.

HDFS is designed to work with large files, typically in the range of gigabytes to terabytes. It is also optimized for streaming reads, which makes it well suited for big data analytics and batch processing workloads.

What is HDFS?


HDFS Client
NameNode
DataNode
DataNode
DataNode
DataNode
DataNode
Local Disk
Local Disk
Local Disk
Local Disk
Local Disk

What is HDFS?

.HDFS supports standard file system operations such as create, read, write, delete, and rename. It also provides a web-based user interface, called the NameNode web UI, that allows users to view the file system and perform basic operations.

HDFS is a popular storage platform in the big data ecosystem and can be used in conjunction with other big data tools such as Hadoop MapReduce, Apache Spark, and Apache Hive.

Distributed Processing with MapReduce

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. The MapReduce model was inspired by the map and reduce functions commonly used in functional programming,

The MapReduce model consists of two main tasks: the map task and the reduce task. The map task takes a set of input data and converts it into a set of key-value pairs, also known as intermediate data. The reduce task then takes the intermediate data and combines it into a smaller set of key-value pairs, which are the final output.

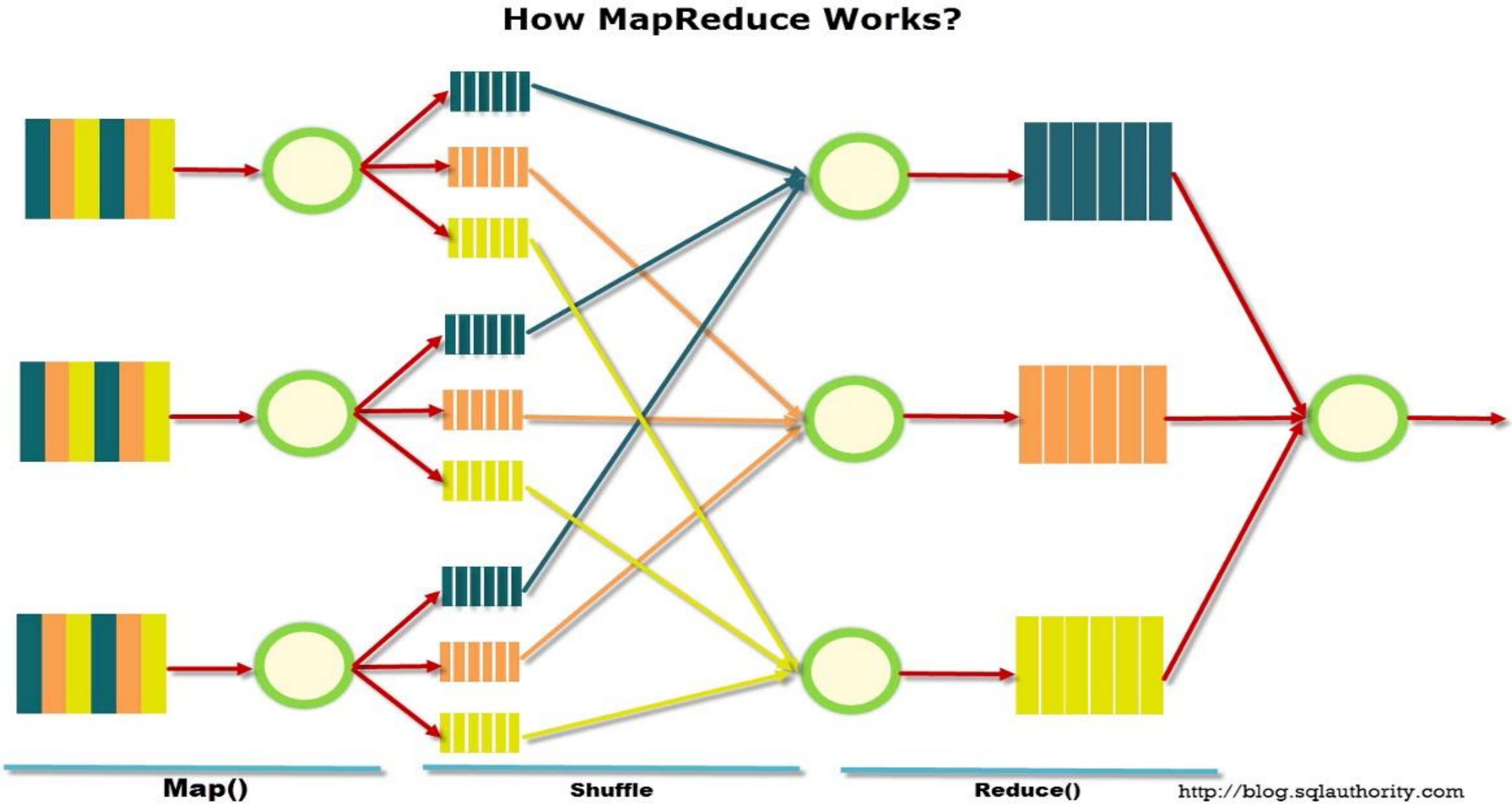
The Algorithm

- Generally MapReduce paradigm is based on sending the computer to where the data resides!
- MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.
  - Map stage − The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
  - Reduce stage − This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

The Algorithm

- During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.
- The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.
- Most of the computing takes place on nodes with data on local disks that reduces the network traffic.
- After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.

The Algorithm



Key Terms

Output Format:Types of output format in MapReduce are

1. TextOutputFormat
2. SequenceFileOutputFormat
3. SequenceFileAsBinaryOutputFormat
4. MapFileOutputFormat
5. MultipleOutputs
6. LazyOutputFormat
7. DBOutputFormat

Key Terms

Partitioners:Partitioners are used in distributed systems such as MapReduce, Spark, and Kafka to ensure data is distributed in a way that maximizes parallelism and load balancing.

Combiners:-Combiners, also known as intermediate reducers, are a technique used in distributed computing to reduce the amount of data that needs to be shuffled across the network during the reduce phase of a MapReduce job. They are applied to the output of the map task, before the data is shuffled and sorted, and they work by locally reducing the data on each individual node.

- A combiner does not have a predefined interface and it must implement the Reducer interface’s reduce() method.
- A combiner operates on each map output key. It must have the same output key-value types as the Reducer class.
- A combiner can produce summary information from a large dataset because it replaces the original Map output.

Key Terms

Shuffle:It refers to the process of redistributing data across the nodes of a cluster. The shuffle phase of a MapReduce job takes place after the map task has produced intermediate key-value pairs, and it is responsible for organizing and redistributing the data so that it can be processed by the reduce task.