Big Data Analysis

Lecture 5

2019/11/28

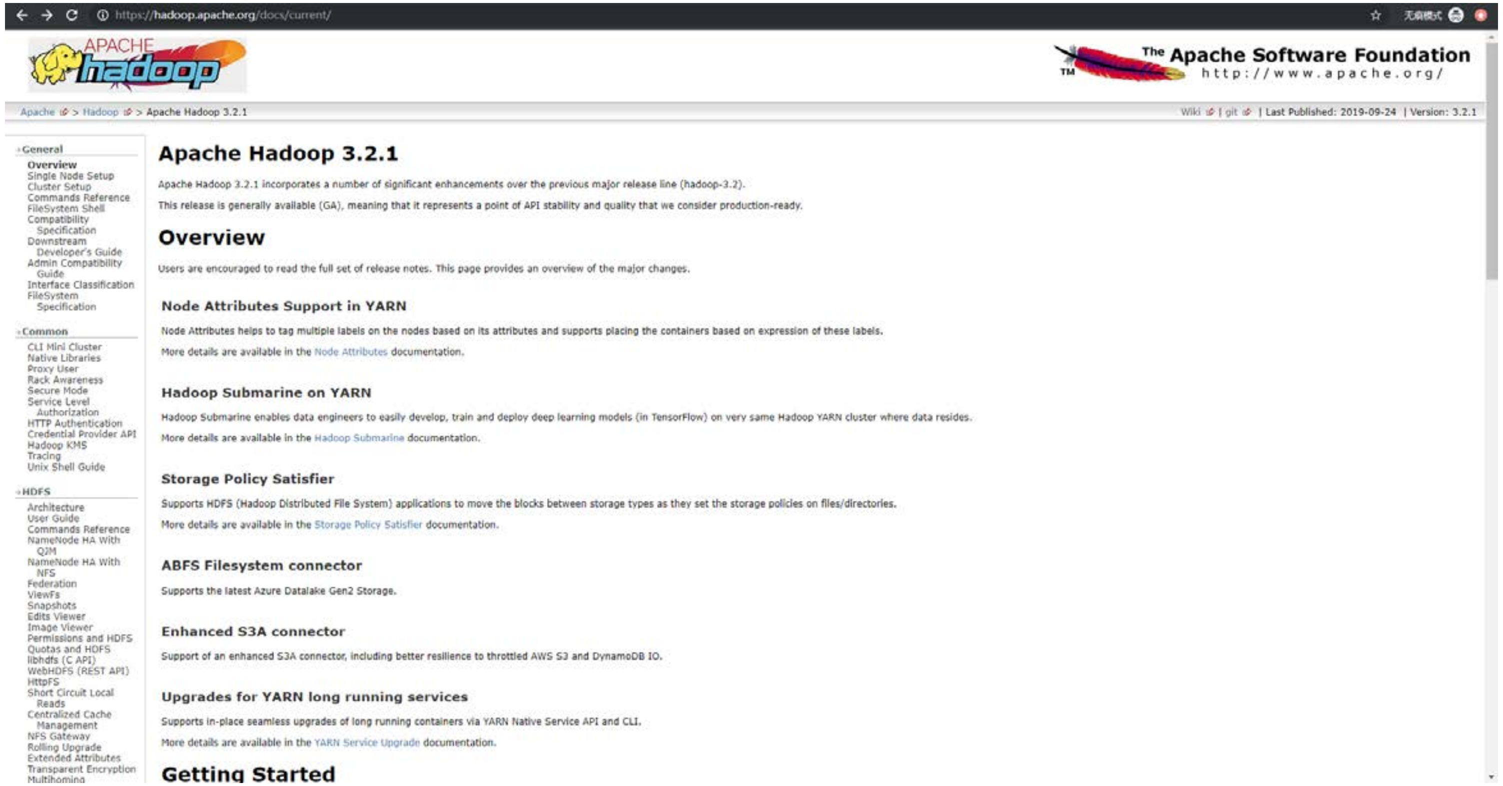
Hadoop

- https://hadoop.apache.org/

- The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.
- The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
- It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Hadoop Documents

- https://developer.yahoo.com/hadoop/tutorial/
- https://hadoop.apache.org/docs/current/



Apache Hadoop 3.2.1

Apache Hadoop 3.2.1 incorporates a number of significant enhancements over the previous major release line (hadoop-3.2).

This release is generally available (GA), meaning that it represents a point of API stability and quality that we consider production-ready.

Overview

Users are encouraged to read the full set of release notes. This page provides an overview of the major changes.

Node Attributes Support in YARN

Node Attributes helps to tag multiple labels on the nodes based on its attributes and supports placing the containers based on expression of these labels.

More details are available in the Node Attributes documentation.

Hadoop Submarine on YARN

Hadoop Submarine enables data engineers to easily develop, train and deploy deep learning models (in TensorFlow) on very same Hadoop YARN cluster where data resides.

More details are available in the Hadoop Submarine documentation.

Storage Policy Satisfier

Supports HDFS (Hadoop Distributed File System) applications to move the blocks between storage types as they set the storage policies on files/directories.

More details are available in the Storage Policy Satisfier documentation.

ABFS Filesystem connector

Supports the latest Azure Datalake Gen2 Storage.

Enhanced S3A connector

Support of an enhanced S3A connector, including better resilience to throttled AWS S3 and DynamoDB IO.

Upgrades for YARN long running services

Supports in-place seamless upgrades of long running containers via YARN Native Service API and CLI.

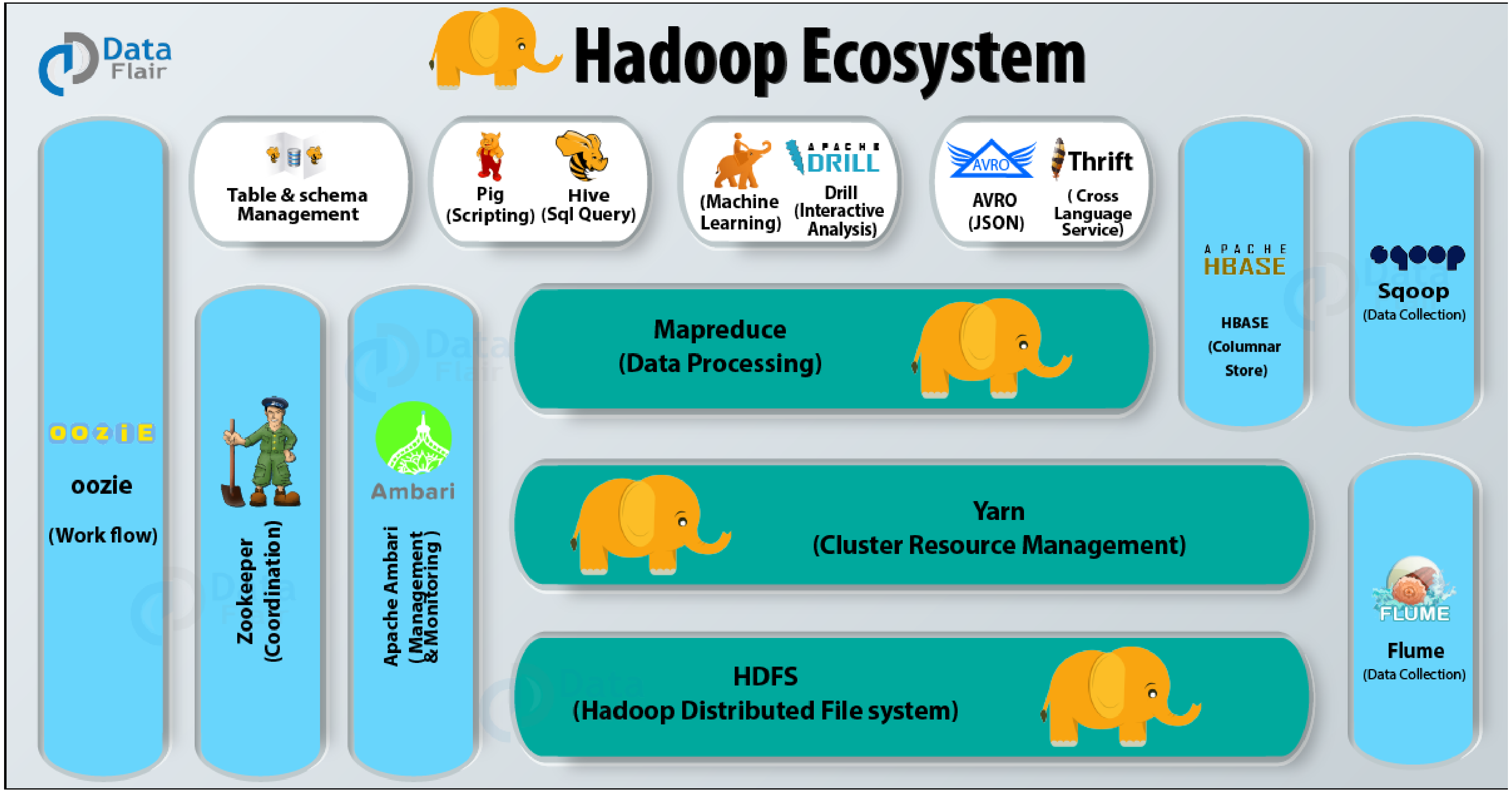
More details are available in the YARN Service Upgrade documentation.

Getting Started

Hadoop Core Components

- HDFS
  - Distributed storage on clusters of machines
- MapReduce
  - Distributed data processing locally on cluster machines
- Yarn
  - Resource management

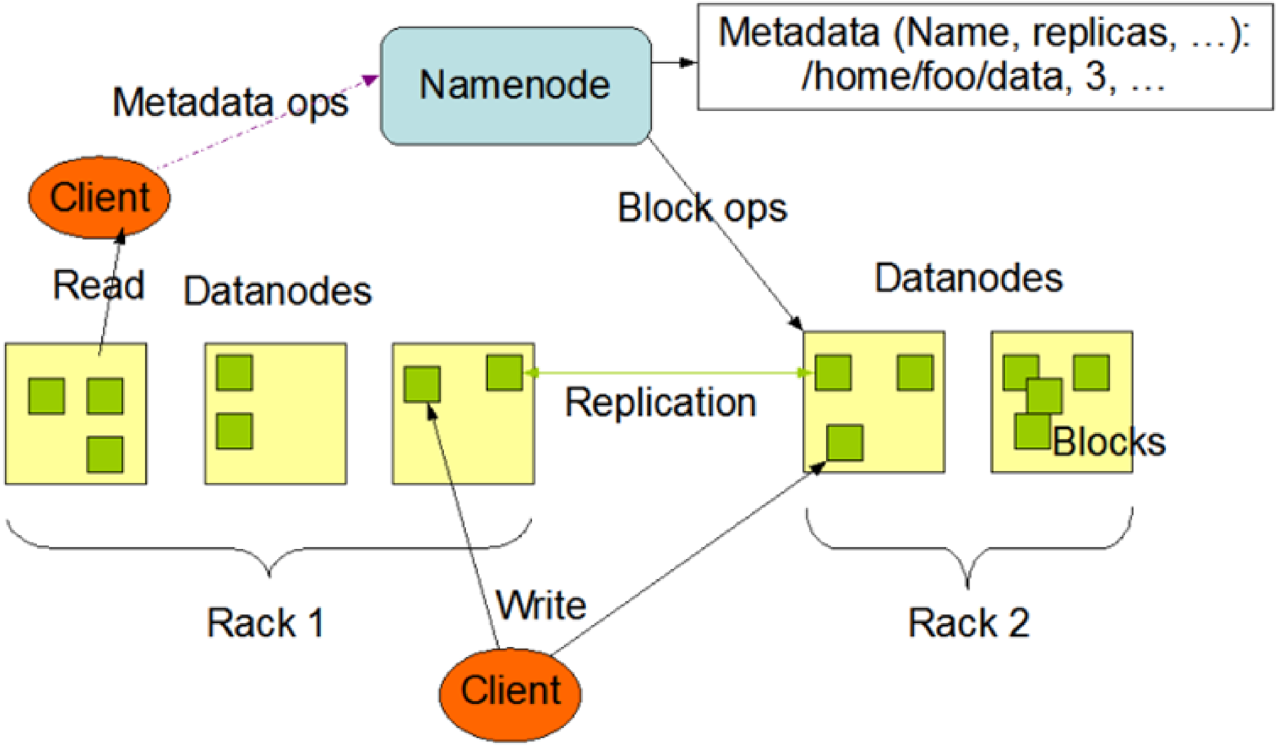
Hadoop Ecosystem



https://data-flair.training/blogs/hadoop-ecosystem/

HDFS

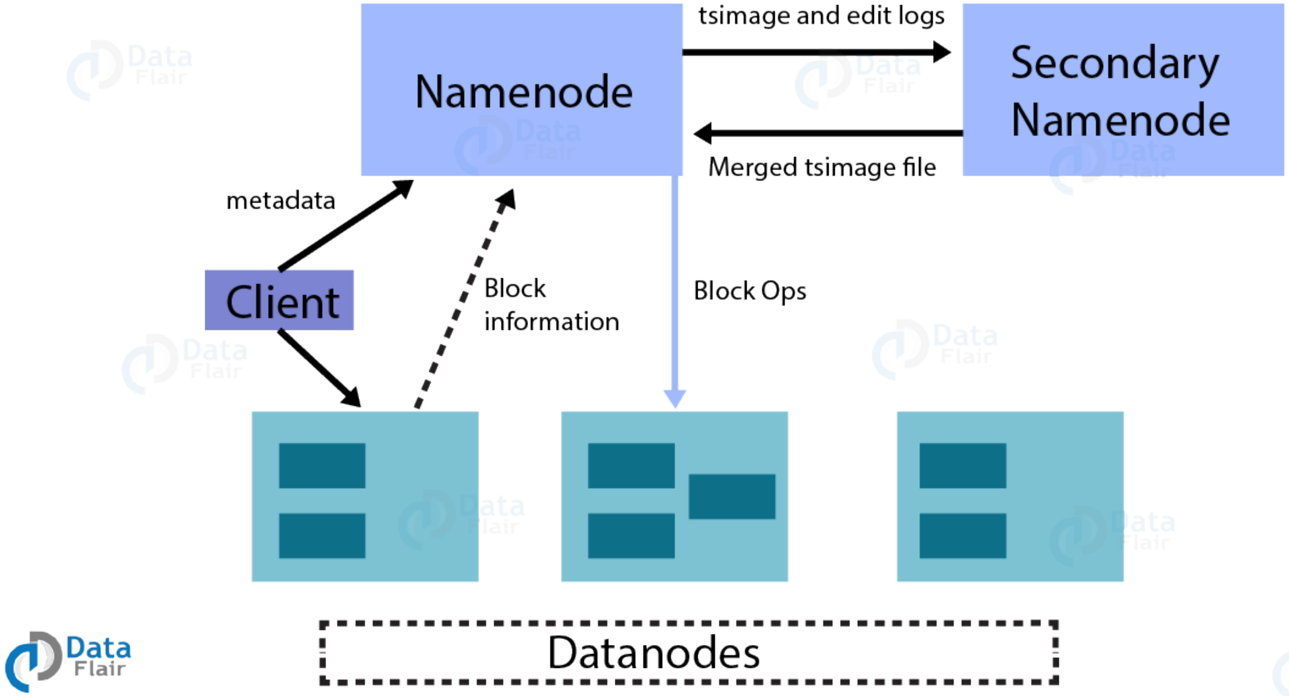
- Files get partitioned into blocks
- Block size configurable *dfs.block.size*
- Replicated *dfs.replication* times
- Namenode: master
- Datanode: slave



Apache.org

HDFS

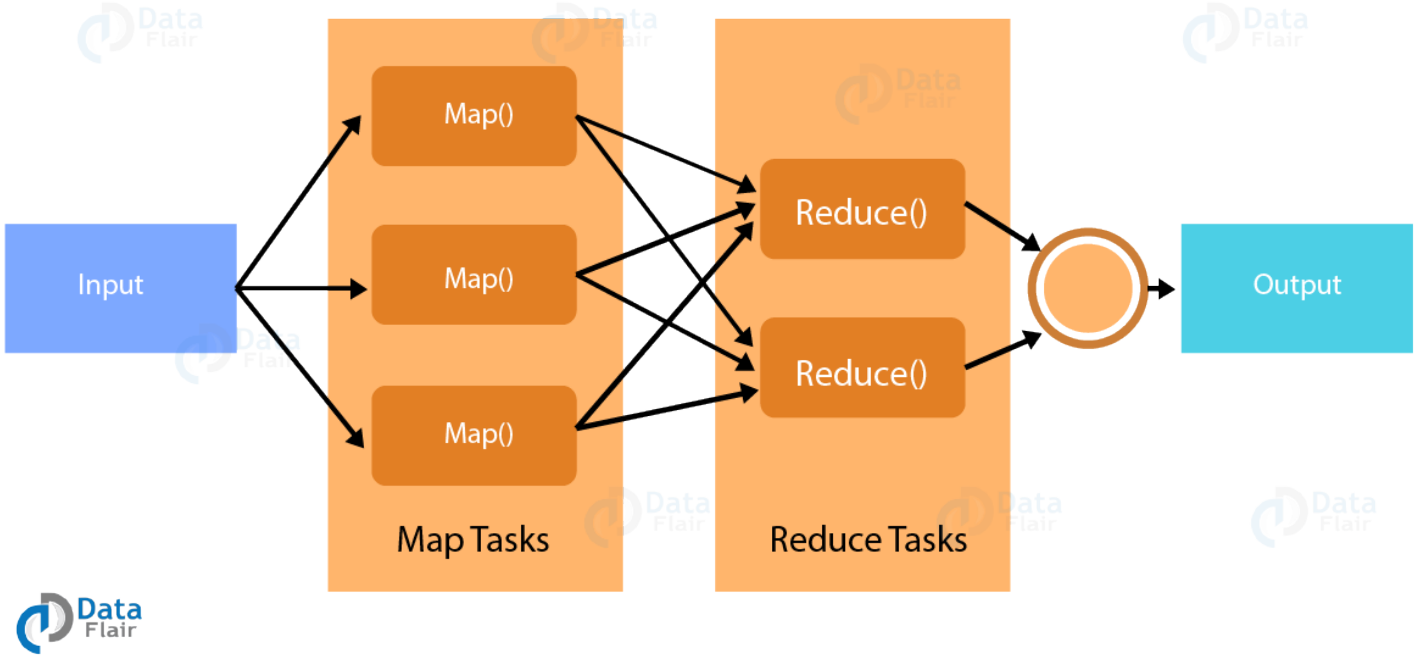
- Files writing process
- Files reading process
- Secondary Namenode
- Standby Namenode



https://data-flair.training/blogs/hadoop-ecosystem/

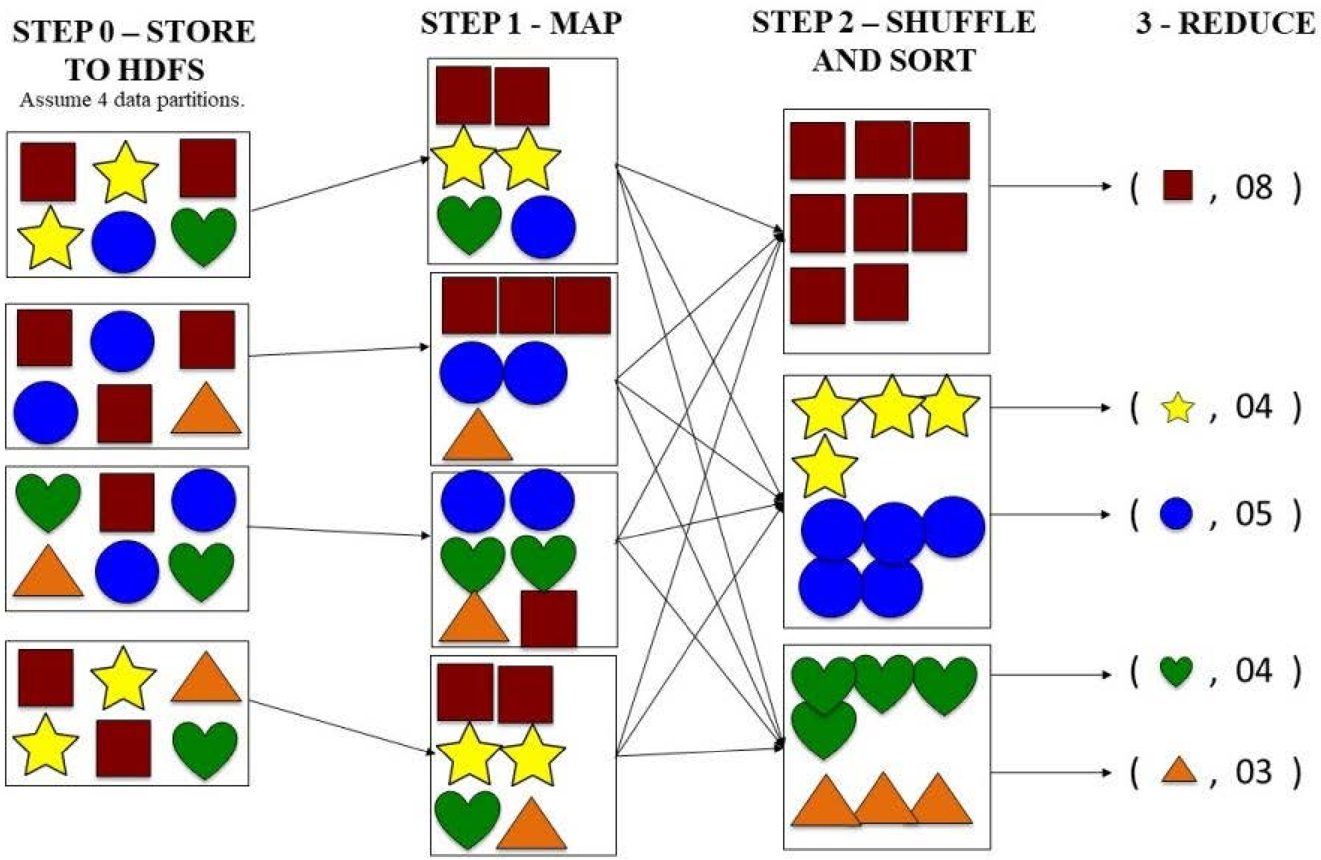
MapReduce

- Processing in parallel
- Intermediate records, basically key-value pairs
- Shuffle and sort based on keys
- Merge results



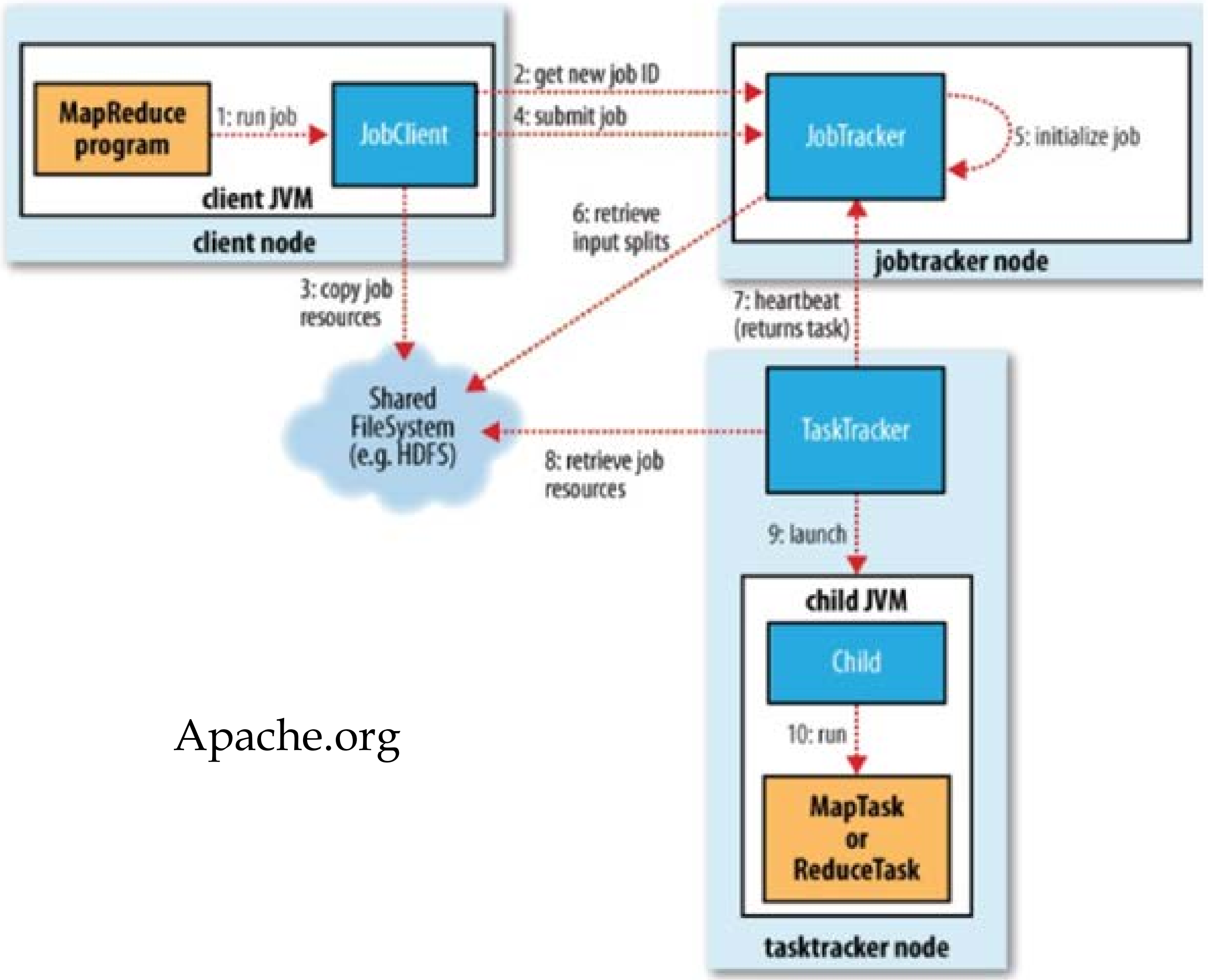
https://data-flair.training/blogs/hadoop-ecosystem/

MapReduce – Example



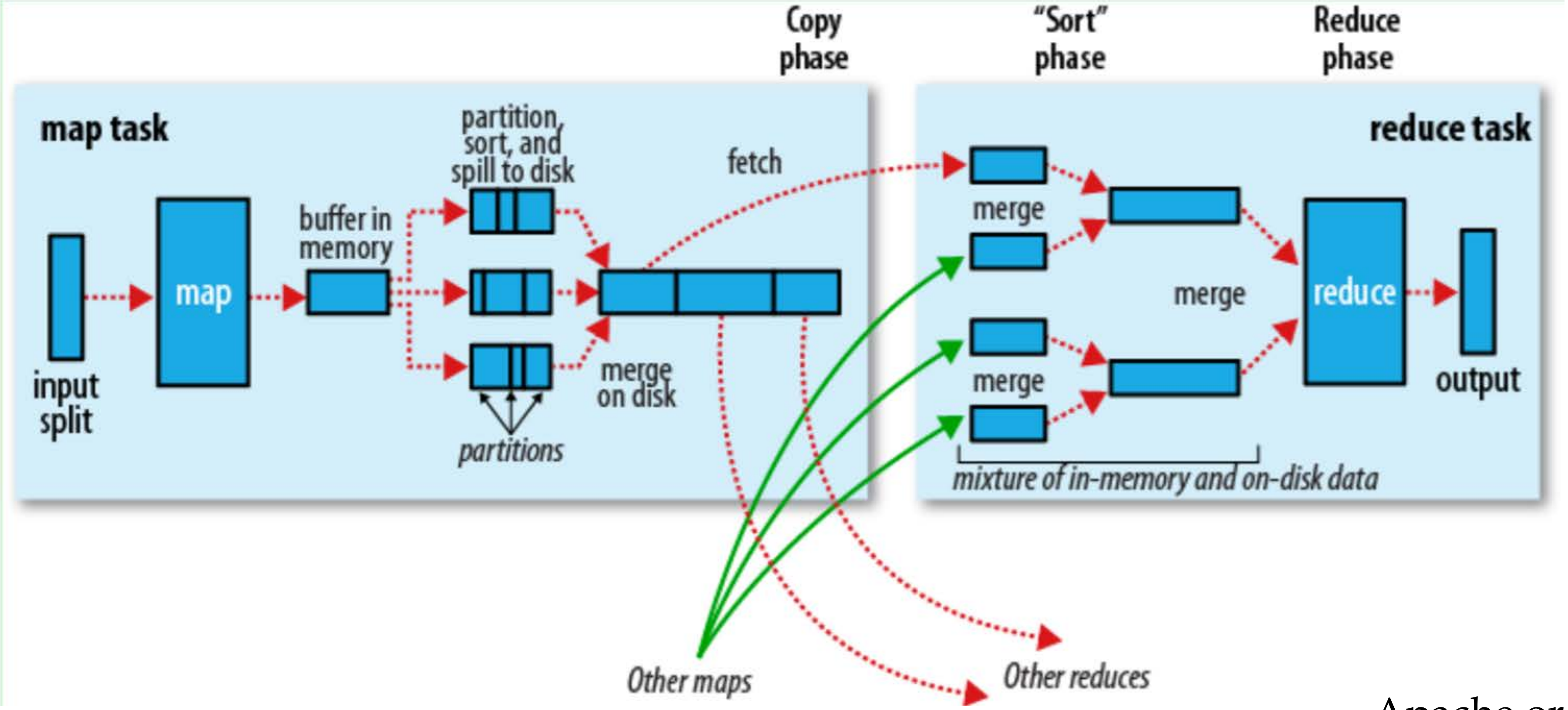
https://www.digitalvidya.com/blog/hdfs-and-yarn-explained/

MapReduce Daemons



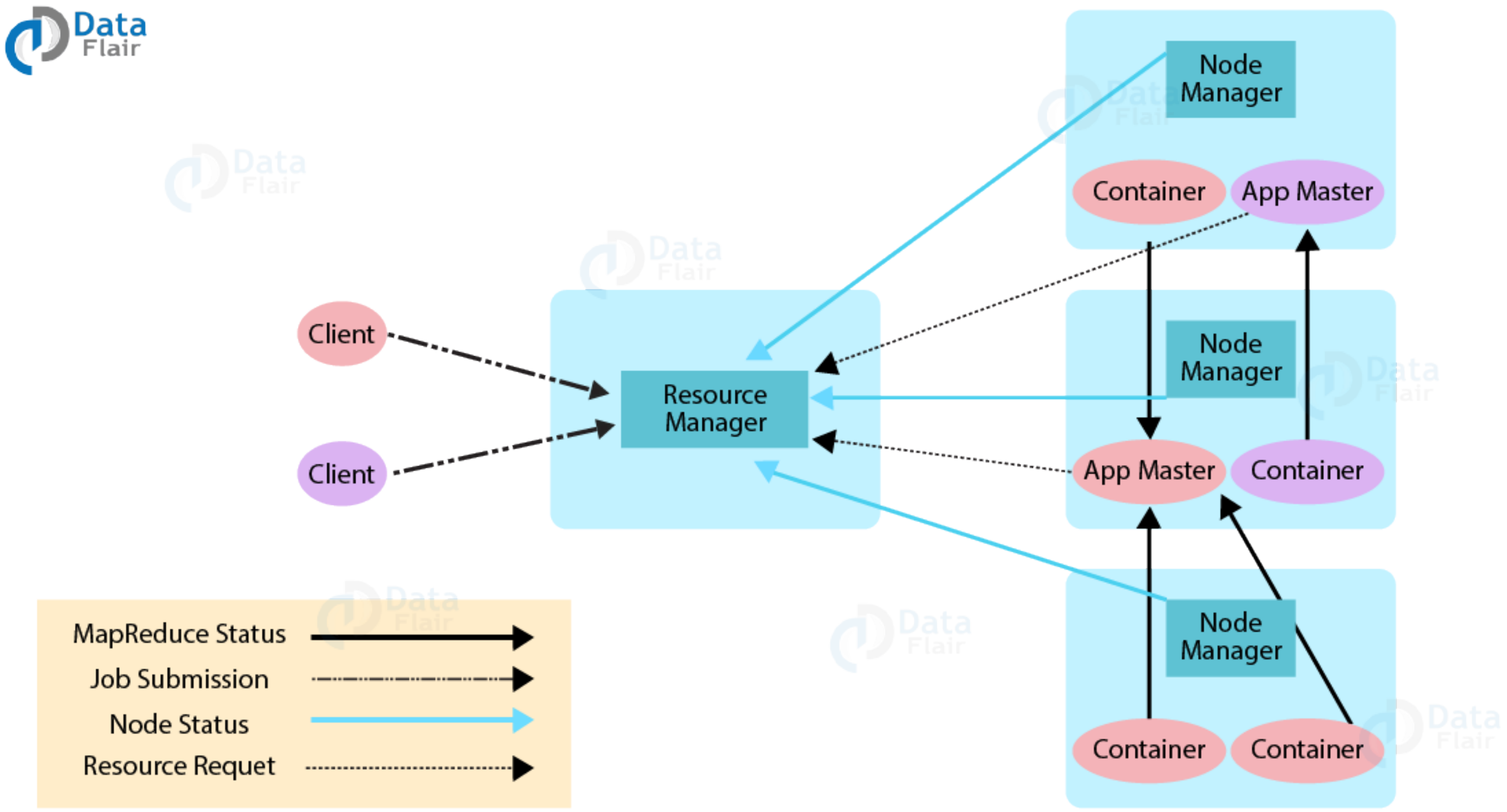
Apache.org

MapReduce Complete Workflow



Apache.org

Yarn

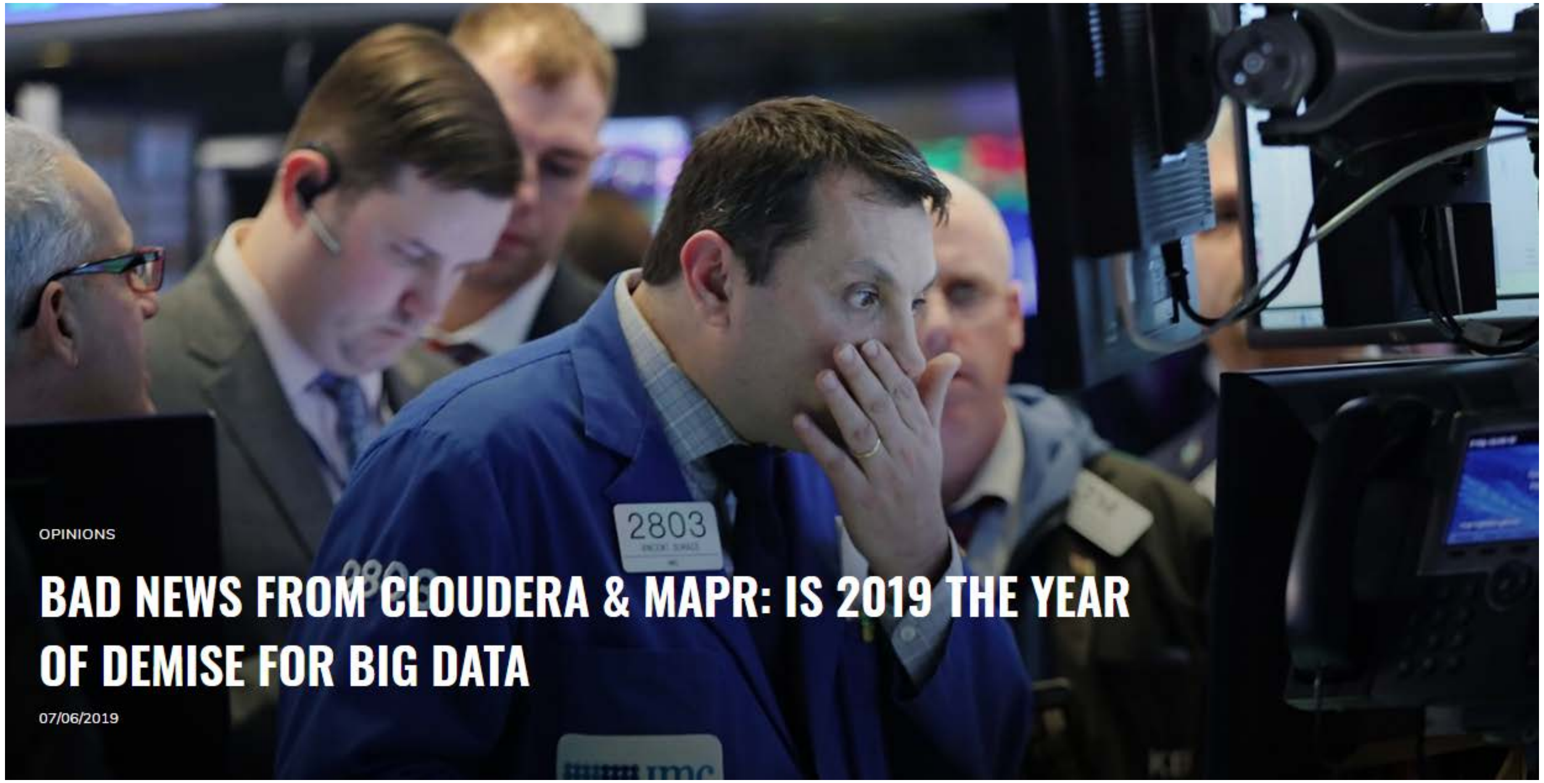


https://data-flair.training/blogs/hadoop-ecosystem/

Demos

- Discussions
    - Cons and pros
- Patterns
    - Filtering: sampling
    - Summarization: counting, statistics
    - Structural patterns: combination

Future of Hadoop



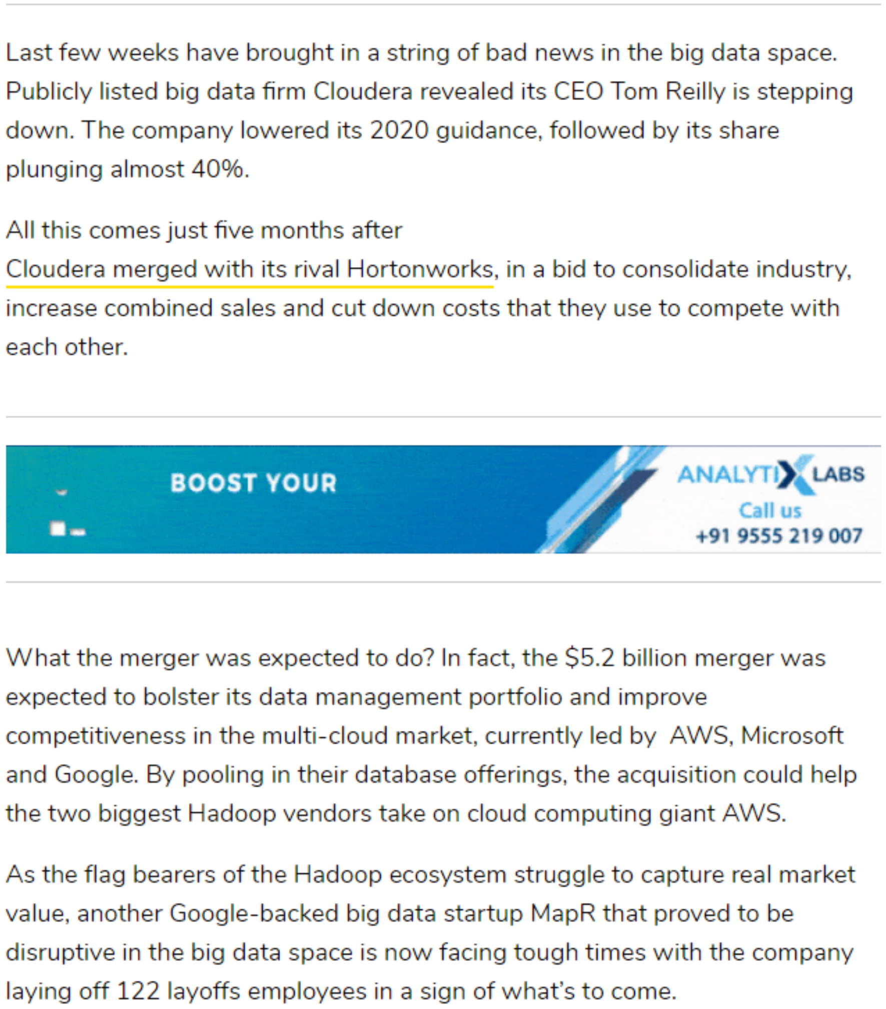
OPINIONS

BAD NEWS FROM CLOUDERA & MAPR: IS 2019 THE YEAR OF DEMISE FOR BIG DATA

07/06/2019

https://analyticsindiamag.com/bad-news-from-cloudera-mapr-is-2019-the-year-of-demise-for-big-data/

Future of Hadoop

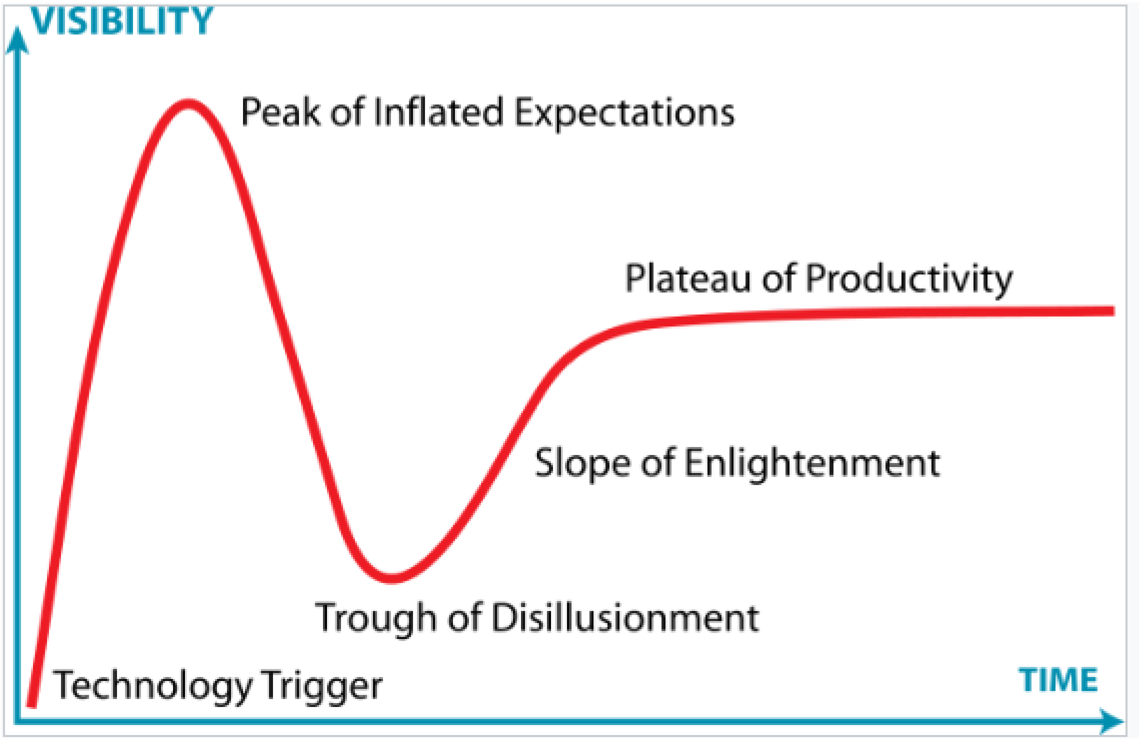


Last few weeks have brought in a string of bad news in the big data space. Publicly listed big data firm Cloudera revealed its CEO Tom Reilly is stepping down. The company lowered its 2020 guidance, followed by its share plunging almost 40%.

All this comes just five months after Cloudera merged with its rival Hortonworks, in a bid to consolidate industry, increase combined sales and cut down costs that they use to compete with each other.

What the merger was expected to do? In fact, the $5.2 billion merger was expected to bolster its data management portfolio and improve competitiveness in the multi-cloud market, currently led by AWS, Microsoft and Google. By pooling in their database offerings, the acquisition could help the two biggest Hadoop vendors take on cloud computing giant AWS.

As the flag bearers of the Hadoop ecosystem struggle to capture real market value, another Google-backed big data startup MapR that proved to be disruptive in the big data space is now facing tough times with the company laying off 122 layoffs employees in a sign of what's to come.



Hype cycle

https://analyticsindiamag.com/bad-news-from-cloudera-mapr-is-2019-the-year-of-demise-for-big-data/

Cloud Computing vs. Hadoop

- Cloud Computing is economical as it has lower maintenance costs centralized platform no upfront cost and disaster safe implementation. Whereas, Big data is highly scalable, robust ecosystem, and cost-effective.
- Services on demand vs. distributed storage and processing of large sets of data
- Cloud computing easier to use while Hadoop with more flexibility to adjust for customization
- Open source success?