

Hadoop and Yarn

Md Sahil
0017010501029

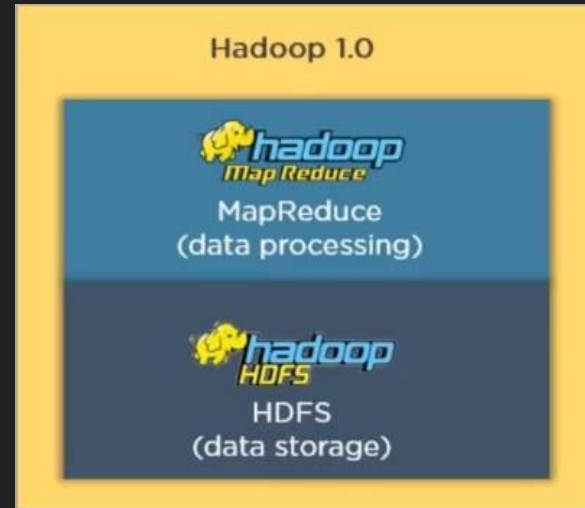
Priyank Lohariwal
001710501055

*Department of Computer Science and Engineering
Jadavpur University*

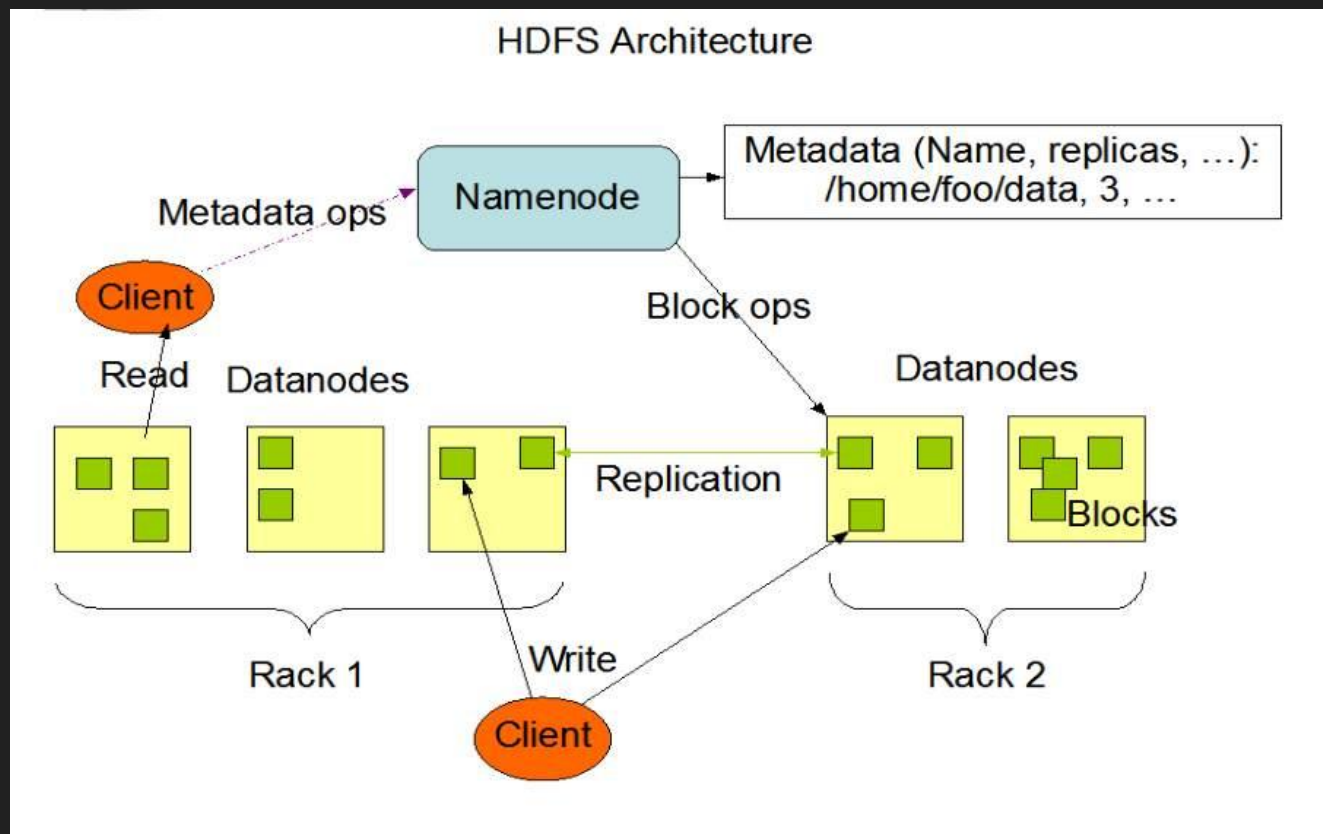
What is Hadoop?

Hadoop 1.x

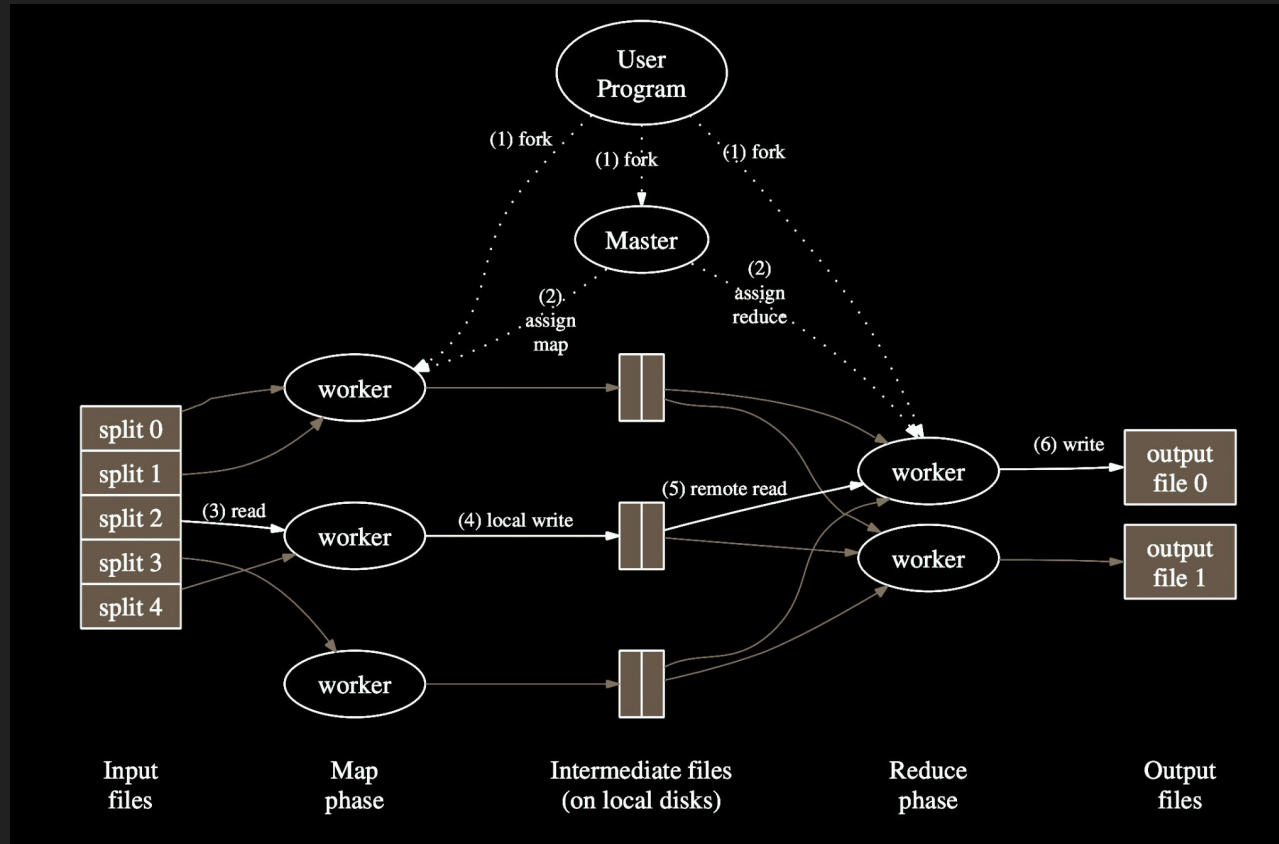
- HDFS
- MapReduce

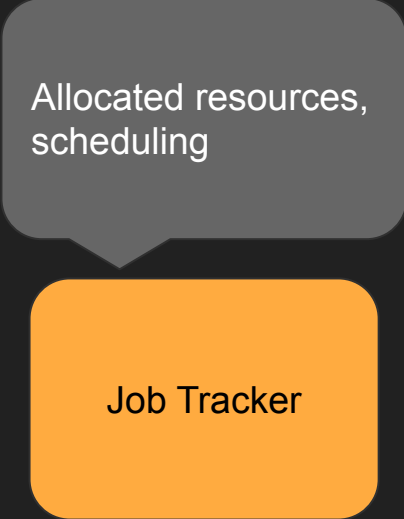


HDFS



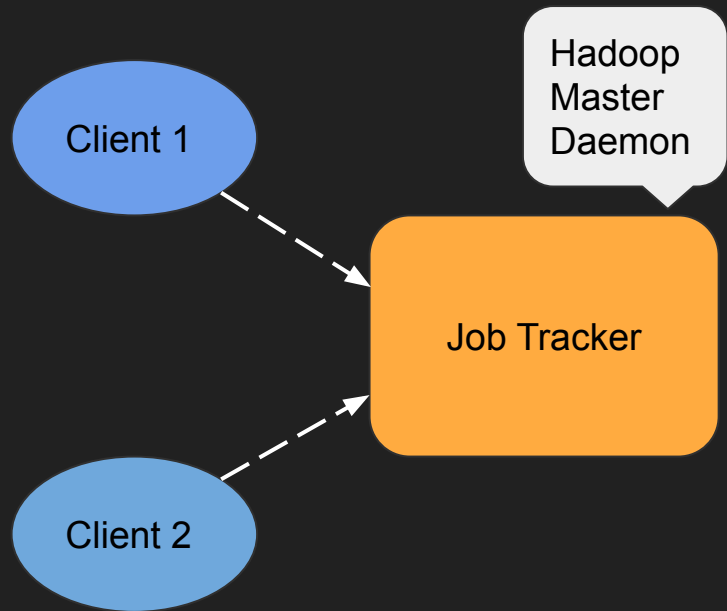
MapReduce

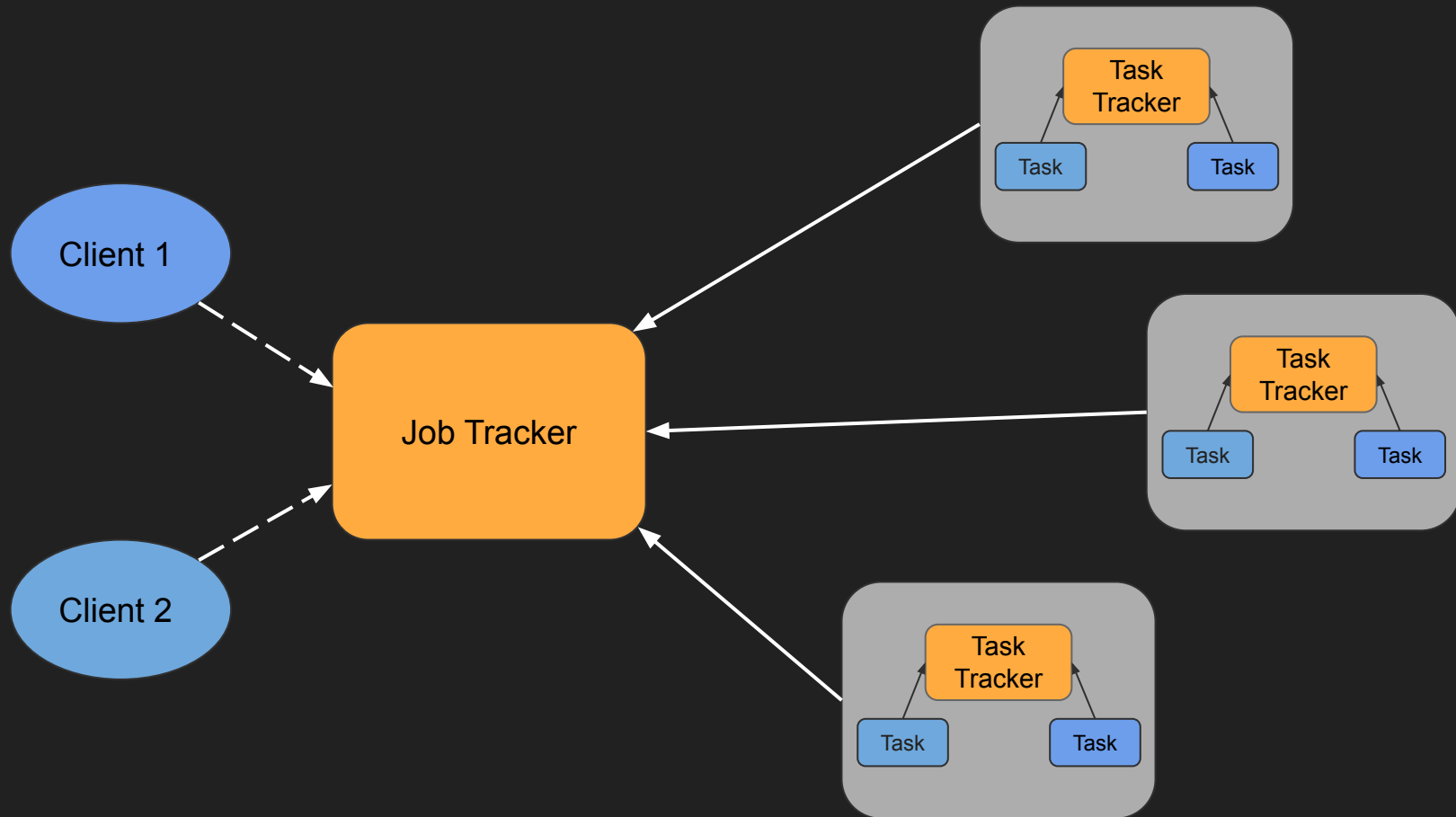




Allocated resources,
scheduling

Job Tracker

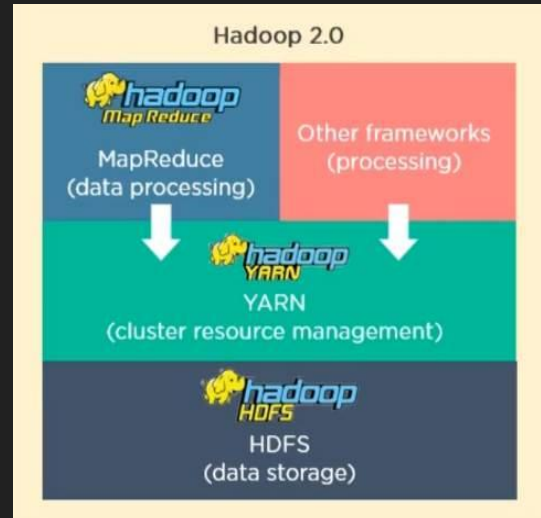
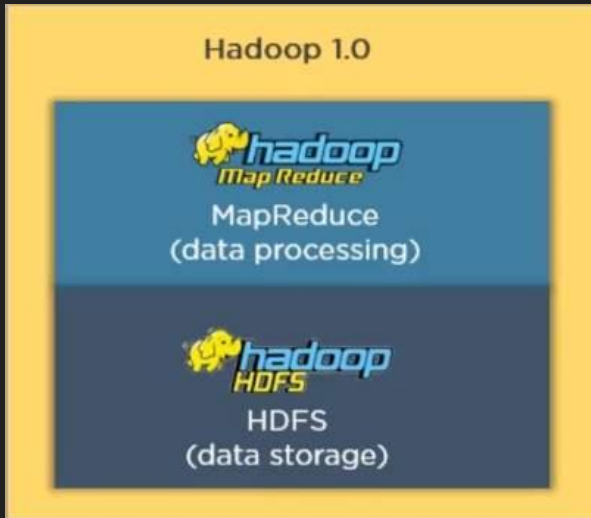




Limitations of Hadoop 1

- Scalability
- Availability
- Resource Utilization
- Limited Programming models

Hadoop 2.X

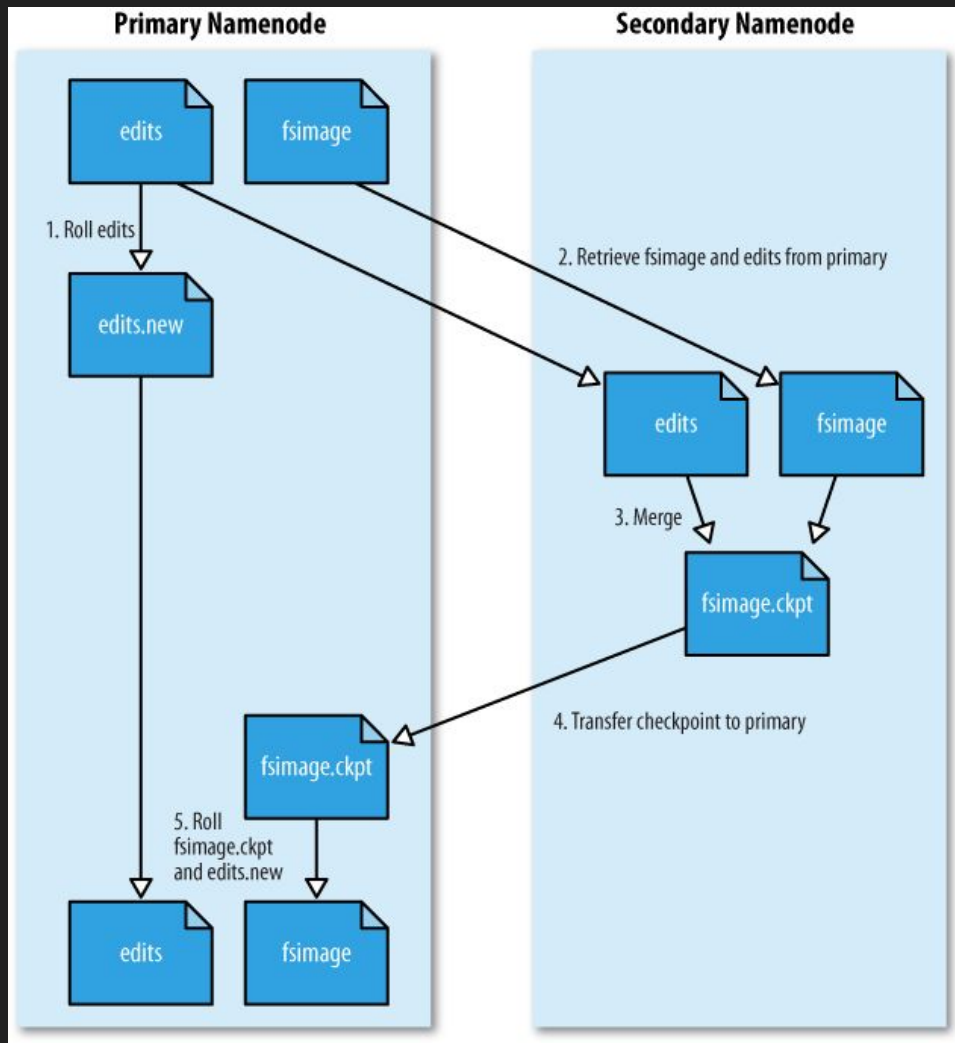


HDFS v2

- Master-Slave Architecture
- NameNode
- DataNode
- Secondary NameNode
- Stand by NameNode
- Federation
- Heartbeats

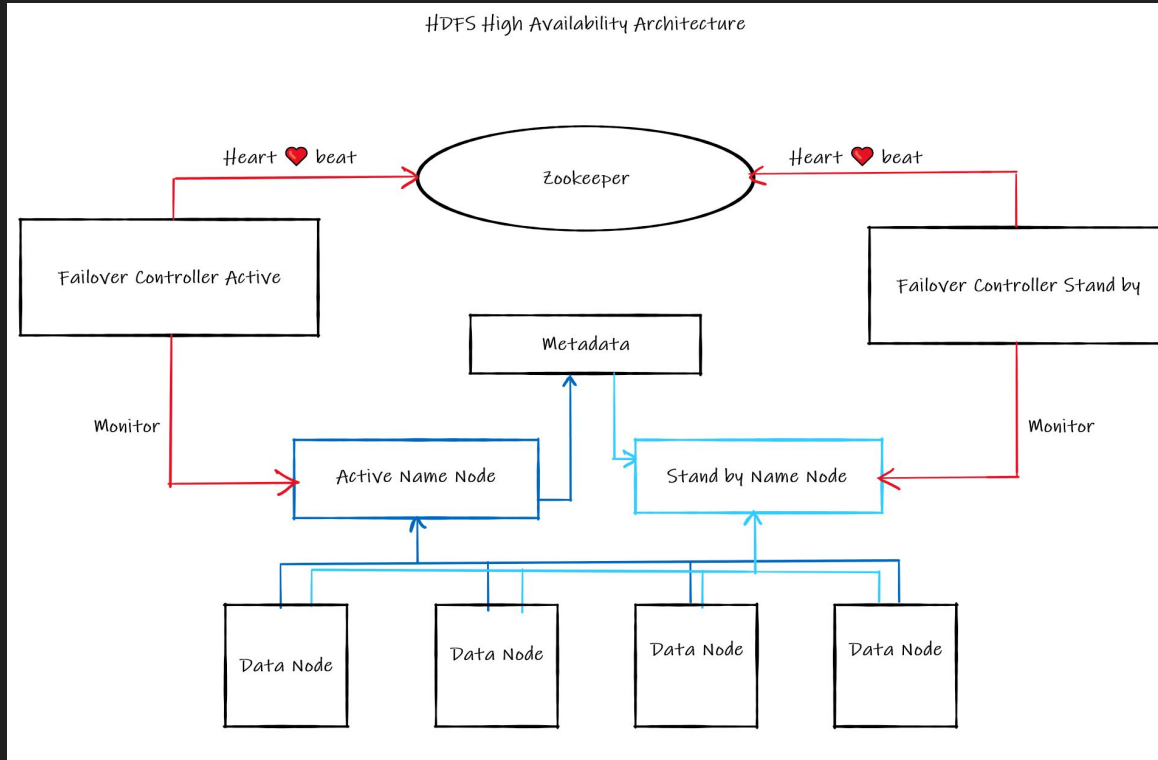
HDFS v2

Secondary NameNode



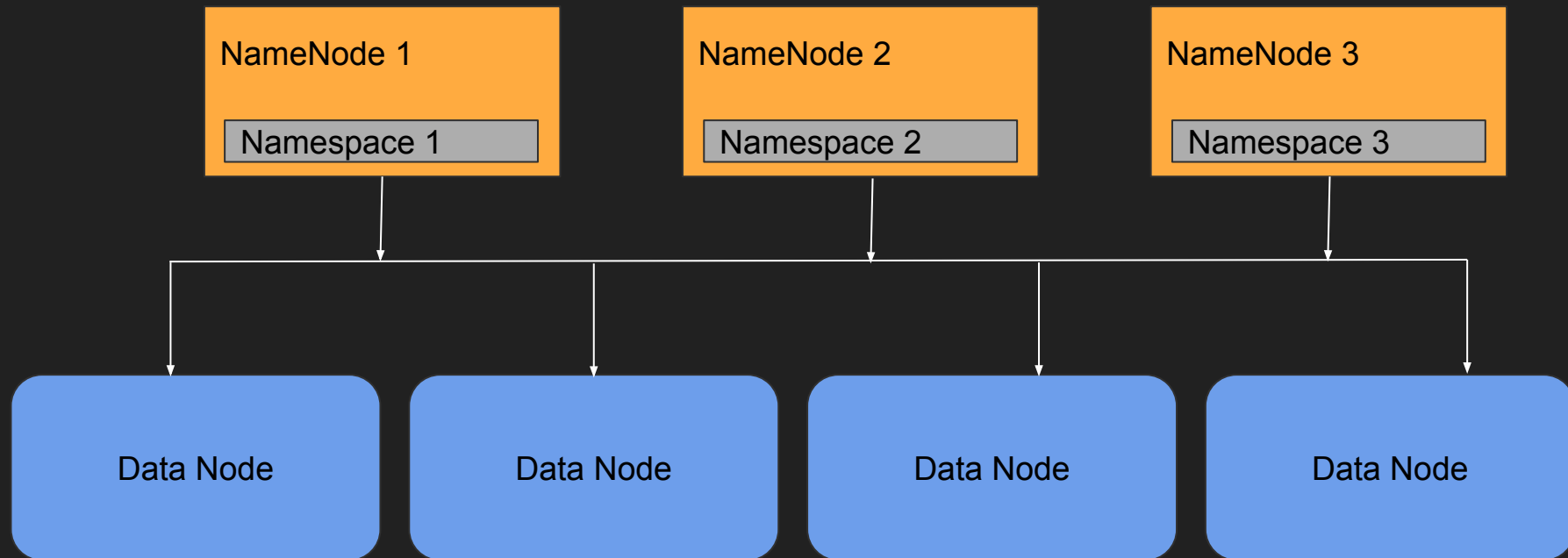
HDFS v2

Stand by NameNode - High Availability



HDFS v2

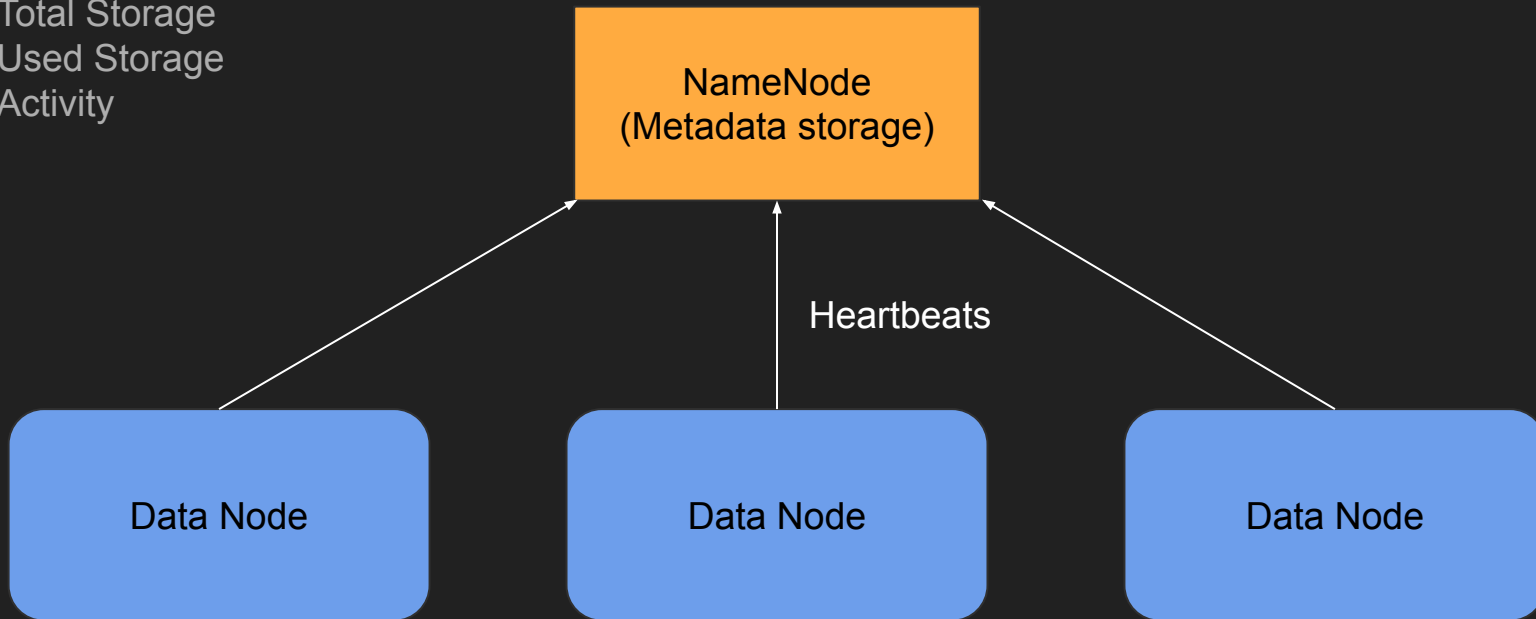
Federation



HDFS v2

Heartbeats

- Total Storage
- Used Storage
- Activity



Yet Another Resource Negotiator

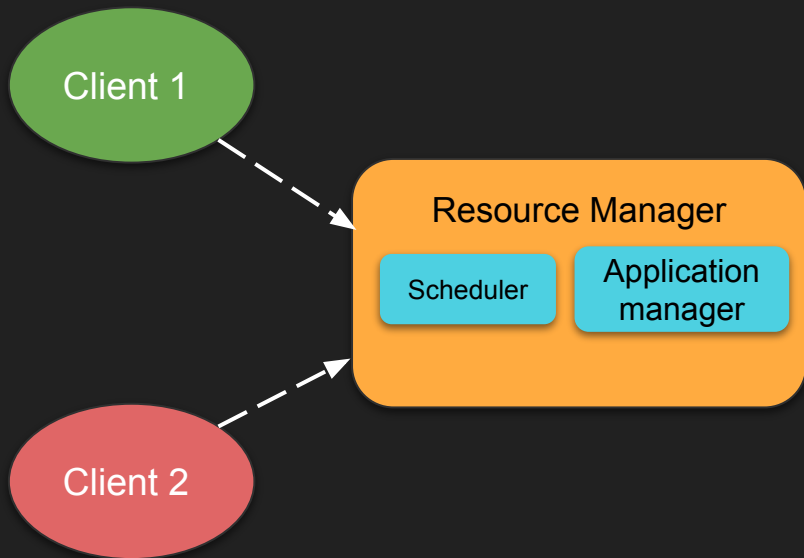
Resource management

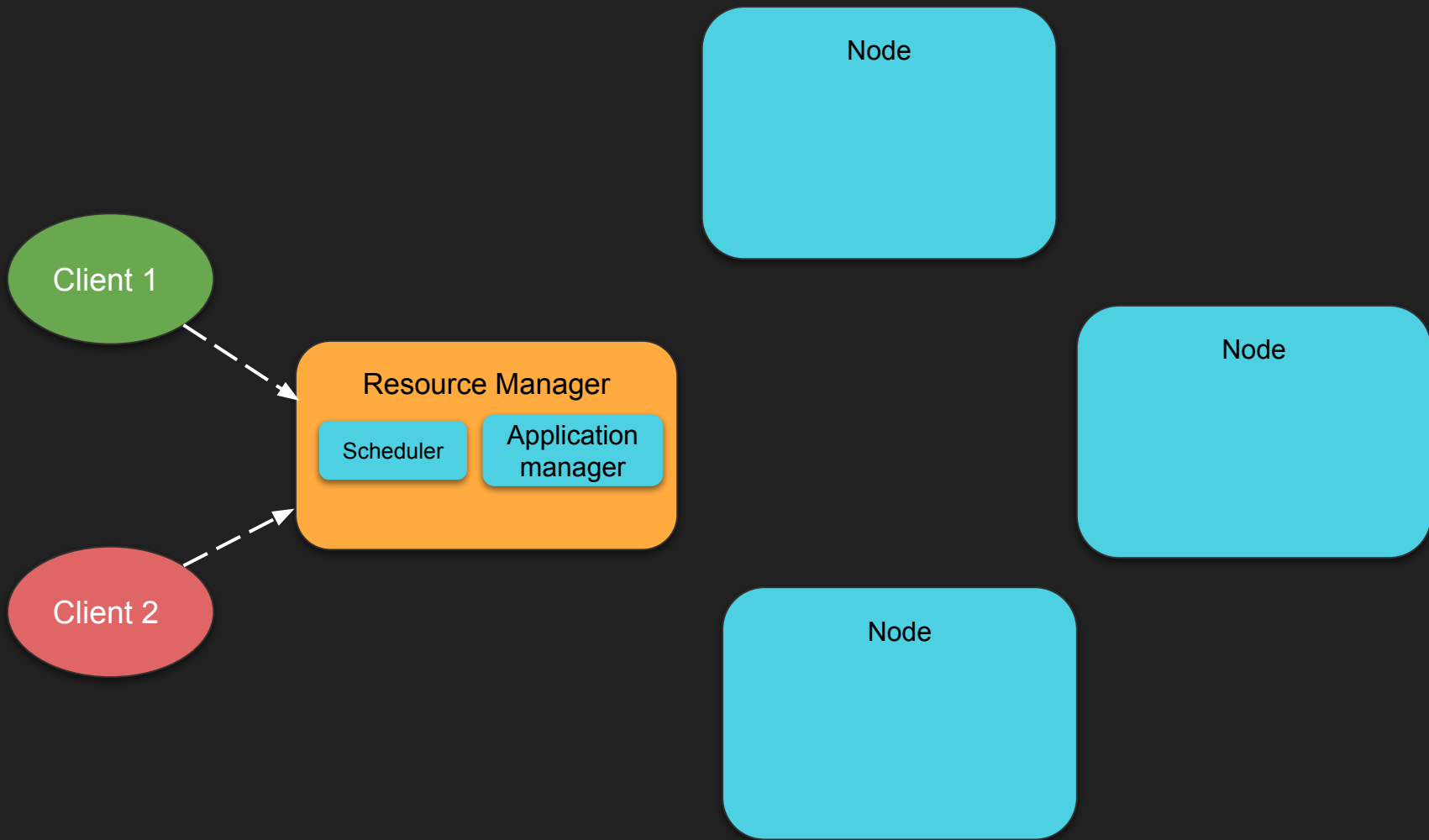
Job Scheduling

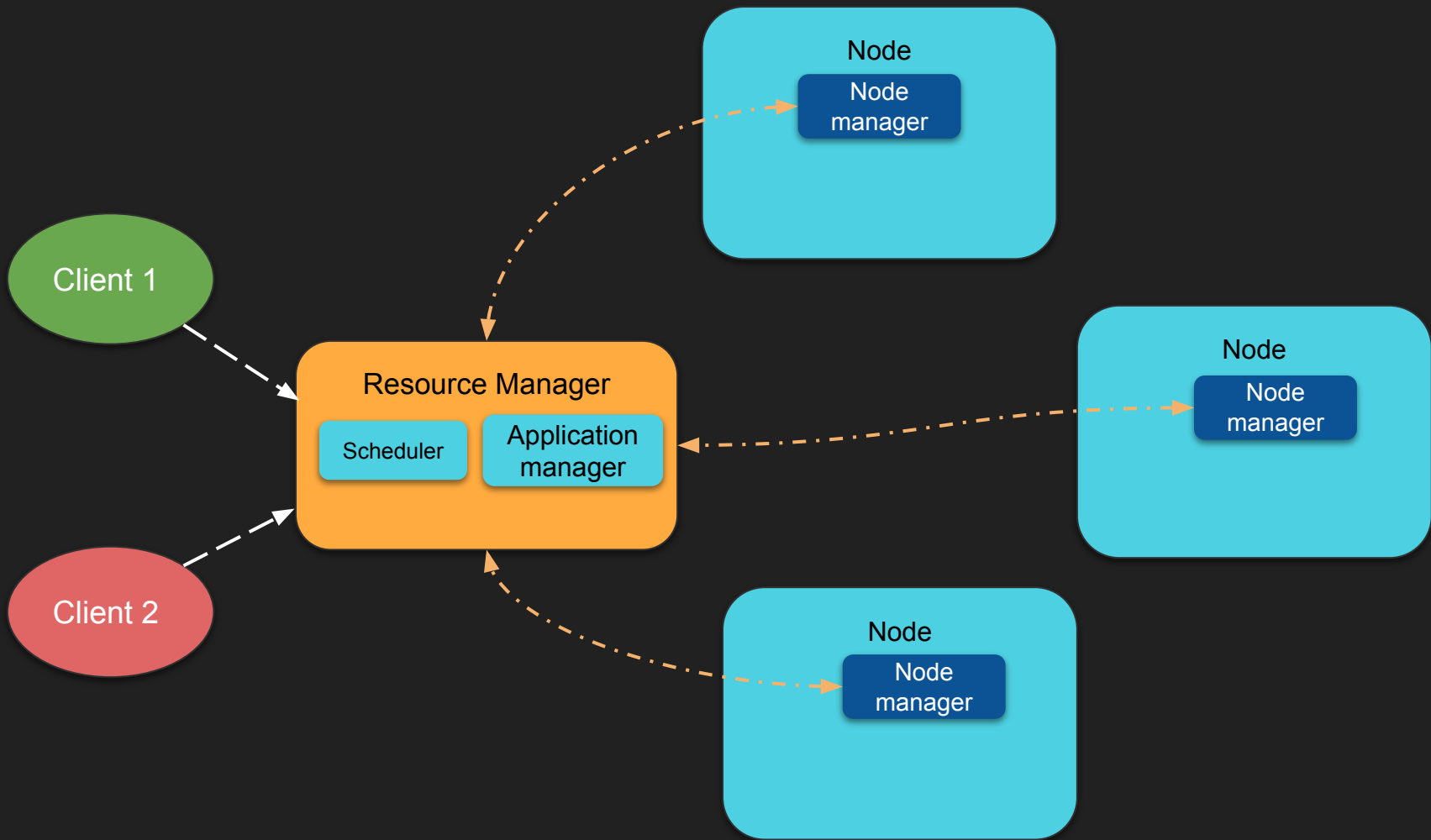
Resource Manager

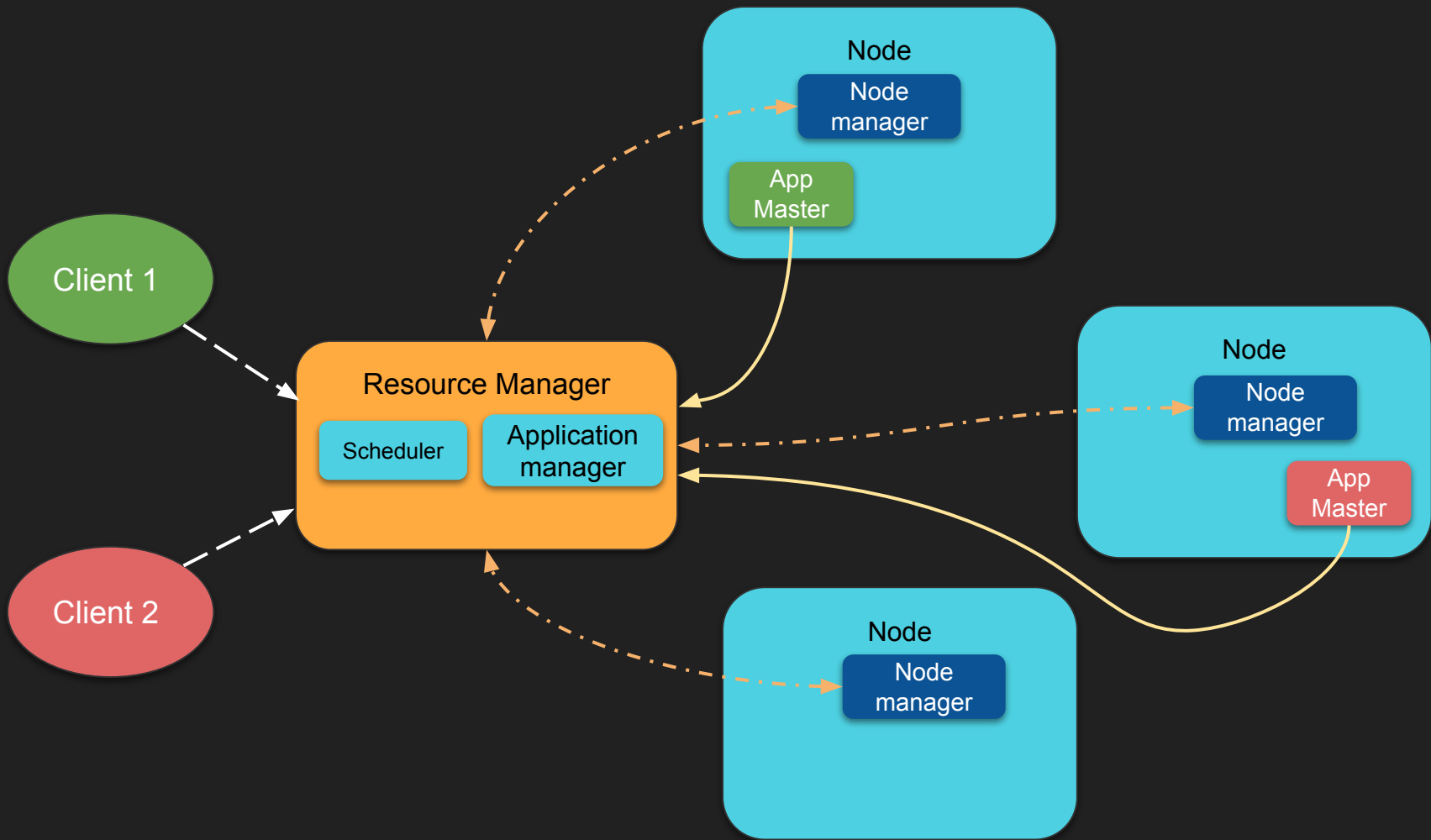
Scheduler

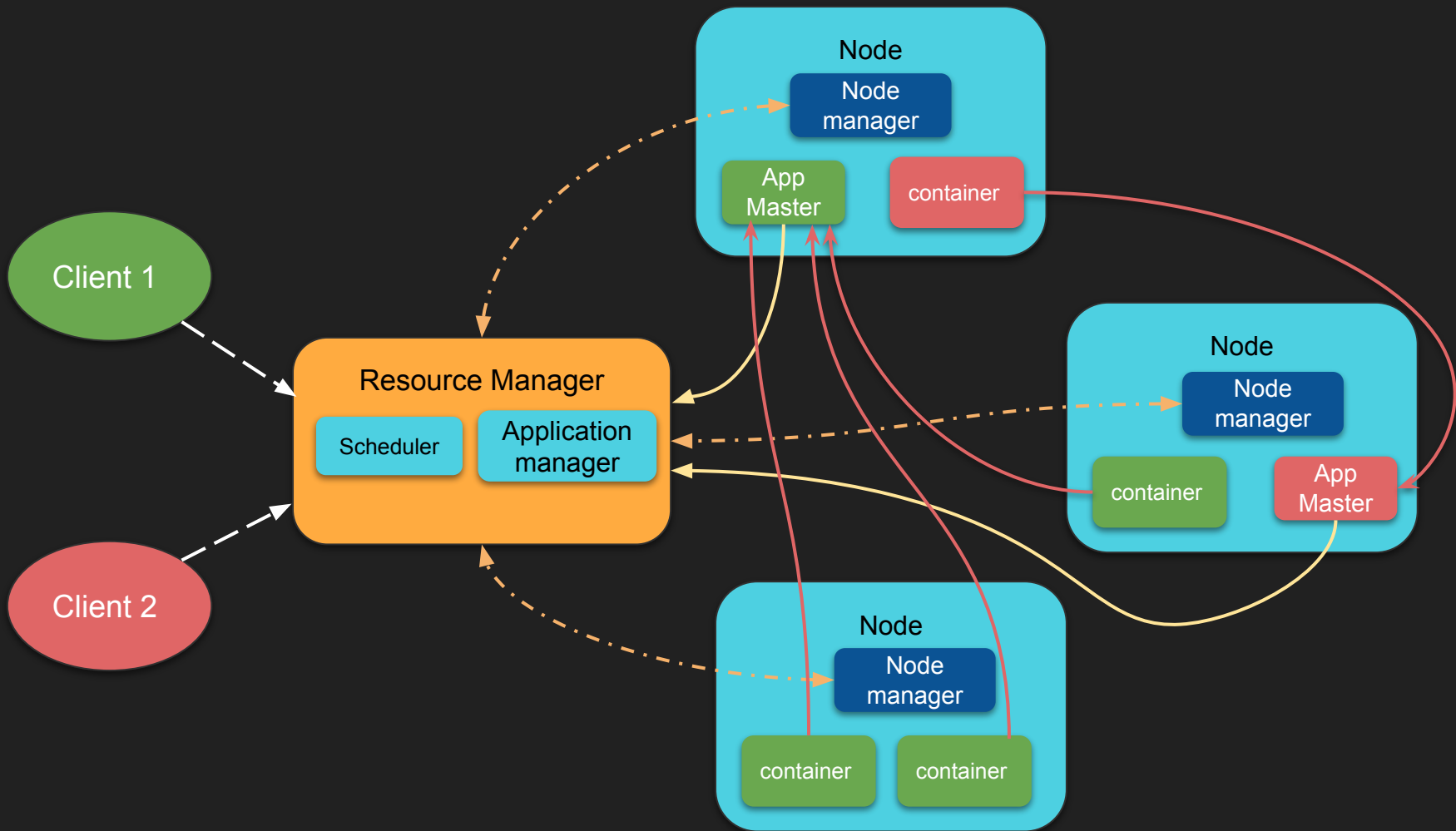
Application
manager



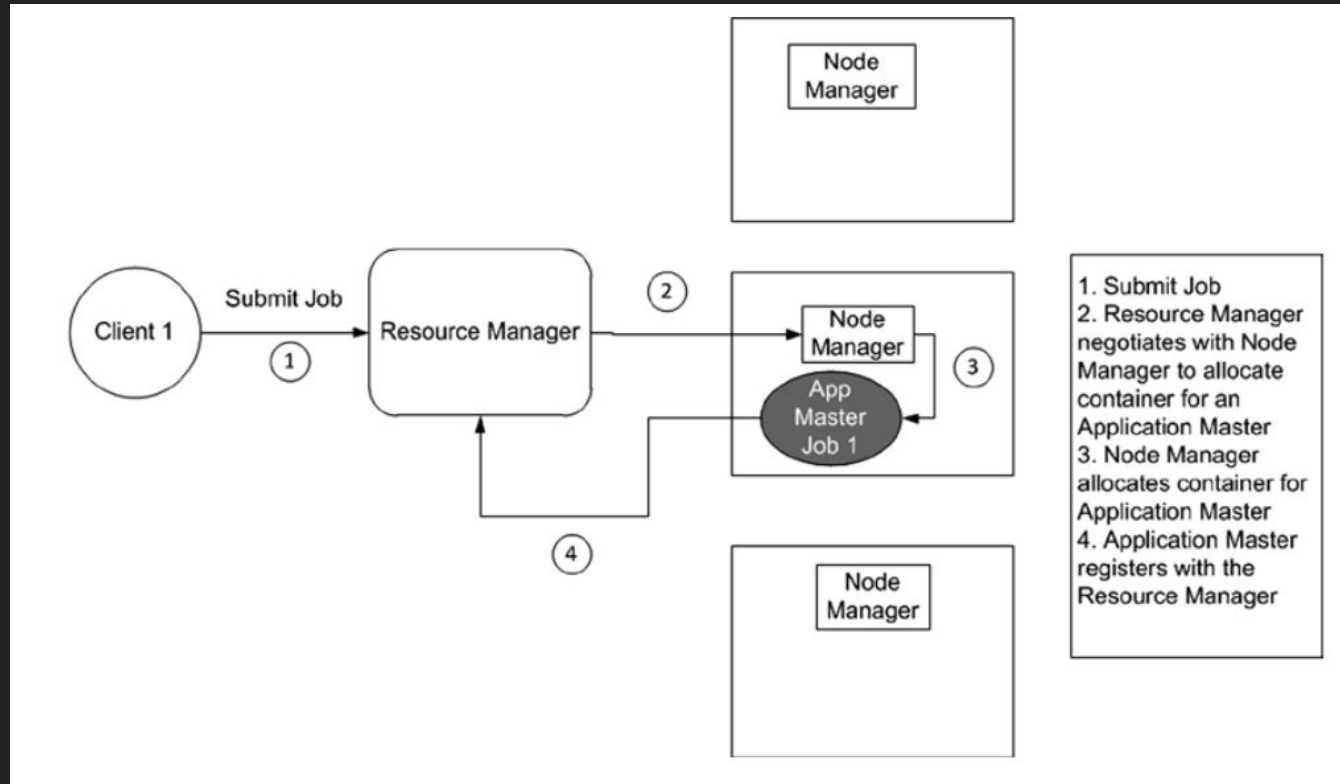




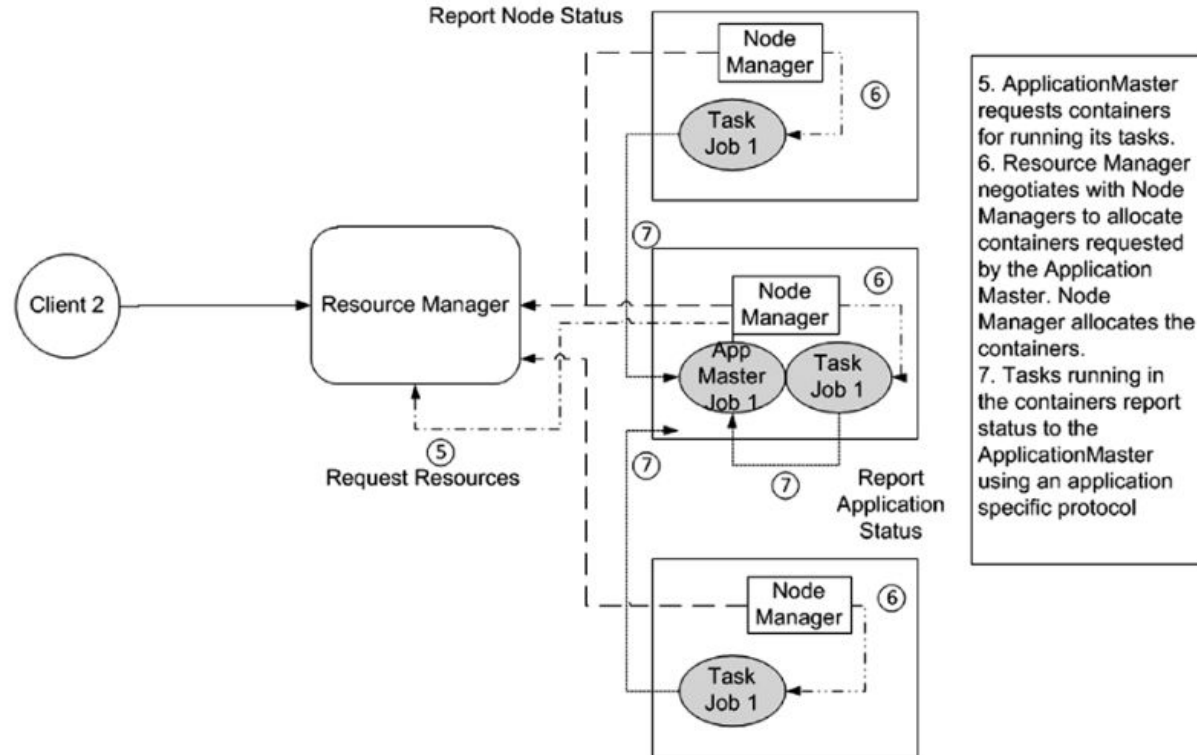




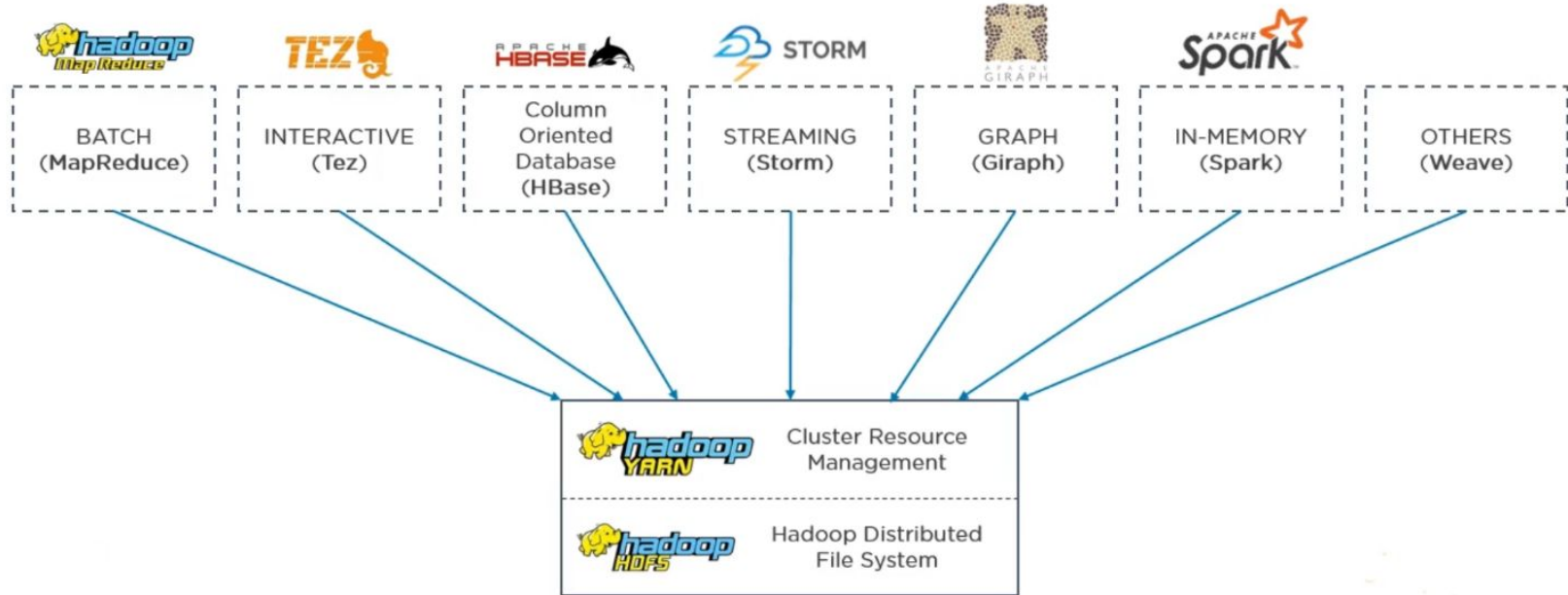
Application master startup



Job resource allocation execution



List of frameworks that runs on top of YARN:



Hadoop 2.x over Hadoop 1.x

- Fault tolerance and Failure recovery
- Multiple namespaces
- Scalability: 10,000 nodes per cluster
- Resource Utilization
- Multiple Programming Models

Future of Hadoop

Future of Hadoop

Hadoop 3.x

- JournalNodes
- Erasure coding for data replication
- Intra-datanode data balancer
- Support for multiple file systems
- Support for opportunistic containers and distributed scheduling

Company	Business	Technical Specs	Uses
Facebook	Social Site	8 cores and 12 TB of storage	Used as a source for reporting and machine learning
Twitter	Social site		Hadoop is used since 2010 to store and process tweets, log files using LZOP compression technique as it is fast and also helps release CPU for other tasks.
LinkedIn	Social site	2X4 and 2X6 cores – 6X2TB SATA 4100 nodes	LinkedIn's data flows through Hadoop clusters. User activity, server metrics, images, transaction logs stored in HDFS are used by data analysts for business analytics like discovering people you may know.
Yahoo!	Online Portal	4500 nodes – 1TB storage, 16 GB RAM	Used for scaling tests

Thank You

References

- <https://hadoop.apache.org/docs/r1.2.1/hdfsdesign.html>
- <http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html>
- <https://www.journaldev.com/8808/hadoop1-architecture-and-how-major-components-works>
- <https://www.ibm.com/analytics/hadoop/hdfs>
- <https://link.springer.com/book/10.1007/978-1-4302-4864-4>
- <https://www.edureka.co/blog/hadoop-yarn-tutorial/>
- <https://hadoop.apache.org/docs/r3.0.0>
- <https://www.dezyre.com/article/top-10-industries-using-big-data-and-121-companies-who-hire-hadoop-developers/69>