Spark:
HDFS, Yarn, Lab

Lecture 04-05

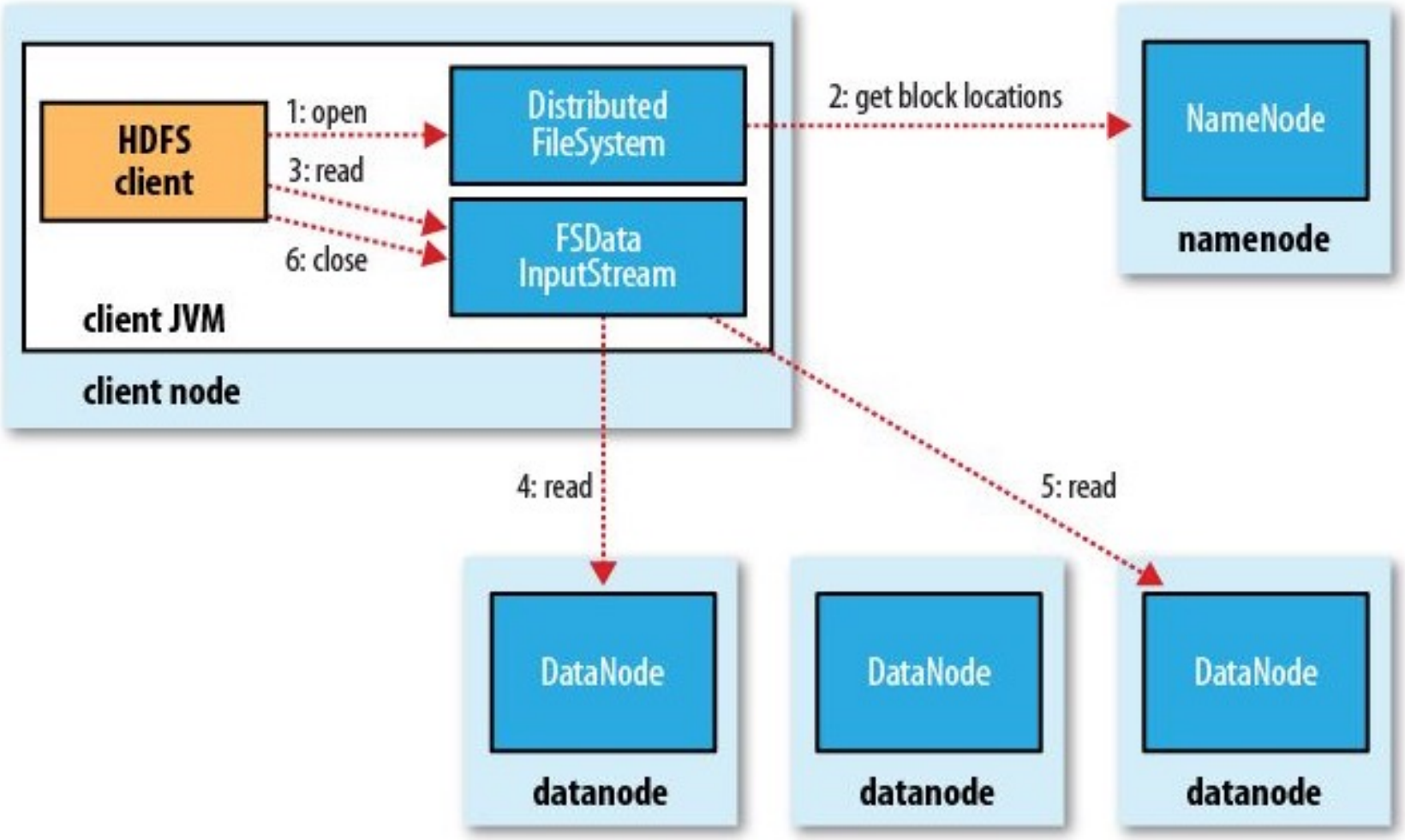
Marilson Campos

UCSC Extension - 2022

Hadoop

HDFS

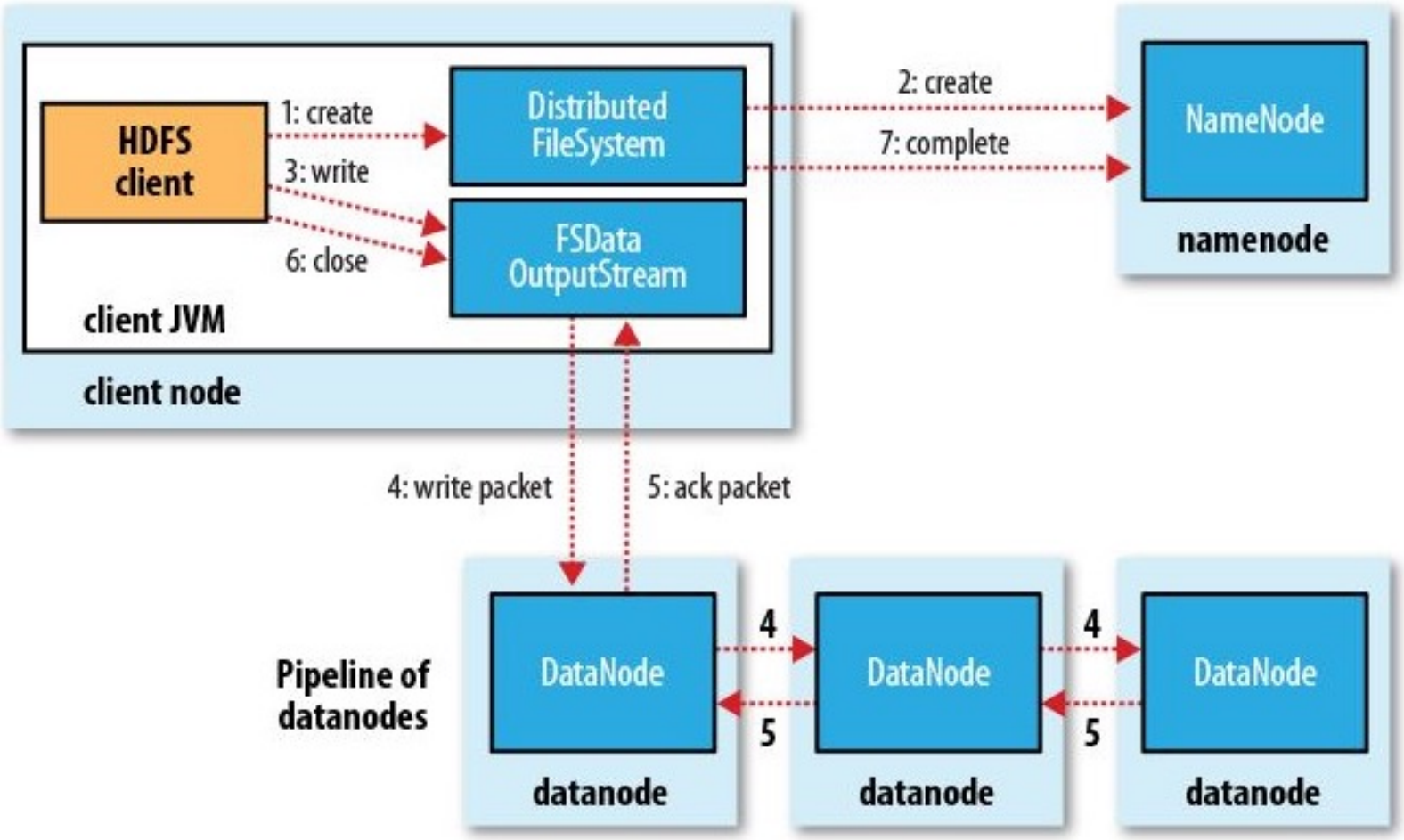
HDFS Read



Hadoop

HDFS

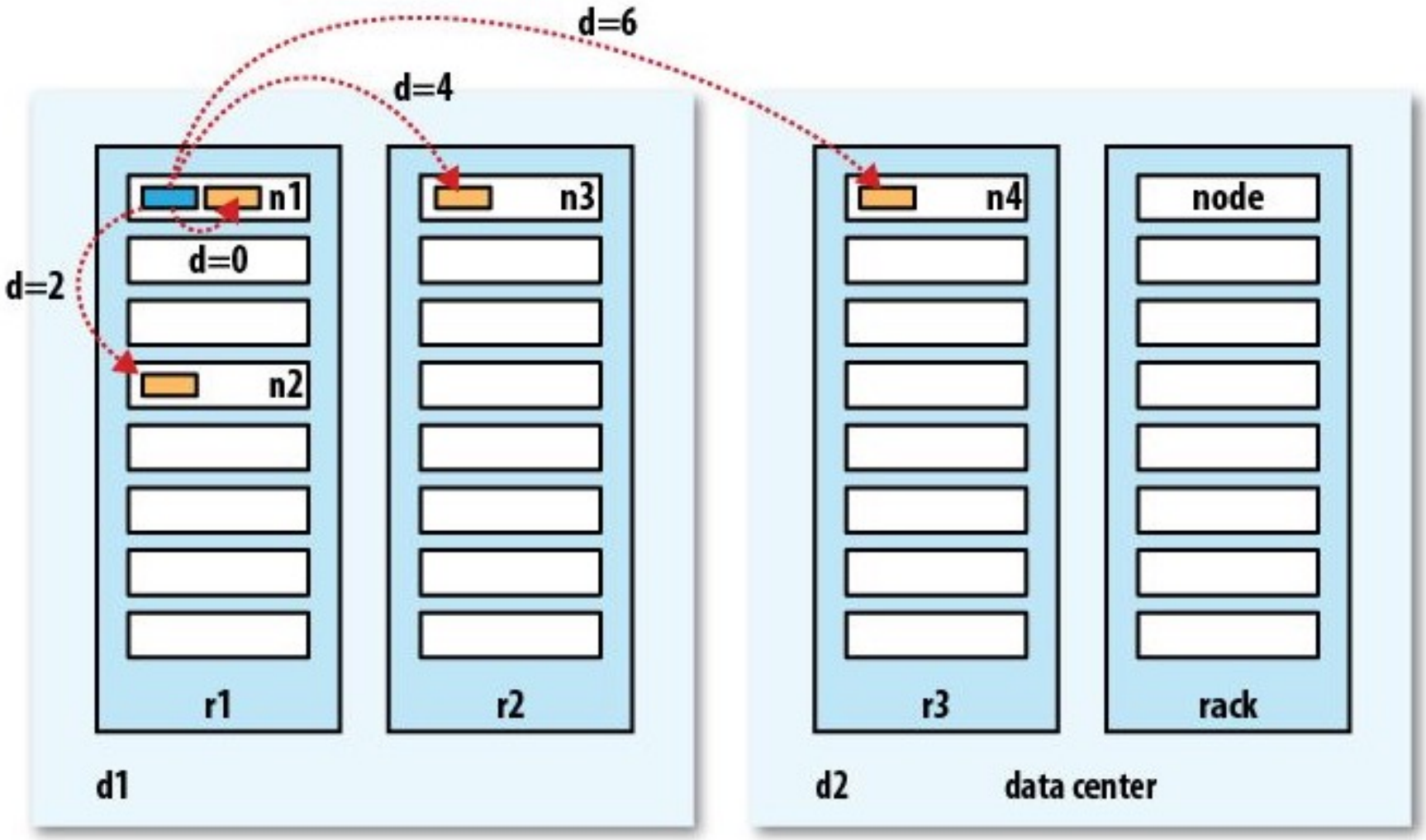
HDFS Write



Hadoop

HDFS

Rack Awareness



Hadoop

HDFS

Network traffic patterns

Website stack traffic

- Many relatively small requests
- Data flows vertically
- “Sharding” data

https://en.wikipedia.org/wiki/Shard_(database architecture)

Hadoop stack traffic

- Large continuous data transfers
- Data flows horizontally
- data natively distributed

Hadoop

HDFS

HDFS Data Integrity

Data can be corrupted during transmission or at rest.

HDFS maintains a CRC-32 for each 512 bytes.
CRC-32 adds an overhead of a little less than 1% of size.

Datanodes verify data coming from clients and other Datanodes during replication.

Also, the “DataNodes” run a background thread called “DataBlockScanner” continuously.

Hadoop

HDFS

HDFS Data Integrity

When the “DataBlockScanner” finds an inconsistency it will use the “replicas” to heal itself.

Hadoop

HDFS Administration

HDFS Data Integrity

Cluster Backups

- Typically, we don't have space for backup all the data.

Solution:

- Backup the “meta-data” of the data from the “NameNodes”.

With popularity of the cloud technology some companies are backing up critical parts using services like Amazon S3 glacier.

Hadoop

HDFS Administration

HDFS Data Integrity

Periodical File System checks

use command “hdfs fsck -report”

There are options to try to ”fix” things.

I recommend run the “report” option to detect errors and review results before issuing commands.

Hadoop

HDFS Administration

HDFS Adding new nodes.

Adding new nodes currently is done using a Web interface provided by each distribution.

One step that is still manual is to make sure you run the Balancer (start-balancer.sh) once you finished to add nodes.

You can add as many nodes and the run the balancer at the end.

Hadoop

HDFS Administration

HDFS Removing nodes.

Removing nodes is also done via Web interface.

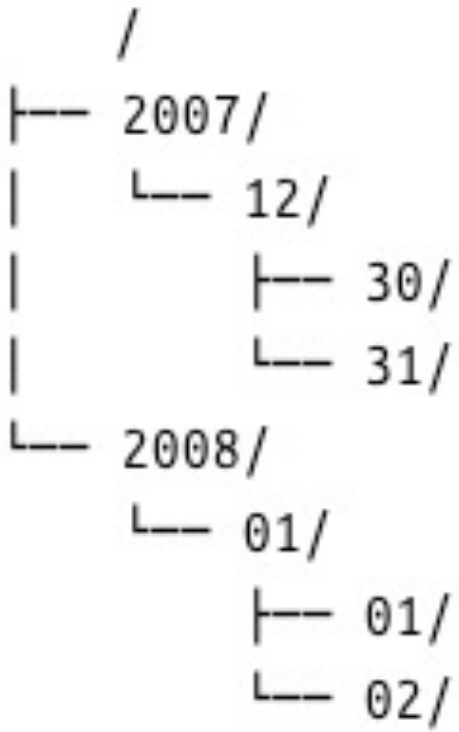
**** Important ****

In the case of removing more than one ”datanode”, Its highly recommended to remove one node at time and run the balancer after the removal.

Hadoop

HDFS

HDFS Globs

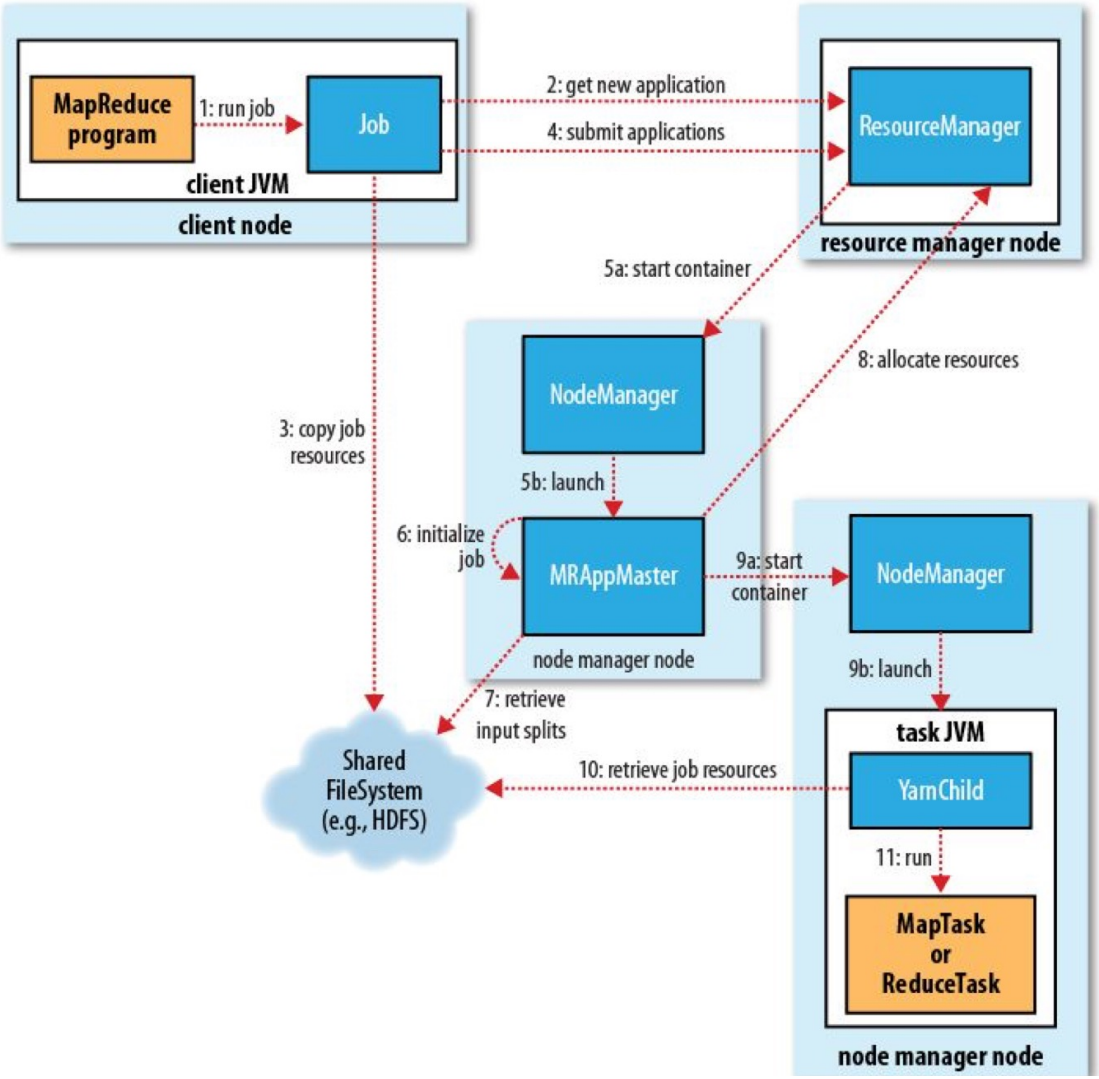


Glob	Expansion
/*	/2007 /2008
/*/*	/2007/12 /2008/01
/*/12/*	/2007/12/30 /2007/12/31
/200?	/2007 /2008
/200[78]	/2007 /2008
/200[7-8]	/2007 /2008
/200[^01234569]	/2007 /2008
/*/*/{31,01}	/2007/12/31 /2008/01/01
/*/*/3{0,1}	/2007/12/30 /2007/12/31
/*/{12/31,01/01}	/2007/12/31 /2008/01/01

Hadoop

Job Submission & Execution

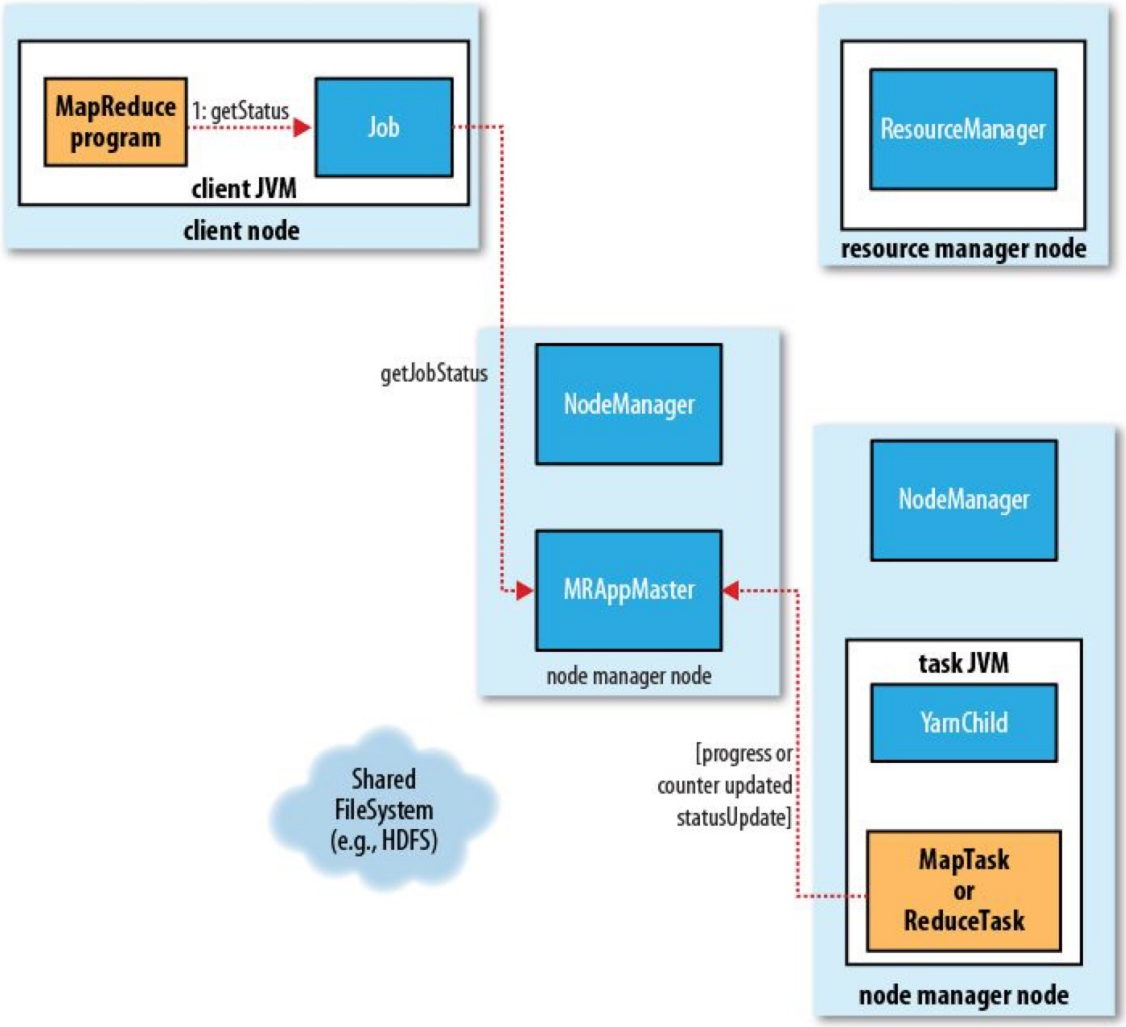
Map/Reduce V2



Hadoop

Job Submission & Execution

Progress V2



Hadoop

Job Submission & Execution

Types of Job Failures

Task Failure

- A task fails, gets restarted and keeps failing until if fails the job.

- Control the number o retries before giving up.

  mapreduce.map.maxattempts

  mapreduce.reduce.maxattempts

- Control acceptable percentage of failure

  mapreduce.map.failures.maxpercent

  mapreduce.reduce.failures.maxpercent

Hadoop

Job Submission & Execution

Types of Job Failures

App Master Failure

- After failure is restarted or recovered.
- Control the number o retries before giving up.

  yarn.resourcemanager.am.max-retries
- Control if retrieve state from previous runs

  yarn.app.mapreduce.am.job.recovery.enable

Hadoop

Job Submission & Execution

Types of Job Failures

Node Manager Failure

- Node manager fails to send heartbeat

- Blacklisted

Hadoop

Job Submission & Execution

Types of Job Failures

Resource Manager Failure

- The most serious

- Property to set a class to handle state

  yarn.resourcemanager.store.class

- Normally kept in memory, but there is also a "Zookeeper" based that retain state.

Job Submission & Execution

Jobs

Job Schedulers

- FIFO – First in First out

  Jobs executed in sequence of scheduling.

  Pros:

  Simple

  Cons:

  One large job blocks everybody.

Hadoop

Job Submission & Execution

Jobs

Job Schedulers

- FIFO – First in First out

  Recent versions allow to set Job priority.

- This only rearranges jobs waiting on the queue.

- If a large job is running other jobs are still blocked.

Job Submission & Execution

Jobs

Job Schedulers

- Fair Scheduler

  Jobs are allocated in pools

- Support preemption of jobs.

  New job comes and forces a running job to give up resources.

- Avoids blocking jobs.

Hadoop

Job Submission & Execution

Jobs

Job Schedulers

- Capacity Scheduler

  Jobs are allocated in “queue”

  Each queue is processed using FIFO w/ Priorities

- Can Create hierarchy.

  Ex: prod/critical prod/normal and office.

- Avoids blocking jobs.

Hadoop

Reduce - Copy

M/R

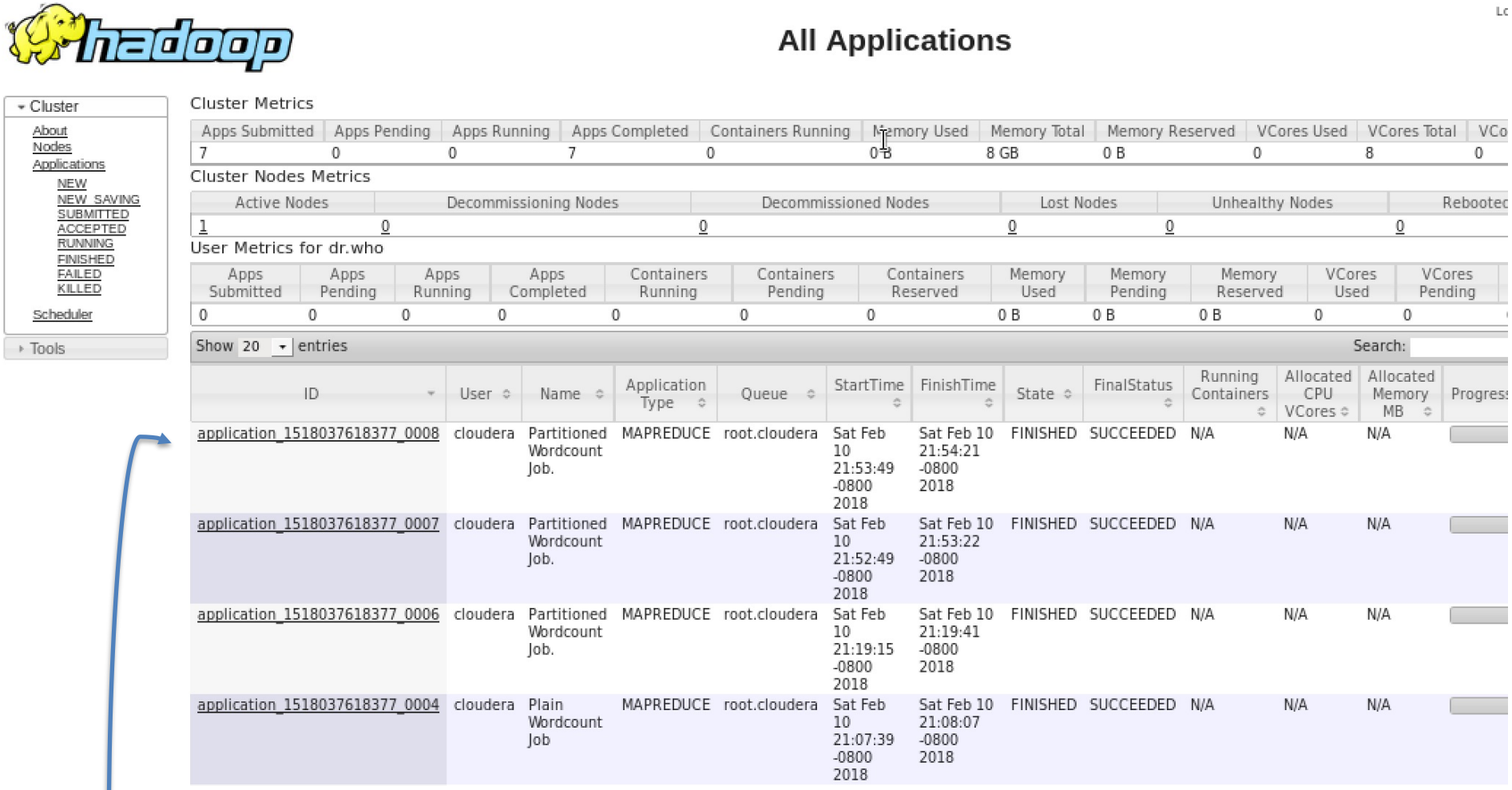
- First stage of reduce phase

Maps complete at different times. So the copy phase starts the before the “Map” phase is complete.

mapred.reduce.parallel.copies: Defines how many threads each reducer daemon can use to pull data from the mapper.

Hadoop navigation

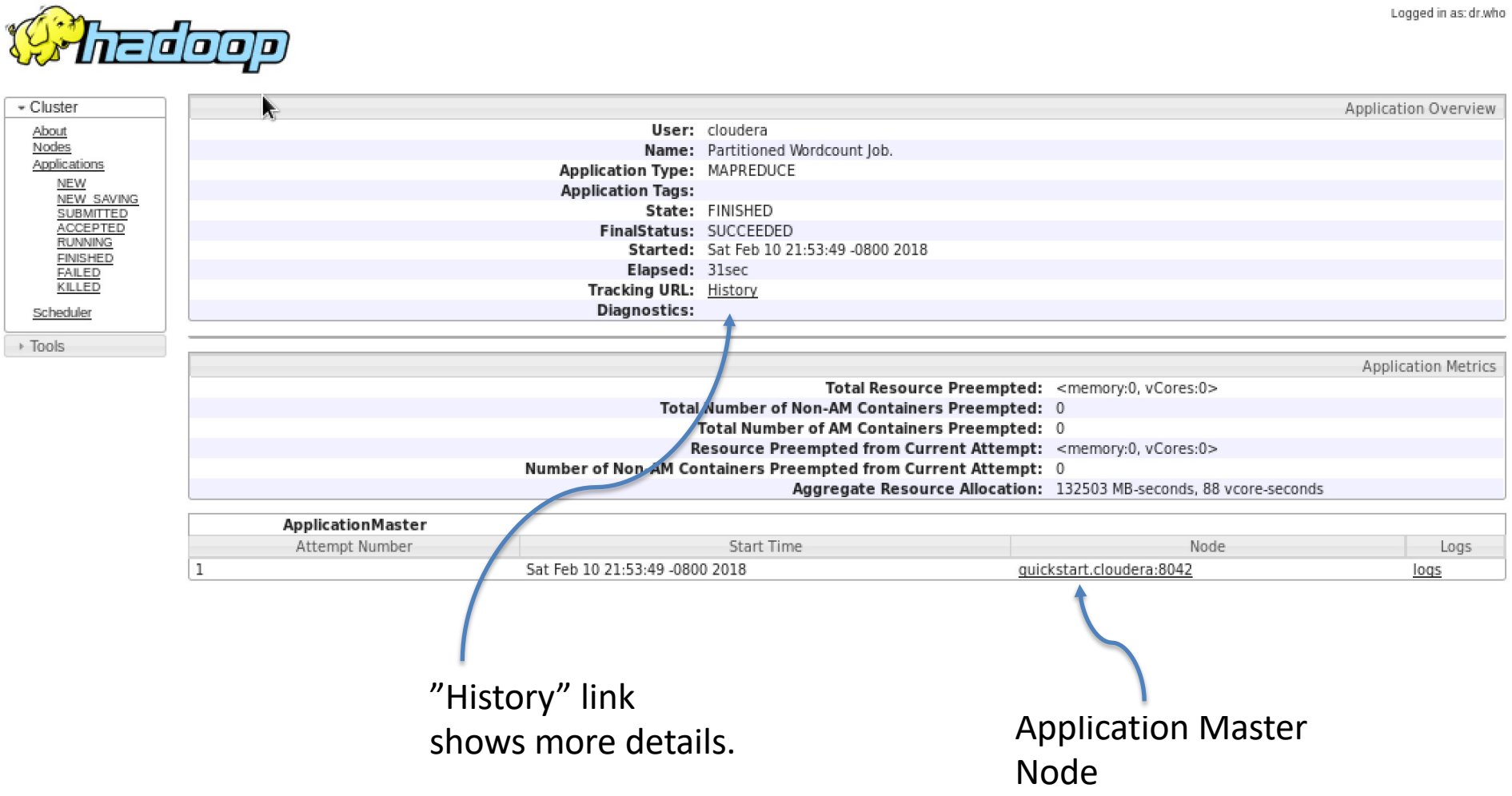
Yarn Web UI



Clicking on one application link takes to the application run page.

Hadoop navigation

Yarn Web UI

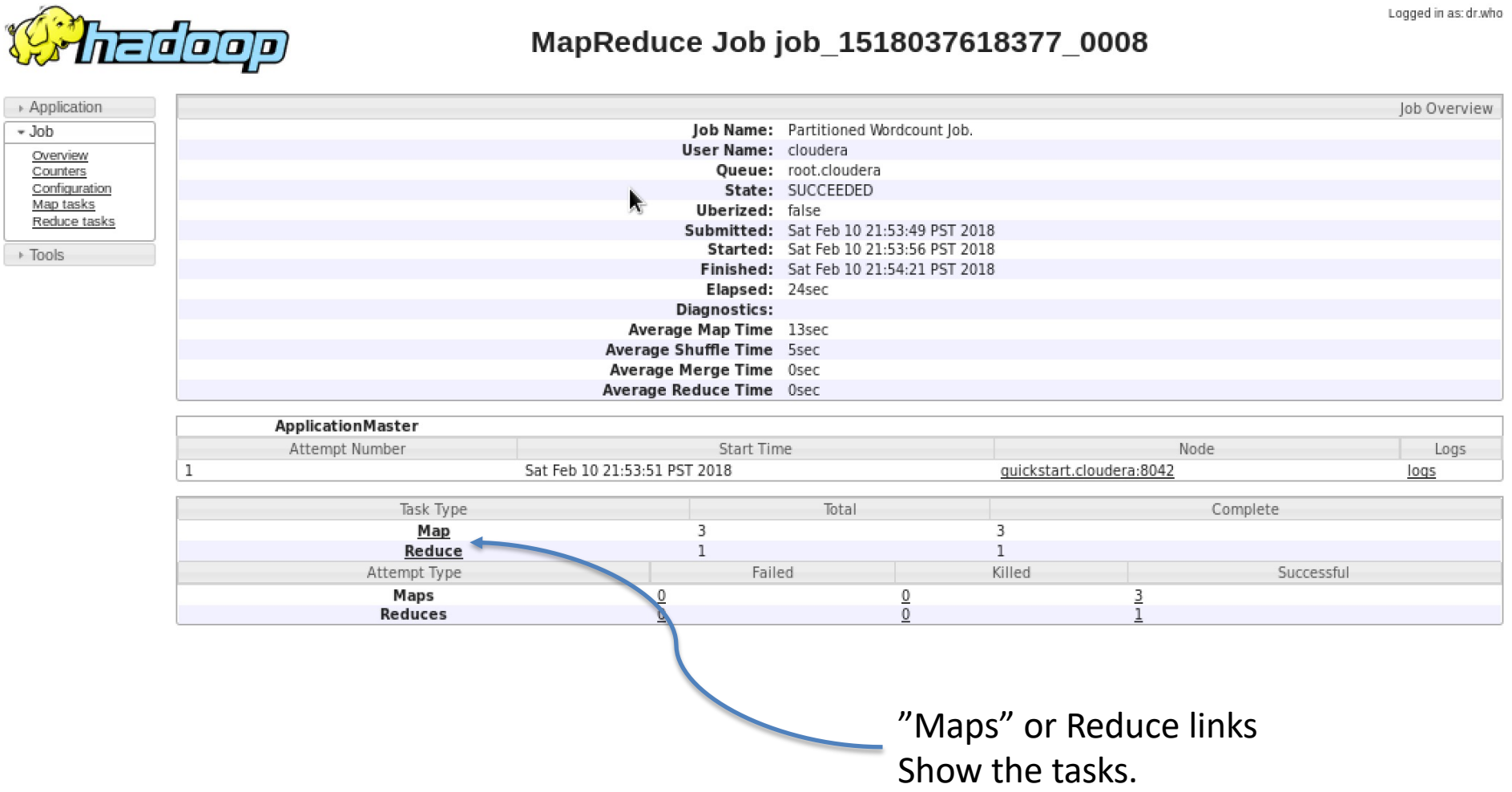


”History” link
shows more details.

Application Master
Node

Hadoop navigation

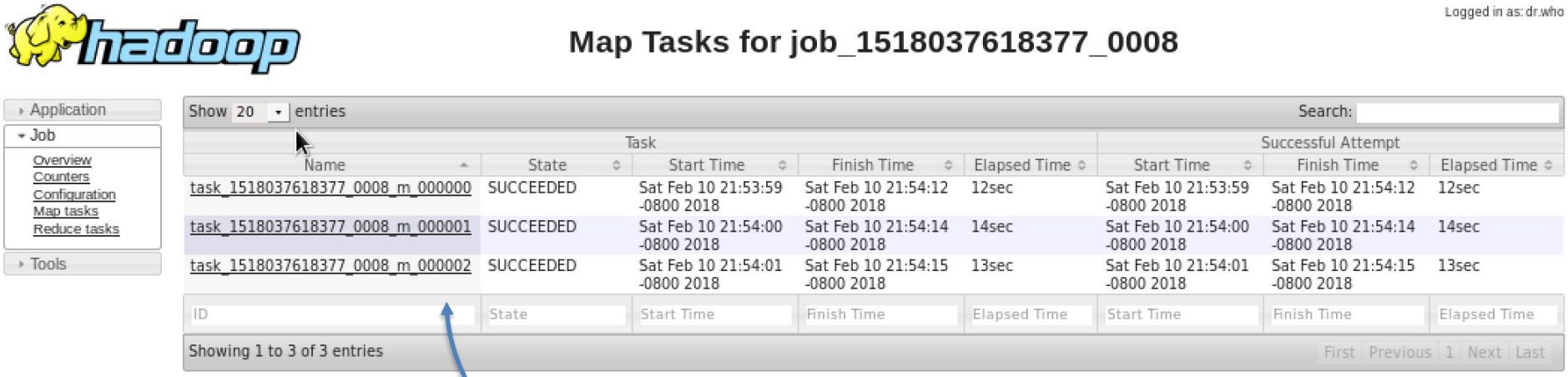
Yarn Web UI



”Maps” or Reduce links
Show the tasks.

Hadoop navigation

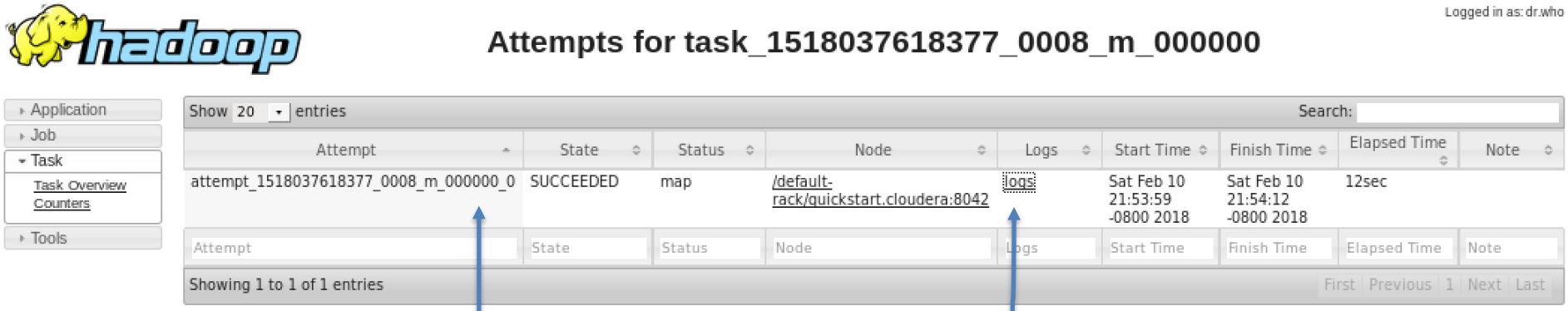
Yarn Web UI



Lists the tasks for a particular map.
Clicking on the “task” link…

Hadoop navigation

Yarn Web UI



We can see the logs for a particular attempt on the ‘logs’ Link.

Shows information about task execution attempts.
In this case the first attempt succeed (end with ‘_0’)

Hadoop navigation

Yarn Web UI

Notice that in a real cluster these logs are being produced in the node that executed that attempt.

Q & A