Hadoop: Past and Present

Andrew Wang | Software Engineer

April 29, 2013

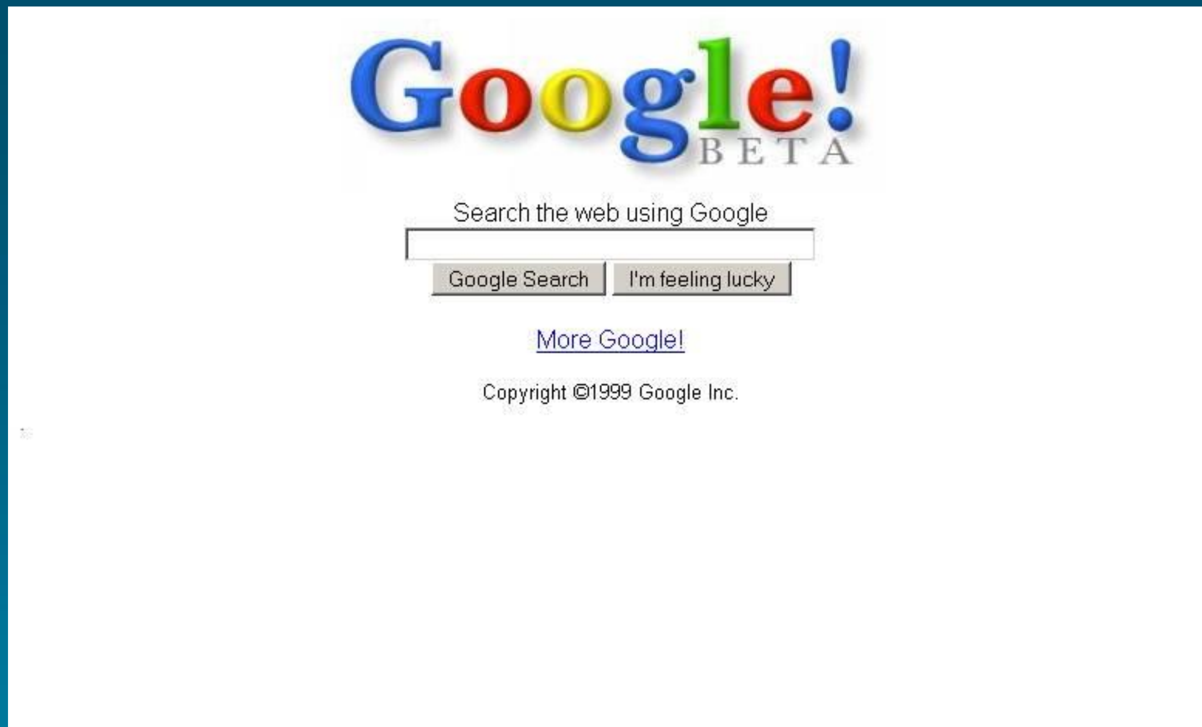
About Me

- 2010: BS CS UVa
- 2012: MS CS UC Berkeley
  - AMP Lab alumni
  - Advised by Ion Stoica
- Now: HDFS team at Cloudera

Outline

- State of databases in 1999
  - Why is Hadoop displacing DB technology?
- Core stack
  - HDFS and MapReduce
- Rest of the Hadoop ecosystem
  - HBase, Pig, Hive, Oozie, Zookeeper, Flume, Impala, …

1999!


Google!
BETA
Search the web using Google
Google Search
I'm feeling lucky
More Google!
Copyright ©1999 Google Inc.

Indexing the Web

- Web is huge
  - Hundreds of millions of pages in 1999
- How do you index it?
  - Crawl all the pages
  - Rank pages based on relevance metrics
  - Build search index of keywords to pages
  - Do it in real-time!

SYBASE

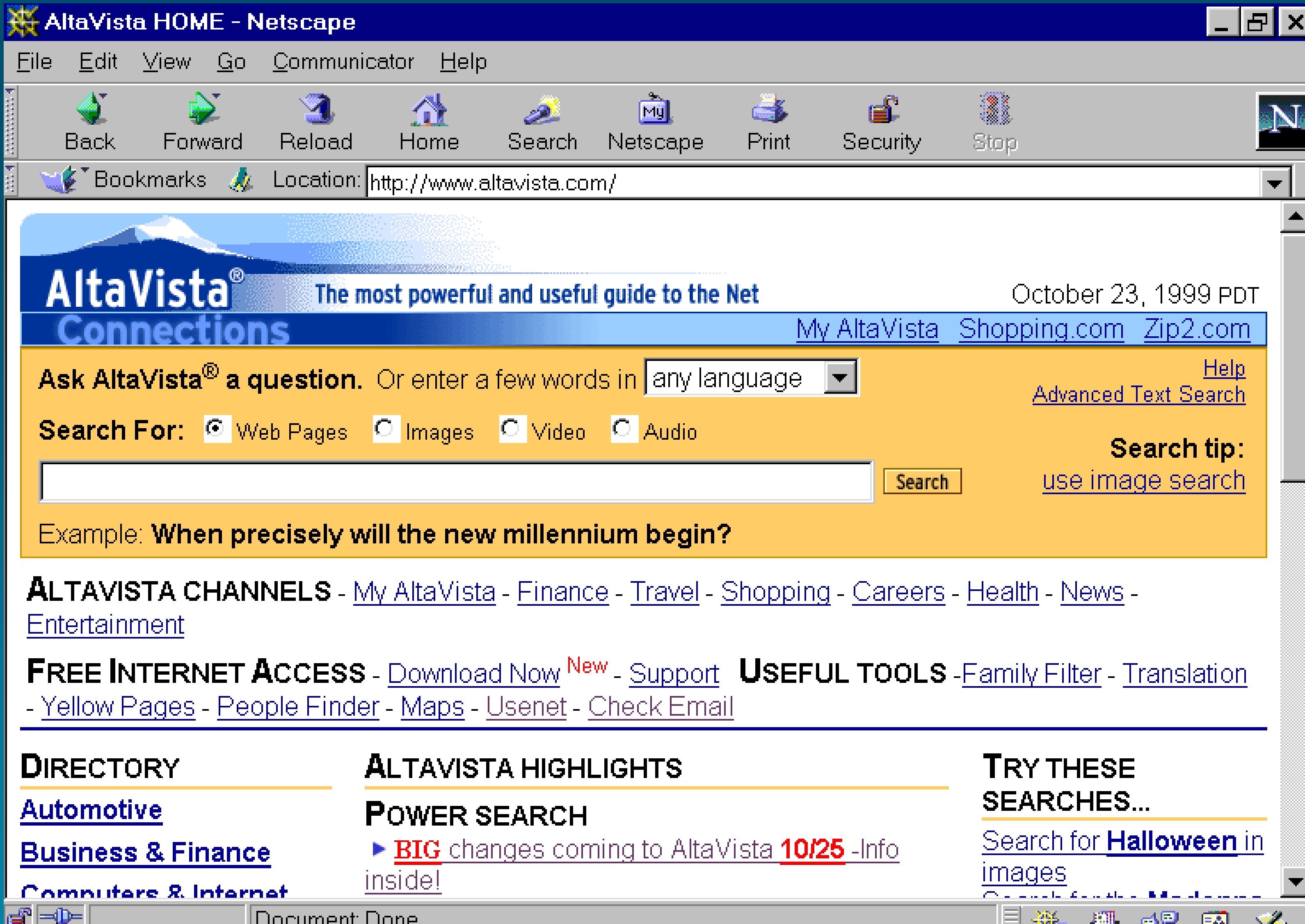
ORACLE

Informix

IBM

Databases in 1999

1. Buy a really big machine
2. Install an expensive DBMS on it
3. Point your workload at it
4. Hope it doesn't fail
5. Ambitious: buy another really big machine as a backup


AltaVista HOME - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: http://www.altavista.com/
AltaVista® Connections
The most powerful and useful guide to the Net
October 23, 1999 PDT
My AltaVista Shopping.com Zip2.com
Ask AltaVista® a question. Or enter a few words in any language
Help
Advanced Text Search
Search For: Web Pages Images Video Audio
Search
Search tip: use image search
Example: When precisely will the new millennium begin?
ALTAVISTA CHANNELS - My AltaVista - Finance - Travel - Shopping - Careers - Health - News - Entertainment
FREE INTERNET ACCESS - Download Now New - Support USEFUL TOOLS -Family Filter - Translation - Yellow Pages - People Finder - Maps - Usenet - Check Email
DIRECTORY
Automotive
Business & Finance
ALTAVISTA HIGHLIGHTS
POWER SEARCH
BIG changes coming to AltaVista 10/25 -Info inside!
TRY THESE SEARCHES...
Search for Halloween in images
Document: Done

Database Limitations

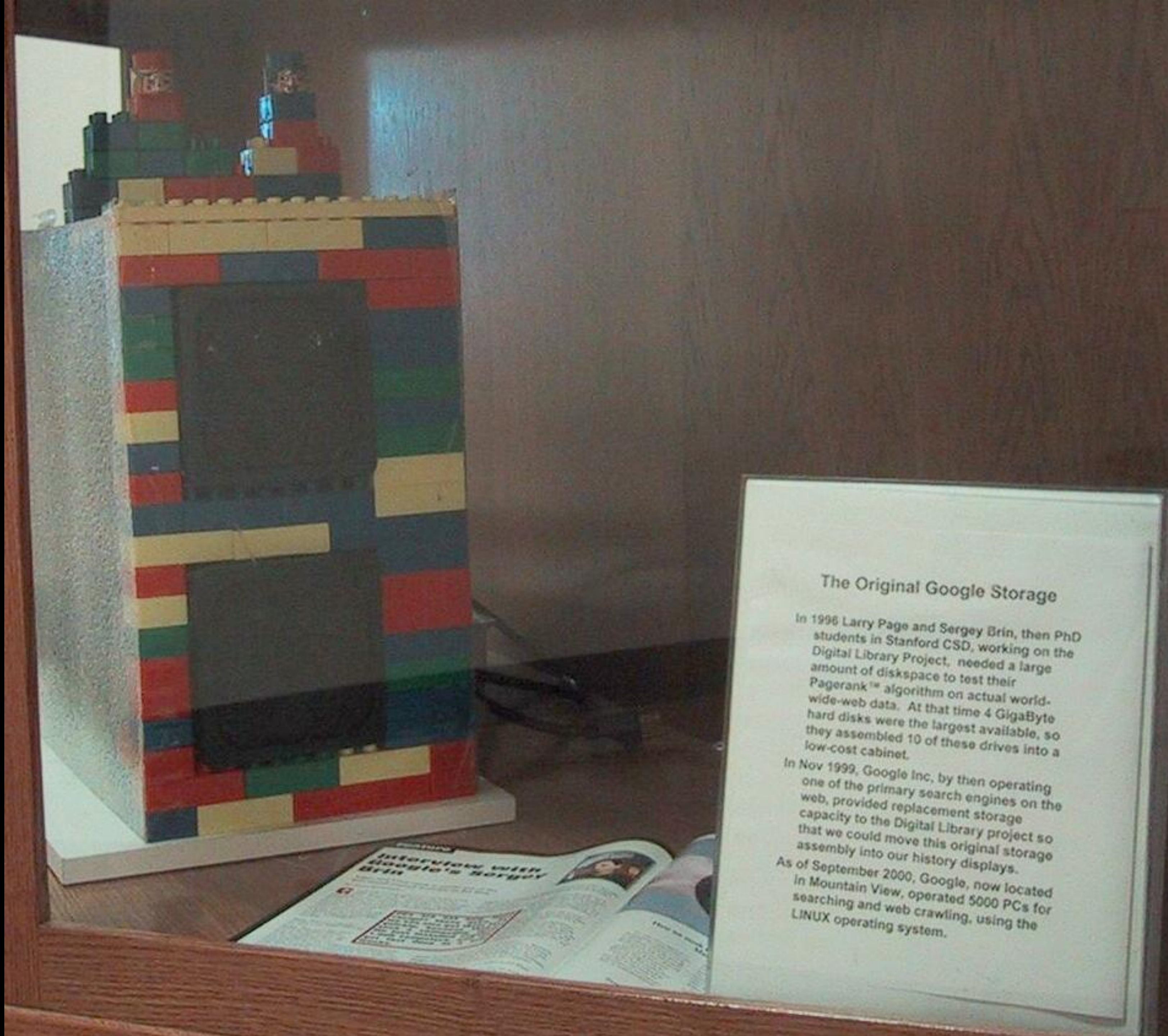
- Didn't scale horizontally
  - High marginal cost ($$$)
- No real fault-tolerance story
- Vendor lock in ($$$)
- SQL unsuited for search ranking
  - Complex analysis (PageRank)
  - Unstructured data

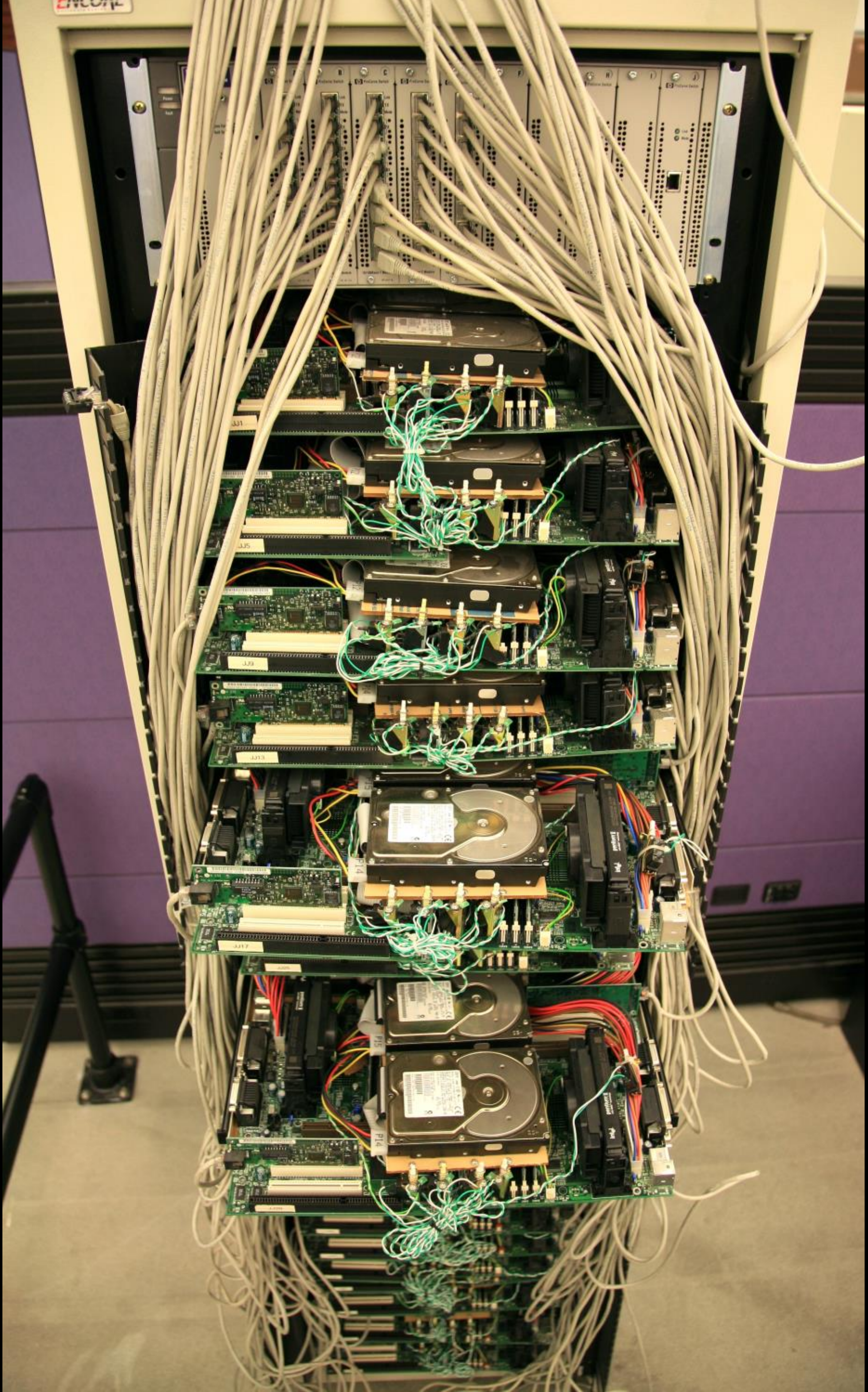
Google Does Something Different

- Designed their own storage and processing infrastructure
  - Google File System and MapReduce
- Goals: KISS
  - Cheap
  - Scalable
  - Reliable

Google Does Something Different

- It worked!
- Powered Google Search for many years
- General framework for large-scale batch computation tasks
- Still used internally at Google to this day


Interview with Google's Sergey Brin
The Original Google Storage
In 1996 Larry Page and Sergey Brin, then PhD students in Stanford CSD, working on the Digital Library Project, needed a large amount of diskspace to test their Pagerank™ algorithm on actual world-wide-web data. At that time 4 GigaByte hard disks were the largest available, so they assembled 10 of these drives into a low-cost cabinet.
In Nov 1999, Google Inc, by then operating one of the primary search engines on the web, provided replacement storage capacity to the Digital Library project so that we could move this original storage assembly into our history displays.
As of September 2000, Google, now located in Mountain View, operated 5000 PCs for searching and web crawling, using the LINUX operating system.




102662167

Google's messages from the future

- Google was benevolent enough to publish
  - 2003: Google File System (GFS) paper
  - 2004: MapReduce paper
- Already mature technologies at this point

Google's messages from the future

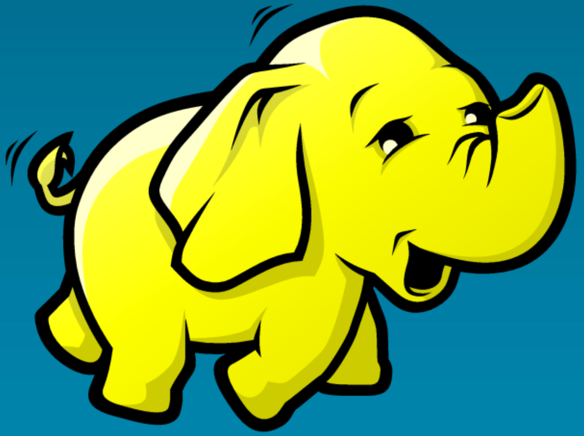
- Community didn't get it immediately
  - DB people thought it was silly
  - Non-Google weren't at the same scale yet
- Google had little interest in releasing GFS and MapReduce
  - Business was ads, not infrastructure

Birth of Hadoop

- Doug Cutting and Mike Cafarella
- Nutch
  - Open-source search platform
- Ran into scaling issues
  - 4 nodes
  - Hard to program
  - Hard to manage
- Immediate application for GFS and MR

Birth of Hadoop

- 2004-2006: Implemented GFS/MR and ported Nutch to it
- 2006: Spun out into Apache Hadoop
- Name of Doug's son's stuffed elephant



Birth of Hadoop

Summary

- The web is huge and unstructured
- Databases didn't fit the problem
  - Didn't scale, expensive, SQL limitations
- Google did their own thing: GFS + MR
- Hadoop is based on the Google papers

HDFS and MapReduce

HDFS

- Based on GFS
- Distributed, fault-tolerant filesystem
- Primarily designed for cost and scale
  - Works on commodity hardware
  - 20PB / 4000 node cluster at Facebook

HDFS design assumptions

- Failures are common
  - Massive scale means more failures
  - Disks, network, node
- Files are append-only
- Files are large (GBs to TBs)
- Accesses are large and sequential

Quick primers

- Filesystems
- Hard drives
- Datacenter networking

Quick filesystem primer

- Same concepts as the FS on your laptop
  - Directory tree
  - Create, read, write, delete files
- Filesystems store metadata and data
  - Metadata: filename, size, permissions, …
  - Data: contents of a file
- Other concerns
  - Data integrity, durability, management

Quick disk primer

- Disk does a seek for each I/O operation
- Seeks are expensive (~10ms)
- Throughput / IOPS tradeoff
  - 100 MB/s and 10 IOPS
  - 10MB/s and 100 IOPS
- Big I/Os mean better throughput

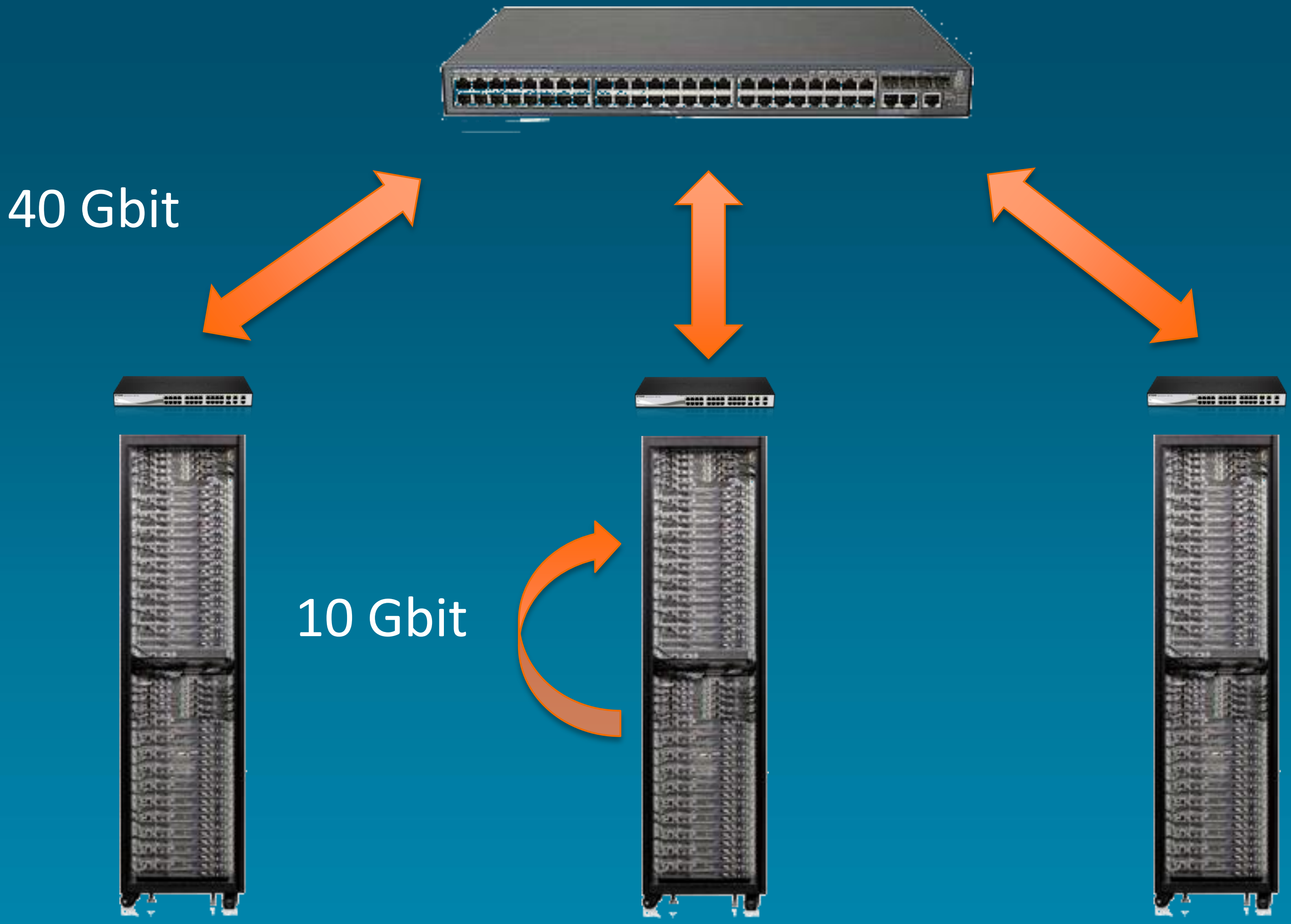
Quick networking primer

Core switch

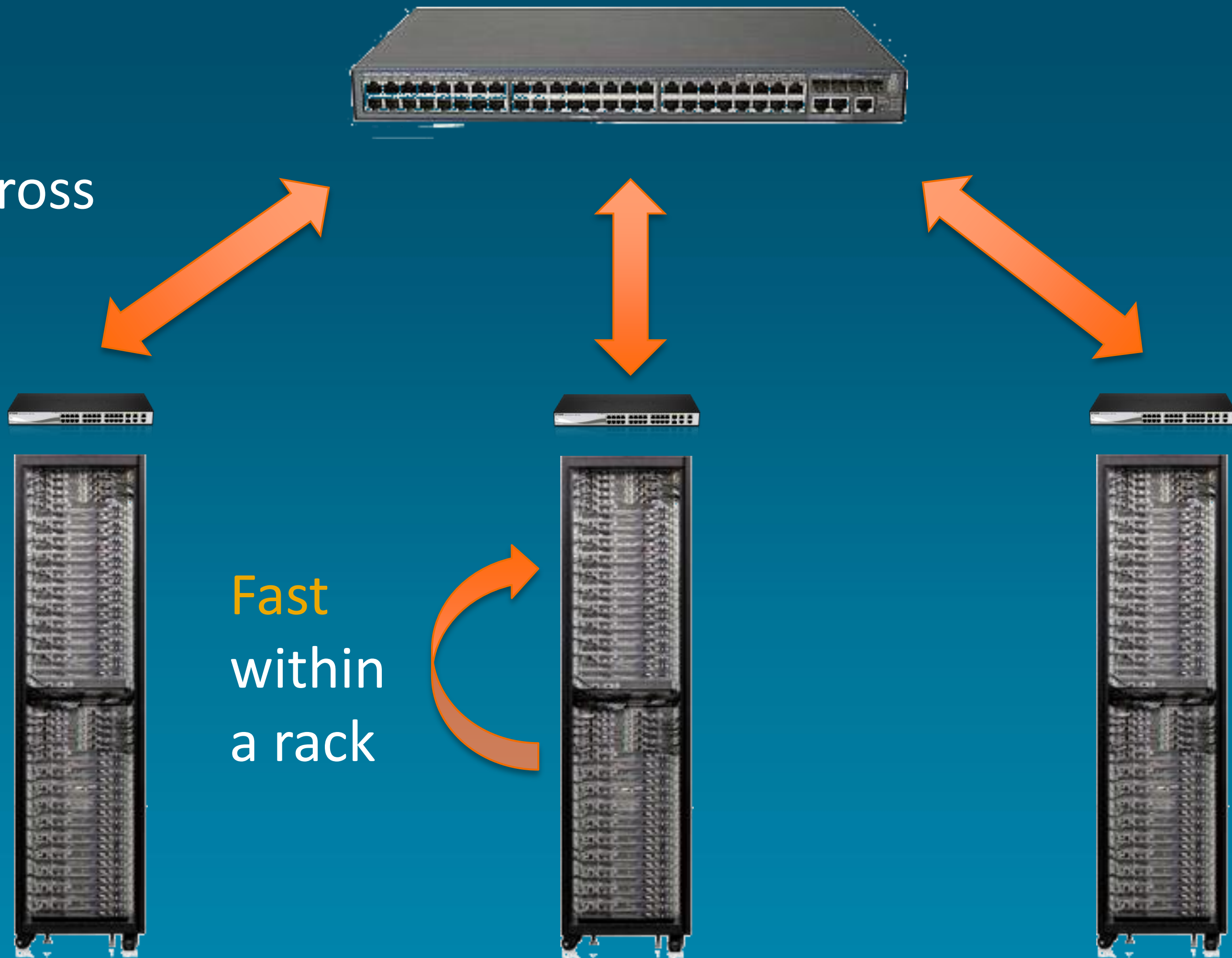
Top-of-rack switch

Rack

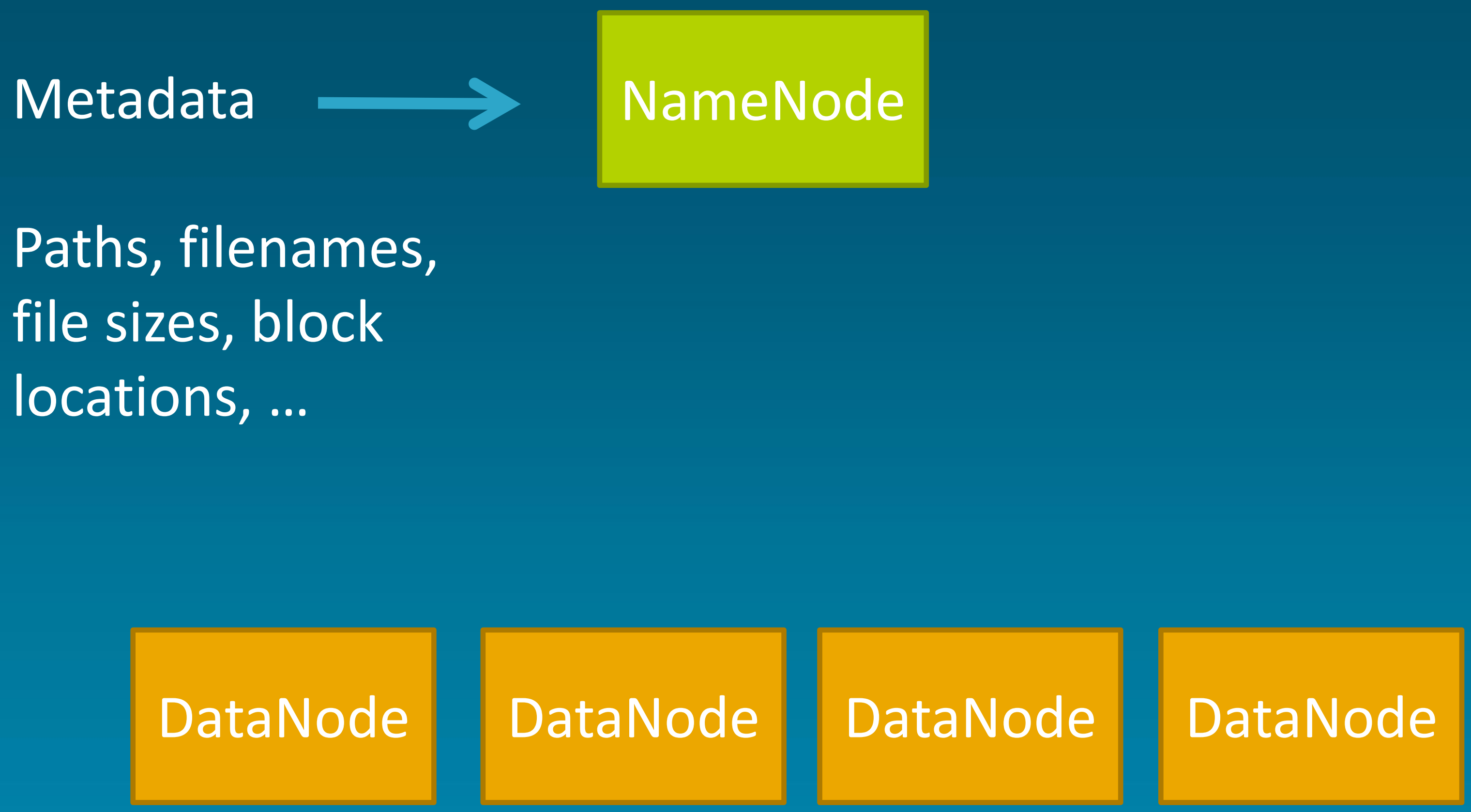
Quick networking primer



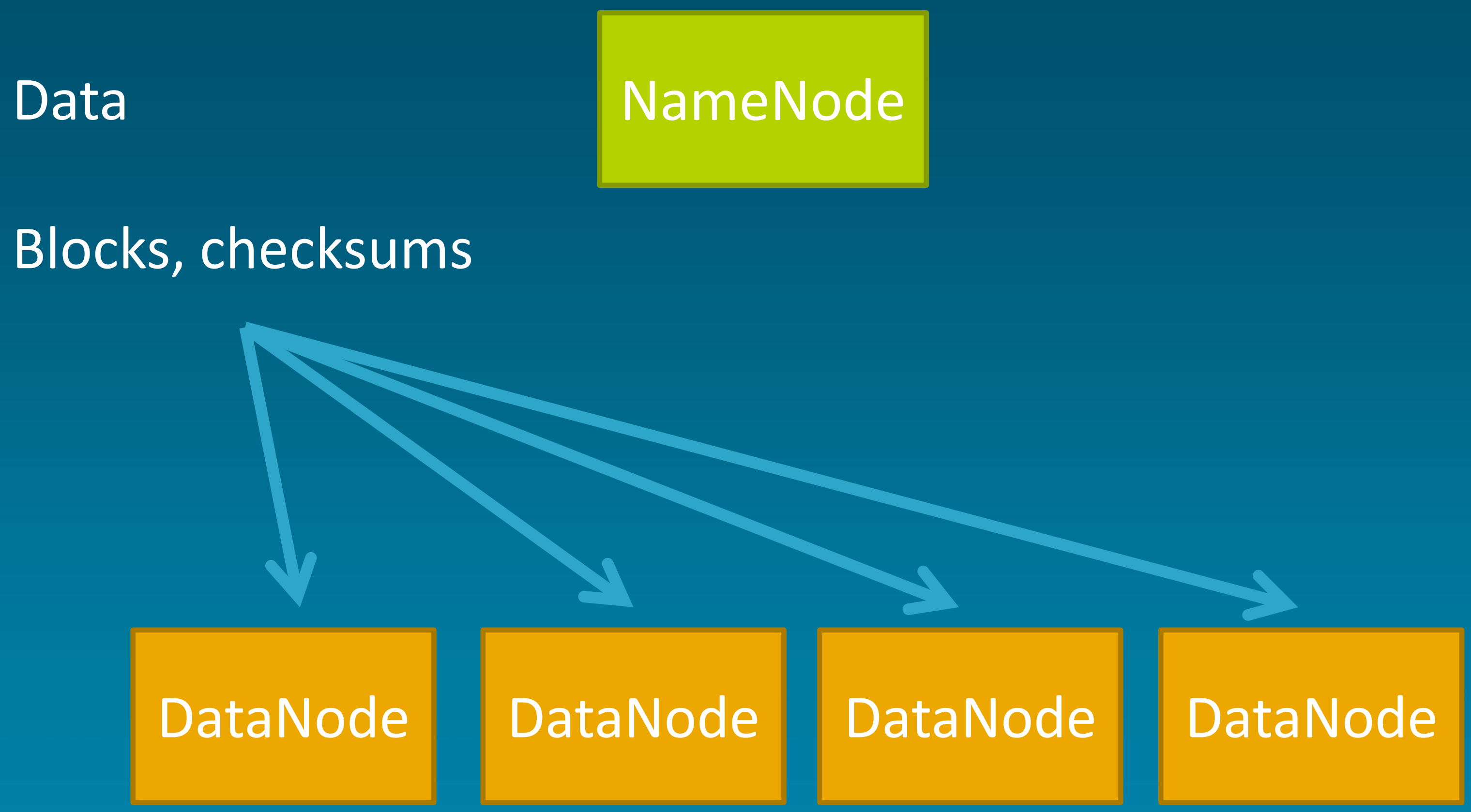
Quick networking primer



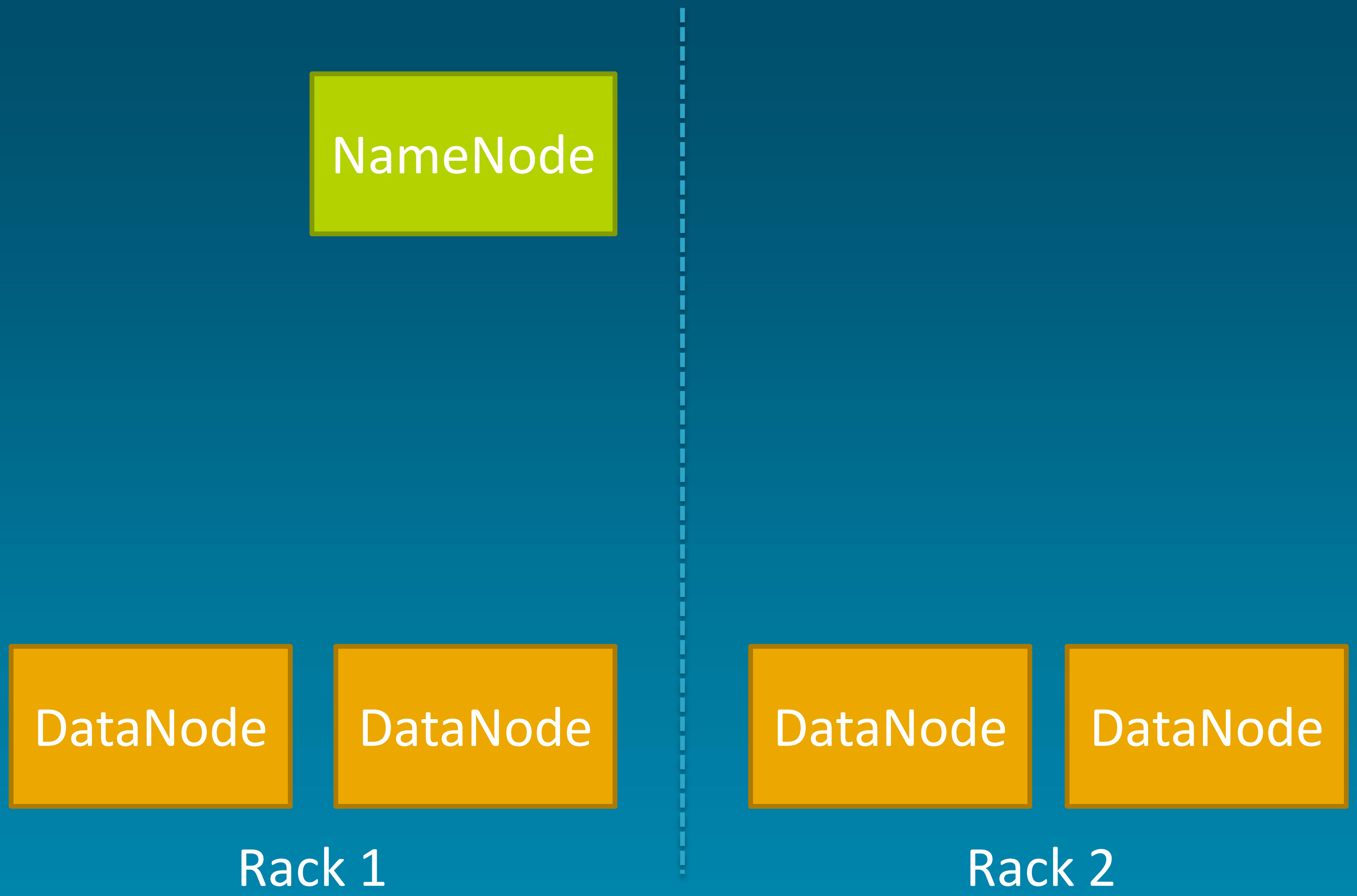
HDFS Architecture



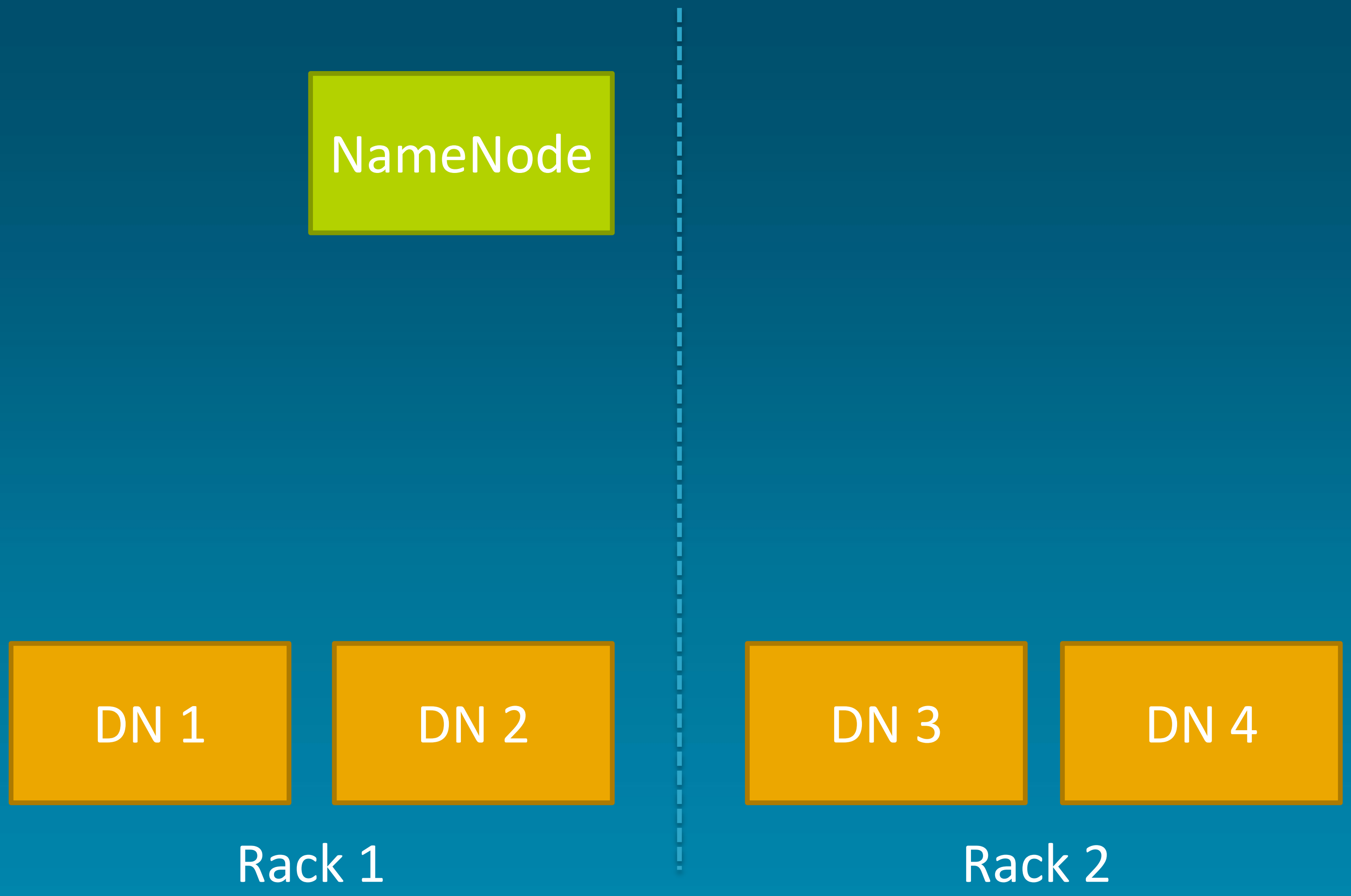
HDFS Architecture



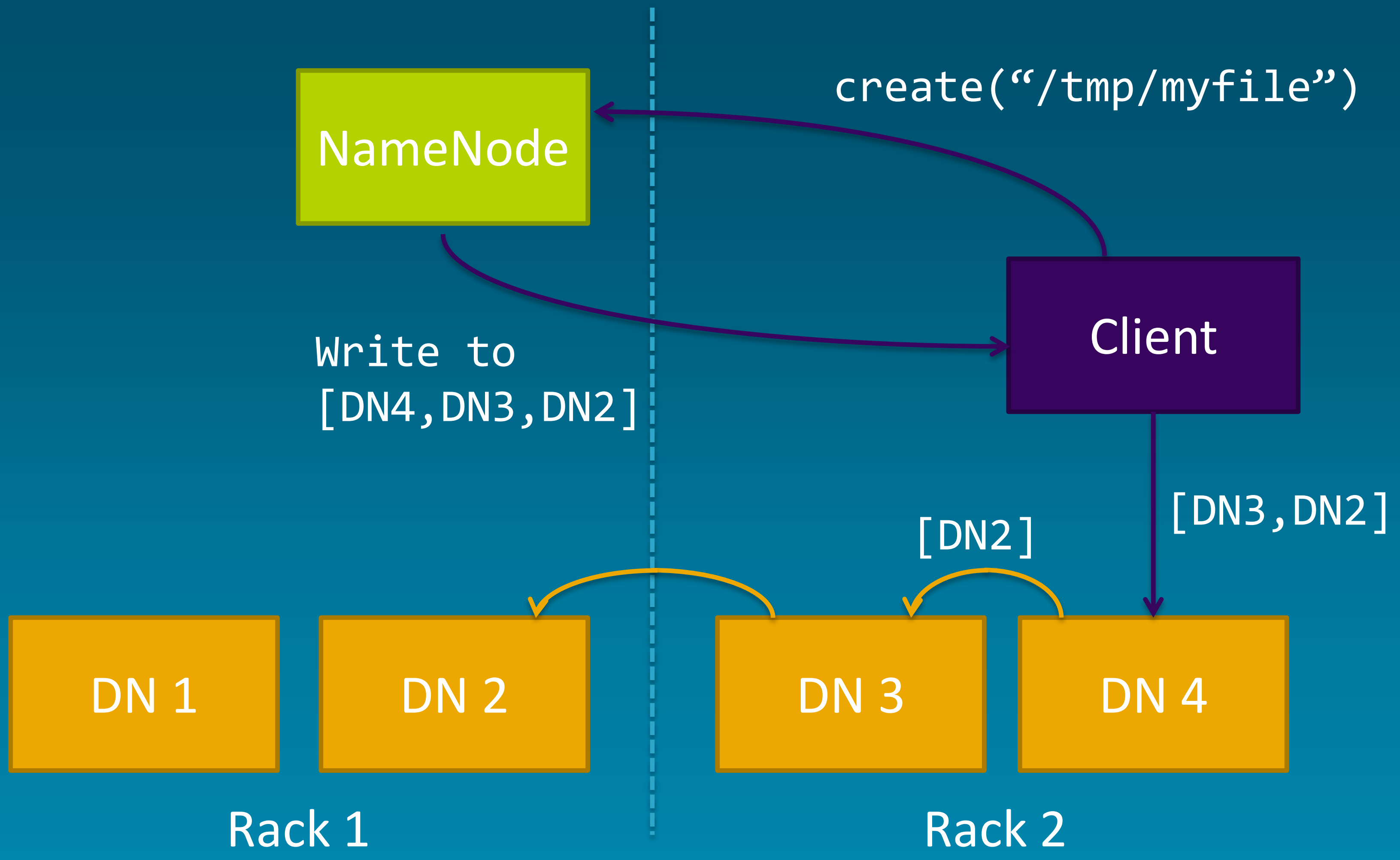
HDFS Architecture



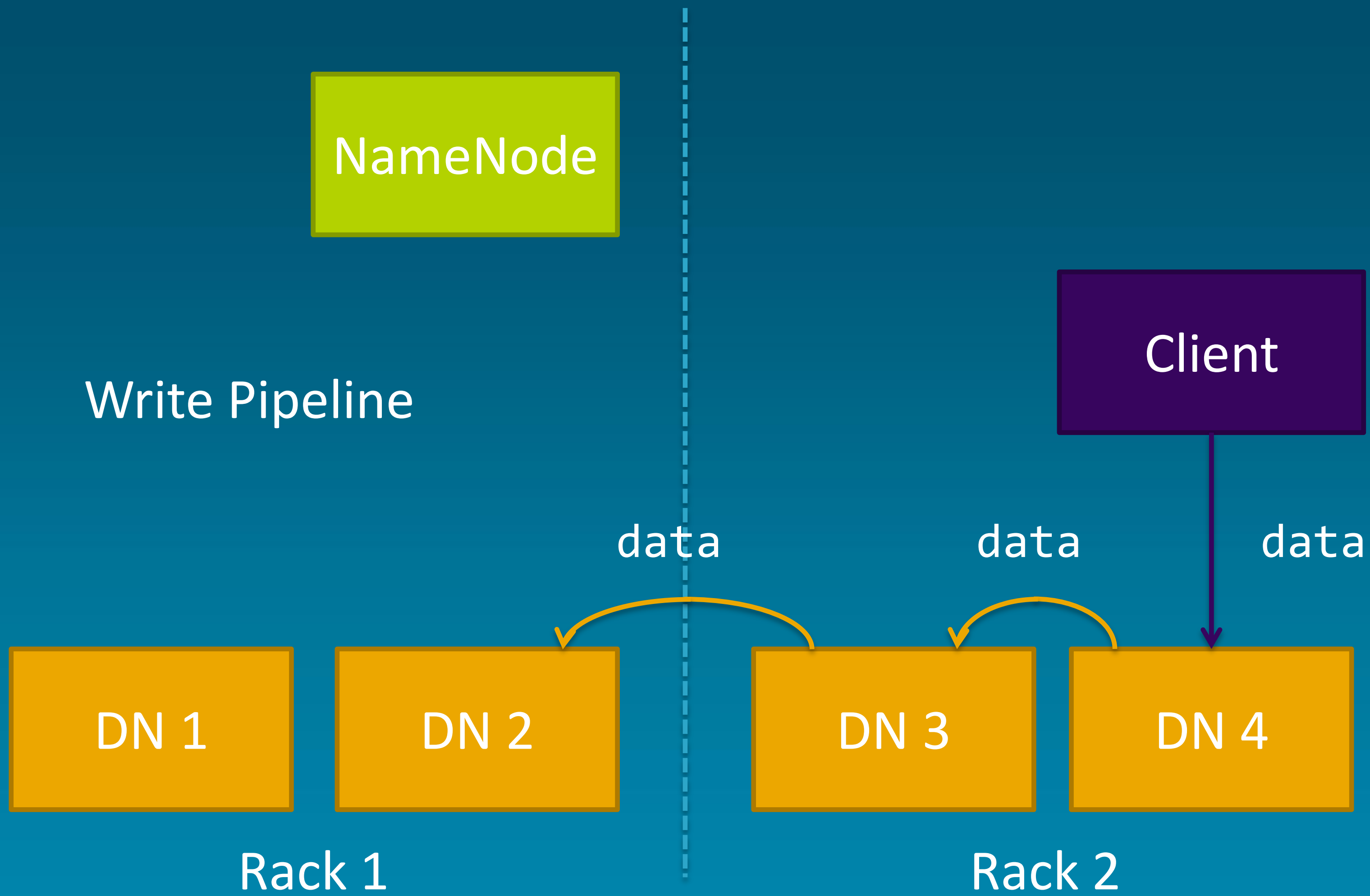
HDFS Architecture



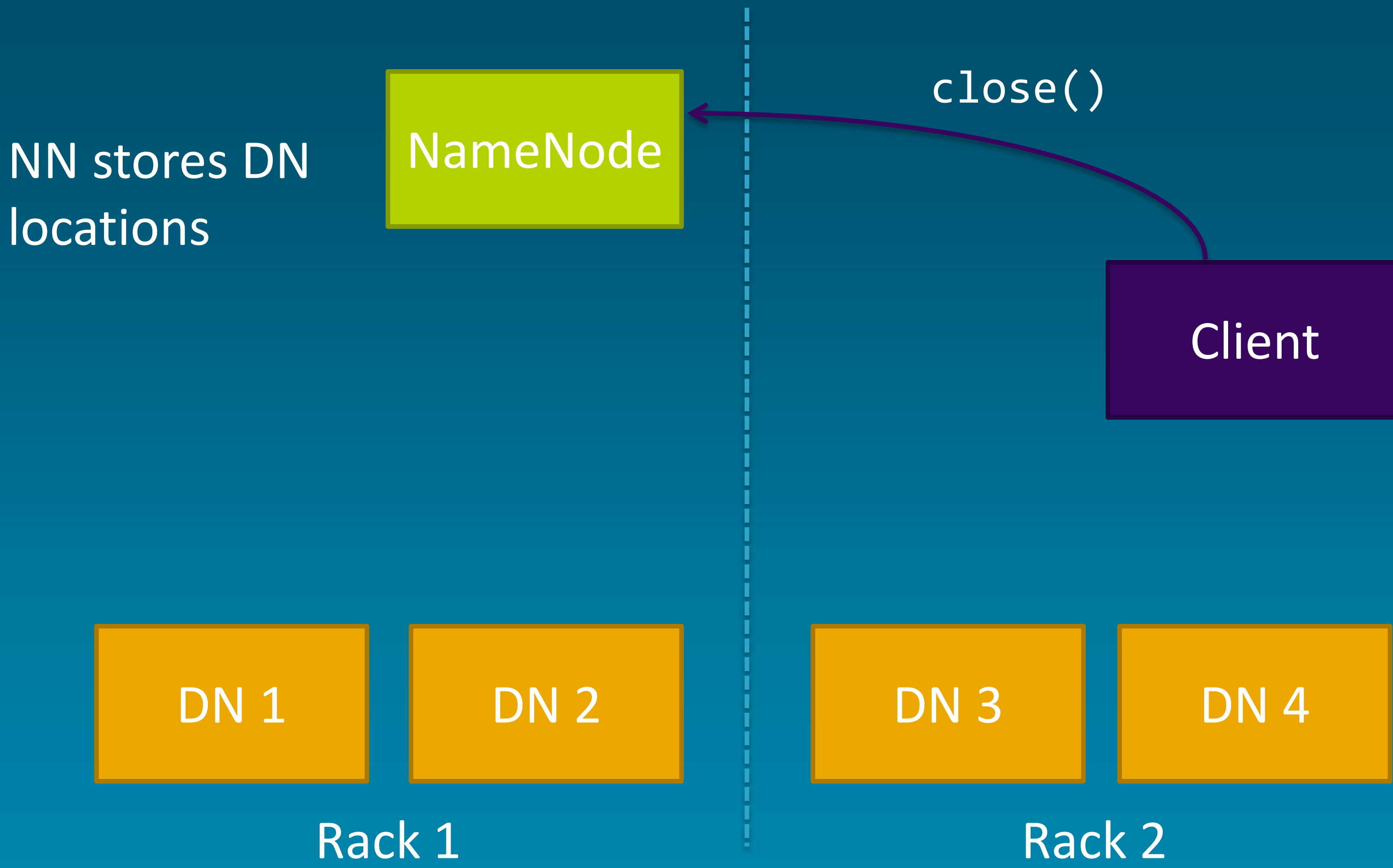
HDFS Write Path



HDFS Write Path



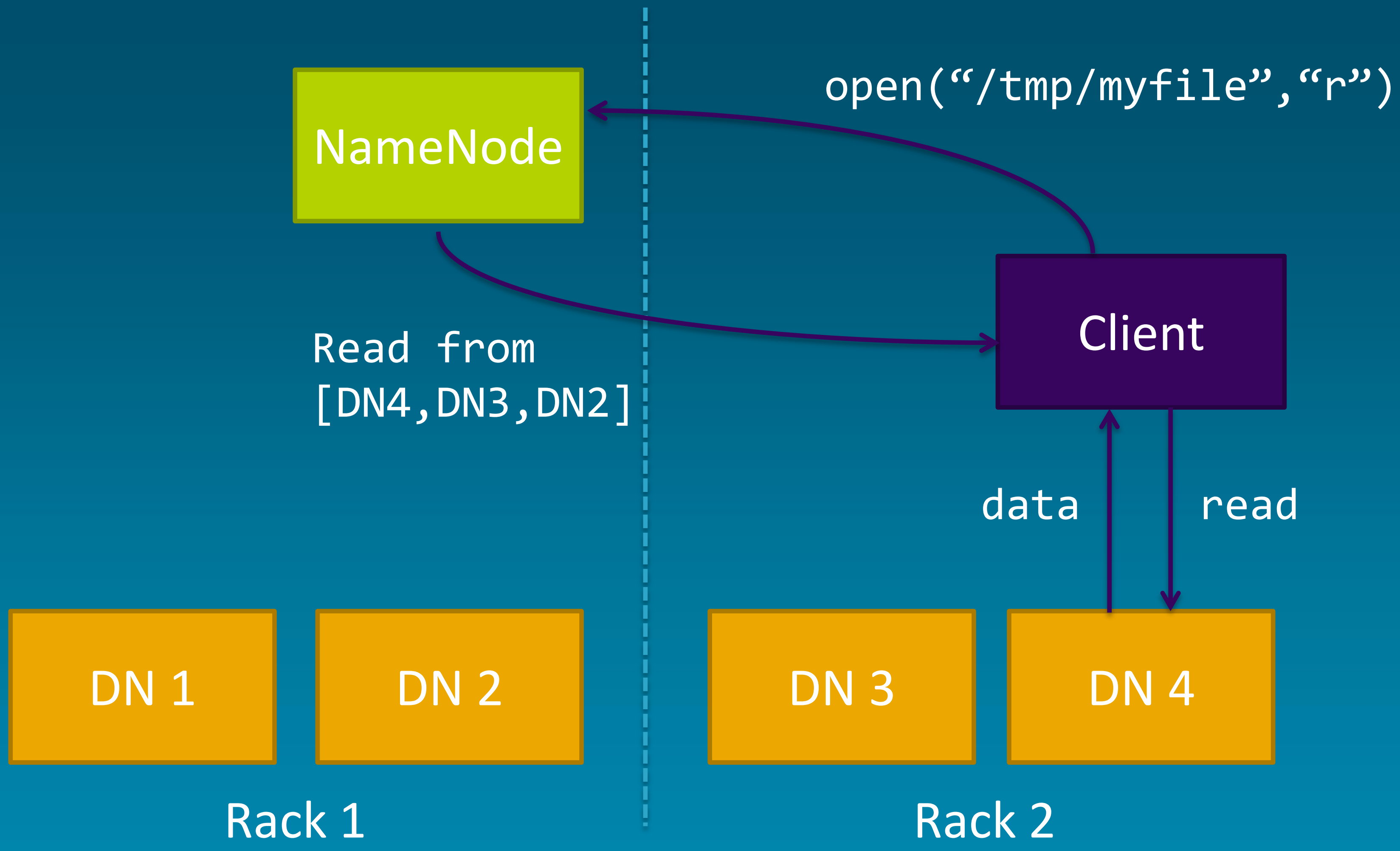
HDFS Write Path



HDFS Write Path

- Talk to NameNode
  - Store metadata for new file
  - Get topology-aware list of DataNodes
- Setup the write pipeline
- Stream data to pipeline
- Tell NameNode when done

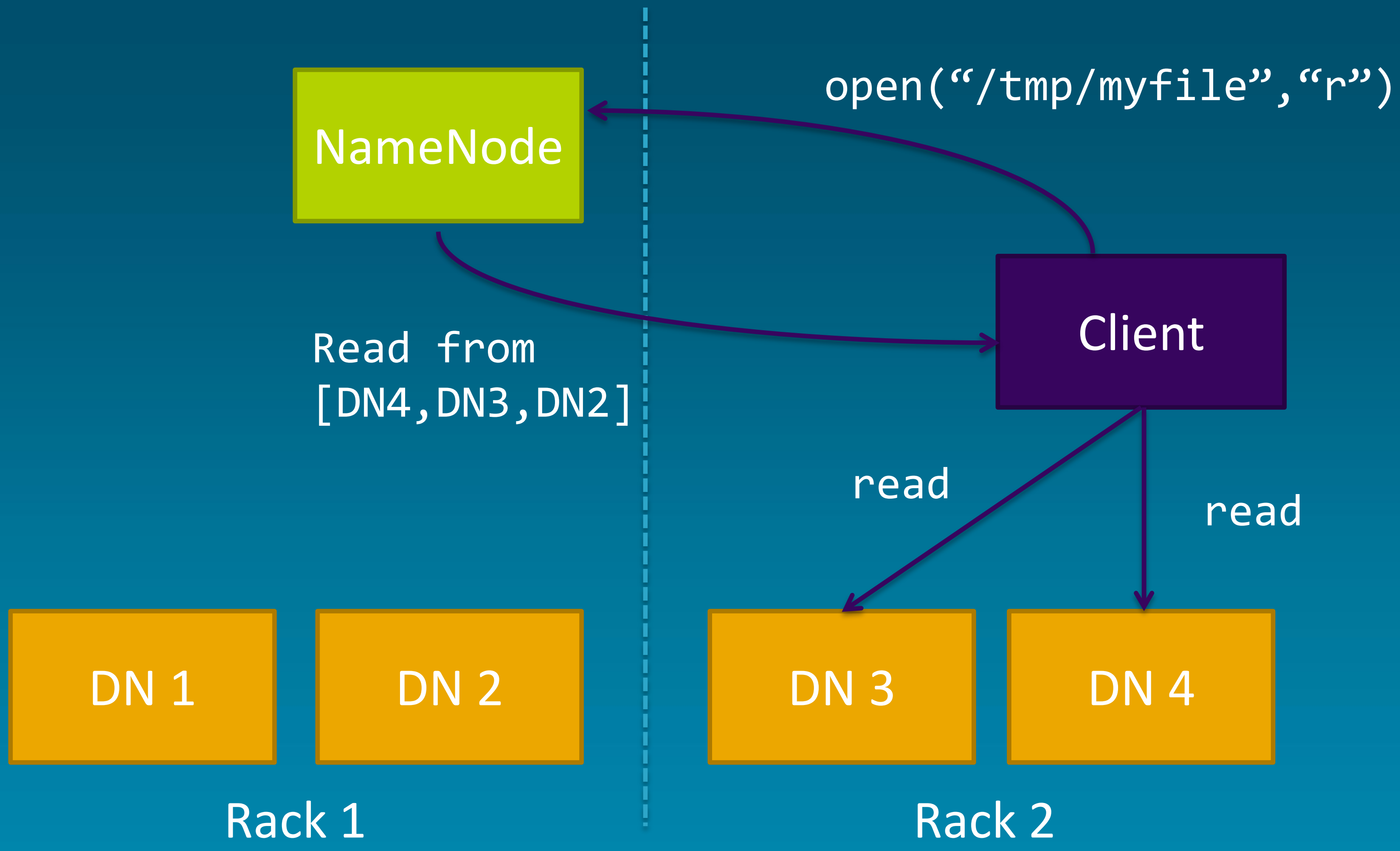
HDFS Read Path



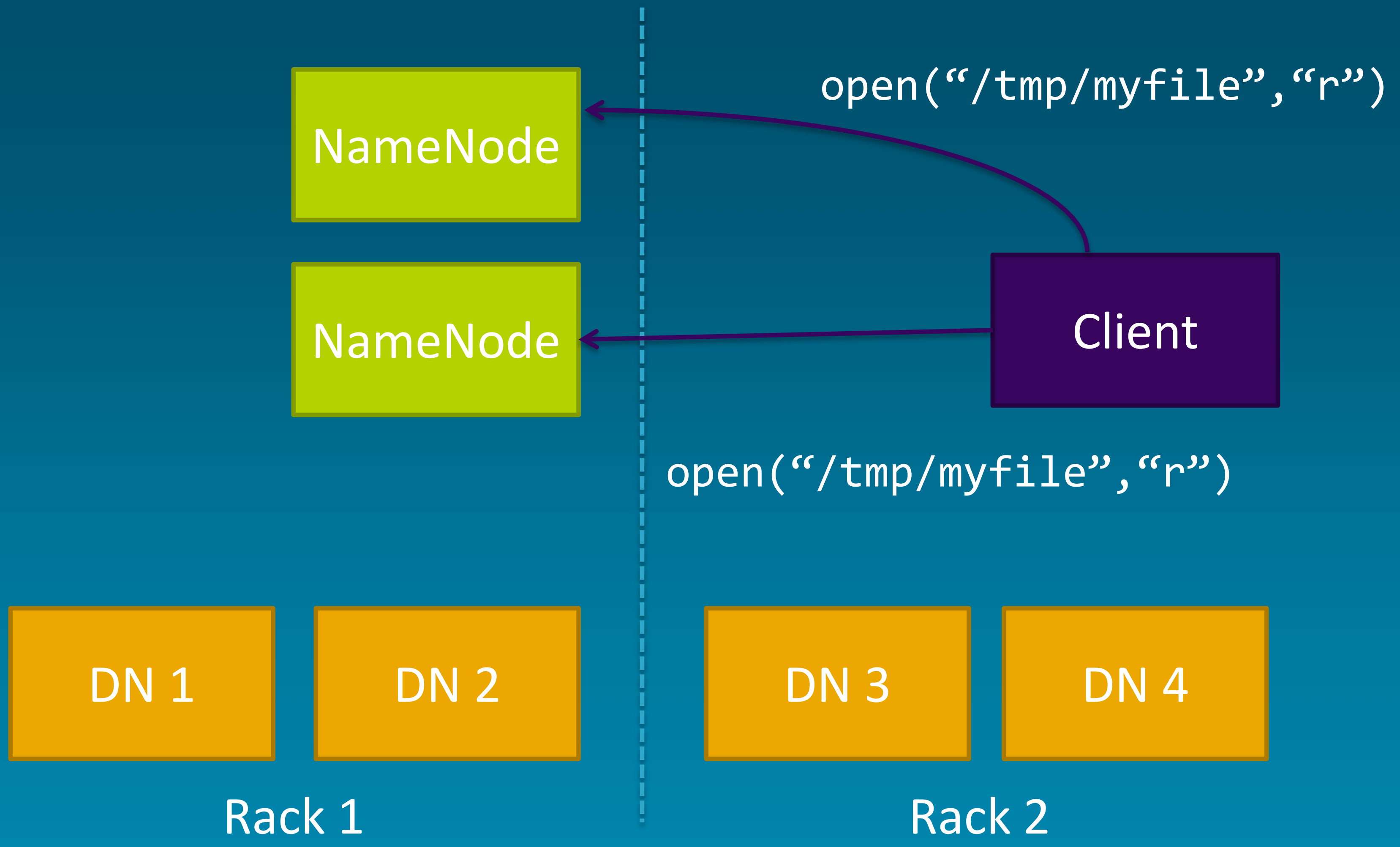
HDFS Fault-tolerance

- Many different failure modes
  - Disk corruption, node failure, switch failure
- Primary concern
  - Data is safe!!!
- Secondary concerns
  - Keep accepting reads and writes
  - Do it transparently to clients

HDFS DataNode Failure



HDFS NameNode Failure



Other HDFS features

- NameNode federation
- Storage block pools
- Snapshots (new!)
- Future
  - Hierarchical storage management
  - Quality-of-Service
  - NameNode and DataNode scalability

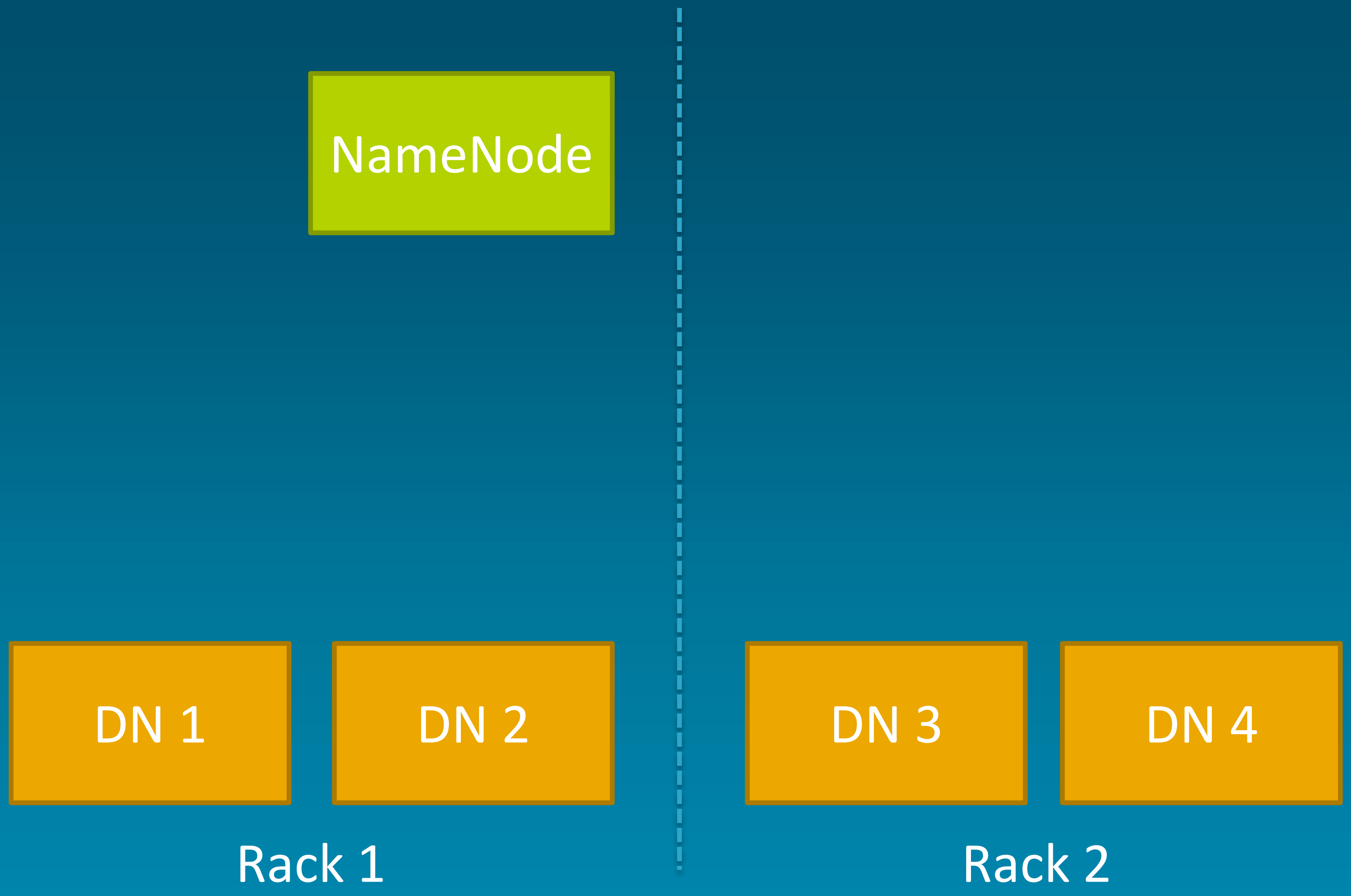
MapReduce

- Programming and execution framework
- Taken from functional programming
  - Map – operate on every element
  - Reduce – combine and aggregate results
- Abstracts storage, concurrency, execution
  - Just write two Java functions
  - Contrast with MPI

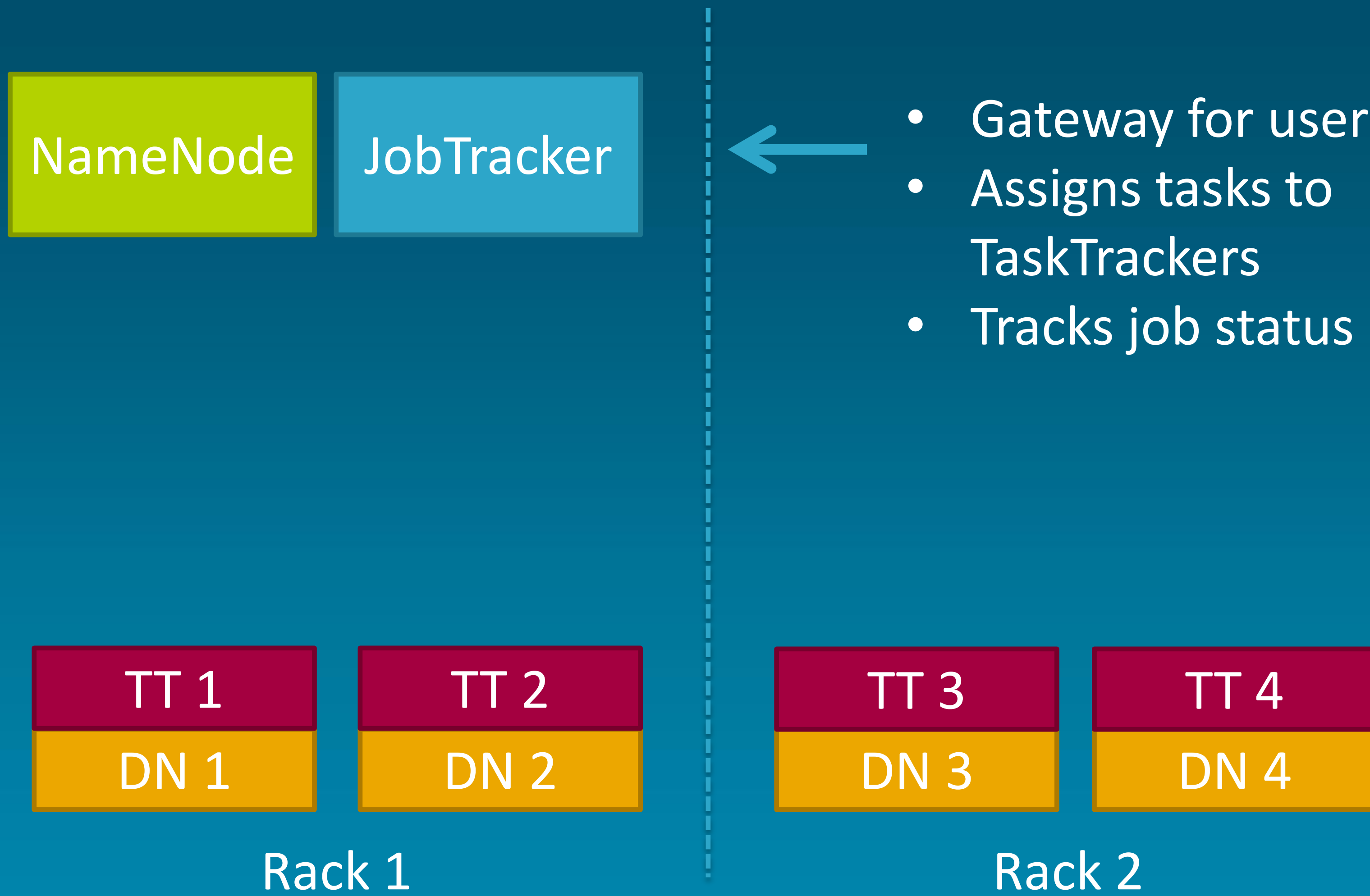
MapReduce

- Constrained, but general
  - Can do custom ML not possible in SQL
  - Not as efficient as a DB for some queries
- No update in place
  - Take data in, transform, write new data out
  - Makes fault-tolerance easier

MapReduce Architcture



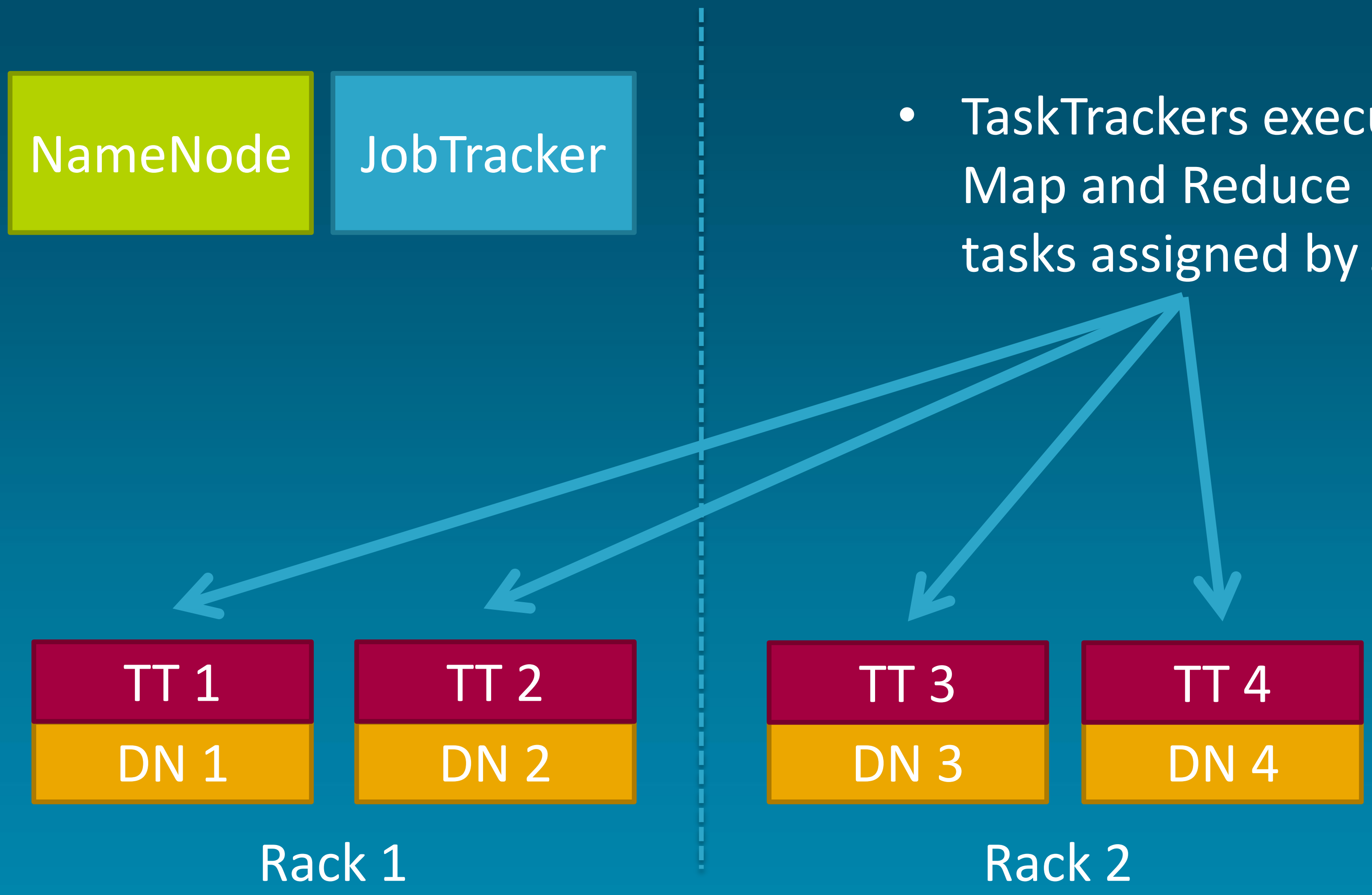
MapReduce Architcture



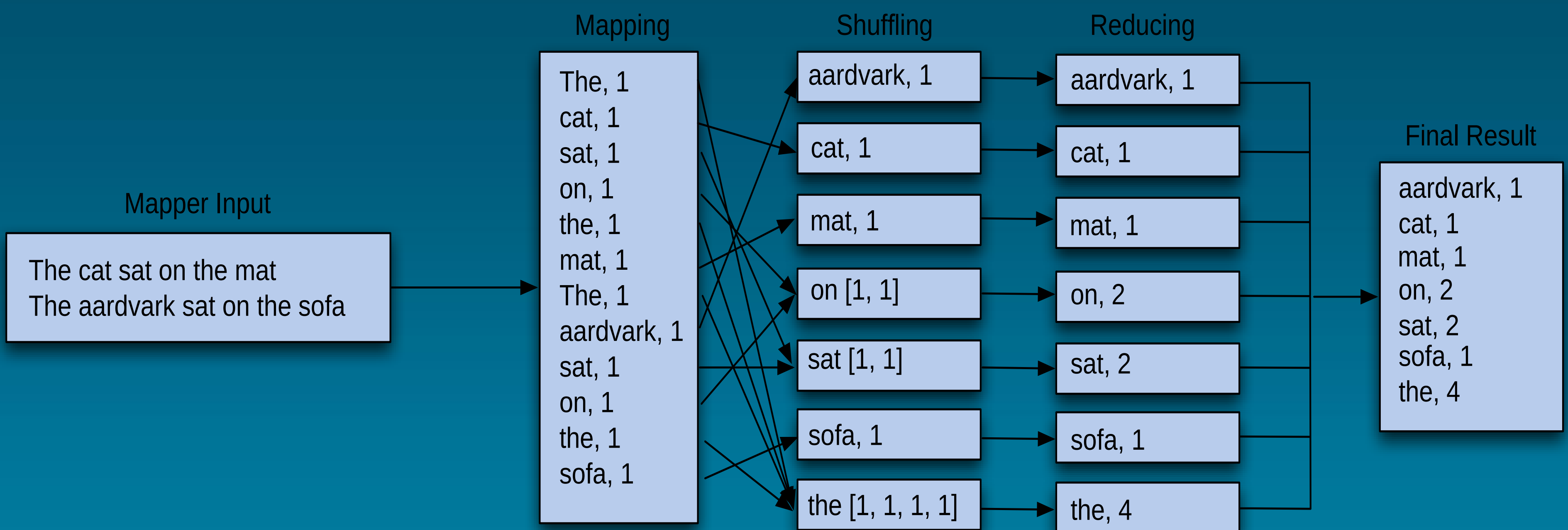
- Gateway for users
- Assigns tasks to TaskTrackers
- Tracks job status

MapReduce Architcture

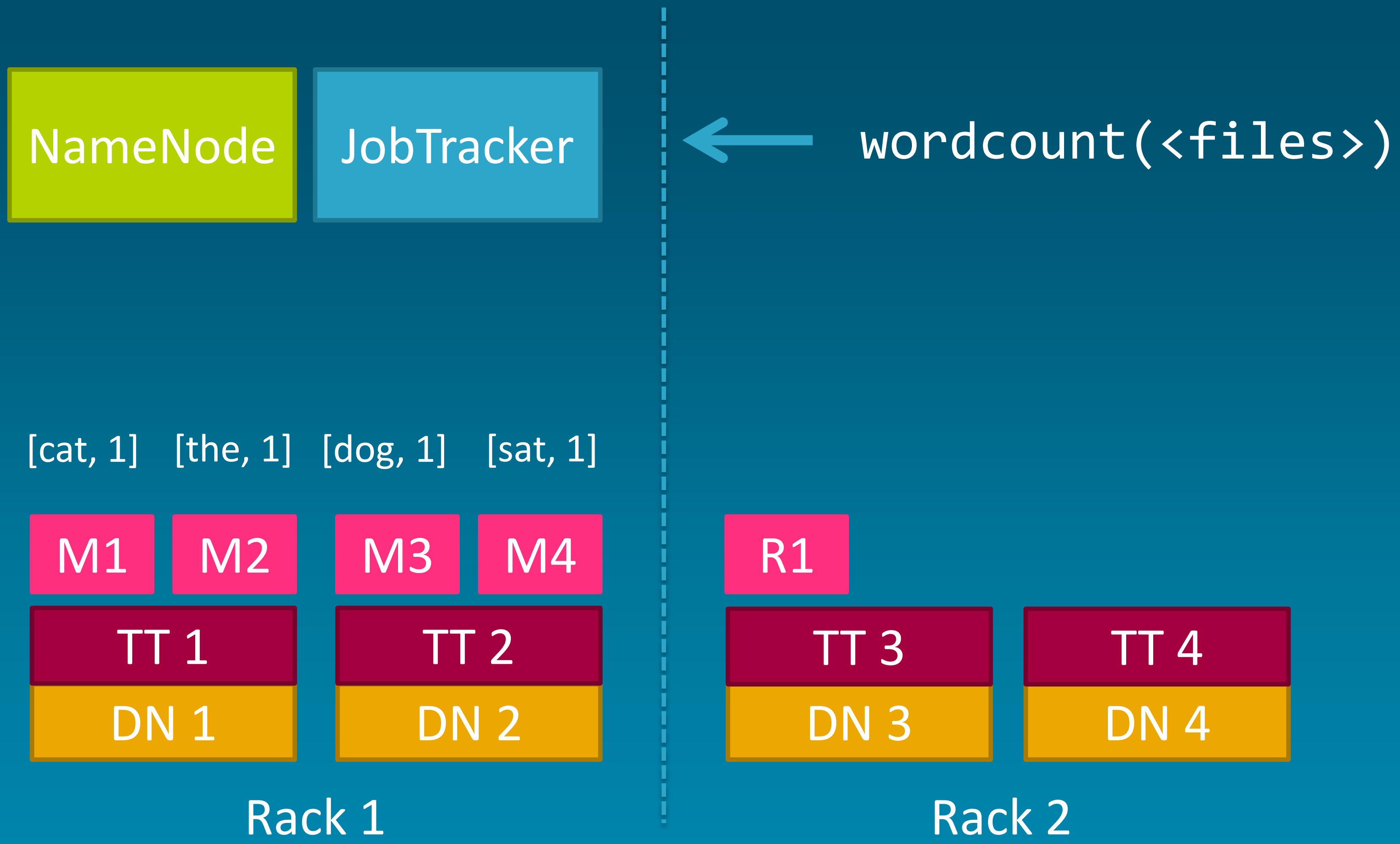
- TaskTrackers execute Map and Reduce tasks assigned by JT



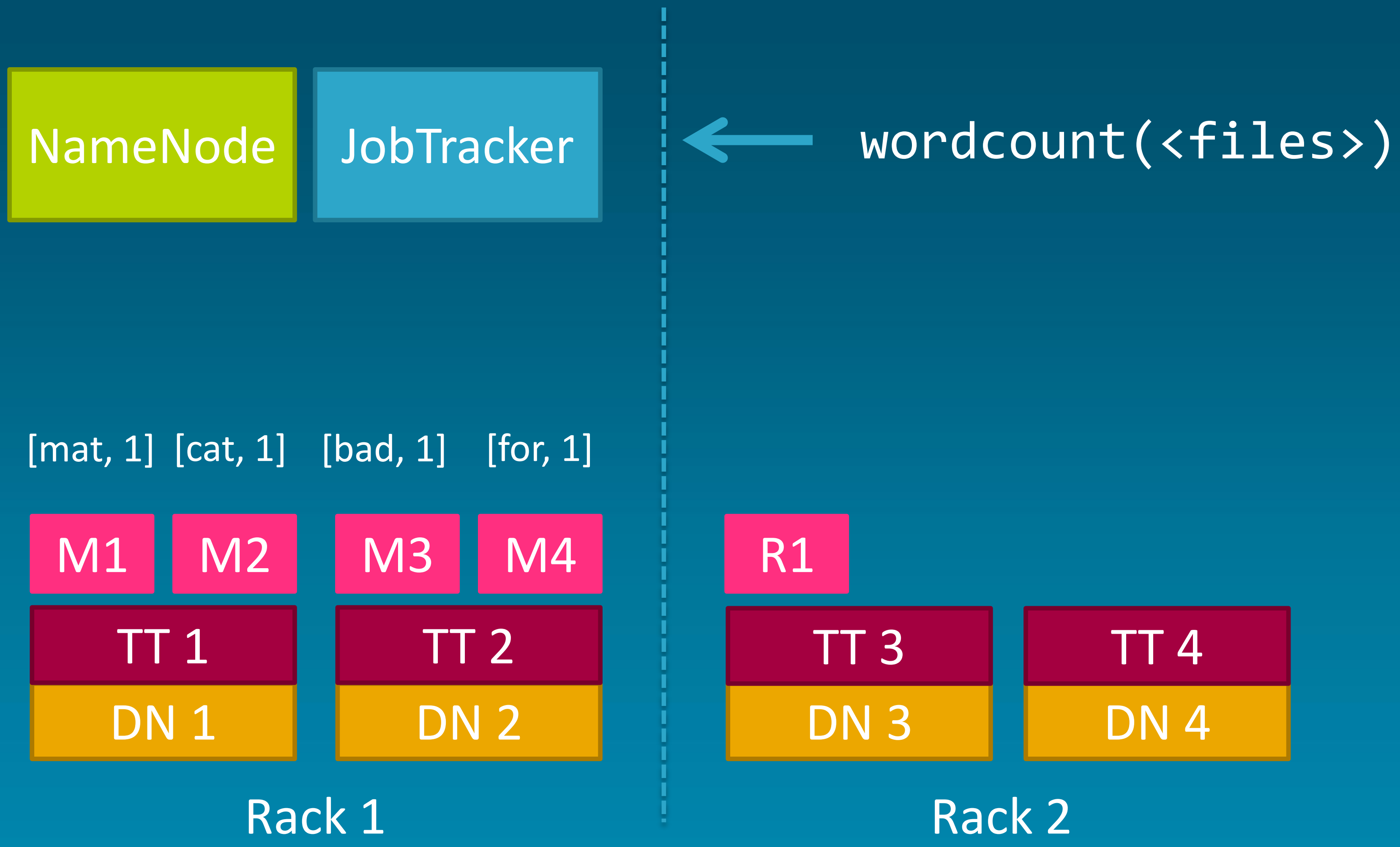
Word Count Example



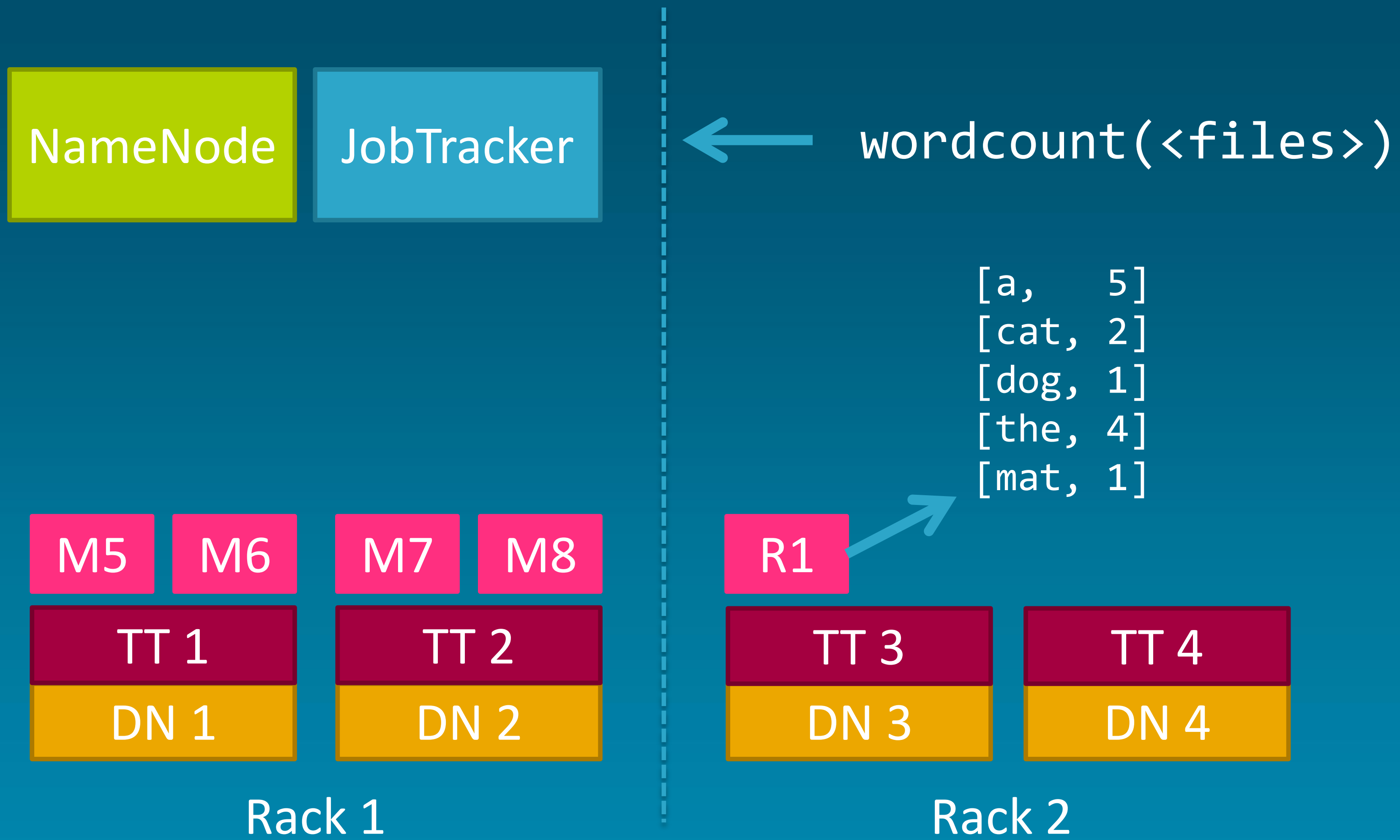
MapReduce Architcture



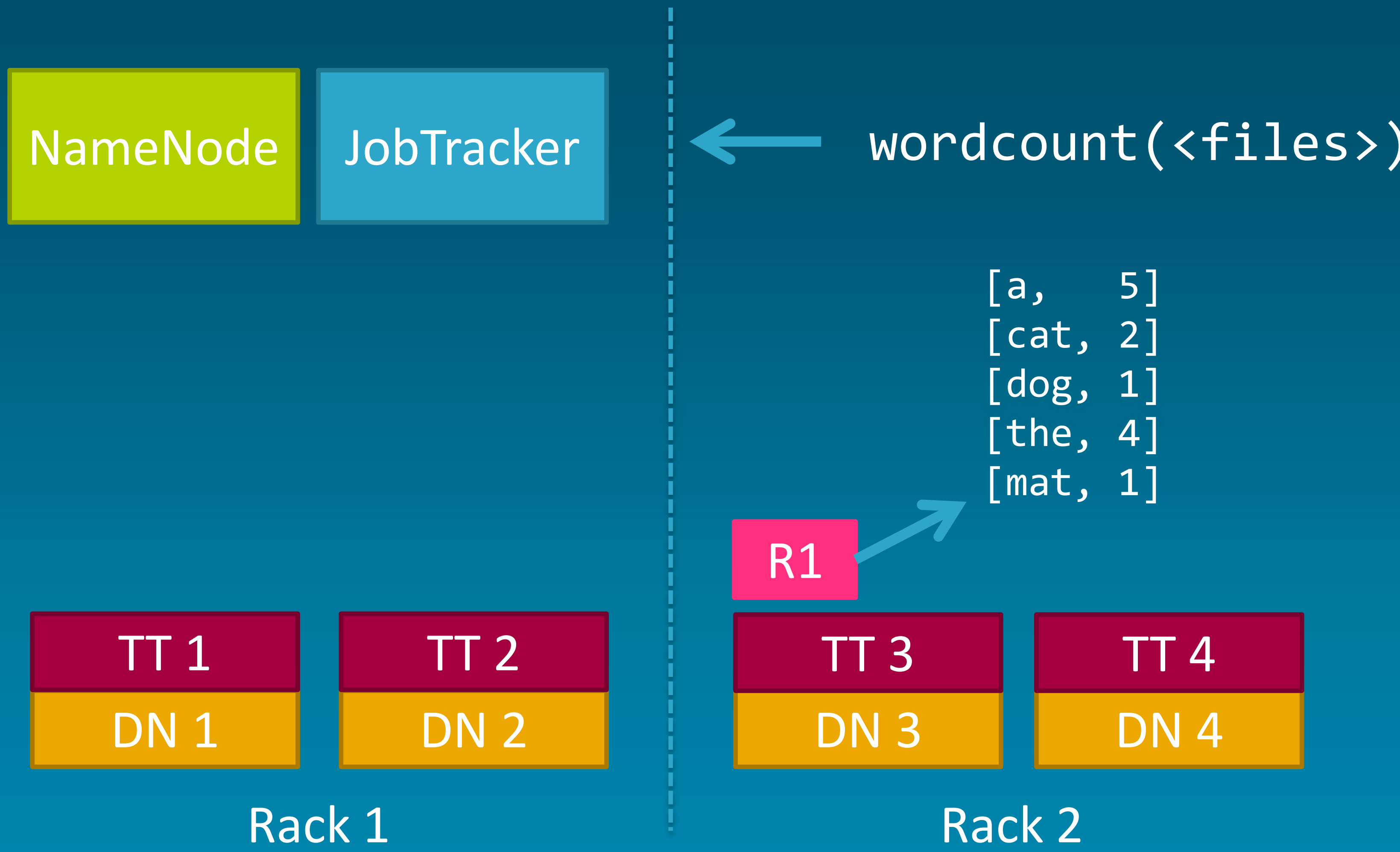
MapReduce Architcture



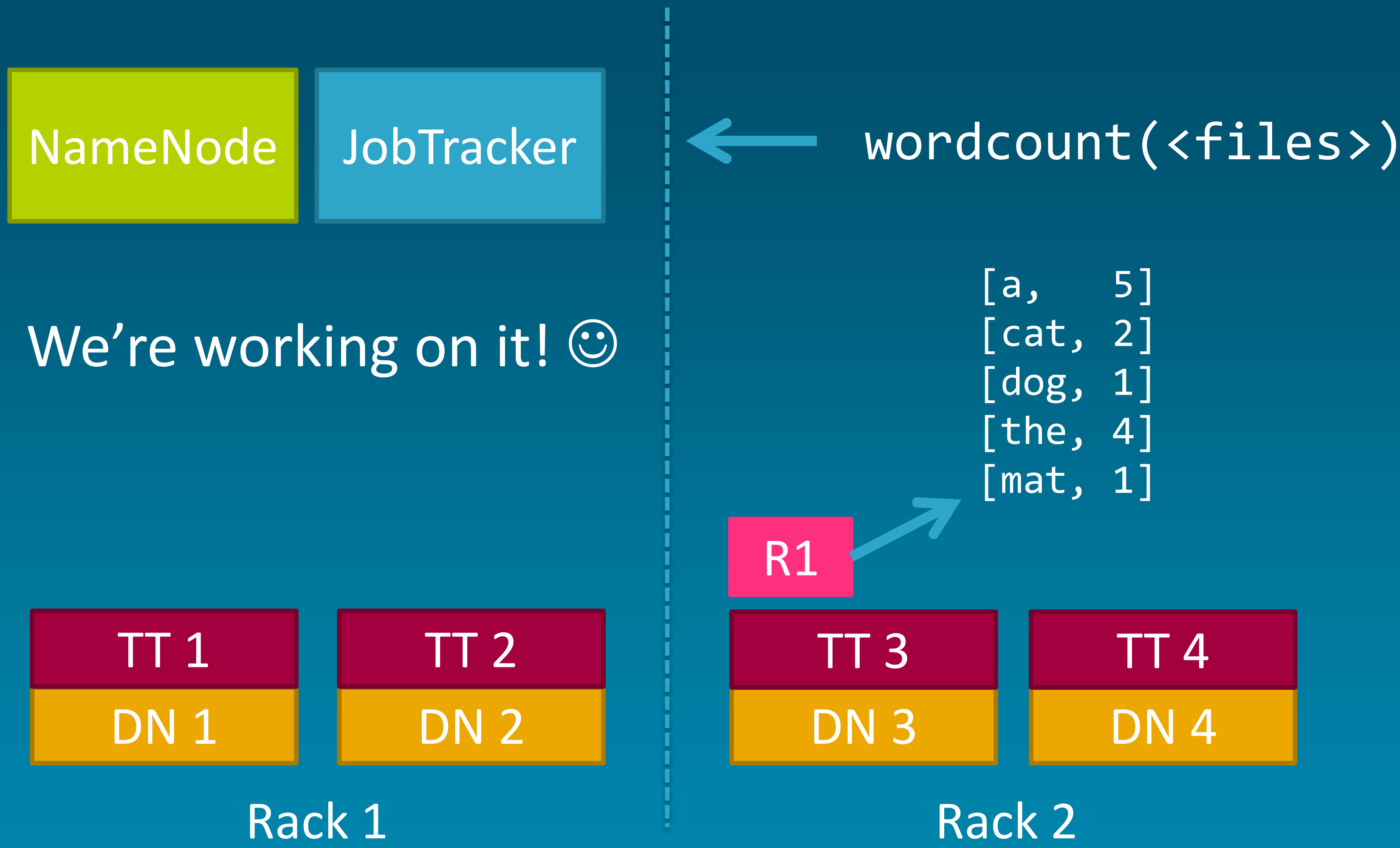
MapReduce Architcture



MapReduce Architcture



MapReduce Architcture

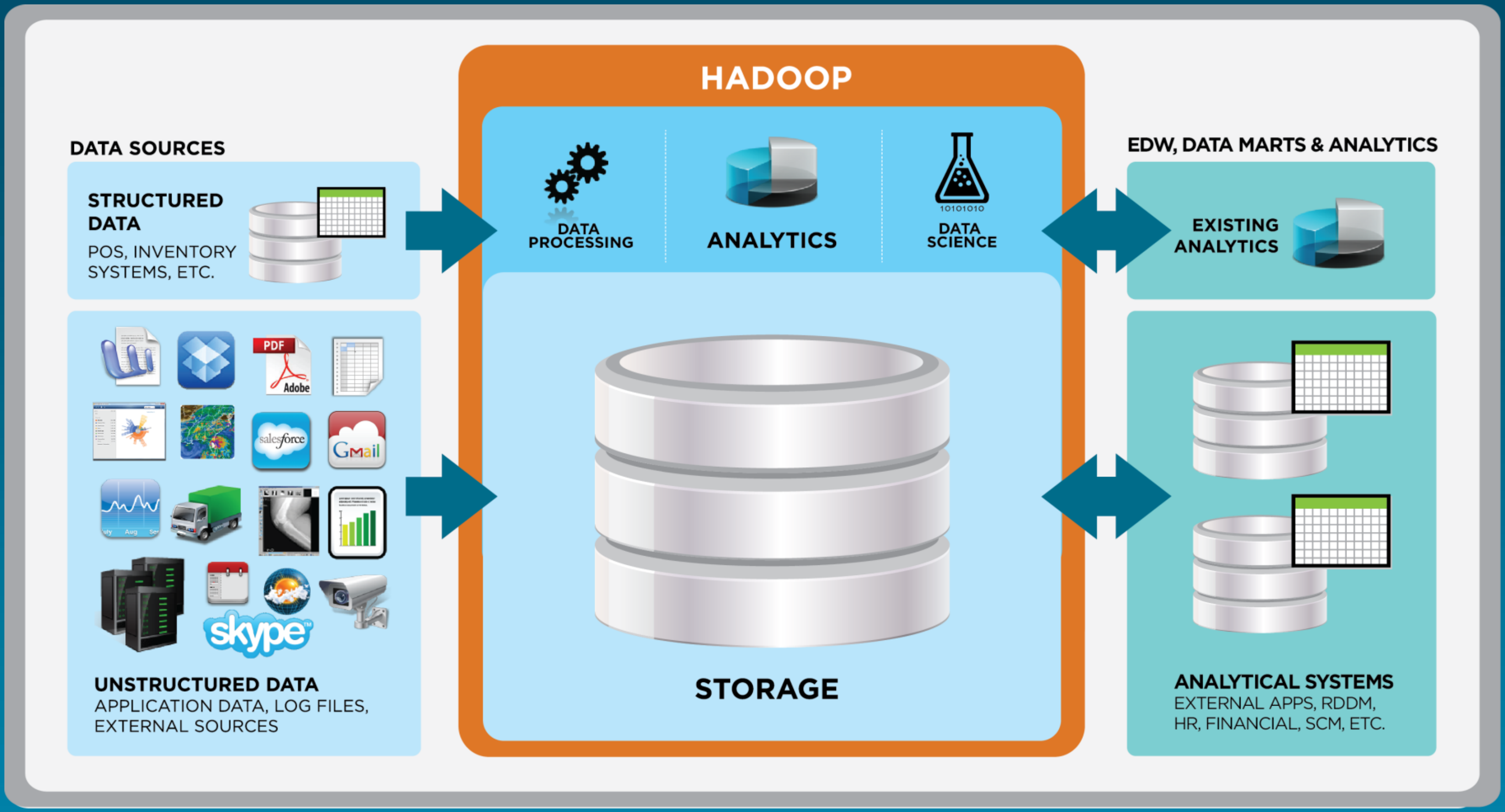


Summary

- GFS and MR co-design
  - Cheap, simple, effective at scale
- Fault-tolerance baked in
  - Replicate data 3x
  - Incrementally re-execute computation
  - Avoid single points of failure
- Held the world sort record (0.578TB/min)

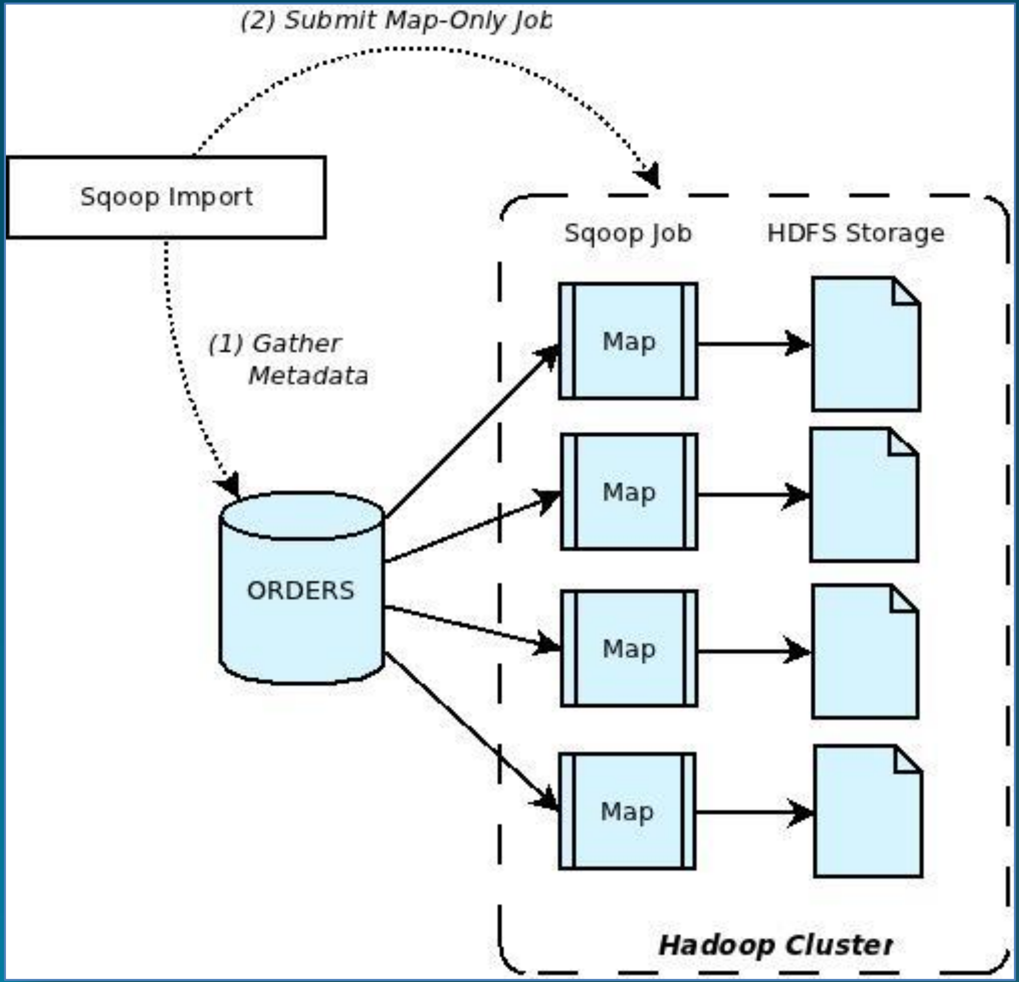
Hadoop ecosystem

Data Processing Pipeline



Sqoop

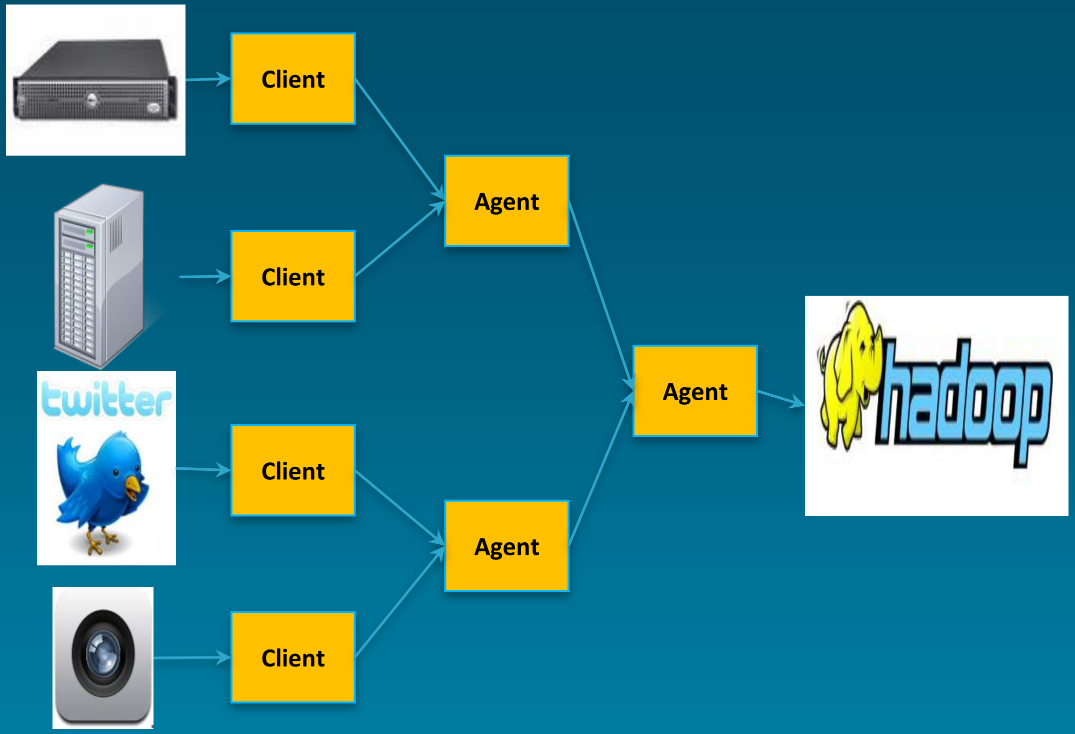
Performs bidirectional data transfers between Hadoop and almost any SQL database with a JDBC driver



cloudera
Ask Bigger Questions

Flume

A streaming data collection and aggregation system for massive volumes of data, such as RPC services, Log4J, Syslog, etc.



Hive

- Relational database abstraction using a SQL like dialect called HiveQL
- Statements are executed as one or more MapReduce Jobs

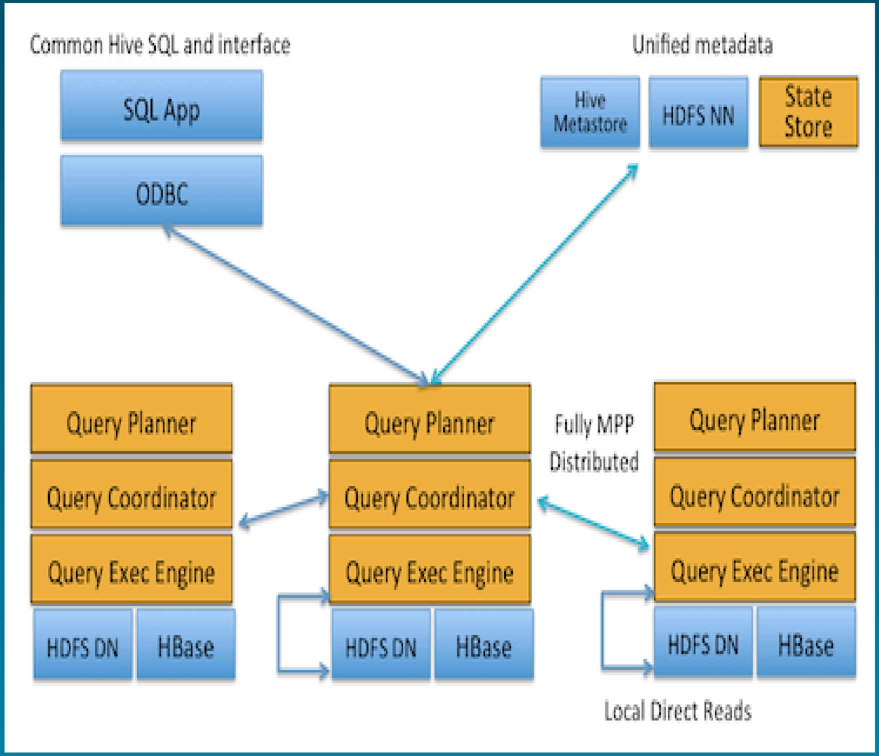
```
SELECT
        s.word, s.freq, k.freq
FROM shakespeare
JOIN ON (s.word= k.word)
WHERE s.freq >= 5;
```

Impala

Modern MPP database built on top of HDFS

Really fast! Written in C++

10-100x faster than Hive



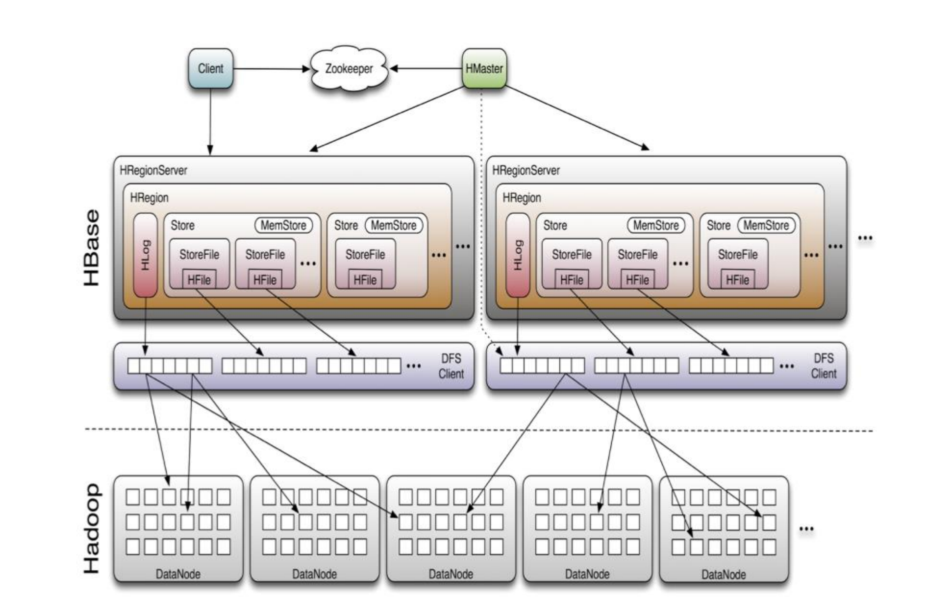
Pig

- High-level scripting language for for executing one or more MapReduce jobs
- Created to simplify authoring of MapReduce jobs
- Can be extended with user defined functions

emps = LOAD 'people.txt' AS (id,name,salary);
rich = FILTER emps BY salary > 200000;
sorted_rich = ORDER rich BY salary DESC;
STORE sorted_rich INTO 'rich_people.txt';

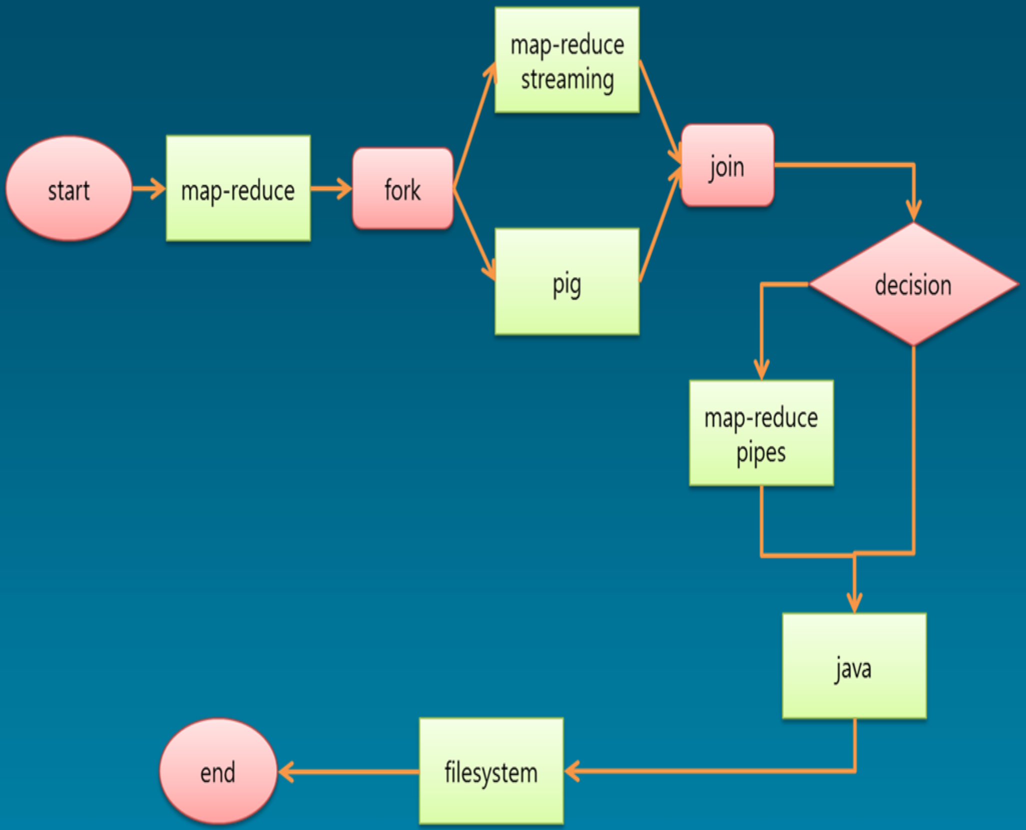
HBase

- Low-latency, distributed, columnar key-value store
- Based on BigTable
- Efficient random reads/writes on HDFS
- Useful for frontend applications



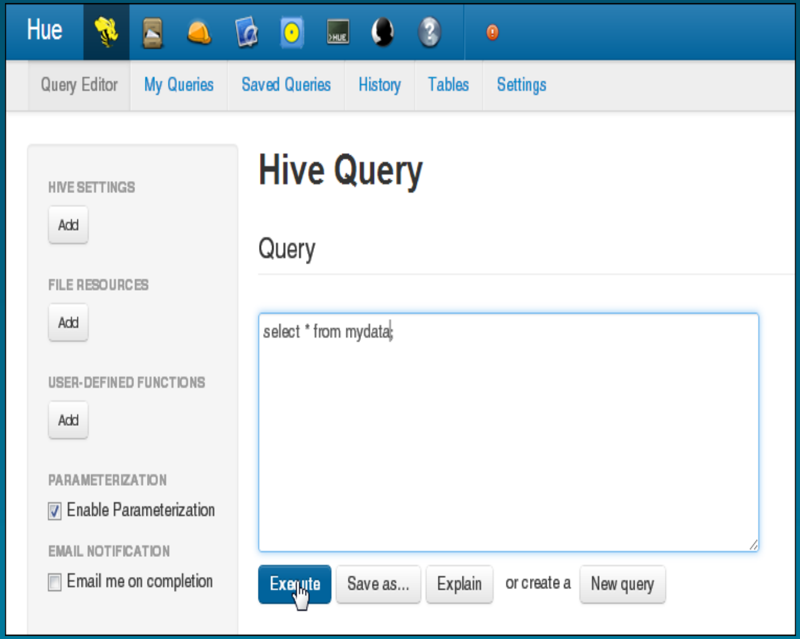
Oozie

A workflow engine and scheduler built specifically for large-scale job orchestration on a Hadoop cluster



Hue

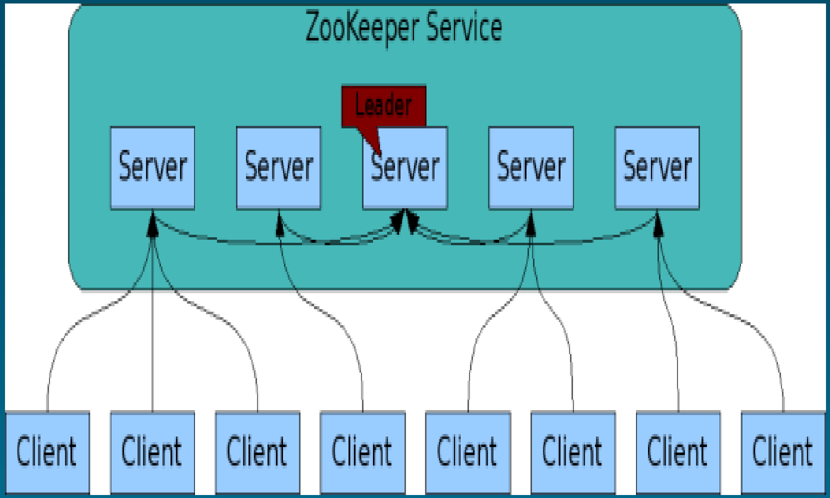
- Hue is an open source web-based application for making it easier to use Apache Hadoop.
- Hue features
  - File Browser for HDFS
  - Job Designer/Browser for MapReduce
  - Query editors for Hive, Pig and Cloudera Impala
  - Oozie



cloudera
Ask Bigger Questions

Zookeeper

- Zookeeper is a distributed consensus engine
- Provides well-defined concurrent access semantics:
  - Leader election
  - Service discovery
  - Distributed locking / mutual exclusion
  - Message board / mailboxes



cloudera
Ask Bigger Questions

Cloudera Manager

End-to-End Administration for CDH

1 Manage
Easily deploy, configure & optimize clusters

2 Monitor
Maintain a central view of all activity

3 Diagnose
Easily identify and resolve issues

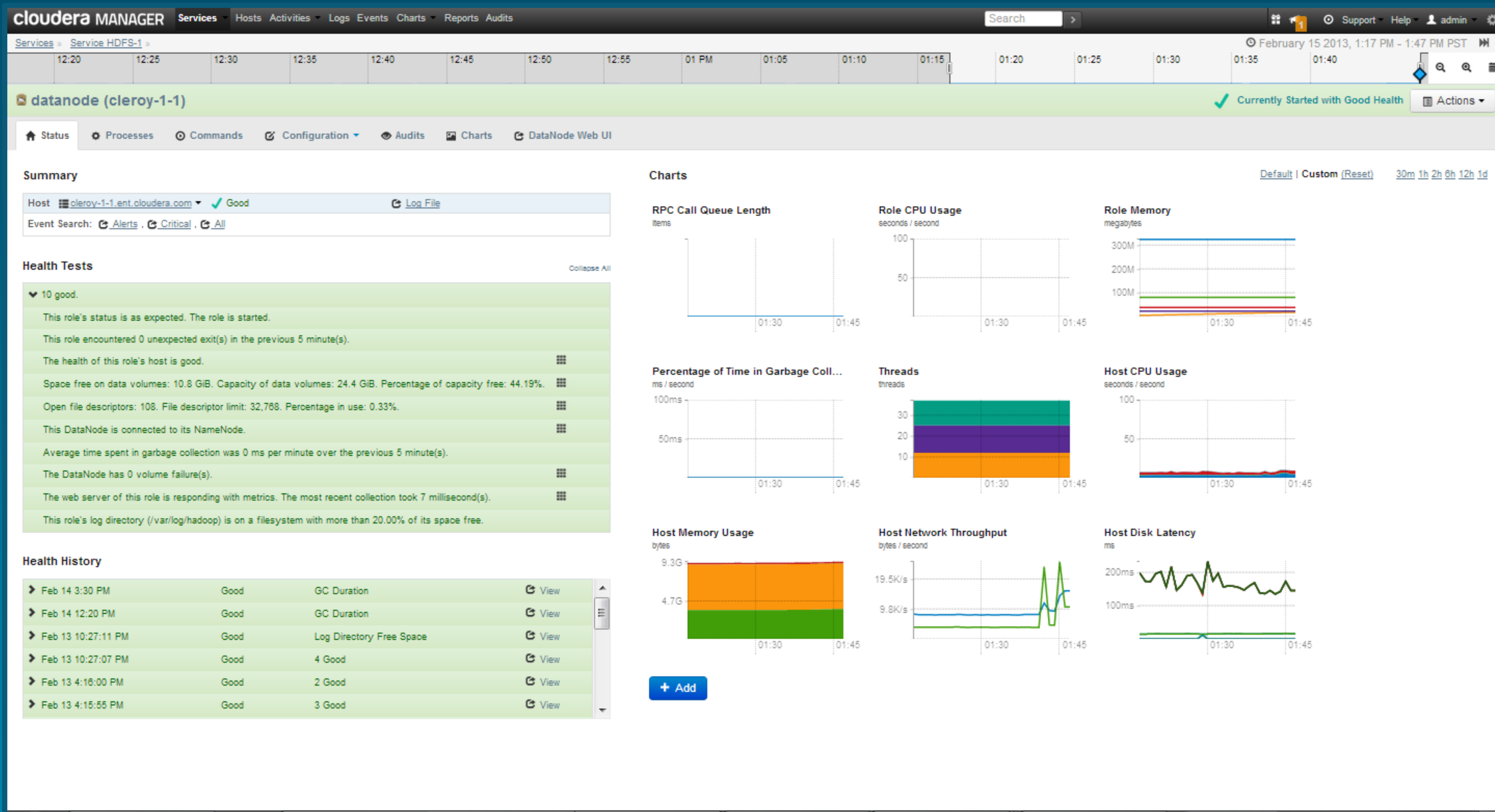
4 Integrate
Use Cloudera Manager with existing tools

Cloudera Manager

	DEPLOYMENT & CONFIGURATION	MONITORING	WORKFLOWS	EVENTS & ALERTS	LOG SEARCH	DIAGNOSTICS	REPORTING	ACTIVITY MONITORING
DO-IT-YOURSELF	puppet labs, Chef, CFengine	Ganglia, Scraping WebUI + cacti	Scripts	Nagios, splunk>	grep, splunk>	Ganglia, Chef, cacti, Nagios, splunk>, grep	Nagios, Ganglia, cacti	JobTracker, WebUI
WITH CLOUDERA	cloudera® MANAGER							

View Service Health & Performance

Cloudera Manager Key Features



cloudera®
Ask Bigger Questions

Thank You!

andrew.wang@cloudera.com

@umbrant