Hortonworks

Apache Hadoop.Next

What it takes and what it means…

Arun C. Murthy
Founder & Architect, Hortonworks
@acmurthy (@hortonworks)

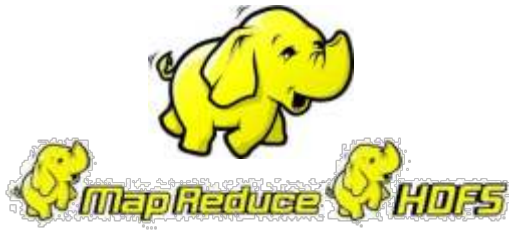
Hello! I'm Arun

- **Founder/Architect at Hortonworks Inc.**
  - Lead, Map-Reduce
  - Formerly, Architect Hadoop MapReduce, Yahoo
  - Responsible for running Hadoop MR as a service for all of Yahoo (50k nodes footprint)
    - Yes, I took the 3am calls! ☺
- **Apache Hadoop, ASF**
  - VP, Apache Hadoop, ASF (Chair of Apache Hadoop PMC)
  - Long-term Committer/PMC member (full time ~6 years)
  - Release Manager - hadoop-0.23 (i.e. Hadoop.Next)

Yahoo!, Apache Hadoop & Hortonworks

http://www.wired.com/wiredenterprise/2011/10/how-yahoo-spawned-hadoop

Yahoo! embraced Apache Hadoop, an open source platform, to
crunch epic amounts of data using an **army of dirt-cheap servers**

YAHOO! → **2006** → 

Hadoop at Yahoo!

40K+ Servers

170PB Storage

5M+ Monthly Jobs

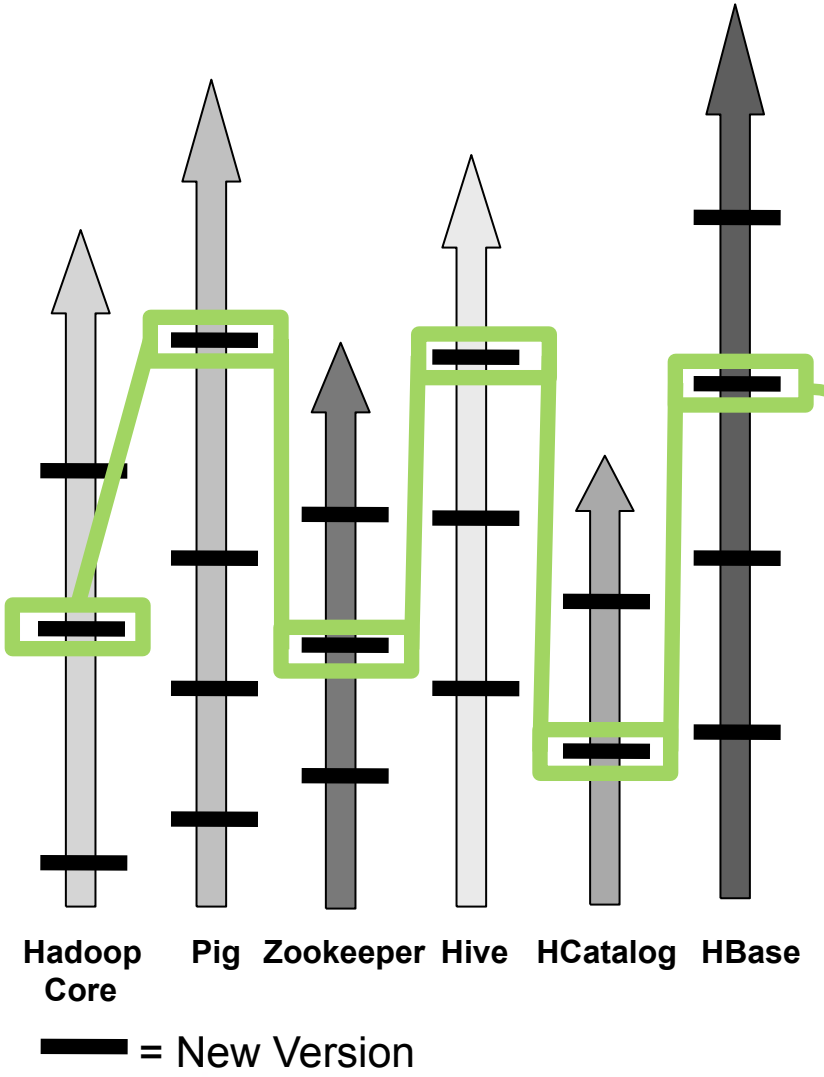
1000+ Active Users

YAHOO! → **2011** → Hortonworks

Yahoo! spun off 22+ engineers into Hortonworks, a company focused on
advancing open source Apache Hadoop for the broader market

Hortonworks Data Platform

Fully Supported Integrated Platform



Challenge:

- Integrate, manage, and support changes across a wide range of open source projects that power the Hadoop platform; each with their own release schedules, versions, & dependencies.
- Time intensive, Complex, Expensive

Solution: Hortonworks Data Platform

- Integrated certified platform distributions
- Extensive Q/A process
- Industry-leading Support with clear service levels for updates and patches
- Continuity via multi-year Support and Maintenance Policy

HDP1: “Hadoop.Now”

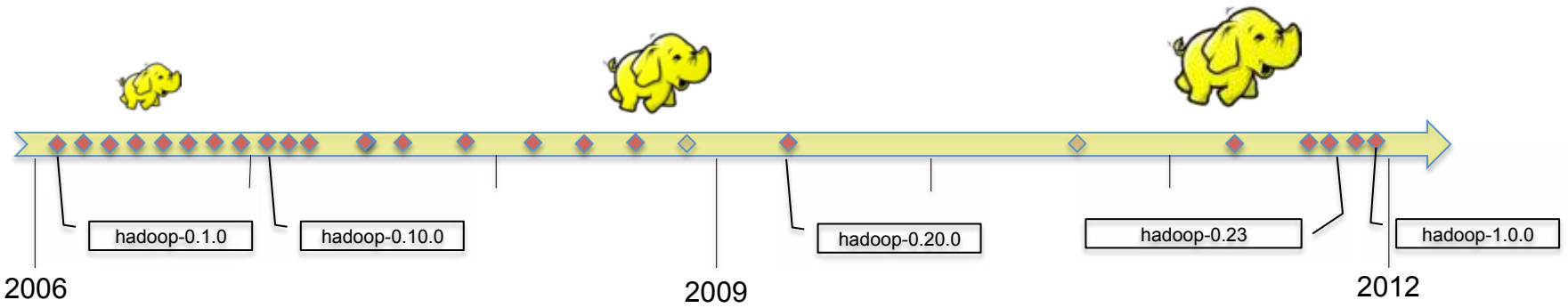
- **Based on Hadoop 1.0 (formerly known as 0.20.205)**
  - Merges various Hadoop 0.20.* branches (security, HBase support, ..)
  - Goal: ensure patching and back-porting happen on one stable line
- **Highlights:**
  - First Apache line that supports Security, HBase, and WebHDFS
  - Common table and schema management via HCatalog (based on Hive)
  - Open APIs for metadata access, data movement, app job management
  - Consumable “standard Hadoop” stack:
    - Hadoop 0.20.205* (HDFS, MapReduce)
    - Pig 0.9.* data flow programming language
    - Hive 0.8.* SQL-like language
    - HBase 0.92.* bigtable datastore
    - HCatalog 0.3.* common table and schema management
    - ZooKeeper 3.4.* coordinator

HDP2: “Hadoop.Next”

- **Based on Hadoop 0.23.***
  - Include latest stable “standard Hadoop” components
  - Goal: ensure next-generation architecture happens on one stable line
- **Highlights:**
  - HDFS Federation
    - Clear separation of Namespace and Block Storage
    - Improved scalability and isolation
  - HDFS HA
  - Next Generation MapReduce architecture
    - New architecture enables other application types to plug in
      - Ex. Streaming, Graph, Bulk Sync Processing, MPI (Message Passing Interface)
    - New Resource Manager enhances enterprise viability
      - Scalability, HA, FT, and SPoF
      - Client/cluster version compatibility
  - Performance

Releases so far…

- Started for Nutch… Yahoo picked it up in early 2006, hired Doug Cutting
- Initially, we did monthly releases (0.1, 0.2 …)
- Quarterly after hadoop-0.15 until hadoop-0.20 in 04/2009…
- **hadoop-0.20 is still the basis of all current, stable, Hadoop distributions**
  - Apache Hadoop 1.0.0
  - CDH3.*
  - HDP1.*
- hadoop-0.20.203 (security) – 05/2011
- **hadoop-1.0.0 (security + append + webhdfs) – 12/2011**

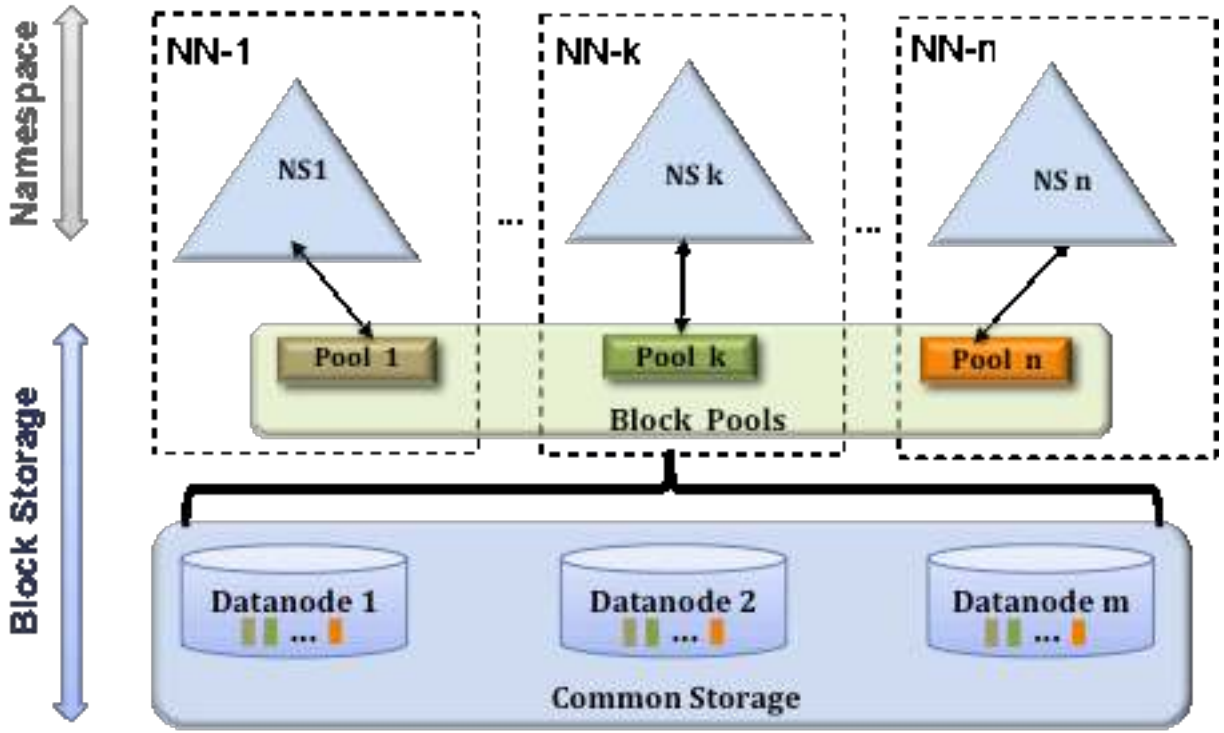


hadoop-0.23

- First stable release off Apache Hadoop trunk in over 30 months…
- Currently alpha quality (*hadoop-0.23.0*)
- Significant ***major*** features
  - HDFS Federation
  - NextGen Map-Reduce (YARN)
  - HDFS HA
  - Wire protocol compatibility
  - Performance

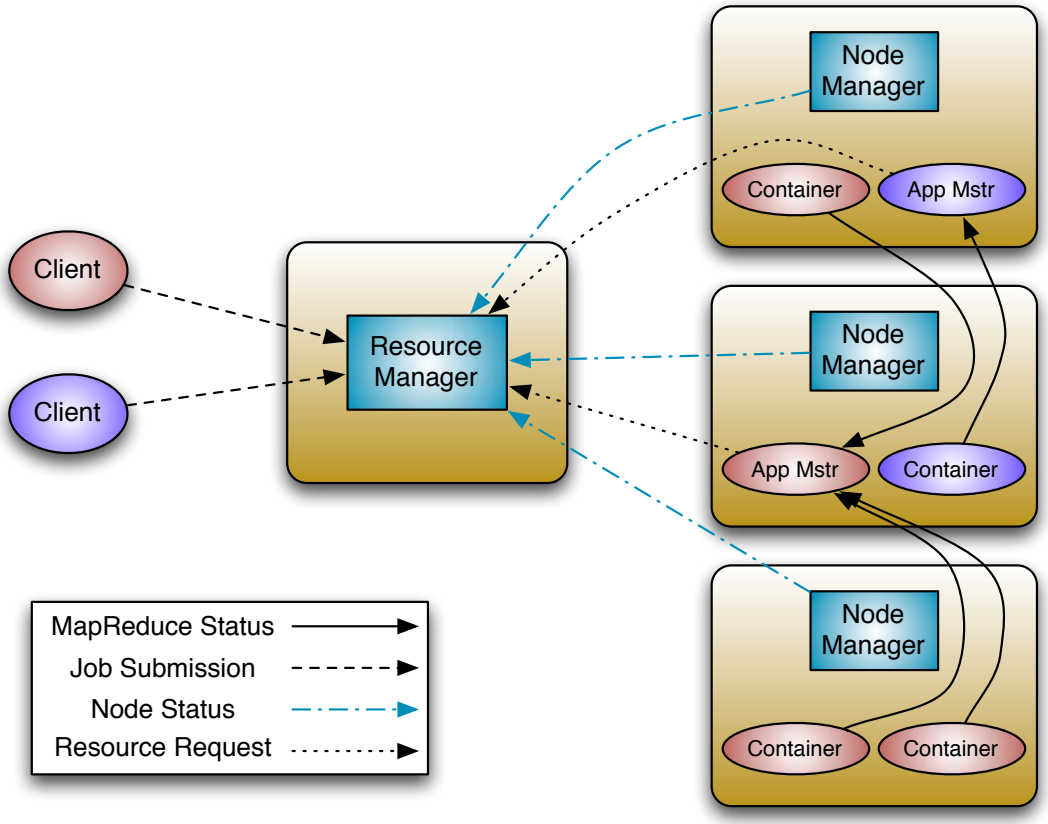
HDFS - Federation

- Significant scaling…
- Separation of Namespace mgmt and Block mgmt



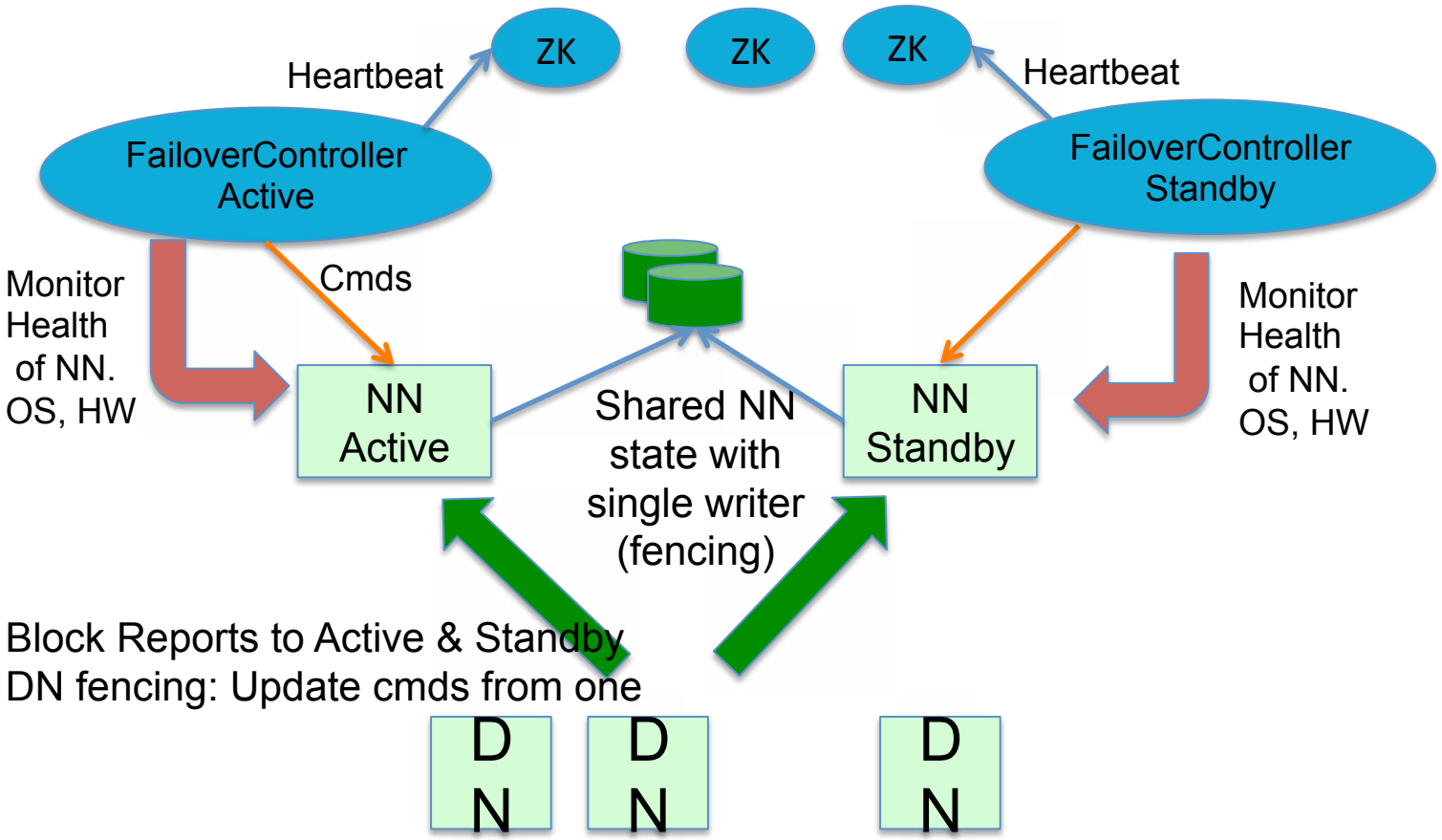
MapReduce - YARN

- NextGen Hadoop Data Processing Framework
- Support MR and other paradigms
- High Availability



HDFS NameNode HA

- Nearly code-complete
  - Automatic failover
  - Multiple-options for failover (shared disk, Linux HA, Zookeeper)



Performance

- 2x+ across the board
- HDFS read/write
  - CRC32
  - fadvise
  - Shortcut for local reads
- MapReduce
  - Unlock lots of improvements from Terasort record (Owen/Arun, 2009)
  - Shuffle 30%+
  - Small Jobs – Uber AM

More…

- Compatible Wire Protocols
  - On the way to rolling upgrades
- HDFS Write pipeline improvements for HBase
  - Append/flush etc.
- Build - Full Mavenization
- EditLogs re-write
  - https://issues.apache.org/jira/browse/HDFS-1073
- Lots more …

Deployment goals

- Clusters of 6,000 machines
  - Each machine with 16+ cores, 48G/96G RAM, 24TB/36TB disks
  - 200+ PB (raw) per cluster
  - 100,000+ concurrent tasks
  - 10,000 concurrent jobs

- Yahoo: 50,000+ machines

What does it take to get there?

- Testing, *lots* of it
- Benchmarks (Every release should be at least as good as the last one)
- Integration testing
  - HBase
  - Pig
  - Hive
  - Oozie
- Deployment discipline

Testing

- Why is it hard?
  - Map-Reduce is, effectively, very wide api
  - Add Streaming
  - Add Pipes
  - Oh, Pig/Hive etc. etc.
- Functional tests
  - Nightly
  - Over1000 functional tests for Map-Reduce alone
  - Several hundred for Pig/Hive etc.
- Scale tests
  - Simulation
- Longevity tests
- Stress tests

Benchmarks

- Benchmark every part of the HDFS & MR pipeline
  - HDFS read/write throughput
  - NN operations
  - Scan, Shuffle, Sort
- GridMixv3
  - Run production *traces* in test clusters
  - Thousands of jobs
  - Stress mode v/s Replay mode

Integration Testing

- Several projects in the ecosystem
  - HBase
  - Pig
  - Hive
  - Oozie
- Cycle
  - Functional
  - Scale
  - Rinse, repeat

Deployment

- Alpha/Test (early UAT)
  - Started in Dec, 2011 - ongoing
  - Small scale (500-800 nodes)
- Alpha
  - Feb, 2012
  - Majority of users
  - ~1000 nodes per cluster, > 2,000 nodes in all
- Beta
  - *Misnomer:* 100s of PB, Millions of user applications
  - Significantly wide variety of applications and load
  - 4000+ nodes per cluster, > 15000 nodes in all
  - Late Q1, 2012
- Production
  - Well, it's production
  - Mid-to-late Q2 2012

HDP2: “Hadoop.Next”

- **Based on Hadoop 0.23.***
  - Include latest stable “standard Hadoop” components
  - Goal: ensure next-generation architecture happens on one stable line
- **Highlights:**
  - HDFS Federation
  - HDFS HA
  - Next Generation MapReduce architecture
  - Performance
- **Apache Release**
  - hadoop-0.23.0 (pre-alpha)
    - http://hadoop.apache.org/common/releases.html
    - http://hadoop.apache.org/common/docs/r0.23.0/
  - hadoop-0.23.1 (alpha) – Jan, 2012

How Hortonworks Can Help

- **Training and Certification**
  - www.hortonworks.com/training/
- **Hortonworks Data Platform and Support**
  - www.hortonworks.com/hortonworksdataplatform/
  - www.hortonworks.com/technology/techpreview/
  - www.hortonworks.com/support/
- **Educational Webinars**
  - www.hortonworks.com/webinars/

Questions?

hadoop-0.23.0 (alpha release):
http://hadoop.apache.org/common/releases.html

Release Documentation:
http://hadoop.apache.org/common/docs/r0.23.0/

Thank You.

@acmurthy