Hortonworks

Extending Hadoop beyond MapReduce

Mahadev Konar
Co-Founder
@mahadevkonar (@hortonworks)

Bio

- Apache Hadoop since 2006 - committer and PMC member
  - Developed and supported Map Reduce @Yahoo!
  - Core member of design and development team on MR Next Gen
- Apache ZooKeeper since 2008 – committer and current PMC chair
  - Lead Apache ZooKeeper development and support @ Yahoo!
- Co-founder @hortonworks

Bio

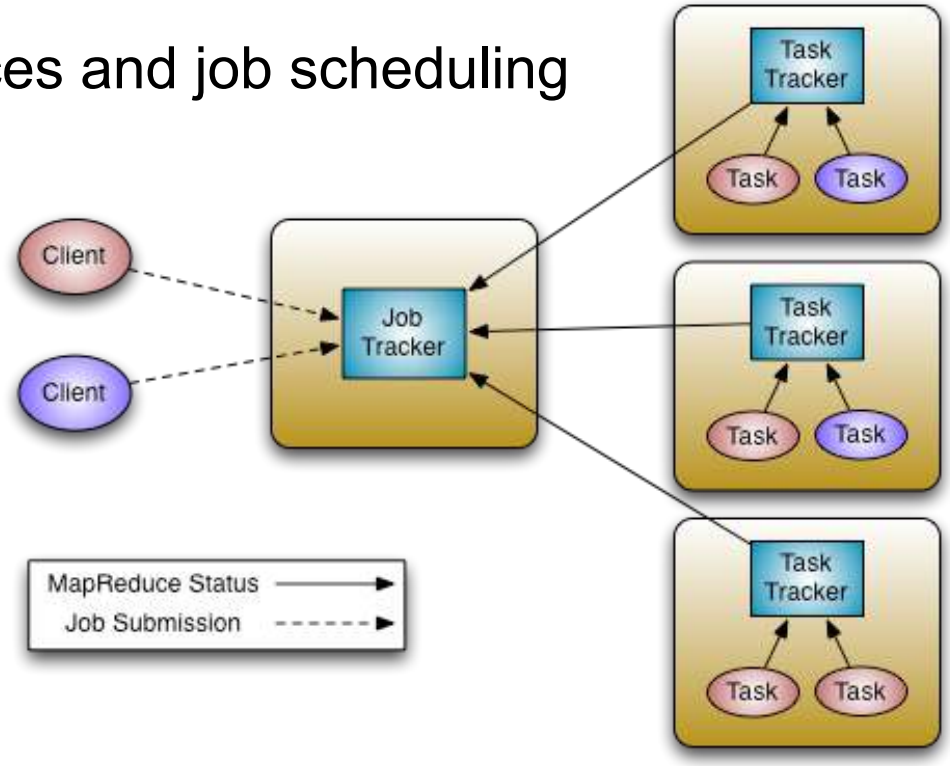
- Apache Hadoop since 2006 - committer and PMC member
  - Developed and supported Map Reduce @Yahoo!
  - Core member of design and development team for MR Next Gen
- Apache ZooKeeper since 2008 – committer and current PMC chair
  - Lead Apache ZooKeeper development and support @ Yahoo!
- Co-founder @hortonworks

Agenda

- Overview
- Current Limitations and Requirements
- Architectures
- Improvements and Updates
- Q&A

Hadoop MapReduce Classic

- JobTracker
  - Manages cluster resources and job scheduling
- TaskTracker
  - Per-node agent
  - Manage tasks



Current Limitations

- Hard partition of resources into map and reduce slots
  - Low resource utilization
- Lacks support for alternate paradigms
  - Iterative applications implemented using MapReduce are 10x slower
  - Hacks for the likes of MPI/Graph Processing
- Lack of wire-compatible protocols
  - Client and cluster must be of same version
  - Applications and workflows cannot migrate to different clusters

Current Limitations

- Utilization
- Scalability
  - Maximum Cluster size – 4,000 nodes
  - Maximum concurrent tasks – 40,000
  - Coarse synchronization in JobTracker
- Single point of failure
  - Failure kills all queued and running jobs
  - Jobs need to be re-submitted by users
- Restart is very tricky due to complex state

Requirements

- Reliability
- Availability
- Utilization
- Wire Compatibility
- Agility & Evolution – Ability for customers to control upgrades to the grid software stack.
- Scalability - Clusters of 6,000-10,000 machines
  - Each machine with 16 cores, 48G/96G RAM, 24TB/36TB disks
  - 100,000+ concurrent tasks
  - 10,000 concurrent jobs

Design Centre

- Split up the two major functions of JobTracker
  - Cluster resource management
  - Application life-cycle management
- MapReduce becomes **user-land** library

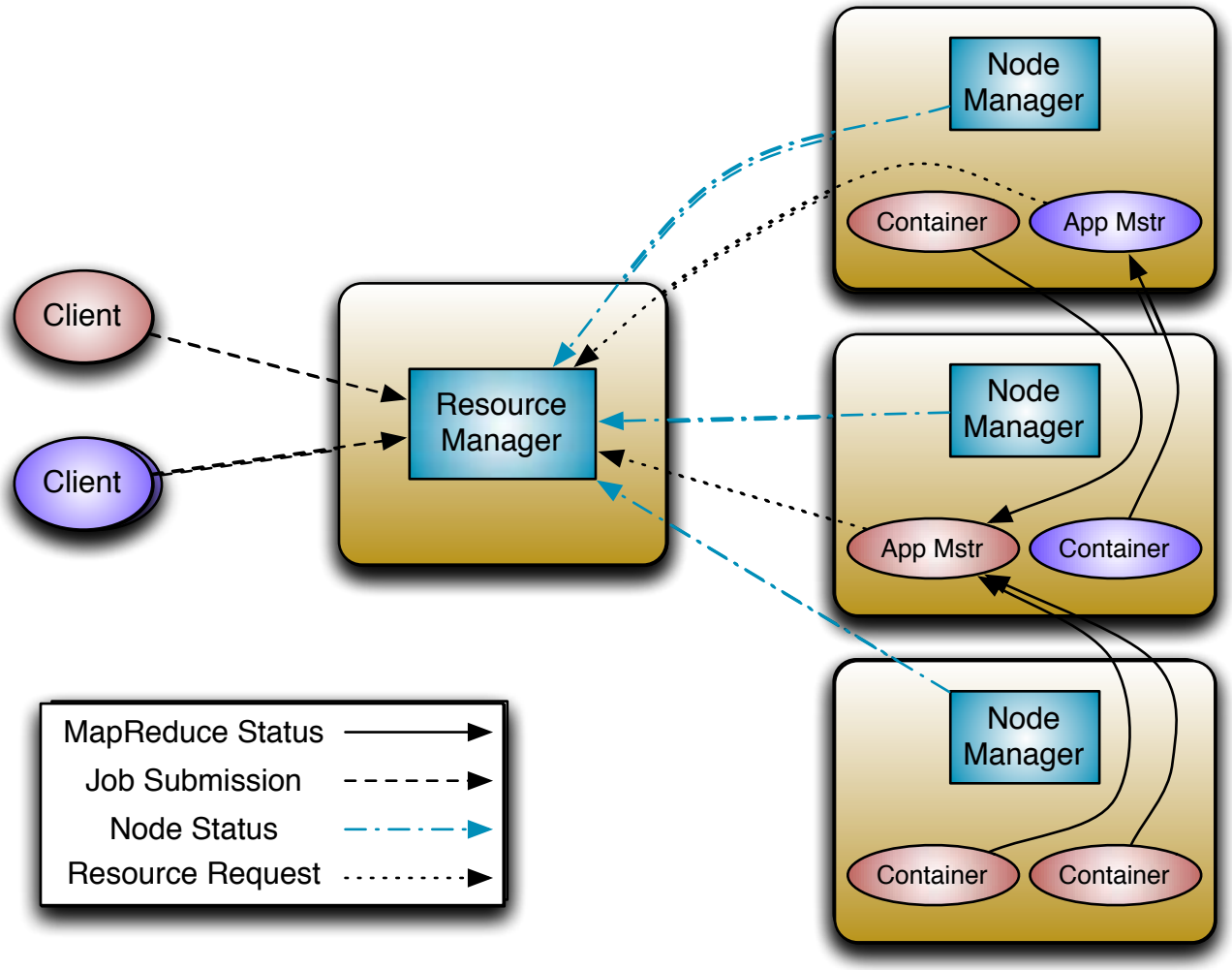
Architecture

- Application
  - Application is a job submitted to the framework
  - Example – Map Reduce Job
- Container
  - Basic unit of allocation
  - Example – container A = 2GB, 1CPU
  - Replaces the fixed map/reduce slots

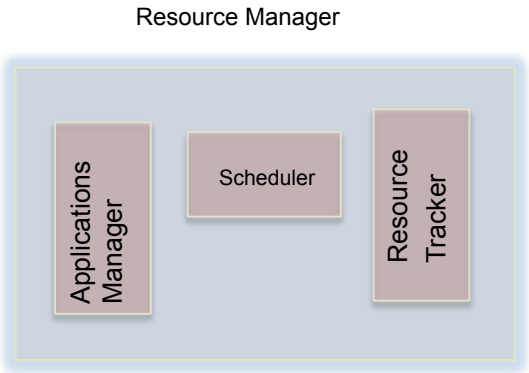
Architecture

- Resource Manager
  - Global resource scheduler
  - Hierarchical queues
- Node Manager
  - Per-machine agent
  - Manages the life-cycle of container
  - Container resource monitoring
- Application Master
  - Per-application
  - Manages application scheduling and task execution
  - E.g. MapReduce Application Master

Architecture



Architecture – Resource Manager



- Applications Manager
  - Responsible for launching and monitoring Application Masters (per Application process)
  - Restarts an Application Master on failure
- Scheduler
  - Responsible for allocating resources to the Application
- Resource Tracker
  - Responsible for managing the nodes in the cluster

Improvements vis-à-vis classic MapReduce

- Utilization
  - Generic resource model
    - Memory
    - CPU
    - Disk b/q
    - Network b/w
  - Remove fixed partition of map and reduce slot
- Scalability
  - Application life-cycle management is very expensive
  - Partition resource management and application life-cycle management
  - Application management is distributed
  - Hardware trends - Currently run clusters of 4,000 machines
    - 6,000 2012 machines > 12,000 2009 machines
    - <16+ cores, 48/96G, 24TB> v/s <8 cores, 16G, 4TB>

Improvements vis-à-vis classic MapReduce

- Fault Tolerance and Availability
  - Resource Manager
    - No single point of failure – state saved in ZooKeeper (coming soon)
    - Application Masters are restarted automatically on RM restart
  - Application Master
    - Optional failover via application-specific checkpoint
    - MapReduce applications pick up where they left off via state saved in HDFS
- Wire Compatibility
  - Protocols are wire-compatible
  - Old clients can talk to new servers
  - Rolling upgrades

Improvements vis-à-vis classic MapReduce

- Innovation and Agility
  - MapReduce now becomes a user-land library
  - Multiple versions of MapReduce can run in the same cluster (a la Apache Pig)
    - Faster deployment cycles for improvements
  - Customers upgrade MapReduce versions on their schedule
  - Users can customize MapReduce

Improvements vis-à-vis classic MapReduce

- Support for programming paradigms other than MapReduce
  - MPI
  - Master-Worker
  - Machine Learning
  - Iterative processing
  - Enabled by allowing the use of paradigm-specific application master
  - Run all on the same Hadoop cluster

Is it released?

- Available in 0.23.1 release
- Coming soon 0.23.2 release

Any Performance Gains?

- 2x+ across the board
- MapReduce
  - Unlock lots of improvements from Terasort record (Owen/Arun, 2009)
  - Shuffle 30%+
  - Small Jobs – Uber AM
  - Re-use task slots (container reuse)

More details: http://hortonworks.com/delivering-on-hadoop-next-benchmarking-performance/

Testing?

- Testing, *lots* of it
- Benchmarks (every release should be at least as good as the last one)
- Integration testing
  - HBase
  - Pig
  - Hive
  - Oozie
- Functional tests
  - Nightly
  - Over1000 functional tests for Map-Reduce alone
  - Several hundred for Pig/Hive etc.
- Deployment discipline

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

Deployment

- Alpha/Test (early UAT) in November 2011
  - Small scale (500-800 nodes)
- Alpha in February 2012
  - Majority of users
  - ~1000 nodes per cluster, > 2,000 nodes in all
- Beta
  - *Misnomer:* 10s of PB of storage
  - Significantly wide variety of applications and load
  - 4000+ nodes per cluster, > 15000 nodes in all
  - Q2, 2012
- Production
  - Well, it's production
  - Mid-to-late Q2 2012

Questions?

hadoop-0.23.1 (alpha release):
http://hadoop.apache.org/common/releases.html

Release Documentation:
http://hadoop.apache.org/common/docs/r0.23.1/

Hortonworks website:
http://hortonworks.com/

Other Resources

- **Hadoop Summit**
  - June 13-14
  - San Jose, California
  - www.Hadoopsummit.org

- **Hadoop Training and Certification**
  - Developing Solutions Using Apache Hadoop
  - Administering Apache Hadoop
  - http://hortonworks.com/training/

Hortonworks UNIVERSITY

- **On-demand Webinars**
  - Available now on Hortonworks website
  - http://hortonworks.com/webinars/

WATCH NOW

Hortonworks

Thank You

@mahadevkonar