
cloudera

Global Information Platforms

Evolving the Data Warehouse

Jeff Hammerbacher
Chief Scientist and Vice President of Products, Cloudera
April 9, 2009

Presentation Outline

- Introductions
- What we've built
  - Short history of Facebook's Data team
  - Hadoop applications at Yahoo!, Facebook, and Cloudera
- Where the world is headed
  - The Unreasonable Effectiveness of Data
- What we're building at Cloudera
  - Cloudera's Distribution for Hadoop
  - Training, Support, and Cloud Services
  - Research problems

Lessons from Facebook

Early 2006: The First Research Scientist

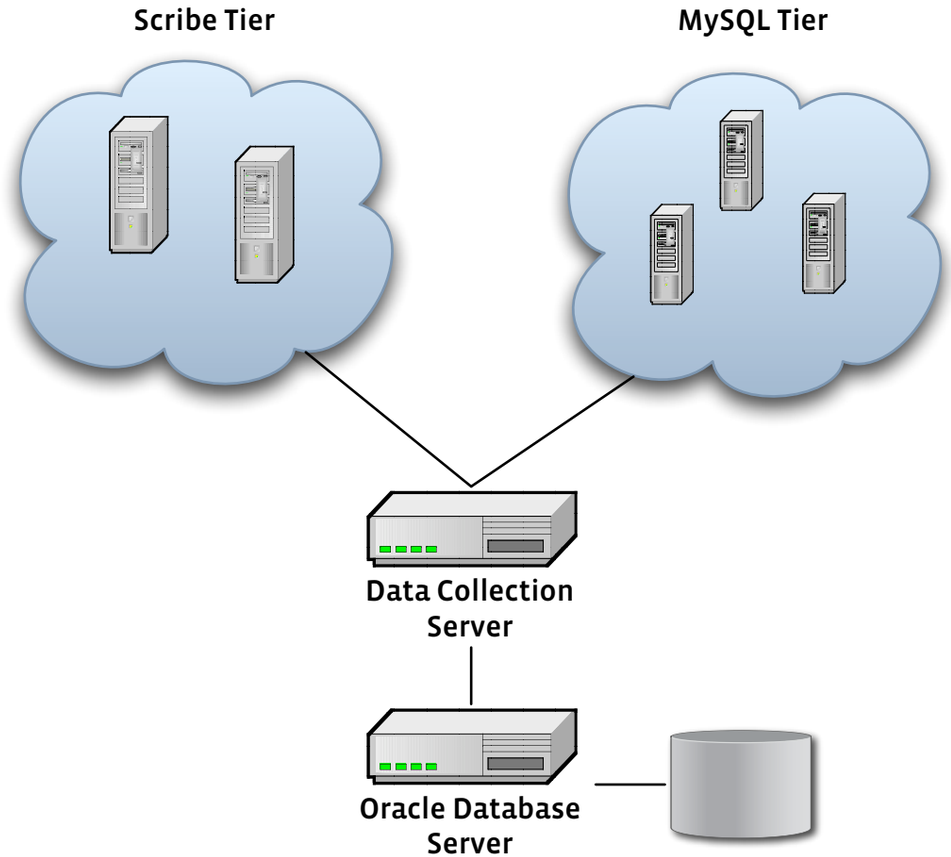
- Source data living on horizontally partitioned MySQL tier
- Intensive historical analysis difficult
- No way to assess impact of changes to the site

- First try: Python scripts pull data into MySQL
- Second try: Python scripts pull data into Oracle

- ...and then we turned on impression logging

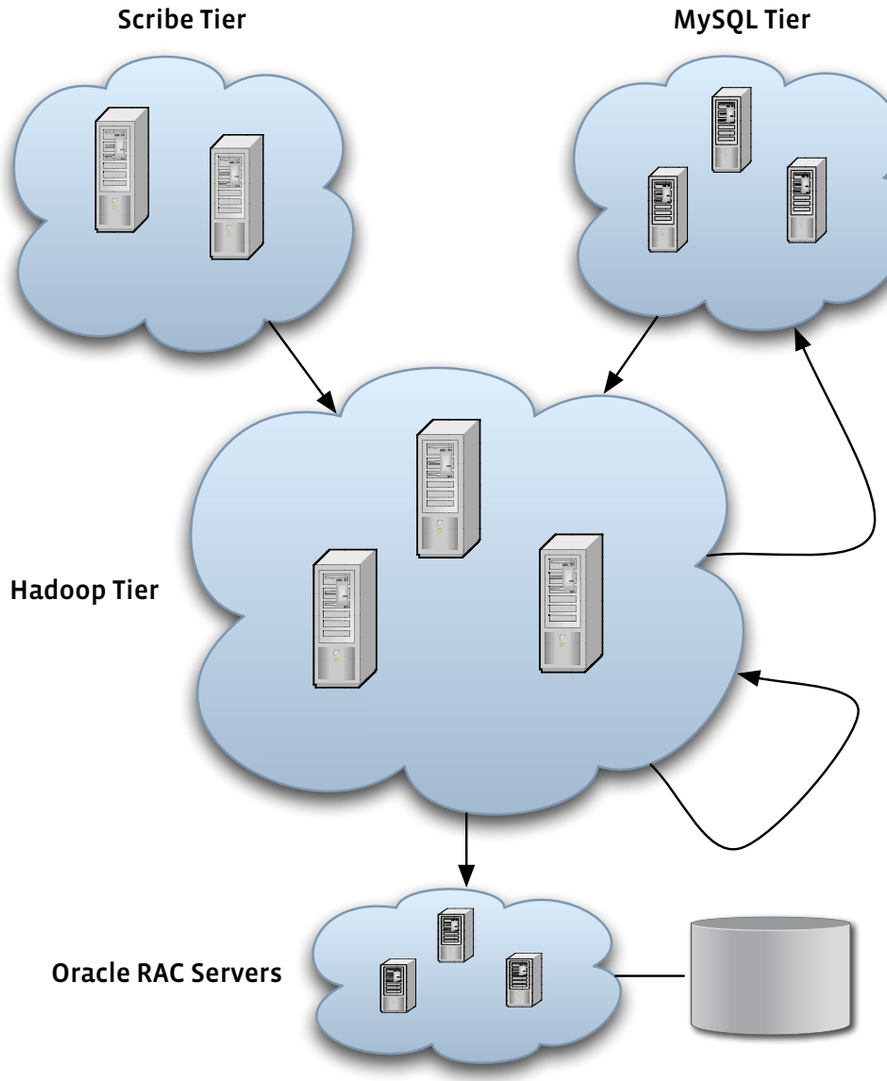
Facebook Data Infrastructure

2007



Facebook Data Infrastructure

2008



Major Data Team Workloads

- Data collection
  - server logs
  - application databases
  - web crawls
- Thousands of multi-stage processing pipelines
  - Summaries consumed by external users
  - Summaries for internal reporting
  - Ad optimization pipeline
  - Experimentation platform pipeline
- Ad hoc analyses

Other Workloads

Keeping the Cluster Busy

- Parameterized queries from business analysts
- Data transformations and data integrity enforcement
- Document indexing
- Feature generation pipelines for machine learning
- Model building and publishing for machine learning
- Storage system bulk loading

Facebook Hardware Statistics

- 4 data centers
  - two on west coast, two on east coast
- Around 20,000 Servers
  - 15,000 Apache/PHP/APC
  - 1,500 MySQL
  - 700 Hadoop
  - 500 Memcache
  - 100 Cassandra
  - Also: Search, Ads, News Feed, etc.

Facebook Workload Statistics

- Relative data volumes
  - Cassandra: 40 TB
  - MySQL: 60 TB
  - Haystack: 1 PB
  - Hadoop: 2.5 PB
- Hadoop Statistics
  - ingests 15 TB per day
  - processes 55 TB per day with 4,000 jobs per day
  - generates 15 TB of intermediate data per day
- Hadoop tier not retiring data!

Hadoop at Yahoo!

- Jan 2006: Hired Doug Cutting
- Apr 2006: Sorted 1.9 TB on 188 nodes in 47 hours
- Apr 2008: Sorted 1 TB on 910 nodes in 209 seconds
- Aug 2008: Deployed 4,000 node Hadoop cluster
- Data Points
  - Over 20,000 nodes running Hadoop
  - Hundreds of thousands of jobs per day
  - Typical HDFS cluster: 1,400 nodes, 2 PB capacity
  - Largest shuffle is 450 TB
  - Workload: 42% Streaming, 28% Pig, 28% Java

Example Hadoop Applications

- Yahoo!
  - Yahoo! Search Webmap
  - Processing news and content feeds
  - Content and ad targeting optimization
- Facebook
  - Fraud and abuse detection
  - Lexicon
- Cloudera
  - Facial recognition for automatic tagging
  - Next-generation genome sequence analysis

The Future of Data Processing

Hadoop, the Browser, and Collaboration

- "The Unreasonable Effectiveness of Data"
- Single namespace for your organization's bits
- Single engine for distributed data processing
- Materialization of structured subsets into optimized stores
- Browser as client interface with focus on user experience
- The system gets better over time using workload information
- Cloning and sharing of common libraries and workflows
- Global metadata store driving collection, analysis, and reporting

Data Points: Global

- 8 million servers shipped per year (IDC)
  - 20% go to web companies (Rick Rashid)
  - 33% go to HPC (Andy Bechtolsheim)
- 2.5 exabytes of external storage shipped per year (IDC)
- Data center costs (James Hamilton)
  - 45% servers
  - 25% power and cooling hardware
  - 15% power draw
  - 15% network
- Jim Gray
  - “Disks will replace tapes, and disks will have infinite capacity. Period.”
  - “Processors are going to migrate to where the transducers are.”

Hadoop is Everywhere

last.fm

A9

nhn

CARRIER iQ

Powerset

Google

quantcast

facebook

Autodesk

YAHOO!

ebay

hp invent

joost

intel

The Historical New York Times Project

AOL

mailtrust
A DIVISION OF RACKSPACE

Integrating Hadoop into the Enterprise

- Configuration: Chef, Puppet, Bcfg2, Cfengine
- Deployment: iClassify, Capistrano, Puppet
- Monitoring and Alerting: Ganglia, Nagios, Cacti, Hyperic
- File System Interfaces: NFS, FUSE, Samba, GridFTP, WebDAV
- ETL: Informatica, Ab Initio, DataStage
- ESB: Mule, XMPP, JMS, WebSphere
- Workflow: Quartz, YAWL
- Databases: DBInputFormat, upcoming Cloudera tools
- BI: MicroStrategy, QlikView, JasperSoft

“MAD” Skills

Hellerstein et al., VLDB 2009

- Magnetic
  - We referred to HDFS as our “gaping maw of bits”: store it all!
  - Disintermediate Data team for persisting data
- Agile
  - Throw out schemas and support diverse serialization formats
- Deep
  - Hive; support for sampling and R/Excel export
  - Libraries for common statistics and machine learning tasks
- Hadoop used like staging and production tier in paper

The Rise of the Data Scientist

Leaders of the Data Revolution

- Data Scientists play four roles
  - Statistician
  - Coder
  - Customer Service Rep
  - Product Manager

- Build data intensive products and services in addition to analyses
- Storage and processing layers should learn from their habits
- Collaboration features should disseminate learned knowledge

Cloudera Founding Team

Turning Data into Awesome since 1986

- Mike Olson
  - Sleepycat Software (Berkeley DB), Illustra (PostgreSQL), and many more
- Amr Awadallah
  - VP of Yahoo! Product Intelligence Engineering
- Jeff Hammerbacher
  - Facebook Data Team: Thrift, Scribe, Hive, Cassandra; SIGMOD, CHI, ICWSM
- Christophe Bisciglia
  - Google Personalized Search, UW Hadoop course, Google/IBM Academic Cluster, NSF CluE Program

Cloudera's Distribution for Hadoop

- Sane packaging for standard Linux service management
- Version matching between Hadoop and related subprojects
- Bundled as AMI with utility scripts for easy prototyping
- Stable release management process

- Future releases
  - Hive server and HBase support
  - Ganglia and Scribe for monitoring and logfile aggregation
  - Improved tools for authoring, debugging, and monitoring jobs
  - Utilities for import and export from RDBMS

Cloudera Training

- Freely available basic training
- Basic and Advanced courses delivered in L.A. in May
- Focused on developing solutions with MapReduce, Pig, and Hive

- At Facebook, internal education was a significant burden
- Cloudera can help design internal curricula to aid in adoption
- We can also develop literature to educate internal executives

Cloudera Support

- Installation and upgrades using our hardened distribution
- Custom integration with ETL and BI tools
- Design reviews
  - Processing pipelines
  - Operations framework: configuration, monitoring, alerting
  - Algorithm development
- Bug fixing and troubleshooting
- Profiling and performance optimizations
- Prioritized feature development for Hadoop Core and CDH
- Regression testing of common workloads

Research Problems

HDFS, part one

- Handle small files
  - Optimize read and write of small objects
  - Partition metadata or page to disk
- Single namespace across data centers
- Access control, encryption, and other security measures
- Hardware optimizations
  - Integration of Flash, low power CPUs
  - Tiered storage
  - Multicore, especially local filesystem optimizations

Research Problems

HDFS, part two

- Global snapshots and recovery
- Pluggable block placement
- High availability
- More granular quality of service, especially for anti-entropy tasks
- Local write optimizations for database workloads
- Multiple-writer appends
- Different file system interfaces: SMB, GridFTP, pNFS, S3, HDF5
- Client statistics and application hints

Research Problems

MapReduce, part one

- Multi-stage MapReduce
- Improved authoring environments
  - Domain-specific libraries and DSLs
  - Testing harness and debugging tools
- Performance
  - Profiling
  - Shuffle stage optimization
  - Pipelining
  - Small job performance

Research Problems

MapReduce, part two

- Job scheduling
  - Memory-aware scheduling
  - Currency-based scheduling (cf. Thomas Sandholm)
  - Adaptive optimization
- Streaming MapReduce
- Separate JobScheduler from JobManager

Research Problems

Hive, part one

- Support for schema evolution
- Iterative construction of complex queries
- Columnar storage
- Statistics collection and cost-based query optimization
- Optimized block placement algorithms
  - Static: schema analysis
  - Dynamic: workload analysis
- Novel join algorithms

Research Problems

Hive, part two

- Learn structure from data, e.g. PADS
- Store source and reporting metadata in MetaStore
- Indexing
- Compression
- Further SQL compliance
- Advanced operators, like cubes and frequent item sets

- (Thanks, Joydeep)

Research Problems

General

- Education of engineers and analysts
  - Tools for mapping existing workloads
  - Tools for integration with existing environments
- Disk and wire format: Thrift, Avro, Protocol Buffers
- Table storage: HBase, HyperTable, Cassandra, Redis, Project Voldemort, Scalaris, CouchDB, MongoDB, Tokyo Cabinet, Drizzle
  - Jeez
- Other higher-order services
- Get better over time