Hadoop Introduction

Sang Shin
JPassion.com
"Learn with Passion!"

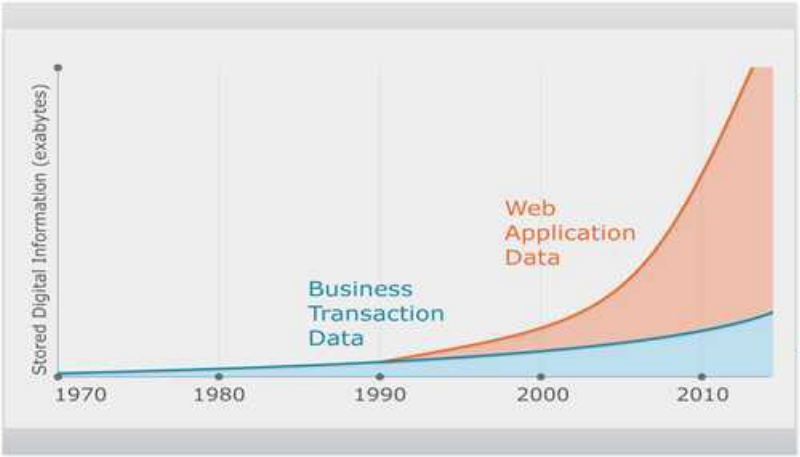
Topics

- Big Data Analytics
- What is and Why Hadoop?
- Comparing Hadoop with other technologies
- Hadoop architecture
- Hadoop ecosystem
- Hadoop usage examples

Big Data Analytics

What is Big Data?

- Big data is a collection of data sets so large and complex that it becomes difficult to process using traditional data processing technologies
  - How do you capture, store, search, share, transfer, analyze, and visualize big data?
- Unstructured data is exploding



Complex, Unstructured
- Text
- Images
- Audio
- Video …

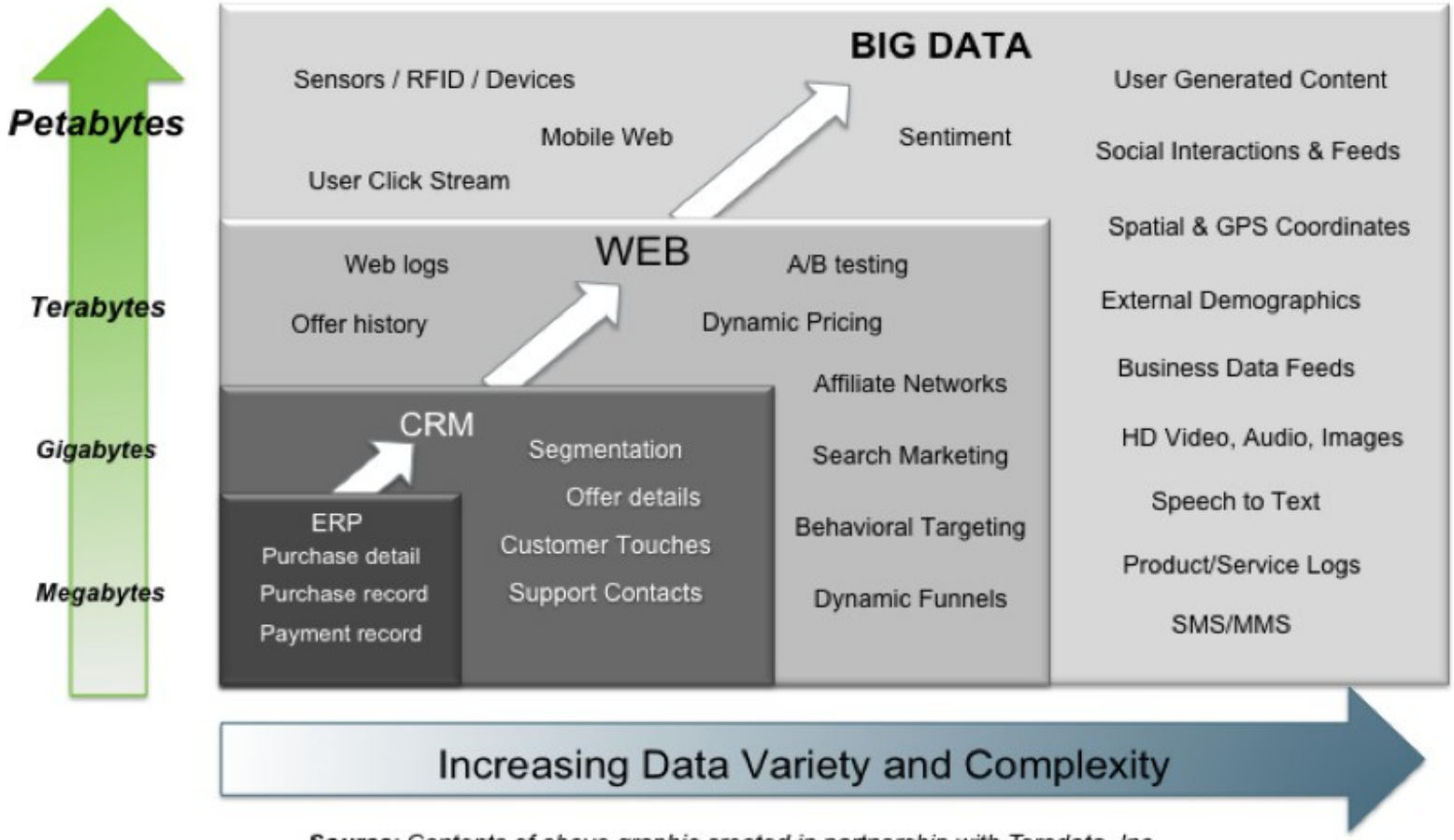
Relational

Big Data Examples

- Facebook
  - Has ~400 terabytes of data
  - ~20 terabytes of new data per day
- New York Stock Exchange (NYSE)
  - Generated 1 terabyte of trade data per day
- Internet Archive
  - Stores ~2 perabytes of data
  - Grows at a rate of 20 terabytes per month
  - http://archive.org/web/web.php
- Skybox imaging (satellite images)
  - 1 terabyte per day
- Ancestry.com
  - Stores around ~2.5 perabytes of data

Big Data Evolution

Big Data = Transactions + Interactions + Observations



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Challenges in Big Data (Using Traditional Data Processing System)

- Slow to process
  - Takes 11 days to read ~100TB on a single computer
- Slow to move data
  - Moving big data over the network is slow
- Can't scale
  - Scaling up vertically (more memory, more powerful hardware) has limitation
- Hard drive capacity is limited
  - A single hard drive cannot accommodate the size of big data
- Unreliable
  - Failures in computing devices are inevitable

Big Data Characteristics

- Very large, distributed aggregations of loosely structured (or unstructured) data
- Petabytes/exabytes of data
- Flat schemas with few complex interrelationships
- Often involving time-stamped events
- Often made up of incomplete data

Challenges in Big Data Handling

- Moving large amount of data from storage cluster to computation cluster is not feasible (cost and performance wise)
  - Instead, moving computing logic to the data is more efficient
- In a large cluster, failure is expected
  - Machines, networks break down
- It is expensive to build reliability in each application
  - Reliability should be handled by the framework

Things to Consider in Big Data Analytics

Examples of Public Data Sets

- U.S. Government
  - http://usgovxml.com/
- Amazon
  - http://aws.amazon.com/public-data-sets/
- Weather data from NCDC
  - http://www.ncdc.noaa.gov/data-access
- Million Song Dataset
  - http://labrosa.ee.columbia.edu/millionsong/

What is and Why Hadoop?

What is Hadoop?

- Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of commodity computers using a simple programming model
- It is designed to scale up from a single node to thousands of nodes, each providing computation and storage

Hadoop Historical Background

- Started as a sub-project of Apache Nutch
  - Nutch is for indexing the web and expose it for searching
  - Open source alternative to Google
  - Started by Doug Cutting
- In 2004, Google publishes Google File System (GFS) and Map Reduce framework papers
- Doug Cutting and Nutch team implemented GFS and Map Reduce in Nutch
- In 2006, Yahoo! hires Doug Cutting to work on Hadoop
- In 2008, Hadoop became Apache Top Level project
  - http://hadoop.apache.org

Why is Hadoop?

- Scalable – by simply adding new nodes to the cluster
- Economical – by using commodity machines
- Efficient – by running tasks where data is located
- Flexible – by handling schema-less, non-structured data
- Fault tolerant – by self-discovering failed nodes and self-healing, by redistributing failed jobs and data reliability by replication
- Simple – by using simple programming model
- Evolving – by having vibrant ecosystem

Hadoop Design Principles

- Store and process large amounts of data
- Scales on performance, storage, processing
- “Computing” (“code”) moves to “Data” (Not the other way around)
- Self-heals – recovery from failures is built-in
- Designed to run on commodity hardware

When to use or not to use Hadoop?

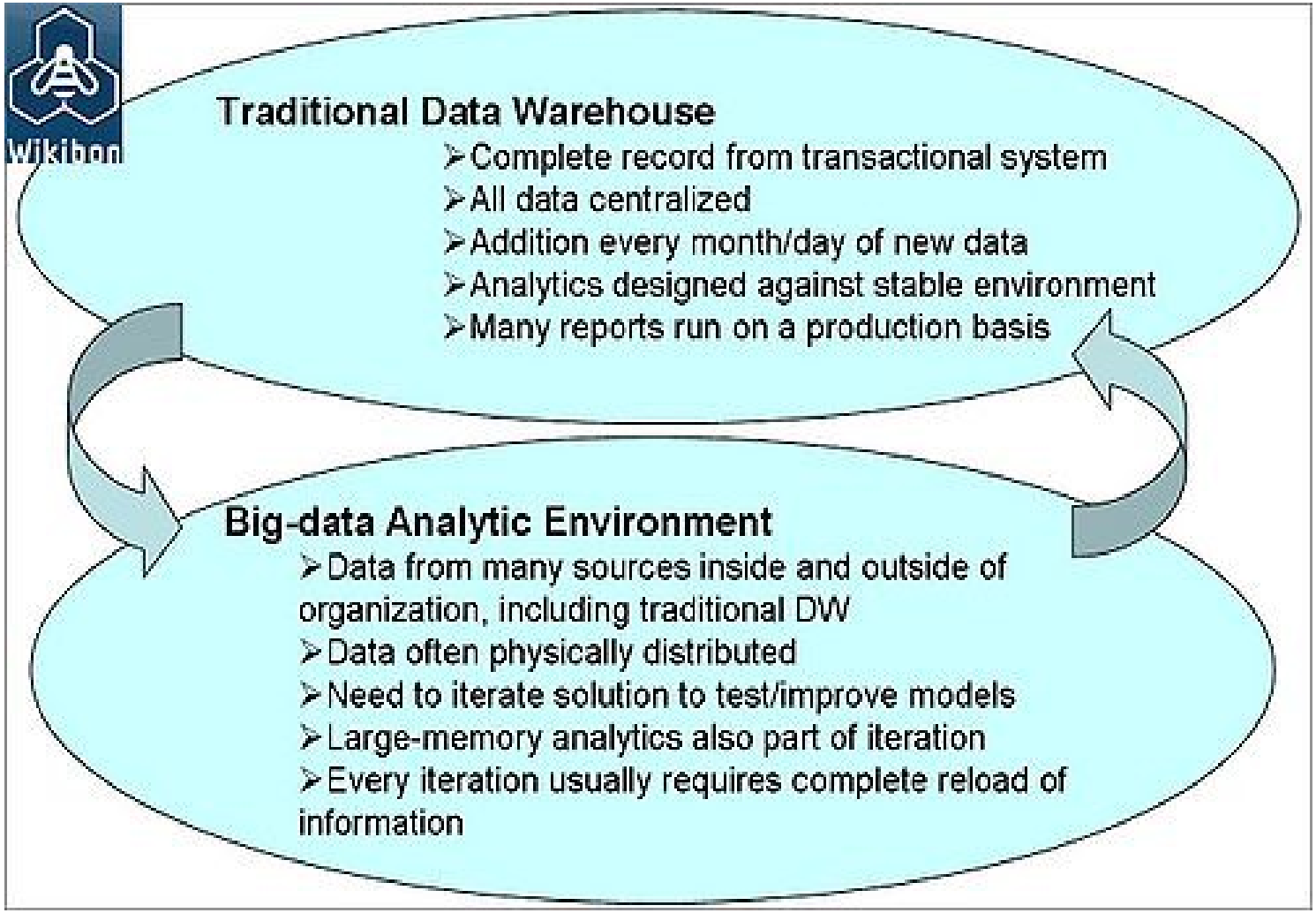
- Hadoop is good for
  - Indexing data
  - Log analysis
  - Image manipulation
  - Sorting large scale data
  - Data mining
- Hadoop is NOT good for
  - Real time processing (Hadoop is batch oriented)
  - Random access (Hadoop is not database)
  - Computation-intensive tasks with little data
- Some limitations of Hadoop are addressed by Hadoop ecosystem technologies, however

Comparing Hadoop with Other Technologies

RDBMS vs. Hadoop

RDBMS	Hadoop
Small data (comp. to Hadoop)	Big data
Strictly structured	Loosely structured
Schema on write	Schema on read
Relational	Non relational
Transaction-oriented	Analytics-oriented
ACID	Non-ACID
Real-time query	Batch-oriented
Frequent write/update	Initial write, No update/delete
Centralized	Distributed
Limited scalability (vertical)	High scalability (horizontal)

Data Warehouse vs. Hadoop



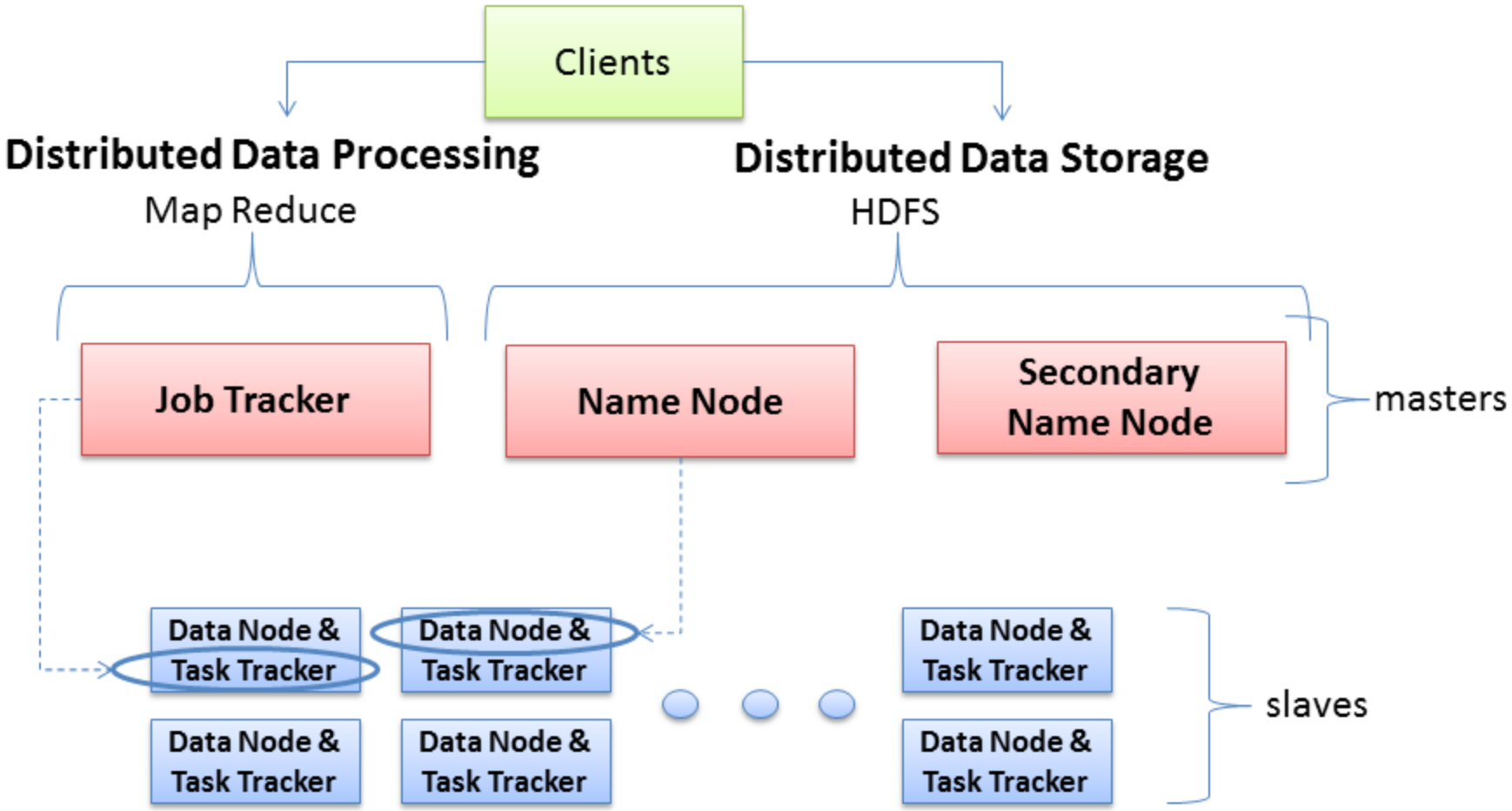
Hadoop Architecture

Hadoop Architecture

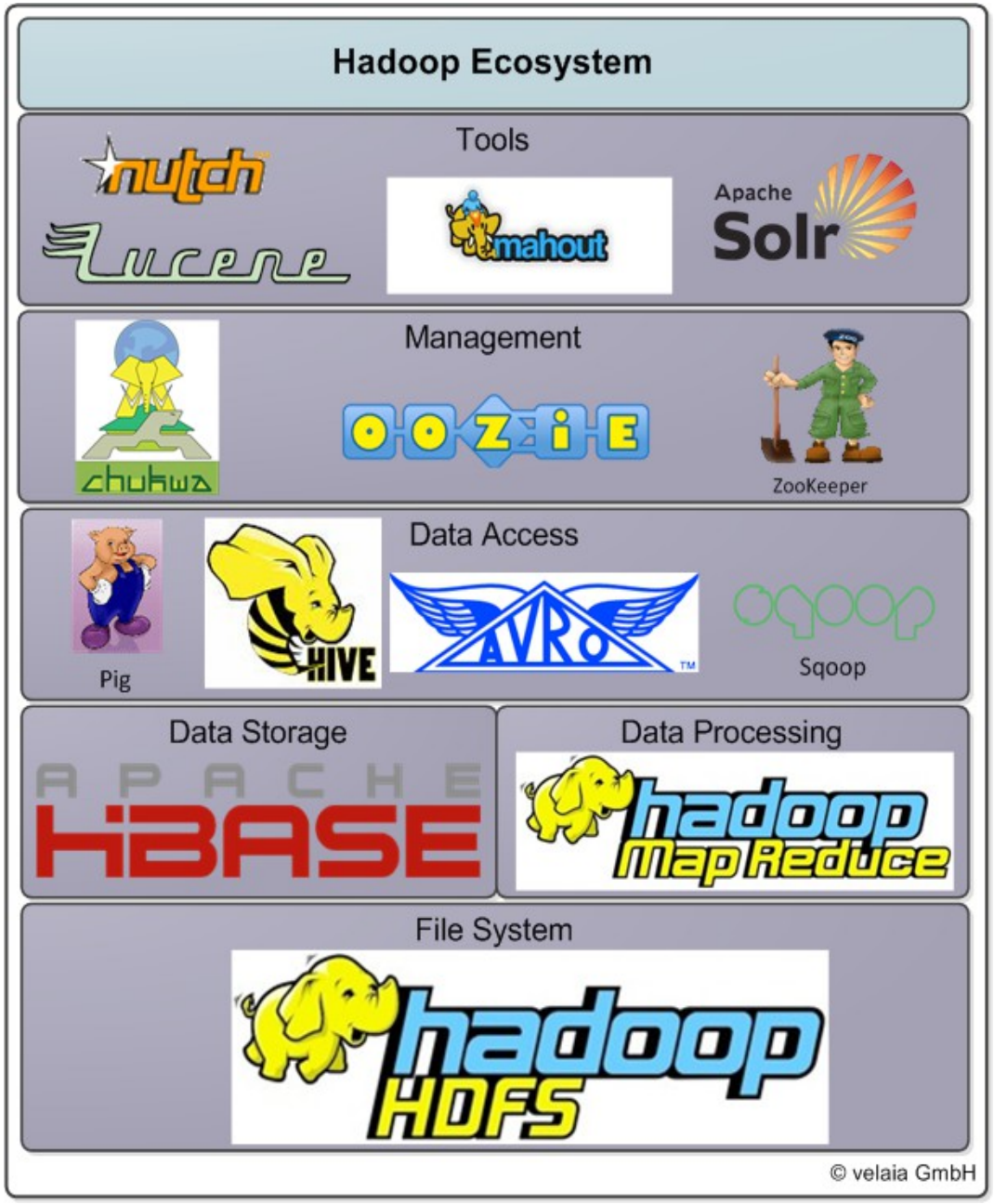
- Hadoop is built on two core technologies: “Map Reduce” for processing and “HDFS” for distributed file system
  - Map Reduce is a framework for performing high performance distributed data processing using the “divide and aggregate” programming paradigm
  - HDFS is a reliable distributed file system that provides high-throughput access to data
- Hadoop has master/slave architecture for both “Map Reduce” and “HDFS”
  - Map Reduce has “Job Tracker” as a master and multiple “Task Trackers” as slaves
  - HDFS has “Name node” as a master and multiple “Data nodes” as slaves

Hadoop Architecture

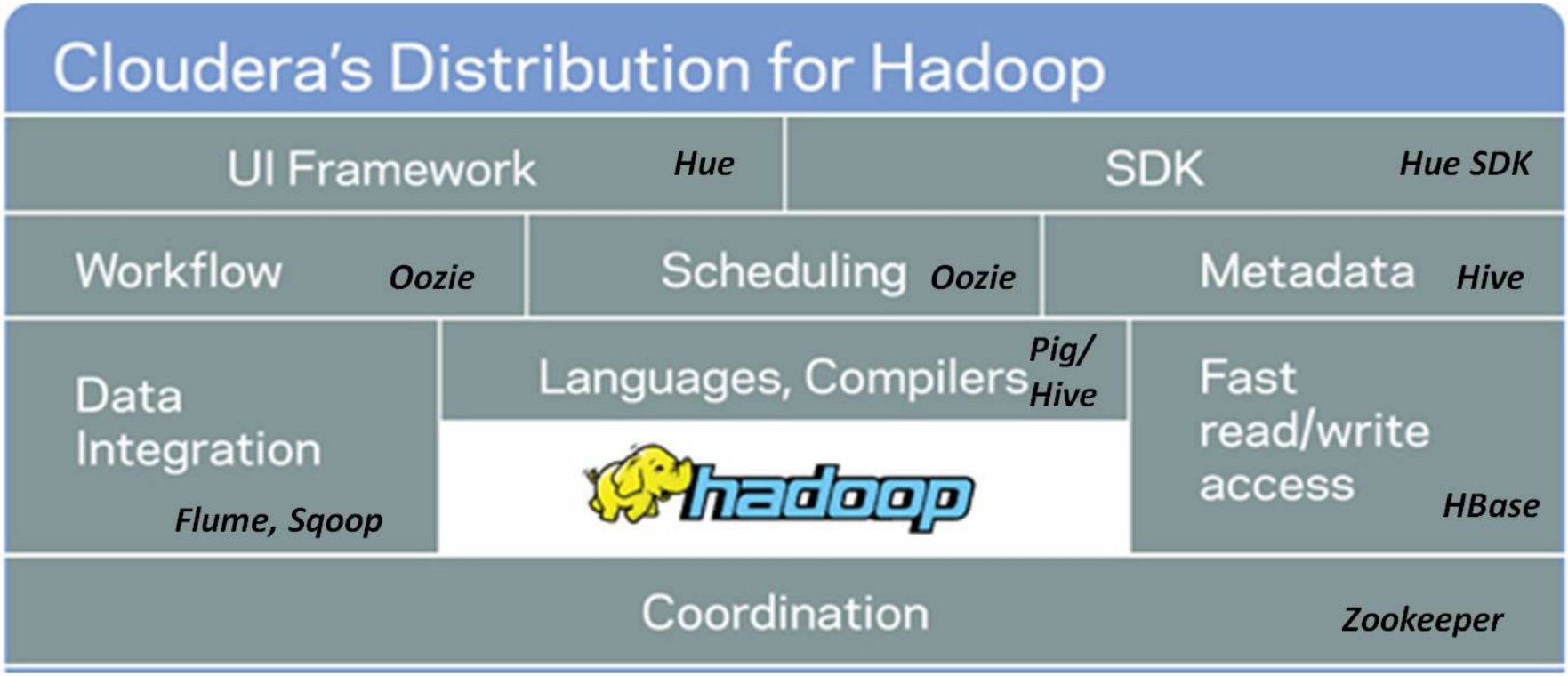
- Hadoop = Map Reduce + HDFS



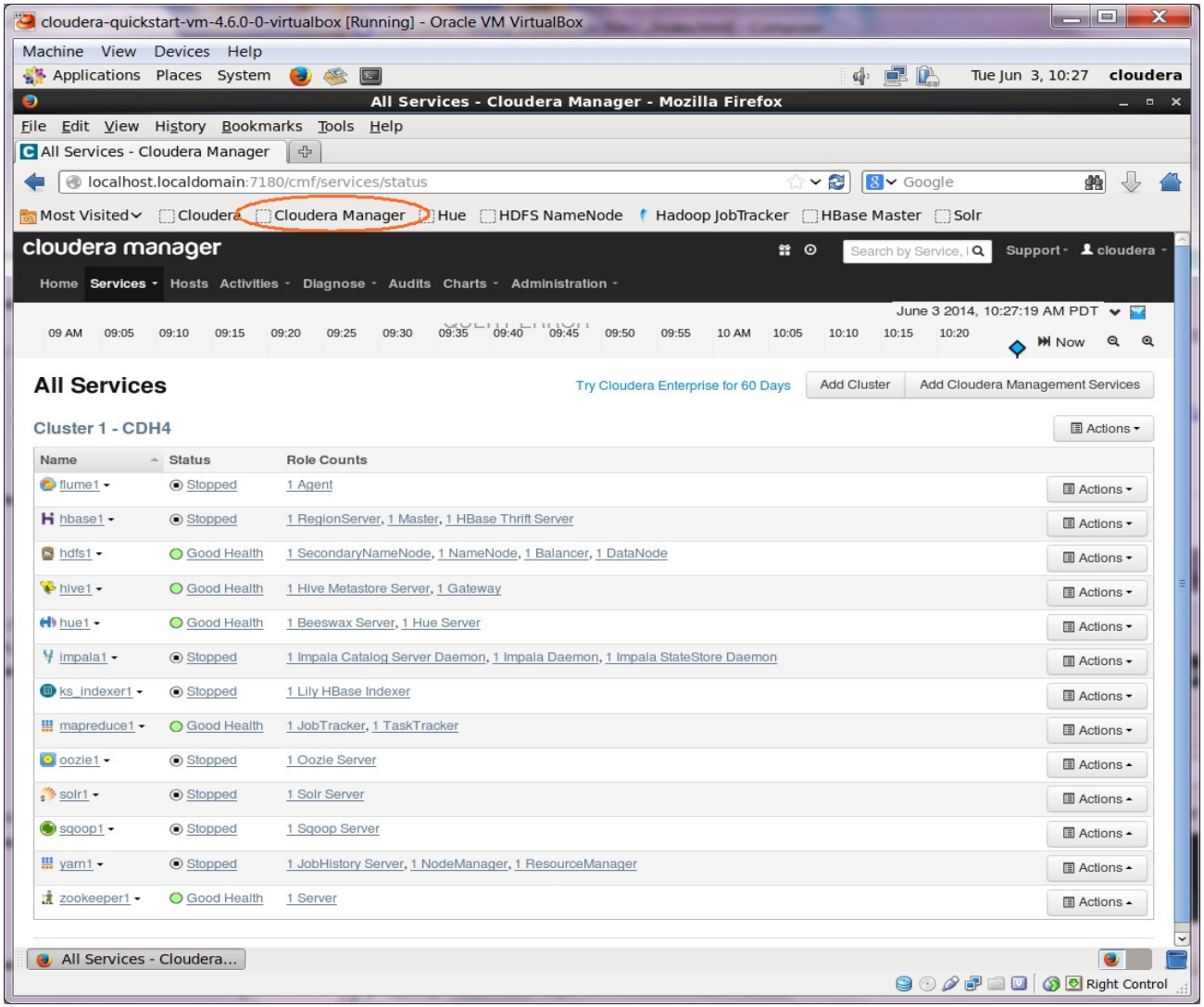
Hadoop Ecosystem


Hadoop Ecosystem
Tools
nutch
lucene
mahout
Apache
Solr
Management
chukwa
OOZIE
ZooKeeper
Data Access
Pig
HIVE
AVRO
Sqoop
Data Storage
APACHE
HBASE
Data Processing
hadoop
Map Reduce
File System
hadoop
HDFS
© velaia GmbH

Cloudera Hadoop Distribution



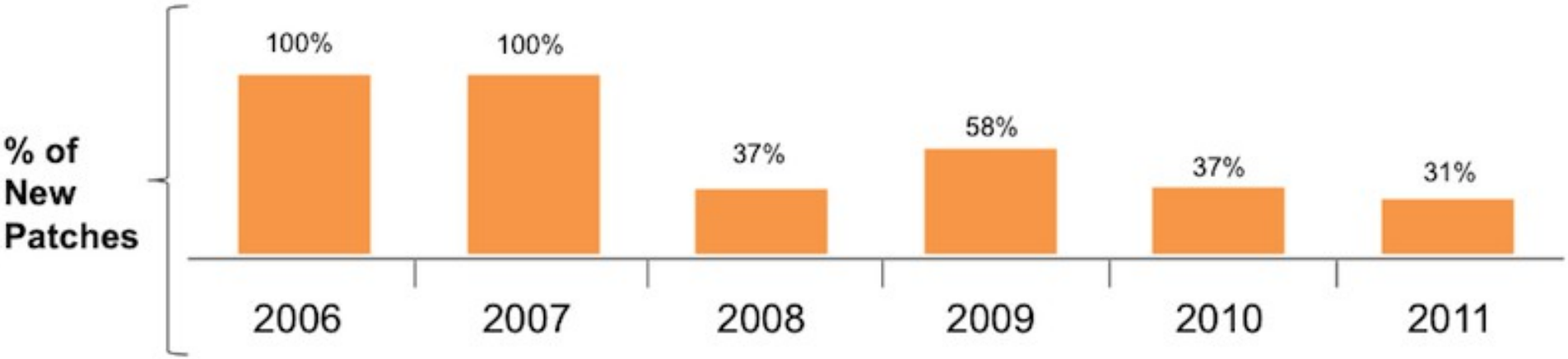
Cloudera Manager



For CDH 4.x.x: Automatically Started

For CDH 5.x.x Needs to be started manually

Hadoop Ecosystem Development



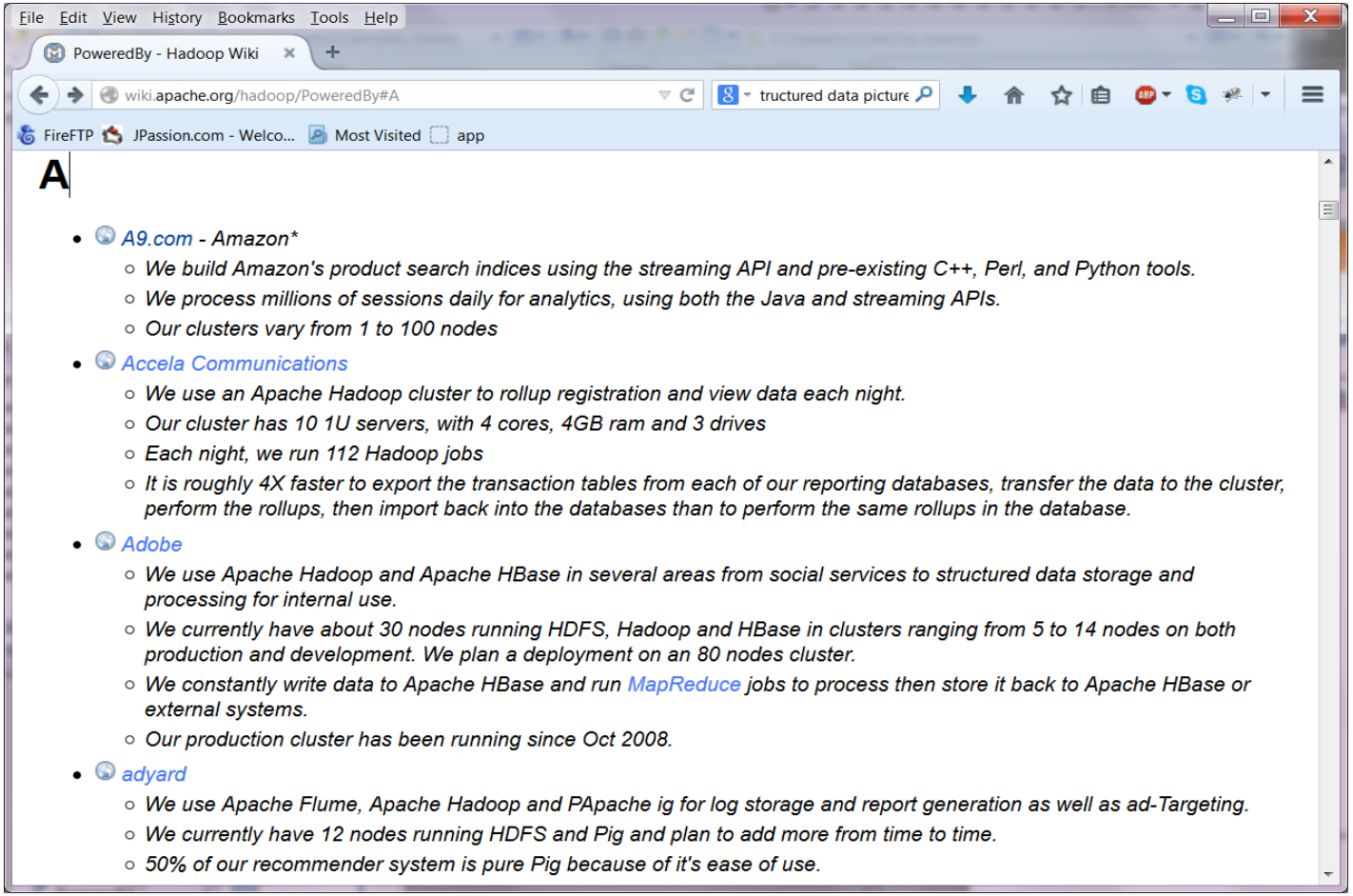
Relevant Projects

2006	2007	2008	2009	2010	2011
• Core Hadoop	• Core Hadoop	• Core Hadoop • HBase • ZooKeeper • Mahout	• Core Hadoop • HBase • ZooKeeper • Mahout • Pig • Hive	• Core Hadoop • HBase • ZooKeeper • Mahout • Pig • Hive • Flume • Avro • Whirr • Sqoop	• Core Hadoop • HBase • ZooKeeper • Mahout • Pig • Hive • Avro • Flume • Whirr • Sqoop • HCatalog • Mrunit • Bigtop • Oozie

Hadoop Usage Examples

Institutions that are using Hadoop

- http://wiki.apache.org/hadoop/PoweredBy



File Edit View History Bookmarks Tools Help

PoweredBy - Hadoop Wiki

wiki.apache.org/hadoop/PoweredBy#A

FireFTP JPassion.com - Welco... Most Visited app

A

- *A9.com - Amazon**
  - *We build Amazon's product search indices using the streaming API and pre-existing C++, Perl, and Python tools.*
  - *We process millions of sessions daily for analytics, using both the Java and streaming APIs.*
  - *Our clusters vary from 1 to 100 nodes*
- *Accela Communications*
  - *We use an Apache Hadoop cluster to rollup registration and view data each night.*
  - *Our cluster has 10 1U servers, with 4 cores, 4GB ram and 3 drives*
  - *Each night, we run 112 Hadoop jobs*
  - *It is roughly 4X faster to export the transaction tables from each of our reporting databases, transfer the data to the cluster, perform the rollups, then import back into the databases than to perform the same rollups in the database.*
- *Adobe*
  - *We use Apache Hadoop and Apache HBase in several areas from social services to structured data storage and processing for internal use.*
  - *We currently have about 30 nodes running HDFS, Hadoop and HBase in clusters ranging from 5 to 14 nodes on both production and development. We plan a deployment on an 80 nodes cluster.*
  - *We constantly write data to Apache HBase and run MapReduce jobs to process then store it back to Apache HBase or external systems.*
  - *Our production cluster has been running since Oct 2008.*
- *adyard*
  - *We use Apache Flume, Apache Hadoop and PApache ig for log storage and report generation as well as ad-Targeting.*
  - *We currently have 12 nodes running HDFS and Pig and plan to add more from time to time.*
  - *50% of our recommender system is pure Pig because of it's ease of use.*

Institutions that are using Hadoop

- Facebook
  - We use Apache Hadoop to store copies of internal log and dimension data sources and use it as a source for reporting/analytics and machine learning.
  - Currently we have 2 major clusters:
    - A 1100-machine cluster with 8800 cores and about 12 PB raw storage
    - A 300-machine cluster with 2400 cores and about 3 PB raw storage
    - Each (commodity) node has 8 cores and 12 TB of storage
  - We are heavy users of both streaming as well as the Java APIs
  - We have built a higher level data warehousing framework using Hive
  - We have also developed a FUSE implementation over HDFS.

Institutions that are using Hadoop

- EBay
  - 532 nodes cluster (8 * 532 cores, 5.3PB).
  - Heavy usage of Java MapReduce, Apache Pig, Apache Hive, Apache HBase
  - Using it for Search optimization and Research

Institutions that are using Hadoop

- Twitter
  - We use Apache Hadoop to store and process tweets, log files, and many other types of data generated across Twitter
  - We store all data as compressed LZO files.
  - We use both Scala and Java to access Hadoop's MapReduce APIs
  - We use Apache Pig heavily for both scheduled and ad-hoc jobs, due to its ability to accomplish a lot with few statements.
  - We employ committers on Apache Pig, Apache Avro, Apache Hive, and Apache Cassandra, and contribute much of our internal Hadoop work to opensource (see hadoop-lzo)

Institutions that are using Hadoop

- Yahoo
  - More than 100,000 CPUs in >40,000 computers running Hadoop
  - Our biggest cluster: 4500 nodes (2*4cpu boxes w 4*1TB disk & 16GB RAM)
    - Used to support research for Ad Systems and Web Search
    - Also used to do scaling tests to support development of Apache Hadoop on larger clusters
  - Our Blog - Learn more about how we use Apache Hadoop.
  - >60% of Hadoop Jobs within Yahoo are Apache Pig jobs.

Learn with Passion!
JPassion.com