An Introduction to Big Data, Apache Hadoop, and Cloudera

Ian Wrigley, Curriculum Manager, Cloudera
The Motivation for Hadoop
Traditional Large-Scale Computation

- Traditionally, computation has been processor-bound
  - Relatively small amounts of data
  - Significant amount of complex processing performed on that data

- For decades, the primary push was to increase the computing power of a single machine
  - Faster processor, more RAM
The Data Explosion

1.8 trillion gigabytes of data was created in 2011...

- More than 90% is unstructured data
- Approx. 500 quadrillion files
- Quantity doubles every 2 years

Source: IDC 2011
Current Solutions

Current Database Solutions are designed for structured data.

- Optimized to answer known questions quickly
- Schemas dictate form/context
- Difficult to adapt to new data types and new questions
- Expensive at Petabyte scale
Why Use Hadoop?

Move beyond rigid legacy frameworks

Hadoop handles any data type, in any quantity
- Structured, unstructured
- Schema, no schema
- High volume, low volume
- All kinds of analytic applications

Hadoop grows with your business
- Proven at petabyte scale
- Capacity and performance grow simultaneously
- Leverages commodity hardware to mitigate costs

Hadoop is 100% Apache® licensed and open source
- No vendor lock-in
- Community development
- Rich ecosystem of related projects

Hadoop helps you derive the complete value of all your data
- Drives revenue by extracting value from data that was previously out of reach
- Controls costs by storing data more affordably than any other platform

1 2 3
The Origins of Hadoop

2002
- Open source web crawler project created by Doug Cutting

2007
- Publishes MapReduce and GFS Paper
- Open Source MapReduce and HDFS project created by Doug Cutting
- Runs 4,000-node Hadoop cluster
- Hadoop wins Terabyte sort benchmark

2012
- Launches SQL support for Hadoop
- Releases CDH and Cloudera Enterprise
Self-healing, high bandwidth clustered storage

HDFS breaks incoming files into blocks and stores them redundantly across the cluster.
Processes large jobs in parallel across many nodes and combines the results.
Hadoop and Databases

You need **both tools**

**Relational Database**

- Interactive OLAP Analytics (<1sec)
- Multistep ACID Transactions
- 100% SQL Compliance

**Hadoop**

- Structured or Not (Flexibility)
- Scalability of Storage/Compute
- Complex Data Processing
Typical Datacenter Architecture

Enterprise web site

Interactive database

Data export

OLAP load

Oracle, SAP...

Business intelligence apps
Adding Hadoop To The Mix

Enterprise web site

Interactive database

New data

Hadoop

Dynamic OLAP queries

Recommendations, etc...

Business intelligence apps

Oracle, SAP...
Why Cloudera?
Cloudera is...

#1 in Customers and Users

3 of the top 5 organizations in banking, telecommunications, mobile services, defense & intelligence, media and retail depend on Cloudera

More enterprise customers and users than all other Hadoop systems combined

#1 in Integrated Partners

More than 100 partners across hardware, platforms, database and business intelligence (BI)

The industry’s only certification program for hardware, platforms, software and services

#1 in Training and Certification

More than 7,000 developers, administrators and managers trained on 6 continents since 2009

The industry’s only certification program for developers, administrators and managers

#1 in Nodes Under Management

#1 in Open Source Contributions

#1 in Data Science
Experienced and Proven Across Hundreds of Deployments
The Only Vendor With a Complete Solution

Cloudera’s Distribution Including Apache Hadoop (CDH)
Big Data storage, processing and analytics platform based on Apache Hadoop – 100% open source

Cloudera Enterprise 4.0

Cloudera Manager
End-to-end management application for the deployment and operation of CDH

Production Support
Our team of experts on call to help you meet your Service Level Agreements (SLAs)

Partner Ecosystem
250+ partners across hardware, software, platforms and services

Cloudera University
Equipping the Big Data workforce – 12,000+ trained

Professional Services
Use case discovery, pilots, process & team development
Solving Problems with Hadoop
<table>
<thead>
<tr>
<th></th>
<th>Eight Common Hadoop-able Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Modeling true risk</td>
</tr>
<tr>
<td>2.</td>
<td>Customer churn analysis</td>
</tr>
<tr>
<td>3.</td>
<td>Recommendation engine</td>
</tr>
<tr>
<td>4.</td>
<td>PoS transaction analysis</td>
</tr>
<tr>
<td>5.</td>
<td>Analyzing network data to predict failure</td>
</tr>
<tr>
<td>6.</td>
<td>Threat analysis</td>
</tr>
<tr>
<td>7.</td>
<td>Search quality</td>
</tr>
<tr>
<td>8.</td>
<td>Data “sandbox”</td>
</tr>
</tbody>
</table>
1. Modeling True Risk

Challenge:

- How much risk exposure does an organization really have with each customer?
  - Multiple sources of data and across multiple lines of business

Solution with Hadoop:

- Source and aggregate disparate data sources to build data picture
  - e.g. credit card records, call recordings, chat sessions, emails, banking activity

- Structure and analyze
  - Sentiment analysis, graph creation, pattern recognition

Typical Industry:

- Financial Services (banks, insurance companies)
2. Customer Churn Analysis

Challenge:

- Why is an organization really losing customers?
  - Data on these factors comes from different sources

Solution with Hadoop:

- Rapidly build behavioral model from disparate data sources
- Structure and analyze with Hadoop
  - Traversing
  - Graph creation
  - Pattern recognition

Typical Industry:
- Telecommunications, Financial Services
3. Recommendation Engine/Ad Targeting

Challenge:

- Using user data to predict which products to recommend

Solution with Hadoop:

- Batch processing framework
  - Allow execution in parallel over large datasets

- Collaborative filtering
  - Collecting ‘taste’ information from many users
  - Utilizing information to predict what similar users like

Typical Industry
- Ecommerce, Manufacturing, Retail
- Advertising
4. Point of Sale Transaction Analysis

Challenge:

- Analyzing Point of Sale (PoS) data to target promotions and manage operations
  - Sources are complex and data volumes grow across chains of stores and other sources

Solution with Hadoop:

- Batch processing framework
  - Allow execution in parallel over large datasets

- Pattern recognition
  - Optimizing over multiple data sources
  - Utilizing information to predict demand

Typical Industry:

- Retail
5. Analyzing Network Data to Predict Failure

Challenge:

- Analyzing real-time data series from a network of sensors
  - Calculating average frequency over time is extremely tedious because of the need to analyze terabytes

Solution with Hadoop:

- Take the computation to the data
  - Expand from simple scans to more complex data mining

- Better understand how the network reacts to fluctuations
  - Discrete anomalies may, in fact, be interconnected

- Identify leading indicators of component failure

Typical Industry:

- Utilities, Telecommunications, Data Centers
6. Threat Analysis/Trade Surveillance

**Challenge:**

- Detecting threats in the form of fraudulent activity or attacks
  - Large data volumes involved
  - Like looking for a needle in a haystack

**Solution with Hadoop:**

- Parallel processing over huge datasets
- Pattern recognition to identify anomalies,
  - i.e., threats

**Typical Industry:**

- Security, Financial Services,
  General: spam fighting, click fraud
7. Search Quality

**Challenge:**
- Providing real time meaningful search results

**Solution with Hadoop:**
- Analyzing search attempts in conjunction with structured data
- Pattern recognition
  - Browsing pattern of users performing searches in different categories

**Typical Industry:**
- Web, Ecommerce
8. Data “Sandbox”

Challenge:

- Data Deluge
  - Don’t know what to do with the data or what analysis to run

Solution with Hadoop:

- “Dump” all this data into an HDFS cluster
- Use Hadoop to start trying out different analysis on the data
- See patterns to derive value from data

Typical Industry:
- Common across all industries
Orbitz: Major Online Travel Booking Service

- **Challenge:**
  - Orbitz performs millions of searches and transactions daily, which leads to hundreds of gigabytes of log data every day
  - Not all of that data has value (i.e., it is logged for historic reasons)
  - Much is quite valuable
  - Want to capture even more data

- **Solution with Hadoop:**
  - Hadoop provides Orbitz with efficient, economical, scalable, and reliable storage and processing of these large amounts of data
  - Hadoop places no constraints on how data is processed
Before Hadoop

- Orbitz’s data warehouse contains a full archive of all transactions
  - Every booking, refund, cancellation etc.
- Non-transactional data was thrown away because it was uneconomical to store

| Non-transactional Data (e.g., Searches) | Transactional Data (e.g., Bookings) | Data Warehouse |
After Hadoop

- Hadoop was deployed late 2009/early 2010 to begin collecting this non-transactional data
  - Orbitz has been using CDH for that entire period with great success.

- Much of this non-transactional data is contained in Web analytics logs
What Now?

- **Access to this non-transactional data enables a number of applications...**
  - Optimizing hotel search
    - E.g., optimize hotel ranking and show consumers hotels more closely matching their preferences
  - User specific product Recommendations
  - Web page performance tracking
  - Analyses to optimize search result cache performance
  - User segments analysis, which can drive personalization
    - Lots of press coverage in June 2012: company discovered that people using Macs are willing to spend 30% more on hotels that PC users
    - Mac users are now presented with pricier hotels first in the list
Major National Bank

- **Background**
  - 100M customers
  - Relational data: 2.5B records/month
    - Card transactions, home loans, auto loans, etc.
    - Data volume growing by hundreds of TB/year
  - Needs to incorporate non-relational data as well
    - Web clicks, check images, voice data

- **Uses Hadoop to**
  - Identify credit risk, fraud
  - Proactively manage capital
Financial Regulatory Body

- **Stringent data reliability requirements**
  - Must store seven years of data

- **850TB of data collected from every Wall Street trade each year**
  - Data volumes growing at 40% each year

- **Replacing EMC Greenplum + SAN with CDH**
  - Goal is to store data from years two to seven in Hadoop
  - Will have 5PB of data in Hadoop by the end of 2013

- **Cost savings predicted to be 10s of millions of dollars**

- **Application performance testing is showing speed gains of 20x in some cases**
Leading North American Retailer

- **Storing 400TB of data in CDH cluster**
  - Capture and analysis of data on individual customers and SKUs across 4,000 locations

- **Using Hadoop for:**
  - Loyalty program analytics and personal pricing
  - Fraud detection
  - Supply chain optimization
  - Marketing and promotions
  - Locating and pricing overstocked items for clearance
Digital Media Company

- Needs to quickly and reliably process high volume clickstream and pageview data
- Experienced database bottlenecks and reliability issues
- Now using CDH
  - A cluster of just 20 nodes
- Ingesting 75 million clickstream, page view, and user profile events per day
  - 15GB of data
- Processes 430 million records from six million users in 11 minutes
  - Alternative solution would have required 10x more investment in database software, high-end servers, developer time
Hundreds of customers need unique views of the data

Were using Netezza; unable to run more than 2-3 big jobs per day
  - Too expensive to scale

Now using CDH
  - Processing hundreds of jobs concurrently
    - 200-300GB/hour per job
  - Ingesting 10TB of data per day
  - Moving data between CDH, Netezza, and Vertica
Netflix

- **Before Hadoop**
  - Nightly processing of logs
  - Imported into a database
  - Analysis/BI

- As data volume grew, it took more than 24 hours to process and load a day’s worth of logs

- Today, an hourly Hadoop job processes logs for quicker availability to the data for analysis/BI

- Currently ingesting approximately 1TB of data per day
Hadoop as Cheap Storage

- **Yahoo**
  - Before Hadoop: $1 million for 10TB storage
  - With Hadoop: $1 million for 1 PB of storage

- **Other Large Company**
  - Before Hadoop: $5 million to store data in Oracle
  - With Hadoop: $240K to store the data in HDFS

- **Facebook**
  - Hadoop as unified storage
Hadoop Jobs
The Roles People Play

- System Administrators
- Developers
- Analysts
- Data Stewards
System Administrators

- **Required skills:**
  - Strong Linux administration skills
  - Networking knowledge
  - Understanding of hardware

- **Job responsibilities**
  - Install, configure and upgrade Hadoop software
  - Manage hardware components
  - Monitor the cluster
  - Integrate with other systems (e.g., Flume and Sqoop)
Developers

- **Required skills:**
  - Strong Java or scripting capabilities
  - Understanding of MapReduce and algorithms

- **Job responsibilities:**
  - Write, package and deploy MapReduce programs
  - Optimize MapReduce jobs and Hive/Pig programs
Data Analyst/Business Analyst

- **Required skills:**
  - SQL
  - Understanding data analytics/data mining

- **Job responsibilities:**
  - Extract intelligence from the data
  - Write Hive and/or Pig programs
Data Steward

- **Required skills:**
  - Data modeling and ETL
  - Scripting skills

- **Job responsibilities:**
  - Cataloging the data (analogous to a librarian for books)
  - Manage data lifecycle, retention
  - Data quality control with SLAs
Combining Roles

- **System Administrator + Steward analogous to DBA**

- **Required skills:**
  - Data modeling and ETL
  - Scripting skills
  - Strong Linux administration skills

- **Job responsibilities:**
  - Manage data lifecycle, retention
  - Data quality control with SLAs
  - Install, configure and upgrade Hadoop software
  - Manage hardware components
  - Monitor the cluster
  - Integrate with other systems (e.g., Flume and Sqoop)
Finding The Right People

- **Hiring Hadoop experts**
  - Strong Hadoop skills are scarce and expensive
  - Hadoop User Groups
  - Key words
    - Developers: MapReduce, Cloudera Certified Developer for Apache Hadoop (CCDH)
    - System Admins: distributed systems (e.g., Teradata, RedHat Cluster), Linux, Cloudera Certified Administrator for Apache Hadoop (CCAH)

- **Consider cross-training, especially system administrators and data librarians**
Cloudera’s Academic Partnership Program
Cloudera’s Academic Partnerships: Overview

- **Cloudera’s Academic Partnerships (CAP)**
  - An essential component of Cloudera’s strategy to provide comprehensive Apache Hadoop training to current and future data professionals

- **Designed to be a mutually beneficial relationship**
  - Universities are enabled to deliver new and relevant areas of study to their students
  - Cloudera is able to help fill the demand for qualified data professionals to help the market continue is explosive growth

- **With CDH and Cloudera Manager available for free, and our curriculum and Virtual Machine, we provide universities the foundation to start experimenting with Hadoop and developing expertise among their students**
Cloudera’s Academic Partnerships: Goals

- Introduce students to Apache Hadoop
- Provide students and instructors with quality course materials and virtual machine images to complete hands-on labs
- Grant 50% discount on certification costs to students associated with the program who are interested in attempting Cloudera's industry leading Hadoop certification exams
  - Highly recommended they take the class and attempt the certification exam
- Allow academic institutions options to augment their degree program requirements
Cloudera’s Academic Partnerships: Financial Overview

- Cloudera does not currently charge Academic Partners for usage of the training materials
  - This is a program designed solely to facilitate students learning of an emerging technology
  - Our reward is helping the industry grow, and ideally the exposure to Cloudera is a positive one which will be remembered when the students we service today are making decisions for their business tomorrow

- Instructors who are delivering the Cloudera courses are eligible for a 50% discount to commercial training courses delivered by Cloudera
  - We want to make sure the folks leading the classes have the skillset to help their students be successful

- Normally we provide universities with courses focused on the roles of Hadoop Developer or Administrator