What is Hadoop?
How does this stuff work?
Cloudera Technology
Making Hadoop Fast, Easy, and Secure for the Modernized Architecture

Hadoop is a new kind of data platform.
- One place for unlimited data
- Unified data access

Cloudera makes it:
- Fast for business
- Easy to manage
- Secure without compromise
MapReduce: A great tool for its day

The original scalable, general, processing engine of Hadoop ecosystem
- Useful across diverse problem domains
- Fueled initial ecosystem explosion
Enter Apache Spark
Flexible, in-memory data processing for Hadoop

Easier Development
- Rich APIs for Scala, Java, and Python
- Interactive shell

Flexible, Extensible API
- APIs for different types of workloads:
  - Batch
  - Streaming
  - Machine Learning
  - Graph

Faster Processing (Batch & Streaming)
- In-Memory processing and caching
The Spark Ecosystem & Hadoop
Back Slides
Hadoop and Spark Essentials and a few other topics

John Hope – Senior Solutions Engineer
hope@cloudera.com
Our relationship with data is changing - forever.
Data can be a powerful strategic asset

...only if...

data helps achieve your business vision.
Data is Transforming Business

Drive Customer Insights +revenue

Improve Product & Services Efficiency -costs

Lower Business Risk
Exploring Use Cases

• ETL Offload
  • Too much data, too little time, too costly
• Active Archive
  • Save all data vs. moving to slow archival storage
• Mainframe Migration
  • Move costly CPU loads
• Real-time streaming
  • Fraud detection, patient care, transactions

• Data Discovery
• Search All Types of Data
• Predictive Analytics
• Scalable BI
• 360 View of xyz
  • Eliminate siloed data
• Anomaly detection
• PB-scale platform for cyber security
• Behavioral analytics
• Multi-tenant / shared resources
Data Management

Yesterday and Today
80% of data available to most companies is not used to make data driven decisions.

Existing solutions are not enough.
- Only capture structured data.
- Challenging to adapt to unstructured or semi-structured sources and diverse types.
- Optimized for clean, organized data, not uncertain data quality and value.
Discovery & Data Management Lifecycle
Too little time finding insights that change the business

Data Select
Data Prepare
Data Transform

f(x)

? Time

a = A

Knowledge

Traditional Approach
Model First vs. Data First

Insights

80% 20%
Inverting Data Access Cycles

What if we could make data preparation 20% of the effort so you can focus 80% of your time on executing and improving your business?
Adopt an Agile Approach
Successful projects start small, fail often, and iterate to success

1. **Get data** you already have, or create new data.

2. **Explore and analyze**, quickly.

3. **Deploy** your application.
   ...and repeat

**Add:**
new data sources, more users, more use cases, more complex analytics, go real-time
Discovery & Data Management Lifecycle
Move quickly to business changing insights

Data
Select Data
Prepare Data
Transform Data
See Patterns
Interpret & Evaluate
Knowledge

Cloudera + Open Source Frameworks
Data First vs Model First

a = A

\( f(x) \)

20%
80%

© Cloudera, Inc. All rights reserved.
Current Data Management Architectures

- Limited Data
- Structured Only
- Slow Performance
- Restricted Use
- Difficult Redundancy
- Sometimes SPOF
- Not Real-Time
- Many more.....

20% of Available Data Used Today
The Cloudera Solution

Fast, Easy and Secure
Cloudera Enterprise Data Hub
Unlimited data. Diverse access. One platform.

Yesterday

ETL  MODEL  REPORT

Today

OPERATIONS
- PROCESS, ANALYZE, SERVE
- UNIFIED SERVICES
- STORE
- INTEGRATE

DATA MANAGEMENT
Cloudera Enterprise Data Hub
Unlimited data. Diverse access. One platform.
Cloudera Enterprise Data Hub

How does this stuff work?

Distributed Computing

Shared Processing

Shared Data
Cloudera Enterprise Data Hub
Making Hadoop Fast, Easy, and Secure for the Modernized Architecture

Hadoop is a new kind of data platform.
- One place for unlimited data
- Unified data access

Cloudera makes it:
- Fast for business
- Easy to manage
- Secure without compromise
MapReduce: A great tool for its day

The original scalable, general, processing engine of Hadoop ecosystem
- Useful across diverse problem domains
- Fueled initial ecosystem explosion
Enter Apache Spark
Flexible, in-memory data processing for Hadoop

Easier Development
- Rich APIs for Scala, Java, and Python
- Interactive shell

Flexible, Extensible API
- APIs for different types of workloads:
  - Batch
  - Streaming
  - Machine Learning
  - Graph

Faster Processing (Batch & Streaming)
- In-Memory processing and caching
The Spark Ecosystem & Hadoop

- Spark Streaming
- MLlib
- SparkSQL
- GraphX
- Dataframes
- SparkR

- Spark
- Impala
- Search
- MR
- Others

RESOURCE MANAGEMENT
YARN

STORAGE
HDFS, HBase

Business Value

- Increase Revenue
- Reduce Costs
- Mitigate Risks

Cloudera Enterprise: Fast, Easy, Secure

Technology Use Cases

- Data Engineering
- Data Discovery & Analytics
- Real-Time Data Applications
Apache Hadoop

... This will change before we leave the room.
# Hadoop Isn’t Just Hadoop Anymore

<table>
<thead>
<tr>
<th>Core Hadoop (HDFS, MR)</th>
<th>HBase</th>
<th>ZooKeeper</th>
<th>Core Hadoop</th>
<th>Core Hadoop</th>
<th>Core Hadoop</th>
<th>Core Hadoop</th>
<th>Core Hadoop + YARN</th>
<th>Core Hadoop + YARN</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006-07</td>
<td>2008</td>
<td>2009</td>
<td>2010</td>
<td>2011</td>
<td>2012</td>
<td>Present</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Kudu**
- RecordService
- Spark
- SparkSQL
- Parquet
- Sentry
- SparkSQL
- Tez
- Impala
- Kafka
- Flume
- Bigtop
- Oozie
- MRUnit
- HCatalog
- Sqoop
- Whirr
- Avro
- Hive
- Mahout
- HBase
- ZooKeeper
- Core Hadoop
- Core Hadoop
- Core Hadoop
- Core Hadoop + YARN
- Core Hadoop + YARN

**Cloudera**

© Cloudera, Inc. All rights reserved.
Apache Hadoop

Scalable
Flexible
Open (future proof)
Cost-Effective