What’s Holding Up Your Hadoop?

Eddie Garcia - Security Architect, Office of the CTO
What’s Holding Up Your Hadoop?

“It’s too complicated.”
“I don’t have the right resources.”
“Security and compliance are never going to approve this.”
5 steps to a successful Hadoop project from POC to production

Solving business problems in a repeatable and measurable fashion
Step 1: Identify a **real** Hadoop business use case
Hey... no one's using Big Data/NoSQL stuff around here, right?
Usage of Hadoop – today

What are you currently using your Hadoop infrastructure for?

- ETL processing: 71.4%
- Data warehouse optimization: 57.1%
- Data archival: 42.9%
- Advanced analytics: 64.3%
- Log management: 14.3%
- Search application: 42.9%

*Results are based on internal Cloudera sampled survey of a CDH install base users
Usage of Hadoop – next 6-12 months

How do you anticipate using your Hadoop infrastructure for in the next 6-12 months?

- ETL processing: 64.3%
- Data warehouse optimization: 85.7%
- Data archival: 57.1%
- Advanced analytics: 85.7%
- Log management: 50.0%
- Search application: 50.0%

*Results are based on internal Cloudera sampled survey of a CDH install base users
Journey to Achieve Full Potential

Operational Efficiency
- Active Archive
- ETL Acceleration
- EDW Optimization

Information Advantage
- Agile Exploration
- Data Science
- Consolidation 360° View

IT Business
Step 2: Get ahead of the potential challenges
Top Challenges Hadoop Operations

1. Integration with other IT management tools
2. Diagnostics and troubleshooting
3. Configuration Management
Top Challenges
Organization Issues

1. Too entrenched in existing technology
2. Lack of skills - operational
3. Lack of understanding of Hadoop and its business benefits
Top Challenges Data Warehouse Optimization

1. Lack of SQL compliance to ANSI standards
2. Lack of supported data types
3. Lack of workload resource management
Top Challenges
Advanced Analytics

1. Lack of robust analytic tools integration
2. Lack of workload resource management
3. Lack of data analytics skilled resources
Top Challenges
Security and Compliance

1. Lack of unified security model for Hadoop
2. Native Hadoop audit and access management
3. Kerberos setup and management
Step 3: Secure your data assets
SECURITY ASSERTION

Protect ALL the data.
Data Breaches Happen

Home Depot confirmed on Monday evening that it is investigating a possible data breach in its retail stores, which could affect up to 40 million customers. The breach is believed to have occurred between April 21 and July 19, 2014. Home Depot said that it is working with law enforcement and other authorities to determine the extent of the breach and to prevent future incidents. The company has also hired a team of experts to review its security systems and procedures.

In other news, Adobe Systems Inc. announced that it has discovered a data breach that affected approximately 145 million customers. The company said that the breach occurred between April and November 2013, and that it was caused by a SQL injection attack on the company’s payment gateway. Adobe said that it has taken steps to improve its security and prevent future incidents.

Credit card breach at 33 P.F. Chang's locations

A breach at 33 P.F. Chang’s locations in the United States has been identified as more extensive than previously disclosed. The breach occurred between July 14, 2014, and July 16, 2014, and affected an estimated 11 million customers. The company said that it has worked with law enforcement and other authorities to determine the extent of the breach and to prevent future incidents. P.F. Chang’s has also hired a team of experts to review its security systems and procedures.

(Reuters) - Adobe Systems Inc said on Tuesday that the scope of a cyber-security breach disclosed nearly a month ago was far bigger than initially reported, with attackers obtaining data on more than 38 million customer accounts.
Regulations Dictate Data

A set of laws, guidelines, and specifications that dictate how a business must operate
### Common Pain Points of Hadoop Security

<table>
<thead>
<tr>
<th>Perimeter</th>
<th>Access</th>
<th>Visibility</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Manual setup of separate Kerberos cluster</td>
<td>• Redundant security policies for every access path</td>
<td>• No unified audit trail with manual point-in-time user access history</td>
<td>• Risk of data breach</td>
</tr>
<tr>
<td>• Manual, error-prone configuration</td>
<td>• Manual mirroring</td>
<td>• Cannot meet core data governance requirements with limited lineage</td>
<td>• Limited integration with existing systems</td>
</tr>
<tr>
<td>• Ongoing maintenance</td>
<td>• Threat of rogue insiders</td>
<td>• Cost of encryption and key management</td>
<td></td>
</tr>
</tbody>
</table>
Cloudera - Comprehensive Hadoop Security

**Perimeter**
Guarding access to the cluster itself
- Authentication
  - Cloudera Manager

**Access**
Defining what users and applications can do with data
- Authorization
  - Sentry

**Visibility**
Reporting on where data came from and how it’s being used
- Audit
  - Cloudera Navigator

**Data**
Protecting data in the cluster from unauthorized visibility
- Compliance
  - Navigator Encrypt and Key Trustee
WHY IT’S DIFFERENT?

Data scientists.
Real data = real insights.
Step 4: A Big Data Center of Excellence
Why Build a Center of Excellence?
Silos Obstruct Scale, Prevent Flexibility, and Drive Up Costs

- Novice Skill Level
- Redundant Infrastructure
- Incompatible Systems
- Disparate Data
Big Data COE Program Roles
Staff Centrally and Train to Scale

Management & Leadership
- Program Manager
- Executive Sponsor
- Big Data Visionary

Business & Data
- Lead Data Scientist
- Data Wranglers
- Lead Business Analyst
- LOB Rep
- LOB Rep
- LOB Rep

Technology & Ops
- Data Warehouse Specialist
- Developers
- Architects
- IT & InfoSec
Big Data as a Service Drives Success
Coordinate Workloads to Distribute Advantages Widely and Seamlessly

Insights & Best Practices
Hadoop Experts
Broadly Accessible Big Data
Technology & Equipment

©2014 Cloudera, Inc. All rights reserved.
Step 5: Plan time to production and blockers
Percentage of time spent to Production-Ready

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Install &amp; config</td>
<td>13.73</td>
</tr>
<tr>
<td>Training</td>
<td>13.42</td>
</tr>
<tr>
<td>User onboarding</td>
<td>9.82</td>
</tr>
<tr>
<td>Coding apps</td>
<td>18.27</td>
</tr>
<tr>
<td>Data migration</td>
<td>16.09</td>
</tr>
<tr>
<td>Workload migration</td>
<td>9.00</td>
</tr>
<tr>
<td>Integration</td>
<td>9.60</td>
</tr>
<tr>
<td>Staffing</td>
<td>9.67</td>
</tr>
<tr>
<td>Testing</td>
<td>11.00</td>
</tr>
<tr>
<td>Other</td>
<td>3.33</td>
</tr>
</tbody>
</table>

*Results are based on internal Cloudera sampled survey of a CDH install base users.*
Blockers to Production-Ready

*Results are based on internal Cloudera sampled survey of a CDH install base users

<table>
<thead>
<tr>
<th>Identify Use Case</th>
<th>Business Sponsor</th>
<th>Install &amp; config</th>
<th>Security Compliance</th>
<th>Data Mgmt</th>
<th>Integration with tools</th>
<th>SQL features</th>
<th>IT Skills</th>
<th>Data Science skills</th>
<th>Other</th>
</tr>
</thead>
</table>
5 Steps for a Production-Ready Hadoop

1. Identify a real Hadoop business use case
2. Get ahead of the potential challenges
3. Secure your data assets
4. A Big Data Center of Excellence
5. Plan time to production and blockers
Step 6: Ask Bigger Questions
Thank You!

Eddie Garcia

eddie@cloudera.com