The Future of Data Management

with Hadoop and the Enterprise Data Hub

Amr Awadallah (@awadallah) | Cofounder and CTO
<table>
<thead>
<tr>
<th>Cloudera Snapshot</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Founded</strong></td>
</tr>
<tr>
<td><strong>Employees Today</strong></td>
</tr>
</tbody>
</table>
| **World Class Support** | 24x7 Global Staff  
Pro-active & Predictive Support Programs |
| **Mission Critical** | Thousands of Enterprise Users  
Over 500+ Paying Subscription Customers |
| **The Largest Ecosystem** | Over 1200+ Partners |
| **Cloudera University** | Over 100,000+ Trained |
| **Open Source Leaders** | Cloudera Employees are Leading Developers & Contributors |
| **Total Capital Raised** | $1B+ (from Intel, Google, Dell, T. Rowe Price, Accel, Greylock) |
| **Mission** | Help Organizations Leverage the Power of **All Their Data** to **Ask Bigger Questions**. |
Why is Big Data Happening Now?

Instrumentation
Everything that can be measured will be measured.

Consumerization
Employees and customers expect more personal interactions, but not at the cost of their privacy.

Experimentation
The most innovative companies embrace experimentation and agility.
Big Data is Only Getting Bigger

1.8 trillion gigabytes of data was created in 2011*

- More than 90% is unstructured data
- Data volume doubles every year

* Source: IDC 2011

© Cloudera, Inc. All rights reserved.
And It Isn’t Just About Web 2.0 / Social

**AUTOMOTIVE**
Auto sensors reporting location, problems

**HIGH TECHNOLOGY / INDUSTRIAL MFG.**
Mfg quality
Warranty analysis

**COMMUNICATIONS**
Location-based advertising

**LIFE SCIENCES**
Clinical trials
Genomics

**OIL & GAS**
Drilling exploration sensor analysis

**CONSUMER PACKAGED GOODS**
Sentiment analysis of what’s hot, customer service

**RETAIL**
Consumer sentiment
Optimized marketing

**MEDIA / ENTERTAINMENT**
Viewers / advertising effectiveness

**TRAVEL & TRANSPORTATION**
Sensor analysis for optimal traffic flows
Customer sentiment

**FINANCIAL SERVICES**
Risk & portfolio analysis
New products

**UTILITIES**
Smart Meter analysis for network capacity

**EDUCATION & RESEARCH**
Experiment sensor analysis

**HEALTH CARE**
Patient sensors, monitoring, EHRs Quality of care

**ON-LINE SERVICES / SOCIAL MEDIA**
People & career matching
Website optimization

**HIGH TECHNOLOGY / INDUSTRIAL MFG.**

**LAW ENFORCEMENT & DEFENSE**
Threat analysis, Social media monitoring, Photo analysis

**OIL & GAS**
Drilling exploration sensor analysis

**OIL & GAS**
Drilling exploration sensor analysis

**OIL & GAS**
Drilling exploration sensor analysis

**OIL & GAS**
Drilling exploration sensor analysis

**OIL & GAS**
Drilling exploration sensor analysis

**OIL & GAS**
Drilling exploration sensor analysis

© Cloudera, Inc. All rights reserved.
Expanding Data Requires A New Approach

What we do
Copy Data to Applications

Process-centric businesses use:
- Structured data mainly
- Internal data only
- “Important” data only
- Multiple copies of data

Data
Information-centric businesses use all Data:
Multi-structured, Internal & external data of all types

What we should do
Bring Applications to Data
Hadoop Changes the Game: Storage & Compute Together

The Old Way

- Compute (RDBMS, EDW)
- Data Storage (SAN, NAS)
- Expensive, Special purpose, “Reliable” Servers
- Expensive Licensed Software
- Hard to scale
- Network is a bottleneck
- Only handles relational data
- Difficult to add new fields & data types
- Expensive & Unattainable
  - $30,000+ per TB

The Hadoop Way

- Compute (CPU)
- Memory
- Storage (Disk)
- Commodity “Unreliable” Servers
- Hybrid Open Source Software
- Scales out forever
- No bottlenecks
- Easy to ingest any data
- Agile data access
- Affordable & Attainable
  - $300-$1,000 per TB
The Old Way: Bringing Data to Applications

1. Can’t Retain Valuable Data
   - Leaving data behind
   - Risk and compliance
   - High cost of storage

2. Can’t Meet ETL SLAs
   - Up-front modeling
   - Transforms slow
   - Transforms lose data

3. Can’t Ask New Questions
   - Existing systems strained
   - No agility
   - “BI backlog”

4. Can’t Get a 360 View
   - Many special-purpose systems
   - Moving data around
   - No complete views
The New Way: Bringing Applications to Data

**Consolidated Architecture**
- Bring applications to data
- Combine different workloads on common data (i.e. SQL + Search)
- True analytic agility

**Agile Exploration**
- Simple search + BI tools
- “Schema on read” agility
- Reduce BI user backlog requests

**Scalable Transformations**
- One source of data for all analytics
- Persist state of transformed data
- Significantly faster & cheaper

**Active Archive**
- Full fidelity original data
- Indefinite time, any source
- Lowest cost storage
Core Benefits of the Enterprise Data Hub

- Full-Fidelity Active Archive
- Accelerate Time to Insight (Scale)
- Unlock Agility and Exploration
- Consolidate Silos for 360° View
- Enable Pervasive Analytics
Cloudera Enterprise powered by Apache Hadoop

A new kind of data platform
• One place for unlimited data
• Unified, multi-framework data access

Key Advantages:
• Leading performance
• Enterprise system and data management
• Fundamentally secure
• Open source, open standards
One Platform, Many Workloads

Batch, Interactive, and Real-Time.
Leading performance and usability in one platform.

- End-to-end analytic workflows
- Access more data
- Work with data in new ways
- Enable new users

Process
- Ingest: Sqoop, Flume, Kafka
- Transform: MapReduce, Hive, Pig, Spark

Discover
- Analytic Database: Impala
- Search: Solr

Model
- Machine Learning: SAS, R, Spark, Mahout, Oryx

Serve
- NoSQL Database: HBase
- Streaming: Spark Streaming

Security and Administration
- YARN, Cloudera Manager, Cloudera Navigator

Unlimited Storage
- HDFS, HBase
Complement Existing Investments and Skills

BI Integration

• Seamlessly integrate into data analytic vendors
• Push heavy workloads down to Cloudera using analytic SQL capabilities
The Modern Information Architecture

Data Architects
- META DATA / ETL TOOLS
- CLOUDERA MANAGER

System Operators
- DATA ARCHITECTS
- SYSTEM OPERATORS
- ENGINEERS
- DATA SCIENTISTS
- ANALYSTS
- BUSINESS USERS

Data Architects
- META DATA / ETL TOOLS

System Operators
- CLOUDERA MANAGER

Engineers
- CONVERGED APPLICATIONS

Data Scientists
- MACHINE LEARNING

Analysts
- BI / ANALYTICS

Business Users
- ENTERPRISE REPORTING

Enterprisve Data Warehouse

Online Serving System

Web/Mobile Applications

Customers & End Users

META DATA / ETL TOOLS

CLOUDERA MANAGER

CONVERGED APPLICATIONS

MACHINE LEARNING

BI / ANALYTICS

ENTERPRISE REPORTING

ENTERPRISE DATA WAREHOUSE

ONLINE SERVING SYSTEM

WEB/MOBILE APPLICATIONS

WEB LOGS

FILES

RDBMS

SYS LOGS
Hadoop Administration Made Easy

Cloudera Manager
Focus on the solution, not the cluster, with the only complete, zero-downtime administration tool for Apache Hadoop.

Unique Capabilities:
• Unified configuration, management and monitoring across all services
• Online installation and upgrades
• Direct connection to Cloudera Support
• 3rd Party Extensibility
Big Data Meets Data Governance

Cloudera Navigator
Minimize risk and maintain compliance with the only native end-to-end data governance solution for Apache Hadoop.

Unique Capabilities:
• Auditing
• Lineage
• Metadata Tagging and Discovery
• Lifecycle Management
A High Level View of the Journey

Operational Efficiency
(Faster, Bigger, Cheaper)

Cheap Storage → ETL Acceleration → EDW Optimization

Transformative Applications
(New Business Value)

Agile Exploration → Not Only SQL → Pervasive Analytics

IT → Business
Premier analyzes $41 billion in healthcare spend, driving recommendations that help providers get better products at lower costs.
Allstate Builds A Universal Data Archive

The Challenge:
• Data silos spread across company with 80+ years historical data; only some digitized
• Analysis on one state’s data takes 24 hours; can’t analyze all 50 states at once

Allstate optimizes offers and pricing with a comprehensive view of individual risk.

The Solution:
• Universal data archive on Cloudera Enterprise spans enterprise-wide systems
• 3 use cases: storage, ETL, applied math
• Analyze all 50 states in 16 hours using Hive; 500X speed-up
Why Cloudera?

✓ Enterprise-Grade Hadoop
  Differentiated performance, security, management, and governance.

✓ Expertise
  No one knows Hadoop better than Cloudera.

✓ Enablement
  Support, Training, and Professional Services enable and deliver success.

✓ Ecosystem
  Cloudera ensures that Hadoop works with the platforms, tools, and integrators you rely on.

✓ Sustainable Innovation
  Our hybrid open source model delivers the benefits of open source and what the enterprise requires, while enabling us to invest in the future for our customers.