Evolving a New Analytical Platform
What Works and What’s Missing

Jeff Hammerbacher
Chief Scientist, Cloudera
October 10, 2010
My Background
Thanks for Asking

- hammer@cloudera.com
- Studied Mathematics at Harvard
- Worked as a Quant on Wall Street
- Conceived, built, and led Data team at Facebook
  - Nearly 30 amazing engineers and data scientists
  - Several open source projects and research papers
- Founder of Cloudera
  - Chief Scientist
  - Also, check out the book “Beautiful Data”
Presentation Outline

1. Defining the Platform
   - BI: Science for Profit
   - Need tools for whole research cycle
   - SQL Server 2008 R2: defining the platform

2. State of the Platform Ecosystem

3. Foundations for a New Implementation
   - HDFS and MapReduce
   - Evolution of Hadoop

4. Future Developments

Questions and Discussion
1. Defining the Platform
BI is looking more like science (for profit)
Jim Gray: Science entering Fourth Paradigm

“We have to do better at producing tools to support the whole research cycle”
RDBMS only a small part of this tool set
Example: SQL Server 2008 R2
RDBMS: SQL Server
ETL: SQL Server Integration Services

RDBMS: SQL Server
ETL: SQL Server Integration Services

RDBMS: SQL Server

Reporting: SQL Server Reporting Services
ETL: SQL Server Integration Services

RDBMS: SQL Server

Reporting: SQL Server Reporting Services

Analysis: SQL Server Analysis Services
ETL: SQL Server Integration Services
RDBMS: SQL Server
Reporting: SQL Server Reporting Services
Analysis: SQL Server Analysis Services
Search: Full-Text Search
CEP: StreamInsight
ETL: SQL Server Integration Services
RDBMS: SQL Server
Reporting: SQL Server Reporting Services
Analysis: SQL Server Analysis Services
Search: Full-Text Search
CEP: StreamInsight
ETL: SQL Server Integration Services
RDBMS: SQL Server
Reporting: SQL Server Reporting Services
Analysis: SQL Server Analysis Services
Search: Full-Text Search
OLAP: PowerPivot
MDM: Master Data Services
CEP: StreamInsight
ETL: SQL Server Integration Services
RDBMS: SQL Server
Reporting: SQL Server Reporting Services
Analysis: SQL Server Analysis Services
Search: Full-Text Search
OLAP: PowerPivot

Wednesday, October 6, 2010
RDBMS: SQL Server
ETL: SQL Server Integration Services
Reporting: SQL Server Reporting Services
Analysis: SQL Server Analysis Services
Search: Full-Text Search
OLAP: PowerPivot

MDM: Master Data Services
CEP: StreamInsight
Collaboration: SharePoint
What do we call this unified suite?
For today: Analytical Data Platform
For today: Analytical Data Platform

LAMP Stack for Analytical Data Management
2. The State of the Platform Ecosystem
Who makes up the platform ecosystem?
Platform Providers
Infrastructure Providers

Platform Providers
Content Providers
Infrastructure Providers
Platform Providers
Application Developers
Content Providers
Infrastructure Providers
Platform Providers
Application Developers
End Users
What is new about the ecosystem today?
Content Providers

1. > 95% of enterprise data is unstructured
2. Data volumes growing rapidly
Infrastructure Providers

1. Cloud
2. Warehouse-Scale Computers
Platform Providers

1. Open source
2. Driven by consumer web properties
Application Developers

1. Data Scientists
2. Diversity of languages
End Users

1. Browser is the client
2. Tell a story about the business
3. Foundations for a New Implementation
New foundations: HDFS and MapReduce
2005: Doug/Mike start project inside Nutch
2006: Doug joins Yahoo!
2007: Make Hadoop scale
2007: Make Hadoop scale

Yahoo! makes Pig open source
Jim Gray’s “Fourth Paradigm” lecture

2007: Make Hadoop scale
Yahoo! makes Pig open source
Randy Bryant’s “DISC” lecture
Jim Gray’s “Fourth Paradigm” lecture

2007: Make Hadoop scale
Yahoo! makes Pig open source
Randy Bryant’s “DISC” lecture
Jim Gray’s “Fourth Paradigm” lecture

2007: Make Hadoop scale
Yahoo! makes Pig open source
Powerset makes HBase open source
2008: Make Hadoop fast
2008: Make Hadoop fast

Yahoo! wins Daytona terabyte sort benchmark
First Hadoop Summit

2008: Make Hadoop fast

Yahoo! wins Daytona terabyte sort benchmark
First Hadoop Summit

2008: Make Hadoop fast

Yahoo! wins Daytona terabyte sort benchmark
Yahoo! builds production webmap with Hadoop
Facebook makes Hive open source
First Hadoop Summit

2008: Make Hadoop fast
Yahoo! wins Daytona terabyte sort benchmark
Yahoo! builds production webmap with Hadoop
2008: Make Hadoop fast

Yahoo! wins Daytona terabyte sort benchmark

Yahoo! builds production webmap with Hadoop
2009: Insert Hadoop into the enterprise
2009: Insert Hadoop into the enterprise

Cloudera releases CDH
2009: Insert Hadoop into the enterprise
Cloudera releases CDH
2009: Insert Hadoop into the enterprise

Yahoo! sorts a petabyte with Hadoop
First Hadoop World NYC

Cloudera releases CDH
Yahoo! sorts a petabyte with Hadoop
First Hadoop World NYC

2009: Insert Hadoop into the enterprise
Cloudera releases CDH
Cloudera adds training, support, services
“The Unreasonable Effectiveness of Data”
Yahoo! sorts a petabyte with Hadoop
First Hadoop World NYC

2009: Insert Hadoop into the enterprise
Cloudera releases CDH
Cloudera adds training, support, services

Wednesday, October 6, 2010
2010: Integrate Hadoop into the enterprise
2010: Integrate Hadoop into the enterprise

IBM announces InfoSphere BigInsights
2010: Integrate Hadoop into the enterprise

Yahoo! completes enterprise-class security

IBM announces InfoSphere BigInsights
Yahoo! completes enterprise-class security

2010: Integrate Hadoop into the enterprise
IBM announces InfoSphere BigInsights
Datameer and Karmasphere funded
2010: Integrate Hadoop into the enterprise

IBM announces InfoSphere BigInsights

Datameer and Karmasphere funded

Quest, Talend, Netezza, and more integrate

Yahoo! completes enterprise-class security
2010: Integrate Hadoop into the enterprise

IBM announces InfoSphere BigInsights
Datameer and Karmasphere funded

Cloudera releases Cloudera Enterprise
Quest, Talend, Netezza, and more integrate
Yahoo! completes enterprise-class security
Hadoop will be an Analytical Data Platform
Cloudera’s Distribution for Hadoop

UI Framework
- Workflow
  - Oozie
- Data Integration
  - Flume, Sqoop

Hue
- Scheduling
  - Oozie

SDK
- Languages, Compilers
  - Pig/Hive
- Metadata
  - Hive
- Fast read/write access
  - HBase

Coordination
- Zookeeper
4. Future Developments
Capture: Web and Intranet Documents
Curate: Unified Metadata
Curate: Workflow and Scheduling
Curate: Indexes and Materialized Views
Curate: Learn Structure from Data
Analyze: Mesos-enabled frameworks
Analyze: Link working set and historical data
Analyze: Iterative in-memory analysis
Analyze: Low-latency queries on Avro data
All behind a single user interface
Hue
Making Many Computers Feel Like One