Datenverwaltung im Wandel

- Building an Enterprise Data Hub with Cloudera

Bernard Doering
Regional Director, Central EMEA, Cloudera
Cloudera – Your Hadoop Experts

Founded
2008, by former employees

Employees
>800

Global 24x7 Support
Follow-the-sun capability; Pro-active & Predictive Support Programs
Dedicated Support Engineers; Support Centers in NA, Europe & Asia

Professional Services
World class services delivery teams worldwide

Mission Critical
Thousands of enterprise customers rely on Cloudera
50% of the Fortune 50; 65% of the Fortune 500
Top Defense & Intelligence Agencies

The Largest Ecosystem
Over 1,300 Members of our Partner Program, ClouderaConnect

Cloudera University
Over 40,000 people trained around the world

Open Source Leaders
Cloudera employees are founders of most of the Apache Hadoop ecosystem projects, and leading contributors to all of them
Data drives innovation – Internet of Things

40 Zettabytes of data will be generated WW in 2020\(^1\)

2.8 Zettabytes of data generated WW in 2012\(^1\)

SMART CLIENTS
- Richer user experiences

INTELLIGENT CLOUD
- Richer data to analyze

INTELLIGENT THINGS
- Richer data from devices

Sources: (1) IDC Digital Universe 2020, (2) IDC
Big Data is All Data and All Paradigms

**Transactional & Application Data**
- Volume
- Structured
- Throughput

**Machine Data**
- Velocity
- Semi-structured
- Ingestion

**Social Data**
- Variety
- Highly unstructured
- Veracity

**Enterprise Content**
- Variety
- Highly unstructured
- Volume
Big Data has outgrown the Database Systems

1.8 trillion gigabytes of data was created in 2011*

- More than 80% is unstructured data
- Quantity doubles every 2 years

* Source: IDC 2011
Hadoop is the Platform for Big Data

- Open Source Software
- Store, Process, Analyze Data
- Large Data Sets, stored as raw files
- Structured, unstructured, semi-structured...all data types
Internet of Things is driving Big Data

- Market leader in silicon
- Long & successful history of investment and collaboration with software platforms: Linux, VMware
- Global reach; market leading Hadoop distribution in China

- Market leader in Hadoop
- Largest base of paid customers & free users
- Consistently delivering industry-leading capabilities around Apache Hadoop
- Laser focus on delivering Enterprise-grade Hadoop capabilities and innovation of Apache Hadoop
Expanding Big Data Requires a New Approach

1980s
Bring Data to Compute

Process-centric businesses use:
• Structured data mainly
• Internal data only
• “Important” data only

Data
Relative size & complexity

Now
Bring Compute to Data

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

Compute
Data
Compute
Data
Compute
Data
Compute
Data
Compute
Data
Compute
Two Categories of Hadoop Use Cases

Innovation and Advantage
Ask Bigger Questions: Gain value from *all* your data

- Data Processing: ETL and/or ELT offload
- Data Storage: Active Archive ("data lake")

Operational Efficiency
Perform existing workloads faster, cheaper, better

Most Companies Start Here

Operational
Efficiency

Business Intelligence
Advanced Analytics
Applications
The Old Way: Moving Data to Compute
Huge Investment in Specialized Systems that Treat Data as a Commodity

**Major Challenges**

**Missing Data**
- Leaving data behind
- Risk and compliance
- High cost of storage

**Time to Data**
- Up-front modeling
- Transforms slow
- Transforms lose data

**Cost of Analytics**
- Existing systems strained
- No agility
- “BI backlog”

**Complex Architecture**
- Many special-purpose systems
- Moving data around
- No complete views
The Old Way: Siloed Business Functions
Lack of Coordination Increases Opportunity Costs and Decreases Data Availability

Major Challenges

- Poor Visibility
- Inefficiency
- Extreme Cost
- Complexity
Enterprise Data Hub adoption is a progression

**Multi-workload analytic platform**
- Bring applications to data
- Combine different workloads on common data (i.e. SQL + Search)
- **True BI agility**

**Self-service exploratory BI**
- Simple search + BI tools
- “Schema on read” agility
- **Reduce BI user backlog requests**

**Data management, transformations**
- One source of data for all analytics
- Persisted state of transformed data
- **Significantly faster & cheaper**

**Active archive**
- Full fidelity original data
- Indefinite time, any source
- **Lowest cost storage**

ERP, CRM, RDBMS, Machines, Images, Video, Logs, Clickstreams, External Data Sources
The New Way: Bringing Compute to Data
Maximize Benefit from All Your Data for Mission-Critical Jobs and Innovation

Major Benefits

Active Compliance Archive
- Full fidelity original data
- Indefinite time, any source
- Lowest cost storage

Persistent Storage
- One source of data for all analytics
- Persist state of transformed data
- Significantly faster & cheaper

Self-Service Exploratory BI
- Simple search + BI tools
- “Schema on read” agility
- Reduce BI user backlog requests

Diverse Analytic Platform
- Bring applications to data
- Combine different workloads on common data (i.e. SQL + Search)
- True analytic agility
The New Way: Bring Business Functions to Data
Consolidate Relevant Services and Data in Multi-tenant Environment

**Major Benefits**

- Compliant
- Centralized
- Self-Service
- Multiple Workloads
The Modern Information Architecture

Data Architects
- META DATA / ETL TOOLS
- CLOUDERA MANAGER

System Operators
- WEB LOGS
- FILES
- RDBMS

Engineers
- CONVERGED APPLICATIONS
- MACHINE LEARNING

Data Scientists
- BI / ANALYTICS

Analysts
- ENTERPRISE REPORTING

Business Users
- ENTERPRISE DATA WAREHOUSE
- ONLINE SERVING SYSTEM
- WEB/MOBILE APPLICATIONS

Customers & End Users
Enterprise Data Hub: Key Attributes

1. Secure & Compliant
   - Robust access controls
   - Data encryption options
   - Shared security policies

2. Enterprise Data Governance
   - Meta data management
   - Data lineage/tethering
   - Audit histories

3. Unified & manageable
   - Common storage & resource management
   - On-premise & cloud
   - Highly available (including DR)

4. Open Architecture
   - Open source platform
   - APIs & engines for multiple workloads
   - Extensible for 3rd parties
Customer Example
Insurer: Real-time Enterprise Data Hub

Comprehensive Risk Analysis

• Structured and unstructured data includes customer data, economic trends, telematic sensors, weather, public data
• Integrated with mainframes and EDWs
• Before Hadoop, could analyze only one state, took 24 hours
• With Cloudera, can analyze risk across all 50 states, in 16 hours (500x improvement)
• First 3 use cases: Data hub, ETL offload, advanced analytics
Customer Example
Mobile Telecom Provider: ETL Engine

From 1% sampling to 100% analysis

- Exponential growth in data, generated by new consumer devices
- ETL and storage constraints limited analytics to 1% sample
- Now combined data warehouse and Cloudera Hadoop delivers analytics on 100% of data (half a PB per day!)
- Query times reduced dramatically (i.e. from 4 days to 53 minutes)
- 90% reduction of ETL code base
MasterCard

Cloudera: The first PCI-Certified Hadoop Platform

Challenge: All applications, databases, or file systems that have the potential to handle personal account-related data must undergo full PCI certification.

Solution: MasterCard’s Cloudera environment fully conforms to the PCI-DSS V 2.0 security standards so it can host PCI datasets and potentially integrate with other internal systems.

“Data privacy and protection is a top priority for MasterCard. As we maximize the most advanced technologies from partners and vendors, they must meet the rigorous security standards we’ve set. With Cloudera’s commitment to the same standards, we now have additional options in how we manage our data center.”

Gary VonderHaar
Chief Technology Officer, Architecture
MasterCard
The Most Complete Partner Ecosystem

Applications

Operational Tools

Data Systems

Infrastructure

More than 1,300 partners ensure compatibility with existing investments, lower skill barriers, and help maximize value from your data.
Why Cloudera?

We deliver long-term production success with enterprise Hadoop.

- **Open Source Innovation**
  No one knows Hadoop better than Cloudera.
  Cloudera leads development of enterprise Hadoop and offers the best support, training, and services.

- **Powerful Enterprise Tools**
  Cloudera extends open source Hadoop with capabilities required by the largest enterprises.

- **Ecosystem**
  Cloudera partners with industry leaders to ensure Hadoop works with the platforms, tools, and integrators our customers rely on.
  
  - **Enterprise Security**
    Meet compliance requirements and reduce risk exposure from storing sensitive data.
  
  - **Data Governance**
    Enable compliance and maximize analyst productivity.
  
  - **Complete Management**
    Deliver optimum system utilization and meet SLA commitments, on-premises or in the cloud, with minimum effort.
Thank You

Bernard Doering, Cloudera EMEA
bdoering@cloudera.com
Tel. + 49 172 692 9837
From Hadoop to an Enterprise Data Hub

CLOUDERA’S ENTERPRISE DATA HUB

- BATCH PROCESSING
  - MAPREDUCE
- ANALYTIC SQL
  - IMPALA
- SEARCH ENGINE
  - SOLR
- MACHINE LEARNING
  - SPARK
- STREAM PROCESSING
  - SPARK STREAMING
- 3RD PARTY APPS

WORKLOAD MANAGEMENT
- YARN

STORAGE FOR ANY TYPE OF DATA
- UNIFIED, ELASTIC, RESILIENT, SECURE

- FILESYSTEM
- ONLINE NOSQL

©2014 Cloudera, Inc. All rights reserved.