cloudera

Ask Bigger Questions

Becoming Information-Driven

Introduction to the Enterprise Data Hub

NEWS – all the Hadoop you can handle

- Strata + Hadoop World
- Partnerships
- Workload News
- Other New topics?
- What is Hadoop? Where does Cloudera Fit?
- Actual Demo – Real Hadoop environment. Real Data.
- Hear from someone who just deployed Hadoop

Strata Hadoop World

- Wednesday October 15th – 17th
- Javits Center in Manhattan NYC
- 5,000 folks in attendance – 7,000 tickets sold
- Extended partnership – now Global
  - All Strata events are now Strata + Hadoop World
  - Access to global venues ~8 additional

Cloudera and RedHat

- **Cloudera** and **Red Hat** have partnered on joint enterprise software solutions for big data with a key integration focused on Sahara, the Hadoop- or Spark-on-OpenStack component in Juno, the latest release of the open source cloud software suite.
- In another Hadoop-OpenStack marriage, the partners plan to provide cloud-ready big data platforms that combine both companies’ development tools with Apache Hadoop at the core. They will integrate Red Hat’s OpenStack distribution, including Sahara, with Cloudera Director and Cloudera Enterprise, all managed by Red Hat CloudForms.

Teradata

- **SAN DIEGO and PALO ALTO, Calif. – October 9, 2014 –** Teradata Corp. (NYSE: TDC), the big data analytics and marketing applications leader, and Cloudera, the leader in enterprise analytic data management powered by Apache Hadoop™, today announced an expanded partnership that covers broad technology integration and development road map alignment, and a unified go-to-market, sales and support offering. The two companies are optimizing the integration between Teradata's integrated data warehouse and Cloudera's enterprise data hub offering to facilitate access to multiple data sources through the Teradata Unified Data Architecture™.
- "The hallmark of the Unified Data Architecture is processing any analytic on any data, and this capability requires multiple technologies to be integrated. Adding Cloudera to the UDA offers our customers additional choices. The Unified Data Architecture combines the enterprise data hub and the integrated data warehouse into one large analytic ecosystem. Our respective product roadmaps are now better aligned to make this a reality," said Scott Gnau, president of Teradata Labs.

Ozone

- “a new initiative called Ozone, an object store, which **extends HDFS beyond a file system, toward a more complete enterprise storage layer**. Below we first cover a few common object store use cases, and then describe our approach…”
- Object stores are popular in cloud environments to provide a persistent storage to services running on virtual machines with ephemeral local storage. **Amazon’s S3 and Azure Storage are examples of popular object stores** in public clouds
- **Ozone stores objects identified by keys**. The keys and objects are organized into independent collections called buckets. Each bucket has a user provided name. The keys are unique within a bucket

Updates on Impala 2.0

- Analytic Window Functions
  - Analytic functions are frequently used in fields such as finance and science to provide trend, outlier, and bucketed analysis for large data sets
- Impala 2.0 now supports the following analytic query clauses and pure analytic functions:
- OVER Clause
- Window Clause
- DENSE_RANK() Function
- FIRST_VALUE() Function
- LAG() Function
- LAST_VALUE() Function
- LEAD() Function
- RANK() Function
- ROW_NUMBER() Function

Updates on Impala 2.0

- New Data Types
  - New data types in Impala 2.0 provide greater compatibility with source code from traditional database systems:
  - VARCHAR is like the STRING data type, but with a maximum length. See VARCHAR Data Type for details.
  - CHAR is like the STRING data type, but with a precise length. Short values are padded with spaces on the right. SeeCHAR Data Type for details.

Updates on Impala 2.0

- Subquery Types
  - Impala 2.0 also supports a number of subquery enhancements including:
  - Subqueries in the WHERE clause (for example, with the IN operator).
  - EXISTS and NOT EXISTS operators (always used in conjunction with subqueries).
  - The IN and NOT IN queries on the result set form a subquery, not just a hardcoded list of values.
  - Uncorrelated subqueries let you compare against one or more values for equality, IN, and EXISTS comparisons. For example, you might use WHERE clauses such as WHERE column = (SELECT MAX(some_other_column FROM table) or WHERE column IN (SELECT some_other_column FROM table WHERE conditions).
  - Correlated subqueries let you cross-reference values from the outer query block and the subquery.
  - Scalar subqueries let you substitute the result of single-value aggregate functions such as MAX(), MIN(), COUNT(), or AVG(), where you would normally use a numeric value in a WHERE clause.

Updates on Impala 2.0

- Disk Spillage
  - Certain memory-intensive operations now write temporary data to disk (known as “spilling to disk”) when Impala is close to exceeding its memory limit for a particular node.
  - For example, when large tables are joined, Impala keeps the distinct values of the join columns from one table in memory, to compare them to incoming values from the other table. When a query uses a GROUP BY clause for columns with millions or billions of distinct values, Impala keeps a similar number of temporary results in memory, to accumulate the aggregate results for each value in the group. When a large result set is sorted by the ORDER BYclause, each node sorts its portion of the result set in memory. The DISTINCT and UNION operators also build in-memory data structures to represent all values found so far, to eliminate duplicates as the query progresses.

CDH 5.2 Released

- Cloudera Enterprise 5.2 (comprising CDH 5.2, Cloudera Manager 5.2, Cloudera Director 1.0, and Cloudera Navigator 2.1

New in CDH 5.2

- Sentry Update
  - Apache Sentry (incubating) provides centralized authorization for services and applications in the Apache Hadoop ecosystem, allowing administrators to set up granular, role-based protection on resources, and to review them in one place.
  - You can use Hive or Impala to grant privileges using the GRANT ... WITH GRANT OPTION SQL statement
- GRANT
- priv_type [, priv_type] ...
- ON table_or_view_name
- TO principal_specification [, principal_specification] ...
- [WITH GRANT OPTION]

Security

- Via Apache **Sentry** (incubating) 1.4, GRANT and REVOKE statements in Impala and Hive can now include WITH GRANT OPTION, for delegation of granting and revoking privileges (joint work with Intel under Project Rhino). (**Learn more**.)
- **Hue** has a new Sentry UI that supports policy management for visually creating/editing roles in Sentry and permissions on files in HDFS.
- **Kerberos** authentication is now supported in Apache Accumulo.
- **Impala**, authentication can now be done through a combination of Kerberos and LDAP.

Data Management and Governance

- Cloudera **Navigator 2.1** features a brand-new auditing UI that is unified with lineage and discovery, so you now have access to all Navigator functionality from a single interface.
- **Navigator 2.1** includes role-based access control so you can restrict access to auditing, metadata and policy management capabilities.
- We're also shipping a beta policy engine in **Navigator 2.1**. Targeted to GA by year-end, the policy engine allows you to set up rules and notifications so you can classify data as it arrives and integrate with data preparation and profiling tools. Try it out and let us know what you think!
- And we've added lots of top-requested enhancements, such as **Sentry auditing for Impala and integration with Hue**

Cloud Deployment

- Cloudera Director is a simple and reliable way to deploy, scale, and manage Hadoop in the cloud (initially for AWS) in an enterprise-grade fashion. It’s free to download and use, and supported by default for Cloudera Enterprise customers. See the **User Guide** for more details.

Real time Architecture

- Re-base on Apache HBase 0.98.6
- Re-base on Apache Spark/Streaming 1.1
- Re-base on Impala 2.0
- Apache Sqoop now supports import into Apache Parquet (incubating) file format
- Apache Kafka integration with CDH is now incubating in Cloudera Labs; a Kafka-Cloudera Labs parcel (unsupported) is available for installation. Integration with Flume via special Source and Sink have also been provided.

First Hadoop Customer with PCI Compliance

- **PALO ALTO, Calif. October 22, 2014** – Cloudera, the leader in enterprise analytic data management powered by Apache Hadoop™, today announced that its Cloudera Enterprise platform is fully certified as compliant with Payment Card Industry (PCI) Data Security Standards, with its first instance of this newly certified platform being used with **MasterCard**.
- **MasterCard** requires that any technology handling its applications or payment card data files must have full PCI certification. Receiving this important certification allows MasterCard the opportunity to integrate Hadoop datasets with other environments that are already PCI-certified.

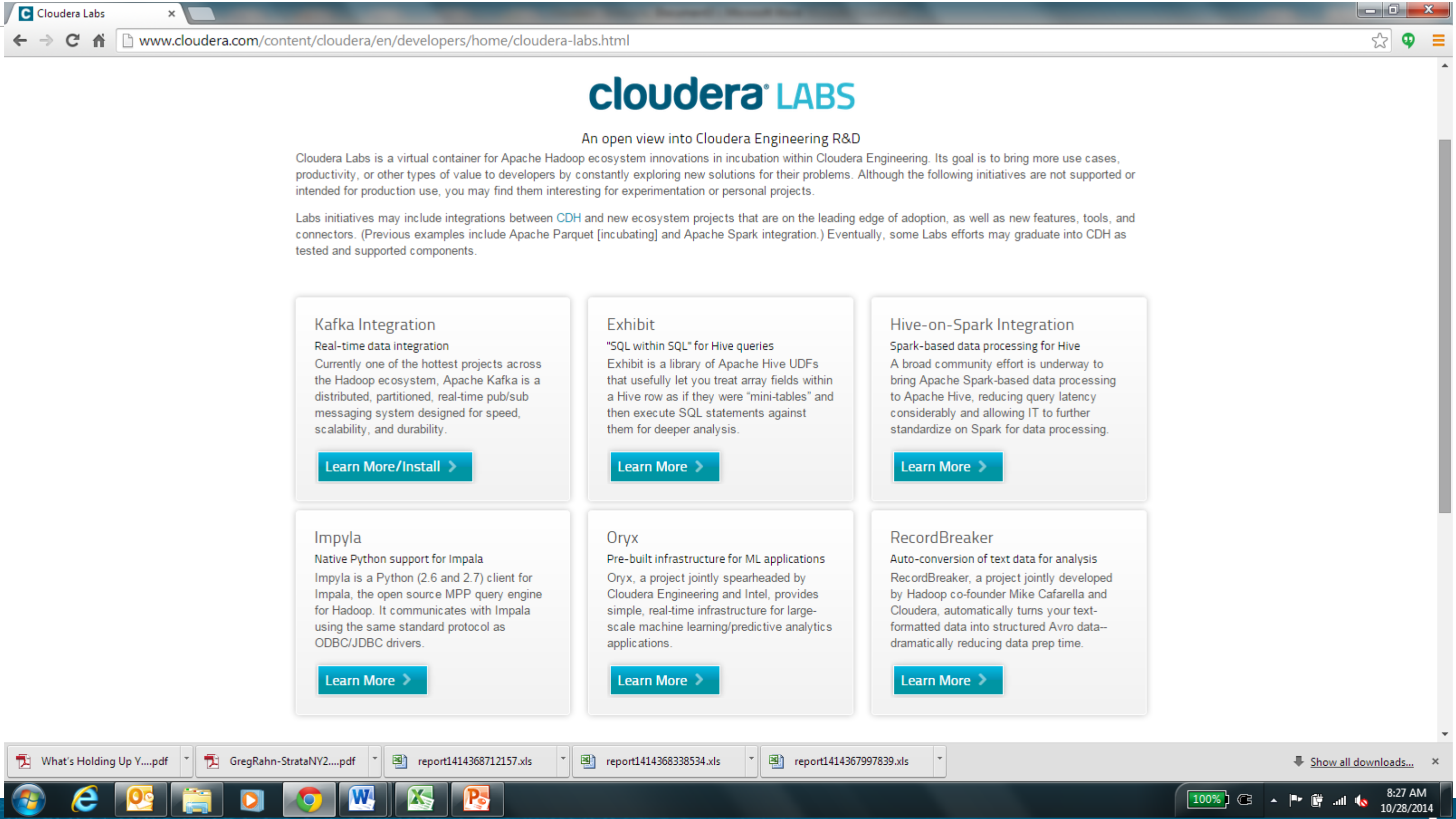
Cloudera and EMC Partnerships

- **SEATTLE, Washington and PALO ALTO, Calif., October 13, 2014** – Cloudera, the leader in enterprise analytic data management powered byApache Hadoop™, together with EMC Corporation (NYSE: EMC), today announced joint delivery of a new data management platform designed to unlock silos of enterprise application data for business analytics. Combining the latest Cloudera Enterprise with EMC® Isilon® scale-out storage, the solution enables customers to use, access and analyze data in an agile environment.
- "The economics of Hadoop for capturing data from anywhere, in any format, for future analysis dramatically changed the conversation around data storage," said Sam Grocott, Senior Vice President of Marketing, Emerging Technologies Division, EMC. "**Scale-out storage is a critical strategy for managing Big Data, and this integration with Cloudera ups the ante further by harnessing the analytic and processing power of Hadoop**."

DataPad - Acquihire

- **PALO ALTO, Calif., September 30, 2014** – Cloudera, the leader in enterprise analytic data management powered by Apache Hadoop™, today announced that it has acquired the technology assets of DataPad, an innovator in the exploration and analysis of big data sets. The acquisition will further strengthen Cloudera's enterprise data hub offering by simplifying data processing and analysis on Big Data. DataPad's Python-based framework will accelerate adoption of Cloudera's leading Big Data management and analytics platform and the team's expertise further expands the breadth of open source committers and contributions from Cloudera. Terms of the deal were not disclosed.
- DataPad co-founders, Wes McKinney and Chang She, well-known open source contributors and system architects, and the DataPad team, will join Cloudera. **McKinney is the creator of the open source project Pandas, the Python open source library and is also the author of the best-selling *Python for Data Analysis***. She, a long-time colleague of McKinney, is also a core developer of Pandas. They will lead the company's efforts to **build high-performance data backends** for business intelligence and analytic use cases, simplifying use of Cloudera's products.

Announcement of Cloudera Labs



cloudera LABS

An open view into Cloudera Engineering R&D

Cloudera Labs is a virtual container for Apache Hadoop ecosystem innovations in incubation within Cloudera Engineering. Its goal is to bring more use cases, productivity, or other types of value to developers by constantly exploring new solutions for their problems. Although the following initiatives are not supported or intended for production use, you may find them interesting for experimentation or personal projects.

Labs initiatives may include integrations between CDH and new ecosystem projects that are on the leading edge of adoption, as well as new features, tools, and connectors. (Previous examples include Apache Parquet [incubating] and Apache Spark integration.) Eventually, some Labs efforts may graduate into CDH as tested and supported components.

Kafka Integration

Real-time data integration

Currently one of the hottest projects across the Hadoop ecosystem, Apache Kafka is a distributed, partitioned, real-time pub/sub messaging system designed for speed, scalability, and durability.

Learn More/Install ›

Exhibit

"SQL within SQL" for Hive queries

Exhibit is a library of Apache Hive UDFs that usefully let you treat array fields within a Hive row as if they were "mini-tables" and then execute SQL statements against them for deeper analysis.

Learn More ›

Hive-on-Spark Integration

Spark-based data processing for Hive

A broad community effort is underway to bring Apache Spark-based data processing to Apache Hive, reducing query latency considerably and allowing IT to further standardize on Spark for data processing.

Learn More ›

Impyla

Native Python support for Impala

Impyla is a Python (2.6 and 2.7) client for Impala, the open source MPP query engine for Hadoop. It communicates with Impala using the same standard protocol as ODBC/JDBC drivers.

Learn More ›

Oryx

Pre-built infrastructure for ML applications

Oryx, a project jointly spearheaded by Cloudera Engineering and Intel, provides simple, real-time infrastructure for large-scale machine learning/predictive analytics applications.

Learn More ›

RecordBreaker

Auto-conversion of text data for analysis

RecordBreaker, a project jointly developed by Hadoop co-founder Mike Cafarella and Cloudera, automatically turns your text-formatted data into structured Avro data--dramatically reducing data prep time.

Learn More ›

Apache Drill / Spark

- MapR – Working on integrating Drill and Spark
- “Integrating Apache Drill and Spark simplifies the development of data pipelines and opens up Drill SQL-based ad-hoc queries on in-memory data,” said M.C. Srivas, CTO and cofounder, MapR Technologies. “Joining forces with Databricks to leverage our combined breadth and depth of technical resources to accelerate innovation is a huge win for customers.”
- Apache Drill provides the flexibility to immediately query complex data in native formats, such as schema-less data, nested data, and data with rapidly-evolving schemas, with minimal IT involvement. Because SQL queries can run directly on various file formats, live data can be explored as it is coming in, versus spending weeks preparing and managing schemas and setting up ETL tasks

Spark – Replacing Map Reduce?

Cloudera
The Leader in Data Management Powered by Apache Hadoop™

The Leading Open Source Distribution of Apache Hadoop

Powerful Suite of System & Data Management Software

Built for the Enterprise

Founded: **2008**

Employees: **700+**

Customers: Over **50% of the Fortune 50** and **65% of the Fortune 500** plus top US intelligence and defense agencies

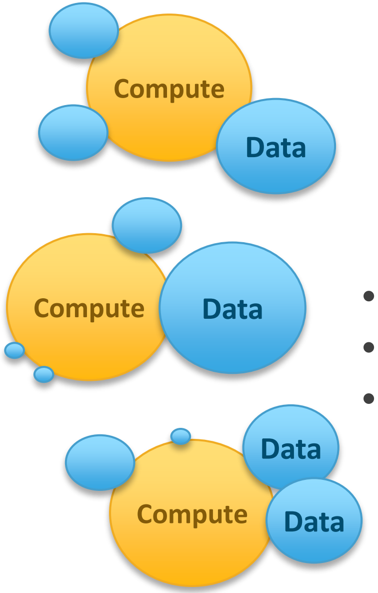
Partners: **700+** in hardware, software, and services

Education: **20,000+** trained annually; developers, admins, analysts, data scientists

Expanding Data Requires A New Approach

1980s

Bring Data to Compute



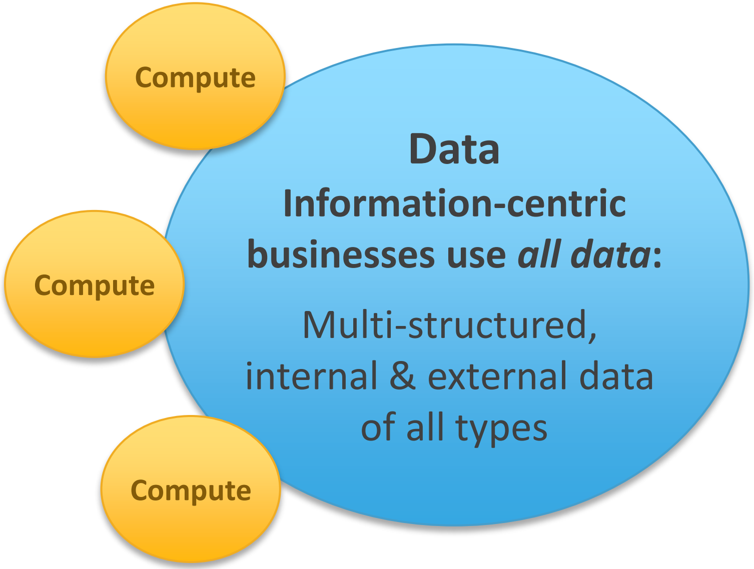
Process-centric businesses use:

- Structured data mainly
- Internal data only
- “Important” data only

Relative size & complexity

Now

Bring Compute to Data



Data

Information-centric businesses use *all data*:

Multi-structured, internal & external data of all types

The Old Way: Bringing Data to Compute

4 Complex Architecture

- Many special-purpose systems
- Moving data around
- No complete views

3 Cost of Analytics

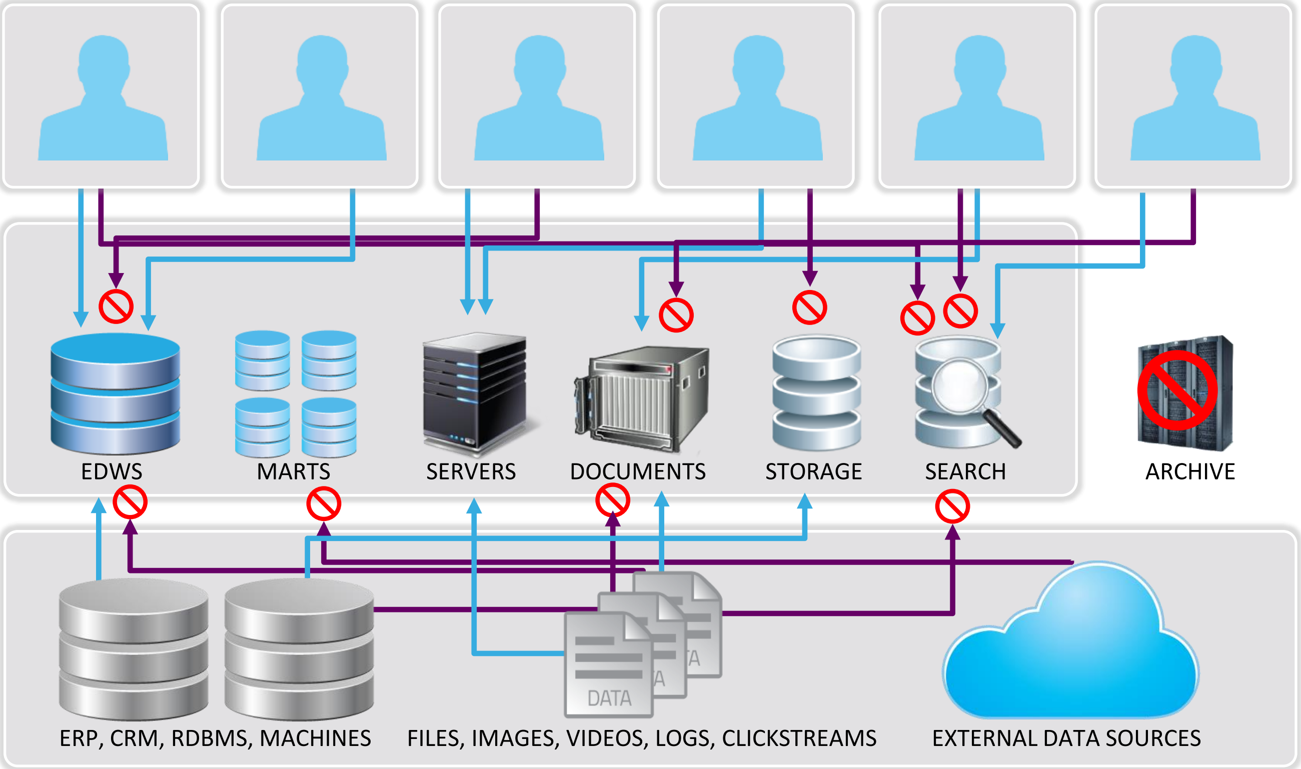
- Existing systems strained
- No agility
- “BI backlog”

2 Time to Data

- Up-front modeling
- Transforms slow
- Transforms lose data

1 Missing Data

- Leaving data behind
- Risk and compliance
- High cost of storage



The New Way: Bringing Compute to Data

4 **Diverse Analytic Platform**

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

3 **Self-Service Exploratory BI**

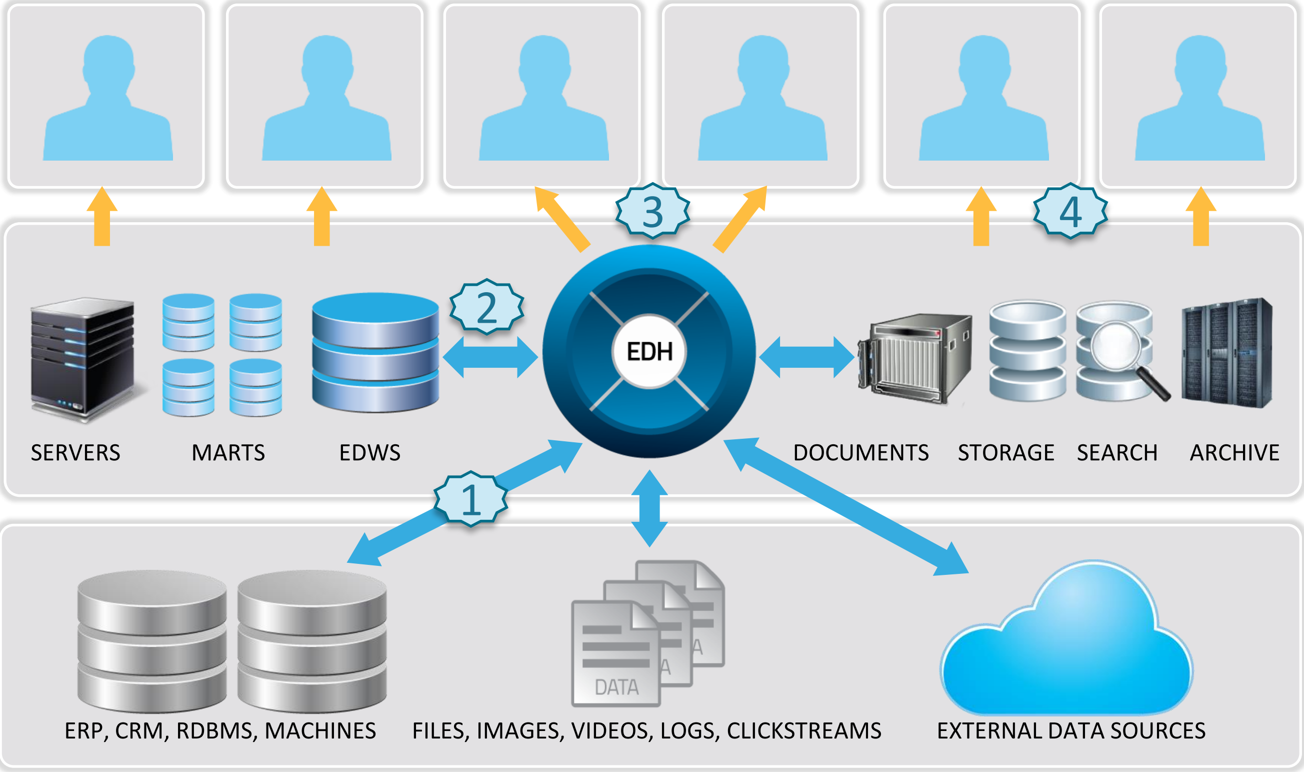
- Simple search + BI tools
- "Schema on read" agility
- ***Reduce BI user backlog requests***

2 **Persistent Staging**

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

1 **Active Compliance Archive**

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



Cloudera’s Enterprise Data Hub

Integration with Over 200 ISVs

- Self-Service BI
- Data Exploration
- Visualization

MicroStrategy tableau QlikView splunk>

Powerful Security Solution

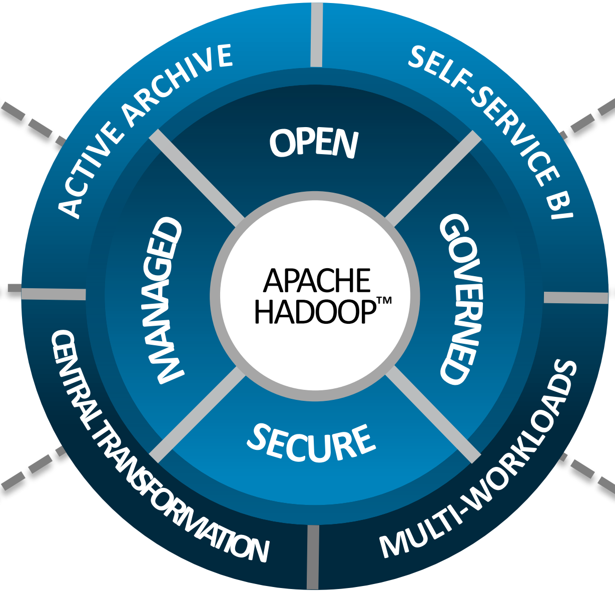
- Risk Analysis
- Fraud Prevention
- Compliance

Gazzang Voltage security DATAGUISE Vormetric ArcSight

Advanced Analytics Engine

- 360° Customer View
- Recommendation Engines
- Processing & Analytics

sas REVOLUTION ANALYTICS IBM COGNOS Paxata SAP Zoomdata



Flexible Deployment Options

- On-Premise or Cloud
- Appliances
- Engineered Systems

ORACLE IBM hp CISCO DELL amazon webservices verizon business

Infinite Analytic Storage

- Multi-Structured Data
- In-place Analytics
- Active Archive

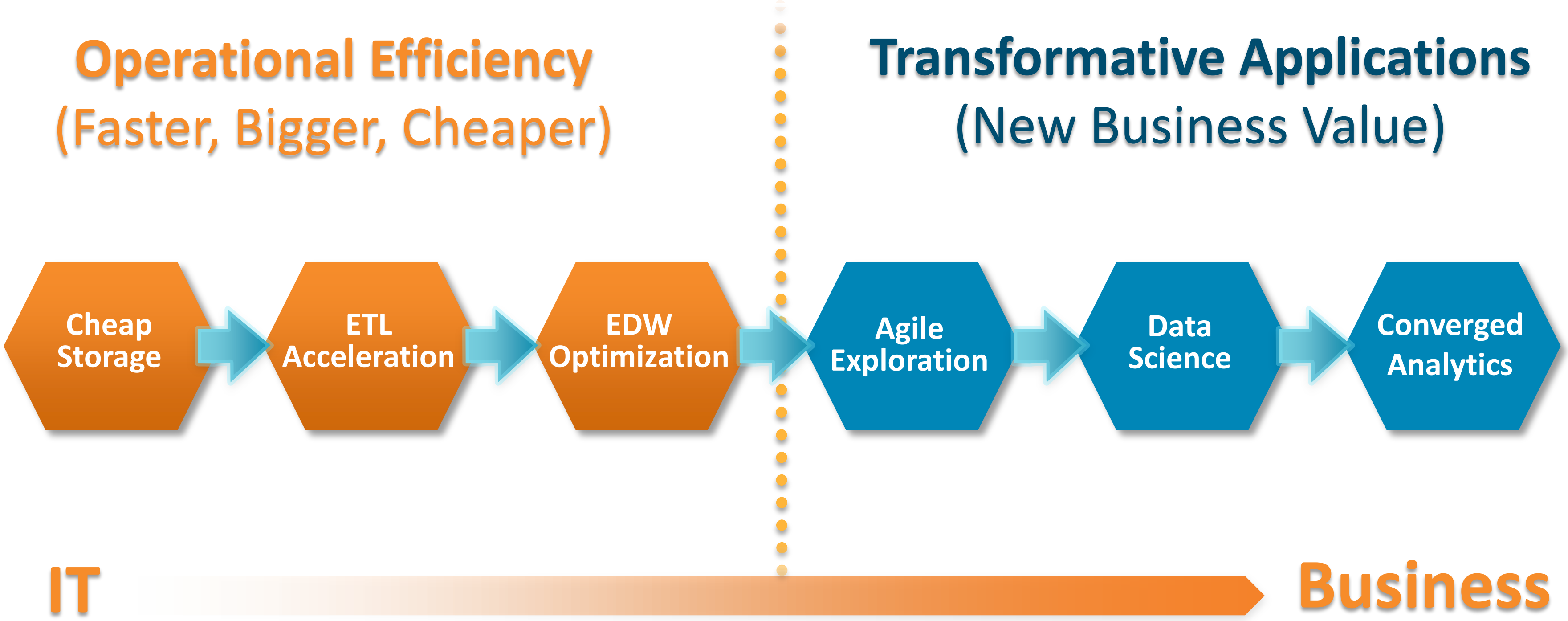
Improve IT Operations

- ETL Acceleration
- EDW Rationalization
- Mainframe Offload

IBM syncsort informatica pentaho appfluent Analytics for IT

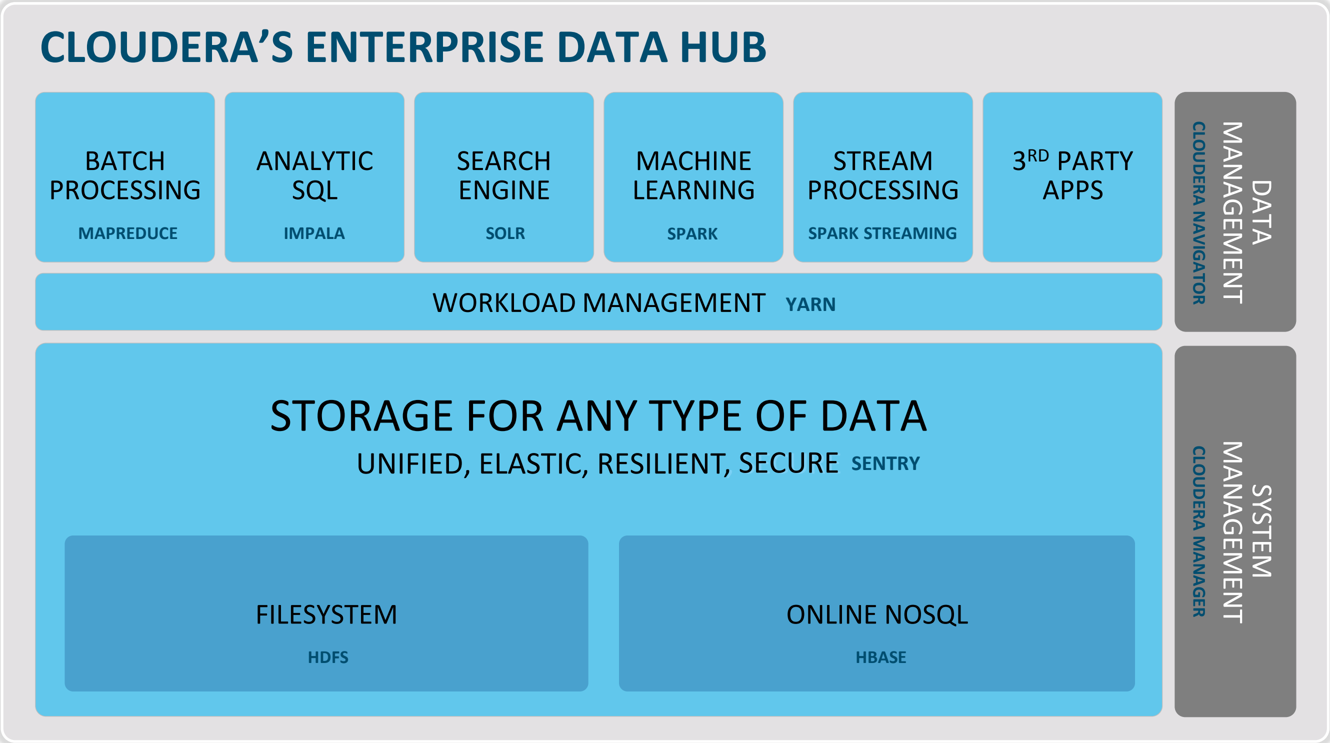
Your Journey to Achieve Full Potential

Advance from Strategy to ROI with Best Practices and Peak Performance



From Hadoop to an Enterprise Data Hub

Open Source **Scalable** **Flexible** **Cost-Effective**	✓
Managed	✓
Open Architecture	✓
Secure and Governed	✓



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 SENTRY

FILESYSTEM
HDFS

ONLINE NOSQL
HBASE

DATA MANAGEMENT
CLOUDERA NAVIGATOR

SYSTEM MANAGEMENT
CLOUDERA MANAGER

cloudera®
Ask Bigger Questions