5th Annual

# Cloudera Government Forum

cloudera

Cloudera Government Forum

# The Essentials of Apache Hadoop

## The What, Why and How to Meet Agency Objectives

Sarah Sproehnle, Vice President, Customer Success

cloudera

Cloudera
Government
Forum

# Introduction

cloudera

# What is Apache Hadoop?

- Hadoop is a software framework for storing, processing, and analyzing “big data”
  - Distributed
  - Scalable
  - Fault-tolerant
  - Open source

# A Large (and Growing) Ecosystem


Impala









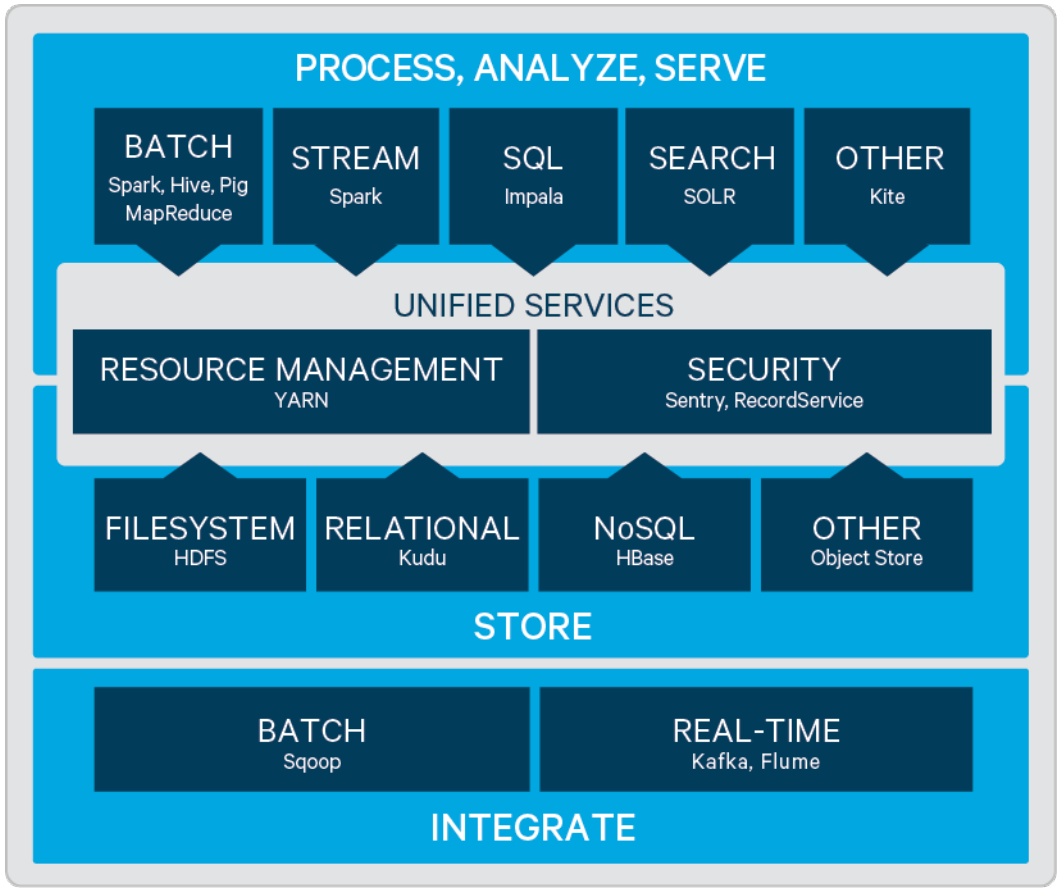
cloudera

# About Cloudera

- The leader in Apache Hadoop-based software and services
- Founded in 2008 by leading experts on Hadoop
  - Over 1000 employees
  - Global operations spanning over 20 countries
- Provides support, consulting, training, and certification for Hadoop users
- Employs committers to virtually every significant Hadoop-related project
- Many authors of industry standard books on Apache Hadoop projects
  - Tom White, Lars George, Kathleen Ting, etc.

# CDH

- CDH (Cloudera's Distribution, including Apache Hadoop)
- 100% open source, enterprise-ready distribution of Hadoop and related projects
- The most complete, tested, and widely-deployed distribution of Hadoop
- Integrates all the key Hadoop ecosystem projects




cloudera

# Vendor Integration

**Software and OEM**

alteryx, Argyle Data, Cask, Cognilytics, Datameer, H2O.ai, informatica, Information Builders, Microsoft, MicroStrategy, Oracle, Paxata, pentaho, platfora, Qlik, rapidminer, rocana, SAP Partner, sas, Securonix, Skytree, Spotfire, StreamSets, syncsort, tableau, talend, Trifacta, Zoomdata

**Data Systems**

IBM, Hewlett Packard Enterprise, Oracle, SAP Partner, Teradata

cloudera

**Platform & Cloud**

amazon web services, CenturyLink Technology Solutions, Cisco, Dell, EMC², Fujitsu, Hewlett Packard Enterprise, Google Cloud Platform, lenovo, Microsoft, NetApp, Oracle, redhat, T-Systems, Teradata, vmware

**System Integration**

accenture, Capgemini, Cognizant, CSC, Deloitte., IBM, Infosys, pwc, TATA, WIPRO

intel

cloudera

cloudera

Cloudera
Government
Forum

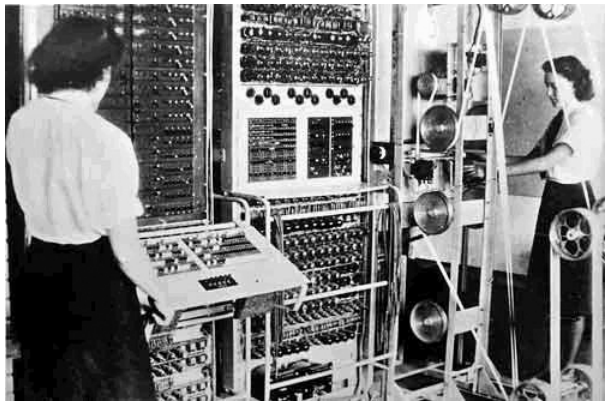
# How it Works

cloudera

# Traditional Large-Scale Computation

Traditionally, computation has been processor-bound

- Relatively small amounts of data
- Lots of complex processing



The early solution: bigger computers

- Faster processor, more memory
- But even this couldn't keep up

cloudera

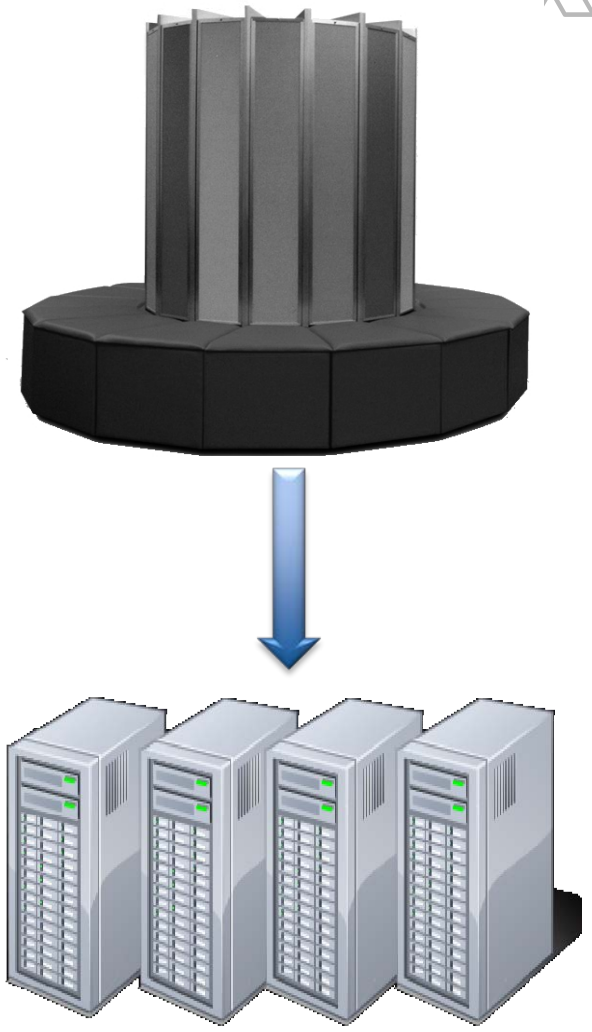
# Distributed Systems

The better solution: more computers

- Distributed systems – use multiple machines for a single job

"In pioneer days they used oxen for heavy pulling, and when one ox couldn't budge a log, we didn't try to grow a larger ox. We shouldn't be trying for bigger computers, but for *more systems* of computers."

– Grace Hopper

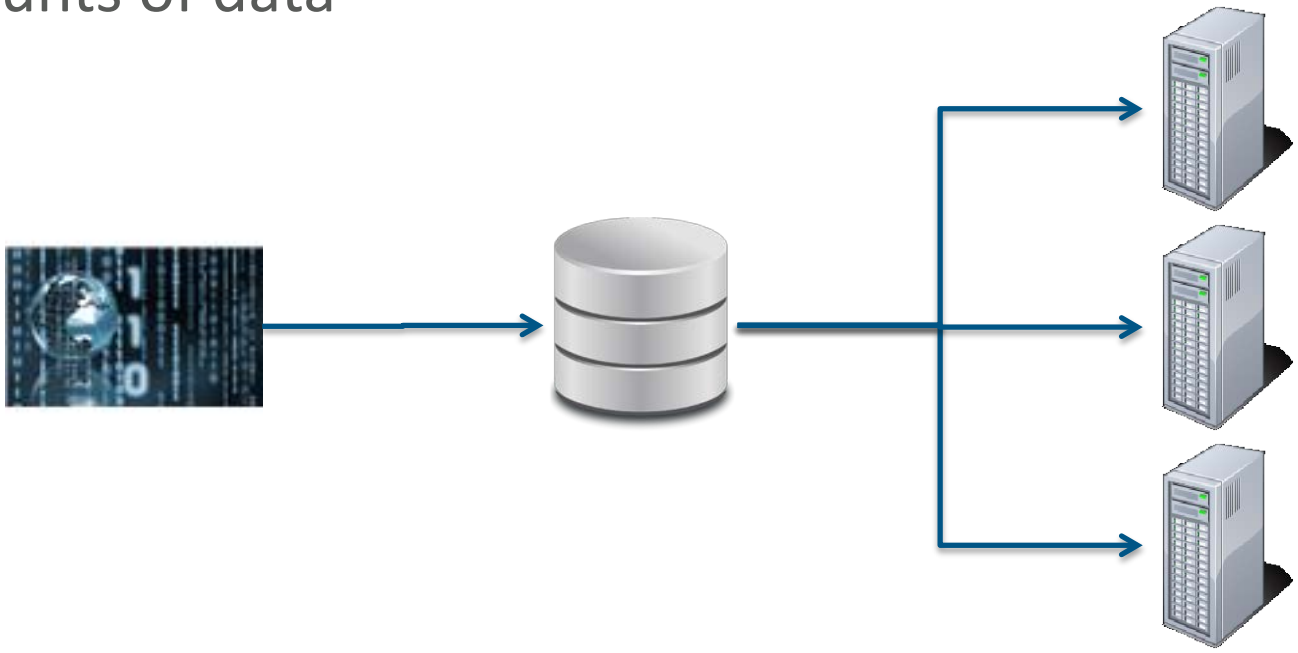


cloudera

# Distributed Systems: The Data Bottleneck (1)

- Traditionally, data is stored in a central location
- Data is copied to processors at runtime
- Fine for limited amounts of data

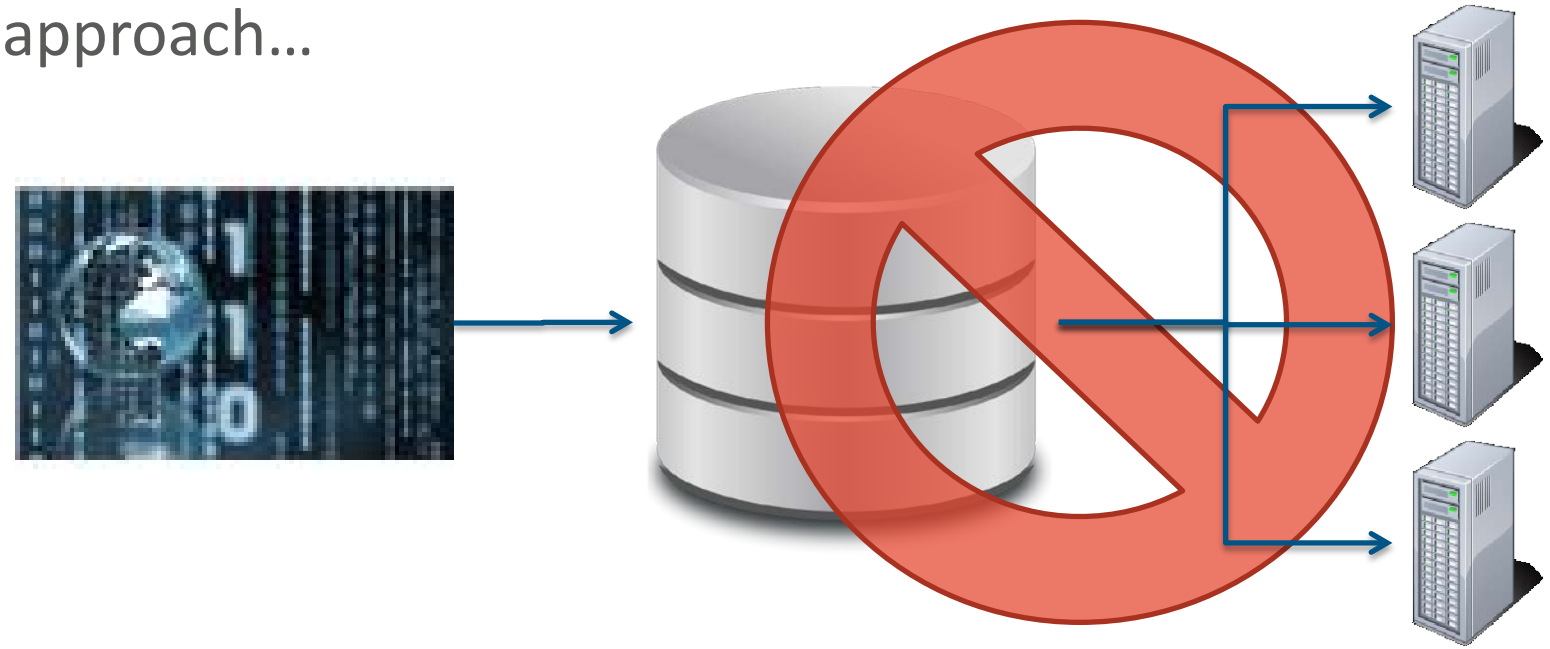




# Distributed Systems: The Data Bottleneck (2)

- Modern systems have much more data
  - terabytes+ a day
  - petabytes+ total
- We need a new approach…





# The Origins of Hadoop

- Hadoop is based on work done at Google in the late 1990s/early 2000s
- Google's problem:
  - Indexing the entire web requires massive amounts of storage
  - A new approach was required to process such large amounts of data
- Google's solution:
  - GFS, the Google File System - described in a paper released in 2003
  - Distributed MapReduce - described in a paper released in 2004
- Doug Cutting and others read these papers and implemented a similar, open source solution
  - This is what would become Hadoop

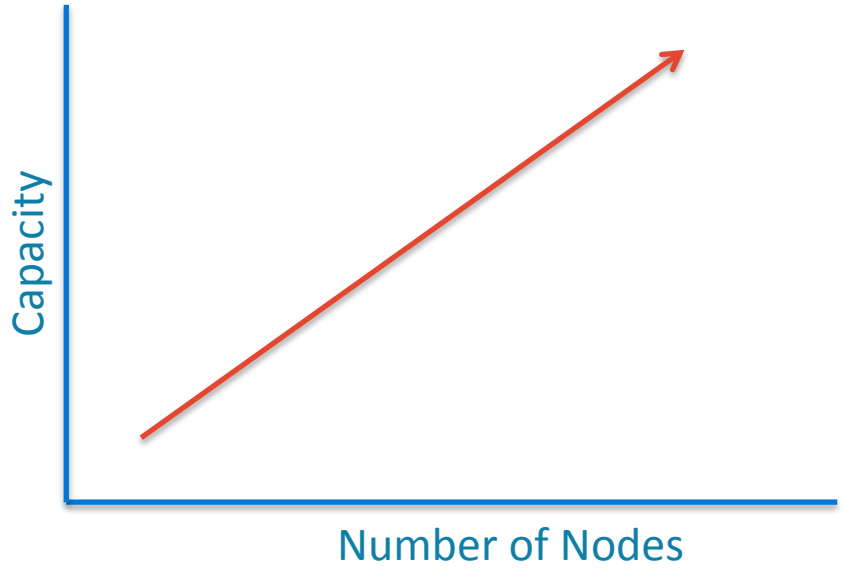
# What is Hadoop?

- Hadoop is a distributed data storage and processing platform
  - Stores massive amounts of data in a very resilient way
  - Distributes the processing to where the data is stored
- Tools built around Hadoop (the 'Hadoop ecosystem') can be configured/extended to handle many different tasks
  - Extract Transform Load (ETL)
  - BI environment
  - Predictive analytics
  - Statistical analysis
  - Machine learning

cloudera

# Hadoop is Scalable

- Adding nodes (machines) adds capacity proportionally
- Increasing load results in a graceful decline in performance
  - Not failure of the system




cloudera

# Hadoop is Fault Tolerant

- Node failure is inevitable
- What happens?
  - System continues to function
  - Master re-assigns work to a different node
  - Data replication means there is no loss of data
  - Nodes which recover rejoin the cluster automatically

# The Hadoop Ecosystem (1)

- Many tools have been developed around ‘Core Hadoop’
  - Known as the Hadoop ecosystem
- Designed to make Hadoop easier to use, or to extend its functionality
- All are open source
- The ecosystem is growing all the time

# The Hadoop Ecosystem (2)

- Examples of Hadoop ecosystem projects (all included in CDH):

| Project | What does it do? |
|---|---|
| Spark | In-memory and streaming processing framework |
| HBase | NoSQL database built on HDFS |
| Hive | SQL processing engine designed for batch workloads |
| Impala | SQL query engine designed for BI workloads |
| Parquet | Very efficient columnar data storage format |
| Sqoop | Data movement to/from RDBMSs |
| Flume, Kafka | Streaming data ingestion |
| Solr | Powerful text search functionality |
| Hue | Web-based user interface for Hadoop |
| Sentry | Authorization tool, providing security for Hadoop |

Cloudera
Government
Forum

# Hadoop in the Real World

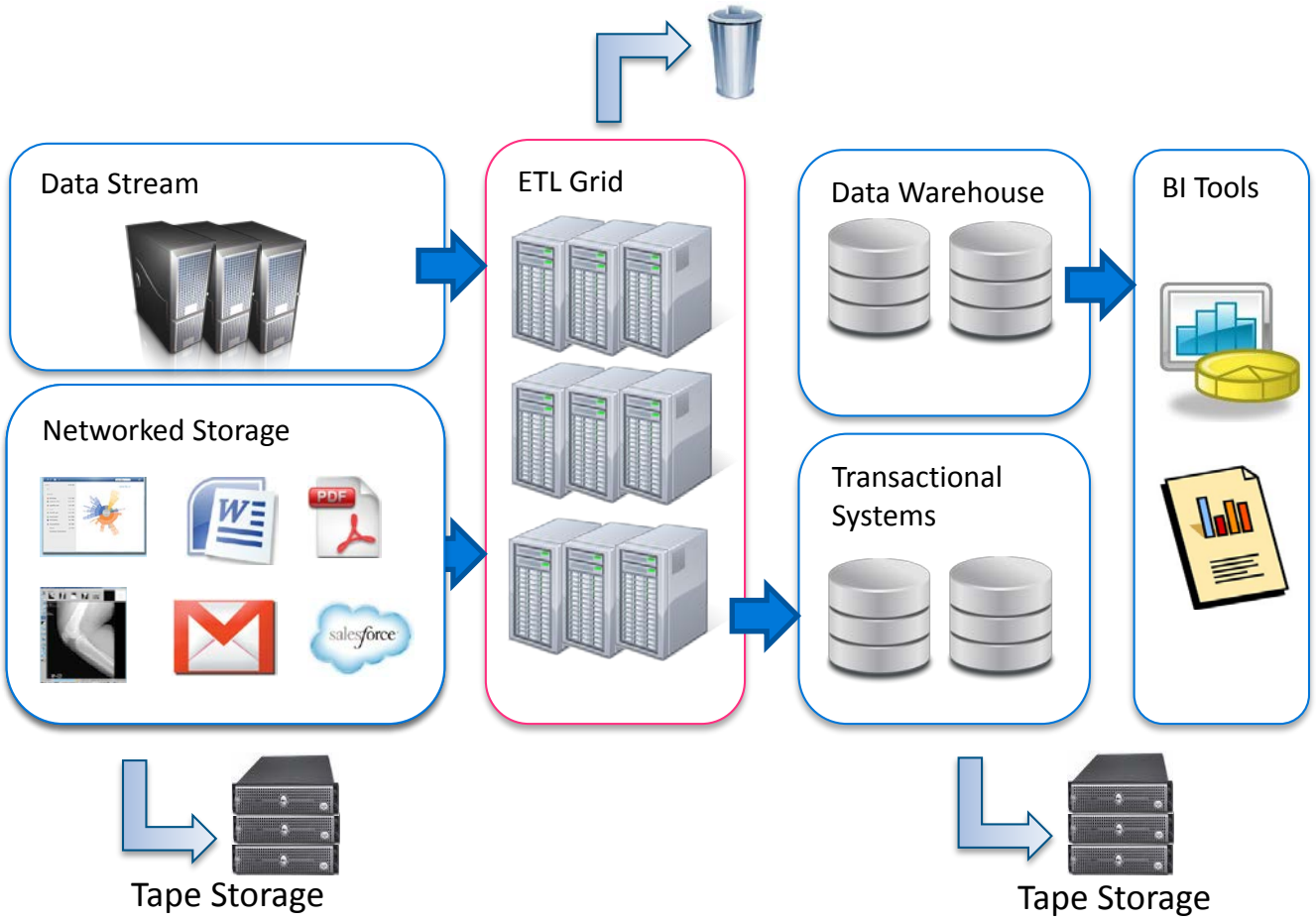
cloudera

# Five Really Popular Use Cases

- ETL Processing
- Business Intelligence
- Predictive Analytics
- Enterprise Data Hub
- Low-cost Storage of Large Data Volumes

cloudera

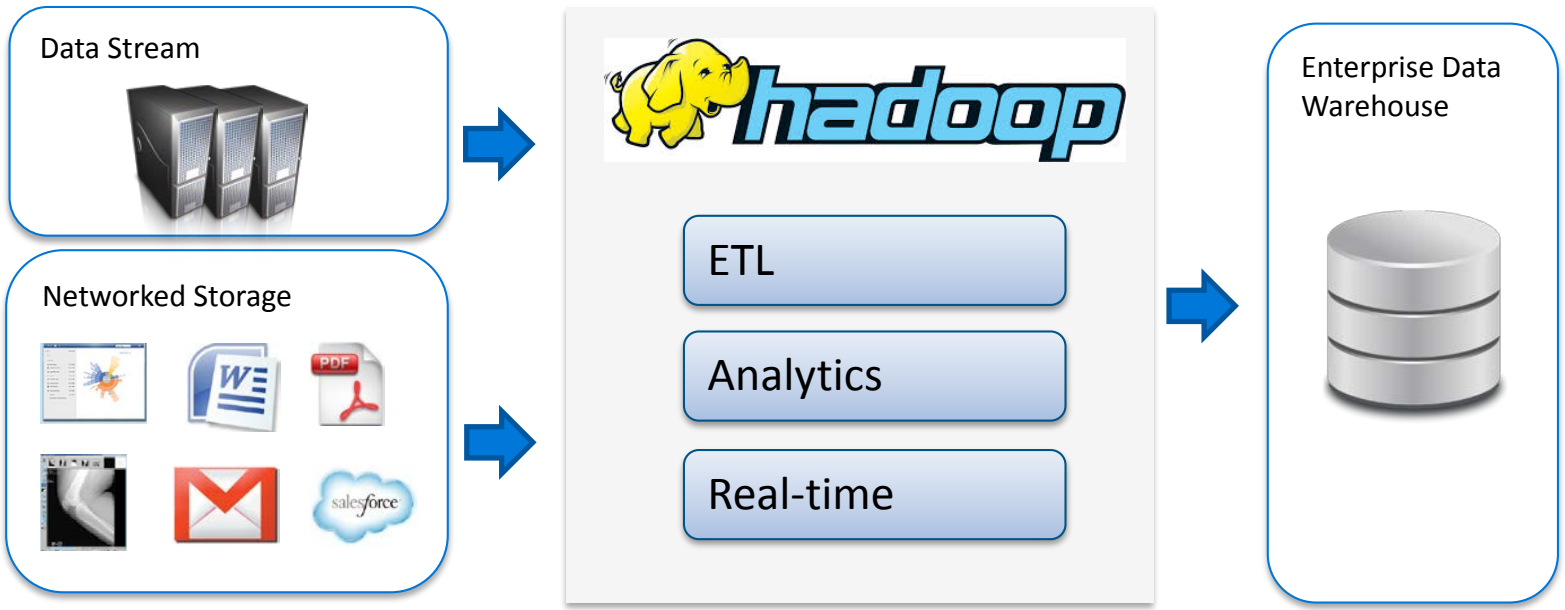
# # 1 – Traditional ETL Processing




- ETL: Extract, Transform, Load
- Challenges:
  - Too much data
  - Takes too long
  - Too costly

cloudera

# # 1 – ETL Processing with Hadoop




- Hadoop cluster is used for ETL
  - Often now ELT: Extract, Load, *then* Transform
- Structured and unstructured data is moved into the cluster
- Once processed, data can be analyzed in Hadoop or moved to the EDW

cloudera

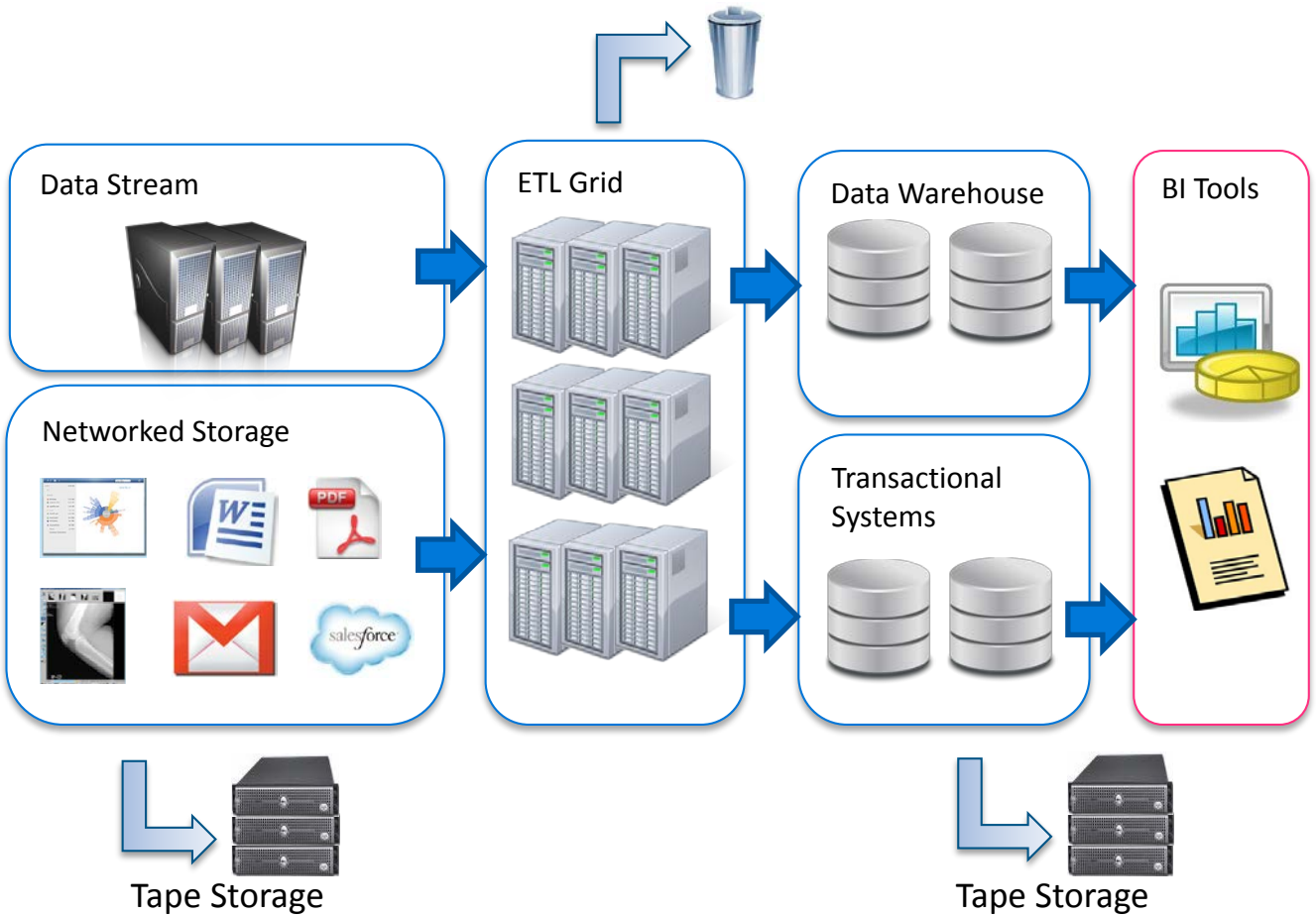
# # 1 – Vendor Integration - ETL


- For more information visit: http://www.cloudera.com/partners/partners-listing.html

cloudera

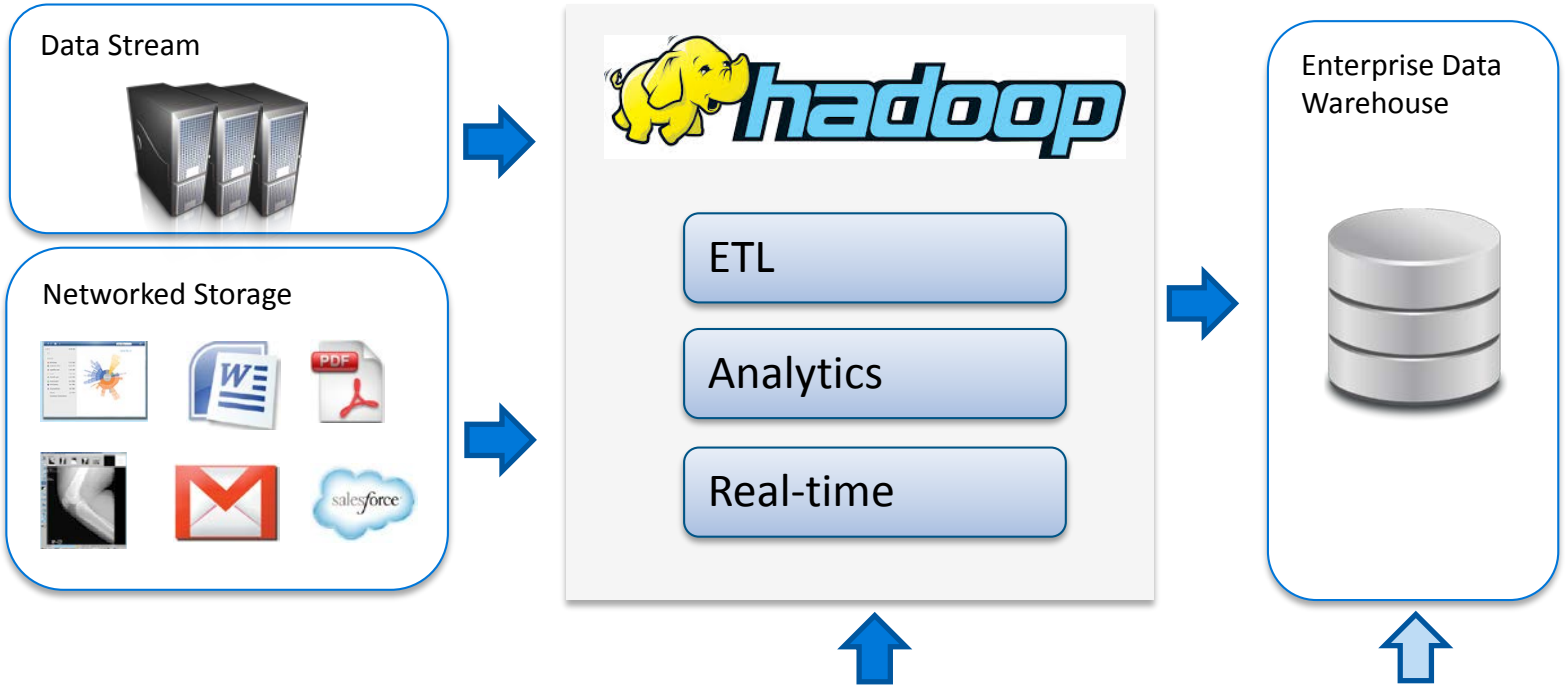
# # 2 – Traditional Business Intelligence




- BI traditionally takes place at the data warehouse layer
- Problem: EDW can't keep up with growing data volumes
  - Performance declines
  - Increasing capacity can be very expensive
  - Archived data is not available for analysis

cloudera

# # 2 – Business Intelligence with Hadoop




- BI tools can use Hadoop for much of their work
  - Analyze *all* the data
- Use the EDW for the tasks for which it is best suited

cloudera

# # 2 – Vendor Integration - BI


- For more information visit: http://www.cloudera.com/partners/partners-listing.html

# # 3 – Predictive Analytics

- Predictive Analytics (Eckerson Group definition)
  - The use of statistical or machine learning models to discover patterns and relationships in data that can help business people predict future behavior or activity
- The Hadoop platform can run analytic workloads on large volumes of diverse data
  - Statistical models can be created and run inside the Hadoop environment
- Entire data sets can be used to create models
  - There is no need to sample data
- Hadoop provides an environment that makes self-service analytics possible
  - No need for ETL developers to stage data for data scientists

# # 3 - Predictive Analytics: Cerner Corporation

**Cerner Corporation**

- Healthcare IT space
- Solutions and Services - Used by 54,000 medical facilities around the world

**The problem**

- Healthcare data is fragmented and lives in silos
- The data was used for historical reporting

**The solution**

- Build a comprehensive view of population health using a single platform
- Use predictive analytics to
  - Improve patient outcomes
  - Increase efficiency / Reduce costs

cloudera

# # 3 – Vendor Integration – Predictive Analytics

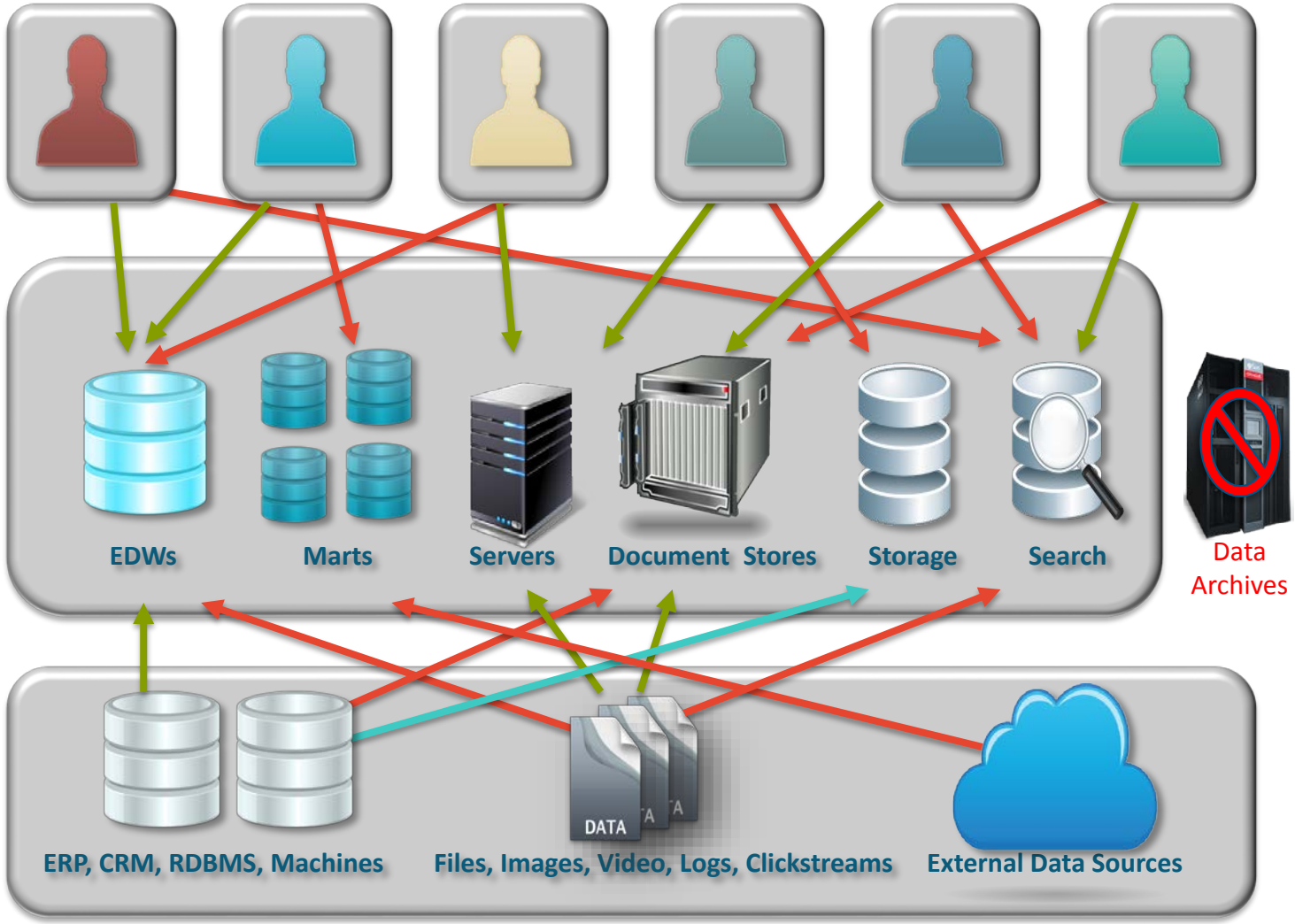

- For more information visit: http://www.cloudera.com/partners/partners-listing.html

# # 4 – The Need for the Enterprise Data Hub

Thousands of Employees & Lots of Inaccessible Information

Heterogeneous Legacy IT Infrastructure

Silos of Multi-Structured Data Difficult to Integrate



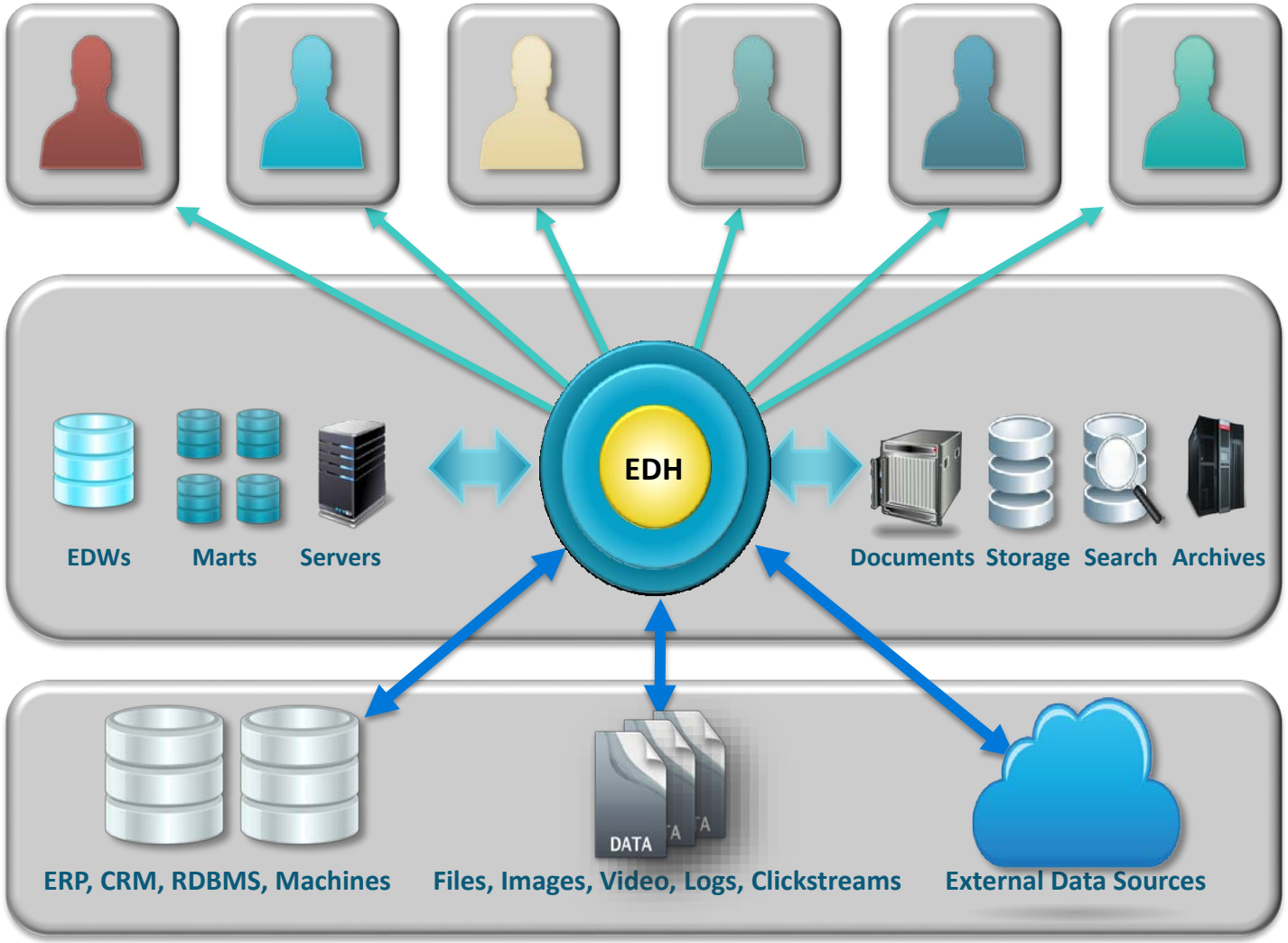

cloudera

# # 4 – The Enterprise Data Hub: One Unified System

Information & data accessible by all for insight using leading tools and apps

Enterprise Data Hub
Unified Data Management Infrastructure

Ingest All Data
Any Type
Any Scale
From Any Source




cloudera

# # 5 - Low-cost Data Storage (1)

- Hadoop combines industry standard hardware and a fault tolerant architecture
  - This combination provides a very cost effective data storage platform
- The data stored on Hadoop is protected from loss by HDFS
  - Data replication ensures that no data is lost
  - The self-healing nature of Hadoop ensures that data is available when you need it
- Hadoop enables users to store data which was previously discarded due to the cost of saving it
  - Transactional
  - Social media
  - Sensor
  - Click stream

cloudera

# # 5 - Low-cost Data Storage (2)

- The low cost of HDFS storage enables the following use-cases:
  - Enterprise Data Hub (EDH)
  - Active data archive
  - Staging area for data warehouses
  - Staging area for analytics store
  - Sandbox for data discovery
  - Sandbox for analytics

cloudera

# In Summary, Why Do You Need Hadoop?

- More data is coming
  - Internet of things
  - Sensor data
  - Streaming
- More data means bigger questions, better answers
- Hadoop easily scales to store and handle all of your data
- Hadoop is cost-effective
  - Provides a significant cost-per-terabyte saving over traditional, legacy systems
- Hadoop integrates with your existing datacenter components
- Answer questions that you previously could not ask

cloudera

Cloudera
Government
Forum

# Thank You