cloudera®

Cloudera Essentials for Apache Hadoop

201608

cloudera®

Introduction

Chapter 1

Course Chapters

- **Introduction**
- Hadoop Basics
- The Hadoop Ecosystem
- An Introduction to Hadoop Architecture
- Hadoop in the Real World
- Managing Hadoop
- Conclusion

Chapter Topics

Introduction

- **About this Course**
- About Apache Hadoop
- About Cloudera
- Course Logistics
- Introductions

Course Objectives

During this course, you will learn

- **Why Hadoop is needed**
- **What type of problems can be solved with Hadoop**
- **The basic concepts of the Hadoop Distributed File System**
- **What components are included in the Hadoop ecosystem**
- **The basics of Hadoop's architecture**
- **Who is using Hadoop**
- **What resources are available to assist in managing your Hadoop deployment**

Chapter Topics

Introduction

- About this Course
- **About Apache Hadoop**
- About Cloudera
- Course Logistics
- Introductions

What is Apache Hadoop?

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

Some Facts About Apache Hadoop

- **Open source**
  - Overseen by the Apache Software Foundation
- **Over 90 active committers to core Hadoop from over 20 companies**
  - Cloudera, Intel, LinkedIn, Facebook, Yahoo!, and more
- **Hundreds more committers on other Hadoop-related projects and tools**
  - Known as the "Hadoop ecosystem"
- **Many hundreds of other contributors writing features, fixing bugs, and so on.**

A Large (and Growing) Ecosystem

Pig

Zookeeper

Vendor Integration

BI/Analytics

ETL

Database

OS/Cloud/ System Management

Hardware

Who uses Hadoop?

Chapter Topics

Introduction

- About this Course
- About Apache Hadoop
- **About Cloudera**
- Course Logistics
- Introductions

About Cloudera (1)

- **The leader in Apache Hadoop-based software and services**
  - Founded by Hadoop experts from Facebook, Yahoo, Google, and Oracle
- **Provides support, consulting, training, and certification for Hadoop users**
  - A global leader with 1,000+ employees spanning more than 20 countries
- **Staff includes committers to virtually all Hadoop projects**
  - Many authors of industry standard books on Apache Hadoop projects

About Cloudera (2)

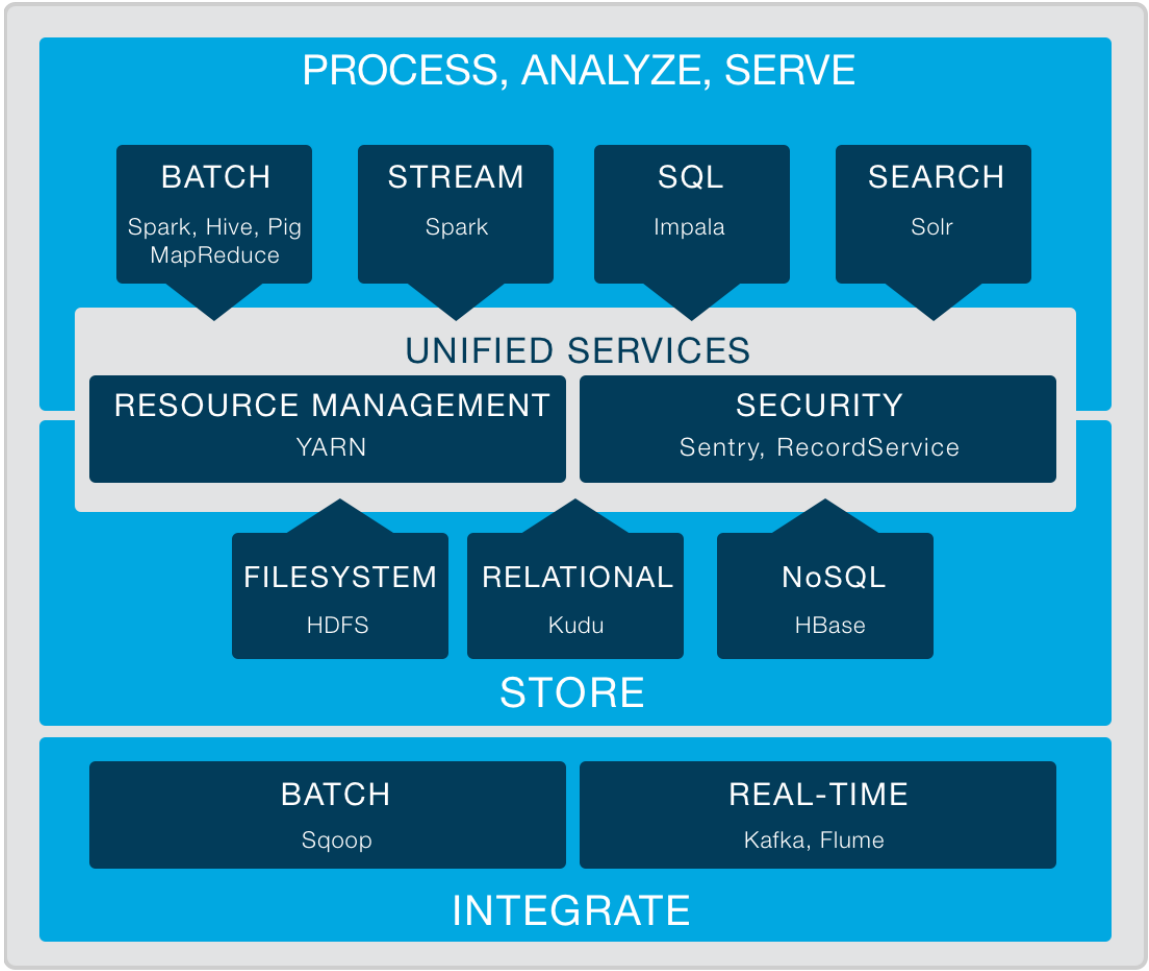
- **Our customers include many key users of Hadoop**
- **We offer several public training courses, such as**
  - *Cloudera Developer Training for Spark and Hadoop*
  - *Cloudera Administrator Training for Apache Hadoop*
  - *Cloudera Data Analyst Training: Using Pig, Hive, and Impala with Hadoop*
  - *Cloudera Search Training*
  - *Data Science at Scale using Spark and Hadoop*
  - *Cloudera Training for Apache HBase*
- **On-site, customized, and online OnDemand trainings are also available**

About Cloudera (3)

- **In addition to our public training courses, Cloudera offers two levels of certifications**
- **Cloudera Certified Professional (CCP)**
  - The industry's most demanding performance-based certification, CCP evaluates and recognizes a candidate's mastery of the technical skills most sought after by employers
  - CCP Data Engineer
  - CCP Data Scientist
- **Cloudera Certified Associate (CCA)**
  - CCA exams validate foundational skills and provide the groundwork for a candidate to achieve mastery under the CCP program
  - CCA Spark and Hadoop Developer
  - Cloudera Certified Administrator for Apache Hadoop (CCAH)

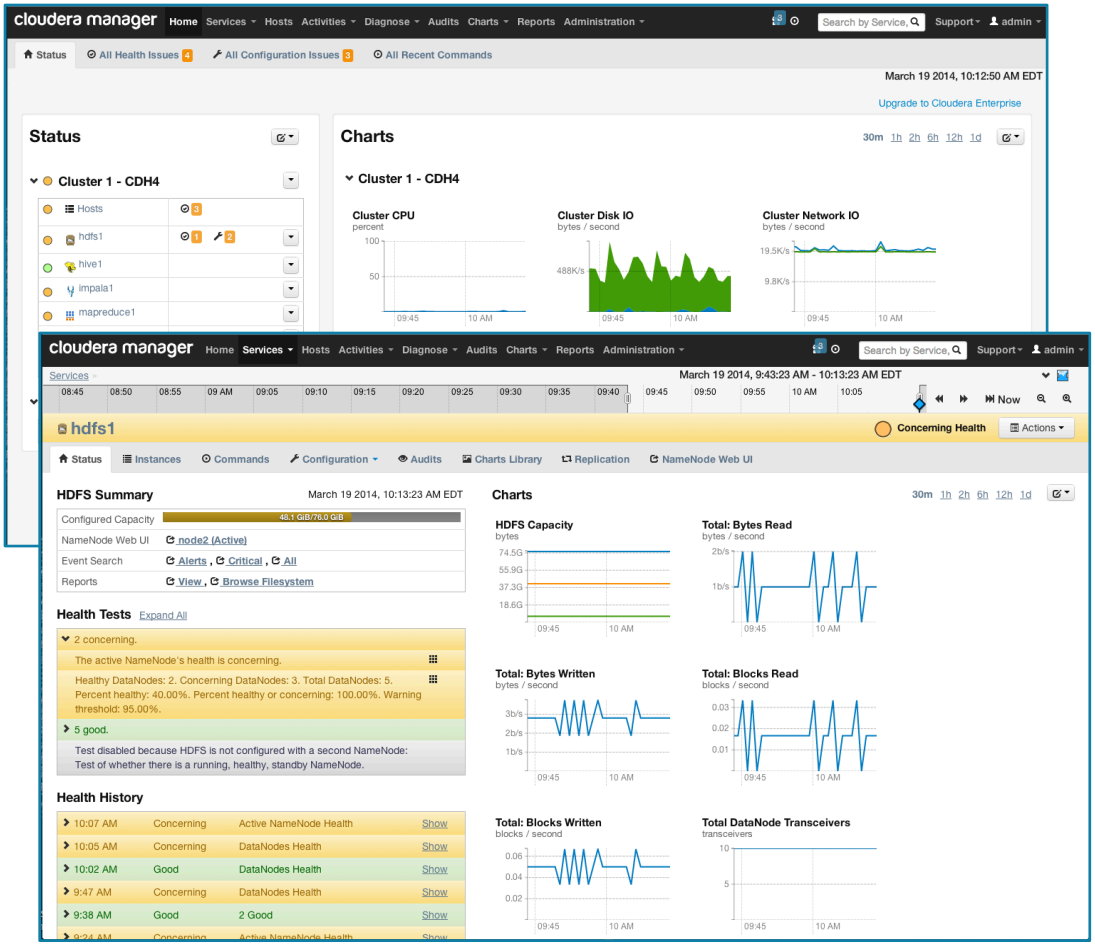
CDH (Cloudera's Distribution including Apache Hadoop)

- 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
- Available in many convenient formats



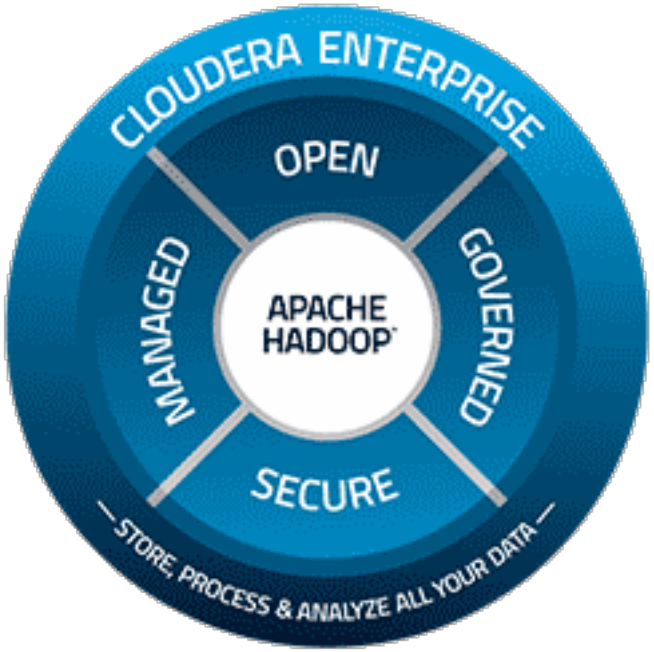
Cloudera Express

- **Cloudera Express**
  - Completely free to download and use
- **The best way to get started with Hadoop**
- **Includes CDH**
- **Includes Cloudera Manager**
  - End-to-end administration for Hadoop
  - Deploy, manage, and monitor your cluster



Cloudera Enterprise

- **Cloudera Enterprise**
  - Subscription product including CDH and Cloudera Manager
- **Includes support**
- **Includes extra Cloudera Manager features**
  - Configuration history and rollbacks
  - Rolling updates
  - LDAP integration
  - SNMP support
  - Automated disaster recovery



- **Extend capabilities with Cloudera Navigator subscription**
  - Event auditing, metadata tagging capabilities, lineage exploration
  - Available in both the Cloudera Enterprise Flex and Data Hub editions

Chapter Topics

Introduction

- About this Course
- About Apache Hadoop
- About Cloudera
- **Course Logistics**
- Introductions

Logistics

- **Class start and finish times**
- **Breaks**
- **Restrooms**
- **Wi-Fi access**

Your instructor will give you details on how to access the course materials and exercise instructions for the class

Chapter Topics

Introduction

- About this Course
- About Apache Hadoop
- About Cloudera
- Course Logistics
- **Introductions**

Introductions

- **About your instructor**
- **About you**
  - Where do you work? What do you do there?
  - What programming languages have you used?
  - Do you have experience with UNIX or Linux?
  - How much Hadoop experience do you have?
  - What do you expect to gain from this course?

cloudera®

Hadoop Basics

Chapter 2

Course Chapters

- Introduction
- **Hadoop Basics**
- The Hadoop Ecosystem
- An Introduction to Hadoop Architecture
- Hadoop in the Real World
- Managing Hadoop
- Conclusion

Hadoop Basics

In this chapter you will learn

- **The insights that led to Hadoop**
- **What Hadoop is**
- **Who uses Hadoop**
- **Why you need Hadoop**

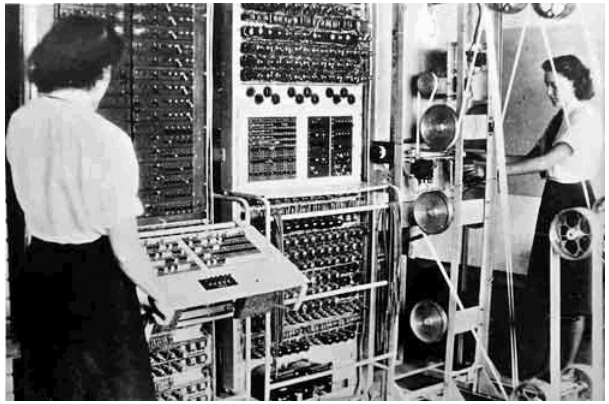
Chapter Topics

Hadoop Basics

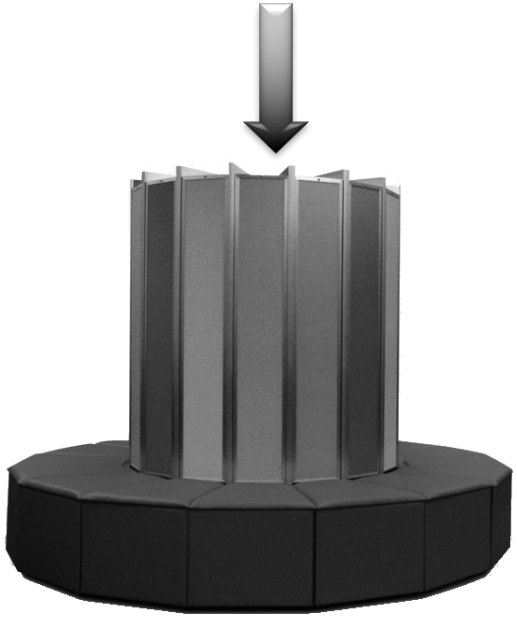
- **The Motivation for Hadoop**
- What Is Hadoop?
- Who Uses Hadoop?
- Why Do You Need Hadoop?
- Essentials Points

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

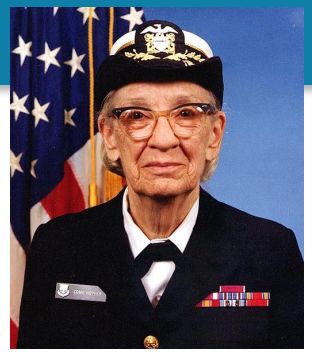


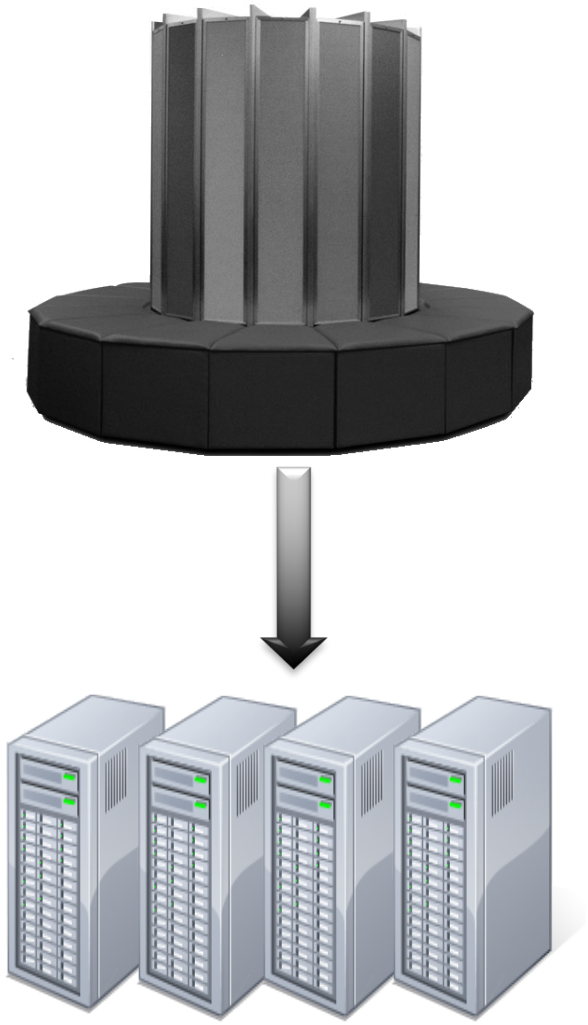
Distributed Systems

- **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



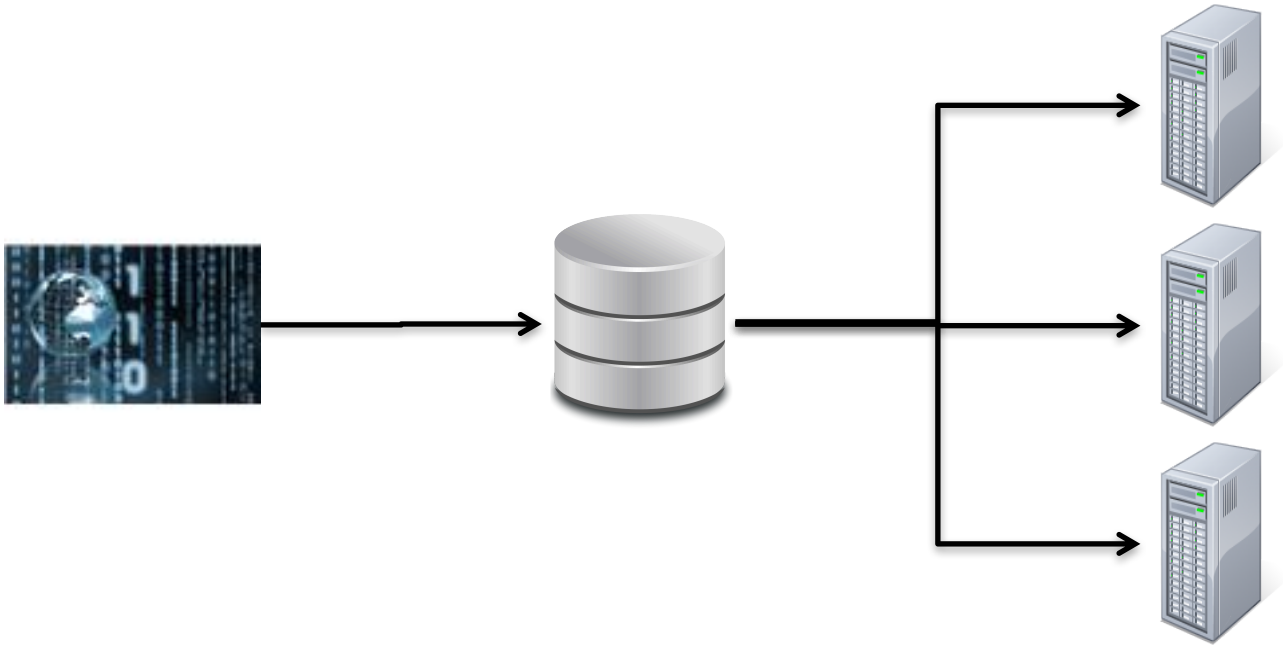


Distributed Systems: Challenges

- **Challenges with distributed systems**
  - Programming complexity
    - Keeping data and processes in sync
  - Finite bandwidth
  - Partial failures

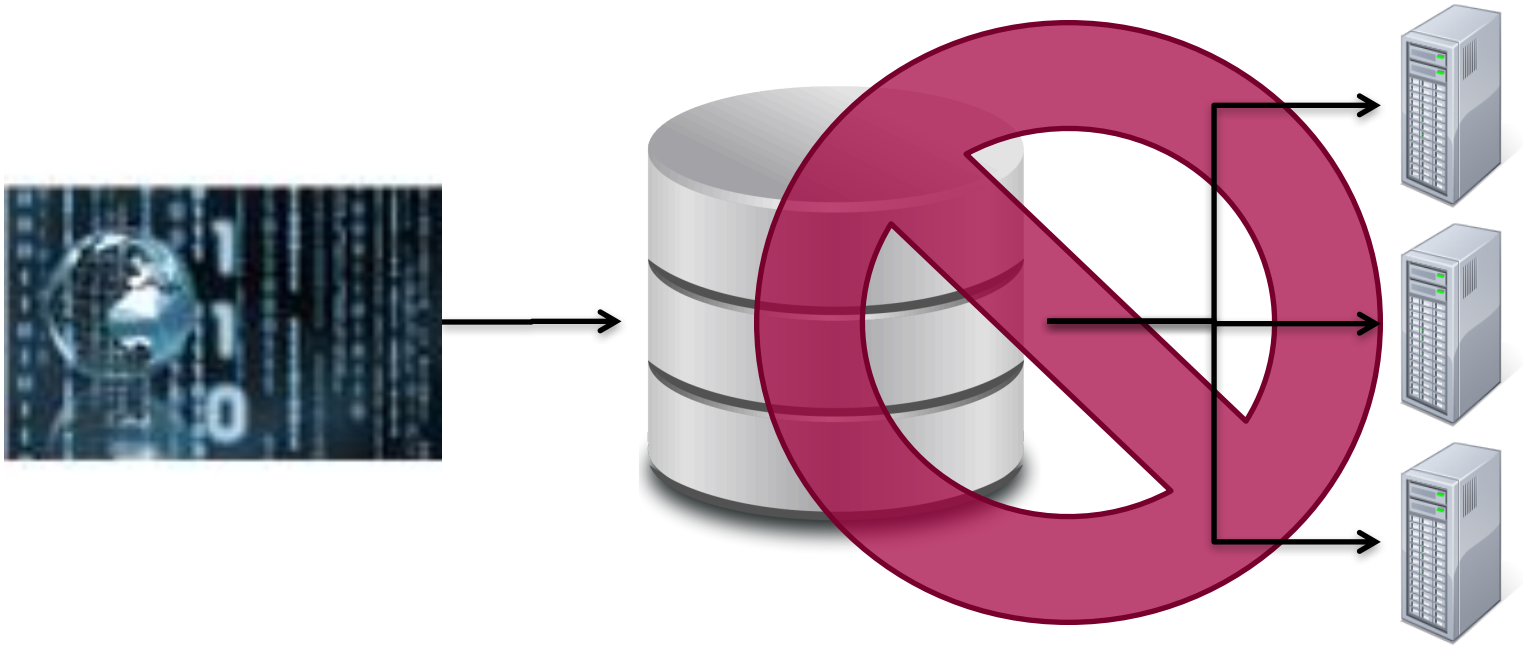
Distributed Systems: The Data Bottleneck (1)

- **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

Chapter Topics

Hadoop Basics

- The Motivation for Hadoop
- **What is Hadoop?**
- Who Uses Hadoop?
- Why Do You Need Hadoop?
- Essentials Points

What is Hadoop?

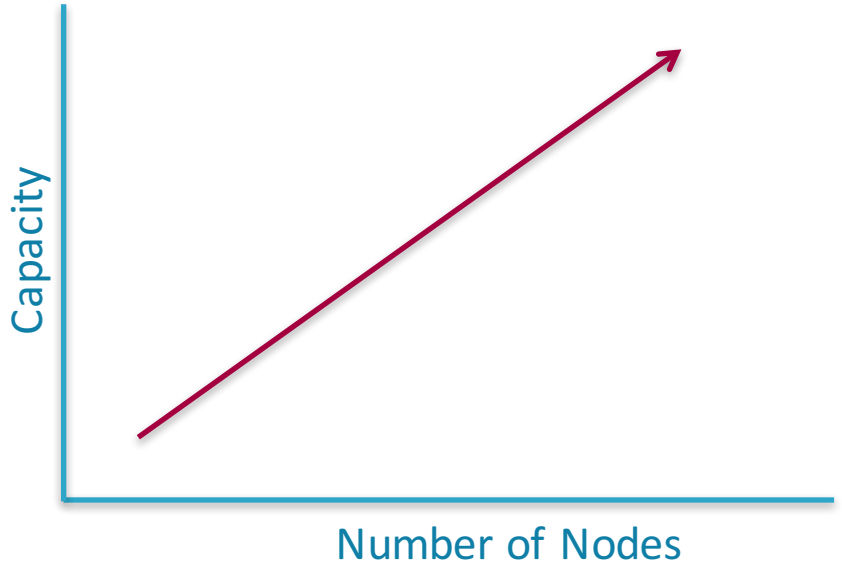
- **Hadoop is a distributed data storage and processing platform**
  - Stores massive amounts of data in a very resilient way
  - Handles low-level distributed system details and enables your developers to focus on the business problems
- **Tools built around Hadoop (the "Hadoop ecosystem") can be configured/extended to handle many different tasks**
  - Extract Transform Load (ETL)
  - Business Intelligence (BI) environment
  - General data storage
  - Predictive analytics
  - Statistical analysis
  - Machine learning
  - …

Core Hadoop is a File System and a Processing Framework

- **The Hadoop Distributed File System (HDFS)**
  - Any type of file can be stored in HDFS
  - Data is split into chunks and replicated as it is written
    - Provides resiliency and high availability
    - Handled automatically by Hadoop
- **YARN (Yet Another Resource Negotiator)**
  - Manages the processing resources of the Hadoop cluster
  - Schedules jobs
  - Runs processing frameworks
- **MapReduce**
  - A distributed processing framework

Hadoop is Scalable

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



Hadoop is Fault Tolerant

- **Node failure is inevitable**
- **What happens?**
  - System continues to function
  - Master reassigns 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 execution 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	Enables users to find the data they need
Hue	Web-based user interface for Hadoop
Oozie	Workflow scheduler used to manage jobs
Sentry	Authorization tool, providing security for Hadoop

Hadoop is Constantly Evolving

- **New features are frequently added to the Hadoop ecosystem**
  - Processing frameworks
  - Support for new file types
  - Performance enhancements
- **Open Source creates an environment which encourages evolution**
- **Examples of recent ecosystem projects**
  - Spark
  - Parquet
  - Sentry
  - Kafka

Chapter Topics

Hadoop Basics

- The Motivation for Hadoop
- What Is Hadoop?
- **Who Uses Hadoop?**
- Why Do You Need Hadoop?
- Essentials Points

Hadoop Users: Financial Services

- **JPMorgan Chase**
  - Fraud detection, anti-money laundering, and self-service applications
  - Now stores and processes data that was previously discarded
- **MasterCard**
  - Credit card industry
  - First Certified Payment Card Industry (PCI) Data Security Standards Hadoop solution
    - In conjunction with Cloudera Manager, Cloudera Navigator
- **Western Union**
  - Cross-border, consumer-to-consumer money transfers and bill payments
  - 360-Degree customer view

Hadoop Users: Insurance

- **Allstate**
  - Fraud detection, Predictive analytics, and reporting
  - Leveraging Hadoop to complement, not replace, existing system
  - They have been able to expedite analytic reports by 500x
- **RelayHealth**
  - Processes healthcare provider-to-payer interactions
  - Creating analytics platforms on Hadoop
- **Markerstudy**
  - Fraud detection and prevention at point-of-sale
  - Uses Hadoop to provide the most appropriate price and product to every customer

Hadoop Users: Telecommunications

- **Telkomsel**
  - Data discovery and analytics
  - Offload extract, transform, and load (ETL) operations from their data warehouse
- **SFR**
  - 360-degree customer view spanning devices and data sources
  - Offload data ingest, processing, and exploration from the data warehouse

Children's Healthcare of Atlanta

- **Improving the care of children in the Neonatal Intensive Care Unit (NICU)**
- **Saves and analyzes sensor data**
  - Determining what procedures cause most stress for infants, so they can be minimized
- **Started with a very small cluster**
  - Six nodes, built from scavenged PCs

Chapter Topics

Hadoop Basics

- The Motivation for Hadoop
- What Is Hadoop?
- Who Uses Hadoop?
- **Why Do You Need Hadoop?**
- Essentials Points

Why Do You Need Hadoop? (1)

- **More data is coming**
  - Internet of things
  - Sensor data
  - Streaming
- **More data means bigger questions**
- **More data means better answers**
- **Hadoop easily scales to store and handle all of your data**
- **Hadoop is cost-effective**
  - Typically provides a significant cost-per-terabyte saving over traditional, legacy systems
- **Hadoop integrates with your existing datacenter components**

Why Do You Need Hadoop? (2)

- **Go from surviving to innovating**
- **Turn cost centers into revenue generators**
- **Hadoop lets you exploit the data you have**
- **Hadoop lets you exploit data you have been throwing away**
- **Hadoop is the foundation for the applications of the future**
  - Hadoop can help you break the 80/20 budget cycle
    - 80% Keep The Lights On (KTLO)
    - 20% for business driven purposes (your 20% goes much further)
- **Answer questions that you previously could not ask**

Chapter Topics

Hadoop Basics

- The Motivation for Hadoop
- What Is Hadoop?
- Who Uses Hadoop?
- Why Do You Need Hadoop?
- **Essential Points**

Essential Points

- **Hadoop arose from a need to store and process massive amounts of data in a cost-effective way**
- **Hadoop is a distributed data processing platform**
  - Parallelism is built-in so your developers can focus on business problems
  - A growing ecosystem provides new features and better ease-of-use
- **Hadoop is used by organizations across a wide range of industries**
  - From financial services to healthcare, oil exploration to social media
- **Using Hadoop offers many benefits for your organization**
  - Lower costs allow you to keep data you would previously have discarded
  - More data means more insights

cloudera®

The Hadoop Ecosystem

Chapter 3

Course Chapters

- Introduction
- Hadoop Basics
- **The Hadoop Ecosystem**
- An Introduction to Hadoop Architecture
- Hadoop in the Real World
- Managing Hadoop
- Conclusion

The Hadoop Ecosystem

In this chapter you will learn

- **How the Hadoop Distributed File System (HDFS) works**
- **The purpose of YARN**
- **What features other parts of the Hadoop ecosystem provide**

Chapter Topics

The Hadoop Ecosystem

- **Introduction**
- Core Hadoop: HDFS, MapReduce, and YARN
- Data Integration: Flume, Kafka, and Sqoop
- Data Processing: Spark
- Data Analysis: Hive and Impala
- Data Exploration: Cloudera Search
- User Interface: Hue
- Data Storage: HBase
- Data Security: Sentry
- Essential Points

Introduction

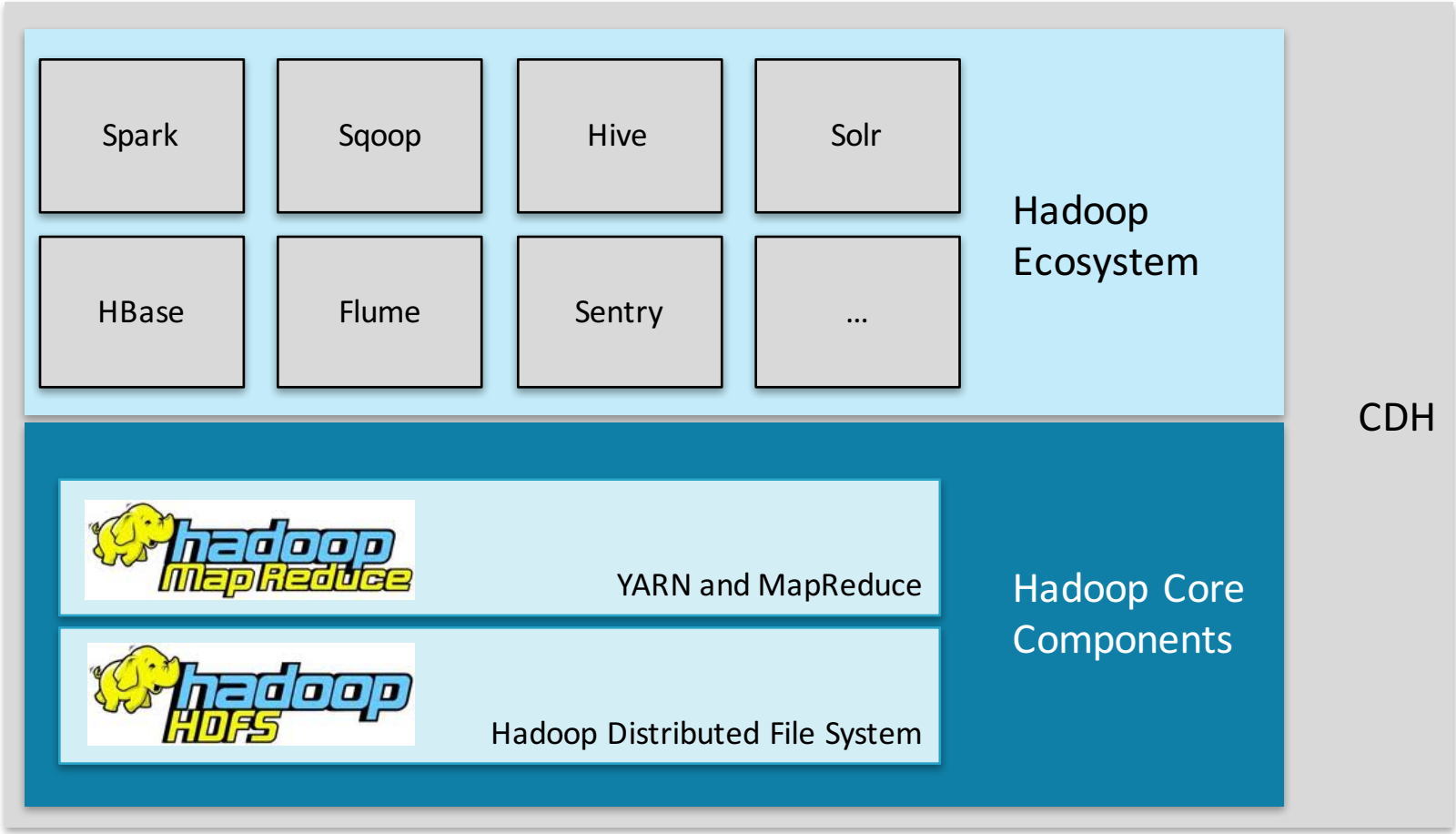
- **Hadoop takes a different approach to distributed computing compared to traditional systems**
- **There are two key concepts**
  - Distribute data when it is loaded into the system
  - Run computation where the data is stored
- **Data is stored on industry-standard hardware**
- **Add capacity by scaling *out* (more machines), not scaling *up* (to a bigger machine)**
- **The Hadoop ecosystem simplifies distributed computing so programmers can focus on the application**

Chapter Topics

The Hadoop Ecosystem

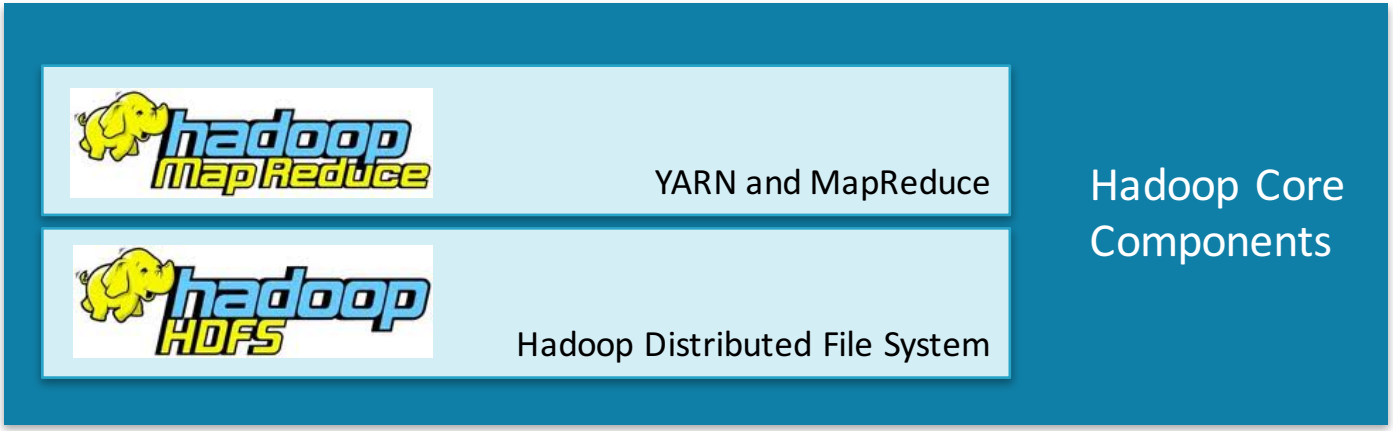
- Introduction
- **Core Hadoop: HDFS, MapReduce, and YARN**
- Data Integration: Flume, Kafka, and Sqoop
- Data Processing: Spark
- Data Analysis: Hive and Impala
- Data Exploration: Cloudera Search
- User Interface: Hue
- Data Storage: HBase
- Data Security: Sentry
- Essential Points

Hadoop Components



Core Components: HDFS, YARN, and MapReduce

- **HDFS (Hadoop Distributed File System)**
  - Stores data on the cluster
- **MapReduce**
  - Processes data on the cluster
- **YARN**
  - Schedules work on the cluster

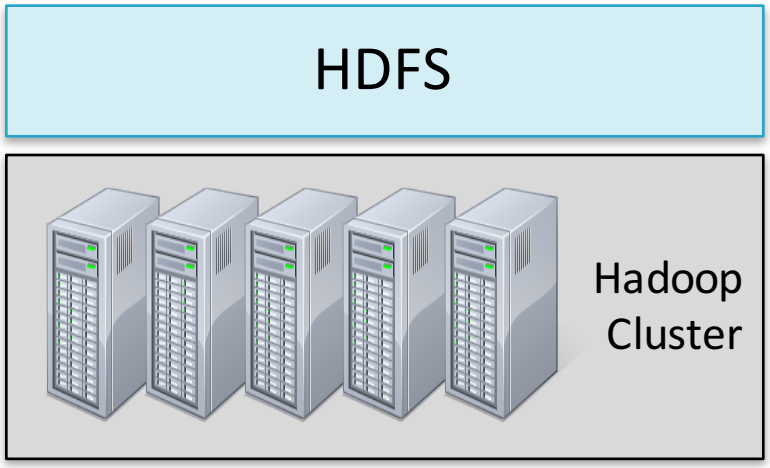


The Hadoop Distributed File System (HDFS)

- **HDFS is the storage layer for Hadoop**
- **A filesystem which can store any type of data**
- **Provides inexpensive and reliable storage for massive amounts of data**
  - Data is replicated across computers
- **HDFS performs best with a "modest" number of large files**
  - Millions, rather than billions, of files
  - Each file typically 100MB or more
- **File in HDFS are "write once"**
  - Appends are permitted
  - No random writes are allowed

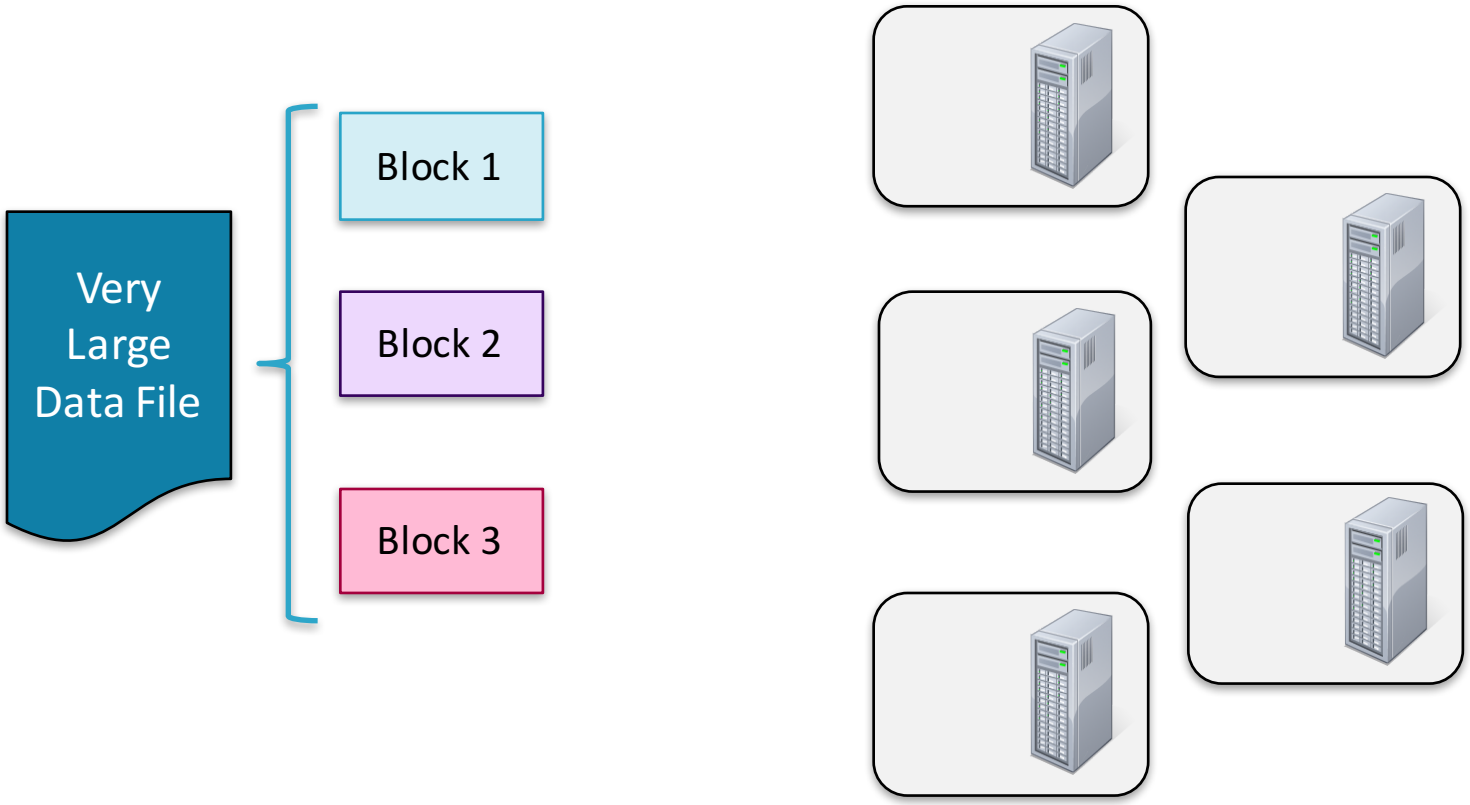
HDFS Basic Concepts

- **HDFS is a filesystem written in Java**
- **Sits on top of a native filesystem**
- **Scalable**
- **Fault tolerant**
- **Supports efficient processing with MapReduce, Spark, and other frameworks**



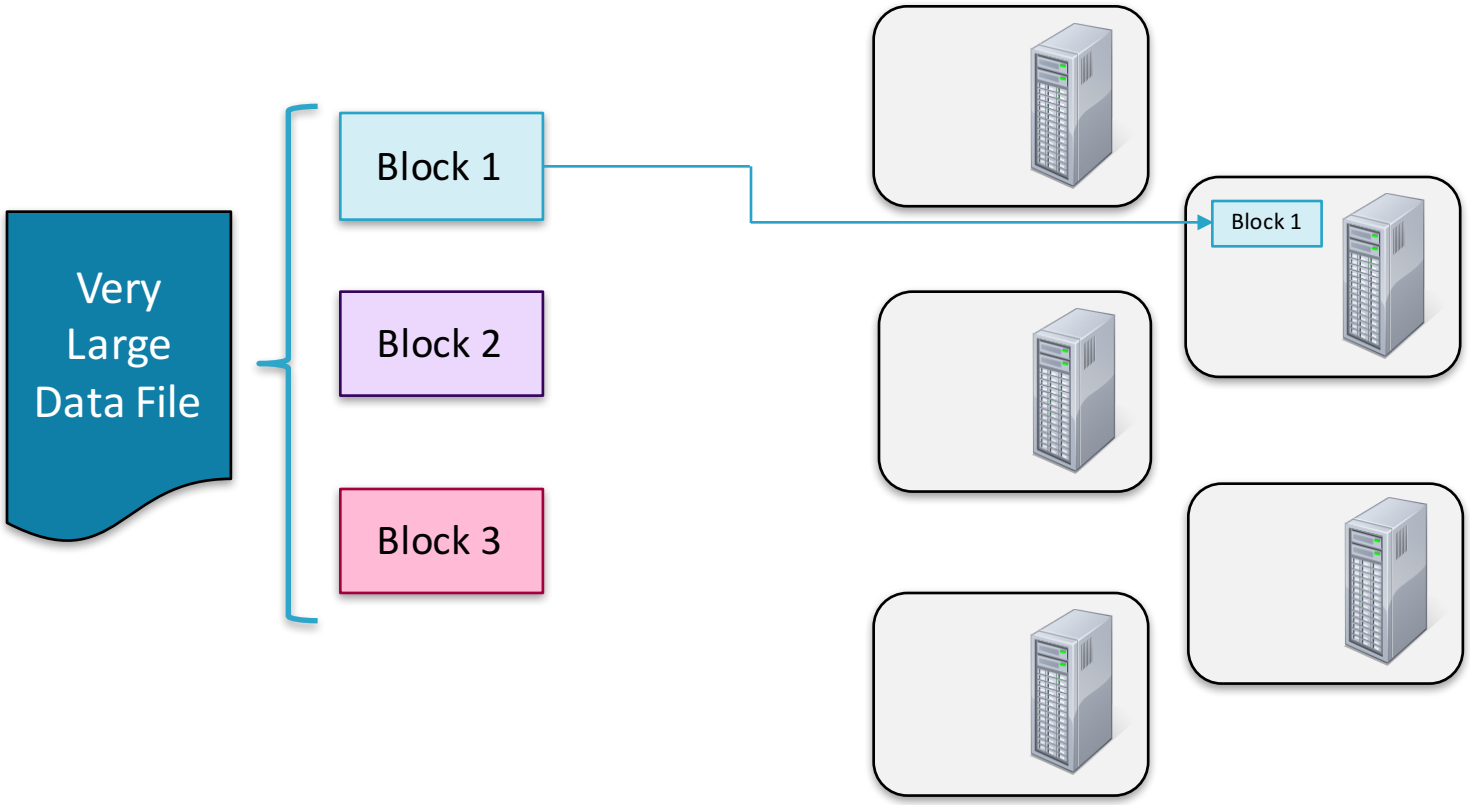
How Files are Stored (1)

- **Data files are split into blocks and distributed to data nodes**



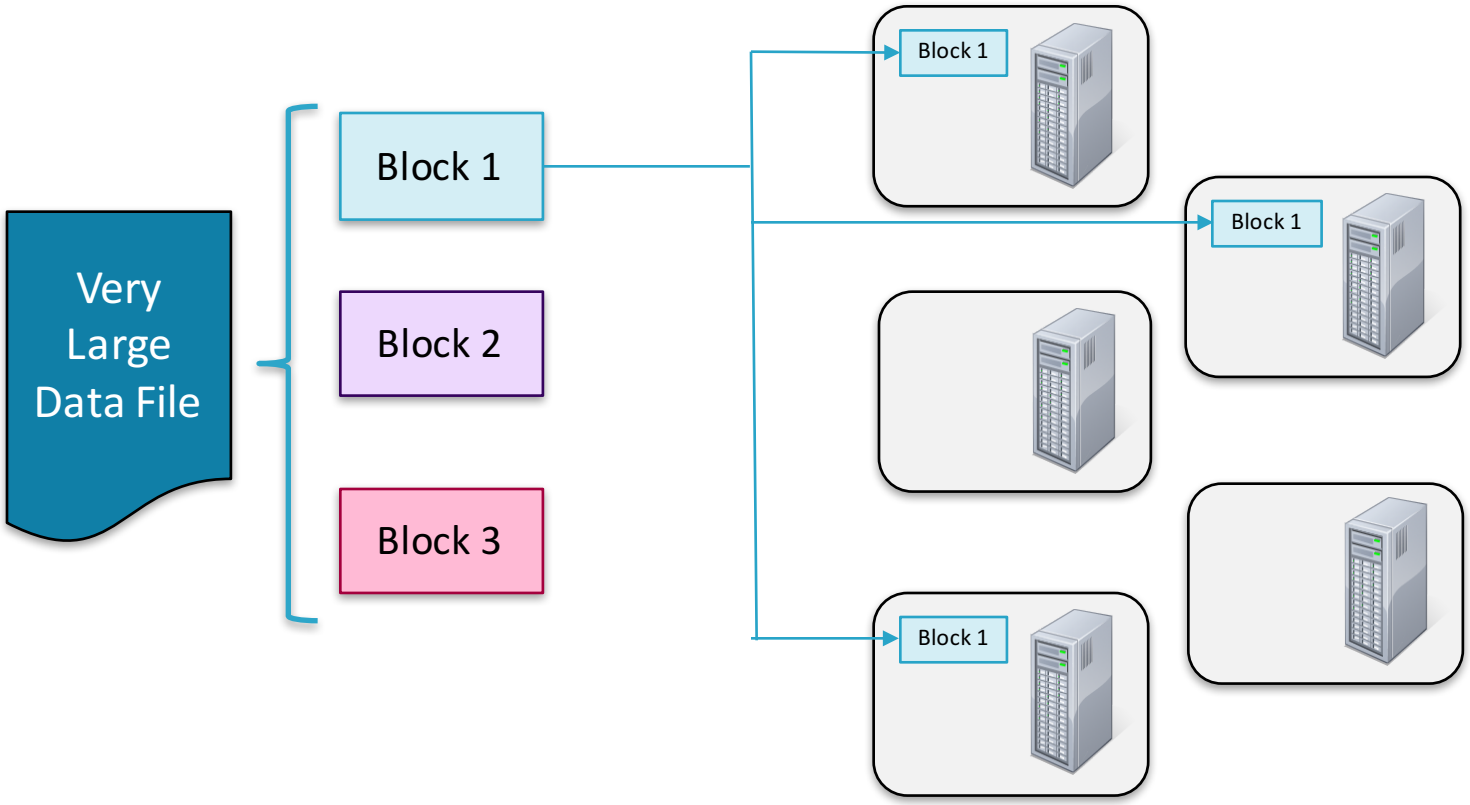
How Files are Stored (2)

- **Data files are split into blocks and distributed to data nodes**



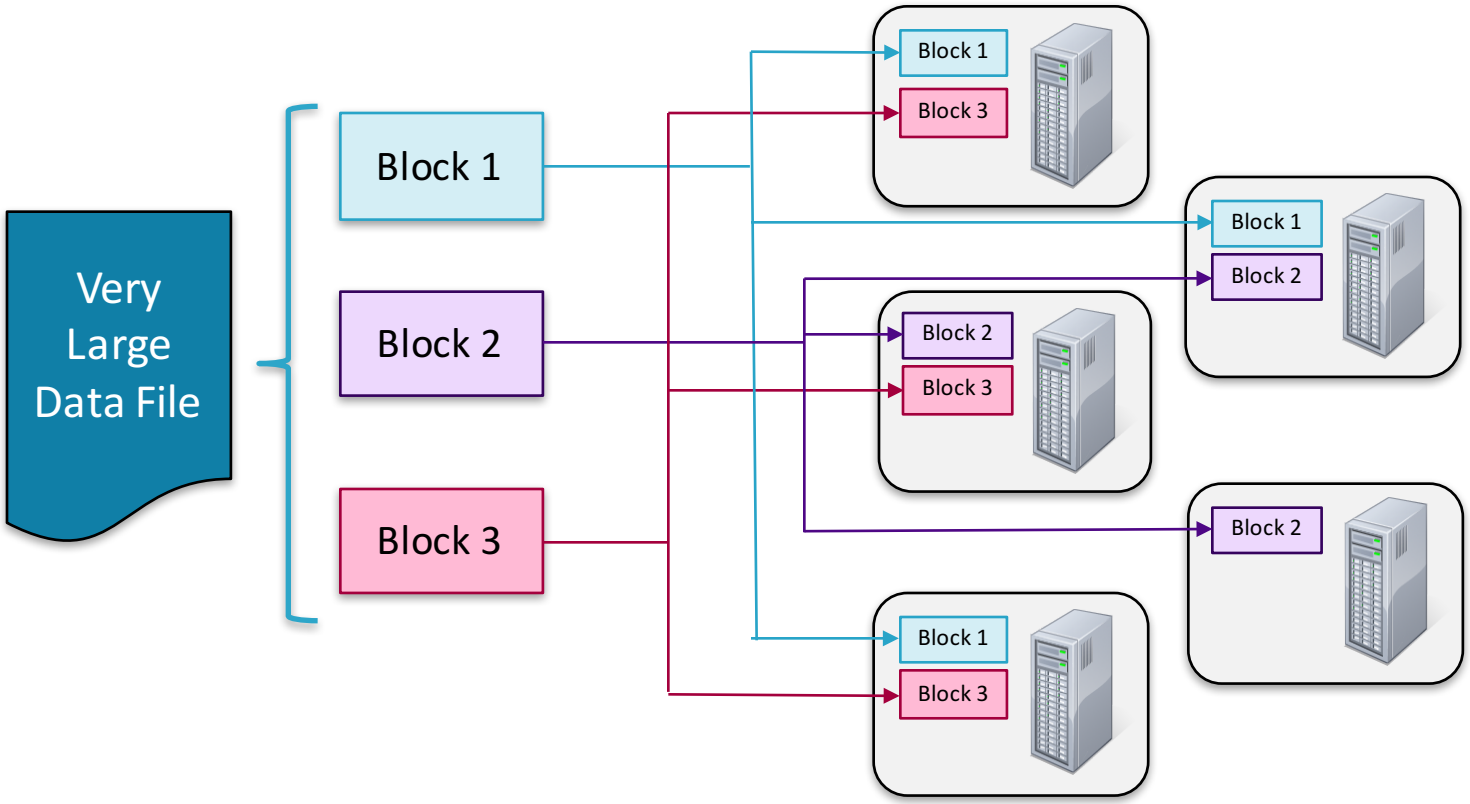
How Files are Stored (3)

- **Data files are split into blocks and distributed to data nodes**
- **Each block is replicated on multiple nodes (default: 3x replication)**



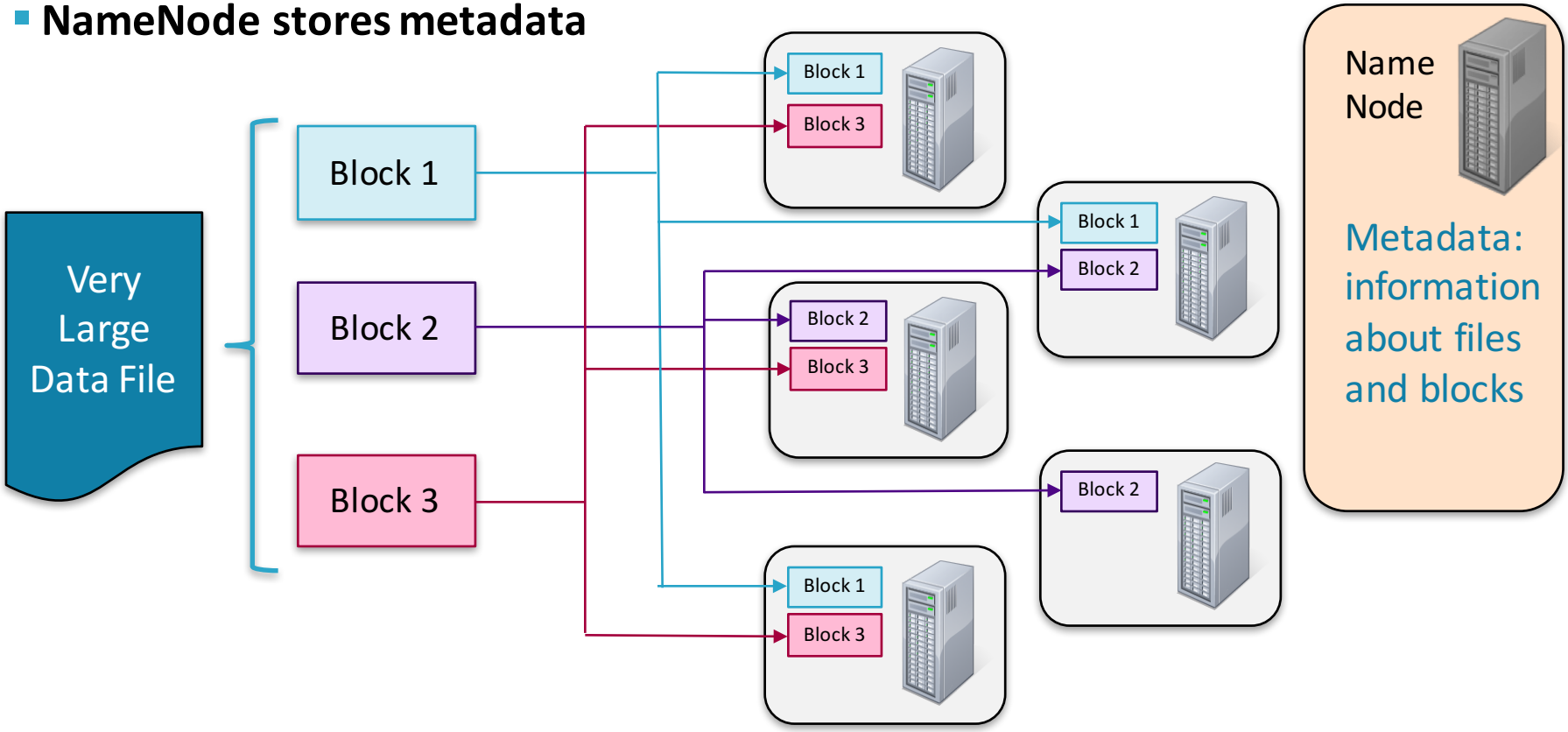
How Files are Stored (4)

- **Data files are split into blocks and distributed to data nodes**
- **Each block is replicated on multiple nodes (default: 3x replication)**



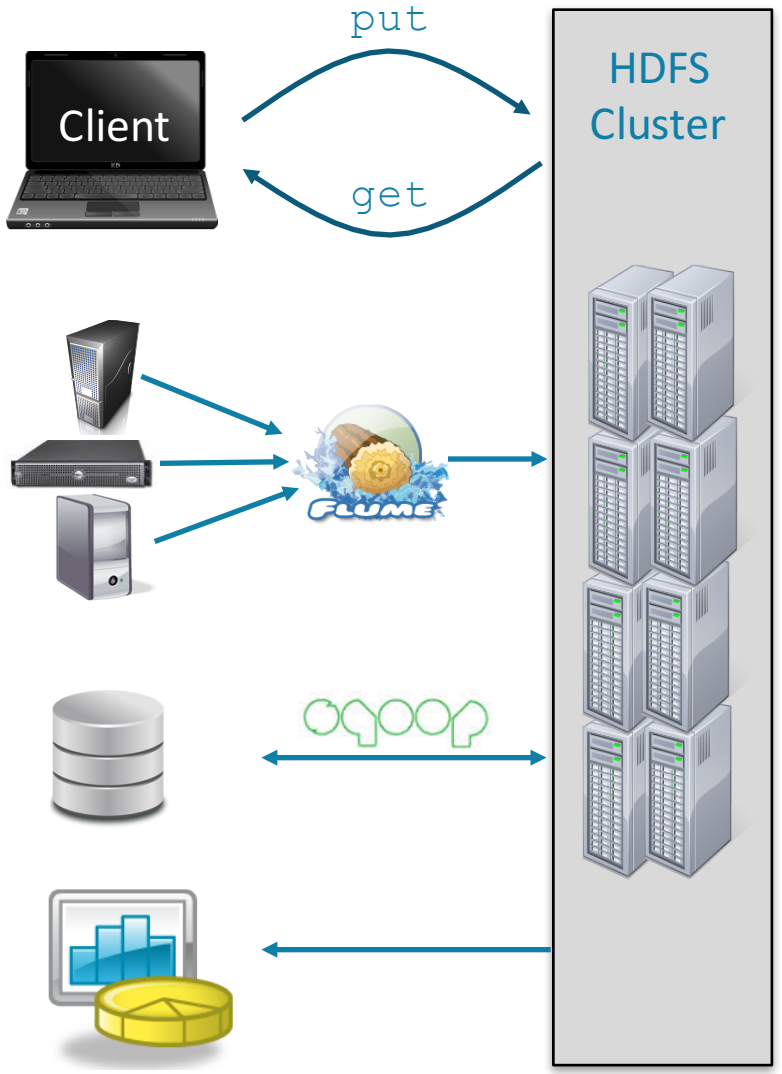
How Files are Stored (5)

- **Data files are split into blocks and distributed to data nodes**
- **Each block is replicated on multiple nodes (default: three-fold replication)**
- **NameNode stores metadata**



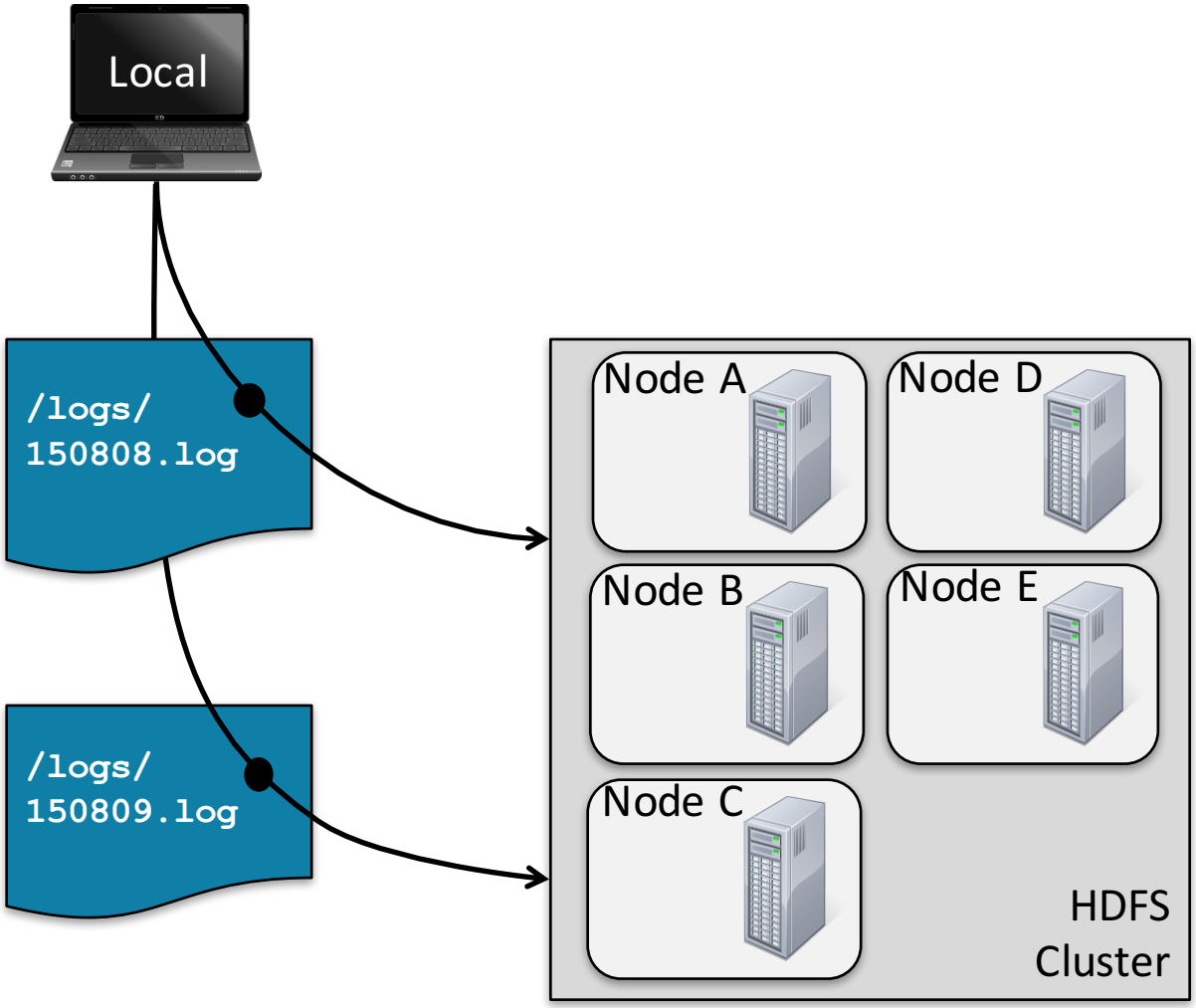
Getting Data In and Out of HDFS

- **Hadoop**
  - Copies data between client (local) and HDFS (cluster)
  - API or command line
- **Ecosystem Projects**
  - Flume
    - Collects data from network sources (such as, websites, system logs)
  - Sqoop
    - Transfers data between HDFS and RDBMSs
- **Business Intelligence Tools**

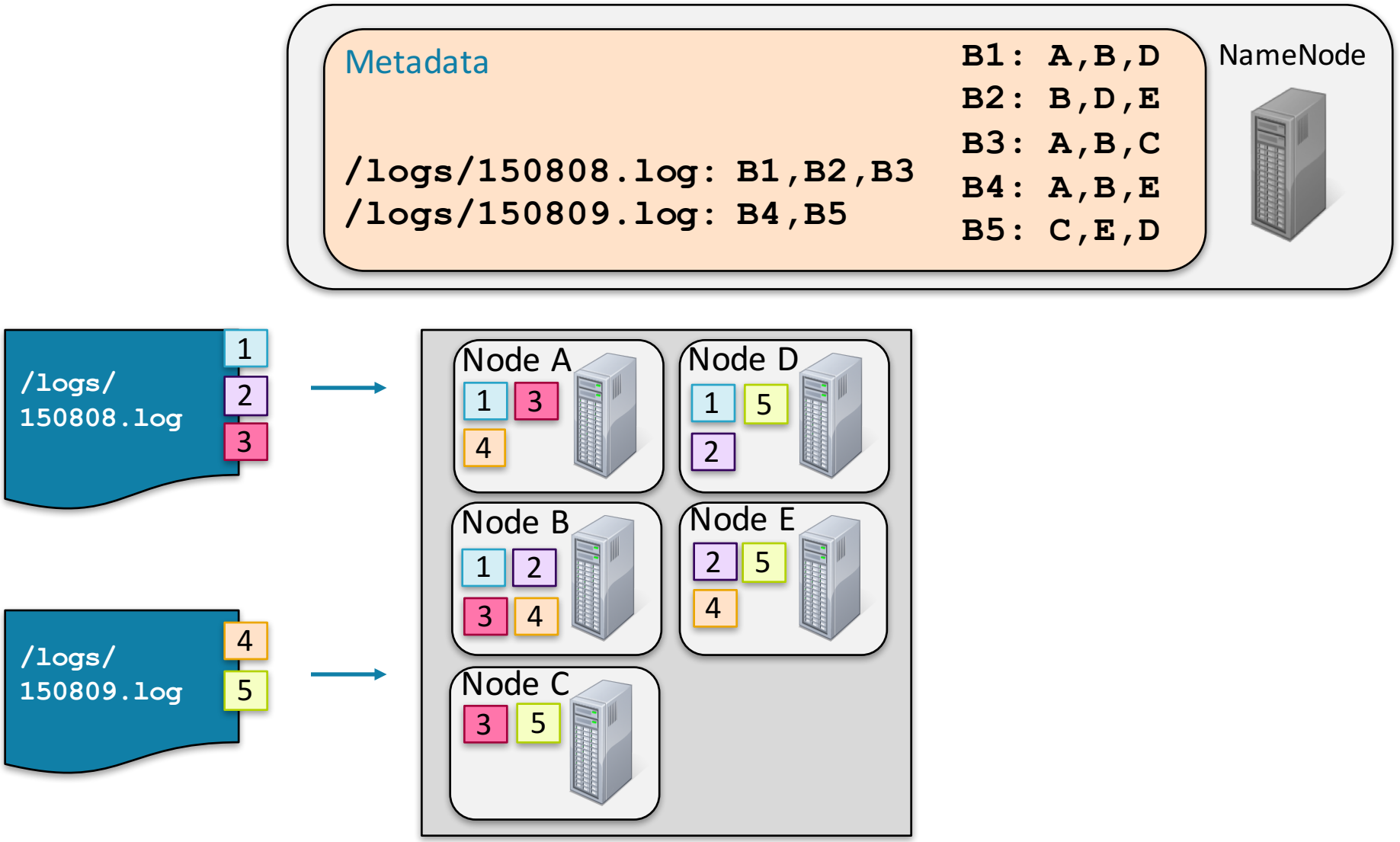


RDBMS: Relational Database Management System

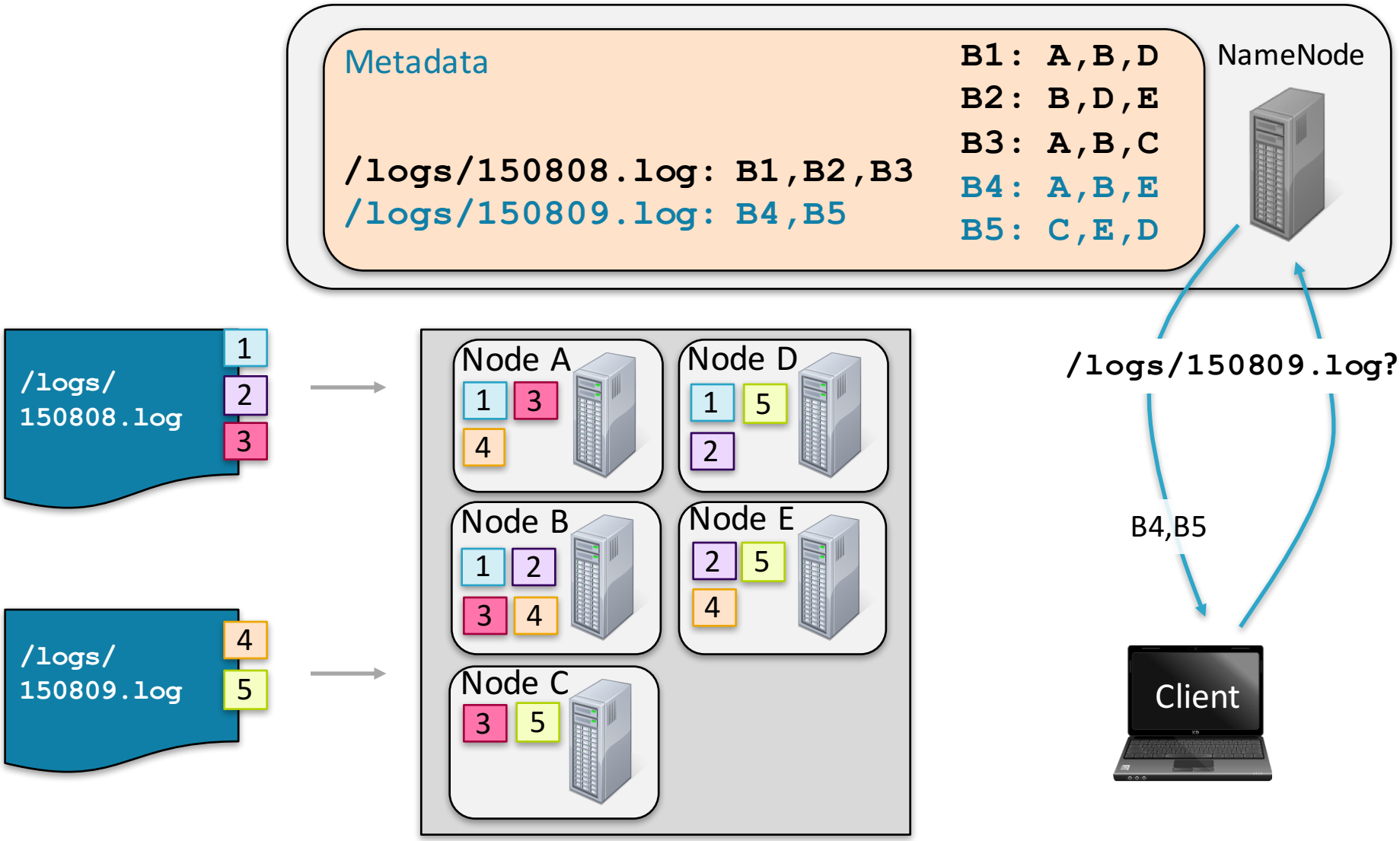
Example: Storing and Retrieving Files (1)



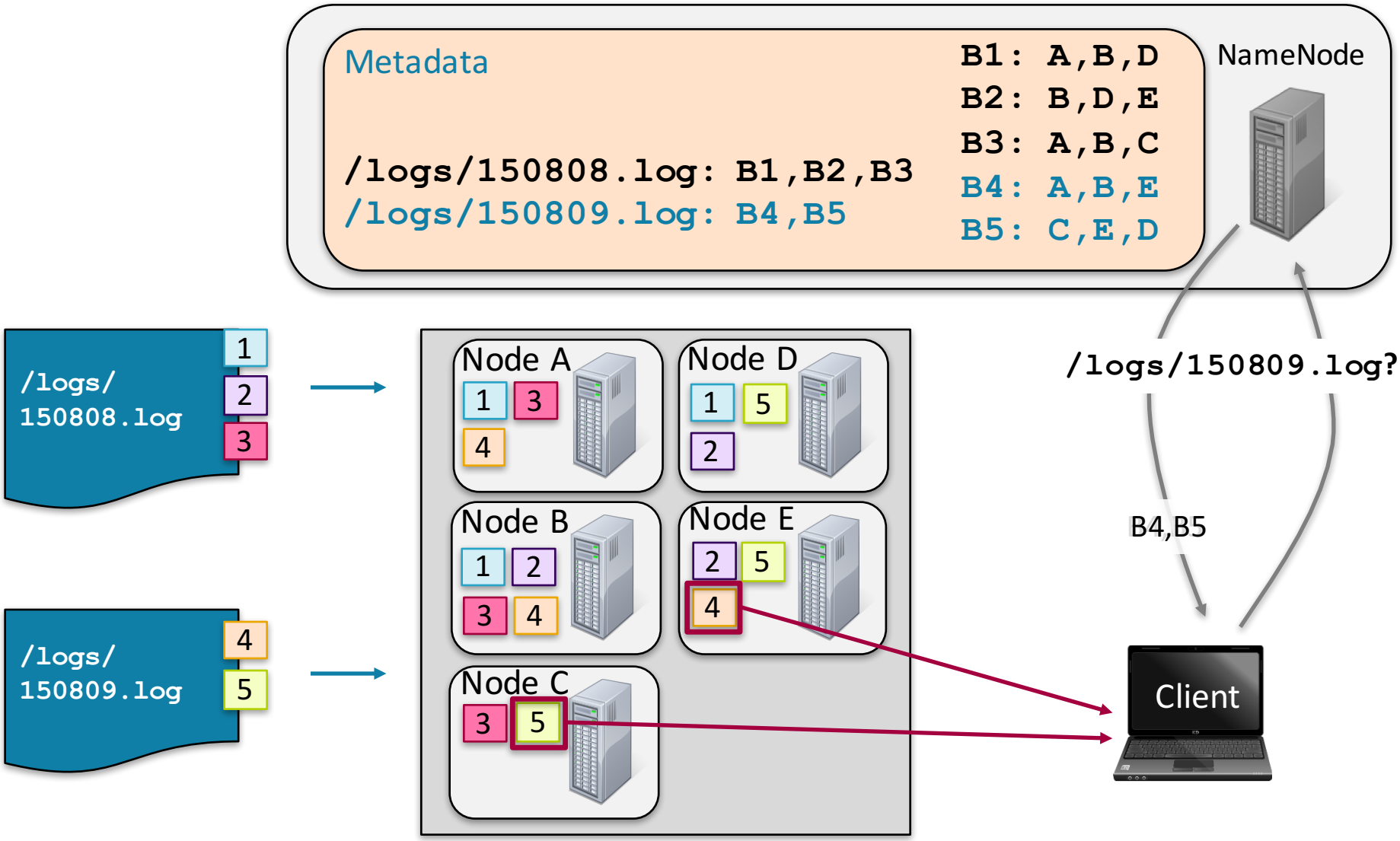
Example: Storing and Retrieving Files (2)



Example: Storing and Retrieving Files (3)



Example: Storing and Retrieving Files (4)



MapReduce: Key Features

- **MapReduce is a programming model**
  - Neither platform- nor language-specific
  - Record-oriented data processing (key and value)
  - Facilitates task distribution across multiple nodes
- **MapReduce was the original processing framework available on Hadoop**
  - Still widely used, although other frameworks are replacing it for many types of workload
- **MapReduce code is typically written in Java**

The Motivation for YARN

- **Originally, Hadoop only supported MapReduce as a processing framework**
  - MapReduce used all of the cluster's processing resources
- **Now, a single cluster may run multiple frameworks, such as MapReduce and Spark**
  - Each framework competes for the nodes' resources
- **YARN (Yet Another Resource Negotiator) helps to manage this contention**
  - Allocates resources to different frameworks based on demand, and on system administrator settings

Chapter Topics

The Hadoop Ecosystem

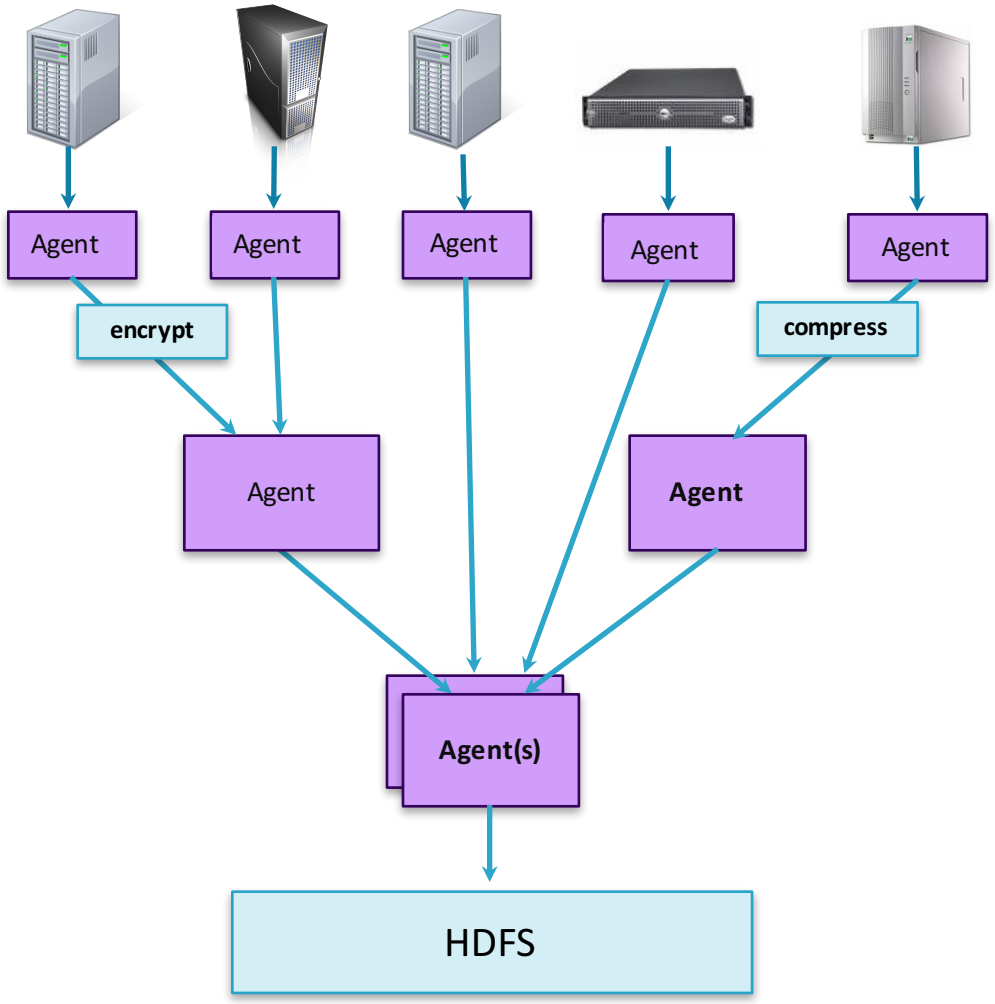
- Introduction
- Core Hadoop: HDFS, MapReduce, and YARN
- **Data Integration: Flume, Kafka, and Sqoop**
- Data Processing: Spark
- Data Analysis: Hive and Impala
- Data Exploration: Cloudera Search
- User Interface: Hue
- Data Storage: HBase
- Data Security: Sentry
- Essential Points

Flume and Kafka: What Are They?

- **Flume and Kafka are tools for ingesting event data into Hadoop as that data is being generated**
  - Log files
  - Sensor data
  - Streaming data from social media, such as Twitter
- **Flume is typically easier to configure, but Kafka provides more functionality**
  - Flume generally provides a path from a data source to HDFS or to a streaming framework such as Spark
  - Kafka uses a "Publish/Subscribe" model
    - Allows data to be consumed by many different systems, including writing to HDFS

Example Flume Pipeline

- Collect data as it is produced
  - Files, syslogs, stdout or custom source
- Process in place
  - Such as encrypt or compress
- Pre-process data before storing
  - Such as transform, scrub or enrich
- Write in parallel
  - Scalable throughput
- Store in any format
  - Text, compressed, binary, or custom sink

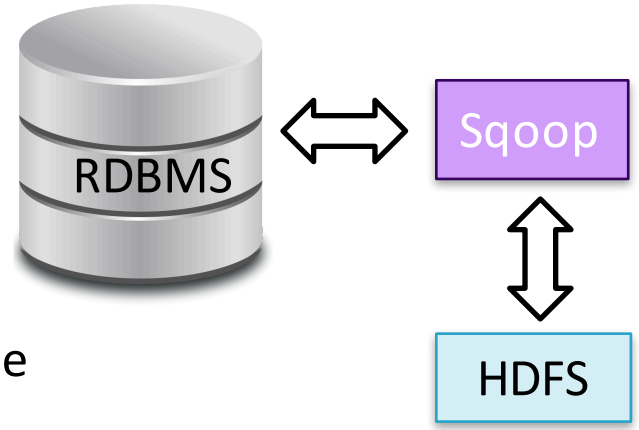


Flume and Kafka: Why Should I Use Them?

- **Flume and Kafka are ideal for aggregating event data from many sources into a centralized location (HDFS)**
- **Well-suited for event driven data**
  - Network traffic
  - Social-media-generated
  - Email messages
  - GPS tracking information
  - Digital sensors
  - Log files
- **Allow you to process streaming data as that data is being generated**
  - Vital for applications such as fraud prevention, threat detection

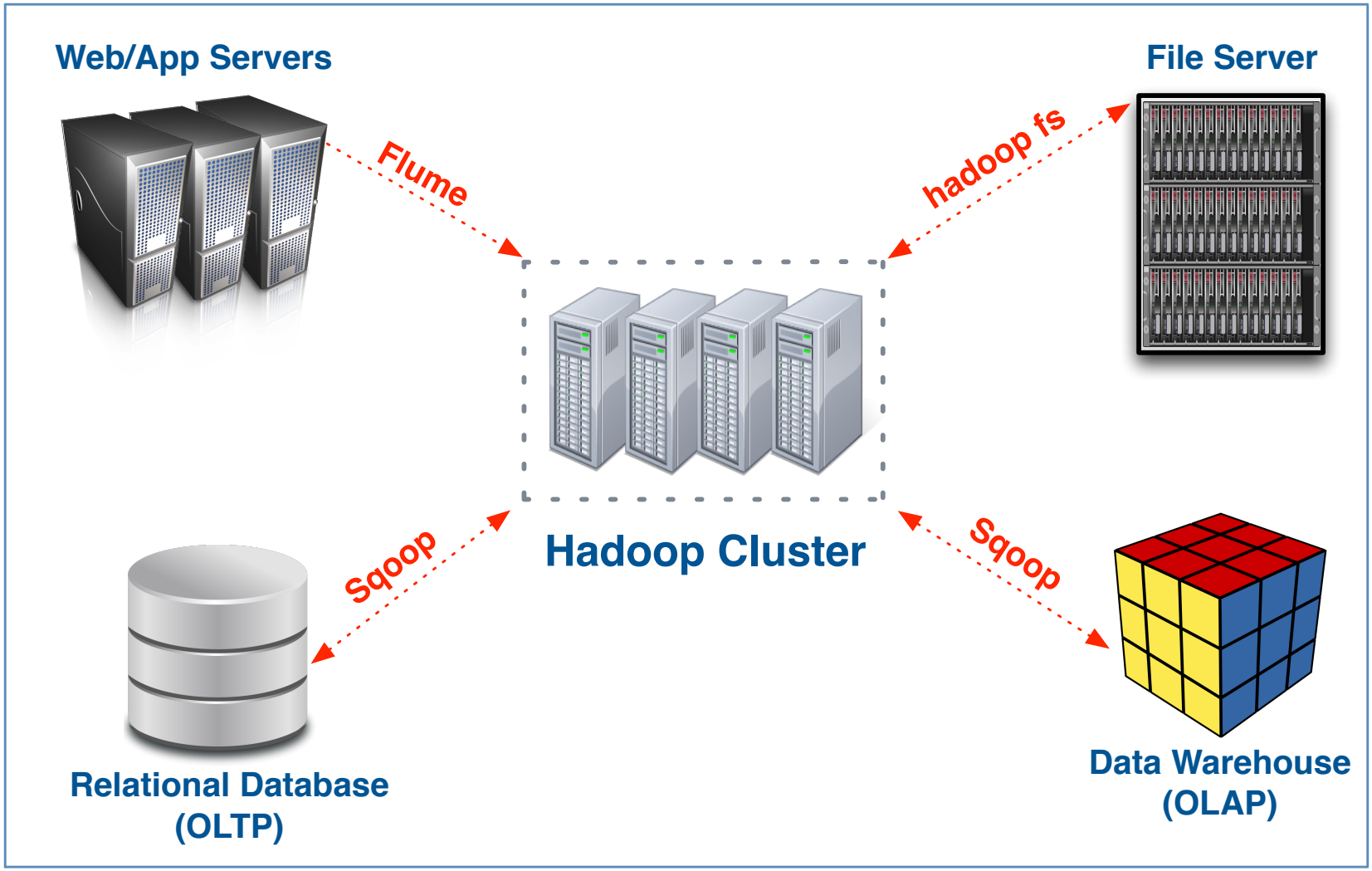
Sqoop: What Is It?

- **Sqoop rapidly moves large amounts of data between relational database management systems (RDBMSs) and HDFS**
  - Import tables (or partial tables) from an RDBMS into HDFS
  - Export data from HDFS to a database table



- **Uses JDBC to connect to the database**
  - Works with virtually all standard RDBMSs
- **Custom "connectors" for some RDBMSs provide much higher throughput**
  - Available for certain databases, such as Teradata and Oracle

Data Center Integration



Chapter Topics

The Hadoop Ecosystem

- Introduction
- Core Hadoop: HDFS, MapReduce, and YARN
- Data Integration: Flume, Kafka, and Sqoop
- **Data Processing: Spark**
- Data Analysis: Hive and Impala
- Data Exploration: Cloudera Search
- User Interface: Hue
- Data Storage: HBase
- Data Security: Sentry
- Essential Points

Spark: What Is It?

- **Apache Spark is large-scale data processing engine**
- **Supports a wide range of workloads**
  - Machine learning
  - Interactive analytics
  - Batch applications
  - Iterative algorithms
  - Business Intelligence
  - And many more
- **Spark Streaming provides the ability to process data as that data is being generated**
  - Typically in conjunction with Flume or Kafka

Spark: Why Should I Use It?

- **Faster than MapReduce**
- **Spark code can be written in Python, Scala, or Java**
  - Easier to develop for than MapReduce

- **Spark is well-suited to iterative processing algorithms such as many of those used in machine learning applications**
- **Spark Streaming provides real-time data processing features**
- **Spark is replacing MapReduce at many organizations**
  - Organizations new to Hadoop will typically start with Spark and never write MapReduce code

Chapter Topics

The Hadoop Ecosystem

- Introduction
- Core Hadoop: HDFS, MapReduce, and YARN
- Data Integration: Flume, Kafka, and Sqoop
- Data Processing: Spark
- **Data Analysis: Hive and Impala**
- Data Exploration: Cloudera Search
- User Interface: Hue
- Data Storage: HBase
- Data Security: Sentry
- Essential Points

Apache Hive: What Is It?

- **Hive is an abstraction layer on top of Hadoop**
  - Hive uses a SQL-like language called HiveQL
- **The Hive interpreter uses MapReduce or Spark to actually process the data**
- **JDBC and ODBC drivers are available**
  - Allows Hive to integrate with BI and other applications


```
SELECT zipcode, SUM(cost) AS total
FROM customers
JOIN orders
ON (customers.cust_id = orders.cust_id)
WHERE zipcode LIKE '63%'
GROUP BY zipcode
ORDER BY total DESC;
```

Hive: Why Should I Use It?

- **Data can be loaded before the table is defined**
  - Schema-on-Read
  - You do not need to know the data's structure prior to loading it

- **Does not require a developer who knows Java, Scala, Python, or other traditional programming languages**
  - Anyone who knows SQL can process and analyze the data on the cluster
- **Well suited for dealing with structured data, or data which can have a structure applied to it**

Comparing Hive to an RDBMS

Feature	RDBMS	Hive
Query language	SQL	SQL
Update and delete records	Yes	Experimental
Transactions	Yes	Experimental
Stored procedures	Yes	No
Index support	Extensive	Limited
Latency	Very low	High
Scalability	Low	Very high
Data format flexibility	Minimal	Very high
Storage cost	Very expensive	Inexpensive

Impala: What Is It?

- **Apache Impala (Incubating) is a high-performance SQL engine**
  - Runs on Hadoop clusters
  - Does not rely on MapReduce
  - Massively-parallel processing (MPP)
  - Inspired by Google's Dremel project
  - Very low latency—typically measured in milliseconds
- **Impala supports a dialect of SQL very similar to Hive's**
- **Impala was developed by Cloudera**
  - 100% open source, released under the Apache software license

Which to Choose: Hive or Impala?

- **Impala and Hive both provide a way to analyze data on the cluster using SQL**
- **Impala is best suited for ad-hoc analytics, and situations where multiple people will be querying the cluster simultaneously**
  - Impala is much faster than Hive
  - Impala deals with multiple simultaneous users much better
- **Hive is well suited for batch processing**
  - ETL
  - Scheduled workloads

Chapter Topics

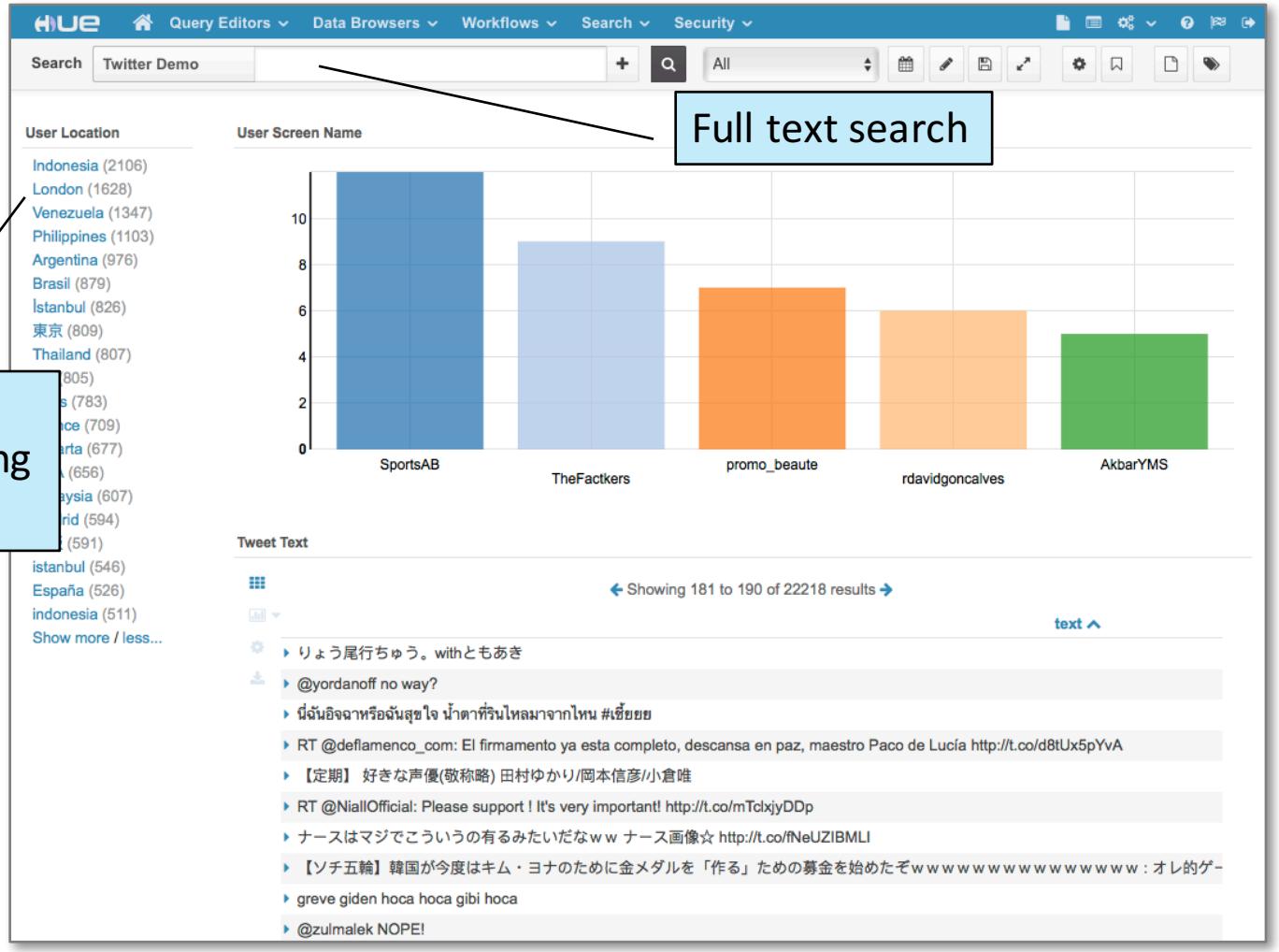
The Hadoop Ecosystem

- Introduction
- Core Hadoop: HDFS, MapReduce, and YARN
- Data Integration: Flume, Kafka, and Sqoop
- Data Processing: Spark
- Data Analysis: Hive and Impala
- **Data Exploration: Cloudera Search**
- User Interface: Hue
- Data Storage: HBase
- Data Security: Sentry
- Essential Points

Cloudera Search: What Is It?

- **Provides interactive full-text search for data in your Hadoop cluster**
- **Supports real-time and batch indexing**
- **Allows non-technical users to access your data**
  - Nearly everyone can use a search engine
- **Cloudera Search enhances Apache Solr**
  - Integrations with HDFS, MapReduce, HBase, Flume, Kafka…
  - Support for file formats widely used with Hadoop
  - Dynamic Web-based dashboard Search interface with Hue
  - Apache Sentry based security
- **Cloudera Search is 100% open source**

Cloudera Search Example: Twitter Feed Search



Iterative search using facets

Cloudera Search: Why Should I Use It?

- **Databases are often used to *analyze data***
  - Search is typically used to *discover data*
- **Databases are designed to join tables based on a key**
  - Search is intended for queries on denormalized (flat) data sets
- **Databases are optimized to find and sort by specific values**
  - Search can match based on specific values, term variants, or ranges
  - Search results are usually sorted by relevance
- **As with a database, Cloudera Search is primarily a backend tool**
  - End users usually interact with it through user interfaces you create
  - APIs are available for application development in multiple languages

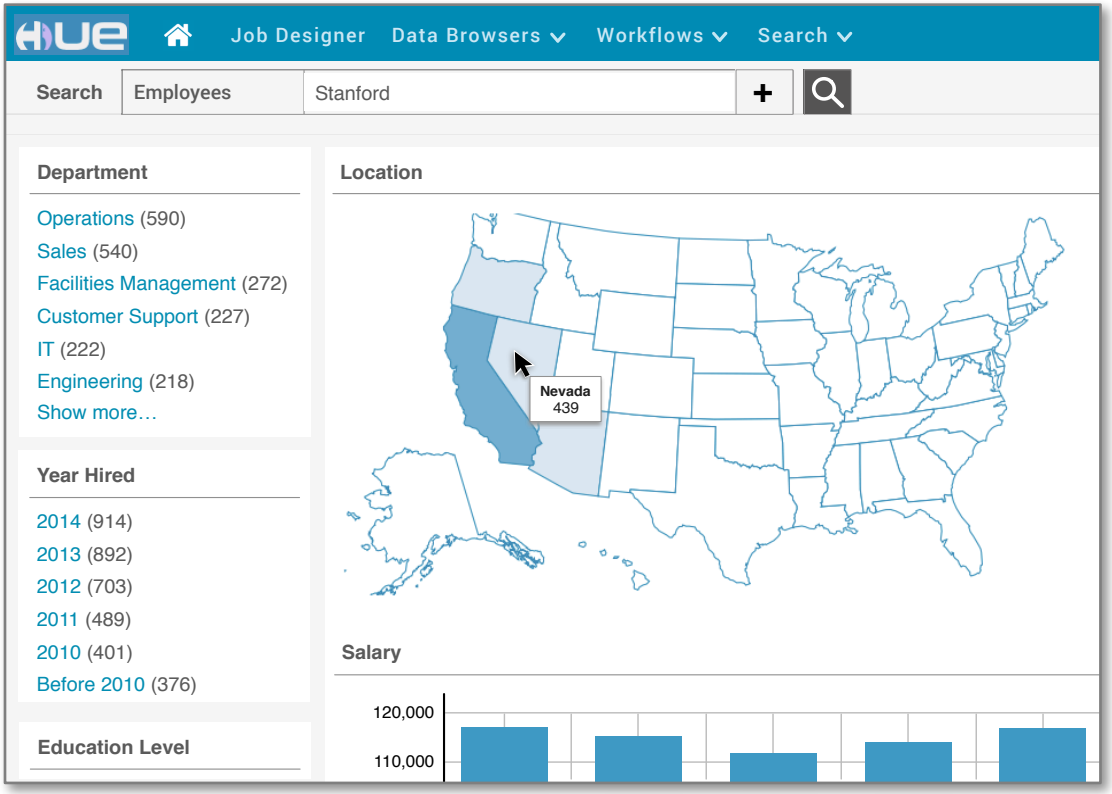
Chapter Topics

The Hadoop Ecosystem

- Introduction
- Core Hadoop: HDFS, MapReduce, and YARN
- Data Integration: Flume, Kafka, and Sqoop
- Data Processing: Spark
- Data Analysis: Hive and Impala
- Data Exploration: Cloudera Search
- **User Interface: Hue**
- Data Storage: HBase
- Data Security: Sentry
- Essential Points

Hue: What Is It?

- **Hue provides a Web front-end to a Hadoop**
  - Upload data
  - Browse data
  - Query tables in Impala and Hive
  - Search
  - And much more
- **Provides access control for the cluster by requiring users to log in before they can use the system**
- **Makes Hadoop easier to use**



Chapter Topics

The Hadoop Ecosystem

- Introduction
- Core Hadoop: HDFS, MapReduce, and YARN
- Data Integration: Flume, Kafka, and Sqoop
- Data Processing: Spark
- Data Analysis: Hive and Impala
- Data Exploration: Cloudera Search
- User Interface: Hue
- **Data Storage: HBase**
- Data Security: Sentry
- Essential Points

HBase: What Is It?

- **HBase is a NoSQL distributed database**
- **Stores data in HDFS**
- **Scales to support very high throughput for both reads and writes**
  - Millions of inserts or updates per second
- **A table can have many thousands of columns**
  - Handles sparse data well
- **Designed to store very large amounts of data (Petabytes+)**

Comparing HBase to a Relational Database

Feature	RDBMS	HBase
Data layout	Row- or column-oriented	Column Family-oriented
Transactions	Yes (ACID)	Single row only
Query language	SQL	get/put/scan
Indexes	Yes	Row-key only *
Max data size	Terabytes	Petabytes
Throughput limits	1000s of queries/second	Millions of queries/second

* Limited support for secondary indexes

HBase: When Should I Used It?

- **Use HBase if…**
  - You need random reads
  - You need random writes
  - You need to do thousands of operations per second on terabytes of data
  - Your access patterns are simple and well-known
- **Don't use HBase if…**
  - You only append to your dataset and typically read the entire table
  - You primarily perform ad-hoc analytics (ill-defined access patterns)
  - Your data easily fits on one large node

Chapter Topics

The Hadoop Ecosystem

- Introduction
- Core Hadoop: HDFS, MapReduce, and YARN
- Data Integration: Flume, Kafka, and Sqoop
- Data Processing: Spark
- Data Analysis: Hive and Impala
- Data Exploration: Cloudera Search
- User Interface: Hue
- Data Storage: HBase
- **Data Security: Sentry**
- Essential Points

Apache Sentry: What Is It?

- **Sentry provides fine-grained access control (authorization) to various Hadoop ecosystem components**
  - Impala
  - Hive
  - Cloudera Search
  - The HDFS command-line shell (upcoming)
- **In conjunction with Kerberos authentication, Sentry authorization provides a complete cluster security solution**

Sentry: Why Should I Use It?

- **Hadoop has long supported Kerberos for authentication**
  - "Prove you are who you say you are"
- **Typically, a production cluster also requires authorization controls**
  - "This person is allowed to do only these things"
  - This is especially true for clusters in regulated environments such as financial services, healthcare, and the like.
- **Sentry allows an administrator to grant fine-grained access rights to individuals**
  - For example, permission to view only certain columns in a given Hive table
- **Sentry is a key component of a secure Hadoop cluster**

Chapter Topics

The Hadoop Ecosystem

- Introduction
- Core Hadoop: HDFS, MapReduce, and YARN
- Data Integration: Flume, Kafka, and Sqoop
- Data Processing: Spark
- Data Analysis: Hive and Impala
- Data Exploration: Cloudera Search
- User Interface: Hue
- Data Storage: HBase
- Data Security: Sentry
- **Essential Points**

Essential Points (1)

- **HDFS is the storage layer of Hadoop**
  - Data replication ensures that you do not lose data when your nodes fail
  - Data locality enhances performance by reducing the need to move data between nodes
- **YARN manages the cluster's processing and memory resources**
  - Ensures that different frameworks, and different applications running on those frameworks, can run simultaneously
- **Data integration**
  - Flume, Kafka
  - Sqoop
- **Data processing**
  - Spark
  - MapReduce

Essential Points (2)

- **Data analysis**
  - Hive
  - Impala
- **Data exploration**
  - Cloudera Search
- **User interface**
  - Hue
- **Data storage**
  - HBase
- **Data security**
  - Sentry

cloudera®

An Introduction to Hadoop Architecture

Chapter 4

Course Chapters

- Introduction
- Hadoop Basics
- The Hadoop Ecosystem
- **An Introduction to Hadoop Architecture**
- Hadoop in the Real World
- Managing Hadoop
- Conclusion

An Introduction to Hadoop Architecture

In this chapter you will learn

- **The differences between master and worker nodes**
- **The hardware requirements for each type of node**
- **How to think about capacity planning for your Hadoop cluster**

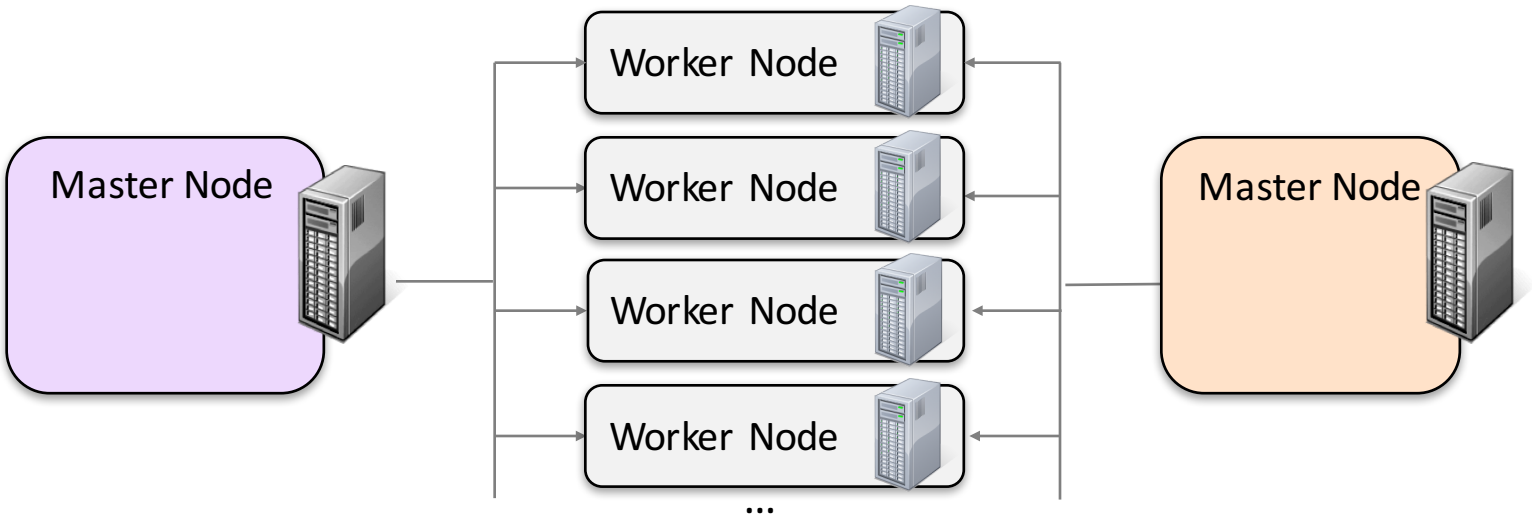
Chapter Topics

An Introduction to Hadoop Architecture

- **Hadoop Cluster Terminology**
- Master Nodes
- Worker Nodes
- Reference Architectures
- Capacity Planning
- Essential Points

Hadoop Cluster Terminology

- **A *cluster* is a group of computers working together**
  - Provides data storage, data processing, and resource management
- **A *node* is an individual computer in the cluster**
  - *Master* nodes manage distribution of work and data to *worker* nodes
- **A *daemon* is a program running on a node**
  - Each performs different functions in the cluster



Hadoop Clusters Contain Master Nodes and Worker Nodes

- **Master nodes manage the work**
  - Master nodes are essential
    - Protect them from failure
  - Daemons running on master nodes ensure that the entire cluster works
  - A failed daemon could cause the entire cluster to become unusable
  - Master nodes are usually configured for high availability (HA)
- **Worker nodes do the work**
  - Worker nodes are expendable (industry standard hardware)
  - Daemons running here handle actual data processing
  - A failed worker node will not bring down the entire cluster

Chapter Topics

An Introduction to Hadoop Architecture

- Hadoop Cluster Terminology
- **Master Nodes**
- Worker Nodes
- Reference Architectures
- Capacity Planning
- Essential Points

Node Failure

- **Master nodes are single points of failure if not configured for high availability (HA)**
  - If the HDFS master node goes down, the cluster is inaccessible
  - If the YARN master node goes down, no new jobs can run on the cluster
- **Hadoop supports HA for all master nodes**
  - Configure for HA when running production workloads
- **Worker nodes are expected to fail**
  - This is an assumption built into Hadoop
- **Master nodes and worker nodes have very different hardware requirements**

Master Node Hardware Recommendations

- **Carrier class hardware**
  - Not industry standard hardware
- **Dual power supplies**
- **Dual Ethernet cards**
  - Bonded to provide failover
- **Hard drives use redundant array of inexpensive disks (RAID) to protect from data loss**
- **Reasonable amount of RAM**
  - 64GB for clusters of 20 nodes or less
  - 96GB for clusters of up to 300 nodes
  - 128GB for larger clusters

Chapter Topics

An Introduction to Hadoop Architecture

- Hadoop Cluster Terminology
- Master Nodes
- **Worker Nodes**
- Reference Architectures
- Capacity Planning
- Essential Points

Worker Node Hardware Recommendations: CPU

- **Hadoop nodes are typically disk- and network I/O-bound**
  - Therefore, top-of-the-range CPUs are not usually necessary
  - You may need greater processing power on your worker nodes if your specific workload requires it
- **Hadoop jobs which may benefit from high-end CPUs**
  - Clustering and classification
  - Complex text mining
  - Natural language processing
  - Feature extraction
  - Image manipulation

Worker Node Hardware Recommendations: RAM

- **New, memory-intensive processing frameworks are being deployed on many Hadoop clusters**
  - Impala
  - Spark
- **Good practice to equip your worker nodes with as much RAM as you can**
  - Memory configurations of 512GB per worker node or even more are not uncommon for workloads with high memory requirements
- **HDFS caching can take advantage of extra RAM on worker nodes**
  - Provides faster access to data

Worker Node Hardware Recommendations: Disk

- **Hadoop's architecture impacts disk space requirements**
  - By default, HDFS data is replicated three times
  - Temporary data storage typically requires 20-30 percent of a cluster's raw disk capacity
- **In general, more spindles (disks) are better**
  - In practice, we see from four to 24 disks (or more) per node
  - For example: 12 x 3TB drives
- **A good practical maximum is 36TB per worker node**
- **7,200 RPM SATA/SATA II drives work well**
  - No need to buy the more expensive 15,000 RPM drives
- **Mechanical hard drives currently provide a significantly better cost/performance ratio than solid-state drives (SSDs)**

Chapter Topics

An Introduction to Hadoop Architecture

- Hadoop Cluster Terminology
- Master Nodes
- Worker Nodes
- **Reference Architectures**
- Capacity Planning
- Essential Points

Cloudera Reference Architectures

- **A Cloudera reference architecture is a document which describes recommended hardware, and best practices, for implementing Hadoop on a vendor's hardware or cloud platformf**
- **You can find reference architectures on the Cloudera Web site for a variety of vendors:**
  - Hardware vendors
  - Virtualization
  - Dedicated appliance
  - Public cloud

Cloud Deployment

- **Hadoop can be deployed to the cloud, using providers such as**
  - Amazon Web Services
  - Microsoft Azure
  - Google Compute Engine
- **Benefits of deploying Hadoop in the cloud include:**
  - Flexible deployment
    - Bypass prolonged infrastructure selection/procurement process
  - Fast ramp-up/ramp-down
    - Easily alter the cluster size based on the current workload
  - Cost savings
    - Deploying in the cloud eliminates the need for dedicated, on-premise computing resources
- **Cloudera Director makes deploying to the cloud easy**

Chapter Topics

An Introduction to Hadoop Architecture

- Hadoop Cluster Terminology
- Master Nodes
- Worker Nodes
- Reference Architectures
- **Capacity Planning**
- Essential Points

Capacity Planning (1)

- **Basing your cluster growth on storage capacity is often a good method to use**
- **Example:**
  - Data grows by approximately 3TB per week/40TB per quarter
  - Hadoop replicates 3 times = 120TB
  - Extra space required for temporary data while running jobs (~30%) = 160TB
  - Assuming machines with 12 x 3TB hard drives
    - 4-5 new machines per quarter
    - Two years of data = 1.3PB requires approximately 36 machines

Capacity Planning (2)

- **New nodes are automatically used by Hadoop**
- **Many clusters start small (less than 10 nodes) and grow as data and processing grows**
- **Hadoop clusters can grow to thousands of nodes**

Chapter Topics

An Introduction to Hadoop Architecture

- Hadoop Cluster Terminology
- Master Nodes
- Worker Nodes
- Reference Architectures
- Capacity Planning
- **Essential Points**

Essential Points (1)

- **Master and worker nodes serve different purposes**
  - Master nodes manage the work and are typically built from carrier-class hardware
  - Worker nodes do the work and typically use lower-cost, industry standard hardware
- **Master nodes are a single point of failure if not configured for high availability**
  - CDH provides HA solutions for all key components
- **Worker nodes are expected to fail**

Essential Points (2)

- **The cloud is a viable platform for running Hadoop**
- **Reference architectures provide hardware and configuration recommendations**
  - These are available on cloudera.com and the hardware vendor's site
- **Typical capacity planning calculations are based on the volume of data to be stored**

cloudera®

Hadoop in the Real World

Chapter 5

Course Chapters

- Introduction
- Hadoop Basics
- The Hadoop Ecosystem
- An Introduction to Hadoop Architecture
- **Hadoop in the Real World**
- Managing Hadoop
- Conclusion

Hadoop in the Real World

In this chapter you will learn

- **What are some common challenges that many organizations face with their existing systems**
- **How organizations are using Hadoop to overcome these challenges**
- **What are the advantages of the Enterprise Data Hub and its unified data storage**

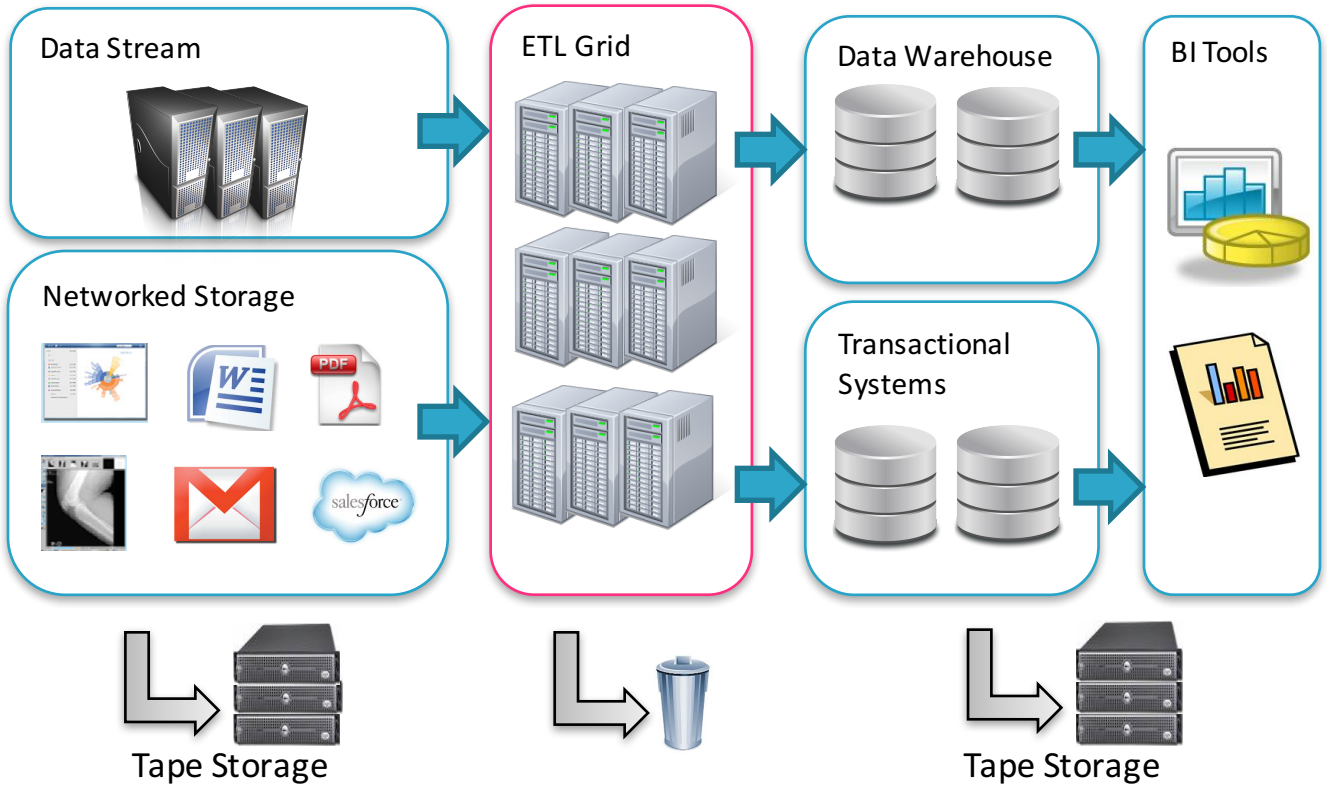
Chapter Topics

Hadoop in the Real World

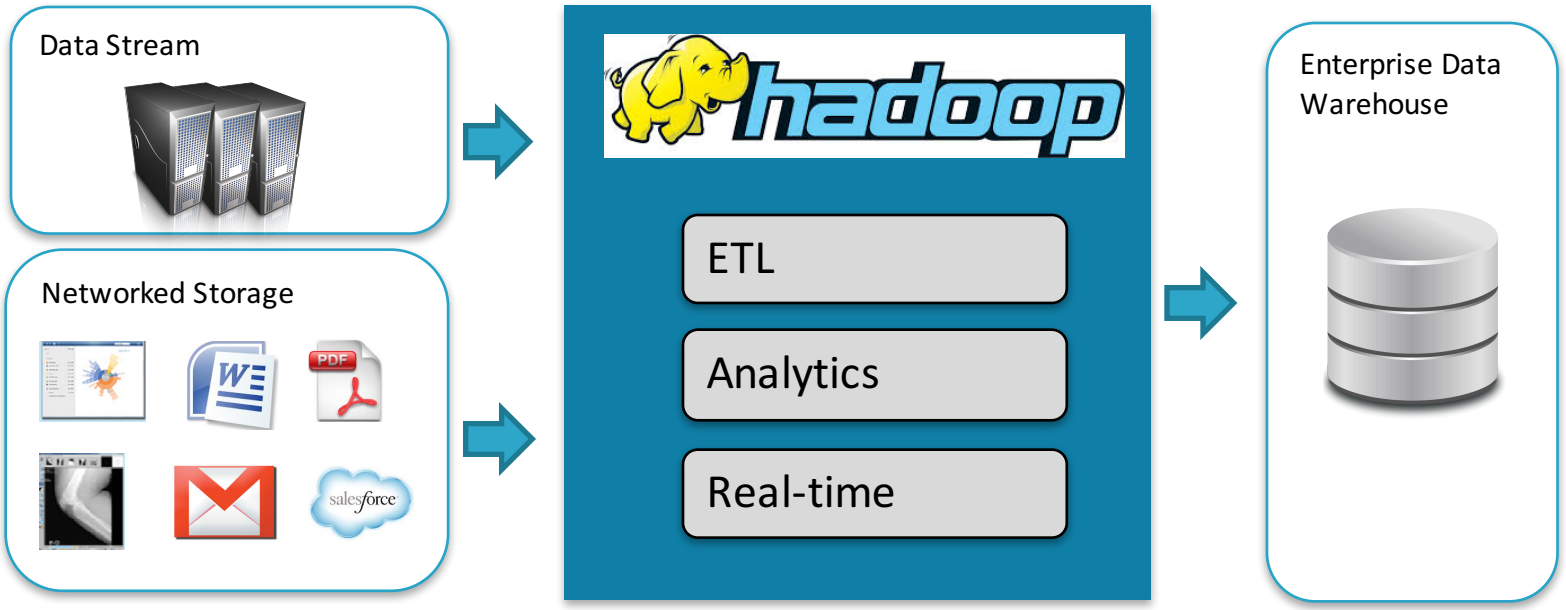
- **ETL Processing**
- Business Intelligence
- Predictive Analytics
- Enterprise Data Hub
- Low-Cost Storage of Large Data Volumes
- Essential Points

Traditional ETL Processing

- **Challenges with ETL (Extract, Transform, and Load) processing:**
  - Too much data, takes too long, and too costly



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**

Vendor Integration—ETL

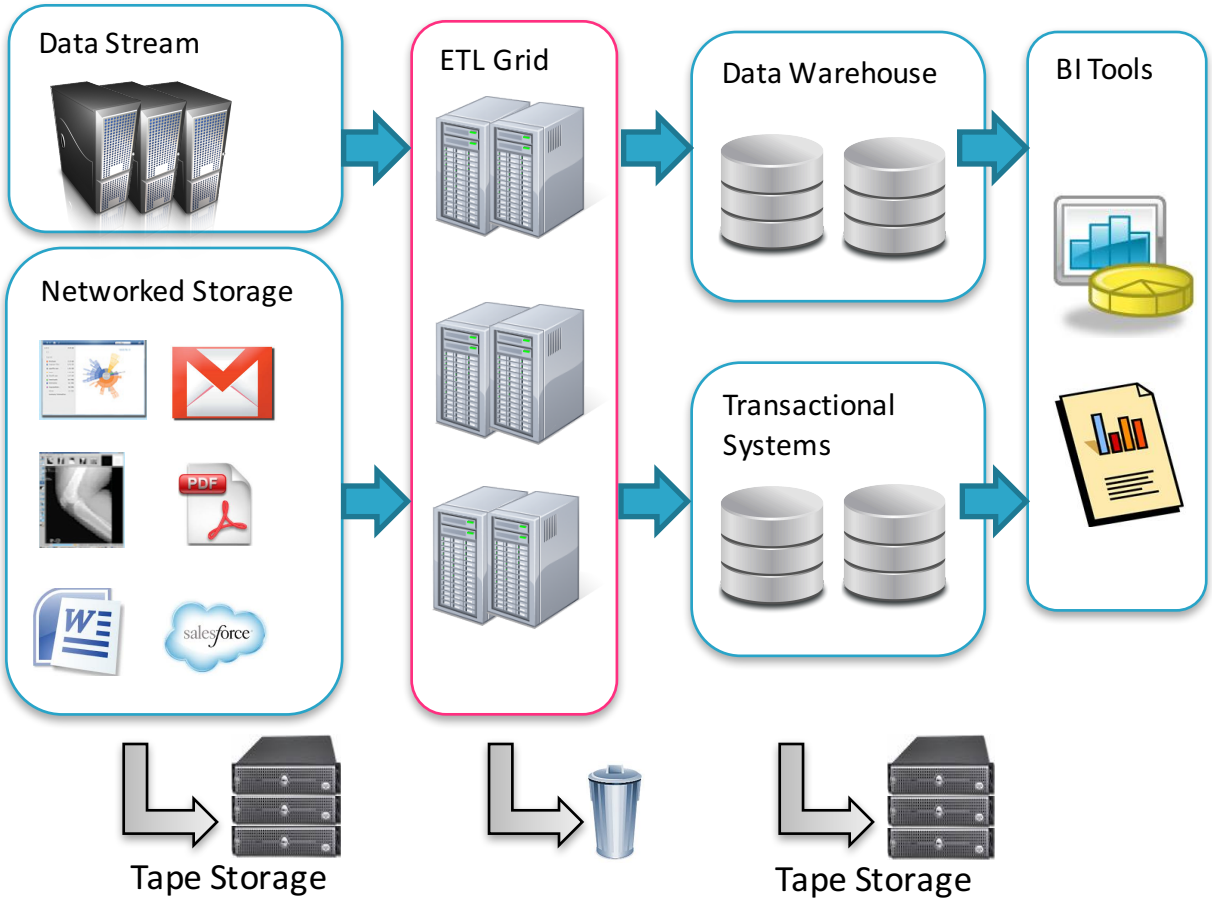
- **For more information visit**
  `http://www.cloudera.com/partners/partners-listing.html`

Chapter Topics

Hadoop in the Real World

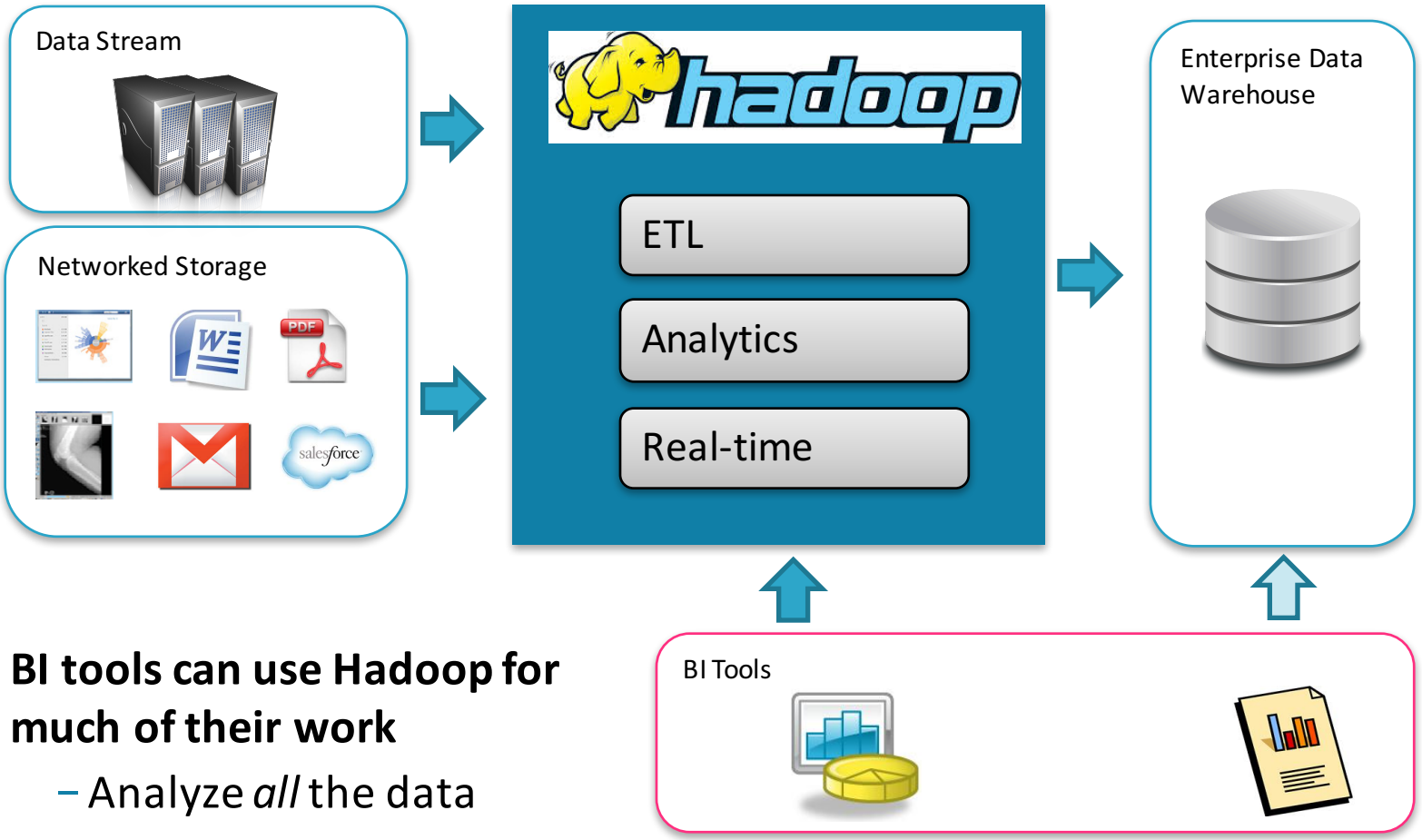
- ETL Processing
- **Business Intelligence**
- Predictive Analytics
- Enterprise Data Hub
- Low-Cost Storage of Large Data Volumes
- Essential Points

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

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**

Vendor Integration—BI

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

360 Degree Customer View: Experian (1)

- **Experian**
  - Leading global information services company
  - Provider of data and analytical services
  - Help their customers
    - Manage credit risk
    - Prevent fraud
    - Marketing automation
- **The problem**
  - Customers are demanding more data for analysis
  - Customers are asking for more frequent updates on consumers'
    - Purchasing behaviors
    - Online browsing patterns
    - Social media activity
  - Addressing this with traditional technology is not cost effective

360 Degree Customer View: Experian (2)

- **Hadoop and HBase are a natural fit for Experian's needs**
  - Rapid processing
  - Large-scale storage
  - Flexible analysis of multi-structured data
- **The impact**
  - They can now process 100 million matches per hour
    - Prior to Hadoop it was 50 million matches per day
  - The cost of their new infrastructure is only a fraction of the legacy environment
- **For more information visit**
  `http://www.cloudera.com/customers/experian.html`

Business Intelligence: Major North American Retailer

- **Legacy system was struggling to process transaction logs in an efficient way**
  - Led to data being discarded
- **Cloudera Enterprise installation now processes transaction logs, streaming data of local price changes made, and other data**
  - Started with a 200TB system, growing at over 50TB per year
- **Dramatically reduced the processing time for common queries**
  - Example: A typical business process required approximately 70 queries on legacy systems and took 16-24 hours; on Cloudera Enterprise it can be performed in a single query, taking 14 minutes
- **Cloudera Enterprise system, including hardware, is 100 times less costly than the legacy systems**

Chapter Topics

Hadoop in the Real World

- ETL Processing
- Business Intelligence
- **Predictive Analytics**
- Enterprise Data Hub
- Low-Cost Storage of Large Data Volumes
- Essential Points

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

Vendor Integration—Predictive Analytics

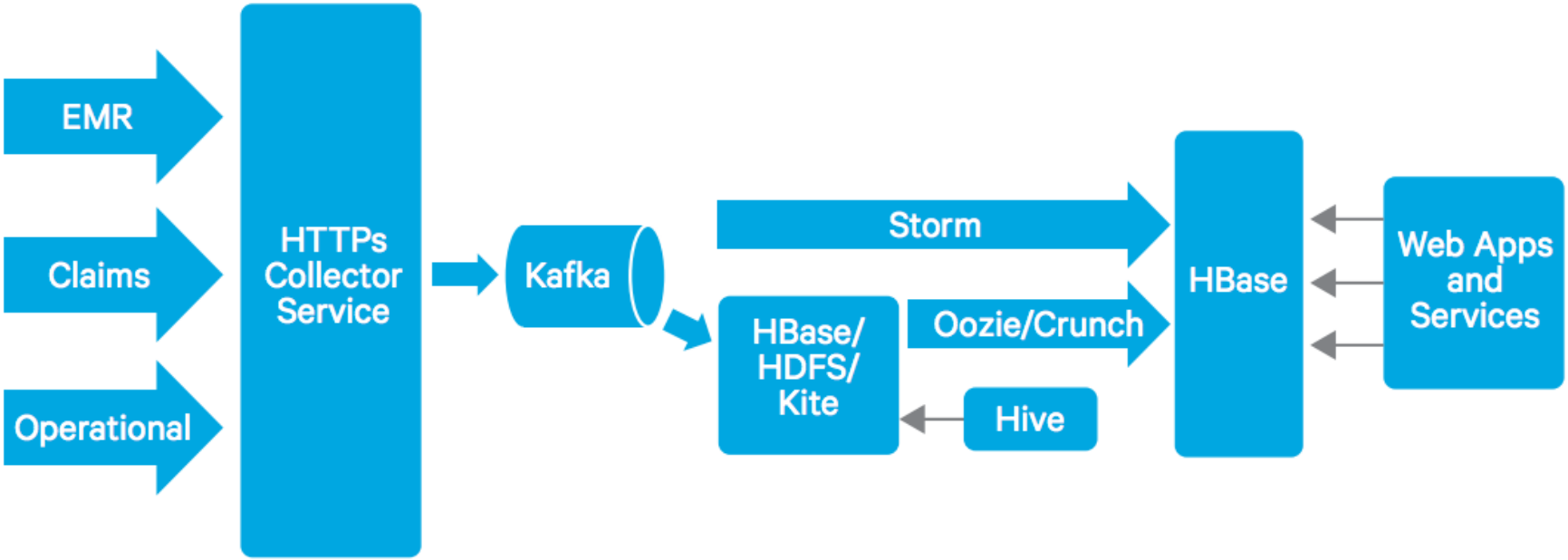
- **For more information visit**
  `http://www.cloudera.com/partners/partners-listing.html`

Predictive Analytics: Cerner Corporation (1)

- **Cerner Corporation**
  - Healthcare IT space
  - Solutions and Services
    - Used by 14,000 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

Predictive Analytics: Cerner Corporation (2)

- **The Hadoop way**
  - Data is streamed and loaded into Hadoop
  - Data is prepared for predictive analytics
  - Some data is loaded to the data warehouse from Hadoop
  - Predictive analytics are performed on the cluster and in the warehouse



Predictive Analytics: Cerner Corporation (3)

- **The impact: Improved insights saves lives**
  - Cerner's EDH brings together an almost unlimited number of sources
  - Builds a more complete picture of any patient, condition, or trend
  - Cerner can accurately predict if a patient has a bloodstream infection
    - This has saved hundreds of lives
- **For more information visit**
  `http://www.cloudera.com/customers/cerner.html`

Predictive Analytics: Large European Auto Insurance Company

- **A large European automotive insurance company uses telematics and predictive analytics to personalize insurance coverage**
- **Installs a "black box" in customers' cars**
  - Contains a GPS device and wireless communications hardware
- **Uses data from the devices to create an individualized prediction of a customer's likelihood to be involved in an accident or other incident**
- **The results:**
  - Significantly reduced fraudulent claims
  - Reduced the number of claims by 30%
  - Attracts and retains low-risk drivers by offering lower premiums

Predictive Analytics: Fraud Detection at FINRA

- **FINRA: Financial Industry Regulatory Authority**
  - Charged with overseeing every brokerage firm and broker doing business in the United States, and monitoring trading on the US stock markets
- **Receives data feeds from stock exchanges and securities firms**
  - 30 billion market events per day
- **Deploys CDH in the cloud to analyze the data**
  - Runs predictive analytics to detect market manipulation and other events
  - Dramatic increase in processing speed
    - A query that took 90 minutes on legacy systems runs in 10 seconds
  - Provides cost savings of $10 million to $20 million per year

Chapter Topics

Hadoop in the Real World

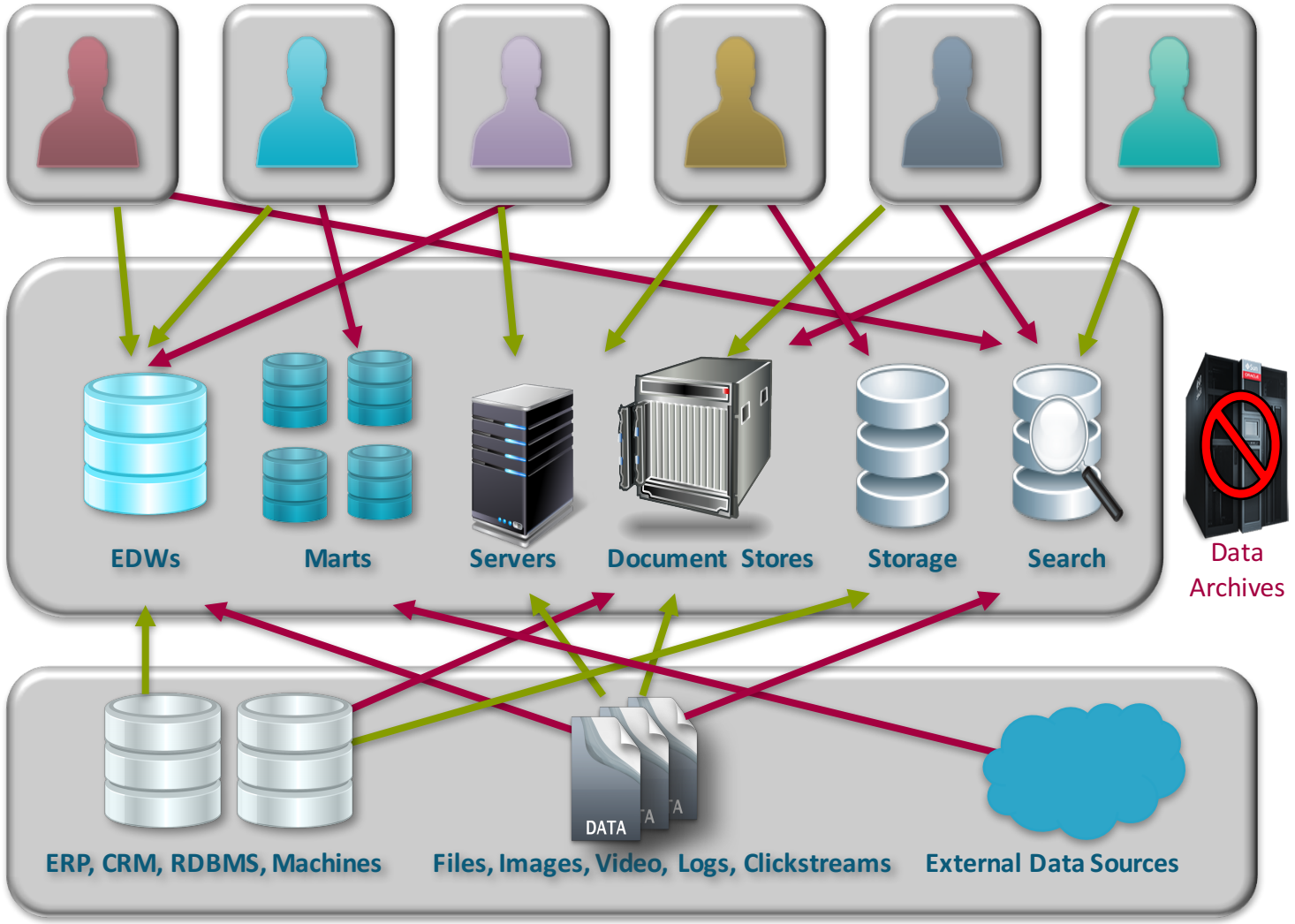
- ETL Processing
- Business Intelligence
- Predictive Analytics
- **Enterprise Data Hub**
- Low-Cost Storage of Large Data Volumes
- Essential Points

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

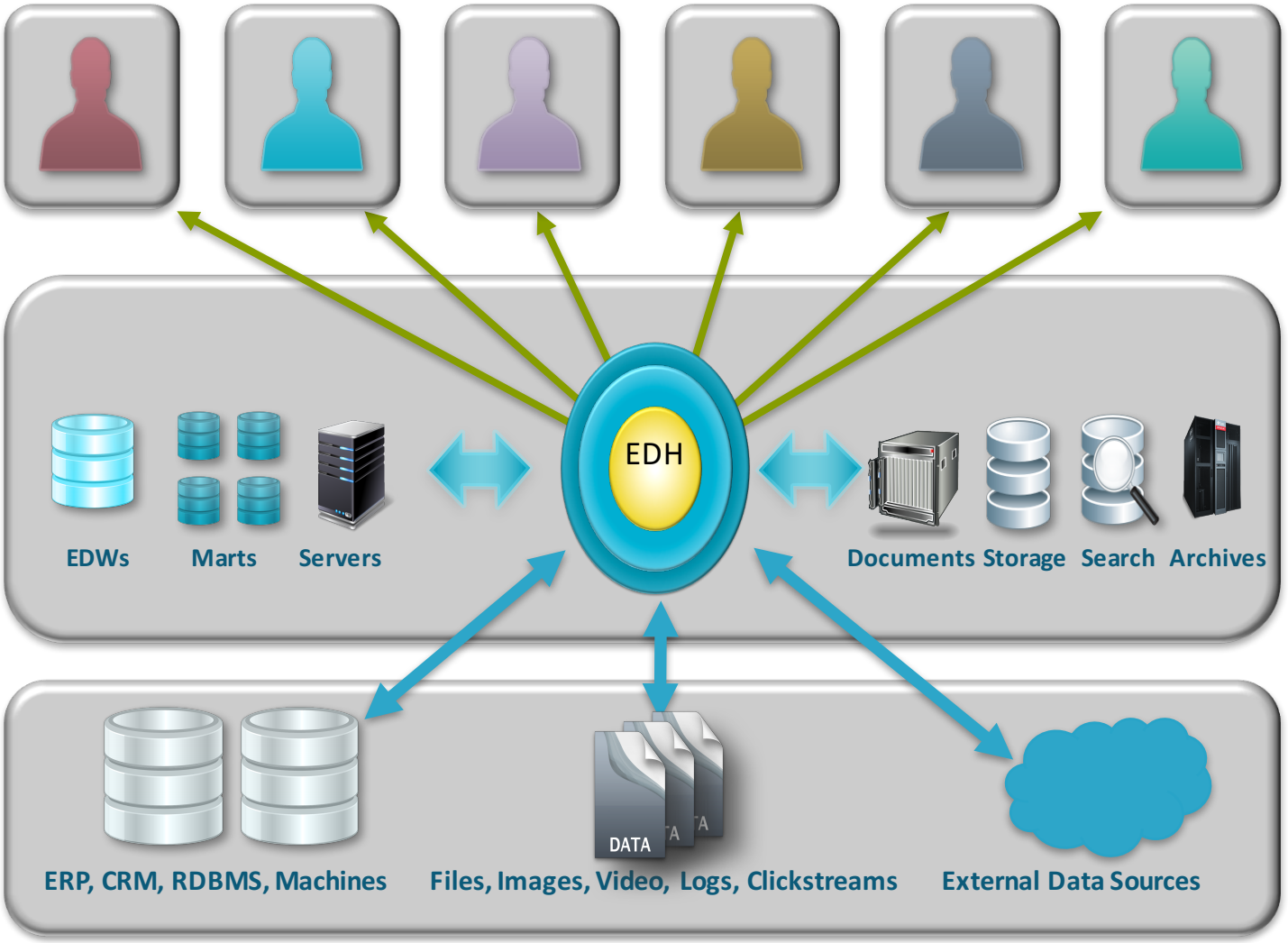


The Enterprise Data Hub: One Unified System

Information and 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



Chapter Topics

Hadoop in the Real World

- ETL Processing
- Business Intelligence
- Predictive Analytics
- Enterprise Data Hub
- **Low-Cost Storage of Large Data Volumes**
- Essential Points

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 your 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

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

Chapter Topics

Hadoop in the Real World

- ETL Processing
- Business Intelligence
- Predictive Analytics
- Enterprise Data Hub
- Low-Cost Storage of Large Data Volumes
- **Essential Points**

Essential Points

- **Hadoop is a flexible platform that can be used to solve real world problems**
  - Combining a storage platform with a powerful processing layer provides great flexibility
- **Common use cases include**
  - ETL and ELT processing
  - Business Intelligence (BI)
  - Predictive analytics
  - Enterprise Data Hub
  - Low-cost storage of large amounts of data

cloudera®

Managing Hadoop

Chapter 6

Course Chapters

- Introduction
- Hadoop Basics
- The Hadoop Ecosystem
- An Introduction to Hadoop Architecture
- Hadoop in the Real World
- **Managing Hadoop**
- Conclusion

Managing Hadoop

In this chapter you will learn

- **What tools are available to help you manage your Hadoop cluster**
- **What issues to think about regarding the security of your Hadoop cluster**
- **Skills and training requirements for a Hadoop team**

Chapter Topics

Managing Hadoop

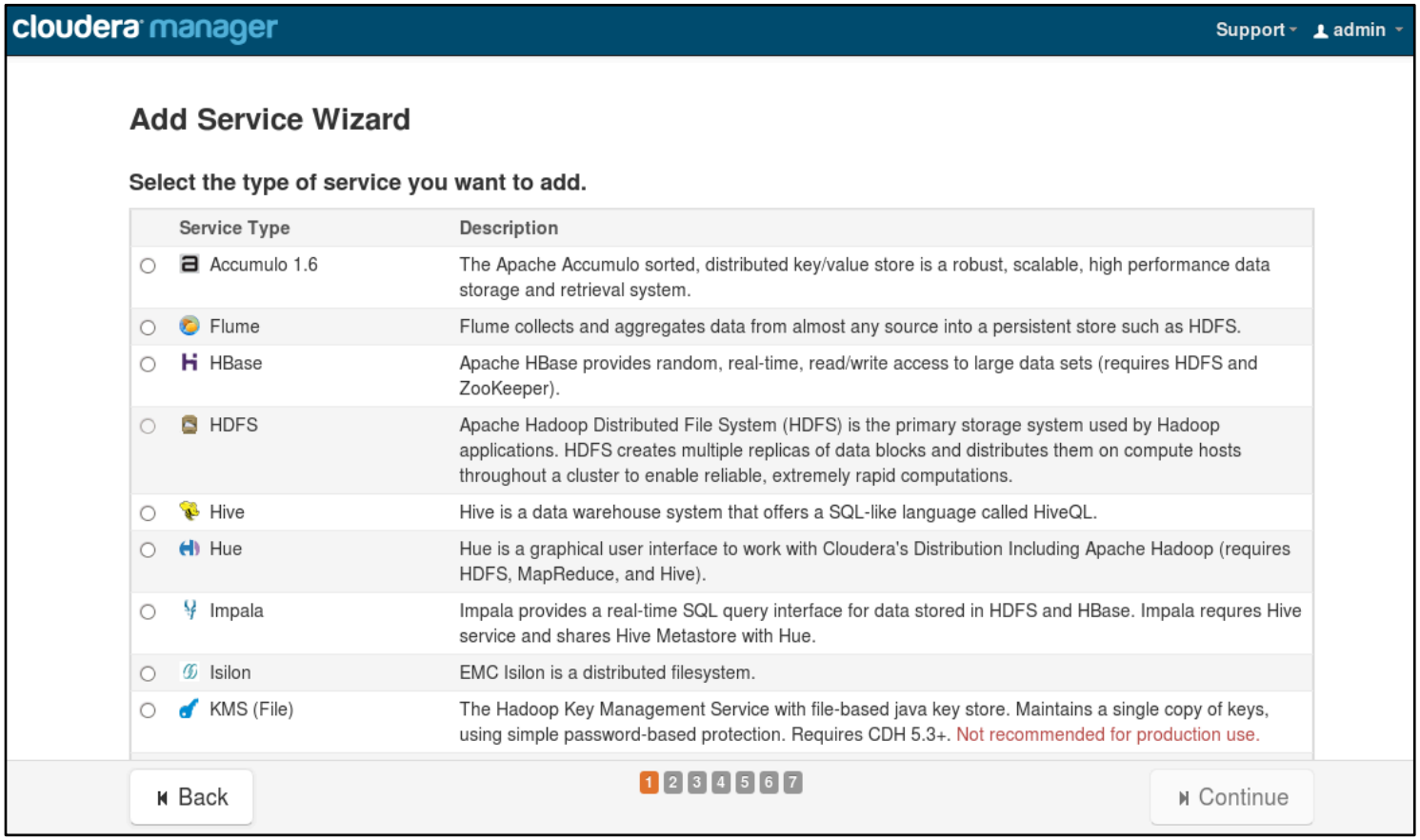
- **Cloudera Manager**
- Cloudera Director
- Cloudera Navigator
- Hadoop Security
- Who Does the Work?
- Essential Points

Motivation for Cloudera Manager

- **Apache Hadoop is a large, complex system**
  - Installing, configuring, monitoring, managing, upgrading, and troubleshooting a Hadoop cluster (or multiple clusters) is non-trivial
- **A manual management approach is error-prone and does not scale or offer the benefits of a fully developed management application**
- **A management tool is essential for any cluster of reasonable size**
- **Cloudera Manager is the preeminent Hadoop management tool**
  - Free to download
  - Free to use
  - No commitment to Cloudera

What Is Cloudera Manager?

- **An application designed to enable administrators to more easily manage an Hadoop cluster**



Cloudera Manager Features

- **Automated deployment**
  - Automatically install and configure Hadoop services on hosts
    - Cloudera Manager sets recommended default parameters
  - Easy to stop and start services on master and worker nodes
- **Manage a wide range of Hadoop and Hadoop ecosystem services**
  - Including HDFS, YARN, MapReduce, Spark, Hive, Pig, Impala, Flume, Oozie, Sqoop, ZooKeeper, Hue, HBase, Cloudera Search, and more
- **Diagnose and resolve issues more quickly**
  - Daemon logging is aggregated and searchable across the cluster
- **Manage user and group access to the cluster(s)**
- **Monitor cluster health and performance, track events**
  - Real-time monitoring, charts, custom reporting, email alerts

Cloudera Manager: High Availability and Disaster Recovery

- **Cloudera Manager makes it easy to enable High Availability for all key Hadoop components**
  - Particularly HDFS and YARN
  - Ensures that the cluster can remain functional even if a master node fails
- **Cloudera Manager also includes Backup and Disaster Recovery features**
  - Automates and schedules copying data to a separate cluster in case of major failure
    - Including metadata for Hive and Impala
  - Provides monitoring and alerting during the process

Cloudera Manager—Two Editions

- **Cloudera Express**
  - Free download
  - Manage a cluster of any size
  - Easy upgrade to Cloudera Enterprise
- **Cloudera Enterprise**
  - A 60-day trial is free
  - Includes support
  - Includes extra features and functionality for Enterprise Data Hubs
    - Rolling upgrades
    - SNMP support
    - LDAP integration
    - Configuration history and rollbacks
    - Operational reports
    - Automated disaster recovery

	Cloudera Express	Cloudera Enterprise
Subscription	Free	Annual
Deployment & Configuration	•	•
Management	•	•
Monitoring	•	•
Diagnostic Tools	•	•
Advanced Mgt Features		•
Support		•

Chapter Topics

Managing Hadoop

- Cloudera Manager
- **Cloudera Director**
- Cloudera Navigator
- Hadoop Security
- Who Does the Work?
- Essential Points

Motivation for Cloudera Director

- **Many organizations are turning to the cloud for their computing infrastructure**
- **Deploying Hadoop to the cloud can be challenging for many reasons**
  - Manual provisioning and management of cloud-based instances does not scale well
  - Determining and adjusting cluster scale elastically without interruption is challenging

What is Cloudera Director?

- **Cloudera Director is a self-service deployment tool for cloud-based CDH clusters**
  - It works as the interface to your cloud provider
- **Cloudera Director enables administrators to**
  - Easily deploy CDH
  - Elastically scale clusters
  - Terminate clusters
- **Once the cluster is deployed, it is managed using Cloudera Manager**

Cloudera Director Features

- **Simplified cluster life cycle management**
  - Self-service spin-up and tear down of clusters
  - Easy to scale the cluster up or down for spiky workloads
  - Simple cloning of clusters
  - Cloud blueprints for repeatable deployments
  - Support for custom, workload-specific deployments
- **Monitoring and metering tools**
  - Multi-cluster health dashboard
  - Instance tracking for account billing

Chapter Topics

Managing Hadoop

- Cloudera Manager
- Cloudera Director
- **Cloudera Navigator**
- Hadoop Security
- Who Does the Work?
- Essential Points

Motivation for Cloudera Navigator

- **Cloudera Navigator is designed for enterprises that must actively manage cluster data and monitor access to the data**
  - Particularly true for regulated industries such as banking, healthcare and so on
- **Cloudera developed Navigator for compliance officers, administrators, and DBAs who generate, use, and oversee vast amounts of cluster data**
- **Navigator provides a single lens for monitoring**
  - All managed cluster data
  - Who is using the data
  - How the data is being used

What Is Cloudera Navigator?

- **Cloudera Navigator is a fully integrated data management tool for the Hadoop platform**
- **Cloudera Navigator enables authorized cluster users to**
  - Audit data access
  - Tag data entities with business metadata that can later be searched
  - Visualize the lineage of data entities

Cloudera Navigator Features

- **Comprehensive, unified auditing across Hadoop**
  - Maintain a full audit history and track access for HDFS, Impala, Hive, HBase, and Sentry
  - Generates reports that meet regulatory requirements
  - Export audit information to global Security Information and Event Management (SIEM) systems
- **Searchable technical and business metadata**
  - Consolidate metadata for Hadoop files and tables
  - Easily track, classify, and locate data to comply with business governance and compliance rules
- **Collect, view, and share lineage**
  - View upstream and downstream column level lineage
  - Quickly identify the origin of a data set and its impact on downstream analysis

Navigator Encrypt and Key Trustee

- **Navigator Encrypt provides high-performance encryption of data on disk**
  - Can be used to encrypt data such as the Hive Metastore, log directories on local disks, and so on
  - Includes process-based access controls
    - Authorized Hadoop processes can access and the data, while preventing admins or superusers from accessing data they don't need to see
- **Navigator Key Trustee is an enterprise-grade key management system**
  - Centralized management of SSL certificates, SSH keys, tokens, passwords, kerberos keytabs etc.
  - Integrates with common HSMs (Hardware Security Modules) from third parties

Chapter Topics

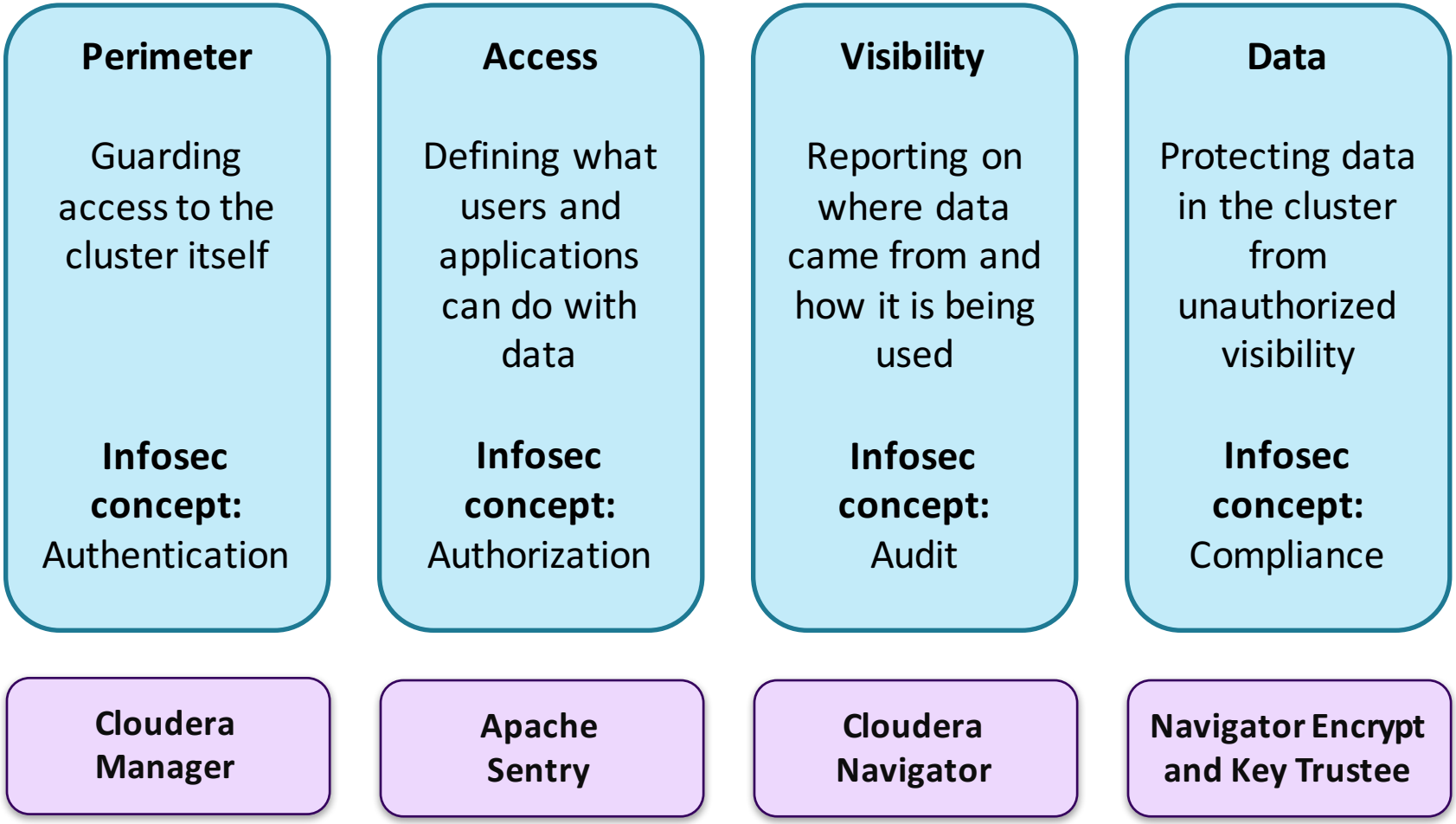
Managing Hadoop

- Cloudera Manager
- Cloudera Director
- Cloudera Navigator
- **Hadoop Security**
- Who Does the Work?
- Essential Points

Hadoop Security: A Full Suite

- **Hadoop has long supported Kerberos**
  - Provides authentication services for Hadoop
    - "Prove you are who you say you are"
- **Sentry provides authorization**
  - Limits the actions people can perform
- **Hadoop supports HDFS encryption**
  - All data in specified directories can be encrypted
  - Cloudera Enterprise includes an enterprise-grade key management solution
    - Also includes volume-level HDFS encryption to protect log files, metadata databases and so on
- **Cloudera Enterprise, including Cloudera Navigator, along with Kerberos and Sentry, provide a full security package for Hadoop**

Authentication, Authorization, Audit, and Compliance



MasterCard PCI Compliance

- **Payment Card Industry Data Security Standards (PCI DSS)**
  - Ensures privacy levels are met when storing, processing and transmitting data
  - Cloudera worked with MasterCard to create the world's first PCI-compliant Hadoop environment
- **Use cases**
  - Fraud detection
  - Consumer behavioral modeling
  - Security analytics
  - And so on

Chapter Topics

Managing Hadoop

- Cloudera Manager
- Cloudera Director
- Cloudera Navigator
- Hadoop Security
- **Who Does the Work?**
- Essential Points

The Roles People Play

- **System Administrators**
- **Developers**
- **Analysts**
- **Data Scientists**
- **Data Stewards**

System Administrators

- **Job responsibilities:**
  - Install, configure, and upgrade Hadoop software
  - Manage hardware components
  - Monitor the cluster
  - Integrate with other systems (such as Flume, Kafka, and Sqoop)
- **Required skills:**
  - Strong Linux administration skills
  - Networking knowledge
  - Understanding of hardware
- **Good candidates:**
  - Current Linux administrators
  - Database administrators
  - Network administrators

Developers

- **Job responsibilities:**
  - Write, package, and deploy Spark or MapReduce programs
- **Required skills:**
  - Strong Scala, Python, or Java capabilities
  - Understanding of distributed processing and algorithms
- **Good candidates:**
  - Developers with database experience

Data Analysts/Business Analysts

- **Job responsibilities:**
  - Extract intelligence from the data
  - Write Impala or Hive queries
- **Required skills:**
  - SQL
  - Understanding of data analytics/data mining
- **Good candidates:**
  - Current data analysts

Data Scientists

- **Job responsibilities:**
  - Answering questions you didn't realize you wanted to ask
  - Building "data products"
    - Products which consume, and generate, data
    - Such as recommender systems
- **Required skills:**
  - Knowledge of statistics
  - Software engineering skills
  - Domain knowledge
- **Good candidates:**
  - Statisticians with experience in data management

Data Stewards

- **Job responsibilities:**
  - Cataloging the data (analogous to a librarian for books)
  - Managing data lifecycle, retention
  - Data quality control with SLAs
- **Required skills:**
  - Data modeling and ETL
  - Scripting skills
- **Good candidates:**
  - Database administrators

Finding the Right People

- **Hiring Hadoop experts**
  - Strong Hadoop skills are scarce and expensive
  - Hadoop User Groups are a good source of talent
  - Key words
    - Developers: Spark, Cloudera Certified Developer for Apache Hadoop (CCDH), Cloudera Certified Professional: Data Engineer (CCP: DE)
    - System Admins: distributed systems (such as Teradata, RedHat Cluster), Linux, Cloudera Certified Administrator for Apache Hadoop (CCAH)
- **Consider cross-training, especially system administrators and data librarians**

Chapter Topics

Managing Hadoop

- Cloudera Manager
- Cloudera Director
- Cloudera Navigator
- Hadoop Security
- Who Does the Work?
- **Essential Points**

Essential Points

- **Management tools**
  - Cloudera Manager—Manage and troubleshoot your Hadoop cluster
  - Cloudera Director—Deploy your Hadoop environment to the cloud
  - Cloudera Navigator—Data management for your Hadoop cluster
- **Plan your human resources**
  - Identify current employees which may be a good fit and train them
  - Identify training and hiring needs
  - Linux knowledge is very important

cloudera®

Conclusion

Chapter 7

Course Chapters

- Introduction
- Hadoop Basics
- The Hadoop Ecosystem
- An Introduction to Hadoop Architecture
- Hadoop in the Real World
- Managing Hadoop
- **Conclusion**

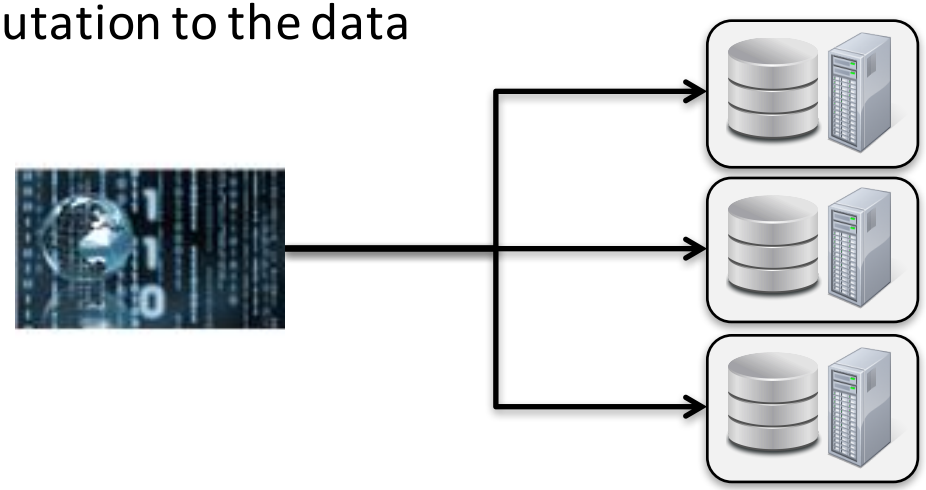
Chapter Topics

Conclusion

- **Review**
- How Can Cloudera Help?

Why Hadoop?

- **Why Hadoop?**
  - Data is growing faster than our ability to store and process it
  - Traditional data storage alternatives are no longer cost effective
  - The Hadoop approach:
    - Distribute data when it is stored
    - Bring the computation to the data



What Is Hadoop? (1)

- **Hadoop is a framework for storage and processing of big data that is**
  - Distributed
  - Scalable
  - Fault tolerant
  - Open source
- **The Hadoop ecosystem consists of two classes of components**
  - Core Hadoop (HDFS, MapReduce, and YARN) are foundational
  - Many other components run on this foundation

What is Hadoop? (2)

- **Hadoop has a large ecosystem**
  - The ability to mix and match components is one of its greatest strengths

Project	What does it do?
Spark	In-memory execution 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	Enables users to find the data they need
Hue	Web-based user interface for Hadoop
Sentry	Authorization tool, providing security for Hadoop

Hadoop is Used by Real Businesses

- **Financial Services**
  - JPMorgan Chase
  - MasterCard
  - Western Union
- **Insurance**
  - Allstate
  - RelayHealth
  - Markerstudy
- **Telecommunications**
  - BT
  - Telkomsel
  - True
  - SFR

Hadoop Can Help Solve Real Problems

- **Data storage**
  - HBase
- **Data integration**
  - Flume, Kafka, and Sqoop
- **Data processing**
  - Spark and MapReduce
- **Data analysis**
  - Impala and Hive
- **Data exploration**
  - Cloudera Search
- **Data security**
  - Sentry

Hadoop Architecture

- **There are two types of nodes**

Master Nodes	Worker Nodes
Manage the work	Do the work
Carrier class hardware	Industry standard hardware
Protect them from failure	Expendable
Configure for high availability (HA)	Do not need to be configured for HA

- **Hadoop can be deployed to hardware in your data center or to the cloud**

CDH Security

- **Authentication: "Prove you are who you say you are"**
  - Kerberos
- **Authorization: "What are you allowed to do?"**
  - HDFS file permissions
  - Sentry
- **Encryption is available for disks, file systems, databases, and applications**
- **Cloudera worked with MasterCard to create the world's first PCI compliant Hadoop environment**

Managing Hadoop

- **Management tools increase productivity**
  - Cloudera Manager
    - Manage and troubleshoot your Hadoop cluster
  - Cloudera Director
    - Deploy your Hadoop environment to the cloud
  - Cloudera Navigator
    - Data management for your Hadoop cluster
- **Plan your physical resources**
  - Start small then grow
  - Anticipate increasing capacity
- **Plan your human resources**
  - Identify current employees who may be a good fit and train them
  - Identify training and hiring needs

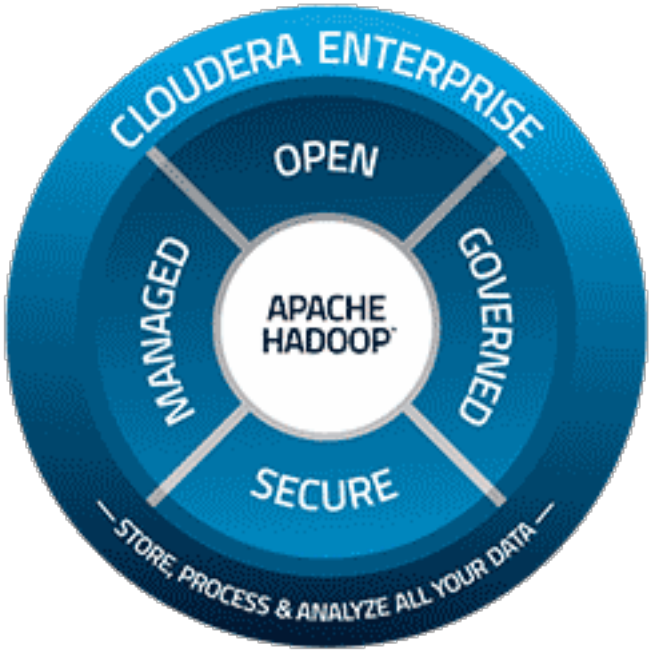
Chapter Topics

Conclusion

- Review
- **How can Cloudera help?**

Cloudera Enterprise

- **Subscription product including CDH and Cloudera Manager**
  - Includes support
  - Includes extra Cloudera Manager features
    - Configuration history and rollbacks
    - Rolling updates
    - LDAP integration
    - SNMP support
    - Automated disaster recovery



Cloudera Professional Services

- **Solutions Architects provide guidance and hands-on expertise**
  - Use case discovery
  - New Hadoop deployment / cluster certification
  - Proof of concept
  - Production pilot
  - Production readiness
  - Process and team development

Cloudera University Training

- ***Cloudera Developer Training for Spark and Hadoop***
  - Ingest and process data using Spark and other ecosystem tools
- ***Cloudera Data Analyst Training: Using Pig, Hive, and Impala with Hadoop***
  - Leverage existing skills to perform data analysis with big data tools
- ***Cloudera Administrator Training for Apache Hadoop***
  - Install, configure, and monitor clusters for optimal performance
- ***Data Science at Scale using Spark and Hadoop***
  - Gain expertise with the tools and techniques used by data scientists
- ***Cloudera Search Training***
  - Index, query, and build interactive dashboards for your data
- ***Cloudera HBase Training***
  - Learn how to use this NoSQL database effectively

Cloudera University Certification

- **Cloudera offers two levels of certifications**
  - Cloudera Certified Professional (CCP) is the industry's most demanding performance-based certification. It certifies a candidate's mastery of the technical skills most sought after by employers
    - CCP Data Engineer
    - CCP Data Scientist
  - Cloudera Certified Associate (CCA) exams test foundational skills and leads a candidate towards achieving mastery under the CCP program
    - CCA Spark and Hadoop Developer
    - Cloudera Certified Administrator for Apache Hadoop (CCAH)
- **For more information on Cloudera training and certifications:**
  - **`http://university.cloudera.com/`**

Course Summary

During this course, you have learned

- Why Hadoop is needed
- What type of problems can be solved with Hadoop
- The basic concepts of the Hadoop Distributed File System (HDFS)
- What projects are included in the Hadoop ecosystem
- The basics of Hadoop's architecture
- Who is using Hadoop
- What resources are available to assist in managing your Hadoop deployment

- **Thank you for attending!**