cloudera®

Introduction to Hadoop and the Hadoop Ecosystem

Chapter 2

Course Chapters

1	Introduction	Course Introduction
2	**Introduction to Hadoop and the Hadoop Ecosystem**	**Introduction to Hadoop**
3	Hadoop Architecture and HDFS	
4	Importing Relational Data with Apache Sqoop	Importing and Modeling Structured Data
5	Introduction to Impala and Hive	
6	Modeling and Managing Data with Impala and Hive	
7	Data Formats	
8	Data File Partitioning	
9	Capturing Data with Apache Flume	Ingesting Streaming Data
10	Spark Basics	Distributed Data Processing with Spark
11	Working with RDDs in Spark	
12	Aggregating Data with Pair RDDs	
13	Writing and Deploying Spark Applications	
14	Parallel Processing in Spark	
15	Spark RDD Persistence	
16	Common Patterns in Spark Data Processing	
17	Spark SQL and DataFrames	
18	Conclusion	Course Conclusion

Introduction to Hadoop and the Hadoop Ecosystem

In this chapter you will learn

- **What Hadoop is and how it addresses big data challenges**
- **The guiding principles behind Hadoop**
- **The major components of the Hadoop Ecosystem**
- **The tools you will be using in the Homework Labs**

Chapter Topics

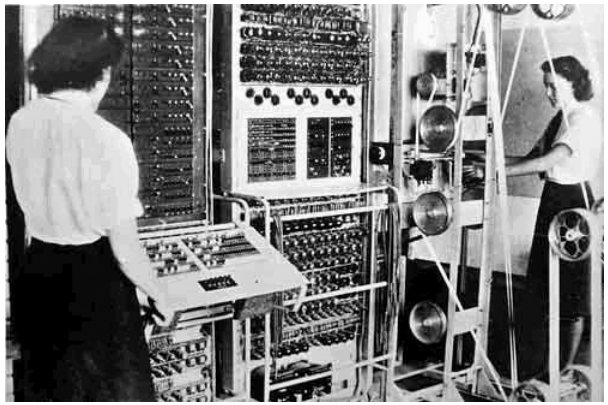
Introduction to Hadoop and the Hadoop Ecosystem

Introduction to Hadoop

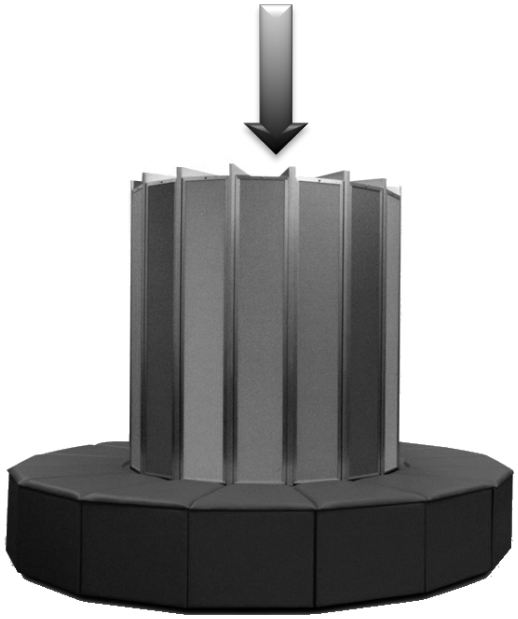
- **Problems with Traditional Large-scale Systems**
- Hadoop!
- Data Storage and Ingest
- Data Processing
- Data Analysis and Exploration
- Other Ecosystem Tools
- Introduction to Homework Labs
- Conclusion

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

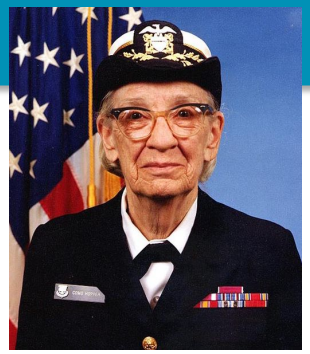


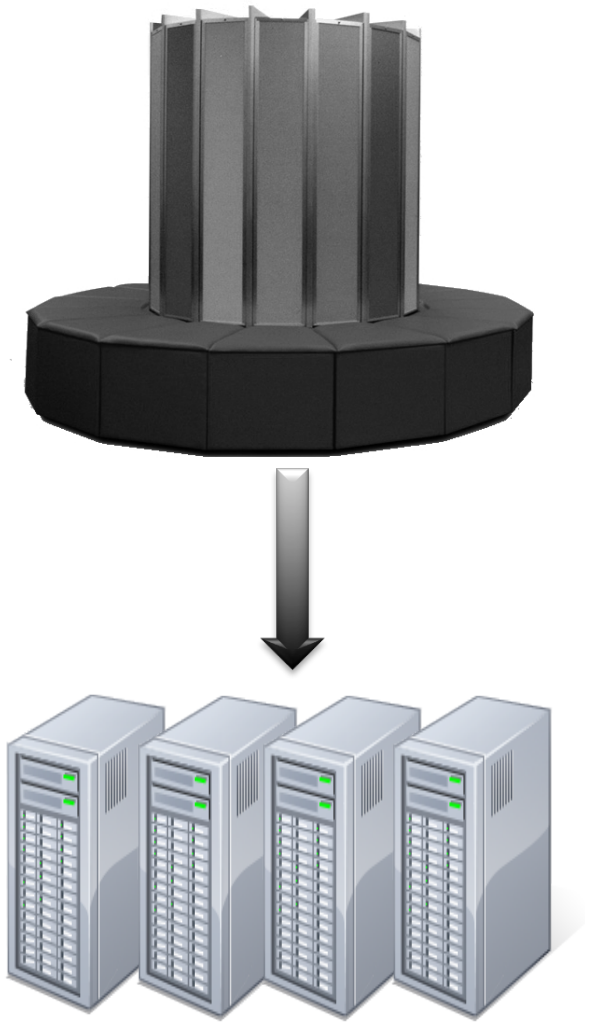
Distributed Systems

- **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





Challenges with Distributed Systems

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

- **The solution?**
  - Hadoop!

Chapter Topics

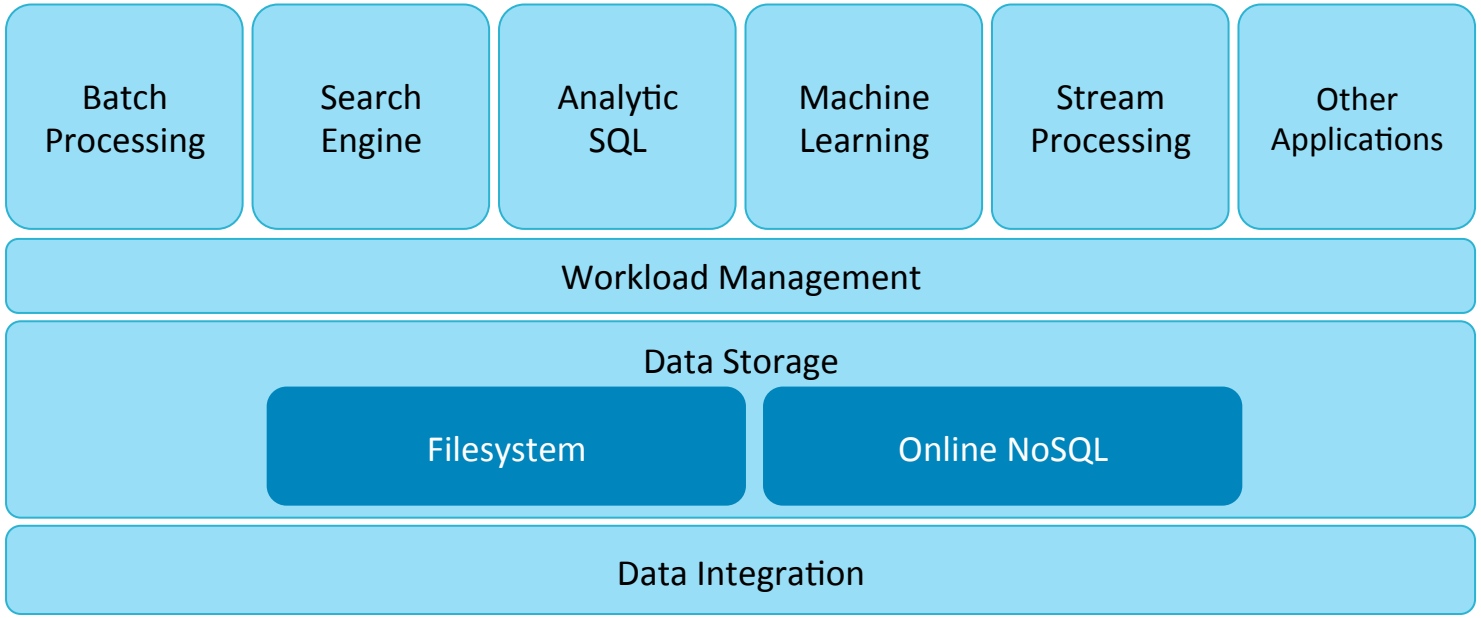
Introduction to Hadoop and the Hadoop Ecosystem

Introduction to Hadoop

- Problems with Traditional Large-scale Systems
- **Hadoop!**
- Data Storage and Ingest
- Data Processing
- Data Analysis and Exploration
- Other Ecosystem Tools
- Introduction to Homework Labs
- Conclusion

What is Apache Hadoop?

- **Scalable and economical data storage, processing and analysis**
  - Distributed and fault-tolerant
  - Harnesses the power of industry standard hardware
- **Heavily inspired by technical documents published by Google**



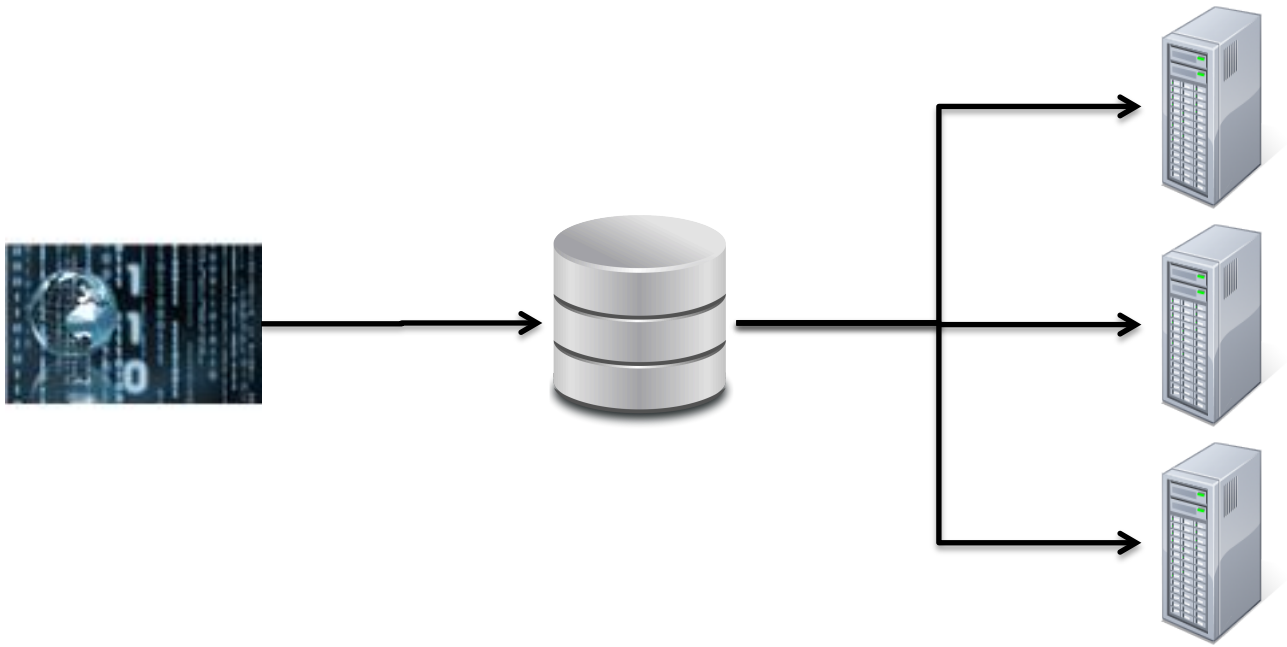
Common Hadoop Use Cases

- **Extract/Transform/Load (ETL)**
- **Text mining**
- **Index building**
- **Graph creation and analysis**
- **Pattern recognition**
- **Collaborative filtering**
- **Prediction models**
- **Sentiment analysis**
- **Risk assessment**

- **What do these workloads have in common? Nature of the data…**
  - Volume
  - Velocity
  - Variety

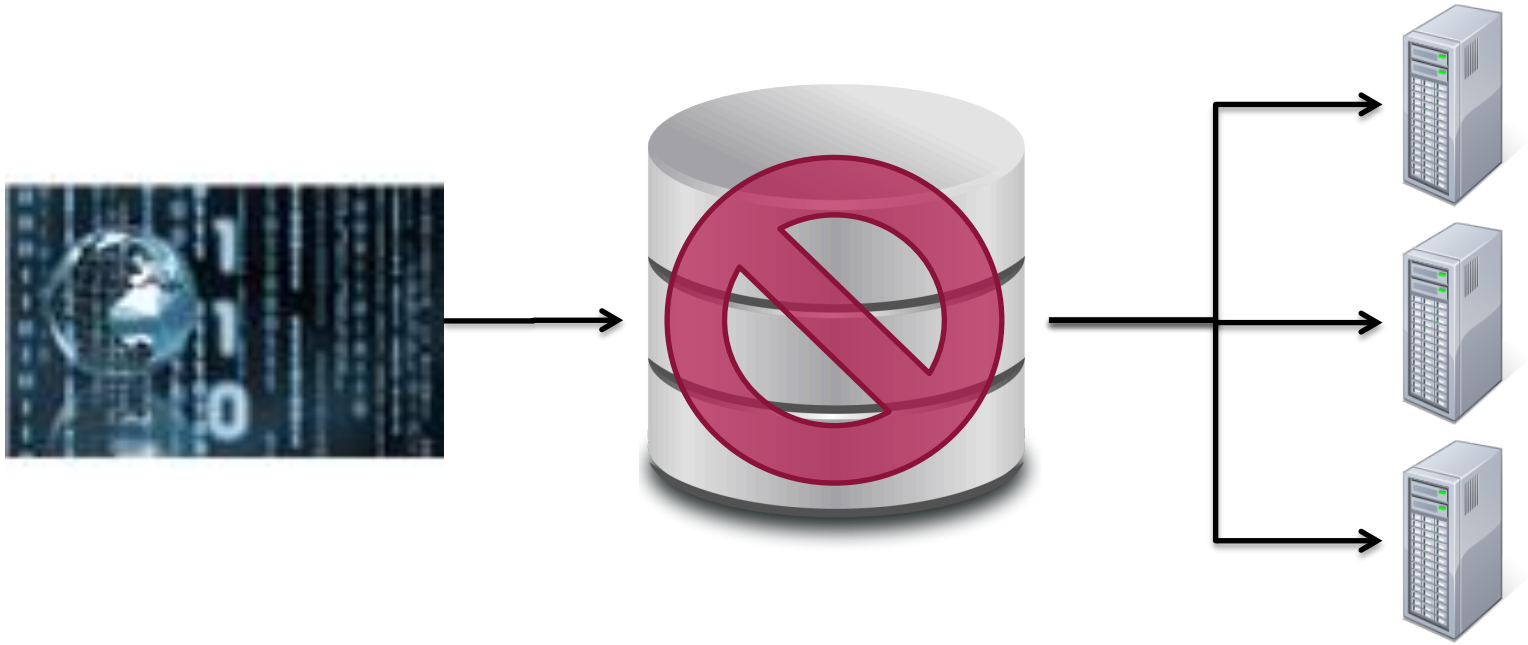
Distributed Systems: The Data Bottleneck (1)

- **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)

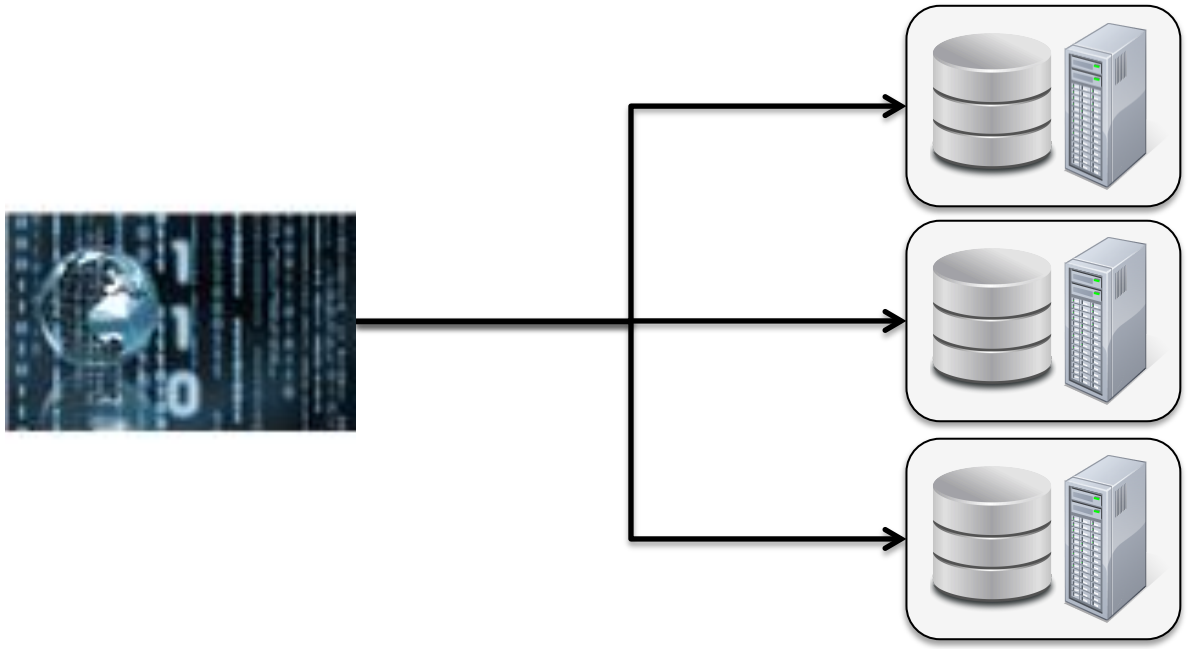
- **Modern systems have much more data**
  - terabytes+ a day
  - petabytes+ total
- **We need a new approach…**



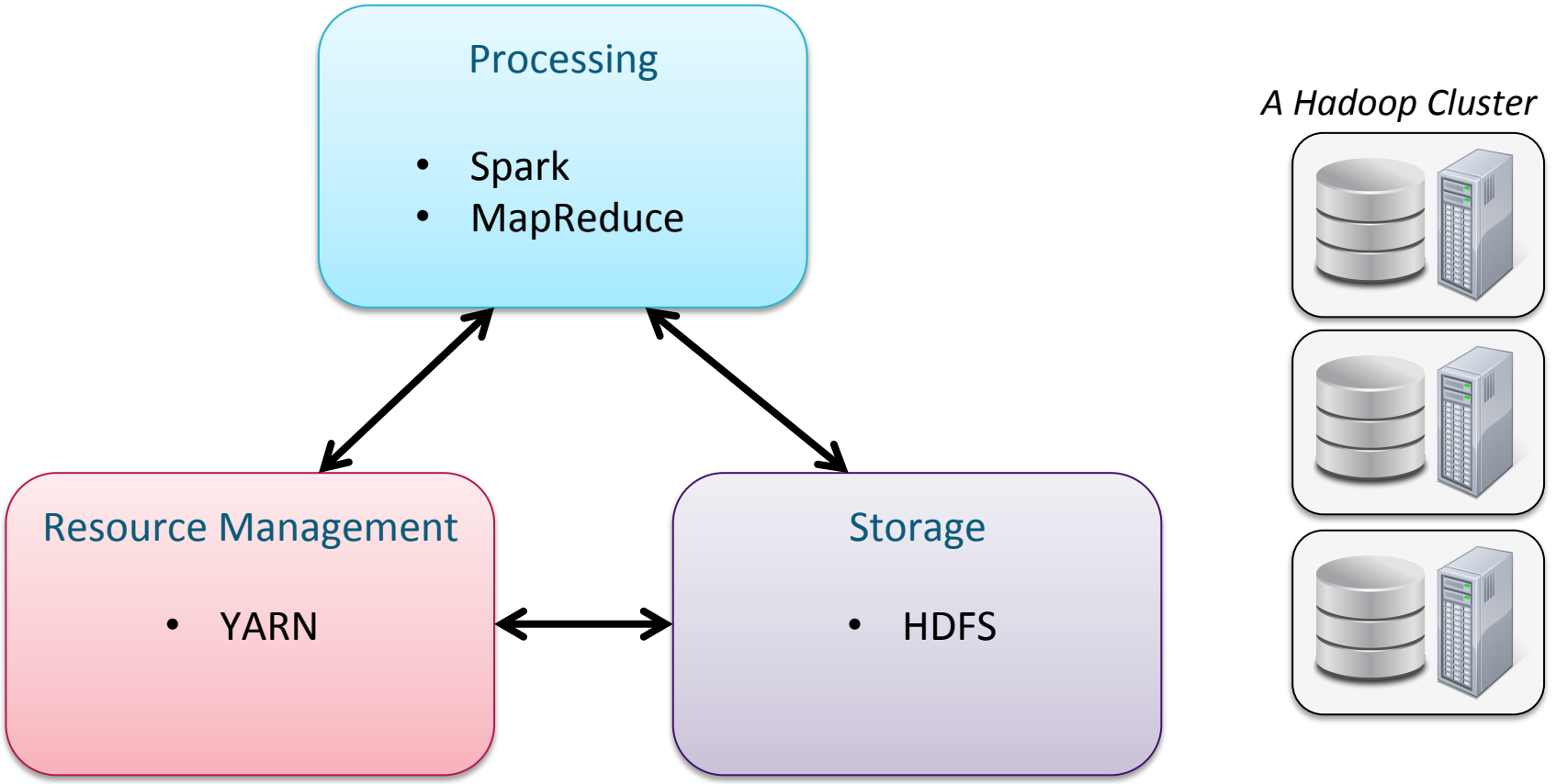
Big Data Processing with Hadoop

- **Hadoop introduced a radical new approach:**
  - Bring the program to the data rather than the data to the program
- **Based on two key concepts**
  - Distribute data when the data is stored
  - Run computation where the data resides

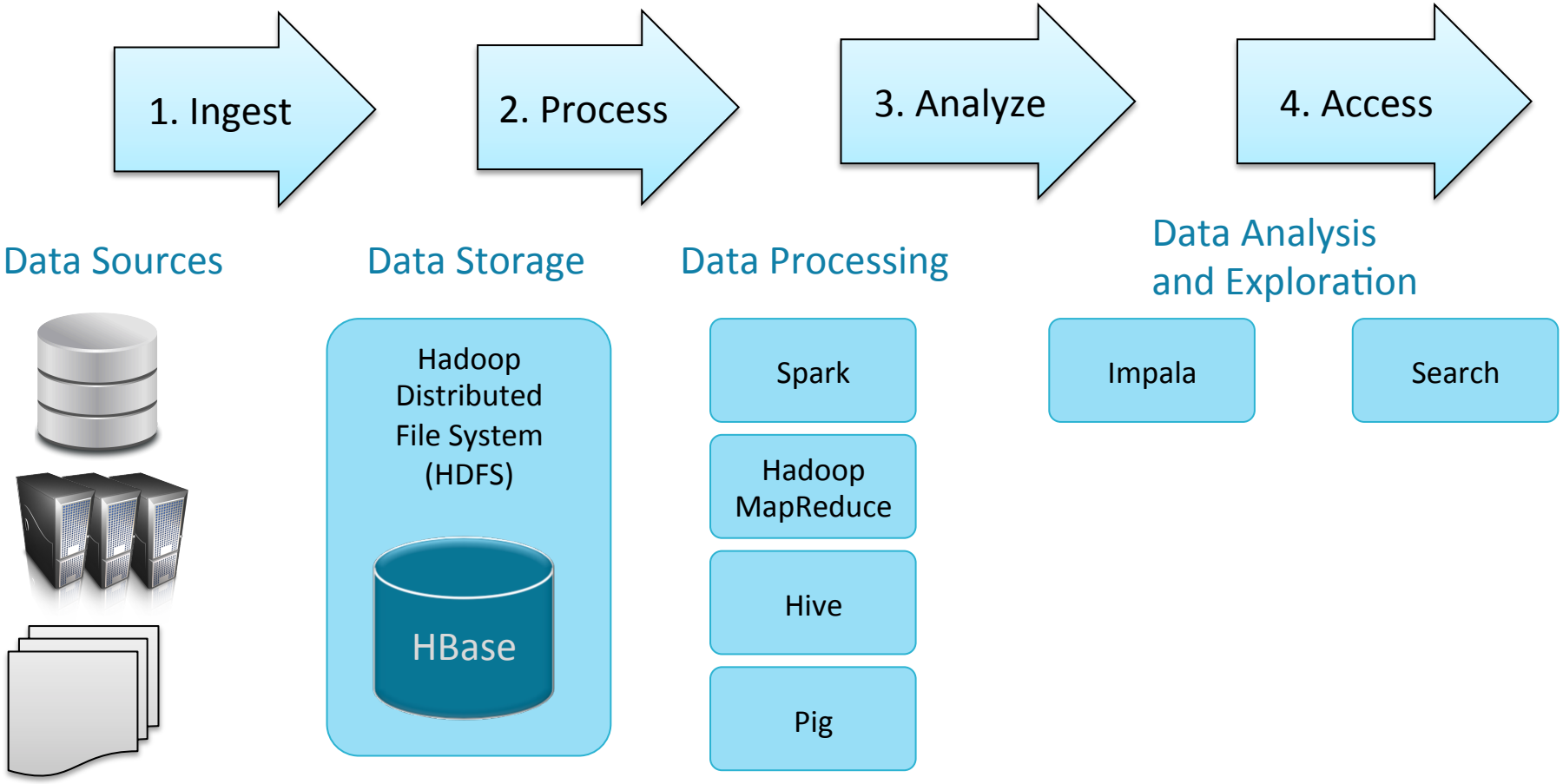
A Hadoop Cluster



Core Hadoop



Big Data Processing



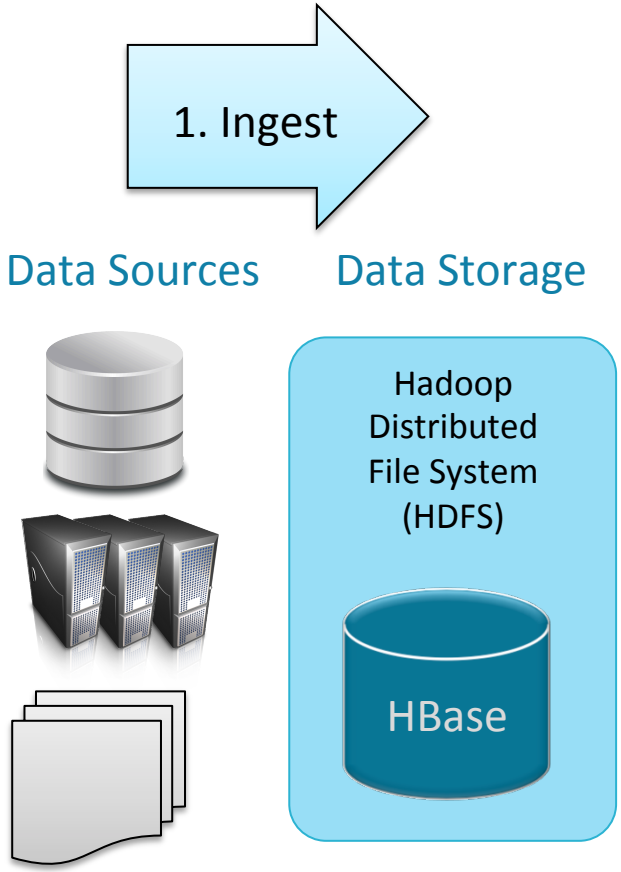
Chapter Topics

Introduction to Hadoop and the Hadoop Ecosystem | **Introduction to Hadoop**

- Problems with Traditional Large-scale Systems
- Hadoop!
- **Data Storage and Ingest**
- Data Processing
- Data Analysis and Exploration
- Other Ecosystem Tools
- Introduction to Homework Labs
- Conclusion

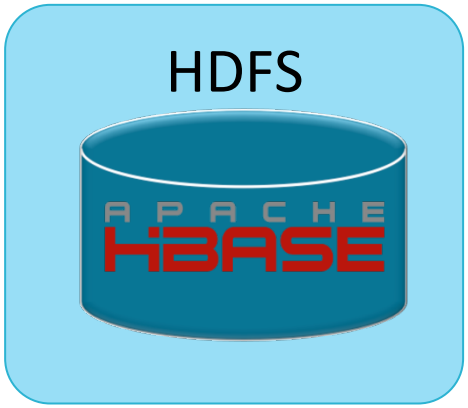
Data Ingest and Storage

- **Hadoop typically ingests data from many sources and in many formats**
  - Traditional data management systems, e.g. databases
  - Logs and other machine generated data (event data)
  - Imported files



Data Storage

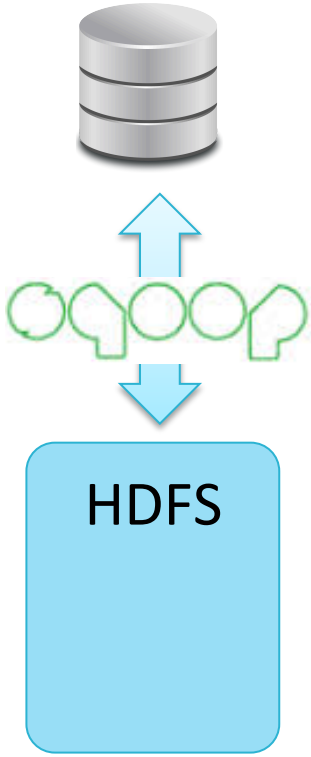
- **Hadoop Distributed File System (HDFS)**
  - HDFS is the storage layer for Hadoop
  - Provides inexpensive reliable storage for massive amounts of data on industry-standard hardware
  - Data is distributed when stored
  - Covered later in this course



- **Apache HBase: The Hadoop Database**
  - A NoSQL distributed database built on HDFS
  - Scales to support very large amounts of data and high throughput
  - A table can have thousands of columns
  - Covered in depth in *Cloudera Training for Apache HBase*

Data Ingest Tools (1)

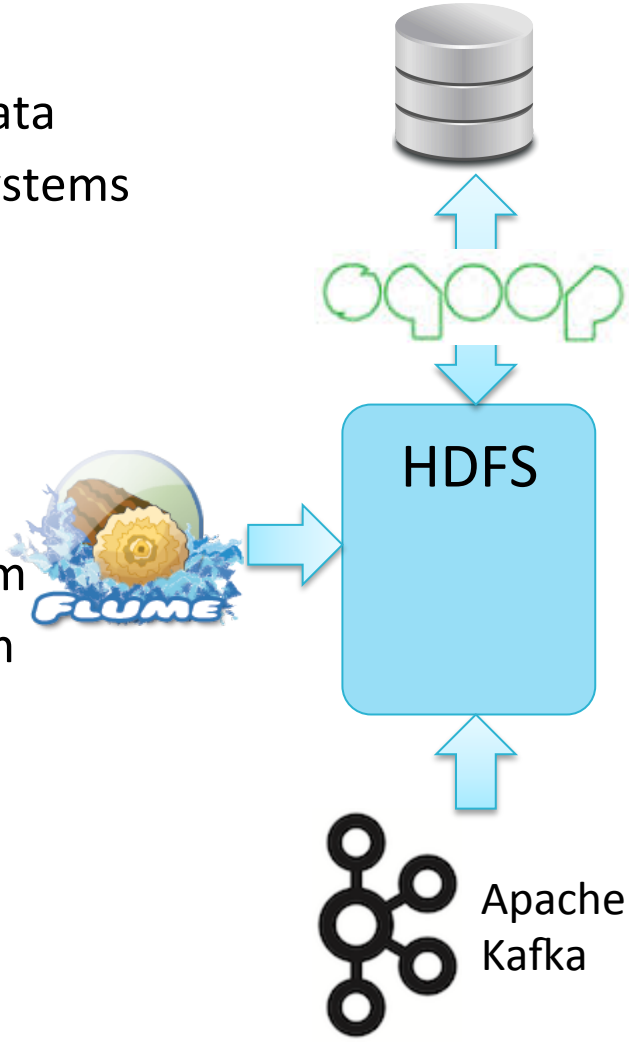
- **HDFS**
  - Direct file transfer
- **Apache Sqoop**
  - High speed import to HDFS from Relationship Database (and vice versa)
  - Supports many data storage systems
    - e.g. Netezza, Mongo, MySQL, Teradata, Oracle
  - Covered later in this course



Data Ingest Tools (2)

- **Apache Flume**
  - Distributed service for ingesting streaming data
  - Ideally suited for event data from multiple systems
    - For example, log files
  - Covered later in this course

- **Kafka**
  - A high throughput, scalable messaging system
  - Distributed, reliable publish-subscribe system
  - Integrates with Flume and Spark Streaming



Chapter Topics

Introduction to Hadoop and the Hadoop Ecosystem | **Introduction to Hadoop**

- Problems with Traditional Large-scale Systems
- Hadoop!
- Data Storage and Ingest
- **Data Processing**
- Data Analysis and Exploration
- Other Ecosystem Tools
- Introduction to Homework Labs
- Conclusion

Apache Spark: An Engine For Large-scale Data Processing

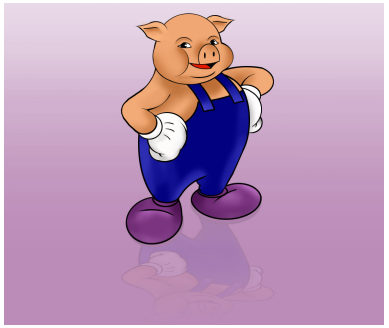
- **Spark is large-scale data processing engine**
  - General purpose
  - Runs on Hadoop clusters and data in HDFS
- **Supports a wide range of workloads**
  - Machine learning
  - Business intelligence
  - Streaming
  - Batch Processing
- **This course uses Spark for data processing**

Hadoop MapReduce: The Original Hadoop Processing Engine

- **Hadoop MapReduce is the original Hadoop framework**
  - Primarily Java based
- **Based on the MapReduce programming model**
- **The core Hadoop processing engine before Spark was introduced**
- **Still the dominant technology**
  - But losing ground to Spark fast
- **Many existing tools are still built using MapReduce code**
- **Has extensive and mature fault tolerance built into the framework**

Apache Pig: Scripting for MapReduce

- **Apache Pig builds on Hadoop to offer high-level data processing**
  - This is an alternative to writing low-level MapReduce code
  - Pig is especially good at joining and transforming data
- **The Pig interpreter runs on the client machine**
  - Turns Pig Latin scripts into MapReduce or Spark jobs
  - Submits those jobs to a Hadoop cluster
  - Covered in Cloudera *Data Analyst Training*



```
people = LOAD '/user/training/customers' AS (cust_id, name);
orders = LOAD '/user/training/orders' AS (ord_id, cust_id, cost);
groups = GROUP orders BY cust_id;
totals = FOREACH groups GENERATE group, SUM(orders.cost) AS t;
result = JOIN totals BY group, people BY cust_id;
DUMP result;
```

Chapter Topics

Introduction to Hadoop and the Hadoop Ecosystem

Introduction to Hadoop

- Problems with Traditional Large-scale Systems
- Hadoop!
- Data Storage and Ingest
- Data Processing
- **Data Analysis and Exploration**
- Other Ecosystem Tools
- Introduction to Homework Labs
- Conclusion

Cloudera Impala: High Performance SQL

- **Impala is a high-performance SQL engine**
  - Runs on Hadoop clusters
  - Data stored in HDFS files
  - Inspired by Google's Dremel project
  - Very low latency – measured in milliseconds
  - Ideal for interactive analysis

- **Impala supports a dialect of SQL (Impala SQL)**
  - Data in HDFS modeled as database tables
- **Impala was developed by Cloudera**
  - 100% open source, released under the Apache software license
- **Impala is used for data analysis in this course**

Apache Hive: SQL on MapReduce

- **Hive is an abstraction layer on top of Hadoop**
  - Hive uses a SQL-like language called HiveQL
  - Similar to Impala SQL
  - Useful for data processing and ETL
    - Impala is preferred for ad hoc analytics

- **Hive executes queries using MapReduce**
  - Hive on Spark is available for early adopters; not yet recommended for production

- **Hive can optionally be used for data analysis in this course**

Cloudera Search: A Platform For Data Exploration

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

Chapter Topics

Introduction to Hadoop and the Hadoop Ecosystem

Introduction to Hadoop

- Problems with Traditional Large-scale Systems
- Hadoop!
- Data Storage and Ingest
- Data Processing
- Data Analysis and Exploration
- **Other Ecosystem Tools**
- Introduction to Homework Labs
- Conclusion

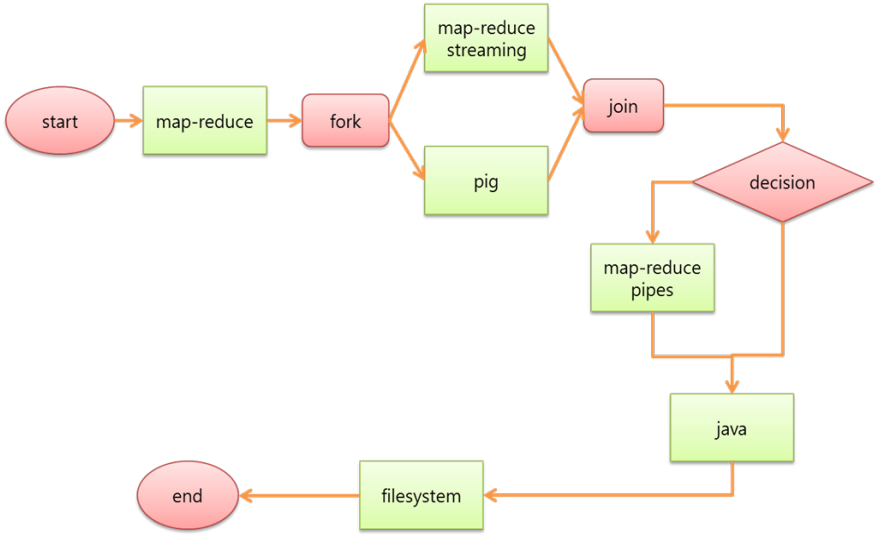
Hue: The UI for Hadoop

- **Hue = Hadoop User Experience**
- **Hue provides a Web front-end to a Hadoop**
  - Upload and browse data
  - Query tables in Impala and Hive
  - Run Spark and Pig jobs and workflows
  - Search
  - And much more
- **Makes Hadoop easier to use**
- **Hue is 100% open-source**
- **Created by Cloudera**
  - Open source, released under Apache license
- **Hue is used throughout this course**

Apache Oozie: Workflow Management

- **Oozie**
  - Workflow engine for Hadoop jobs
  - Defines dependencies between jobs

- **The Oozie server submits the jobs to the server in the correct sequence**



Apache Sentry: Hadoop Security

- **Sentry provides fine-grained access control (authorization) to various Hadoop ecosystem components**
  - Impala
  - Hive
  - Cloudera Search
  - HDFS
- **In conjunction with Kerberos authentication, Sentry authorization provides a complete cluster security solution**
- **Created by Cloudera**
  - Now an open-source Apache project

Chapter Topics

Introduction to Hadoop and the Hadoop Ecosystem

Introduction to Hadoop

- Problems with Traditional Large-scale Systems
- Hadoop!
- Data Storage and Ingest
- Data Processing
- Data Analysis and Exploration
- Other Ecosystem Tools
- **Introduction to the Homework Labs**
- Conclusion

Introduction to the Homework Labs

- **The best way to learn is to *do!***
- **Most topics in this course have a corresponding lab for practicing the skills you have learned in lecture**

Scenario Explanation (1)

- **The Homework Labs are based on a hypothetical scenario**
  - However, the concepts apply to nearly any organization
- **Loudacre Mobile is a (fictional) fast-growing wireless carrier**
  - Provides mobile service to customers throughout western USA

Scenario Explanation (2)

- **Loudacre needs to migrate their existing infrastructure to Hadoop**
  - The size and velocity and their data has exceeded their ability to processing and analyze their data
- **Loudacre data sources**
  - MySQL database – customer account data (name, address, phone numbers, devices)
  - Apache web server logs from Customer Service site
  - HTML files – Knowledge base articles
  - XML files – Device activation records
  - Real-time device status logs
  - Base stations – cell tower locations

Introduction to Homework Labs: Getting Started

- **Instructions are in the Homework Labs**
- **Start with**
  - General Notes
  - Setting Up
    - Run setup script for the course

Introduction to Homework Labs: Classroom Virtual Machine

- **Your virtual machine**
  - Log in as user **`training`** (password **`training`**)
  - Pre-installed and configured with
    - Spark and CDH (Cloudera's Distribution, including Apache Hadoop)
    - Various tools including Firefox, gedit, Emacs, Eclipse, and Maven
- **Training materials: `~/training_materials/dev1` folder on the VM**
  - **`exercises`** – one folder per homework
  - **`scripts`** – course setup scripts
- **Course data: `~/training_materials/data`**

Chapter Topics

Introduction to Hadoop and the Hadoop Ecosystem

Introduction to Hadoop

- Problems with Traditional Large-scale Systems
- Hadoop!
- Data Storage and Ingest
- Data Processing
- Data Analysis and Exploration
- Other Ecosystem Tools
- Introduction to Homework Labs
- **Conclusion**

Essential Points

- **Hadoop is a framework for distributed storage and processing**
- **Core Hadoop includes HDFS for storage and YARN for cluster resource management**
- **The Hadoop ecosystem includes many components for**
  - Ingesting data (Flume, Sqoop, Kafka)
  - Storing data (HDFS, HBase)
  - Processing data (Spark, Hadoop MapReduce, Pig)
  - Modeling data as tables for SQL access (Impala, Hive)
  - Exploring data (Hue, Search)
  - Protecting Data (Sentry)
- **This course introduces most of the key Hadoop infrastructure**
- **Homework Labs let you practice and refine your Hadoop skills!**

Bibliography

The following offer more information on topics discussed in this chapter

- ***Hadoop: The Definitive Guide*** **(published by O'Reilly)**
  - `http://tiny.cloudera.com/hadooptdg`
- ***Cloudera Essentials for Apache Hadoop*** **– free online training**
  - `http://tiny.cloudera.com/esscourse`