Knowledge in motion

Learning Hadoop 2

Randal Scott King


Sensor Data
Transactions
News Feeds
Marketing
Your Company
Web Logs
Customer Service
Weather

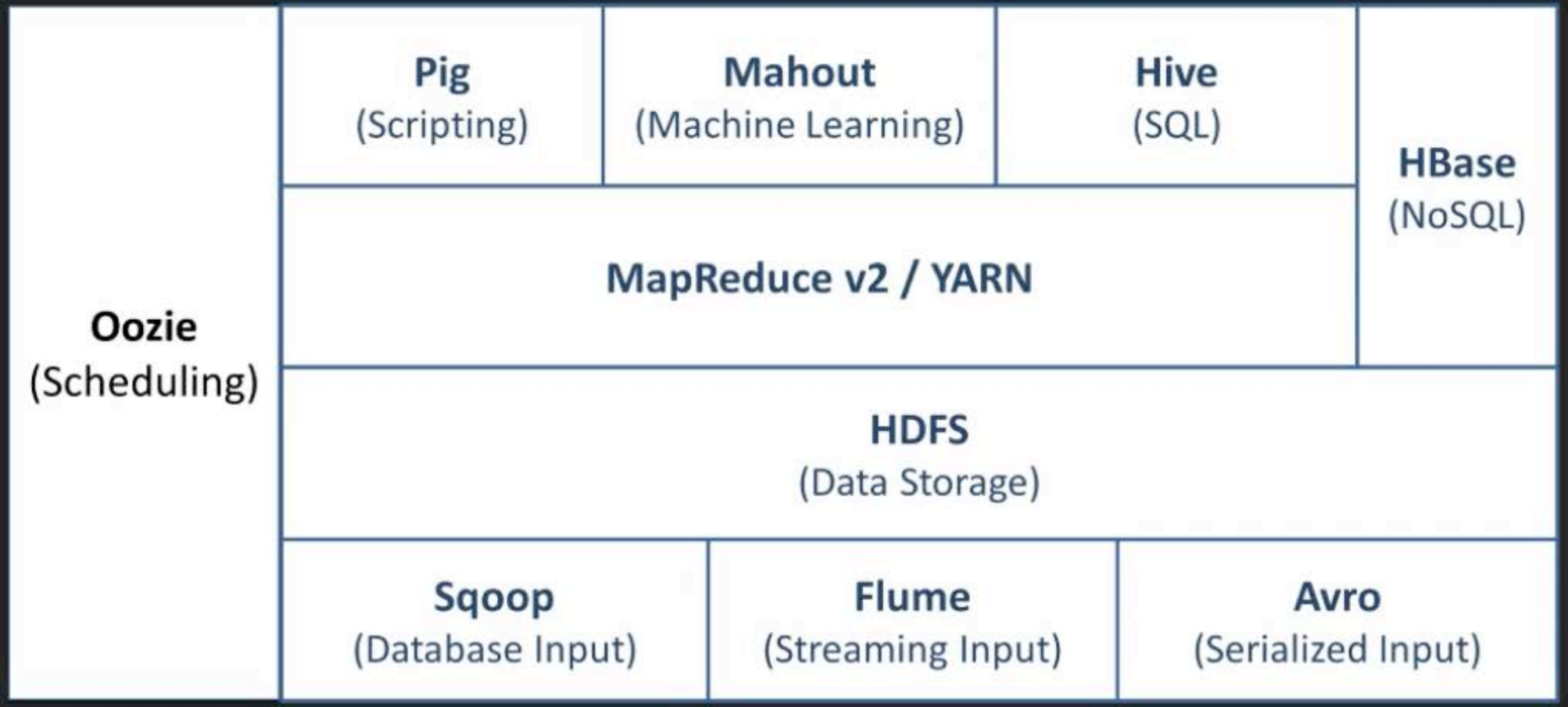
Learning Hadoop 2

Randal Scott King

Video 1.1

The Course Overview

Hadoop Is an Ecosystem



Applications Places System Sun Aug 9, 08:52 cloudera

Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox

Cloudera Live : Welcom...

quickstart.cloudera/#/

Search

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

cloudera LIVE Navigation

Welcome to Your Cloudera QuickStart VM!

Your Cluster

Node	Address
Manager Node	127.0.0.1
Worker Node 1	127.0.0.1

Get Started

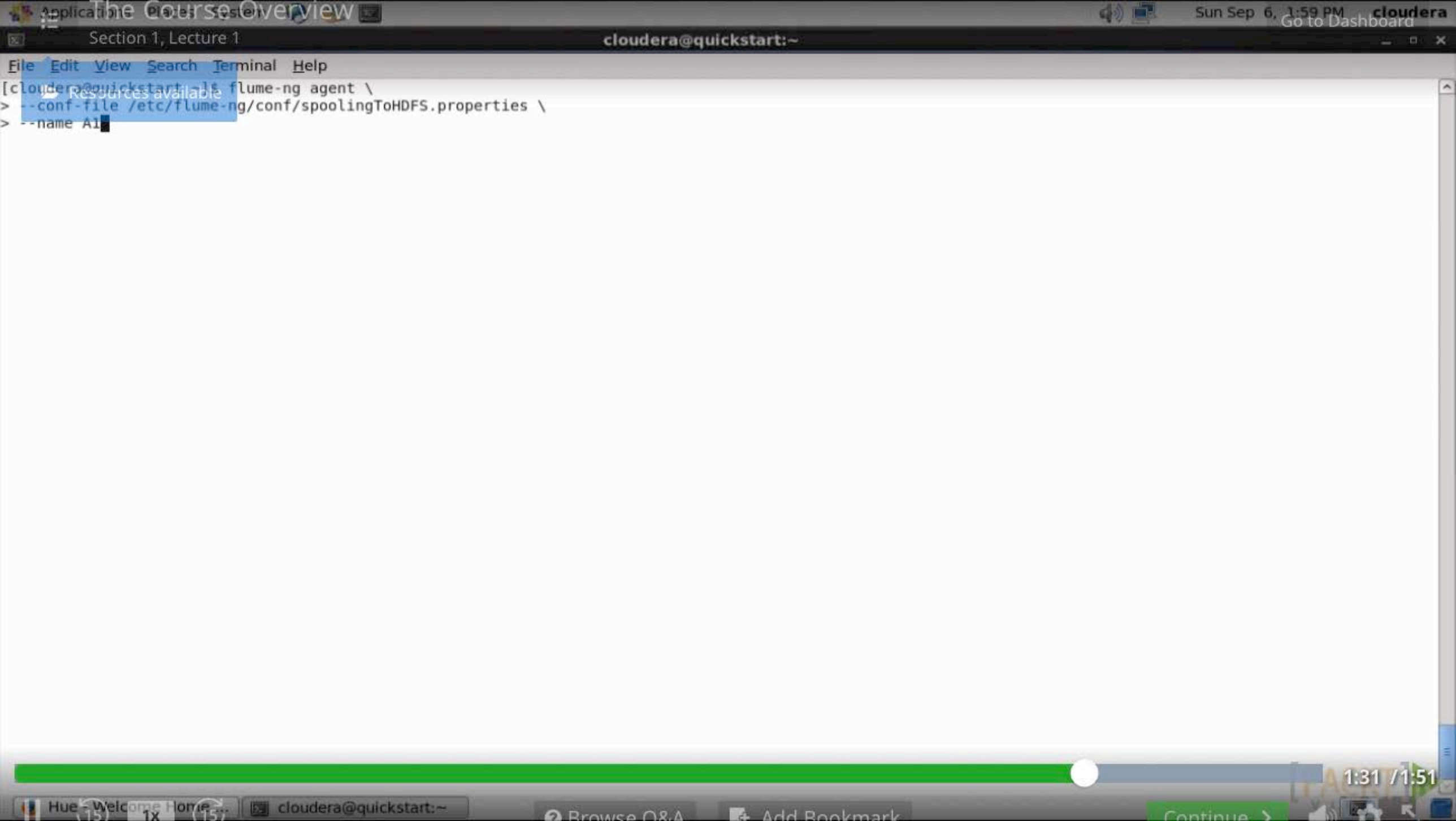
The tutorial below guides you through some analytic use cases, using the most popular open source tools included with CDH (including Cloudera Impala, Cloudera Search, and Hue).

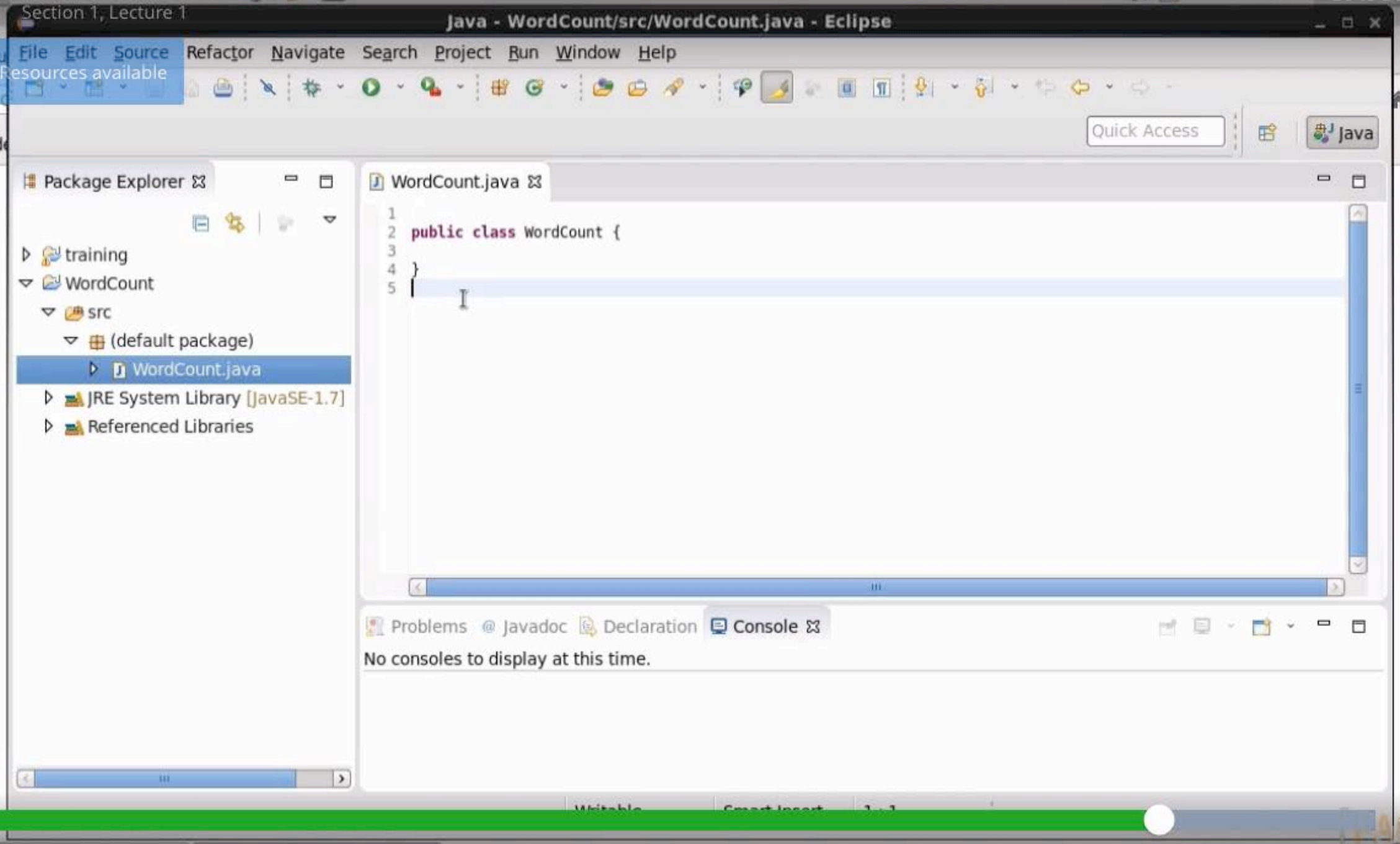
Start Tutorial

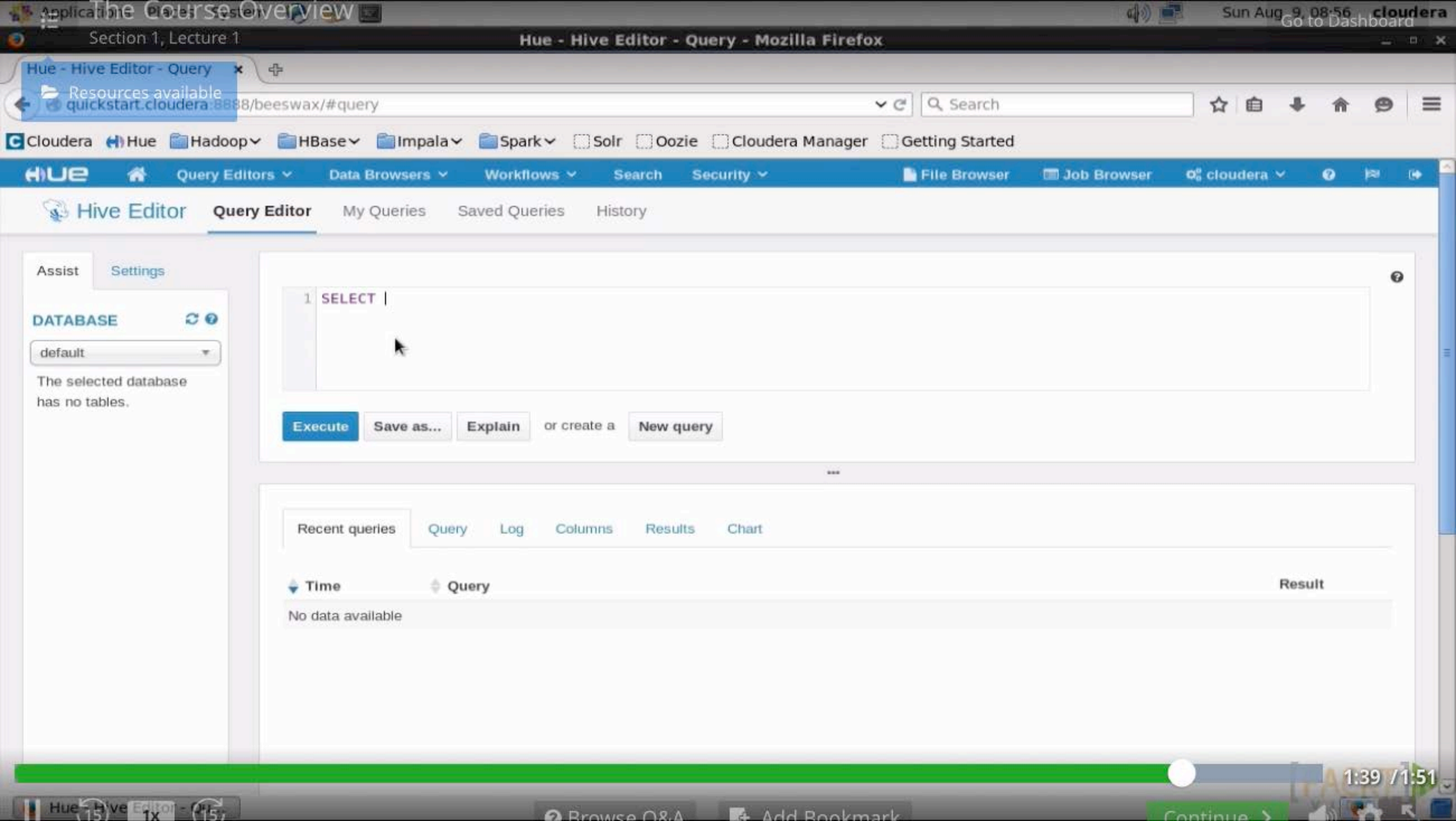
Analyze Your Data

Hue is the open source web interface for Hadoop that lets you analyze your data. Simply load in your data and then easily begin to analyze, search, and visualize it. In the QuickStart VM, the administrative username for Hue is 'cloudera' and the password is

Cloudera Live : Welco...


Applications Places System
Sun Sep 6, 1:59 PM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ flume-ng agent \
> --conf-file /etc/flume-ng/conf/spoolingToHDFS.properties \
> --name A1
Resources available




Applications Places System
Sun Aug 9, 08:56 cloudera
Go to Dashboard
Hue - Hive Editor - Query - Mozilla Firefox
Hue - Hive Editor - Query
Resources available
quickstart.cloudera:8888/beeswax/#query
Search
Cloudera
Hue
Hadoop
HBase
Impala
Spark
Solr
Oozie
Cloudera Manager
Getting Started
HUE
Query Editors
Data Browsers
Workflows
Search
Security
File Browser
Job Browser
cloudera
Hive Editor
Query Editor
My Queries
Saved Queries
History
Assist
Settings
DATABASE
default
The selected database has no tables.
1 SELECT
Execute
Save as...
Explain
or create a
New query
Recent queries
Query
Log
Columns
Results
Chart
Time
Query
Result
No data available

Learning Hadoop 2

Randal Scott King

Section 1

The Hadoop Ecosystem

In this Section, we are going to take a look at...

- Overview of HDFS and YARN
- Overview of Sqoop and Flume
- Overview of MapReduce
- Overview of Pig
- Overview of Hive

Learning Hadoop 2

Randal Scott King

Video 1.2

Overview of HDFS and YARN

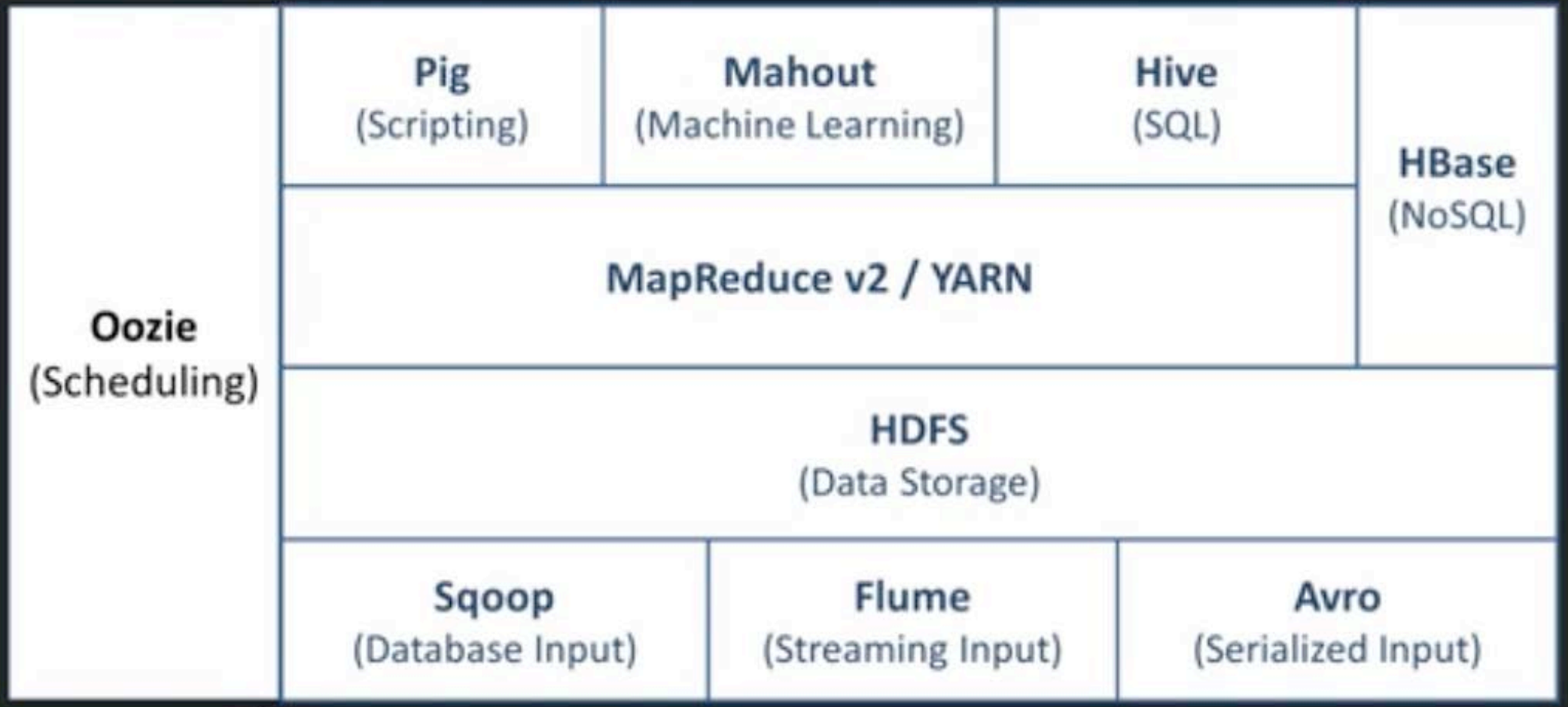
In this Video, we are going to take a look at...

- Distributed computing
- Hadoop Distributed File System (HDFS)
- Yet Another Resource Negotiator (YARN)

Distributed Computing

- Problem: Big Data strains computing resources (CPU, storage)
- Solution: Distribute the load over many servers rather than one
- Yahoo! used the distributed computing model to develop Hadoop

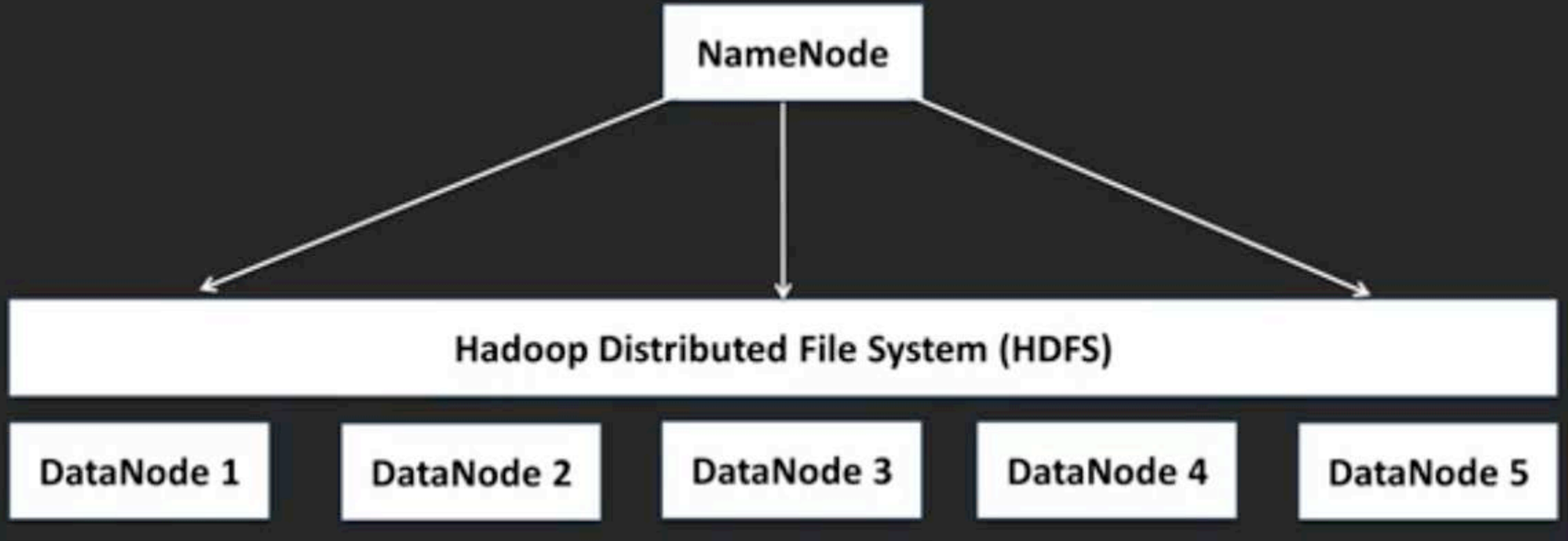
Hadoop Is an Ecosystem



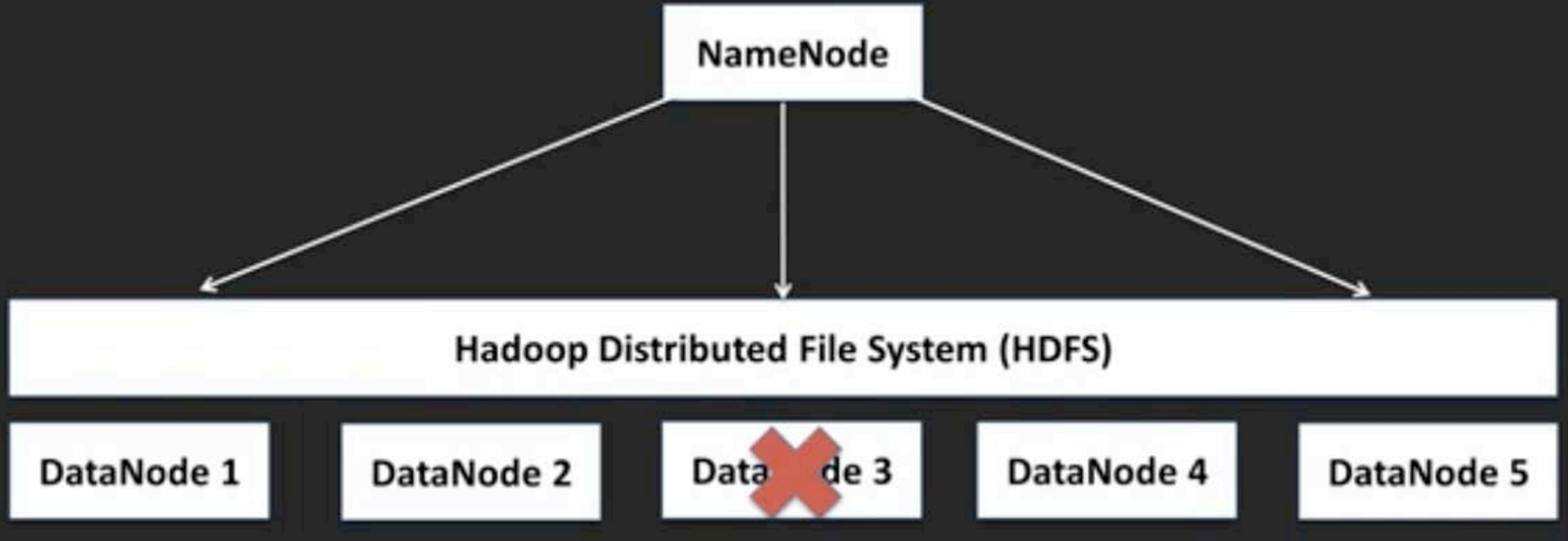
Hadoop Distributed File System (HDFS)

- File system for Hadoop
- Spans all nodes in a cluster
- Stores data in 64Meg chunks on multiple servers

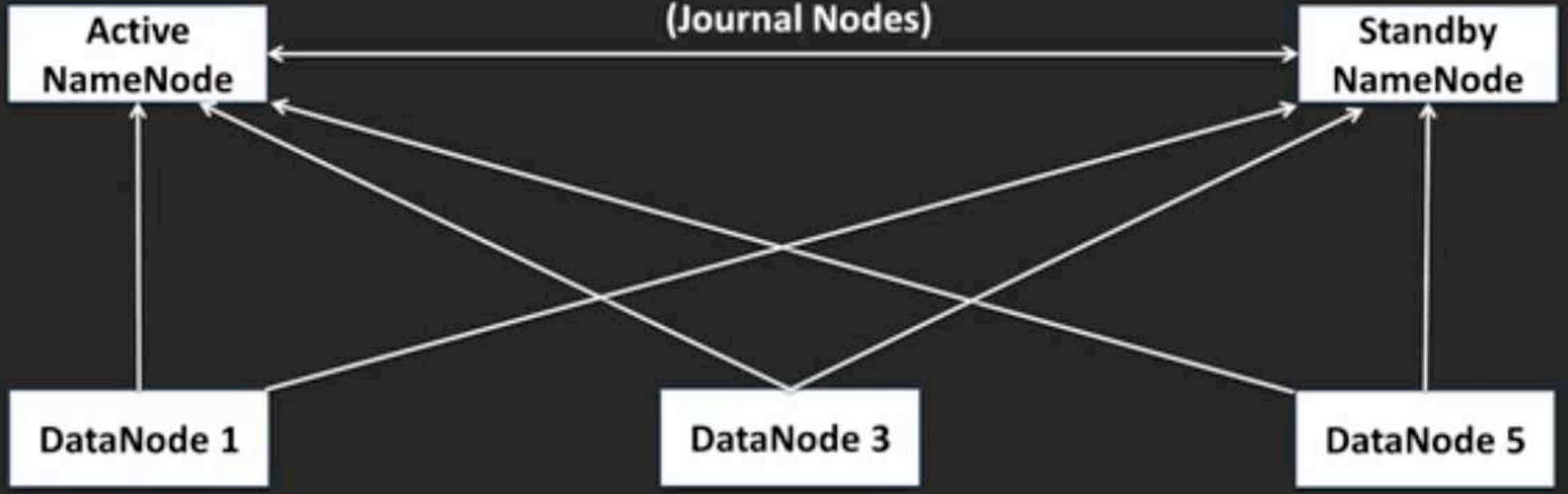
HDFS


NameNode
Hadoop Distributed File System (HDFS)
DataNode 1
DataNode 2
DataNode 3
DataNode 4
DataNode 5

HDFS


NameNode
Hadoop Distributed File System (HDFS)
DataNode 1
DataNode 2
Data de 3
DataNode 4
DataNode 5

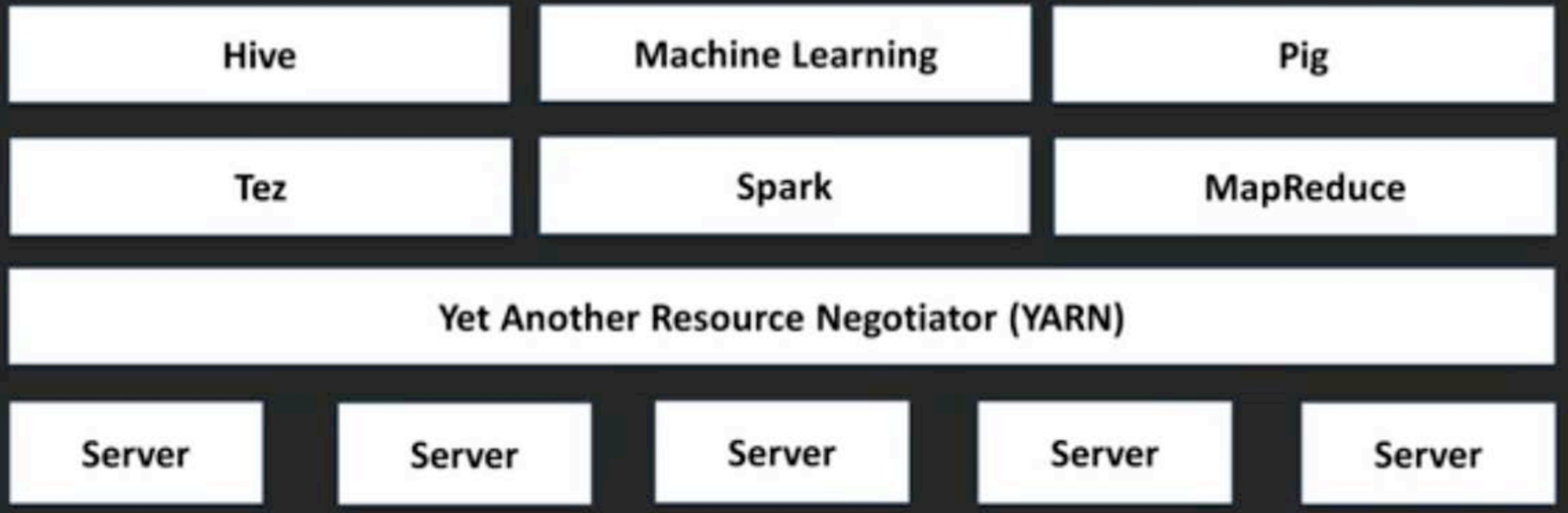
High Availability in HDFS



Yet Another Resource Negotiator (YARN)

- Controls access to cluster resources
- New in Hadoop v2
- Allows multiple compute engines to run (MapReduce, Spark, Tez, and so on)

YARN


Hive
Machine Learning
Pig
Tez
Spark
MapReduce
Yet Another Resource Negotiator (YARN)
Server
Server
Server
Server
Server

Next Video

Overview of Sqoop and Flume

Learning Hadoop 2

Randal Scott King

Video 1.3

Overview of Sqoop and Flume

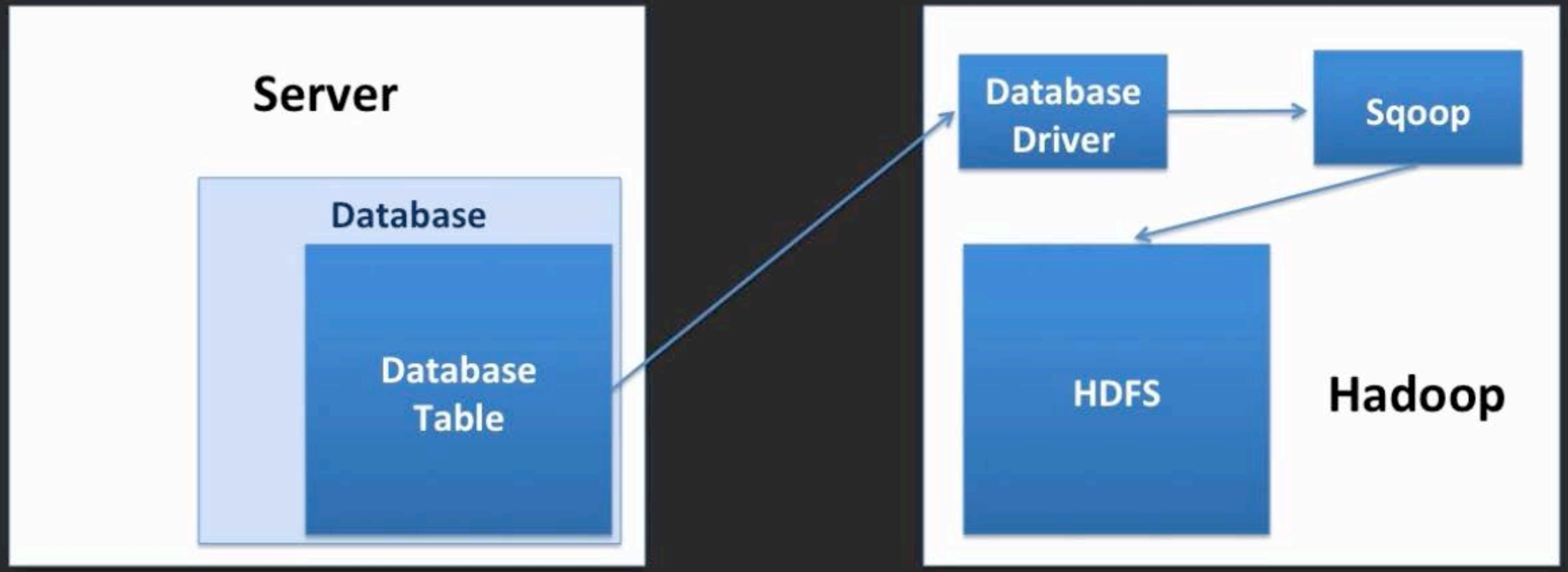
In this Video, we are going to take a look at...

- Introduction to Sqoop
- Introduction to Flume

Sqoop (SQl on hadOOP)

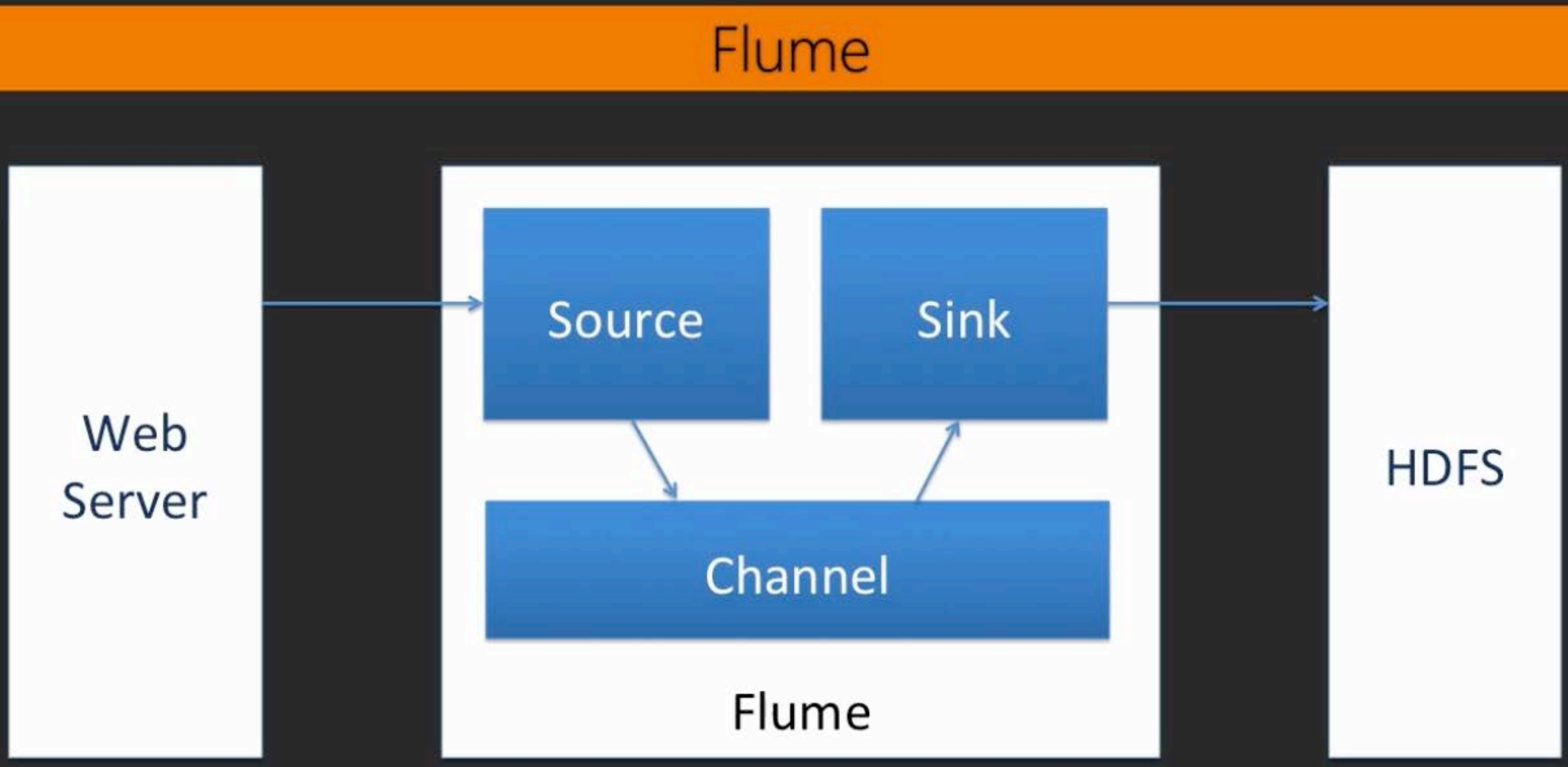
- Pulls data from relational databases (Oracle, PostgreSQL, and so on)
- Stores on HDFS or imports directly to Hive
- Uses drivers that are not included

Sqoop



Flume

- Used to import streaming data (server logs, tweets, and so on)
- Only a transport agent
- Buffered
- Three parts: Source, Channel, and Sink


Flume
Web Server
Source
Sink
Channel
Flume
HDFS

Next Video

Overview of MapReduce

Learning Hadoop 2

Randal Scott King

Video 1.4

Overview of MapReduce

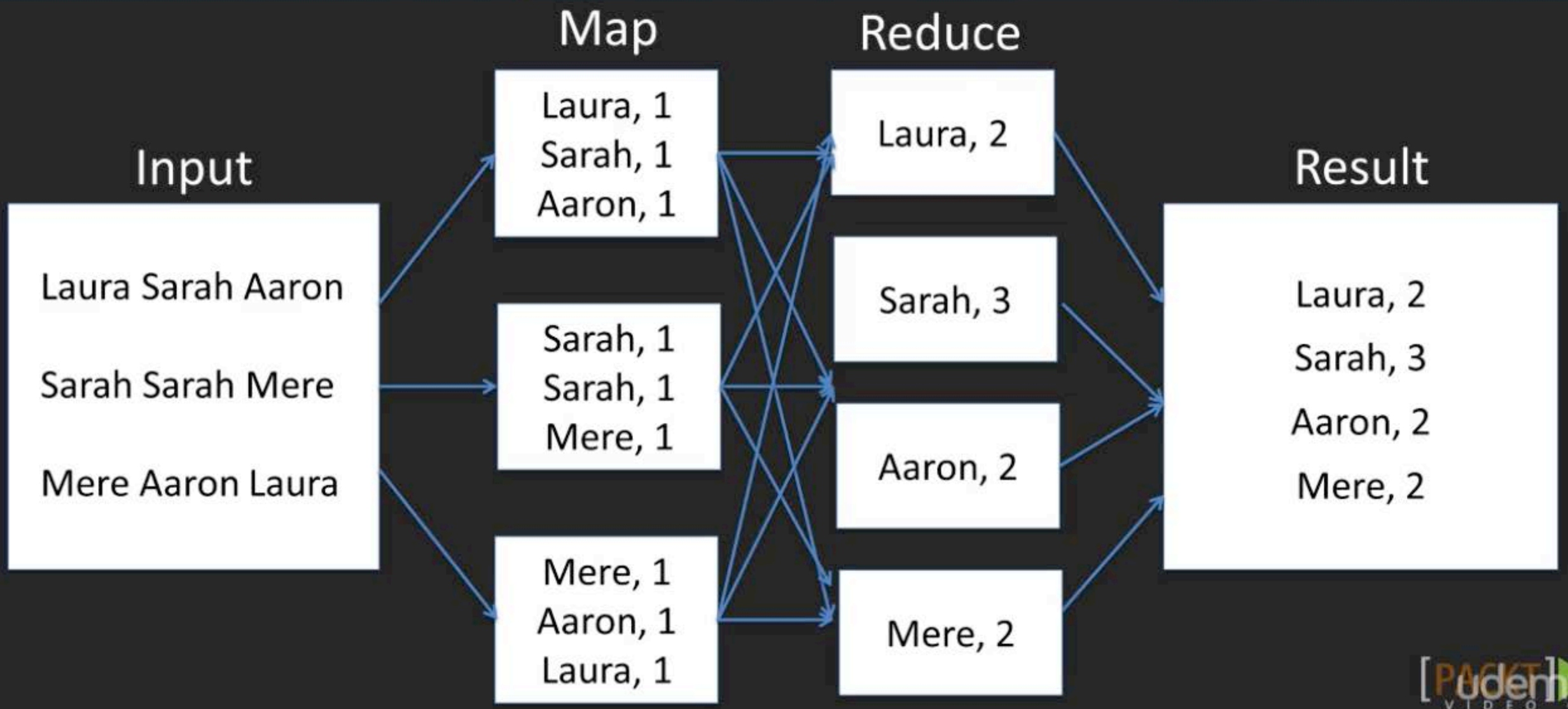
In this Video, we are going to take a look at...

- Introduction to MapReduce
- How MapReduce works (word count example)

Introduction to MapReduce

- Created by Yahoo! from a paper by Google
- Computation engine
- Coded in Java
- Two important functions: Map and Reduce

How MapReduce Works: Word Count



Creating a MapReduce Program

- Create Java code
  - Import packages
  - Map: "Tokenize" words, create key/value pairs
  - Reduce: Sum instances of each word from all lines, create new key/values
  - Results in about 65 lines of code
- Compile and create .jar from Java code
- Add .jar to repository

Next Video

Overview of Pig

Learning Hadoop 2

Randal Scott King

Video 1.5

Overview of Pig

In this Video, we are going to take a look at...

- Introduction to Pig
- How Pig works (word count example)

Introduction to Pig

- Developed by Yahoo! shortly after MapReduce
- Dataflow scripting language
- Builds MapReduce programs from scripts
- User Definable Functions (UDFs)

How Pig Works? (word count)

```
input = LOAD '/path/to/file/' AS(line:Chararray);
words = FOREACH input GENERATE FLATTEN(TOKENIZE(line,' ')) AS word;
grouped = GROUP words BY word;
wordcount = FOREACH grouped GENERATE group, COUNT(words);
dump wordcount;
```

The same program in Java is about 65 lines.

Next Video

Overview of Hive

Learning Hadoop 2

Randal Scott King

Video 1.6

Overview of Hive

In this Video, we are going to take a look at...

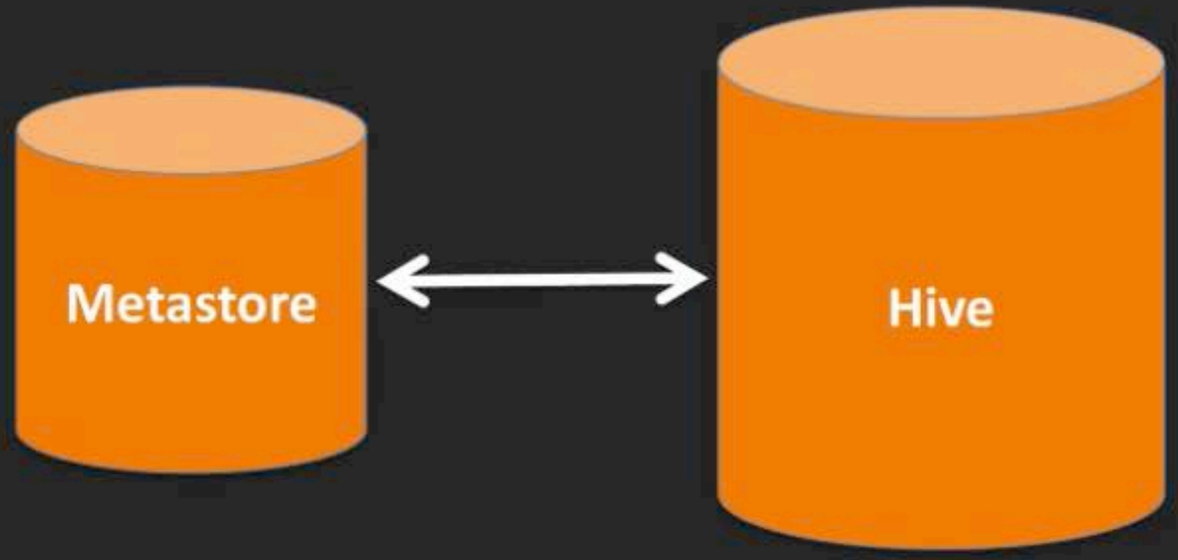
- Introduction to Hive
- Hive and metadata
- Internal versus external tables
- HiveQL

Introduction to Hive

- Data warehousing solution for Hadoop
- Uses tables, just like a traditional database
- HiveQL – SQL-ish query language
- Schema on load
- Uses MapReduce as the engine

Hive and Metadata

- Metadata – data that describes data
- Derby
- Derby versus MySQL

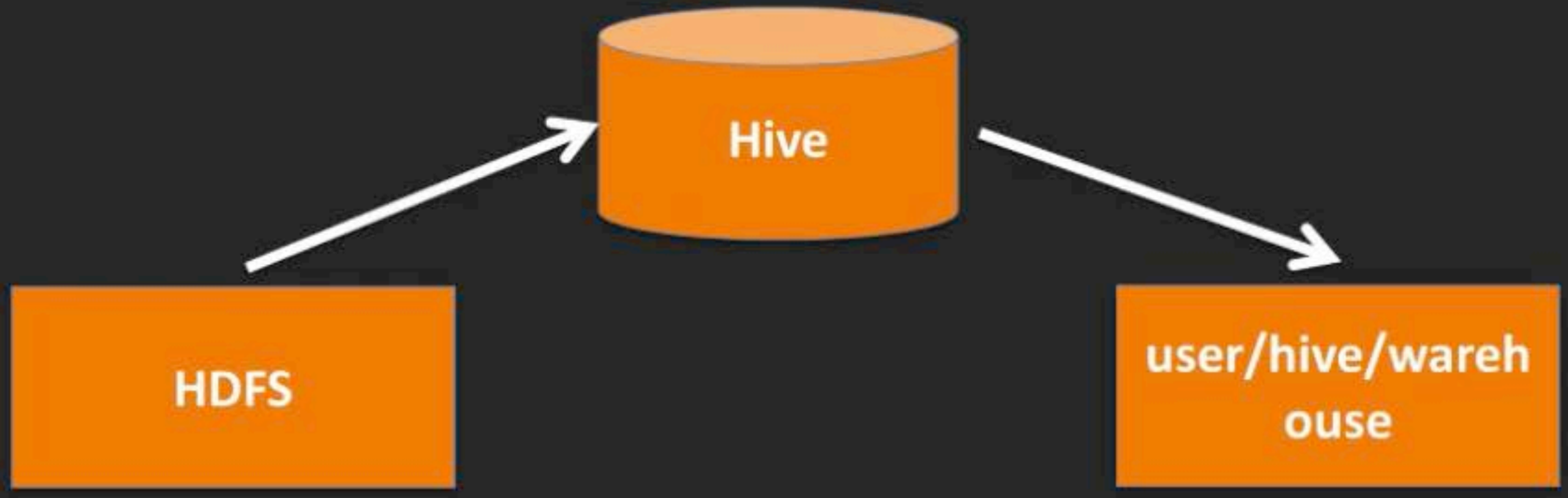


Internal Versus External Tables

- Internal Table (default)
  - Stores data in user/hive/warehouse

Internal Versus External Tables

- Internal Table (default)
  - Stores data in user/hive/warehouse
  - Not accessible to other applications
  - Dropping table deletes data and metadata



Internal Versus External Tables

- External Table
  - Data stays on HDFS
  - Accessible to other applications
  - Dropping table deletes only the metadata

HiveQL

HiveQL	SQL
CREATE DATABASE name;	CREATE DATABASE name;
SELECT * FROM table;	SELECT * FROM table;
SELECT COUNT (*) FROM table;	SELECT COUNT (*) FROM table;

HiveQL

HiveQL

- Default Join is "equi"
- Left Outer Join

SQL

- Default Join is "inner"
- Left Join

Summary

- HDFS and YARN – File system and resource scheduler
- Sqoop and Flume – Automated data import/export
- Mapreduce – The original compute engine of Hadoop
- Pig – Scripting language
- Hive – Data warehousing

Next Section

Installing and Configuring Hadoop