Unit 1.0

Hadoop 101 - Part 1

Admin Items

Before we start: Questions

1. Have you completed the setup?
   Setup the HortonWorks VM.

2. Are you ready to learn about Map/Reduce and the Hadoop ecosystem?

What we’ll do today

1. Setup Hortonworks
2. Learn about SQL and NoSQL datastores
3. Unix and HDFS Commands
4. Introduction to Hadoop
5. Introducing Map/Reduce
6. Simple word count Map/Reduce with Java
7. Using Hive
8. Using Pig

Today’s Class!

Objectives

In today's class we'll be discussing:

- SQL versus NOSQL
- Unix and HDFS Commands
- What is Hadoop?
- What is Map/Reduce?
- Other applications in the Hadoop ecosystem
- Labs!

How Big is Big Data?

Big Big Data

Given that companies data storage needs are growing exponentially, traditional databases are no longer able to meet the challenge.

- Don't send a terabyte to the algorithms, send the algorithms to the terabyte!

Warmup Activity

> YOUR TURN!!

Unix and HDFS File Systems:

1. Work through Unix File System Labs 1-4. (5 mins)
2. Work through Lab Hadoop Commands (5 mins)
3. With a partner, discuss the similarities and differences between the labs. (5 mins)

What is Map/Reduce?

The Challenge

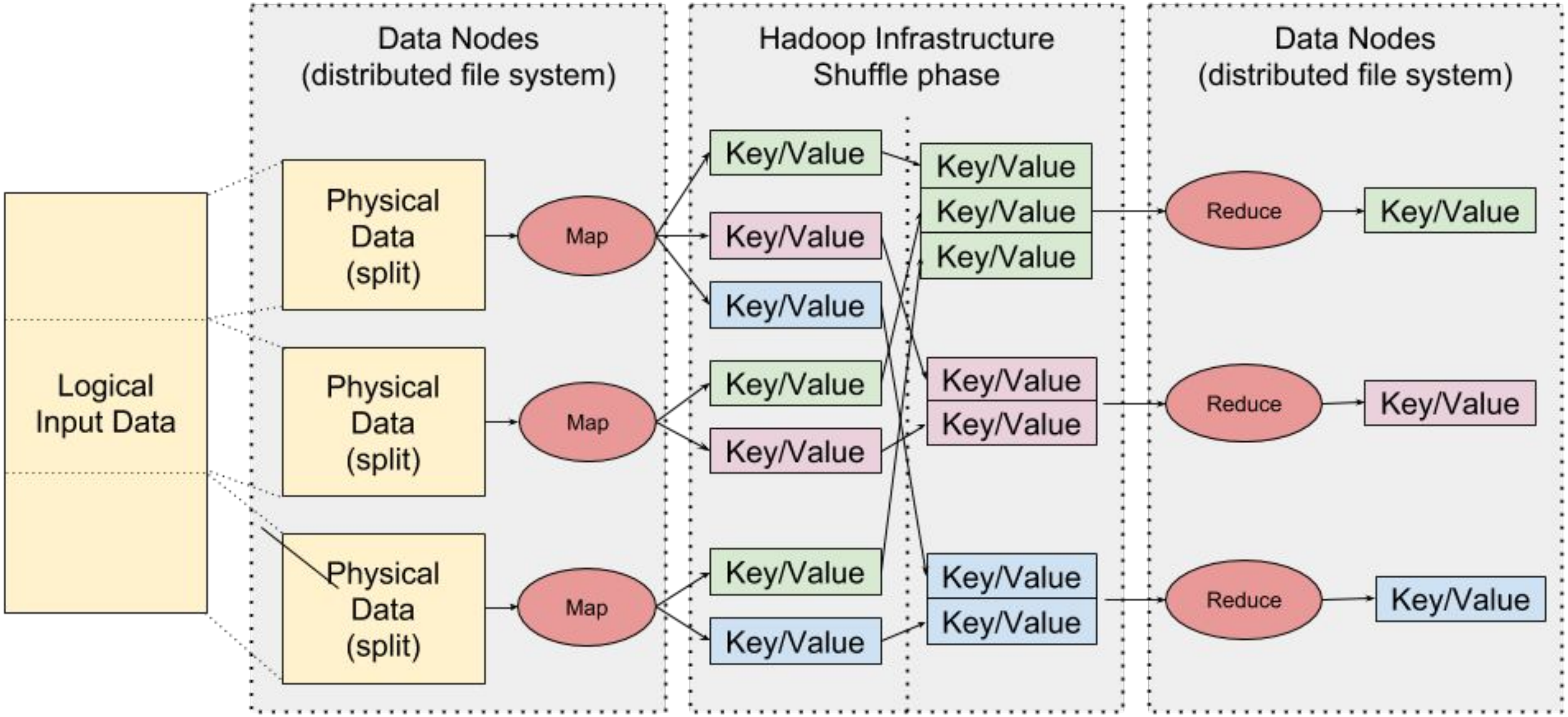
- We've seen how Hadoop stores data on multiple machines. We want to be able to process the data using the computational power of the machines that we use to store the data. How can we do that?

- To take advantage of a distributed cluster of computational nodes, we need to come up with algorithms that allows us to run things in parallel. Also, we would want to make sure that we delegate the processing to the node that actually have the data if possible.

Big Data Needs a New Approach

- Don't send a terabyte to the algorithms, send the algorithms to the terabyte!

Map/Reduce


Logical Input Data
Data Nodes (distributed file system)
Physical Data (split)
Physical Data (split)
Physical Data (split)
Map
Map
Map
Hadoop Infrastructure Shuffle phase
Key/Value
Key/Value
Key/Value
Key/Value
Key/Value
Key/Value
Key/Value
Key/Value
Key/Value
Key/Value
Key/Value
Key/Value
Key/Value
Key/Value
Data Nodes (distributed file system)
Reduce
Reduce
Reduce
Key/Value
Key/Value
Key/Value

> YOUR TURN!!

Simple Map Reduce in Java:

1. Download the file sent to you via slack.
2. In the VM terminal, run the following command.

   ```
   [hdfs@sandbox lesson-160]$ yarn jar wordcount.jar
   com.example.WordCount /tmp/a-tale-of-two-cities.txt
   /tmp/lesson-160
   ```

3. With a partner, open the Java file and discuss the Mapper and Reducer inner classes.

Hive and Pig

Demo Time

Instructor: Demo

(Lab1 WordCount)

What is Hive?

Hive is a data warehouse system built on top of Hadoop.

Hive is great because it offers:

- Easy data summarization
- Ad-hoc queries
- Analysis of large datasets stored in various databases and file systems.

With Hive, you can apply structure to large amounts of unstructured data and then perform batch SQL-like queries on that data.

> YOUR TURN!!

Hive Query:

1. Download the file sent to you via slack.
2. Run Lab1WordCount using instructions in the file.
3. With a partner, review the instructions for Advanced Labs 1-4. Try one if time permits.

What is Pig?

If you want to run complex data transformations in Hadoop but you don't know how to write Java, Apache Pig is for you.

A high level scripting language, Pig enables you to build large, complex apps that use data from structured and unstructured data, and store the results in the Hadoop File System.

> YOUR TURN!!

Pig Latin:

1. Download the file sent to you via slack.
2. Run Lab1HelloWorld using instructions in the file.
3. With a partner, review the instructions for Advanced Labs 1-5. Try one if time permits.

Questions?