SQL on Hadoop - Analyzing Big Data with Hive

Ahmad Alkilani
www.pluralsight.com

pluralsight
hardcore developer training

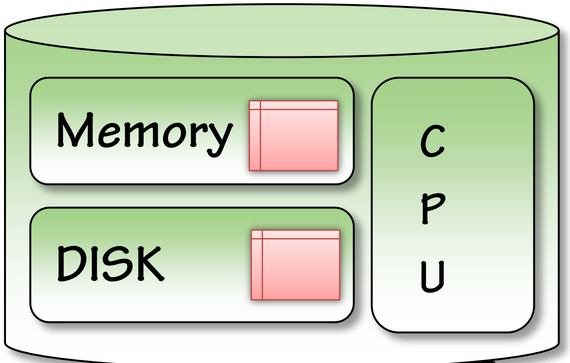
Introduction to Hadoop

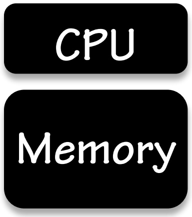
Ahmad Alkilani
www.pluralsight.com

Outline

- Why Hadoop? Motivation
- Hadoop architecture and distributed computing
- HDFS
- MapReduce
- Getting up and running

Motivation for Hadoop

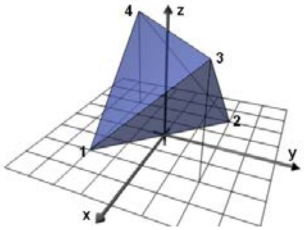


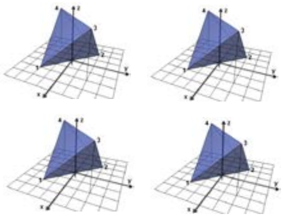


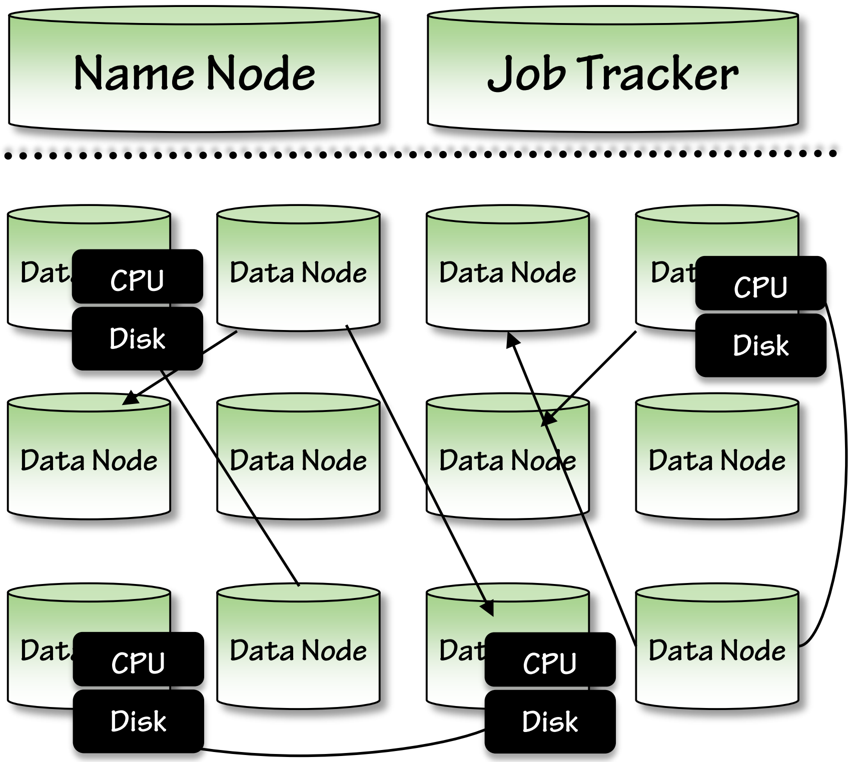
- ~40 Billion Web Pages x 30 KB each = Petabyte
- Today’s average disk speed reads about 120 MB/sec
- Little over 3 months to read the web!
- Approximately 1,000 drives to store and use

Distributed Computing Challenges

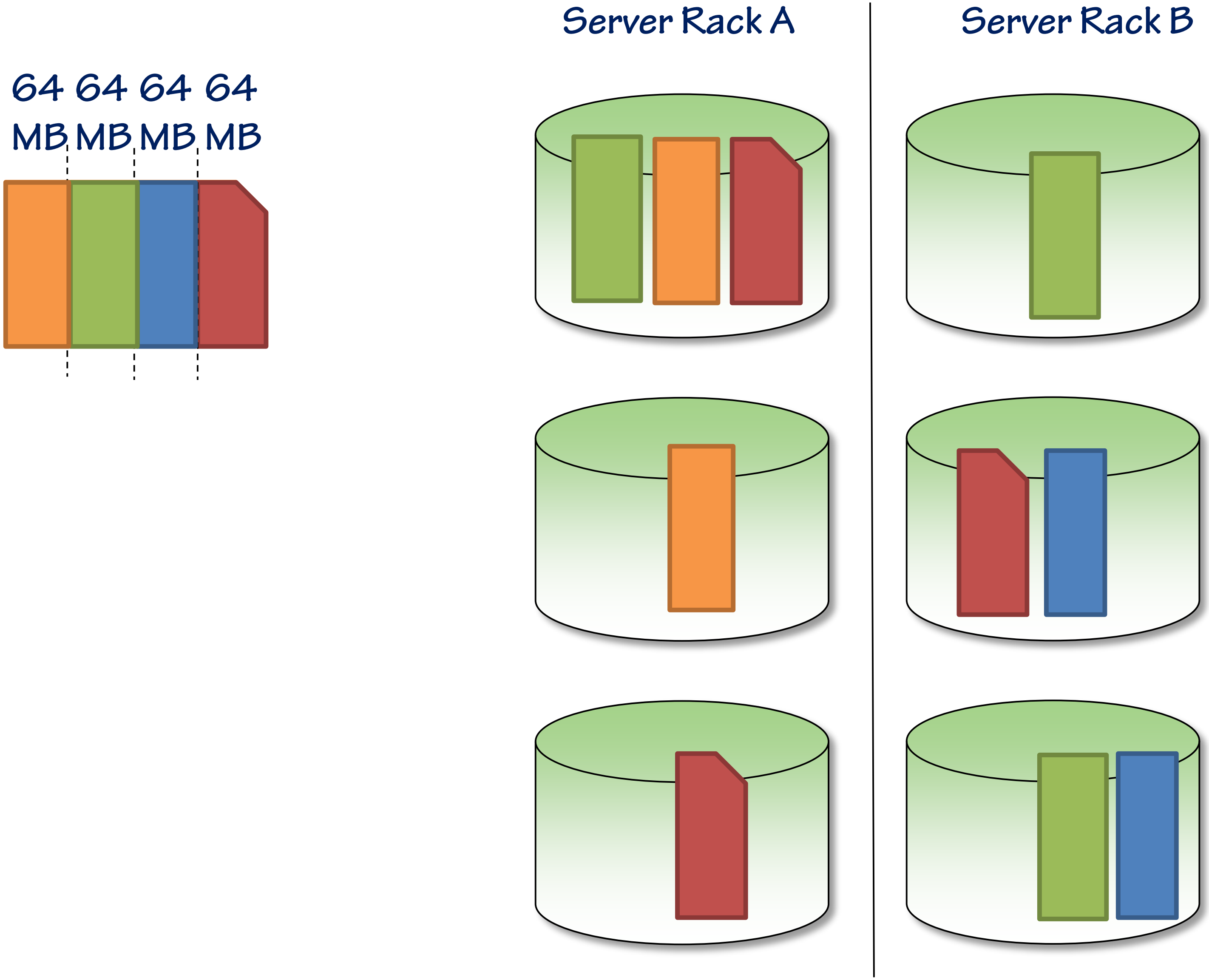
- Scale out with distributed computing
- Hadoop based on Google's implementation
- Volume, Velocity, and Variety
- Recover from failures
- Shared nothing architecture
- Hadoop file system (HDFS)
- MapReduce



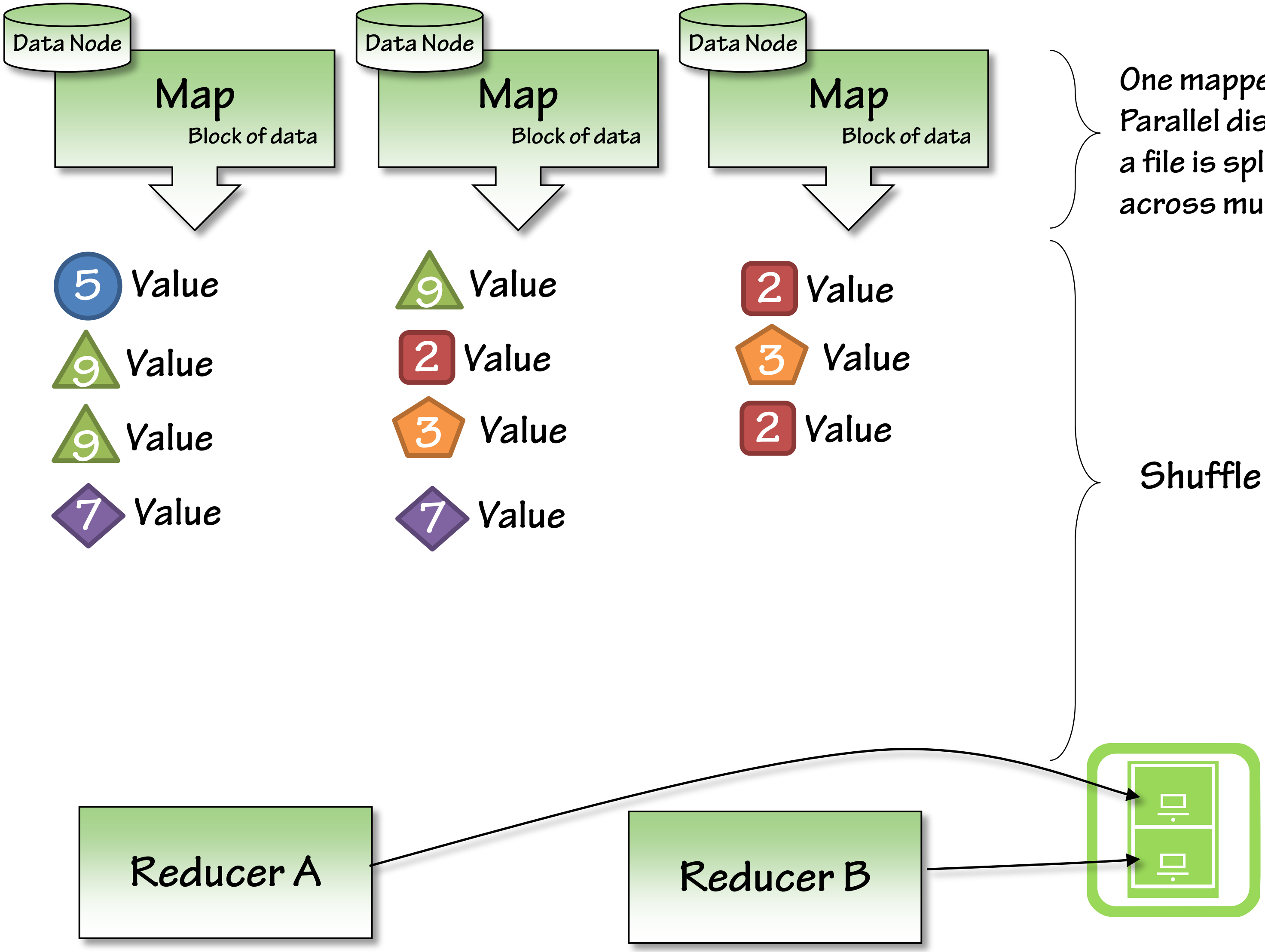




Hadoop File System (HDFS)



MapReduce

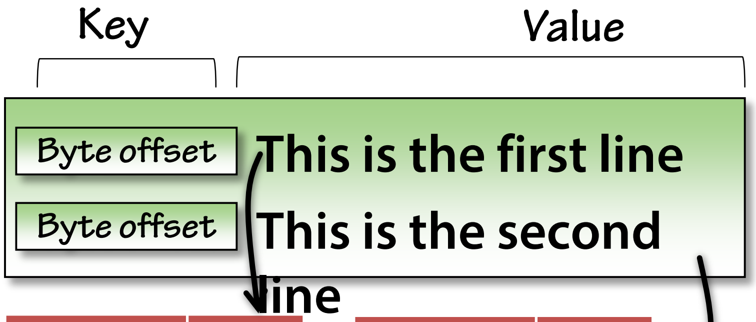


One mapper per block
Parallel distributed processing given
a file is split into blocks
across multiple servers.

Shuffle and Sort

Folder in HDFS

Word Count Example



Byte offset This is the first line

Byte offset This is the second line

```
String line = value.toString();
StringTokenizer tokenizer = new StringTokenizer(line);
while (tokenizer.hasMoreTokens())
{
  word.set(tokenizer.nextToken());
  context.write(word, one);
}
```

Key	Value
This	1
is	1
the	1
first	1
line	1

Key	Value
This	1
is	1
the	1
second	1
line	1

Key	Value
This	1
This	1
the	1
the	1
second	1
first	1

Key	Value
line	1
line	1
is	1
is	1

Reducer A

first	1
second	1
the	2
This	2

Reducer B

is	2
line	2

```
int sum = 0;
for (IntWritable val : values)
{ sum += val.get(); }
context.write(key, new IntWritable(sum));
```

Basic commands using HDFS

Hadoop Demo

Environment Setup

- **Course focus is on development**
- **Use a Virtual Machine image to follow along with examples**
- **Pseudo distributed sandbox**
  - Replication factor set to 1
  - Name Node, Job Tracker, Data Node, and Task Tracker on a single machine
- **Demos using Hortonworks' HDP sandbox**
  - Hive 0.10, 0.11 and above

Summary

- **Distributed computing and scaling out to solve big data problems**
- **Key system characteristics**
  - Built to handle failures
  - Move processing to the data
  - Failures are inevitable. Embracing this allows for solutions built on commodity servers
- **MapReduce**
  - Mapper assigned to each block of data
  - Key-value pairs are both the input to and output of each phase
  - Keys must implement WritableComparable interface
  - Shuffle and Sort plays a key role in solving problem