University of South Australia

Apache Spark and why it's so fast

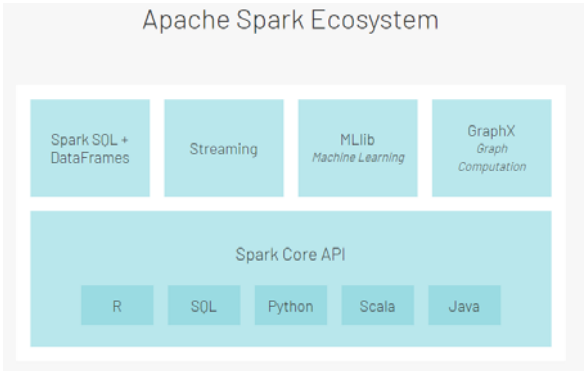
1

What is

MapReduce	Spark	Computing engine
YARN		Resource Manager
HDFS		Storage

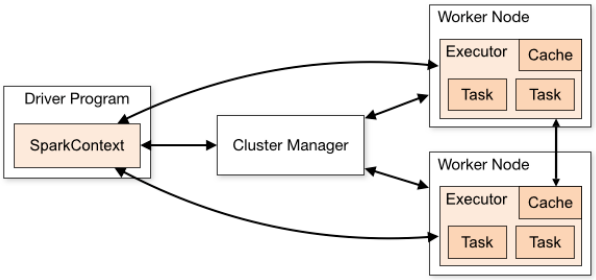
2

Spark is a bit of an



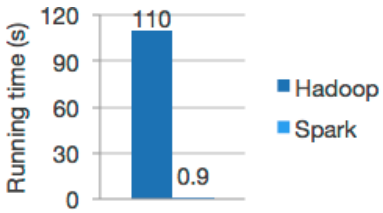
In Spark we have a choice between 3 main cluster managers:

- Spark Standalone Manager
- YARN
- MESOS



3

But why Spark?



Logistic regression in Hadoop and Spark

4

Analogy: Think of your computer as your home office

You're sitting down to do your next assignment.

You have a desk, on top of which is a calculator and next to which is an old filing cabinet. You've got some kind of home-made pulley system for sending files from room to room.

The calculator is your **CPU**.

Data might be on your table, in the cabinet, or coming in on the pulley, but all calculations will end up being done on the calculator.

University of South Australia

5

Analogy: Think of your computer as your home office

You're sitting down to do your next assignment.

You have a desk, on top of which is a calculator and next to which is an old filing cabinet. You've got some kind of home-made pulley system for sending files from room to room.

The desk is your **RAM**.

Data is just being stored here temporarily in the form of pages of paper.

The size of your desk will determine how much data you quickly have access to.

University of South Australia

6

Analogy: Think of your computer as your home office

You're sitting down to do your next assignment.

You have a desk, on top of which is a calculator and next to which is an old filing cabinet. You've got some kind of home-made pulley system for sending files from room to room.

The filing cabinet is your **hard drive**.

It's kind of old and rusty, it can store a lot though!

Analogy: Think of your computer as your home office

You're sitting down to do your next assignment.

You have a desk, on top of which is a calculator and next to which is an old filing cabinet. You've got some kind of home-made pulley system for sending files from room to room.

The pulley system is the **network**.

For data that we don't have readily available it needs to be 'pulled' from the network. It's laborious and time consuming, but our office is only so big, and we can't have instant access to everything.

CPU

3.0 GHz = 3 billion operations per second.

1.2 million bytes per second

As a proportion of your CPU that's 1.2M/3.0B = 0.0004

And most modern machines are at least dual core! (2 CPU's)

RAM

Most of the time your CPU isn't actually processing data, it's waiting for data to arrive from memory.

250 times longer to find and load a byte from within your memory, than it will to process that byte in your CPU.

Storage in RAM is slower than the processing speed of a CPU, but it's still by far the best option currently available.

HDD/SSD

Process 1 hour of tweets
~ 4 gigabytes

30 ms

0.5 sec

4 sec

Network

Slow, avoid if possible.

Processing times can easily be increased by 20x when the data needs to be downloaded across a network first.

Recall rack awareness

9

10