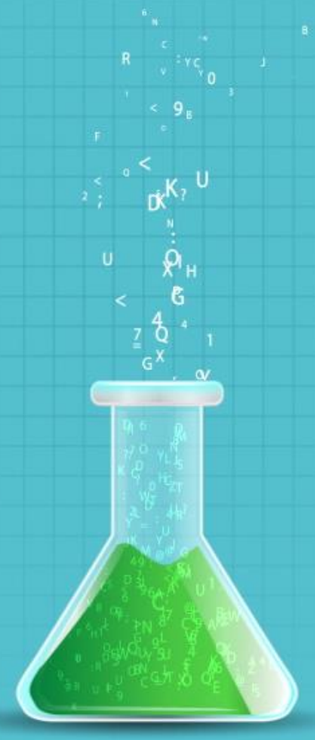
Data Science with Python

Lesson 12—Python Integration with Hadoop MapReduce and Spark

DATA SCIENCE

What You'll Learn

- Why Python should be integrated with Hadoop
- Brief overview of the ecosystem and architecture of Hadoop
- How MapReduce functions
- How Apache Spark functions and what its benefits are
- Write Python programs for Hadoop operations



Quick Recap: Need for Real Time Analytics

Demographics

Orders

Products

Visits/Channels

Social Media Sentiments

Customer Support and Surveys

Real Time Analytics

Improve Conversion

Targeted Marketing

Better Customer Satisfaction

Increased Sales

Quick Recap: Why Python

Data Scientists all over the world prefer to use Python for analytics because of its ease and support to carry out all the aspects of Data Science.

Data Science

Acquire

Wrangle

pandas

Explore

matplotlib

Model

scikit learn

NumPy

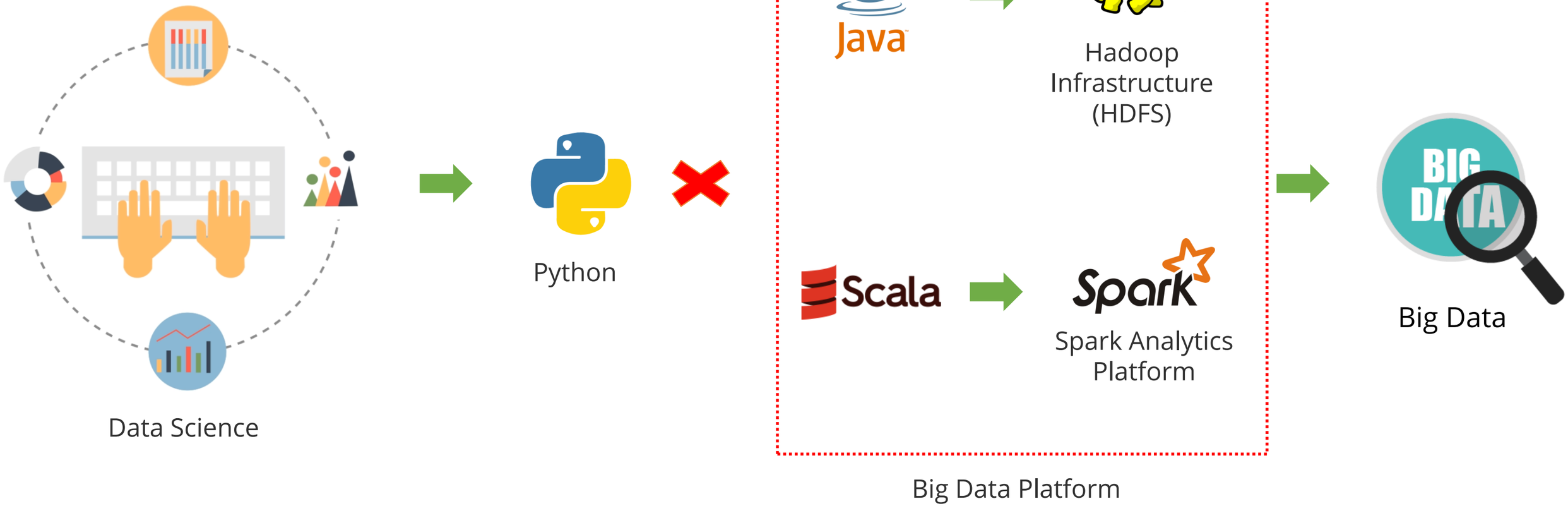
Visualize

Bokeh

SciPy

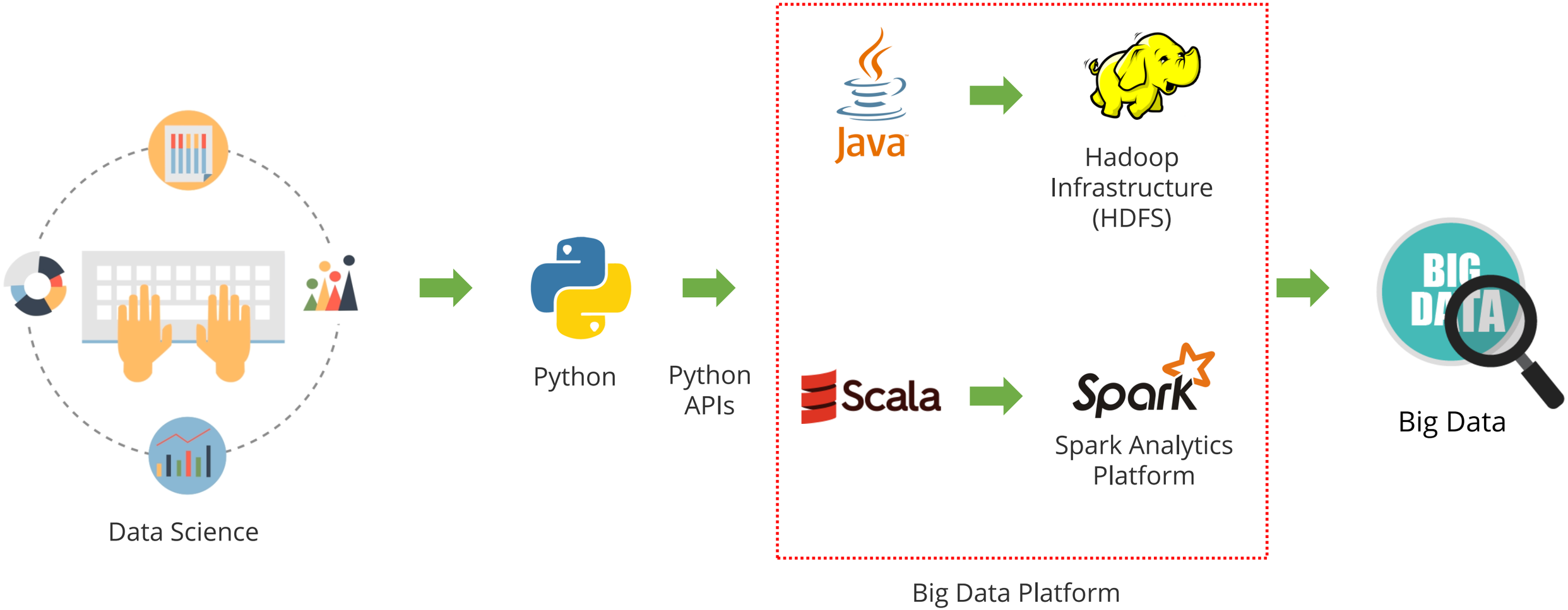
Disparity in Programming Languages

However, Big Data can only be accessed through Hadoop which is completely developed and implemented in Java. Also, analytics platforms are coded in different programming languages.



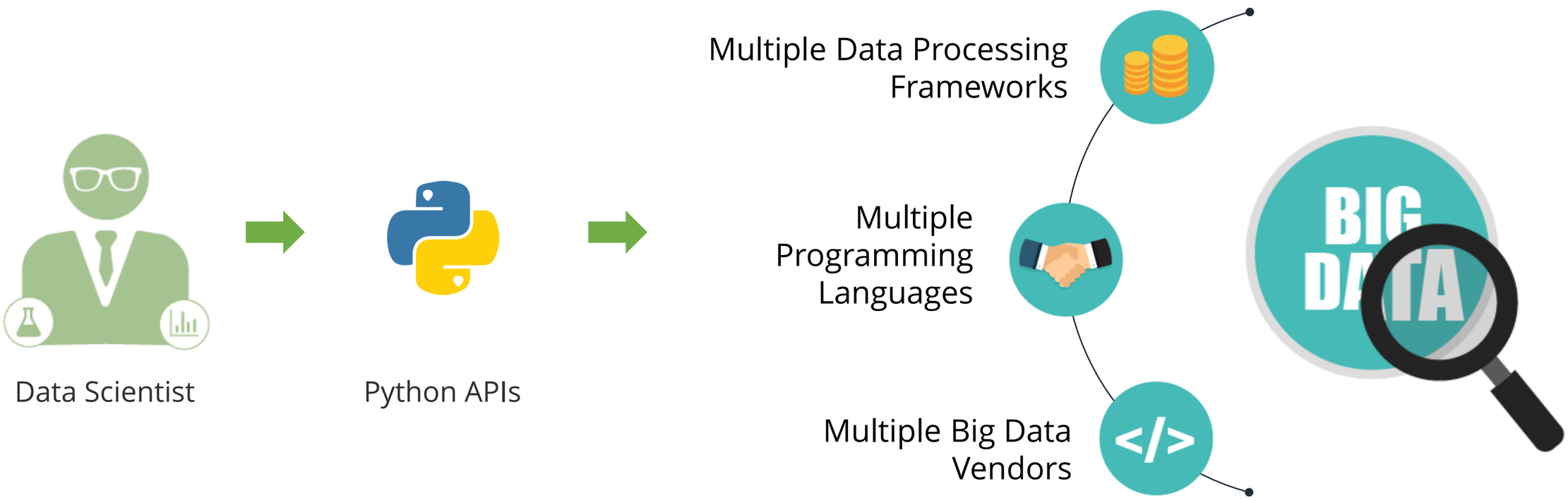
Integrating Python with Hadoop

But as Python is a Data Scientist's first language of choice, both Hadoop and Spark provide Python APIs that allow easy access to the Big Data platform.



Need for Big Data Solutions for Python

There are several reasons for creating Big Data solutions for Python.



Hadoop: Core Components

Hadoop

HDFS

(Hadoop Distributed File System)

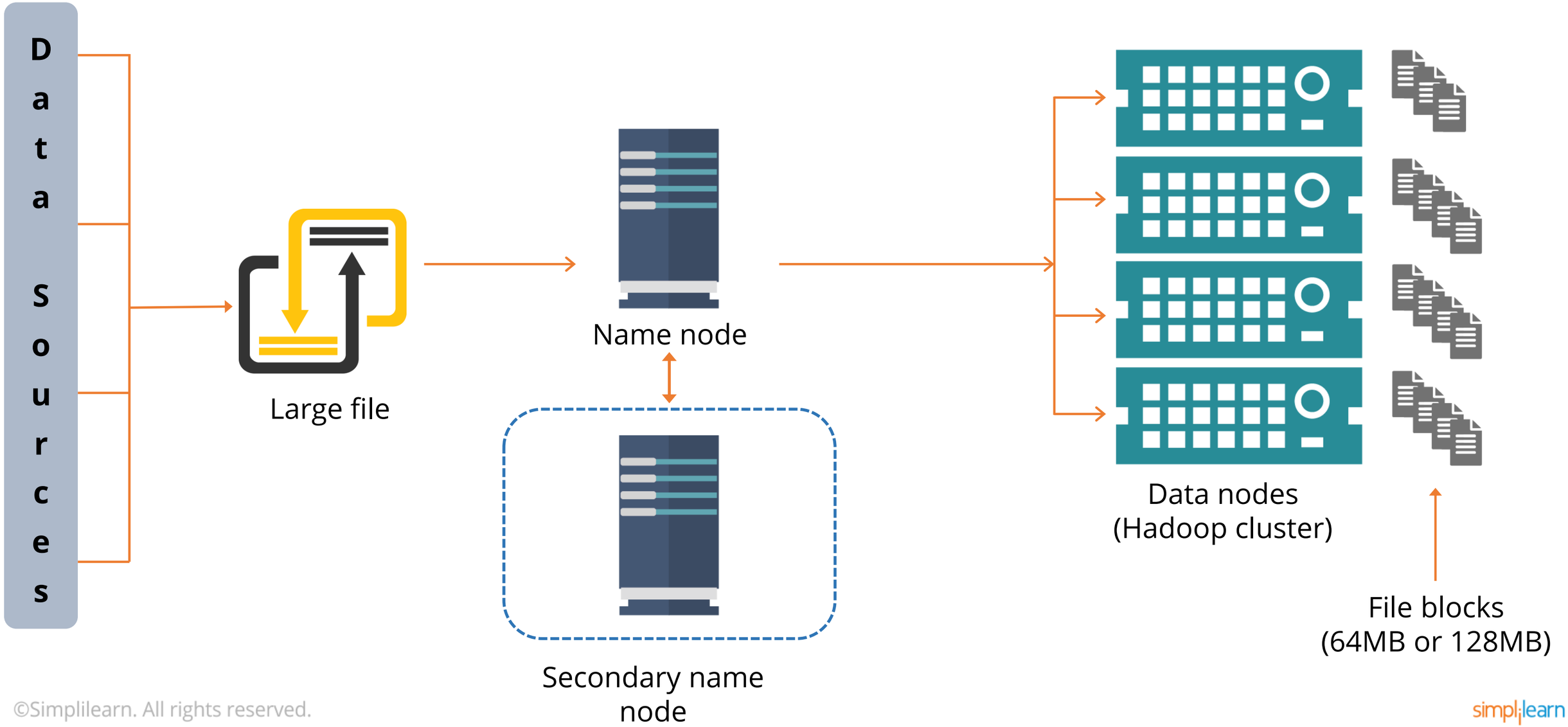
- It is responsible for storing data on a cluster
- Data is split into blocks and distributed across multiple nodes in a cluster
- Each block is replicated multiple times
  - Default is 3 times
  - Replicas are stored on different nodes

MapReduce

- MapReduce is a data processing framework to process data on the cluster
- Two consecutive phases: Map and Reduce
- Each map task operates on discrete portions of data
- After map, reduce works on the intermediate data distributed on nodes

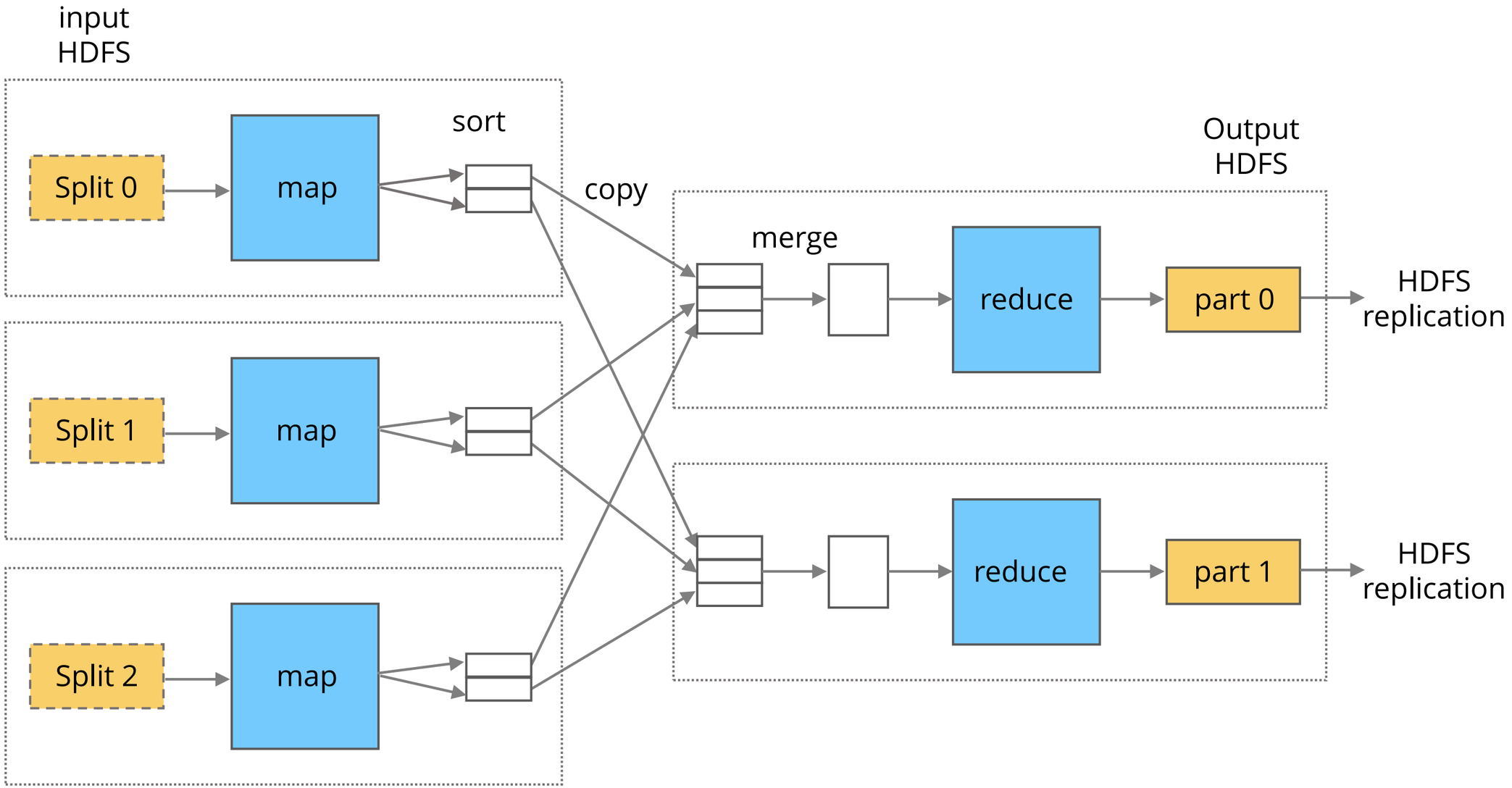
Hadoop: The System Architecture

This example illustrates the Hadoop system architecture and the ways to store data in a cluster.



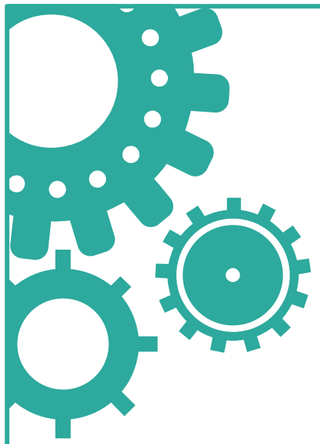
MapReduce

The second core component of Hadoop is MapReduce, the primary framework of the HDFS architecture.



MapReduce: The Mapper and Reducer

Let us discuss the MapReduce functions—mapper and reducer—in detail.



Mapper

- Mappers run locally on the data nodes to avoid the network traffic.
- Multiple mappers run in parallel processing a portion of the input data.
- The mapper reads data in the form of key-value pairs.
- If the mapper writes generates an output, it is written in the form of key-value pairs.

Reducer

- All intermediate values for a given intermediate key are combined together into a list and given to a reducer.
- This step is known as ‘shuffle and sort’.
- The reducer outputs either zero or more final key-value pairs. These are written to HDFS.

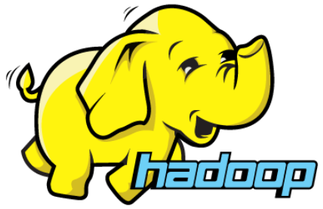
Hadoop Steaming: Python API for Hadoop

Hadoop Streaming acts like a bridge between your Python code and the Java-based HDFS, and lets you seamlessly access Hadoop clusters and execute MapReduce tasks.

Data Scientist

Hadoop Streaming

Python API



Mapper in Python

Python supports map and reduce operations:

Suppose you have list of numbers you want to square =
[1, 2, 3, 4, 5, 6]

Square function is written as follows:

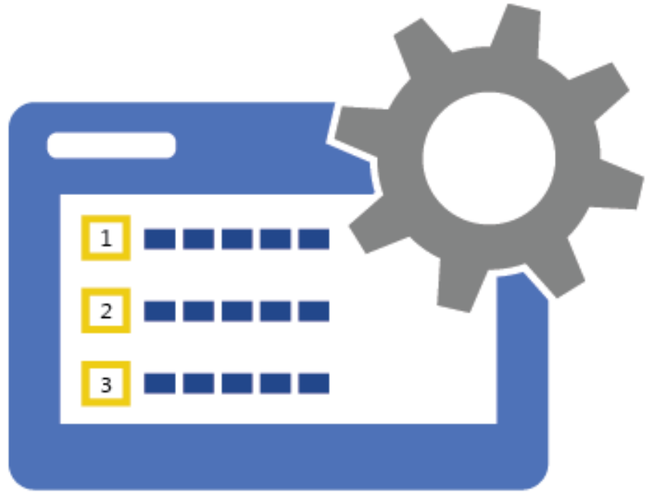
```
def square(num):
    return num * num
```

You can square this list using the following code:

```
squared_nums = map(square, numbers)
```

Output would be:

```
[1, 4, 9, 16, 25, 36]
```



Reducer in Python

Reduce written in Python:

Suppose you want to sum the squared numbers:

[1, 4, 9, 16, 25, 36]

Use the **sum** function to add two numbers

```
def sum(a, b ):
    return a + b
```

You can now sum the numbers using the **reduce** function

```
sum_squared = reduce(sum, squared_nums)
```

Output would be:

[91]

Cloudera QuickStart VM Set Up

Cloudera provides enterprise-ready Hadoop Big Data platform which supports Python as well.
To set up the Cloudera Hadoop environment, visit the Cloudera link:

http://www.cloudera.com/downloads/quickstart_vms/5-7.html

Cloudera recommends that you use 7-Zip to extract these files. To download and install it, visit the link:
http://www.7-zip.org/

Cloudera QuickStart VM: Prerequisites

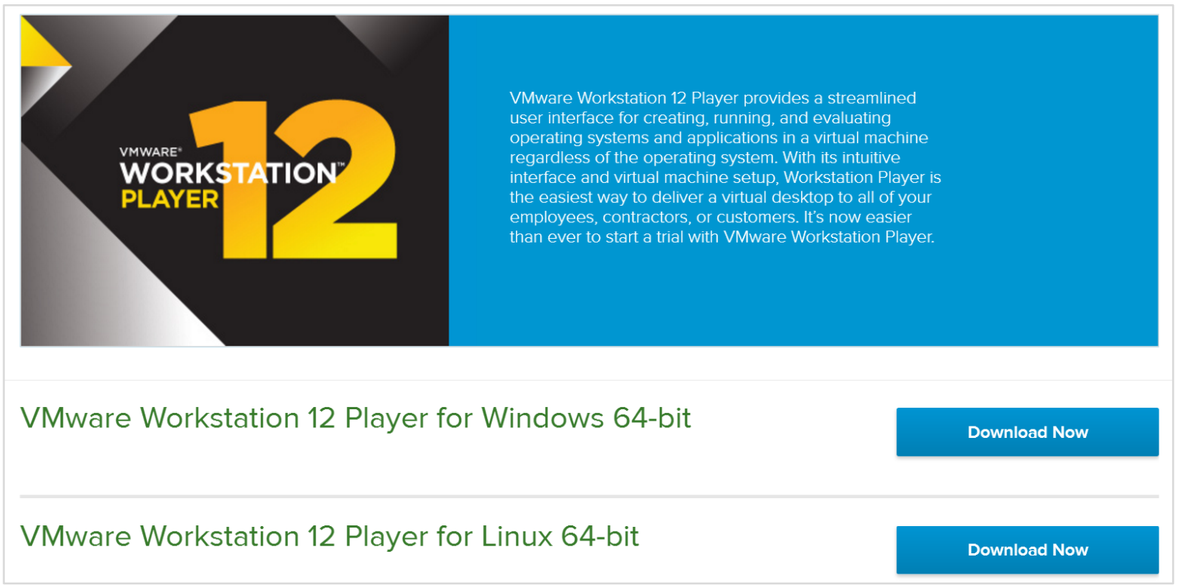
- These 64-bit VMs require a 64-bit host OS and a virtualization product that can support a 64-bit guest OS.
- To use a VMware VM, you must use a player compatible with WorkStation 8.x or higher:
  - Player 4.x or higher
  - Fusion 4.x or higher
- Older versions of WorkStation can be used to create a new VM using the same virtual disk (VMDK file), but some features in VMware Tools are not available.
- The amount of RAM required varies by the run-time option you choose

CDH and Cloudera Manager Version	RAM Required by VM
CDH 5 (default)	4+ GiB*
Cloudera Express	8+ GiB*
Cloudera Enterprise (trial)	10+ GiB*

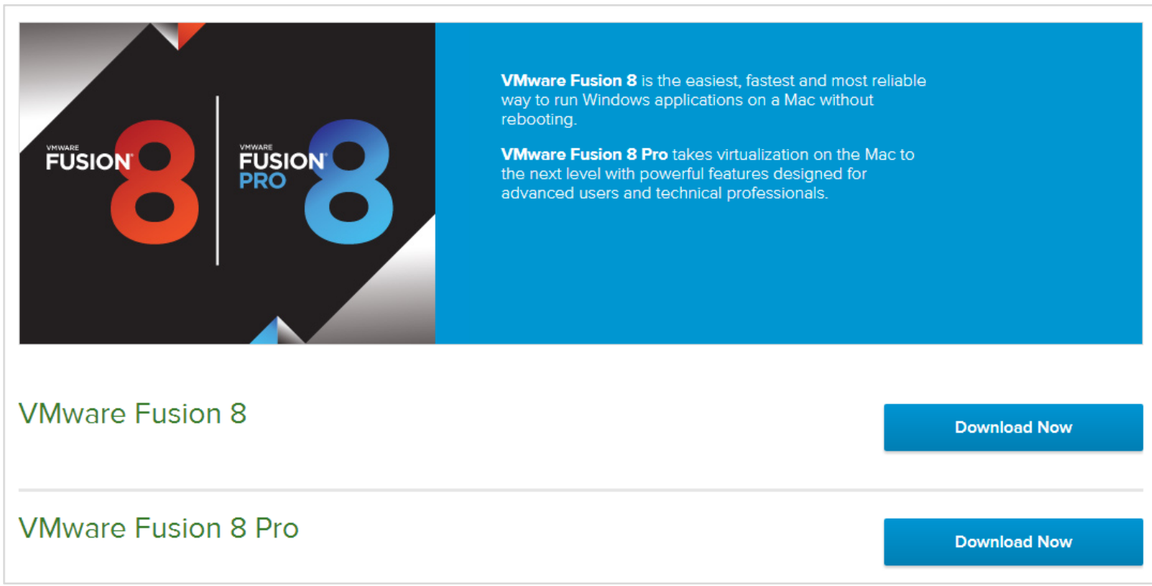
QuickStart VMware Player: Windows, Linux & VMware Fusion: Mac

To launch the VMware, visit the VMware link:

https://www.vmware.com/products/player/playerpro-evaluation.html

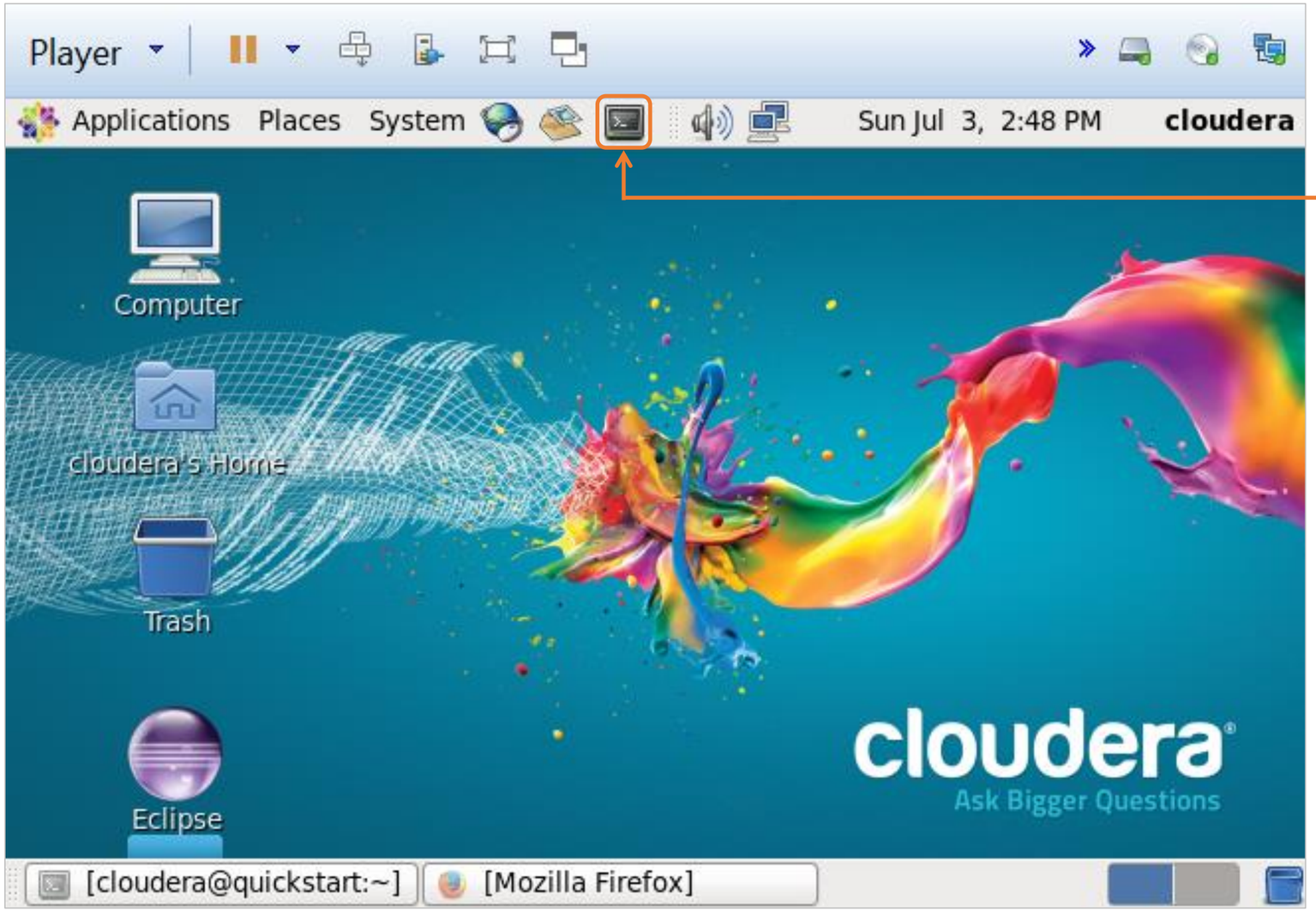


https://www.vmware.com/products/fusion/fusion-evaluation.html



QuickStart VMware Image

Launch VMware player with Cloudera VM



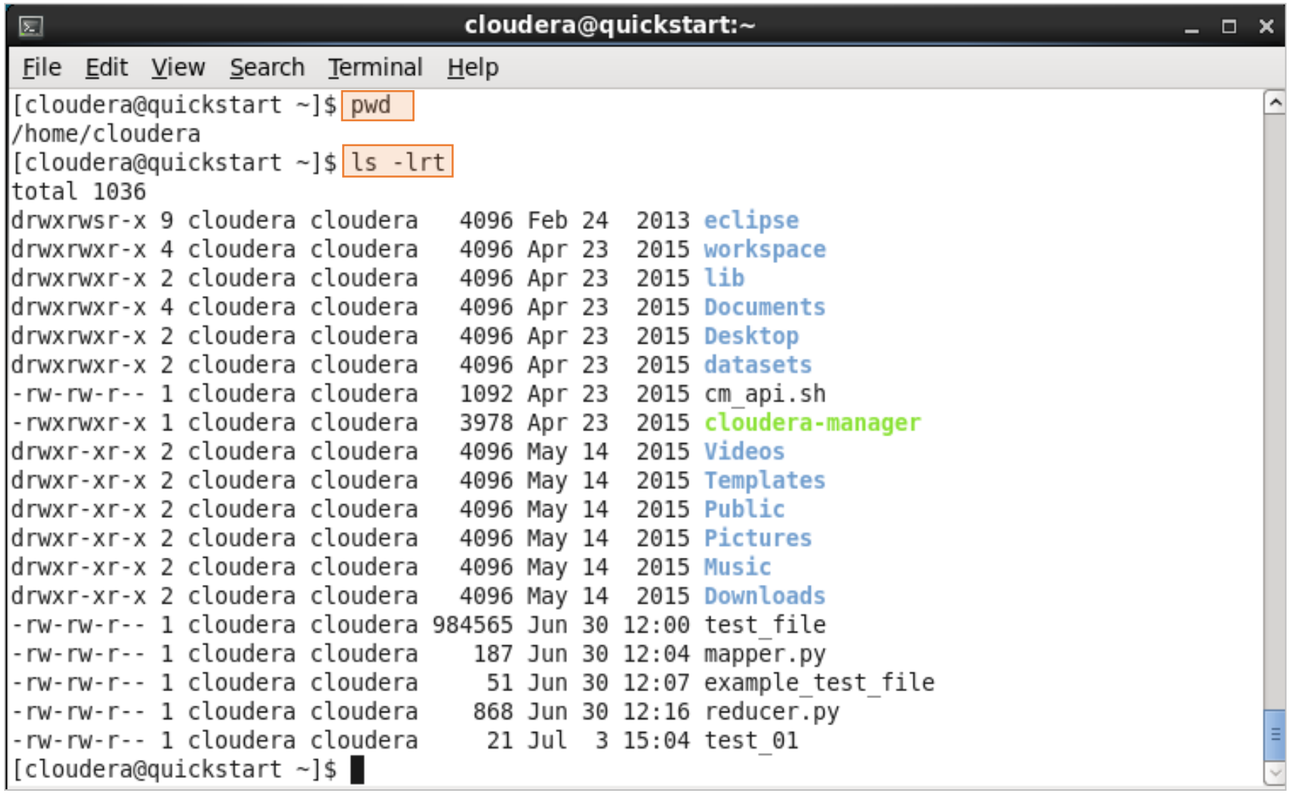
Launch Terminal

Account:
username: cloudera
password: cloudera

QuickStart VM Terminal

Step 01

Step 02



Unix command :

- pwd to verify present working directory
- ls -lrt to list files and directories

Demo 01—Using Hadoop Streaming for Calculating Word Count

Demonstrate how to create a MapReduce program and use Hadoop Streaming to determine the word count of a document

DATA SCIENCE

Knowledge Check

KNOWLEDGE CHECK

What is the usual size of the data block on HDFS?

a. 32 MB

b. 64 MB

c. 100 MB

d. 1 GB

KNOWLEDGE CHECK

What is the usual size of the data block on HDFS

a. 32 MB

b. 64 MB

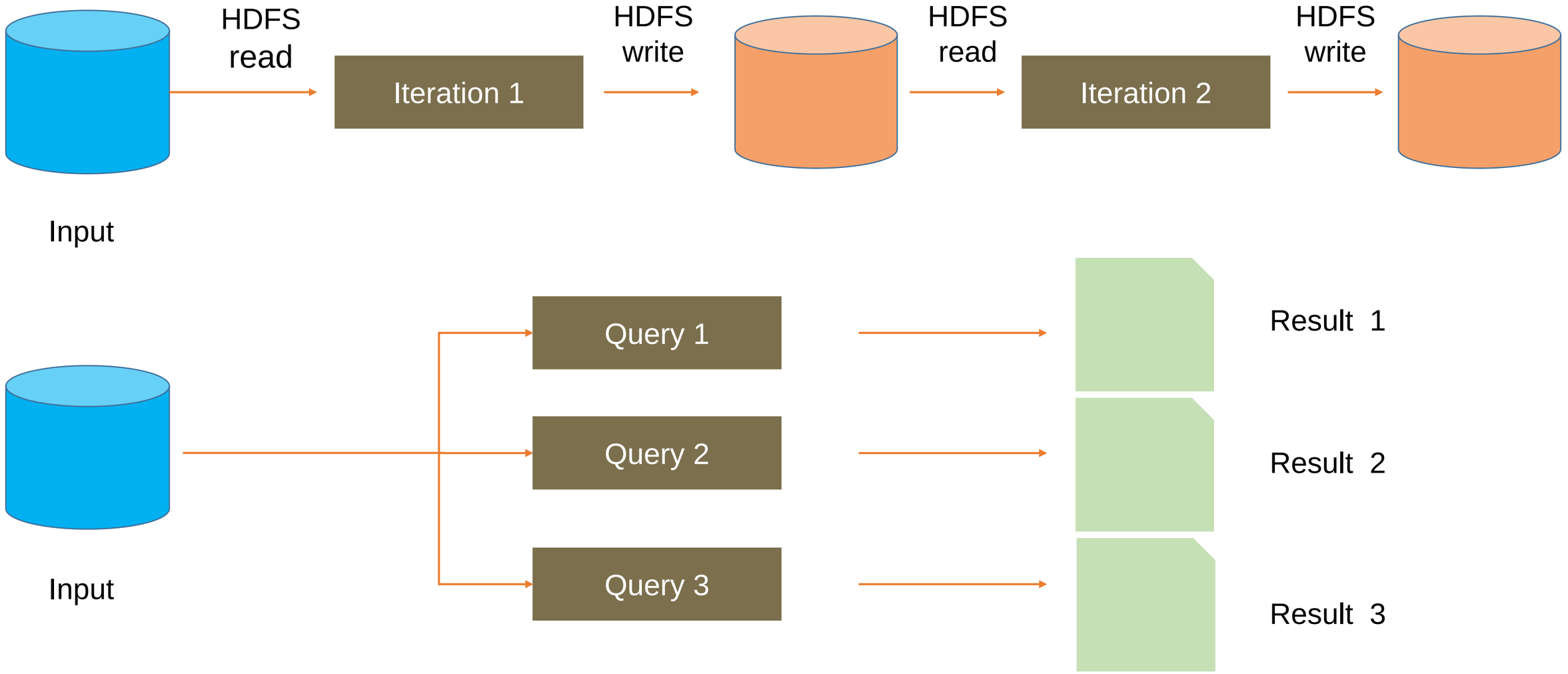
c. 100 MB

d. 1 GB

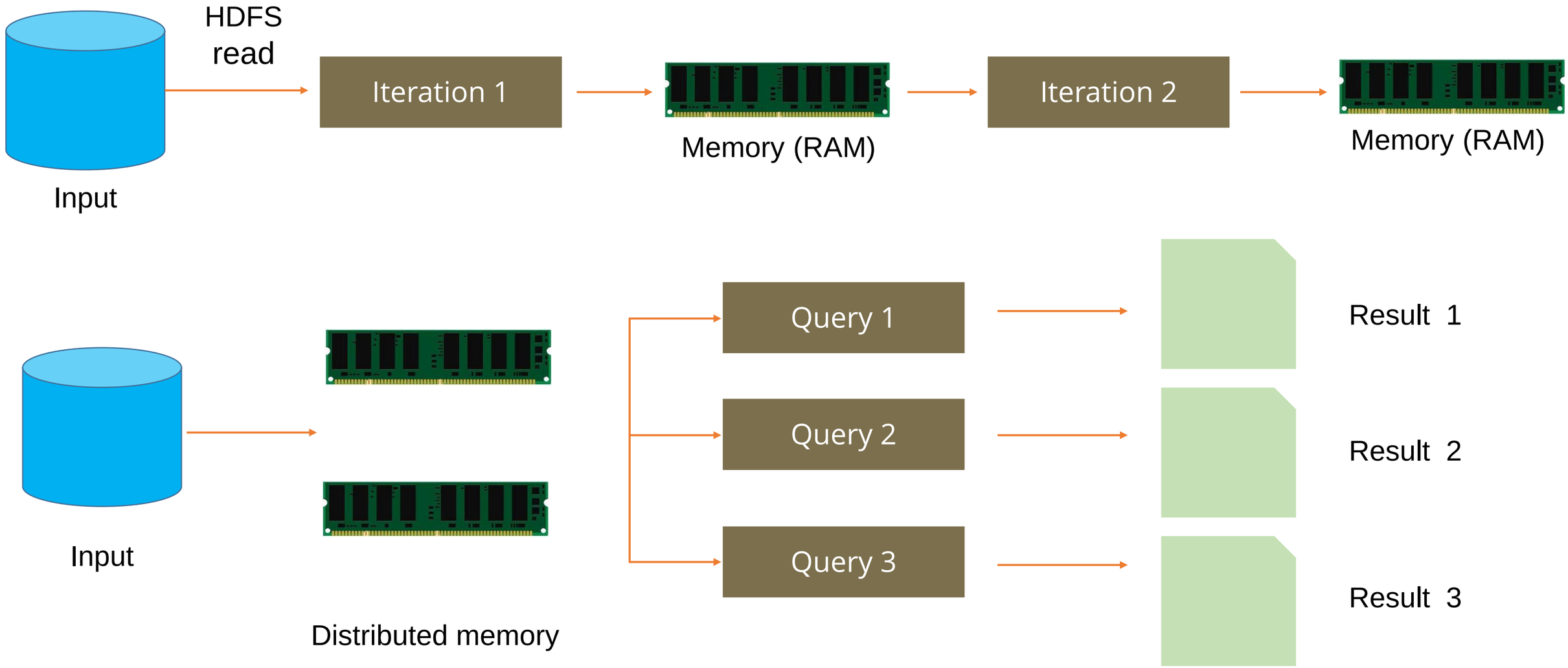
The correct answer is. b.

Explanation The usual data block size on HDFS is 64 MB.

MapReduce Uses Disk I/O Operations

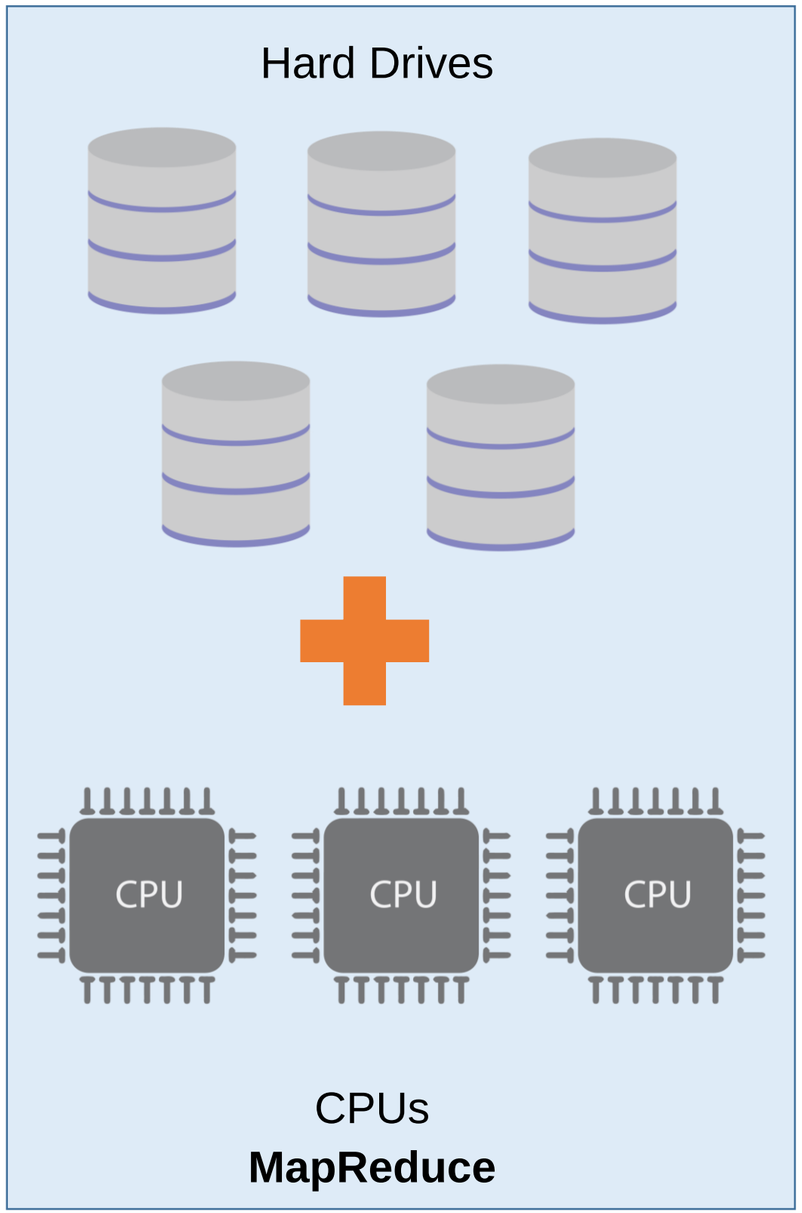


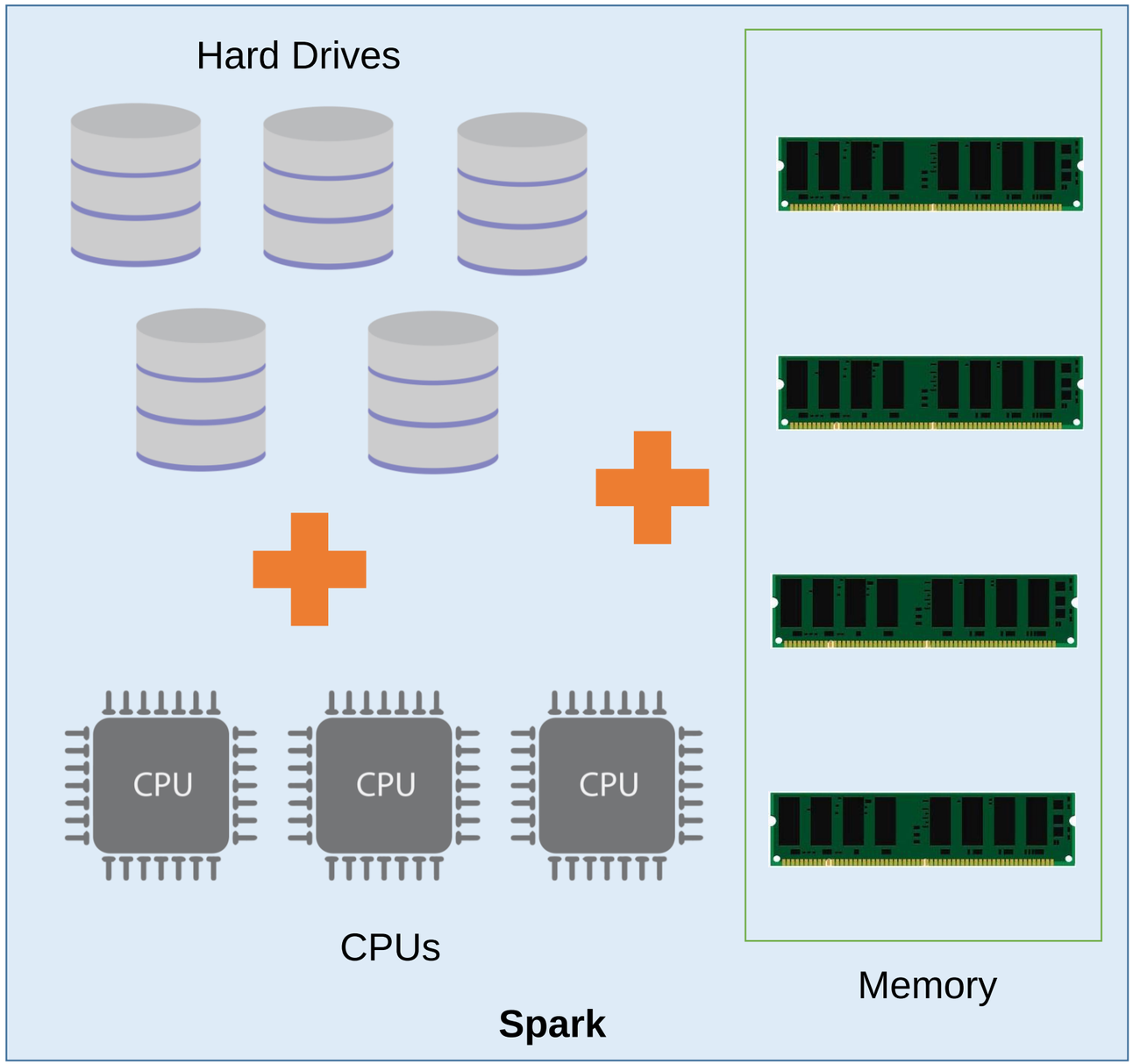
Apache Spark Uses In-Memory Instead of Disk I/O



10-100 X faster than network and disk

Hardware Requirements for MapReduce and Spark





Apache Spark Resilient Distributed Systems (RDD)

Some basic concepts about Resilient Distributed Datasets (RDD) are listed here:

- The main programming approach of Spark is RDD.
- They are fault-tolerant collections of objects spread across a cluster that you can operate on in parallel. They can automatically recover from machine failure.
- You can create an RDD either by copying the elements from an existing collection or by referencing a dataset stored externally.
- RDDs support two types of operations: transformations and actions.
  - Transformations use an existing dataset to create a new one.
    - Example: Map, filter, join
  - Actions compute on the dataset and return the value to the driver program.
    - Example: Reduce, count, collect, save

If the available memory is insufficient, then the data is written to disk.

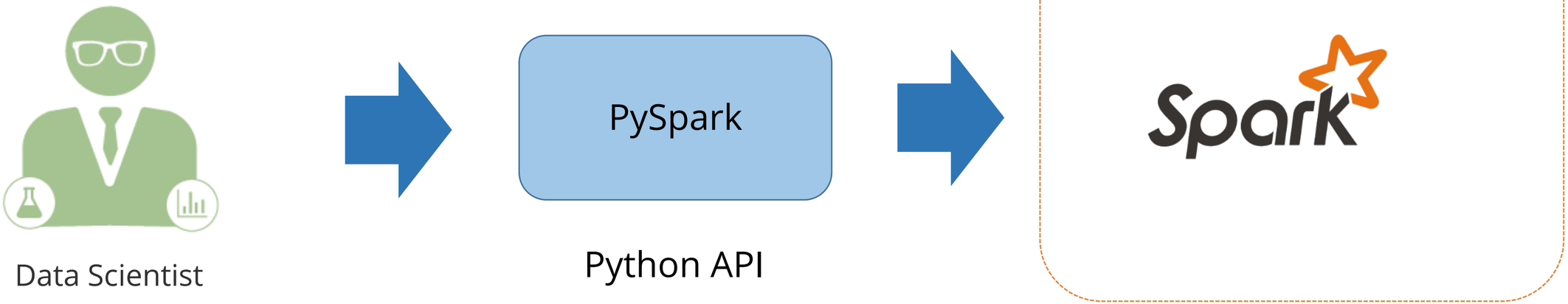
Advantages of Spark

Listed here are some of the advantages of using Spark:

Faster:	10 to 100 times faster than Hadoop MapReduce
Simplified:	• Simple data processing framework • Interactive APIs for Python for faster application development
Efficient:	Has multiple tools for complex analytics operations
Integrated:	Can be easily integrated with existing Hadoop infrastructure

PySpark : Python API for Spark

PySpark is the Spark Python API which enables data scientists to access Spark programming model



PySpark : RDD Transformations and Actions

Transformation

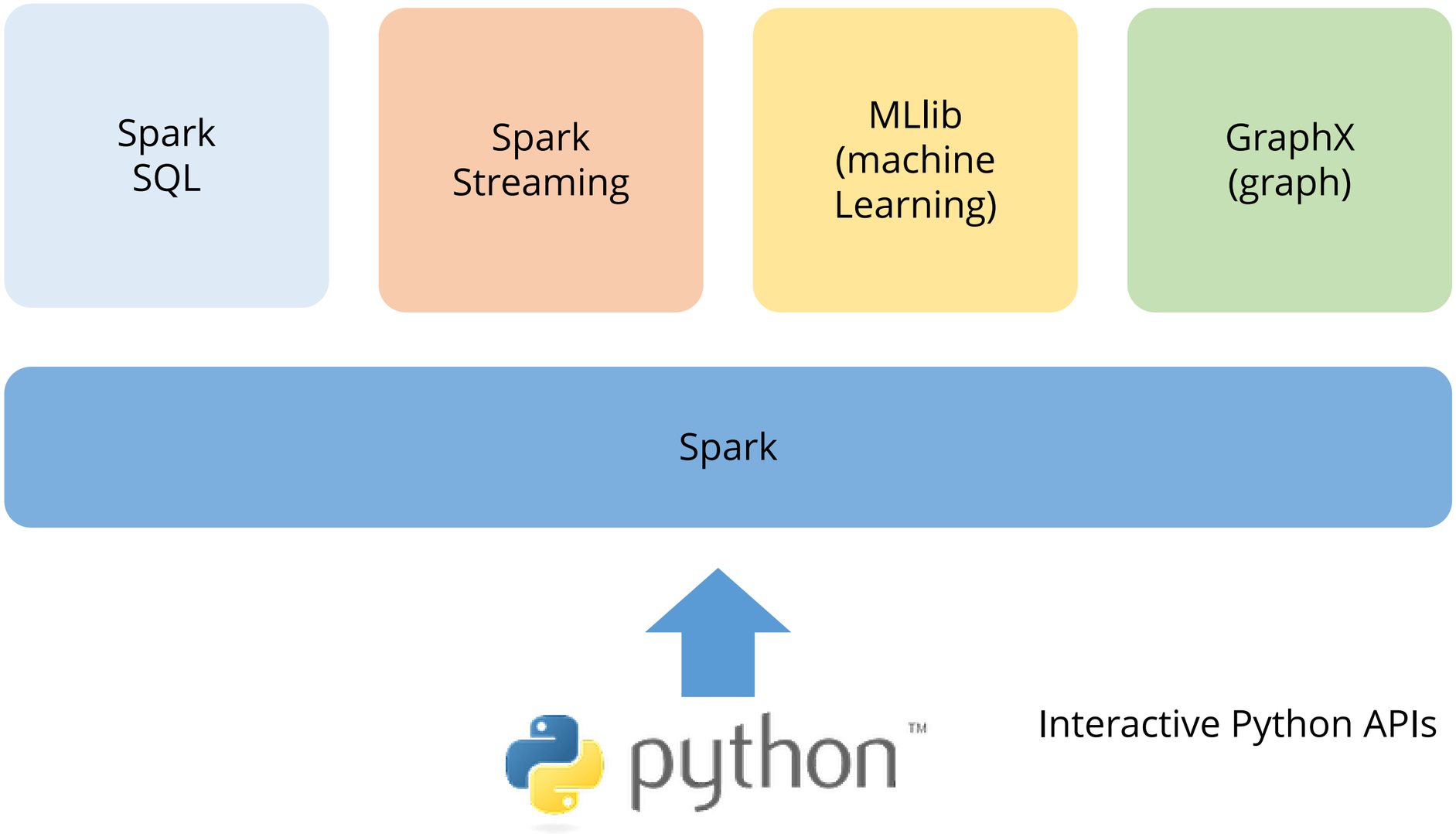
Transformation	Description
map()	Returns RDD, formed by passing data element of the source
filter()	Returns RDD based on selection
flatMap()	Maps items present in the dataset and returns sequence
reduceByKey()	Returns key value pairs where values for which each key is aggregated by value

Action

Action	Description
collect()	Returns all elements of the dataset as an array
count()	Returns the number of elements present in the dataset
first()	Returns the first element in the dataset
take(n)	Returns number of elements (n) as specified by the number in the parenthesis

SparkContext or SC is the entry point to spark for the spark application

Spark Tools

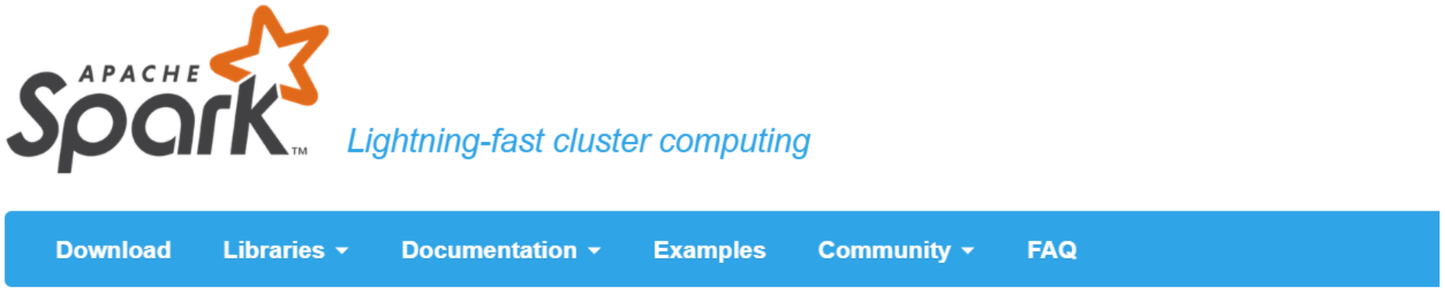


Apache Spark Set Up

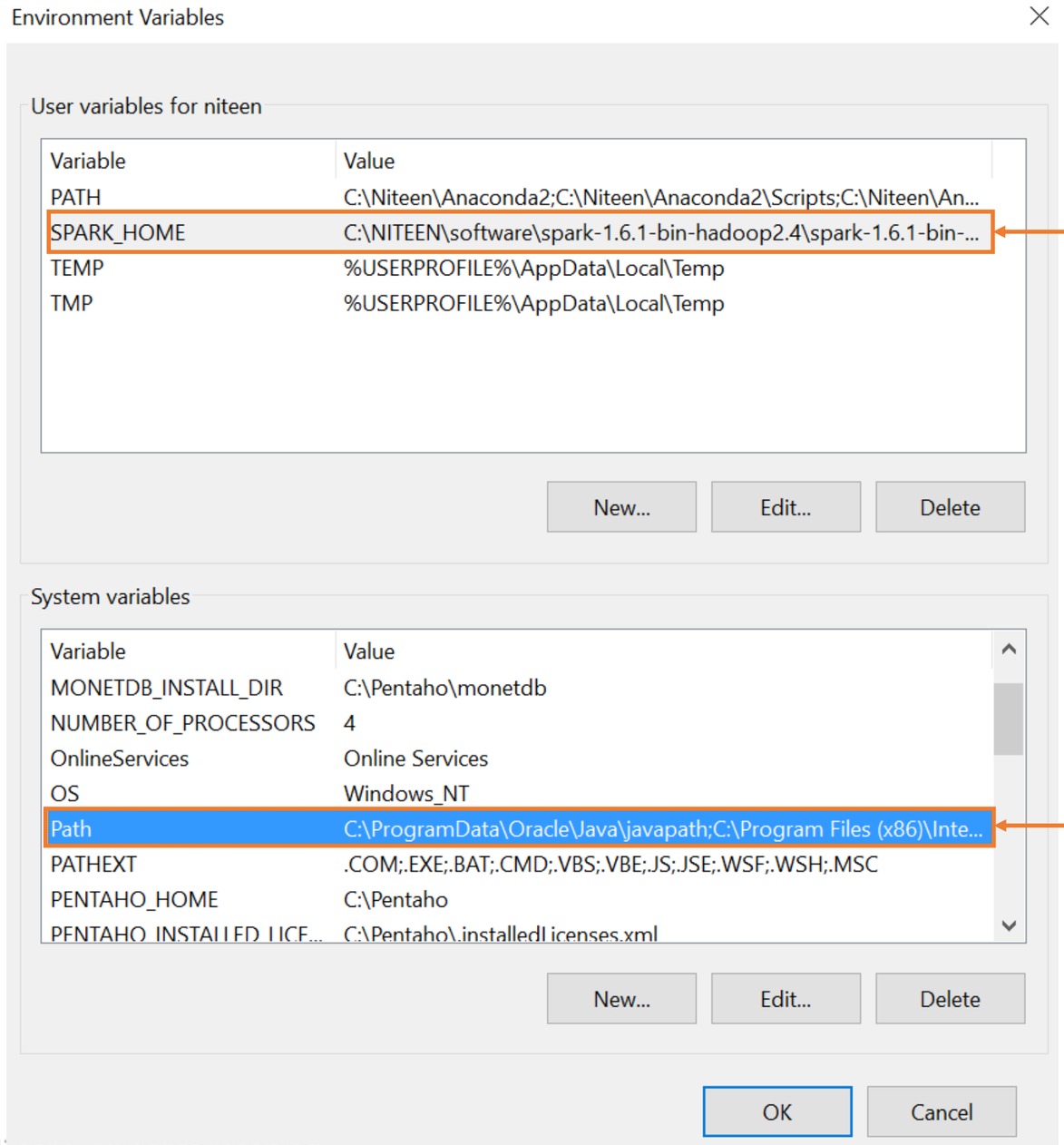
To set up the Apache Spark environment, access the link:

http://spark.apache.org/downloads.html

Please use 7-Zip to extract these files.



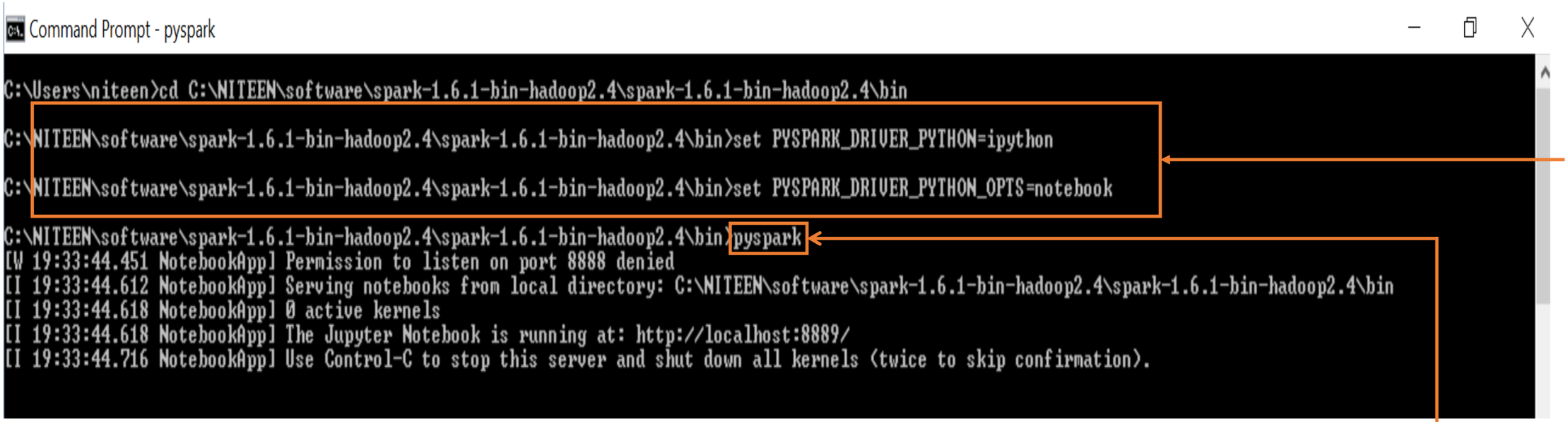
Apache Spark : Environment Variable Set Up



[installed directory]\spark-1.6.1-bin-hadoop2.4\spark-1.6.1-bin-hadoop2.4

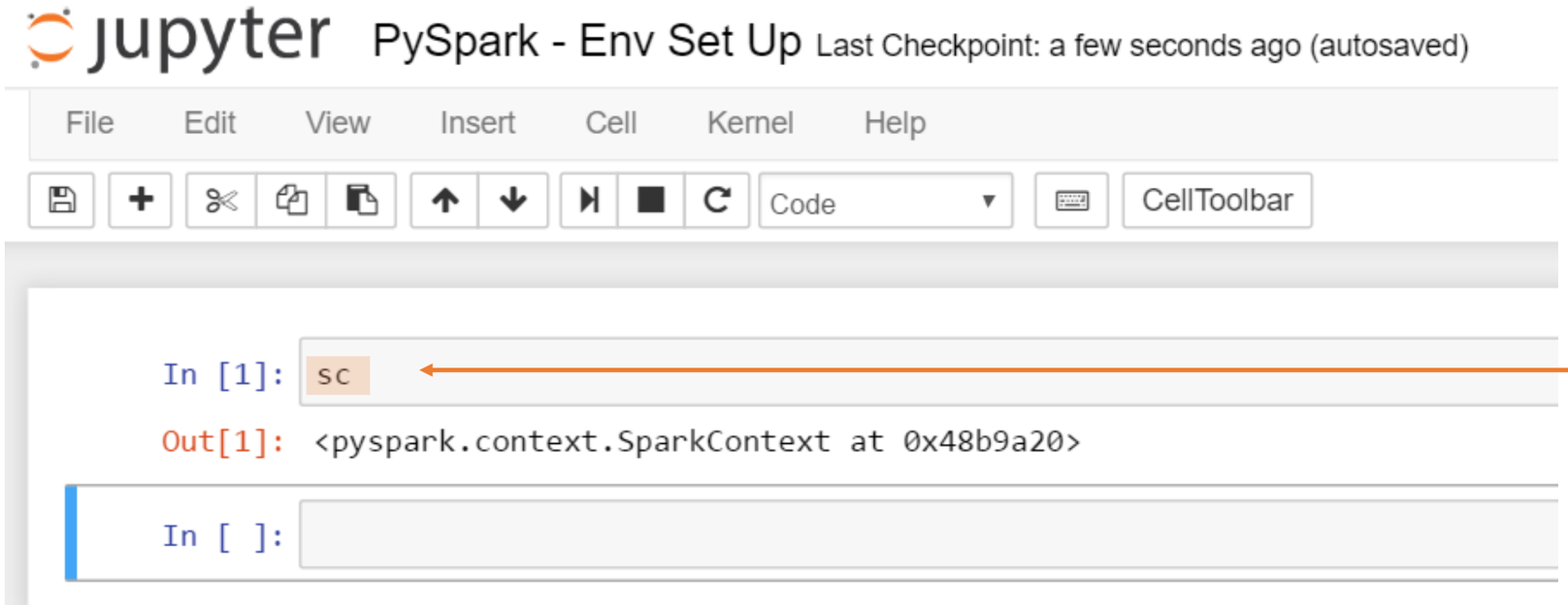
[installed directory] \spark-1.6.1-bin-hadoop2.4\spark-1.6.1-bin-hadoop2.4\bin

Apache Spark: Jupyter Notebook Integration



Setup the pyspark notebook specific variables

Run the pyspark command



Check SparkContext

Demo 02—Using PySpark to Determine Word Count

Demonstrate how to use the Jupyter integrated PySpark API to determine the word count of a given dataset

DATA SCIENCE

Knowledge Check

KNOWLEDGE CHECK

What happens if the available memory is insufficient while performing RDD transformations?

a. The RDD process waits for memory to be available

b. The process is cancelled by scheduler

c. The data is written to the disk

d. The RDD process fails

KNOWLEDGE CHECK

What happens if the available memory is insufficient while performing RDD transformations?

a. The RDD process waits for memory to be available

b. The process is cancelled by scheduler

c. The data is written to the disk

d. The RDD process fails

The correct answer is. c.

Explanation The data is written to the disk in case the memory is insufficient while performing transformations.

Assignment

Assignment 01

Problem | Instructions

To determine the word count of the given Amazon dataset:

- Create a MapReduce program to determine the word count of the Amazon dataset
- Submit the MapReduce task to HDFS and run it
- Verify the output

Click each tab to know more. Click the Resources tab to download the files for this assignment.

Assignment 01

Problem | Instructions

Instructions on performing the assignment:

- Download the “Amazon text dataset.txt” file from the “Resource” tab. Use the QuickStart VM terminal to create a file and copy-paste the Amazon dataset into it.

Special instructions:

- This assignment is done purely on Cloudera’s QuickStart VM. You may need to learn a few basic UNIX commands to operate the program.
- For any cues, refer the Hadoop Streaming demo provided in the lesson.

Assignment

Assignment 02

Problem | Instructions

Use the given dataset to count and display all the airports based in New York using PySpark. Perform the following steps:

- View all the airports listed in the dataset
- View only the first 10 records
- Filter the data for all airports located in New York
- Clean up the dataset, if required

Assignment 02

Problem | Instructions

Instructions on performing the assignment:

- Download the “Airport.csv” file from the “Resource” tab. You can load the saved file to the Jupyter notebook that you would be using to complete the assignment..

Common instructions:

- If you are new to Python, download the “Anaconda Installation Instructions” document from the “Resources” tab to view the steps for installing Anaconda and the Jupyter notebook.
- Download the “Assignment 02” notebook and upload it on the Jupyter notebook to access it.
- Follow the provided cues to complete the assignment.

Quiz

QUIZ 1

What are the core components of Hadoop? *Select all that apply.*

a. MapReduce

b. HDFS

c. Spark

d. RDD

QUIZ 1

What are the core components of Hadoop? *Select all that apply.*

a. MapReduce

b. HDFS

c. Spark

d. RDD

The correct answer is. a & b

Explanation: MapReduce and HDFS are the core components of Hadoop.

QUIZ 2

MapReduce is a data processing framework which gets executed _____.

a. at DataNode

b. at NameNode

c. on client side

d. in memory

QUIZ 2

MapReduce is a data processing framework which gets executed _____.

a. at DataNode

b. at NameNode

c. on client side

d. in memory

The correct answer is. a

Explanation: The MapReduce program is executed at the data node and the output is written to the disk.

QUIZ 3

Which of the following functions is responsible for consolidating the results produced by each of the Map() functions/tasks?

a. Reducer

b. Mapper

c. Partitioner

d. All of the above

QUIZ 3

Which of the following functions is responsible for consolidating the results produced by each of the Map() functions/tasks?

a. Reducer

b. Mapper

c. Partitioner

d. All of the above

The correct answer is. a

Explanation: Reducer combines or aggregates results produced by mappers.

QUIZ 4

What transforms input key-value pairs to a set of intermediate key-value pairs?

a. Mapper

b. Reducer

c. Combiner

d. Partitioner

QUIZ 4

What transforms input key-value pairs to a set of intermediate key-value pairs?

a. Mapper

b. Reducer

c. Combiner

d. Partitioner

The correct answer is. a

Explanation: Mapper processes input data to intermediate key-value pairs which are in turn processed by reducers.

Key Takeaways

- As Python is a Data Scientist's preferred choice of language, it is important to provide Big Data solutions that accommodates it.
- There are two primary components of Hadoop architecture: Hadoop Distributed File System or HDFS and MapReduce.
- Both Hadoop and Spark provide Python APIs to help Data Scientists use the Big Data platform.
- MapReduce has two functions—mapper and reducer.
- MapReduce carries out computations on data through disk I/O operations while Apache Spark carries them out in-memory.
- The main programming approach of Spark is RDD.
- Spark is almost 10 to 100 times faster than Hadoop MapReduce.
- There are mainly four components in Spark tools: Spark SQL, Spark Streaming, Mllib, and GraphX.

This concludes “Python Integration with MapReduce and Spark”.

This is the final lesson of the Data Science with Python Course.

Project

Project 01: Stock Market Data Analysis

After learning about Data Science in depth, it is now time to implement the knowledge gained through this course in real-life scenarios. We will provide you with four scenarios where you need to implement data science solutions. To perform these tasks, you can use the different Python libraries such as NumPy, SciPy, Pandas, scikit-learn, matplotlib, BeautifulSoup, and so on.

You will focus on acquiring stock data information for the companies listed.

The scope of the project is as follows:

Project 01: Stock Market Data Analysis

Problem | Instructions | Solution

- Import the financial data using Yahoo data reader for the following companies:
  - Yahoo
  - Apple
  - Amazon
  - Microsoft
  - Google

- Perform fundamental data analysis
  - Fetch the last one year's data
  - View the values of Apple's stock
  - Display the plot of closing price
  - Display the stock trade by volume
  - Plot all companies' data together for closing prices

Project 01: Stock Market Data Analysis

Problem | Instructions | Solution

- Perform Daily Return Analysis and show the relationship between different stocks
  - Plot the percentage change plot for Apple's stock
  - Show a joint plot for Apple and Google
  - Use PairPlot to show the correlation between all the stocks

- Perform risk analysis

Project 01: Stock Market Data Analysis

Problem | Instructions | Solution

Instructions to perform the project:

- Download the “Anaconda Installation Instructions” document from the “Resources” tab to view the steps to install Anaconda and the Jupyter notebook.
- Download the “Project 01” notebook and upload it on the Jupyter notebook to access it.
- Follow the provided cues to complete the project.

We recommend you to first solve the project and then view the solution to assess your learning.

Problem | Instructions | Solution

Hope you had a good experience working on the project “Stock Market Data Analysis.”

Go to the next screen to assess your performance.

Click **Next** to view the demo.

Project

Project 02: Titanic data set analysis

After learning about Data Science in depth, it is time to implement the knowledge gained through this course in real-life scenarios. We are providing four real-life scenarios where you can implement data science solutions. To develop solutions to these problems, you can use various Python libraries like NumPy, SciPy, Pandas, Scikit-learn, Matplotlib, BeautifulSoup, and so on.
Project details are given below:

Project 02: Titanic Data Set Analysis

Problem | Instructions | Solution

Titanic Dataset Analysis

On April 15, 1912, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This tragedy shocked the world and led to better safety regulations for ships. Here, we ask you to perform the analysis through the exploratory data analysis technique. In particular, we want you to apply the tools of machine learning to predict which passengers survived the tragedy.

The details of these projects and their scope are listed in the following sections.

Click each tab to know more. Click the Resources tab to download the files for this project.

Project 02: Titanic Data Set Analysis

Problem | Instructions | Solution

Titanic Dataset Analysis

- Data acquisition of the Titanic dataset
  - train dataset
  - test dataset

- Perform the Exploratory Data Analysis (EDA) - for train dataset
  - passengers age distribution
  - passengers survival by age
  - passengers survival breakdown
  - passengers class distribution
  - passengers embarkation by locations

Click each tab to know more. Click the Resources tab to download the files for this project.

Project 02: Titanic Data Set Analysis

simplilearn

Problem | Instructions | Solution

Titanic Dataset Analysis

- Perform machine learning to train the machine model and
  - create user defined function to load train data set
  - create user defined function to load test data set
  - create machine model
  - train the machine
  - predict whether a passenger survived the tragedy or not
  - persist the mode for future re-use
-

Click each tab to know more. Click the Resources tab to download the files for this project.

Project 02: Titanic Data Set Analysis

Problem | Instructions | Solution

Instructions to perform the project:

- Download the “Anaconda Installation Instructions” document from the “Resources” tab to view the steps to install Anaconda and the Jupyter notebook.
- Download the “Project 02” notebook and upload it on the Jupyter notebook to access it.
- Follow the provided cues to complete the project.

We recommend you to first solve the project and then view the solution to assess your learning.

Problem | Instructions | Solution

Hope you had a good experience working on the project “Titanic data set analysis.”

Go to the next screen to assess your performance.

Click **Next** to view the demo.

To view the demo for this project, click Next.

Thank You