National University of Computer and Emerging Sciences

Laboratory Manual-08

for

Fundamentals of Big Data Lab

Course Instructor: Dr Iqra Safdar
Lab Instructors: Rida Mahmood, Mr. Muhammad Mazarib
Section: BDS-4A
Date: 18-Apr-2023
Semester: Spring 2023

Department of Computer Science

FAST-NU, Lahore, Pakistan

Big Data processing systems

Hadoop/MapReduce:

Scalable and fault tolerant framework written in Java

Open source

Batch processing

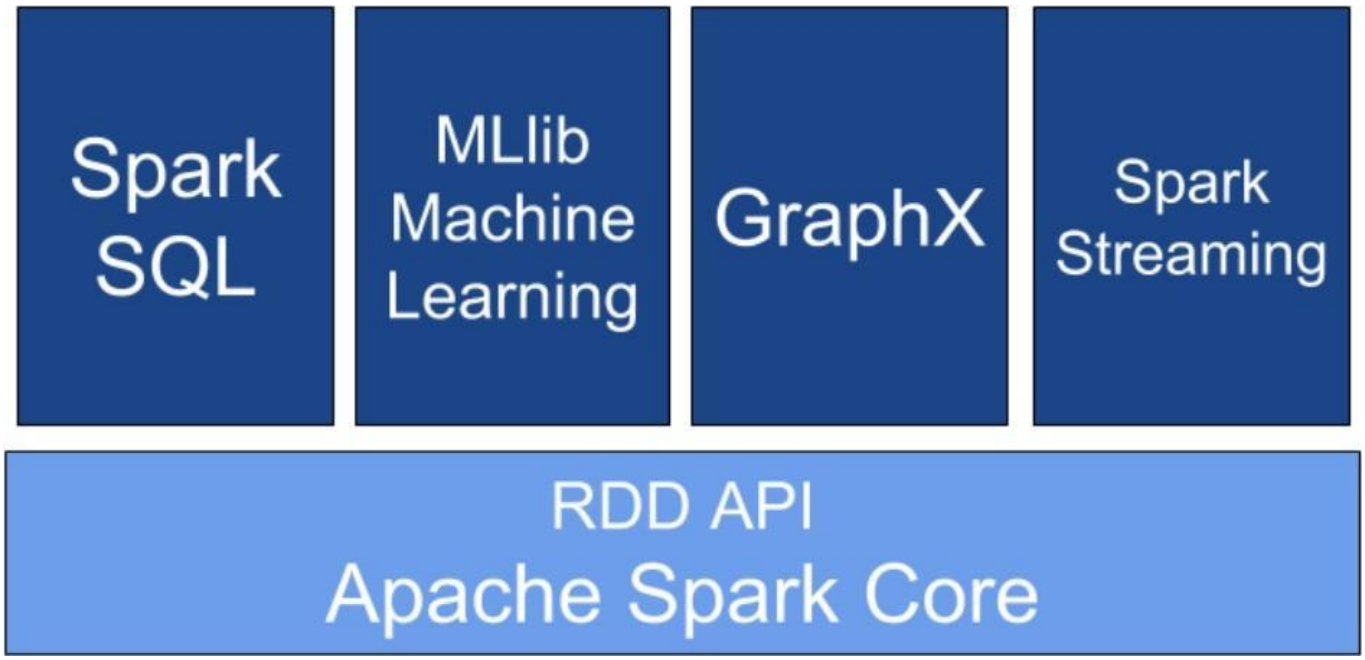
Apache Spark:

General purpose and lightning fast cluster computing system

Open source

Both batch and real-time data processing

Apache Spark Components



Spark modes of deployment

Local mode: Single machine such as your laptop.

Local model convenient for testing, debugging and demonstration

Cluster mode: Set of pre-defined machines

Good for production

Overview of PySpark

Apache Spark is written in Scala

To support Python with Spark, Apache Spark Community released PySpark

Similar computation speed and power as Scala

PySpark APIs are similar to Pandas and Scikit-learn

PySpark Documentation Link : https://spark.apache.org/docs/3.3.2/
Pyspark RDD Documentation Link: https://spark.apache.org/docs/latest/rdd-programming-guide.html

Note: Use google colab or jupyter notebook for PySpark

Configuration of PySpark in System

Install pyspark using the line: !pip install pyspark

Import the following library:

```
from pyspark import SparkContext, SparkConf
```

Configure the PySaprk and start the session:

```
conf = SparkConf().setAppName(appName).setMaster(master)
sc = SparkContext(conf=conf)
```

where appName is your name of your project/lab and "local[*]" is your master if you are working locally.

Understanding SparkContext

A SparkContext represents the entry point to Spark functionality. It's like a key to your car. When we run any Spark application, a driver program starts, which has the main function and your SparkContext gets initiated here.

Use of Lambda function in python - filter()

What are anonymous functions in Python?

Lambda functions are anonymous functions in Python

Very powerful and used in Python. Quite efficient with map() and filter()

Lambda functions create functions to be called later similar to def

It returns the functions without any name (i.e. anonymous)

Inline a function definition or to defer execution of a code

Lambda function syntax

The general form of lambda functions is

```
lambda arguments: expression
```

Example of lambda function is as follow:

```
double = lambda x: x * 2
print(double(3))
```

Difference between def vs lambda functions

Python code to illustrate cube of a number

```
def cube(x):
return x ** 3

g = lambda x: x ** 3
print(g(10))
print(cube(10))
```

No return statement for lambda

Can put lambda function anywhere

Use of Lambda function in Python - map()

map() function takes a function and a list and returns a new list which contains items returned by that function for each item

General syntax of map()

```
map(function, list)
```

Example of map()

```
items = [1, 2, 3, 4]
list(map(lambda x: x + 2 , items))
```

Use of Lambda function in python - filter()

filter() function takes a function and a list and returns a new list for which the function evaluates as true

General syntax of ,filter()

```
filter(function, list)
```

Example of ,filter()

```
items = [1, 2, 3, 4]
list(filter(lambda x: (x%2 != 0), items)
```

LAB TASKS

1. Create my_list which contains number from 1 to 10. Print the list. Square each item in my_list using map() and lambda(). Print the result of map function.
2. Create my_list_2 which contains 20 random numbers. Print the list. Filter the numbers divisible by 5 from my_list2 using filter() and lambda().Print the numbers divisible by 5 from my_list2.
3. Assume you have received a huge user file of user comments and you have to perform some basic statistics on it.

   File format :

   UserName, Comment

   Aliya153, Your website is superb

   Sara2, You need to work on your website design

   Ali45, Good !!!

   Ali45, I will definitely visit again

   Write a PySPARK code to perform the following tasks

   a) Load the file

   b) Find comments given by each user

   c) Determine the number of long comments given by each user where the length of the long comment should be greater than 20 alphabets.

   d) Count the number of UserNames starting with each English alphabet.

   e) Find the user who has given the maximum number of comments

4. We want to remove stop words from the comments of the users in the above dataset.

   Stop Words are those words that do not contain important information for example to, was, do etc. Usually these words are filtered out from search queries.

   Write a PySpark program to input a text file containing stop words (you can get one such file from internet). Use this file to remove stop words from the comment of the users.

Hint: broadcast the stop word file to efficiently removing the stop words.