Aisha Muhammad Nawaz L200921

PySpark Lab 2 8A BSCS MMD

20th February 2024

Instructions: RUN PYSPARK and execute commands done in class

In [1]:
```
# #Running on Colab
!pip install pyspark
!pip install -U -q PyDrive
!apt install openjdk-8-jdk-headless -qq
import os
os.environ['JAVA_HOME'] = '/usr/lib/jvm/java-8-openjdk-amd64'
```

```
Collecting pyspark
  Downloading pyspark-3.5.0.tar.gz (316.9 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 316.9/316.9 MB 2.1 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.1
0.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.5.0-py2.py3-none-any.whl size=317425345 sha256=f8a4b7df2053a6
06ef0c2137e2ff7ae4d7c32eb70196ede1781aa2b37bb2df80
  Stored in directory: /root/.cache/pip/wheels/41/4e/10/c2cf2467f71c678cfc8a6b9ac9241e5e44a01940da8fbb17fc
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.5.0
The following additional packages will be installed:
  libxtst6 openjdk-8-jre-headless
Suggested packages:
  openjdk-8-demo openjdk-8-source libnss-mdns fonts-dejavu-extra fonts-nanum fonts-ipafont-gothic
  fonts-ipafont-mincho fonts-wqy-microhei fonts-wqy-zenhei fonts-indic
The following NEW packages will be installed:
  libxtst6 openjdk-8-jdk-headless openjdk-8-jre-headless
0 upgraded, 3 newly installed, 0 to remove and 35 not upgraded.
Need to get 39.7 MB of archives.
After this operation, 144 MB of additional disk space will be used.
Selecting previously unselected package libxtst6:amd64.
(Reading database ... 121749 files and directories currently installed.)
Preparing to unpack .../libxtst6_2%3a1.2.3-1build4_amd64.deb ...
Unpacking libxtst6:amd64 (2:1.2.3-1build4) ...
Selecting previously unselected package openjdk-8-jre-headless:amd64.
Preparing to unpack .../openjdk-8-jre-headless_8u392-ga-1~22.04_amd64.deb ...
Unpacking openjdk-8-jre-headless:amd64 (8u392-ga-1~22.04) ...
Selecting previously unselected package openjdk-8-jdk-headless:amd64.
Preparing to unpack .../openjdk-8-jdk-headless_8u392-ga-1~22.04_amd64.deb ...
Unpacking openjdk-8-jdk-headless:amd64 (8u392-ga-1~22.04) ...
Setting up libxtst6:amd64 (2:1.2.3-1build4) ...
Setting up openjdk-8-jre-headless:amd64 (8u392-ga-1~22.04) ...
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/orbd to provide /usr/bin/orbd (orbd) in
auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/servertool to provide /usr/bin/servertoo
l (servertool) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/tnameserv to provide /usr/bin/tnameserv
(tnameserv) in auto mode
Setting up openjdk-8-jdk-headless:amd64 (8u392-ga-1~22.04) ...
```

```
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/clhsdb to provide /usr/bin/clhsdb (clhsdb) i
n auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/extcheck to provide /usr/bin/extcheck (extch
eck) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/hsdb to provide /usr/bin/hsdb (hsdb) in auto
mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/idlj to provide /usr/bin/idlj (idlj) in auto
mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/javah to provide /usr/bin/javah (javah) in a
uto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/jhat to provide /usr/bin/jhat (jhat) in auto
mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/jsadebugd to provide /usr/bin/jsadebugd (jsa
debugd) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/native2ascii to provide /usr/bin/native2asci
i (native2ascii) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/schemagen to provide /usr/bin/schemagen (sch
emagen) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/wsgen to provide /usr/bin/wsgen (wsgen) in a
uto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/wsimport to provide /usr/bin/wsimport (wsimp
ort) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/xjc to provide /usr/bin/xjc (xjc) in auto mo
de
Processing triggers for libc-bin (2.35-0ubuntu3.4) ...
/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc_proxy.so.2 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_0.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_5.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc.so.2 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbb.so.12 is not a symbolic link
```

In [2]:
```
#  Import the libraries we will need
import pyspark
from pyspark.sql import *
from pyspark.sql.functions import *
from pyspark import SparkContext, SparkConf
```

In [3]:
```
# Create Spark session and ContextRun PySpark.
# create the session
conf = SparkConf().set("spark.ui.port","4050")
# create the context
sc = pyspark.SparkContext(conf=conf)
spark = SparkSession.builder.appName("DataFrame").config('spark.ui.port', '4050').getOrCreate()
```

In [4]:
```
spark
```

Out[4]:

SparkSession - in-memory

SparkContext

Spark UI (http://f7fd763a46f7:4050)

Version

v3.5.0

Master

local[*]

AppName

pyspark-shell

Execute the following transformations and actions filter, map, flat map, count, saveasfile, collect. union, intersection, distinct, and difference commands learn in class.

In [ ]:
```
nums = sc.parallelize([('a',1),('b',1),('a',6)])
nums.reduceByKey(lambda x,y:x+y).collect() #Summing up values belonging to the same key
```

Out[ ]:
```
[('b', 1), ('a', 7)]
```

In [ ]:
```
nums = sc.parallelize([('a',1),('b',1),('a',6)])
nums.reduce(lambda x,y:x+y) #Summing up values belonging to the same key [NOTE HERE MAP WILL NOT WORK WELL!!!
```

Out[ ]:
```
('a', 1, 'b', 1, 'a', 6)
```

In [ ]:
```
nums = sc.parallelize([('a',1),('b',1),('a',6)])
nums.reduceByKey(lambda x,y:y-x).collect() #Subtracting values belonging to the same key
```

Out[ ]:
```
[('b', 1), ('a', 5)]
```

In [ ]:
```
# Class task1 : Word Count Problem
text=sc.parallelize(['The significant impact of impact of events','makes it of top','to impact the'])
text2=text.flatMap(lambda line : line.split(" "))
text3 = text2.map(lambda word: (word.lower(),1))
text4 = text3.reduceByKey(lambda x,y:x+y).collect()
print(text4)
```

```
[('significant', 1), ('of', 3), ('the', 2), ('impact', 3), ('events', 1), ('makes', 1), ('it', 1), ('top',
1), ('to', 1)]
```

In [ ]:
```
# Class task1 : Word Count Problem version 2
text=sc.parallelize(['The significant impact of impact of events','makes it of top','to impact the'])
text2=text.flatMap(lambda line : line.split(" ")).countByValue()
print(text2)
```

```
defaultdict(<class 'int'>, {'The': 1, 'significant': 1, 'impact': 3, 'of': 3, 'events': 1, 'makes': 1, 'it':
1, 'top': 1, 'to': 1, 'the': 1})
```

In [ ]:
```
# Class task2 : Find Average of Students in Quizes (Each Student can have different number of quizes, find th
eir own respective averages)
text=sc.parallelize([(12,'S1'),(14,'S2'),(8,'S1')]) #marks,studentNo key-value pairs
text2=text.map(lambda x:(x[1],x[0])) #Swap pairs
text3=text2.mapValues(lambda y:(y,1)) #Each student count 1
text4=text3.reduceByKey(lambda x,y:(x[0]+y[0],x[1]+y[1])) #Summing marks and counts
text5=text4.mapValues(lambda y:y[0]/y[1]) #Finding Average
print(text5.collect())
```

```
[('S2', 14.0), ('S1', 10.0)]
```

In [ ]:
```
# Class task3 : Find Average Marks of the assesments (Overall)
text=sc.parallelize([(12,'S1'),(14,'S2'),(8,'S1')]) #marks,studentNo key-value pairs
text2=text.map(lambda x:x[0]).sum() #drop student id not needed! & Then sum up all the values
text3=text2/text.count() #Find Average. Here text.count() returns 3
print(text3)
```

```
11.333333333333334
```

In [ ]:
```
# Map Values
nums = sc.parallelize([('a',1),('b',1),('a',6)])
nums.mapValues(lambda x:x+2).collect()
```

Out[ ]:
```
[('a', 3), ('b', 3), ('a', 8)]
```

In [ ]:
```
# Flat Map Values
nums = sc.parallelize([('a',1),('b',1),('a',6)])
nums.flatMapValues(lambda x:(x+2,)).collect()
```

Out[ ]:
```
[('a', 3), ('b', 3), ('a', 8)]
```

In [ ]:
```
rdd = sc.parallelize([('a', 1), ('b', 2)])
result = rdd.mapValues(lambda x: (x, x + 10))
result.collect()
```

Out[ ]:
```
[('a', (1, 11)), ('b', (2, 12))]
```

In [ ]:
```
rdd = sc.parallelize([('a', 1), ('b', 2)])
result = rdd.flatMapValues(lambda x: (x, x + 10))
result.collect()
```

Out[ ]:
```
[('a', 1), ('a', 11), ('b', 2), ('b', 12)]
```

Explore the SPARKcluster UI (user-interface)

In [ ]:
```
# !pip install pyngrok
```

```
Collecting pyngrok
  Downloading pyngrok-7.1.2-py3-none-any.whl (22 kB)
Requirement already satisfied: PyYAML>=5.1 in /usr/local/lib/python3.10/dist-packages (from pyngrok) (6.0.1)
Installing collected packages: pyngrok
Successfully installed pyngrok-7.1.2
```

In [ ]:
```
from pyngrok import ngrok, conf
import getpass

# Set Ngrok authtoken
print("Enter your authtoken, which can be copied from https://dashboard.ngrok.com/auth")
conf.get_default().auth_token = getpass.getpass()

# Define the port
ui_port = 4050

# Connect to Ngrok and get the public URL
try:
    public_url = ngrok.connect(ui_port).public_url
    print(f" * Ngrok tunnel created: {public_url} -> http://127.0.0.1:{ui_port}")
except Exception as e:
    print(f"Error creating Ngrok tunnel: {e}")

# My Authentication Token 2cSK5j3NB6McxNBPb9wFQQfF2MW_6njVGDJ7hkV1W6e9B7v9F
```

```
Enter your authtoken, which can be copied from https://dashboard.ngrok.com/auth
··········
 * Ngrok tunnel created: https://ffe7-35-221-208-251.ngrok-free.app -> http://127.0.0.1:4050
```

In [ ]:
```

```