Big Data Analytics Lab (23CP309P)

Lab Report Submitted to

Pandit Deendayal Energy University, Gandhinagar

for

Bachelor of Technology

in

Computer Science & Engineering Department

Submitted by

Dhvani PATEL 21BCP116
Semester: VI
Division: 2 , Group: G4

Course Faculty

SOHAM VYAS

PANDIT DEENDAYAL ENERGY UNIVERSITY, GANDHINAGAR - 382007, India
May-2024

List of Practicals

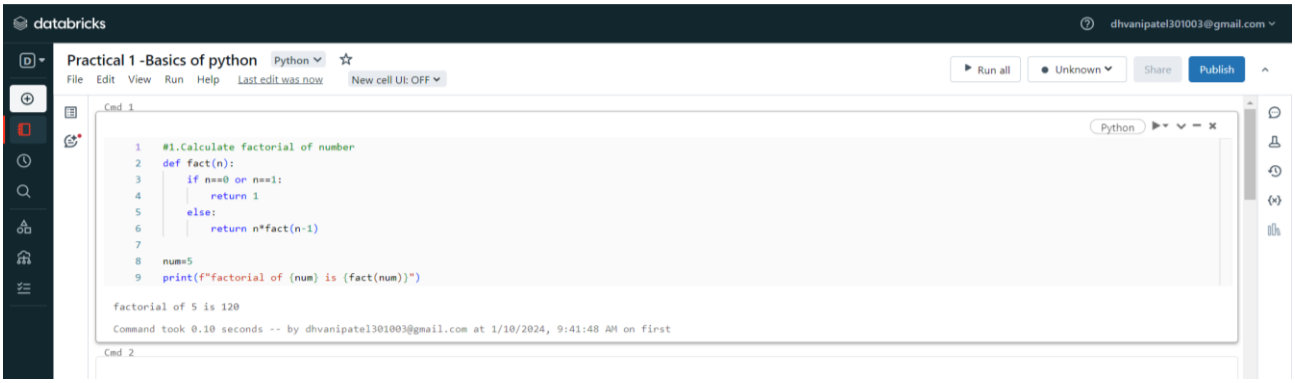
Exp. No.	Experiment Title	Date	Signature
1	Basics of Python	10/01/24	
2	Basics of Scala	17/01/24	
3	Transformation Function	31/01/24	
4	Spark	14/02/24	
5	Spark SQL	21/02/24	
6	Data Processing using Spark	13/03/24	
7	Hadoop, Pig, Kafka, Spark and Hive installation	27/03/24	
8	Implementation of graph data structure using networkx	03/04/24	
9	Linear Regression	10/04/24	

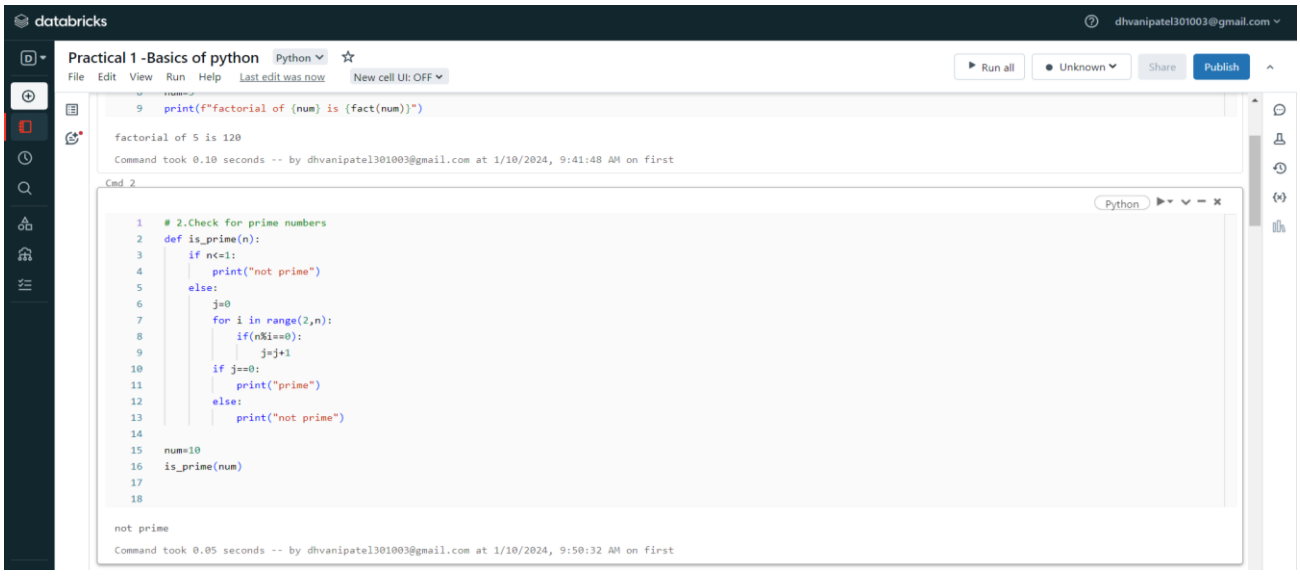
Name: Dhvani Patel

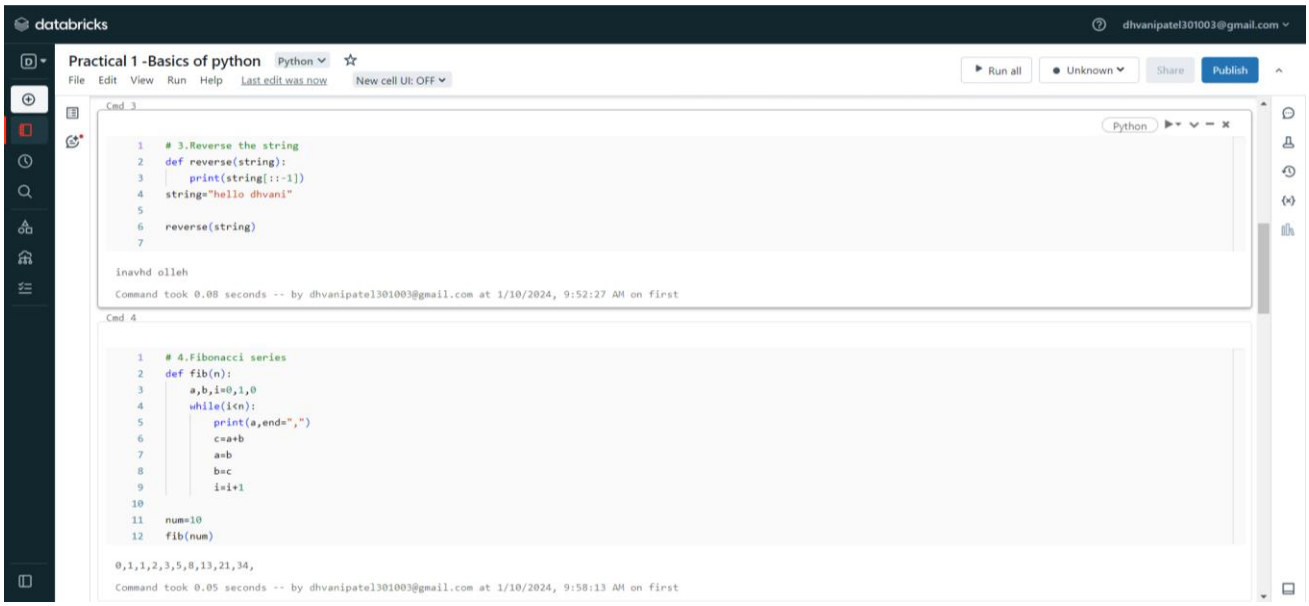
Roll no: 21BCP116

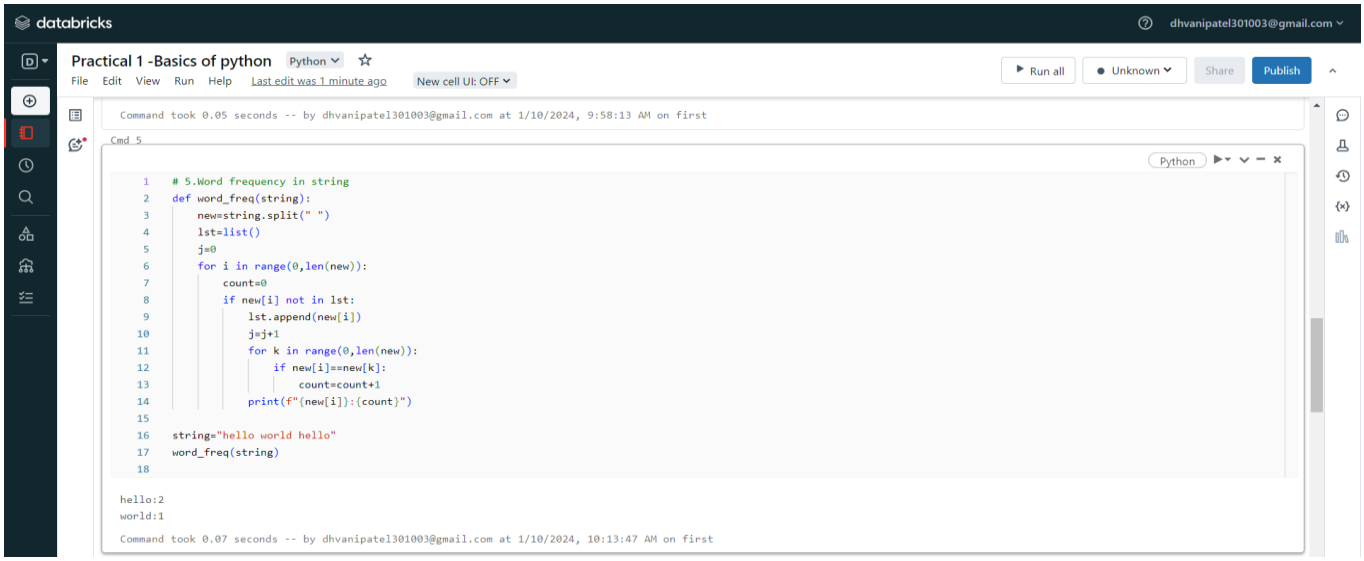
Big Data Analytics Lab

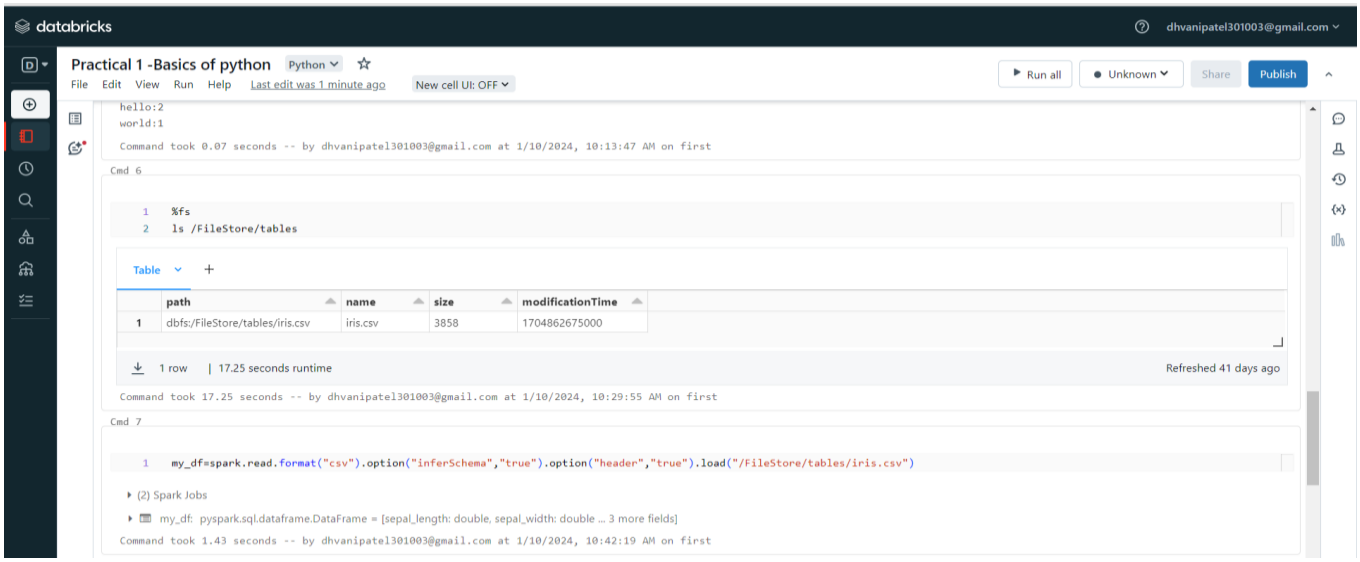
Practical 1 : Basics of Python

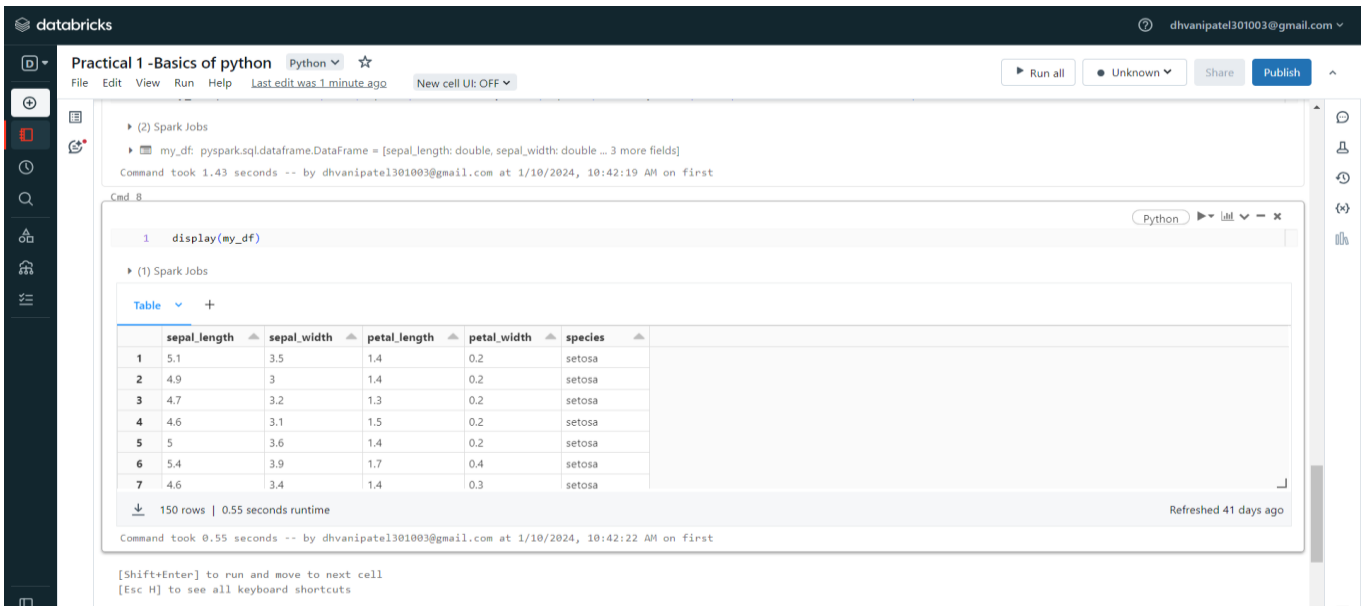







databricks
dhvanipatel301003@gmail.com
Practical 1 -Basics of python
Python
File Edit View Run Help
Last edit was 1 minute ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Command took 0.05 seconds -- by dhvanipatel301003@gmail.com at 1/10/2024, 9:58:13 AM on first
Cmd 5
Python
# 5.Word frequency in string
def word_freq(string):
new=string.split(" ")
lst=list()
j=0
for i in range(0,len(new)):
count=0
if new[i] not in lst:
lst.append(new[i])
j=j+1
for k in range(0,len(new)):
if new[i]==new[k]:
count=count+1
print(f"{new[i]}:{count}")
string="hello world hello"
word_freq(string)
hello:2
world:1
Command took 0.07 seconds -- by dhvanipatel301003@gmail.com at 1/10/2024, 10:13:47 AM on first


databricks
dhvanipatel301003@gmail.com
Practical 1 -Basics of python
Python
File Edit View Run Help
Last edit was 1 minute ago
New cell UI: OFF
Run all
Unknown
Share
Publish
hello:2
world:1
Command took 0.07 seconds -- by dhvanipatel301003@gmail.com at 1/10/2024, 10:13:47 AM on first
Cmd 6
%fs
ls /FileStore/tables
Table
path
name
size
modificationTime
1
dbfs:/FileStore/tables/iris.csv
iris.csv
3858
1704862675000
1 row | 17.25 seconds runtime
Refreshed 41 days ago
Command took 17.25 seconds -- by dhvanipatel301003@gmail.com at 1/10/2024, 10:29:55 AM on first
Cmd 7
my_df=spark.read.format("csv").option("inferSchema","true").option("header","true").load("/FileStore/tables/iris.csv")
(2) Spark Jobs
my_df: pyspark.sql.dataframe.DataFrame = [sepal_length: double, sepal_width: double ... 3 more fields]
Command took 1.43 seconds -- by dhvanipatel301003@gmail.com at 1/10/2024, 10:42:19 AM on first

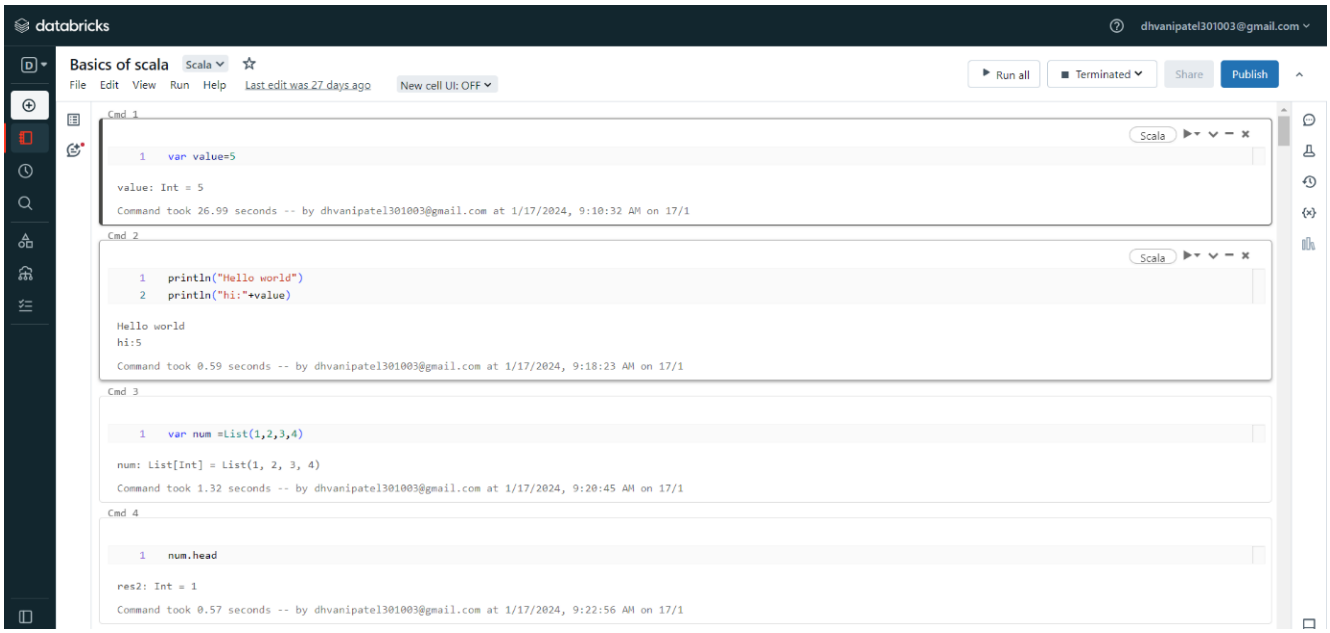

databricks
dhvanipatel301003@gmail.com
Practical 1 -Basics of python
Python
File Edit View Run Help
Last edit was 1 minute ago
New cell UI: OFF
Run all
Unknown
Share
Publish
(2) Spark Jobs
my_df: pyspark.sql.dataframe.DataFrame = [sepal_length: double, sepal_width: double ... 3 more fields]
Command took 1.43 seconds -- by dhvanipatel301003@gmail.com at 1/10/2024, 10:42:19 AM on first
Cmd 8
Python
display(my_df)
(1) Spark Jobs
Table
sepal_length
sepal_width
petal_length
petal_width
species
1
5.1
3.5
1.4
0.2
setosa
2
4.9
3
1.4
0.2
setosa
3
4.7
3.2
1.3
0.2
setosa
4
4.6
3.1
1.5
0.2
setosa
5
5
3.6
1.4
0.2
setosa
6
5.4
3.9
1.7
0.4
setosa
7
4.6
3.4
1.4
0.3
setosa
150 rows | 0.55 seconds runtime
Refreshed 41 days ago
Command took 0.55 seconds -- by dhvanipatel301003@gmail.com at 1/10/2024, 10:42:22 AM on first
[Shift+Enter] to run and move to next cell
[Esc H] to see all keyboard shortcuts

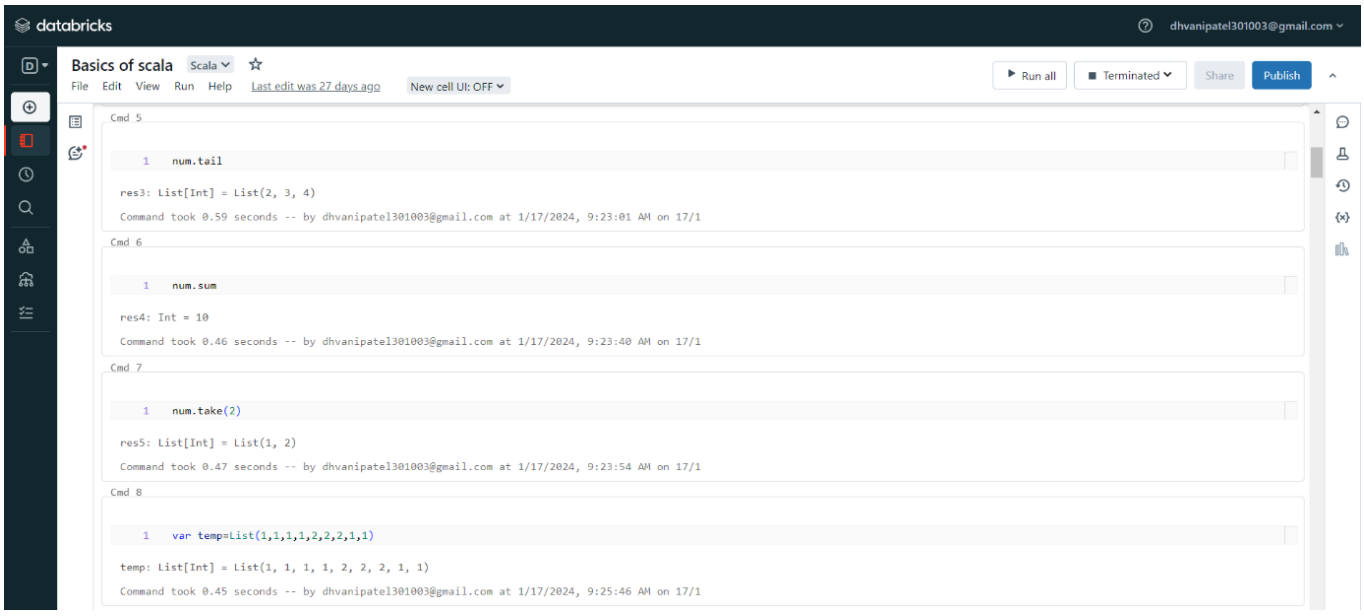
Name: Dhvani Patel

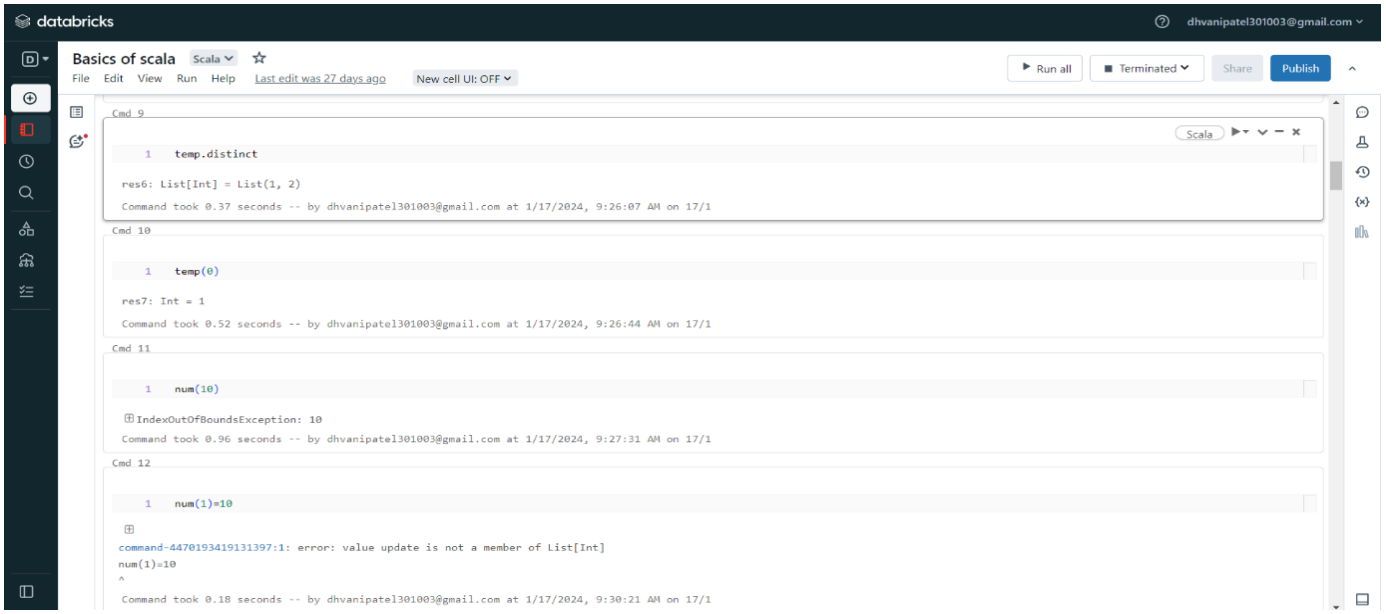
Roll no: 21BCP116

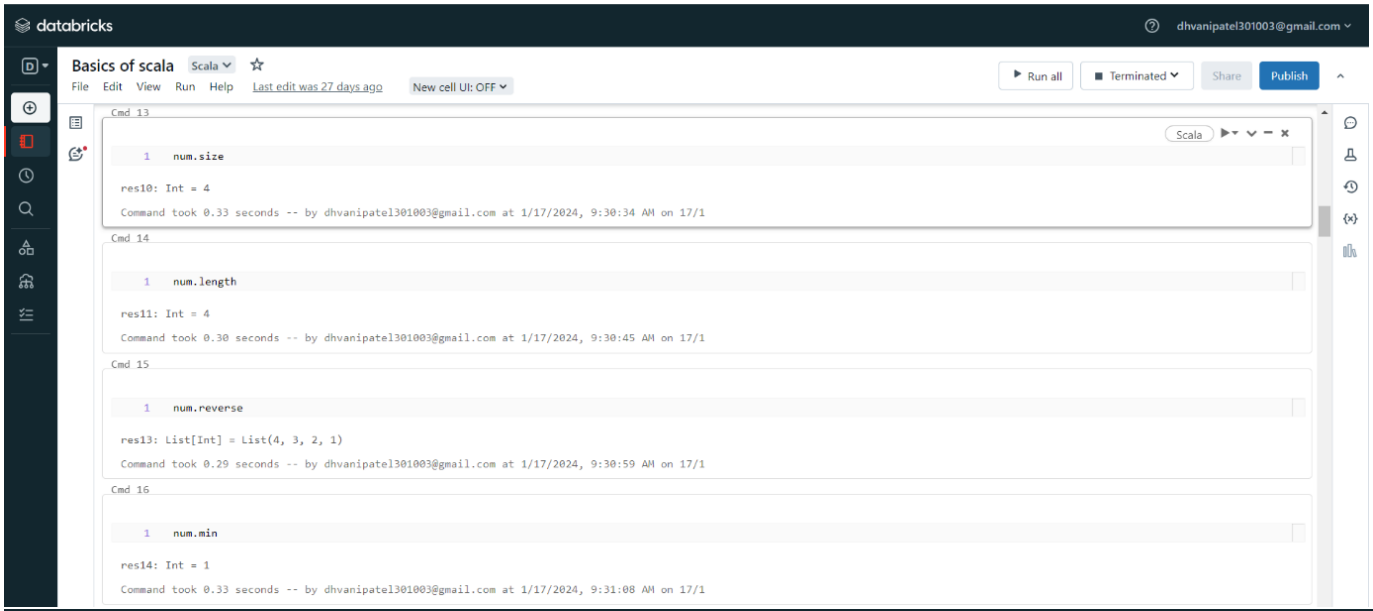
Big Data Analytics Lab

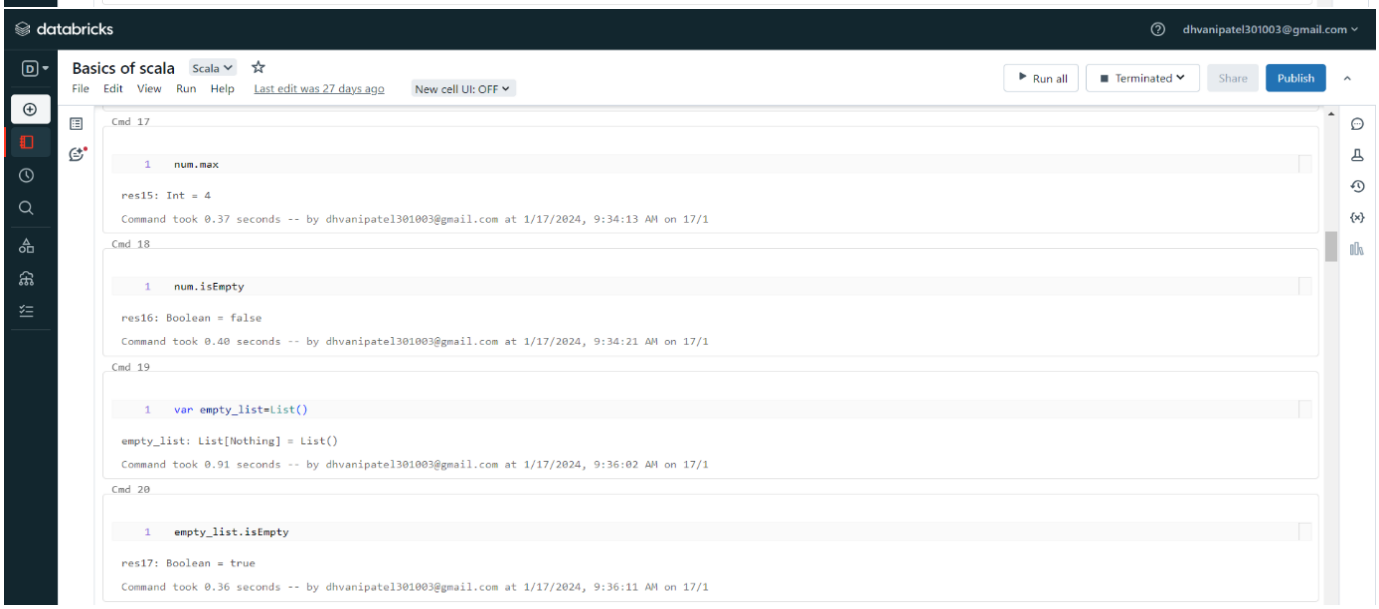
Practical 2 : Basics of Scala

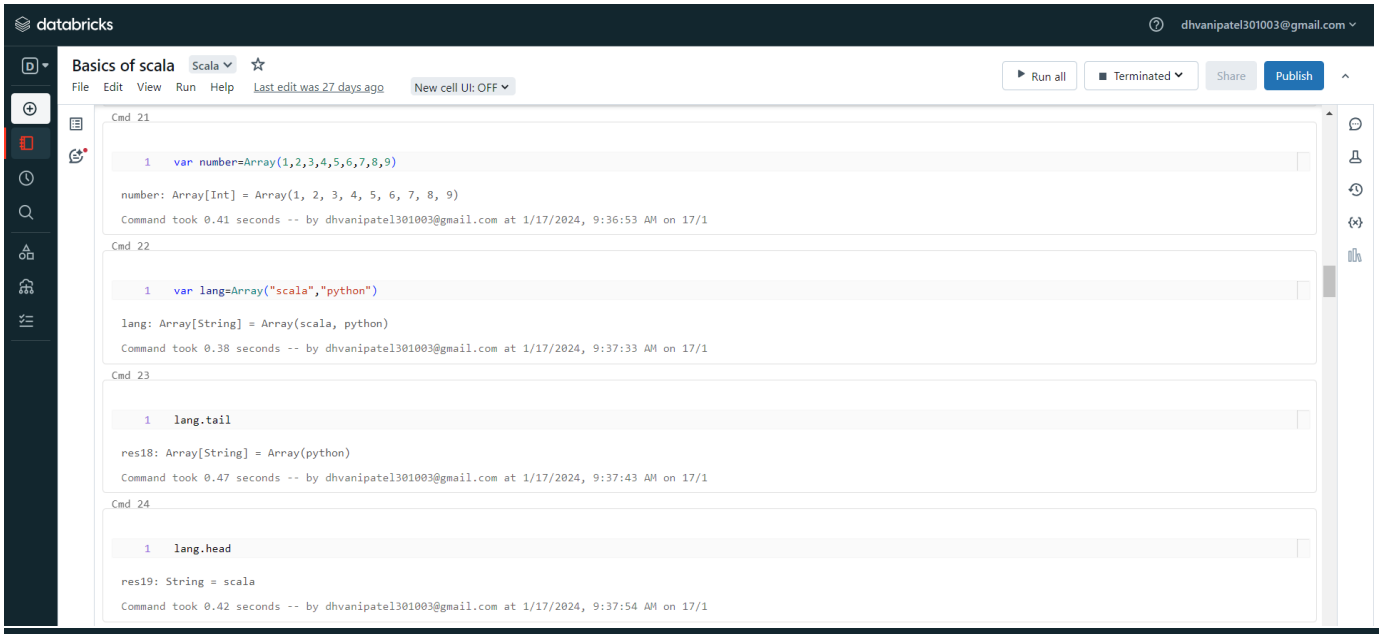


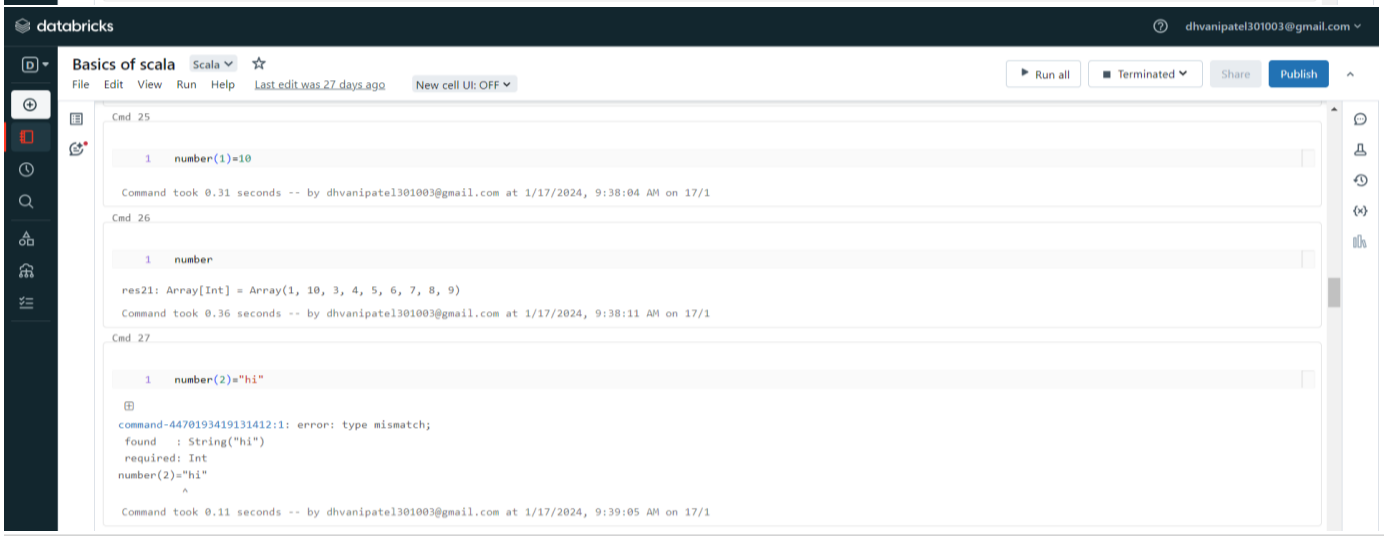


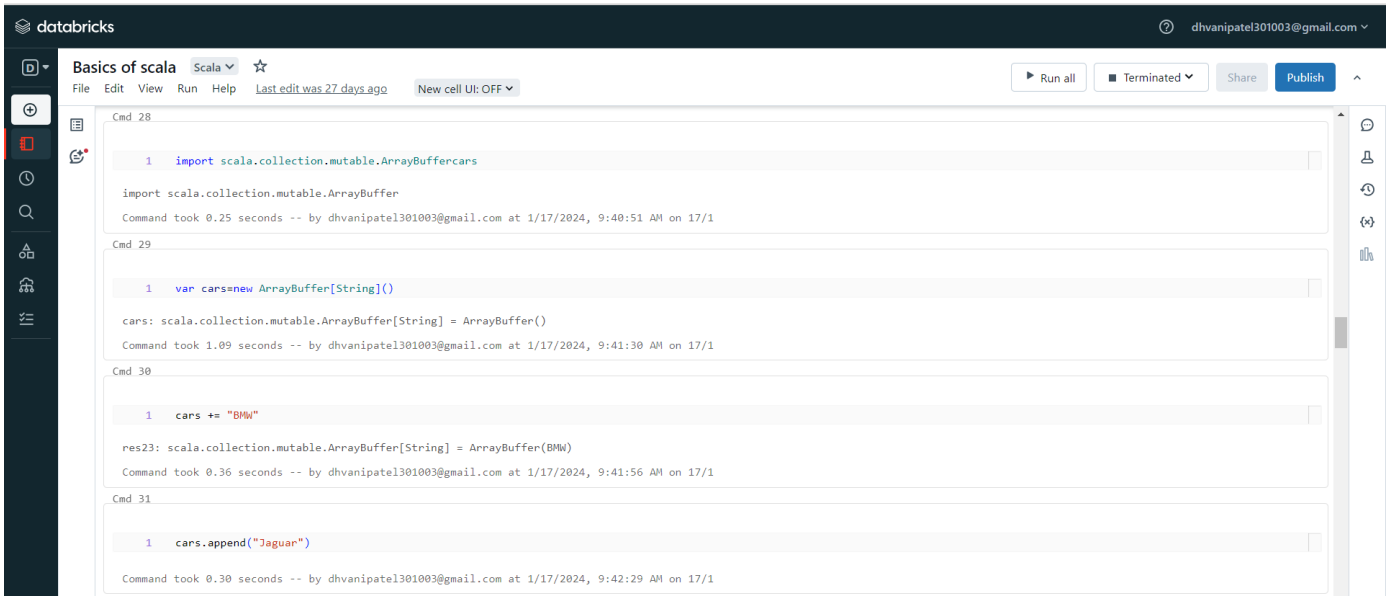

databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 9
1 temp.distinct
res6: List[Int] = List(1, 2)
Command took 0.37 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:26:07 AM on 17/1
Cmd 10
1 temp(0)
res7: Int = 1
Command took 0.52 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:26:44 AM on 17/1
Cmd 11
1 num(10)
IndexOutOfBoundsException: 10
Command took 0.96 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:27:31 AM on 17/1
Cmd 12
1 num(1)=10
command-4470193419131397:1: error: value update is not a member of List[Int]
num(1)=10
^
Command took 0.18 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:30:21 AM on 17/1

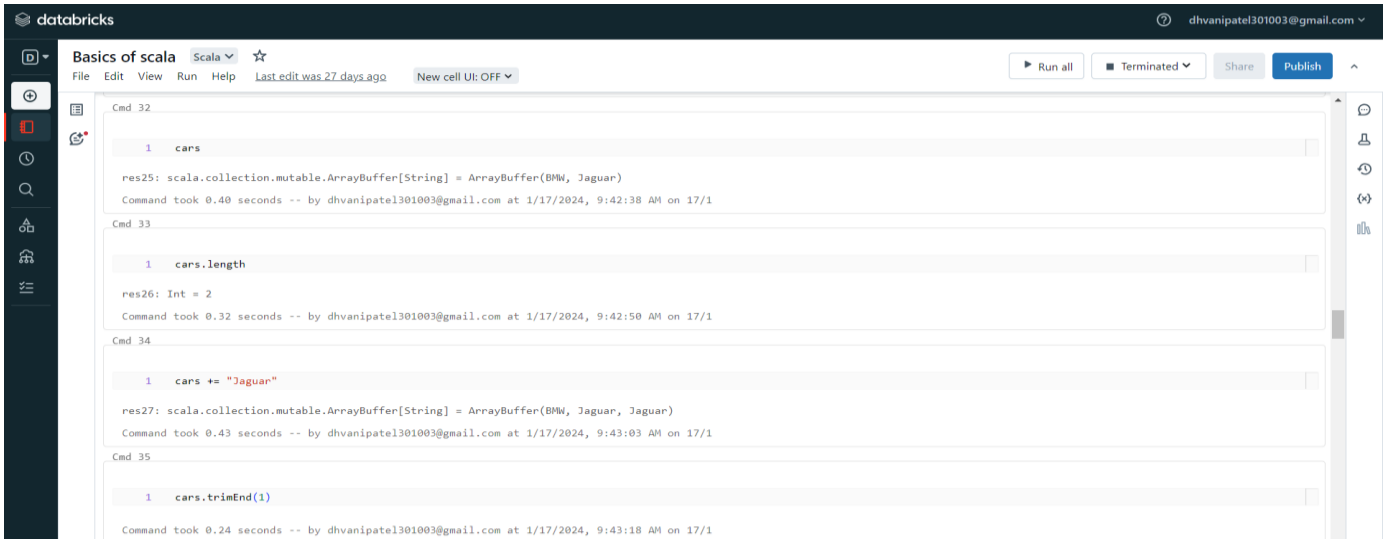

databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 13
1 num.size
res10: Int = 4
Command took 0.33 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:30:34 AM on 17/1
Cmd 14
1 num.length
res11: Int = 4
Command took 0.30 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:30:45 AM on 17/1
Cmd 15
1 num.reverse
res13: List[Int] = List(4, 3, 2, 1)
Command took 0.29 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:30:59 AM on 17/1
Cmd 16
1 num.min
res14: Int = 1
Command took 0.33 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:31:08 AM on 17/1

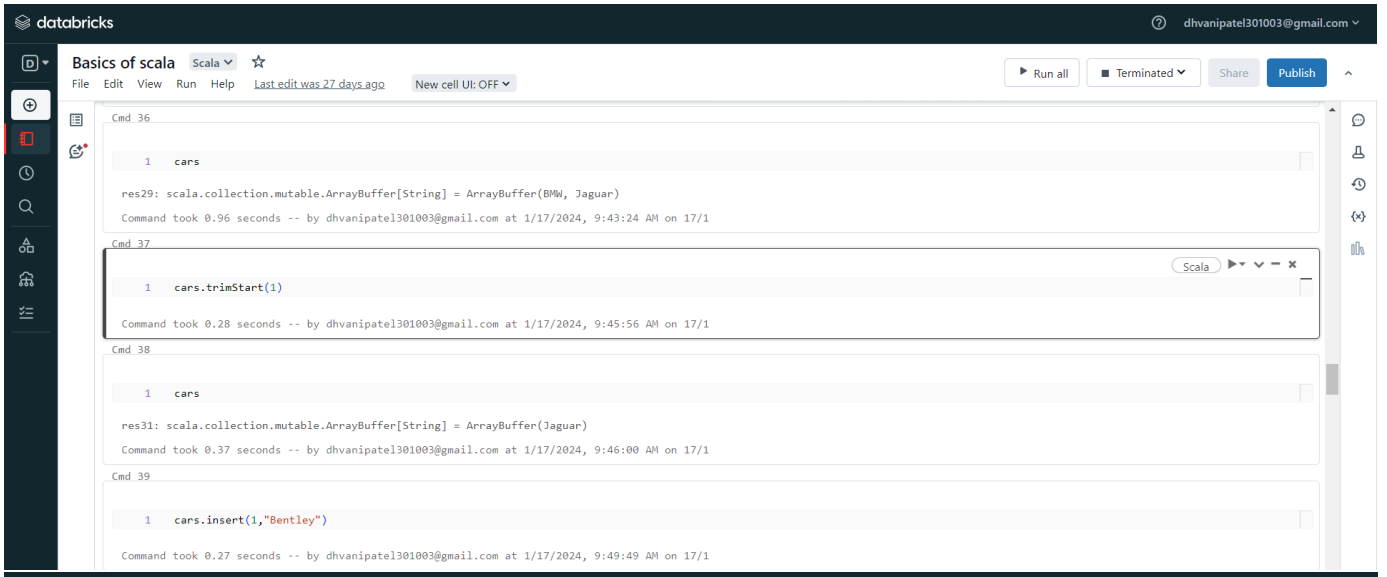

databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 17
1 num.max
res15: Int = 4
Command took 0.37 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:34:13 AM on 17/1
Cmd 18
1 num.isEmpty
res16: Boolean = false
Command took 0.40 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:34:21 AM on 17/1
Cmd 19
1 var empty_list=List()
empty_list: List[Nothing] = List()
Command took 0.91 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:36:02 AM on 17/1
Cmd 20
1 empty_list.isEmpty
res17: Boolean = true
Command took 0.36 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:36:11 AM on 17/1

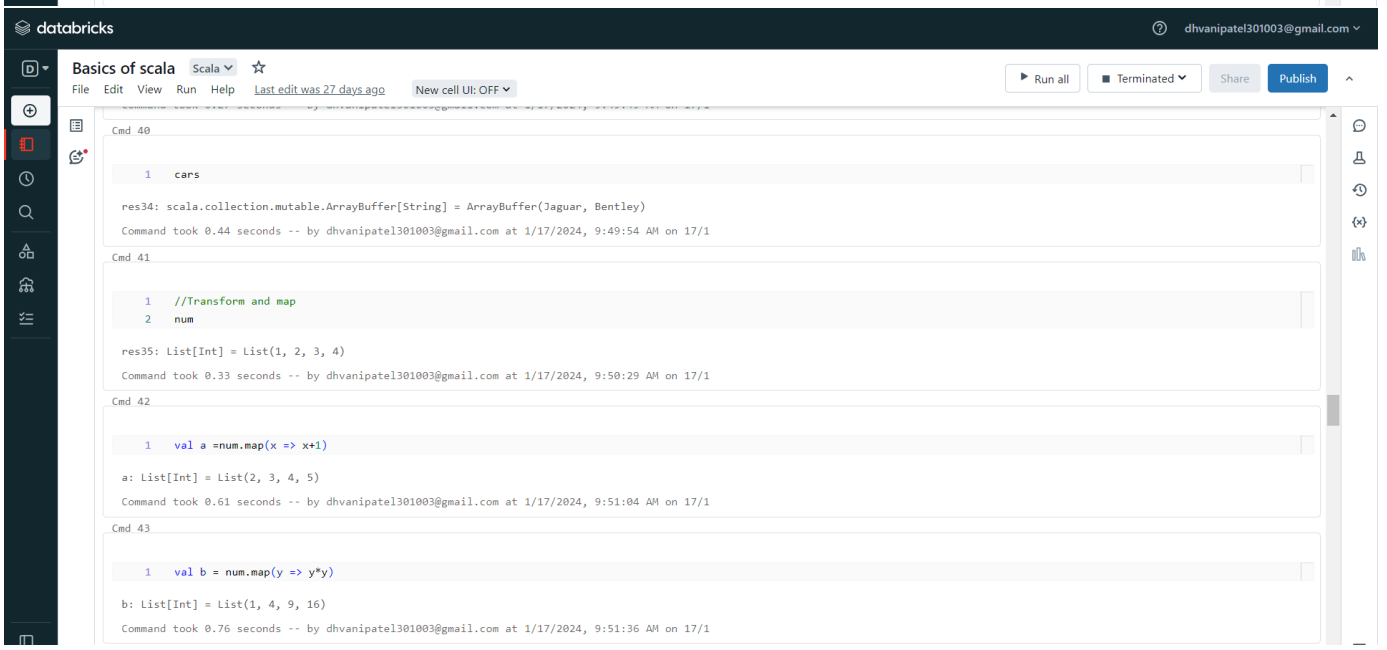

databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 21
1 var number=Array(1,2,3,4,5,6,7,8,9)
number: Array[Int] = Array(1, 2, 3, 4, 5, 6, 7, 8, 9)
Command took 0.41 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:36:53 AM on 17/1
Cmd 22
1 var lang=Array("scala","python")
lang: Array[String] = Array(scala, python)
Command took 0.38 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:37:33 AM on 17/1
Cmd 23
1 lang.tail
res18: Array[String] = Array(python)
Command took 0.47 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:37:43 AM on 17/1
Cmd 24
1 lang.head
res19: String = scala
Command took 0.42 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:37:54 AM on 17/1

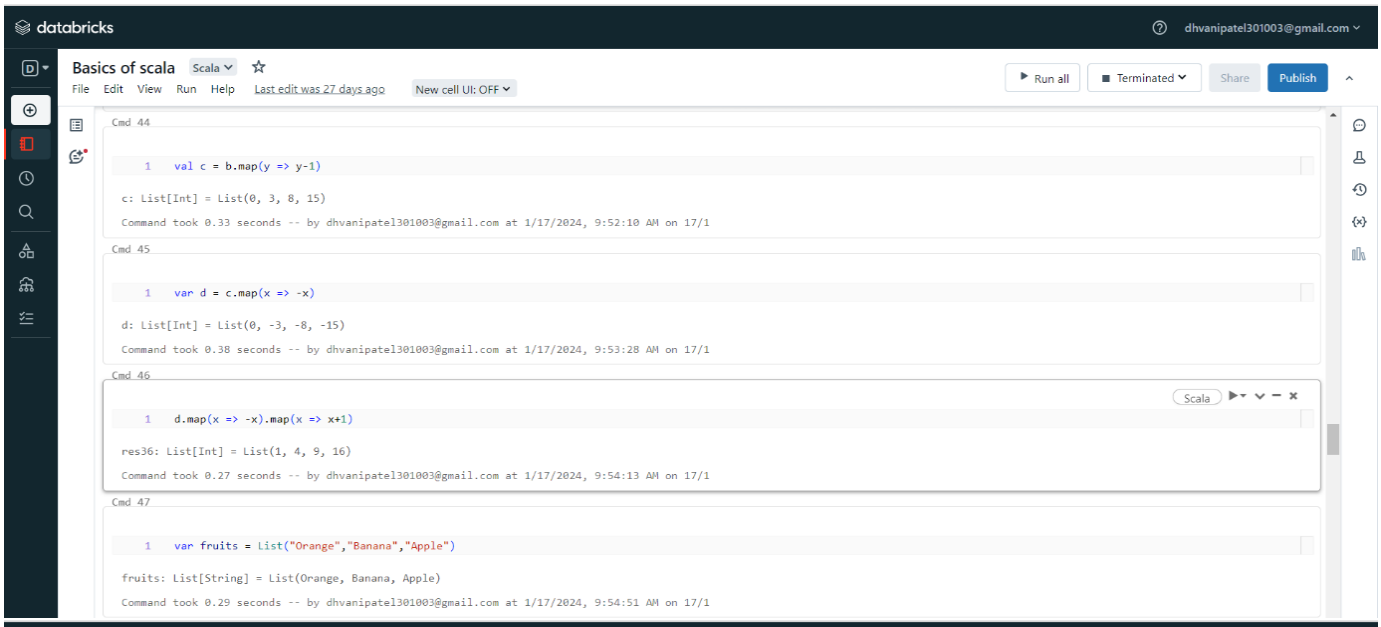

databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 25
1 number(1)=10
Command took 0.31 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:38:04 AM on 17/1
Cmd 26
1 number
res21: Array[Int] = Array(1, 10, 3, 4, 5, 6, 7, 8, 9)
Command took 0.36 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:38:11 AM on 17/1
Cmd 27
1 number(2)="hi"
command-4470193419131412:1: error: type mismatch;
found : String("hi")
required: Int
number(2)="hi"
^
Command took 0.11 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:39:05 AM on 17/1

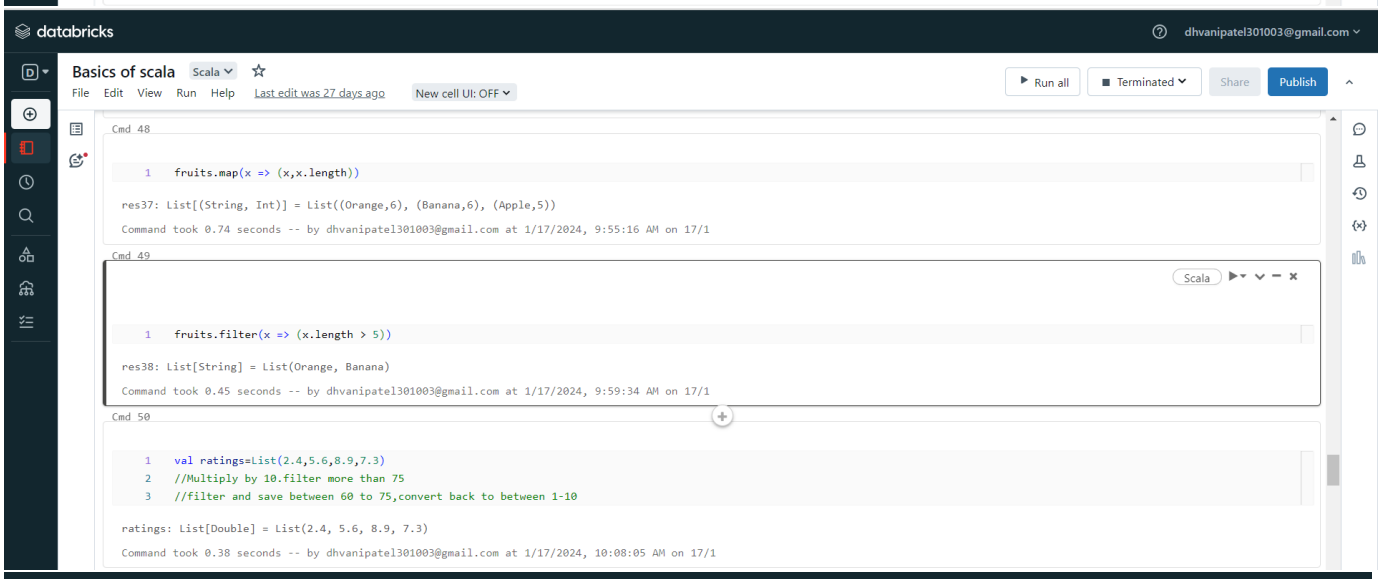

databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 28
1 import scala.collection.mutable.ArrayBuffercars
import scala.collection.mutable.ArrayBuffer
Command took 0.25 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:40:51 AM on 17/1
Cmd 29
1 var cars=new ArrayBuffer[String]()
cars: scala.collection.mutable.ArrayBuffer[String] = ArrayBuffer()
Command took 1.09 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:41:30 AM on 17/1
Cmd 30
1 cars += "BMW"
res23: scala.collection.mutable.ArrayBuffer[String] = ArrayBuffer(BMW)
Command took 0.36 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:41:56 AM on 17/1
Cmd 31
1 cars.append("Jaguar")
Command took 0.30 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:42:29 AM on 17/1

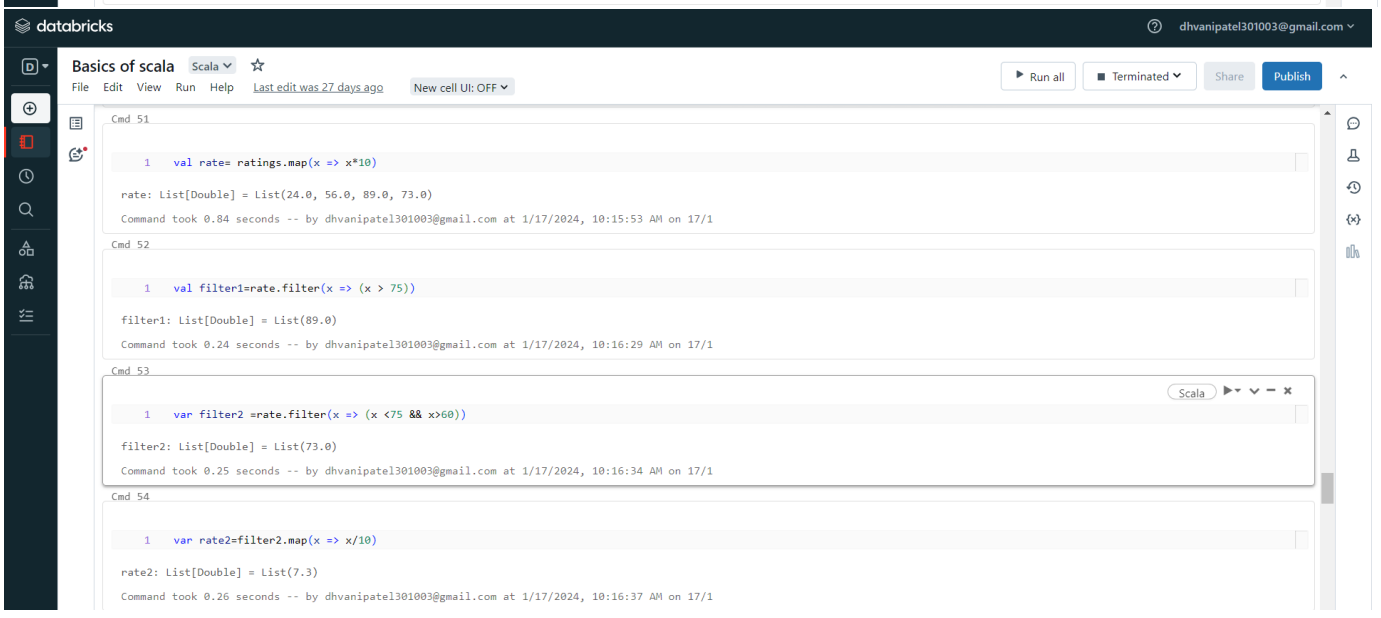

databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 32
1 cars
res25: scala.collection.mutable.ArrayBuffer[String] = ArrayBuffer(BMW, Jaguar)
Command took 0.40 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:42:38 AM on 17/1
Cmd 33
1 cars.length
res26: Int = 2
Command took 0.32 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:42:50 AM on 17/1
Cmd 34
1 cars += "Jaguar"
res27: scala.collection.mutable.ArrayBuffer[String] = ArrayBuffer(BMW, Jaguar, Jaguar)
Command took 0.43 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:43:03 AM on 17/1
Cmd 35
1 cars.trimEnd(1)
Command took 0.24 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:43:18 AM on 17/1

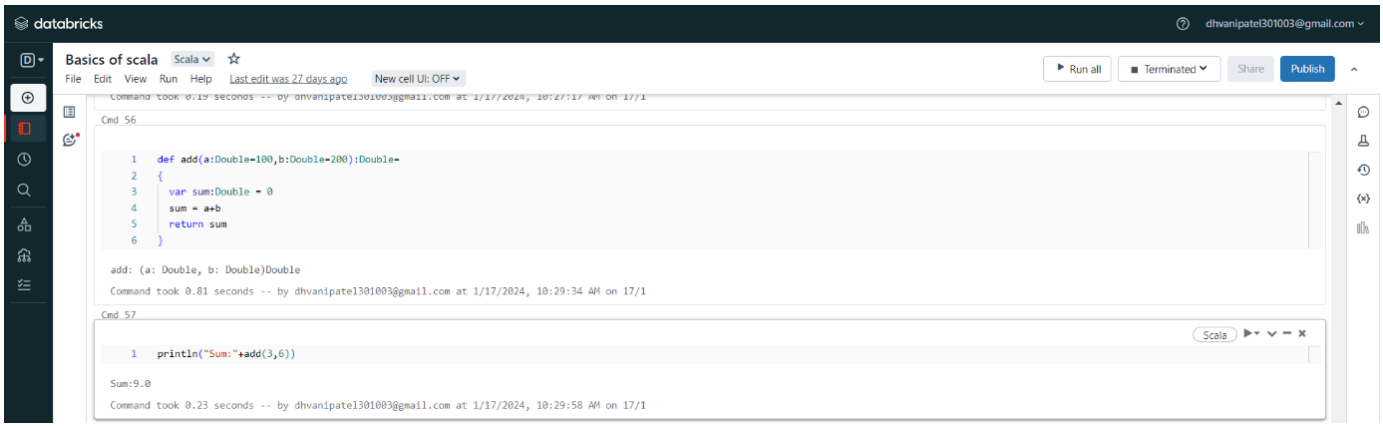

databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 36
1 cars
res29: scala.collection.mutable.ArrayBuffer[String] = ArrayBuffer(BMW, Jaguar)
Command took 0.96 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:43:24 AM on 17/1
Cmd 37
Scala
1 cars.trimStart(1)
Command took 0.28 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:45:56 AM on 17/1
Cmd 38
1 cars
res31: scala.collection.mutable.ArrayBuffer[String] = ArrayBuffer(Jaguar)
Command took 0.37 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:46:00 AM on 17/1
Cmd 39
1 cars.insert(1,"Bentley")
Command took 0.27 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:49:49 AM on 17/1

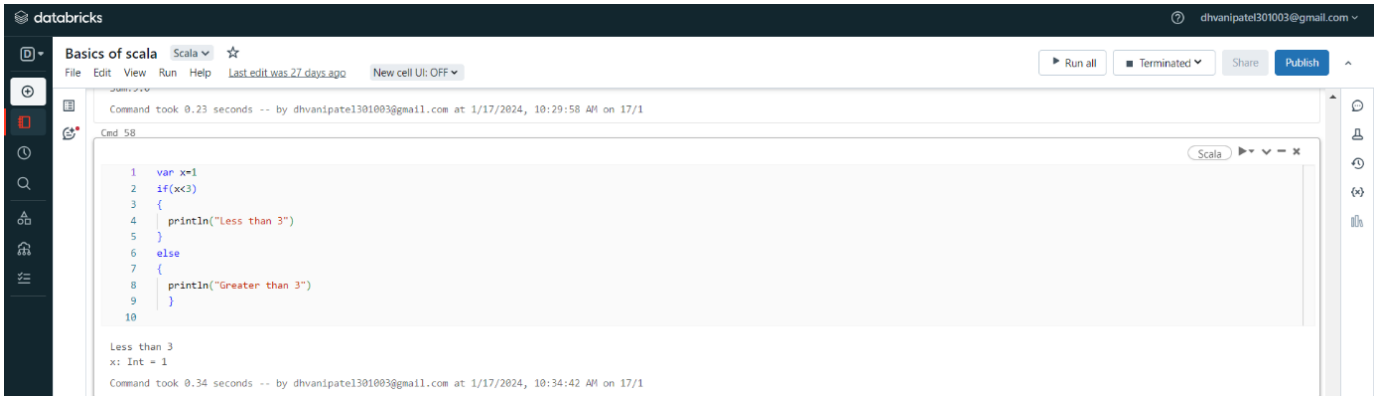

databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 40
1 cars
res34: scala.collection.mutable.ArrayBuffer[String] = ArrayBuffer(Jaguar, Bentley)
Command took 0.44 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:49:54 AM on 17/1
Cmd 41
1 //Transform and map
2 num
res35: List[Int] = List(1, 2, 3, 4)
Command took 0.33 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:50:29 AM on 17/1
Cmd 42
1 val a =num.map(x => x+1)
a: List[Int] = List(2, 3, 4, 5)
Command took 0.61 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:51:04 AM on 17/1
Cmd 43
1 val b = num.map(y => y*y)
b: List[Int] = List(1, 4, 9, 16)
Command took 0.76 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:51:36 AM on 17/1

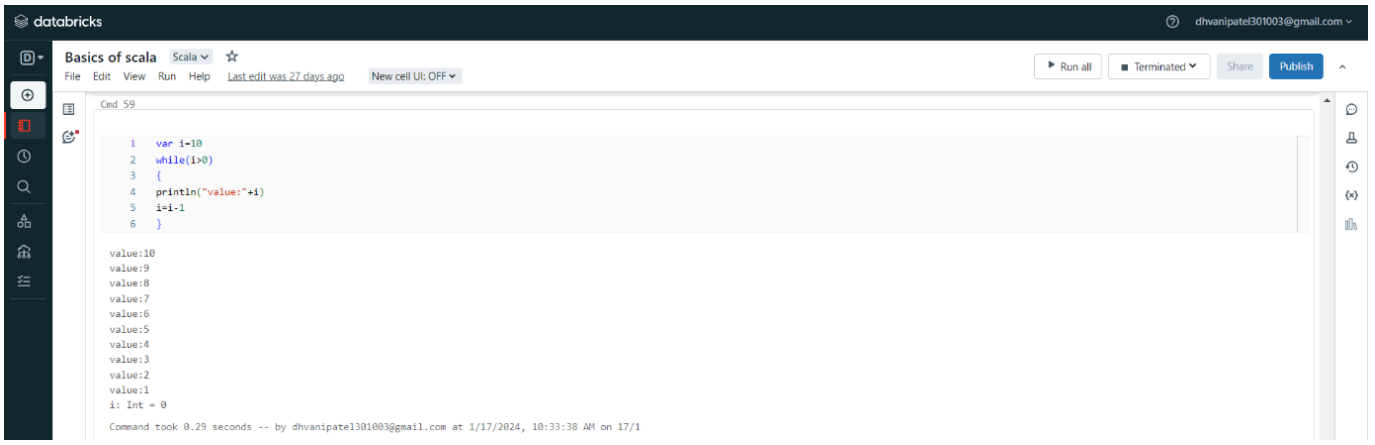

databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 44
1 val c = b.map(y => y-1)
c: List[Int] = List(0, 3, 8, 15)
Command took 0.33 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:52:10 AM on 17/1
Cmd 45
1 var d = c.map(x => -x)
d: List[Int] = List(0, -3, -8, -15)
Command took 0.38 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:53:28 AM on 17/1
Cmd 46
Scala
1 d.map(x => -x).map(x => x+1)
res36: List[Int] = List(1, 4, 9, 16)
Command took 0.27 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:54:13 AM on 17/1
Cmd 47
1 var fruits = List("Orange","Banana","Apple")
fruits: List[String] = List(Orange, Banana, Apple)
Command took 0.29 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:54:51 AM on 17/1

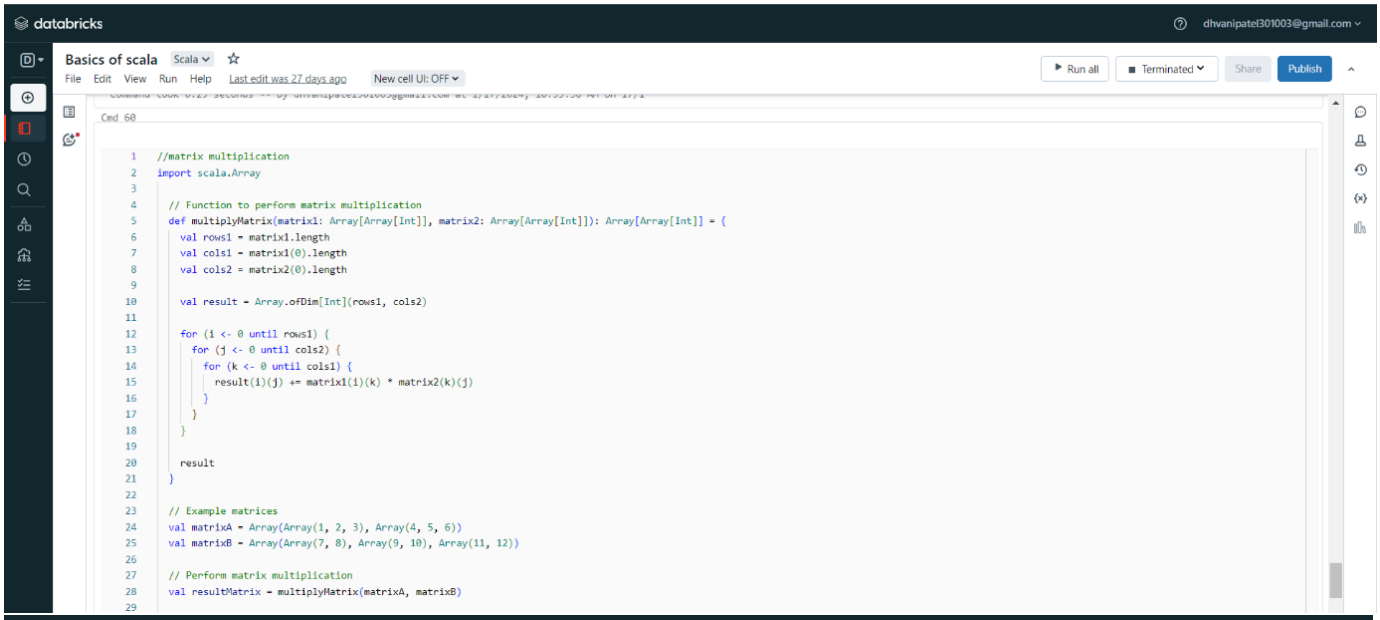

databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 48
1 fruits.map(x => (x,x.length))
res37: List[(String, Int)] = List((Orange,6), (Banana,6), (Apple,5))
Command took 0.74 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:55:16 AM on 17/1
Cmd 49
Scala
1 fruits.filter(x => (x.length > 5))
res38: List[String] = List(Orange, Banana)
Command took 0.45 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 9:59:34 AM on 17/1
Cmd 50
1 val ratings=List(2.4,5.6,8.9,7.3)
2 //Multiply by 10.filter more than 75
3 //filter and save between 60 to 75,convert back to between 1-10
ratings: List[Double] = List(2.4, 5.6, 8.9, 7.3)
Command took 0.38 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 10:08:05 AM on 17/1

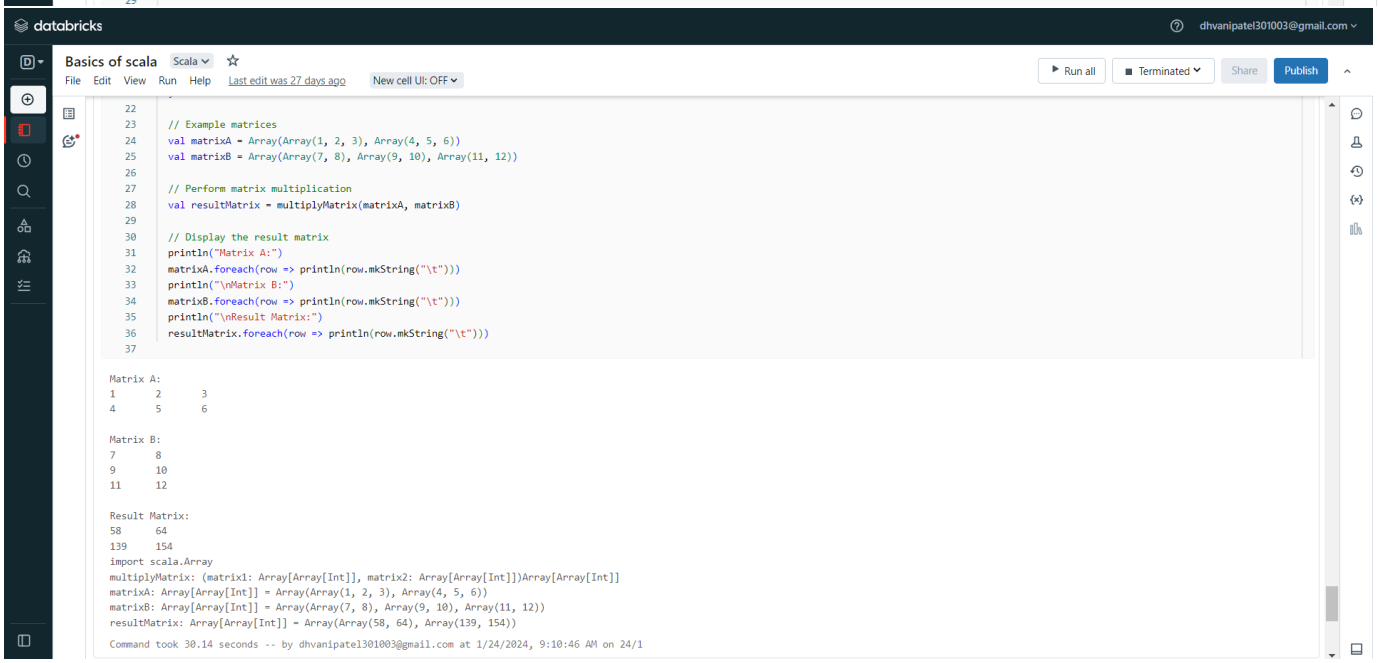

databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 51
1 val rate= ratings.map(x => x*10)
rate: List[Double] = List(24.0, 56.0, 89.0, 73.0)
Command took 0.84 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 10:15:53 AM on 17/1
Cmd 52
1 val filter1=rate.filter(x => (x > 75))
filter1: List[Double] = List(89.0)
Command took 0.24 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 10:16:29 AM on 17/1
Cmd 53
Scala
1 var filter2 =rate.filter(x => (x <75 && x>60))
filter2: List[Double] = List(73.0)
Command took 0.25 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 10:16:34 AM on 17/1
Cmd 54
1 var rate2=filter2.map(x => x/10)
rate2: List[Double] = List(7.3)
Command took 0.26 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 10:16:37 AM on 17/1


databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 56
def add(a:Double=100,b:Double=200):Double=
{
var sum:Double = 0
sum = a+b
return sum
}
add: (a: Double, b: Double)Double
Command took 0.81 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 10:29:34 AM on 17/1
Cmd 57
Scala
println("Sum:"+add(3,6))
Sum:9.0
Command took 0.23 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 10:29:58 AM on 17/1


databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Command took 0.23 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 10:29:58 AM on 17/1
Cmd 58
Scala
var x=1
if(x<3)
{
println("Less than 3")
}
else
{
println("Greater than 3")
}
Less than 3
x: Int = 1
Command took 0.34 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 10:34:42 AM on 17/1


databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 59
var i=10
while(i>0)
{
println("value:"+i)
i=i-1
}
value:10
value:9
value:8
value:7
value:6
value:5
value:4
value:3
value:2
value:1
i: Int = 0
Command took 0.29 seconds -- by dhvanipatel301003@gmail.com at 1/17/2024, 10:33:38 AM on 17/1


databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 60
//matrix multiplication
import scala.Array
// Function to perform matrix multiplication
def multiplyMatrix(matrix1: Array[Array[Int]], matrix2: Array[Array[Int]]): Array[Array[Int]] = {
val rows1 = matrix1.length
val cols1 = matrix1(0).length
val cols2 = matrix2(0).length
val result = Array.ofDim[Int](rows1, cols2)
for (i <- 0 until rows1) {
for (j <- 0 until cols2) {
for (k <- 0 until cols1) {
result(i)(j) += matrix1(i)(k) * matrix2(k)(j)
}
}
}
result
}
// Example matrices
val matrixA = Array(Array(1, 2, 3), Array(4, 5, 6))
val matrixB = Array(Array(7, 8), Array(9, 10), Array(11, 12))
// Perform matrix multiplication
val resultMatrix = multiplyMatrix(matrixA, matrixB)

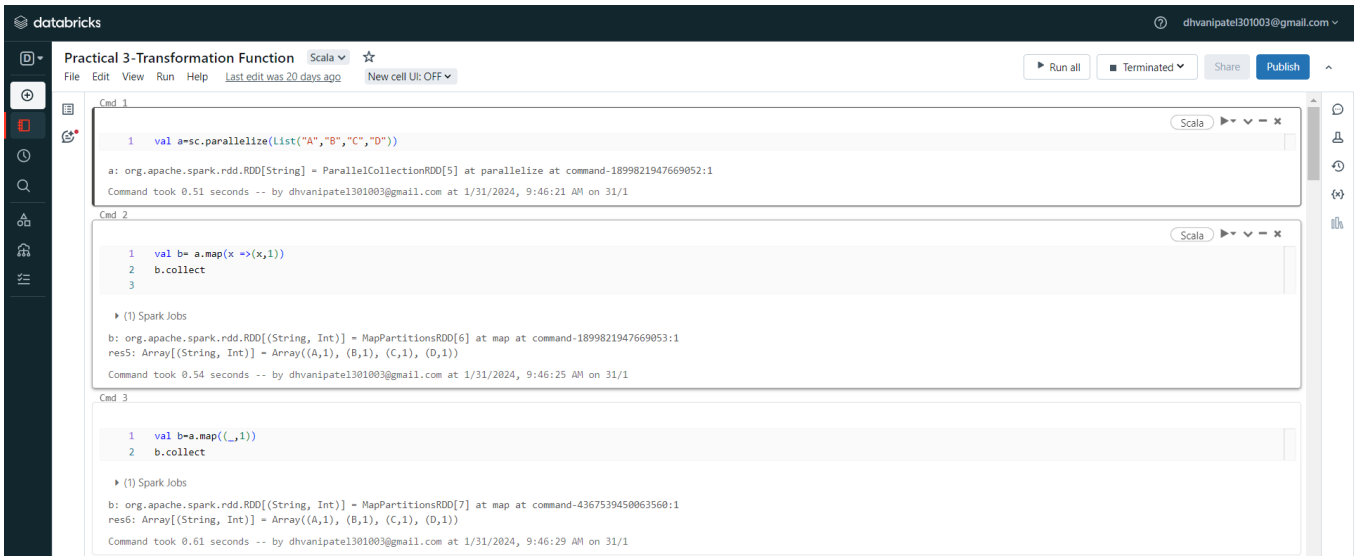

databricks
dhvanipatel301003@gmail.com
Basics of scala
Scala
File Edit View Run Help
Last edit was 27 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
// Example matrices
val matrixA = Array(Array(1, 2, 3), Array(4, 5, 6))
val matrixB = Array(Array(7, 8), Array(9, 10), Array(11, 12))
// Perform matrix multiplication
val resultMatrix = multiplyMatrix(matrixA, matrixB)
// Display the result matrix
println("Matrix A:")
matrixA.foreach(row => println(row.mkString("\t")))
println("\nMatrix B:")
matrixB.foreach(row => println(row.mkString("\t")))
println("\nResult Matrix:")
resultMatrix.foreach(row => println(row.mkString("\t")))
Matrix A:
1 2 3
4 5 6
Matrix B:
7 8
9 10
11 12
Result Matrix:
58 64
139 154
import scala.Array
multiplyMatrix: (matrix1: Array[Array[Int]], matrix2: Array[Array[Int]])Array[Array[Int]]
matrixA: Array[Array[Int]] = Array(Array(1, 2, 3), Array(4, 5, 6))
matrixB: Array[Array[Int]] = Array(Array(7, 8), Array(9, 10), Array(11, 12))
resultMatrix: Array[Array[Int]] = Array(Array(58, 64), Array(139, 154))
Command took 30.14 seconds -- by dhvanipatel301003@gmail.com at 1/24/2024, 9:10:46 AM on 24/1

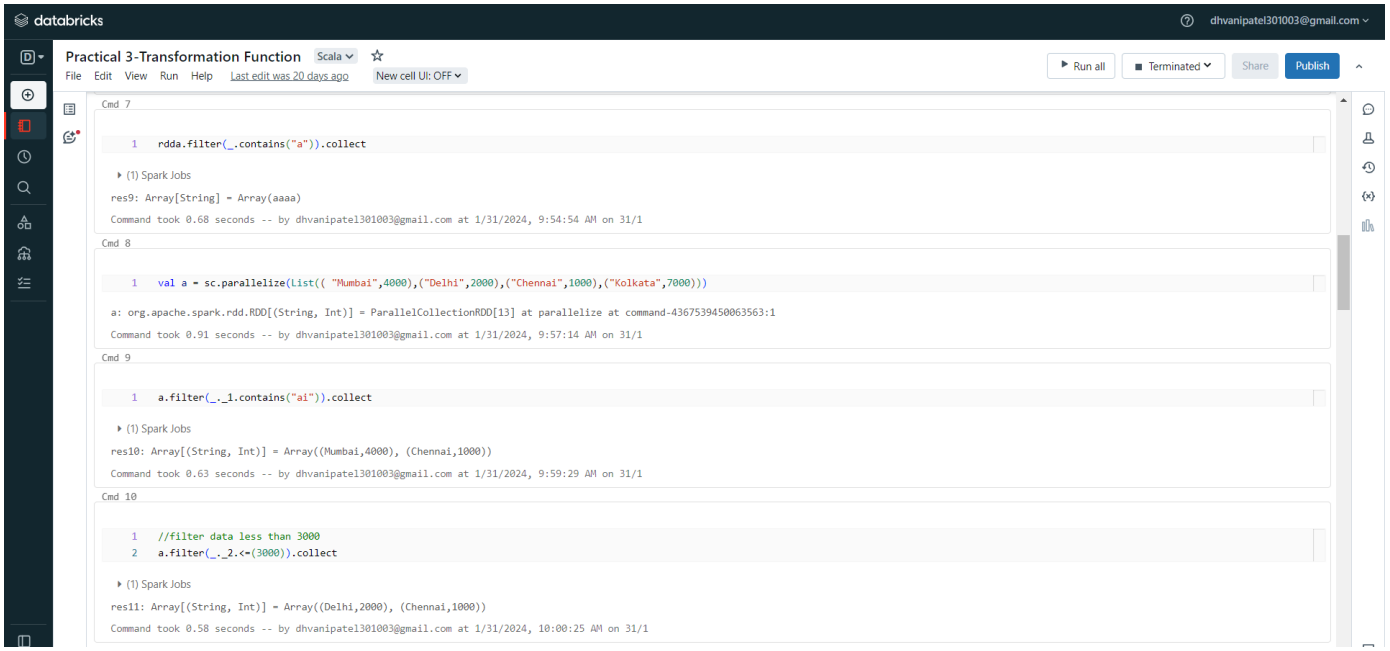
Name: Dhvani Patel

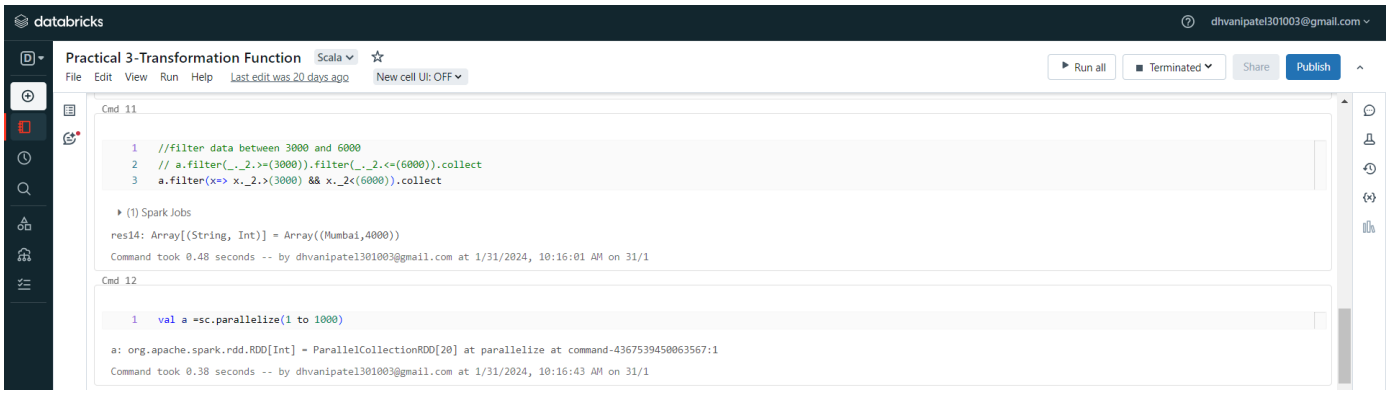
Roll no: 21BCP116

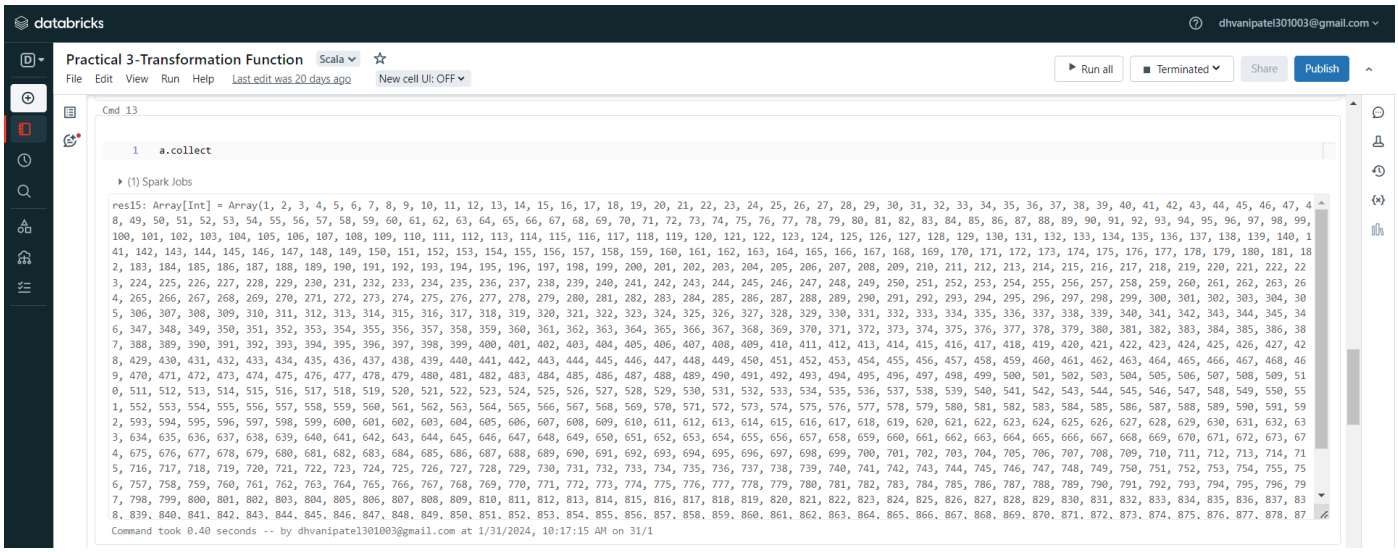
Big Data Analytics Lab

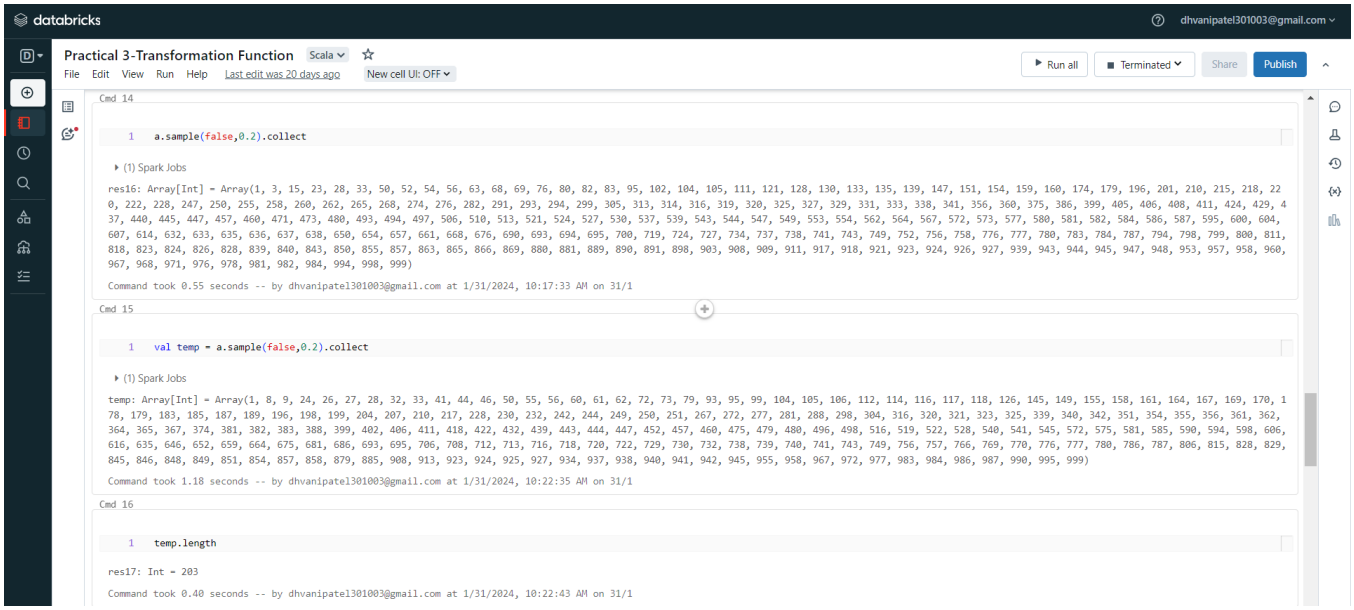
Practical 3 : Transformation Function

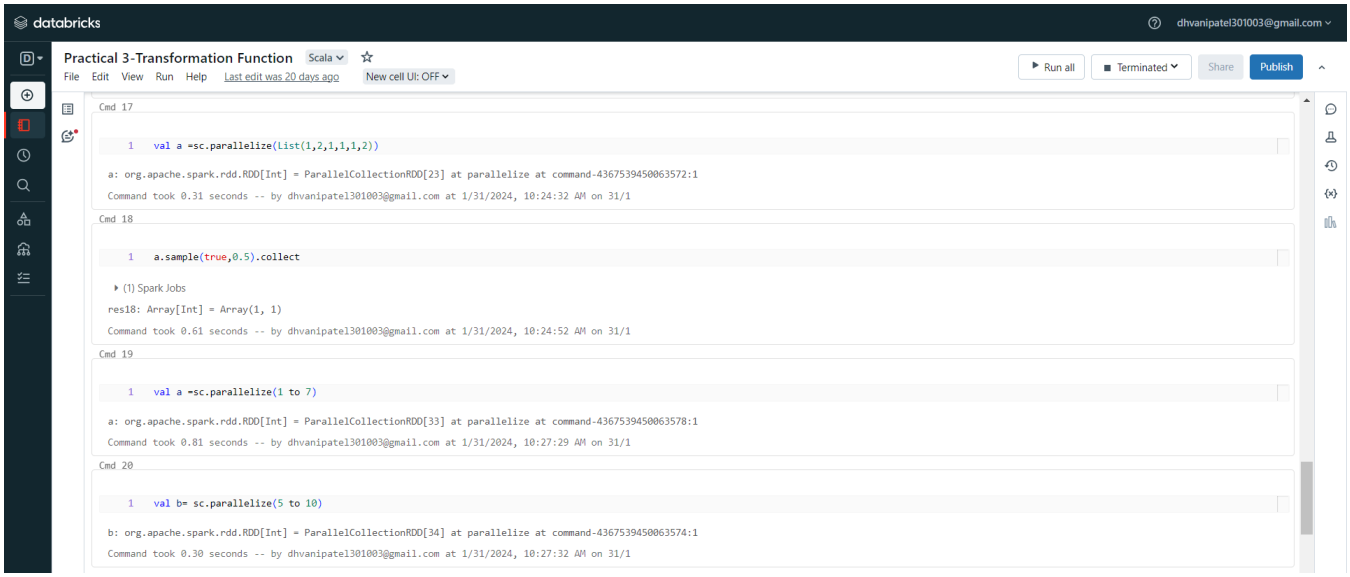


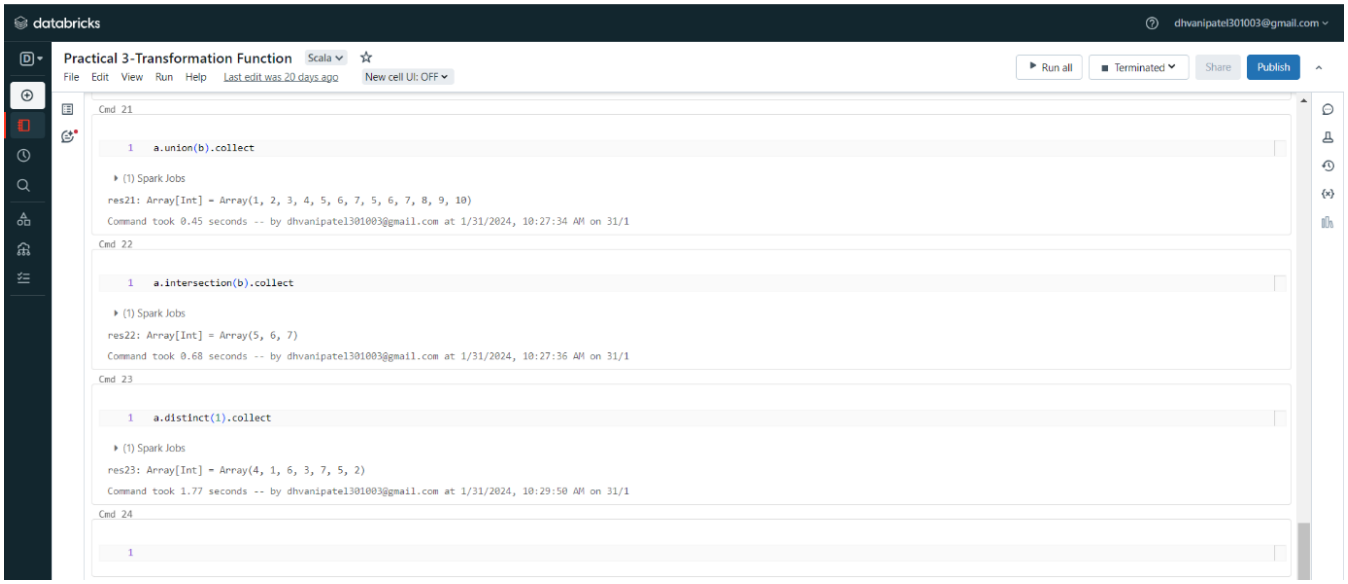

databricks
dhvanipatel301003@gmail.com
Practical 3-Transformation Function
Scala
File Edit View Run Help Last edit was 20 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 7
1 rdda.filter(_.contains("a")).collect
(1) Spark Jobs
res9: Array[String] = Array(aaaa)
Command took 0.68 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 9:54:54 AM on 31/1
Cmd 8
1 val a = sc.parallelize(List(("Mumbai",4000),("Delhi",2000),("Chennai",1000),("Kolkata",7000)))
a: org.apache.spark.rdd.RDD[(String, Int)] = ParallelCollectionRDD[13] at parallelize at command-4367539450063563:1
Command took 0.91 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 9:57:14 AM on 31/1
Cmd 9
1 a.filter(_._1.contains("ai")).collect
(1) Spark Jobs
res10: Array[(String, Int)] = Array((Mumbai,4000), (Chennai,1000))
Command took 0.63 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 9:59:29 AM on 31/1
Cmd 10
1 //filter data less than 3000
2 a.filter(_._2.<=(3000)).collect
(1) Spark Jobs
res11: Array[(String, Int)] = Array((Delhi,2000), (Chennai,1000))
Command took 0.58 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:00:25 AM on 31/1


databricks
dhvanipatel301003@gmail.com
Practical 3-Transformation Function
Scala
File Edit View Run Help Last edit was 20 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 11
1 //filter data between 3000 and 6000
2 // a.filter(_._2.>=(3000)).filter(_._2.<=(6000)).collect
3 a.filter(x=> x._2.>(3000) && x._2<(6000)).collect
(1) Spark Jobs
res14: Array[(String, Int)] = Array((Mumbai,4000))
Command took 0.48 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:16:01 AM on 31/1
Cmd 12
1 val a =sc.parallelize(1 to 1000)
a: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[20] at parallelize at command-4367539450063567:1
Command took 0.38 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:16:43 AM on 31/1


databricks
dhvanipatel301003@gmail.com
Practical 3-Transformation Function
Scala
File Edit View Run Help Last edit was 20 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 13
1 a.collect
(1) Spark Jobs
res15: Array[Int] = Array(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565, 566, 567, 568, 569, 570, 571, 572, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 597, 598, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 659, 660, 661, 662, 663, 664, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 679, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 692, 693, 694, 695, 696, 697, 698, 699, 700, 701, 702, 703, 704, 705, 706, 707, 708, 709, 710, 711, 712, 713, 714, 715, 716, 717, 718, 719, 720, 721, 722, 723, 724, 725, 726, 727, 728, 729, 730, 731, 732, 733, 734, 735, 736, 737, 738, 739, 740, 741, 742, 743, 744, 745, 746, 747, 748, 749, 750, 751, 752, 753, 754, 755, 756, 757, 758, 759, 760, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 772, 773, 774, 775, 776, 777, 778, 779, 780, 781, 782, 783, 784, 785, 786, 787, 788, 789, 790, 791, 792, 793, 794, 795, 796, 797, 798, 799, 800, 801, 802, 803, 804, 805, 806, 807, 808, 809, 810, 811, 812, 813, 814, 815, 816, 817, 818, 819, 820, 821, 822, 823, 824, 825, 826, 827, 828, 829, 830, 831, 832, 833, 834, 835, 836, 837, 838, 839, 840, 841, 842, 843, 844, 845, 846, 847, 848, 849, 850, 851, 852, 853, 854, 855, 856, 857, 858, 859, 860, 861, 862, 863, 864, 865, 866, 867, 868, 869, 870, 871, 872, 873, 874, 875, 876, 877, 878, 879, 880, 881, 882, 883, 884, 885, 886, 887, 888, 889, 890, 891, 892, 893, 894, 895, 896, 897, 898, 899, 900, 901, 902, 903, 904, 905, 906, 907, 908, 909, 910, 911, 912, 913, 914, 915, 916, 917, 918, 919, 920, 921, 922, 923, 924, 925, 926, 927, 928, 929, 930, 931, 932, 933, 934, 935, 936, 937, 938, 939, 940, 941, 942, 943, 944, 945, 946, 947, 948, 949, 950, 951, 952, 953, 954, 955, 956, 957, 958, 959, 960, 961, 962, 963, 964, 965, 966, 967, 968, 969, 970, 971, 972, 973, 974, 975, 976, 977, 978, 979, 980, 981, 982, 983, 984, 985, 986, 987, 988, 989, 990, 991, 992, 993, 994, 995, 996, 997, 998, 999, 1000)
Command took 0.40 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:17:15 AM on 31/1


databricks
dhvanipatel301003@gmail.com
Practical 3-Transformation Function
Scala
File Edit View Run Help Last edit was 20 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 14
1 a.sample(false,0.2).collect
(1) Spark Jobs
res16: Array[Int] = Array(1, 3, 15, 23, 28, 33, 50, 52, 54, 56, 63, 68, 69, 76, 80, 82, 83, 95, 102, 104, 105, 111, 121, 128, 130, 133, 135, 139, 147, 151, 154, 159, 160, 174, 179, 196, 201, 210, 215, 218, 220, 222, 228, 247, 250, 255, 258, 260, 262, 265, 268, 274, 276, 282, 291, 293, 294, 299, 305, 313, 314, 316, 319, 320, 325, 327, 329, 331, 333, 338, 341, 356, 360, 375, 386, 399, 405, 406, 408, 411, 424, 429, 437, 440, 445, 447, 457, 460, 471, 473, 480, 493, 494, 497, 506, 510, 513, 521, 524, 527, 530, 537, 539, 543, 544, 547, 549, 553, 554, 562, 564, 567, 572, 573, 577, 580, 581, 582, 584, 586, 587, 595, 600, 604, 607, 614, 632, 633, 635, 636, 637, 638, 650, 654, 657, 661, 668, 676, 690, 693, 694, 695, 700, 719, 724, 727, 734, 737, 738, 741, 743, 749, 752, 756, 758, 776, 777, 780, 783, 784, 787, 794, 798, 799, 800, 811, 818, 823, 824, 826, 828, 839, 840, 843, 850, 855, 857, 863, 865, 866, 869, 880, 881, 889, 890, 891, 898, 903, 908, 909, 911, 917, 918, 921, 923, 924, 926, 927, 939, 943, 944, 945, 947, 948, 953, 957, 958, 960, 967, 968, 971, 976, 978, 981, 982, 984, 994, 998, 999)
Command took 0.55 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:17:33 AM on 31/1
Cmd 15
1 val temp = a.sample(false,0.2).collect
(1) Spark Jobs
temp: Array[Int] = Array(1, 8, 9, 24, 26, 27, 28, 32, 33, 41, 44, 46, 50, 55, 56, 60, 61, 62, 72, 73, 79, 93, 95, 99, 104, 105, 106, 112, 114, 116, 117, 118, 126, 145, 149, 155, 158, 161, 164, 167, 169, 170, 178, 179, 183, 185, 187, 189, 196, 198, 199, 204, 207, 210, 217, 228, 230, 232, 242, 244, 249, 250, 251, 267, 272, 277, 281, 288, 298, 304, 316, 320, 321, 323, 325, 339, 340, 342, 351, 354, 355, 356, 361, 362, 364, 365, 367, 374, 381, 382, 383, 388, 399, 402, 406, 411, 418, 422, 432, 439, 443, 444, 447, 452, 457, 460, 475, 479, 480, 496, 498, 516, 519, 522, 528, 540, 541, 545, 572, 575, 581, 585, 590, 594, 598, 606, 616, 635, 646, 652, 659, 664, 675, 681, 686, 693, 695, 706, 708, 712, 713, 716, 718, 720, 722, 729, 730, 732, 738, 739, 740, 741, 743, 749, 756, 757, 766, 769, 770, 776, 777, 780, 786, 787, 806, 815, 828, 829, 845, 846, 848, 849, 851, 854, 857, 858, 879, 885, 908, 913, 923, 924, 925, 927, 934, 937, 938, 940, 941, 942, 945, 955, 958, 967, 972, 977, 983, 984, 986, 987, 990, 995, 999)
Command took 1.18 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:22:35 AM on 31/1
Cmd 16
1 temp.length
res17: Int = 203
Command took 0.40 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:22:43 AM on 31/1


databricks
dhvanipatel301003@gmail.com
Practical 3-Transformation Function
Scala
File Edit View Run Help Last edit was 20 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 17
1 val a =sc.parallelize(List(1,2,1,1,1,2))
a: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[23] at parallelize at command-4367539450063572:1
Command took 0.31 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:24:32 AM on 31/1
Cmd 18
1 a.sample(true,0.5).collect
(1) Spark Jobs
res18: Array[Int] = Array(1, 1)
Command took 0.61 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:24:52 AM on 31/1
Cmd 19
1 val a =sc.parallelize(1 to 7)
a: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[33] at parallelize at command-4367539450063578:1
Command took 0.81 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:27:29 AM on 31/1
Cmd 20
1 val b= sc.parallelize(5 to 10)
b: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[34] at parallelize at command-4367539450063574:1
Command took 0.30 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:27:32 AM on 31/1

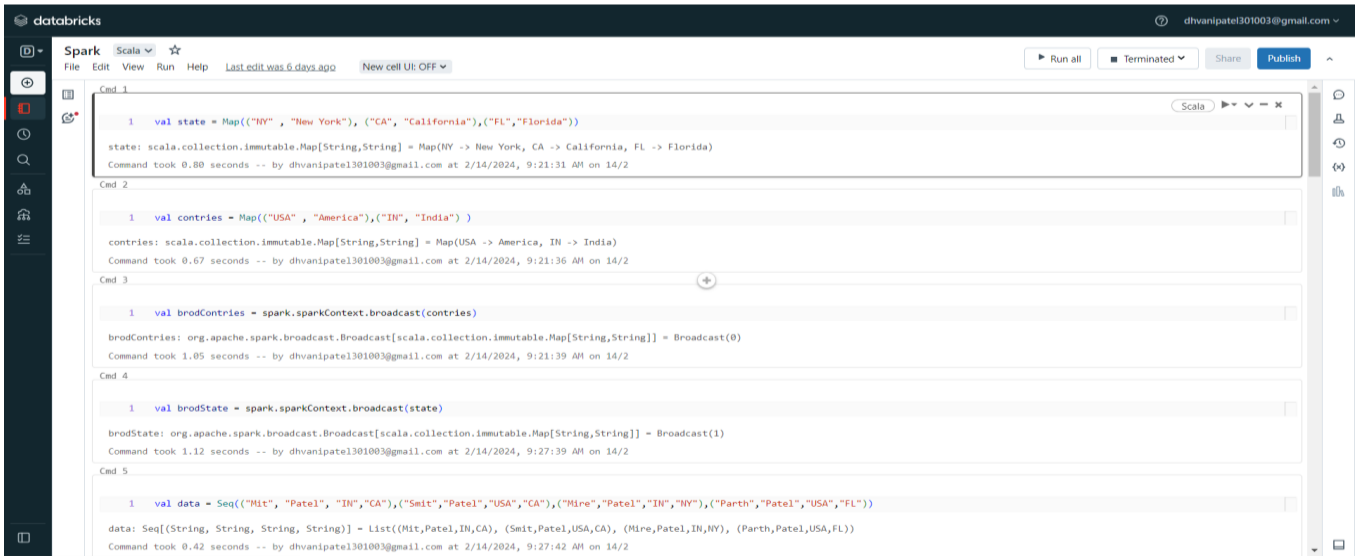

databricks
dhvanipatel301003@gmail.com
Practical 3-Transformation Function
Scala
File Edit View Run Help Last edit was 20 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 21
1 a.union(b).collect
(1) Spark Jobs
res21: Array[Int] = Array(1, 2, 3, 4, 5, 6, 7, 5, 6, 7, 8, 9, 10)
Command took 0.45 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:27:34 AM on 31/1
Cmd 22
1 a.intersection(b).collect
(1) Spark Jobs
res22: Array[Int] = Array(5, 6, 7)
Command took 0.68 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:27:36 AM on 31/1
Cmd 23
1 a.distinct(1).collect
(1) Spark Jobs
res23: Array[Int] = Array(4, 1, 6, 3, 7, 5, 2)
Command took 1.77 seconds -- by dhvanipatel301003@gmail.com at 1/31/2024, 10:29:50 AM on 31/1
Cmd 24
1

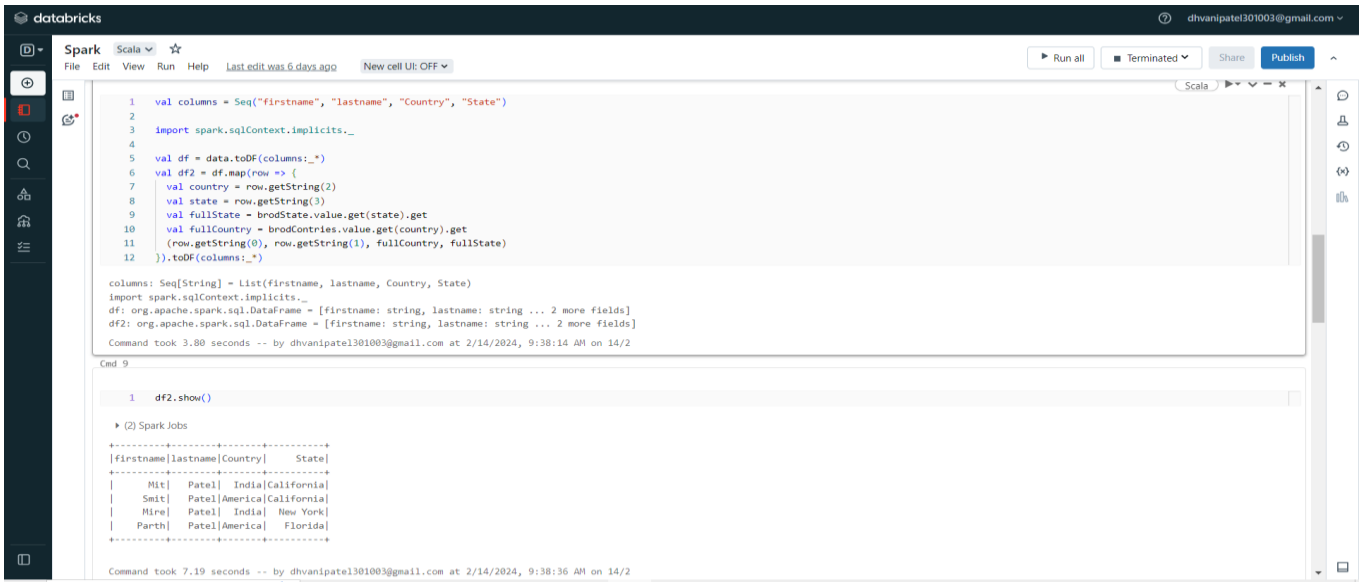
Name: Dhvani Patel

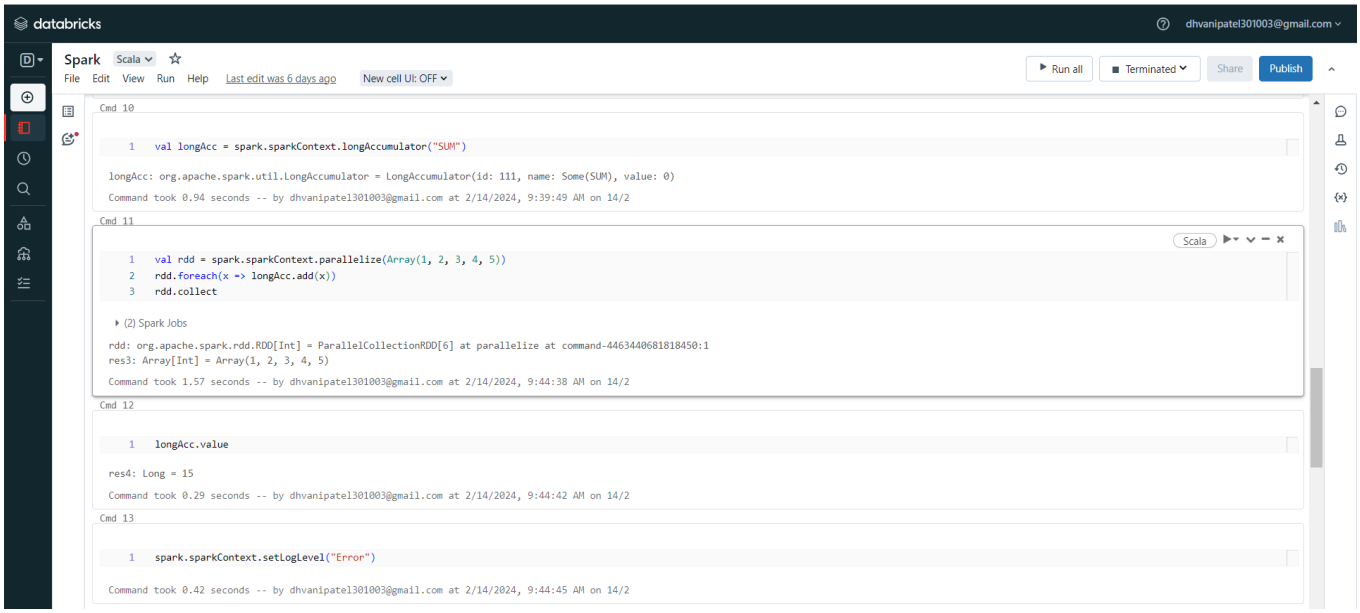
Roll no: 21BCP116

Big Data Analytics Lab

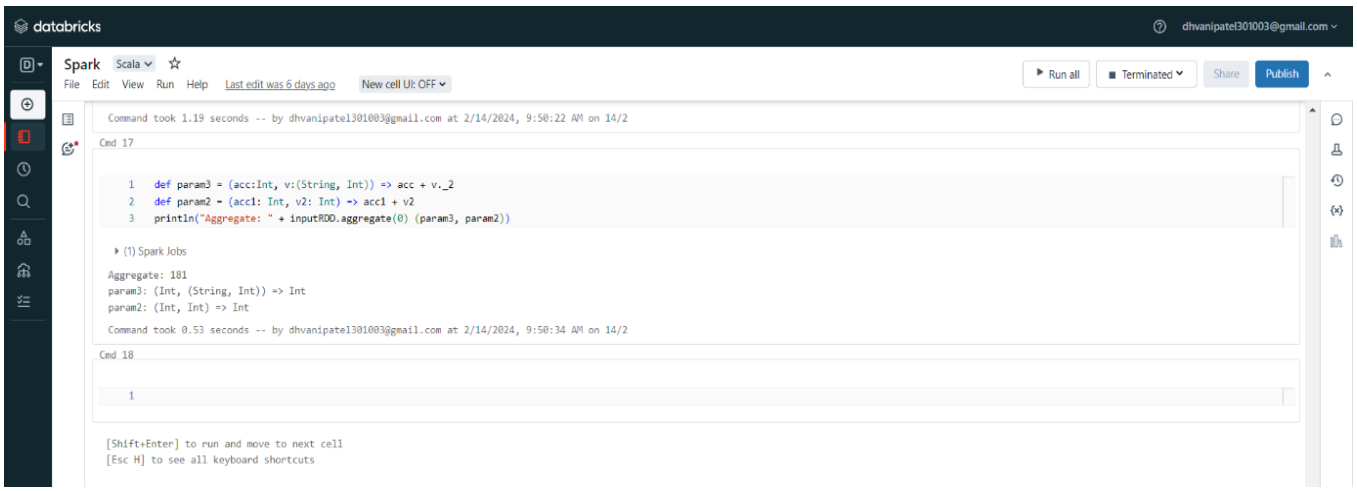
Practical 4 : Spark






databricks
dhvanipatel301003@gmail.com
Spark
Scala
File Edit View Run Help
Last edit was 6 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 10
1 val longAcc = spark.sparkContext.longAccumulator("SUM")
longAcc: org.apache.spark.util.LongAccumulator = LongAccumulator(id: 111, name: Some(SUM), value: 0)
Command took 0.94 seconds -- by dhvanipatel301003@gmail.com at 2/14/2024, 9:39:49 AM on 14/2
Cmd 11
Scala
1 val rdd = spark.sparkContext.parallelize(Array(1, 2, 3, 4, 5))
2 rdd.foreach(x => longAcc.add(x))
3 rdd.collect
(2) Spark Jobs
rdd: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[6] at parallelize at command-4463440681818450:1
res3: Array[Int] = Array(1, 2, 3, 4, 5)
Command took 1.57 seconds -- by dhvanipatel301003@gmail.com at 2/14/2024, 9:44:38 AM on 14/2
Cmd 12
1 longAcc.value
res4: Long = 15
Command took 0.29 seconds -- by dhvanipatel301003@gmail.com at 2/14/2024, 9:44:42 AM on 14/2
Cmd 13
1 spark.sparkContext.setLogLevel("Error")
Command took 0.42 seconds -- by dhvanipatel301003@gmail.com at 2/14/2024, 9:44:45 AM on 14/2


databricks
dhvanipatel301003@gmail.com
Spark
Scala
File Edit View Run Help
Last edit was 6 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Cmd 14
1 val inputRDD = spark.sparkContext.parallelize(List(("2", 1), ("B", 30), ("A", 20), ("B", 30), ("C", 40), ("B", 60)))
inputRDD: org.apache.spark.rdd.RDD[(String, Int)] = ParallelCollectionRDD[9] at parallelize at command-4463440681818449:1
Command took 0.33 seconds -- by dhvanipatel301003@gmail.com at 2/14/2024, 9:47:24 AM on 14/2
Cmd 15
1 val listRDD = spark.sparkContext.parallelize(List(1, 2, 3, 4, 5, 2, 3))
listRDD: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[10] at parallelize at command-4463440681818451:1
Command took 0.31 seconds -- by dhvanipatel301003@gmail.com at 2/14/2024, 9:47:27 AM on 14/2
Cmd 16
1 def param0 = (acc:Int, v:Int) => acc + v
2 def param1 = (acc1:Int, acc2:Int)=>acc1+acc2
3 println("Aggregate: " + listRDD.aggregate(0) (param0, param1))
(1) Spark Jobs
Aggregate: 20
param0: (Int, Int) => Int
param1: (Int, Int) => Int
Command took 1.19 seconds -- by dhvanipatel301003@gmail.com at 2/14/2024, 9:50:22 AM on 14/2

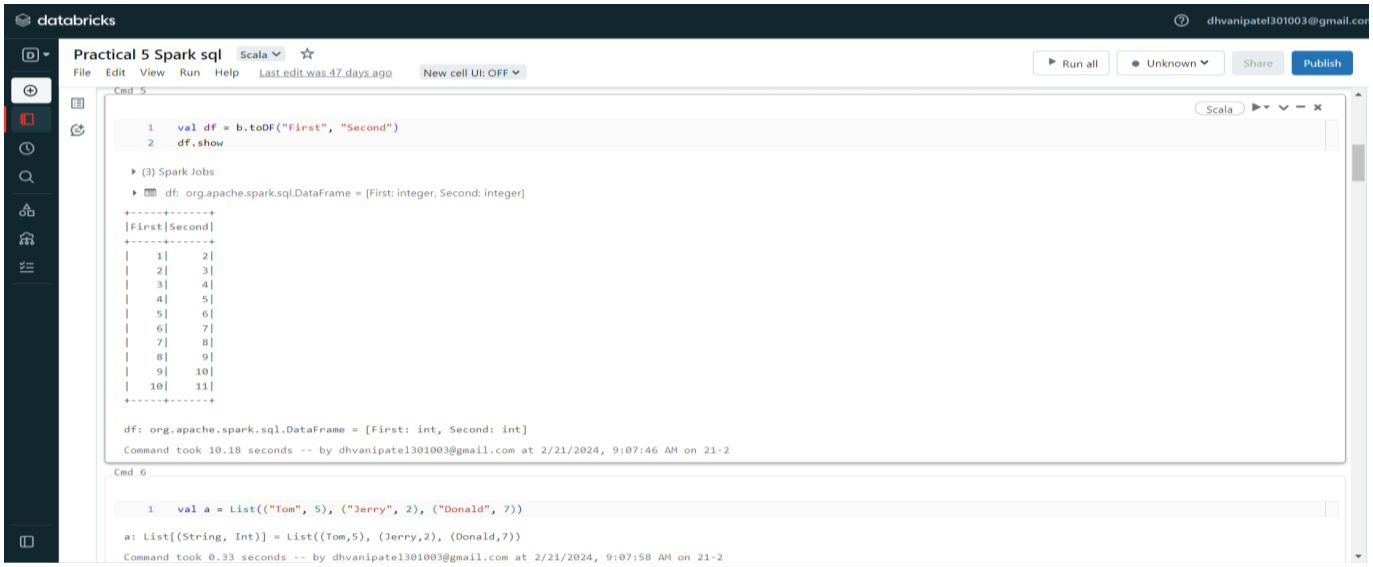

databricks
dhvanipatel301003@gmail.com
Spark
Scala
File Edit View Run Help
Last edit was 6 days ago
New cell UI: OFF
Run all
Terminated
Share
Publish
Command took 1.19 seconds -- by dhvanipatel301003@gmail.com at 2/14/2024, 9:50:22 AM on 14/2
Cmd 17
1 def param3 = (acc:Int, v:(String, Int)) => acc + v._2
2 def param2 = (acc1: Int, v2: Int) => acc1 + v2
3 println("Aggregate: " + inputRDD.aggregate(0) (param3, param2))
(1) Spark Jobs
Aggregate: 181
param3: (Int, (String, Int)) => Int
param2: (Int, Int) => Int
Command took 0.53 seconds -- by dhvanipatel301003@gmail.com at 2/14/2024, 9:50:34 AM on 14/2
Cmd 18
1
[Shift+Enter] to run and move to next cell
[Esc H] to see all keyboard shortcuts

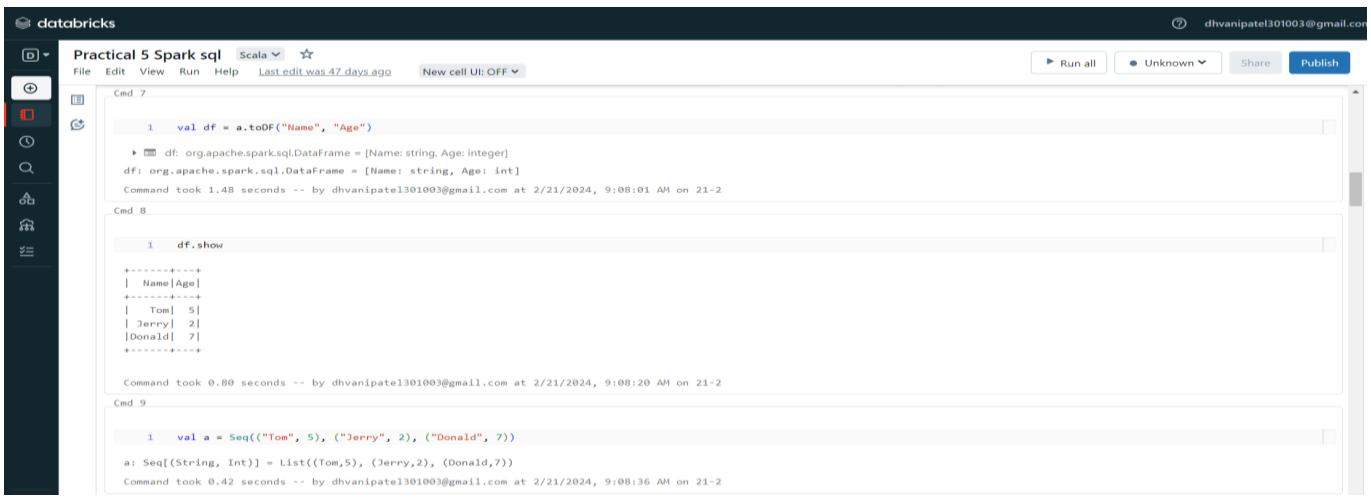
Name: Dhvani Patel

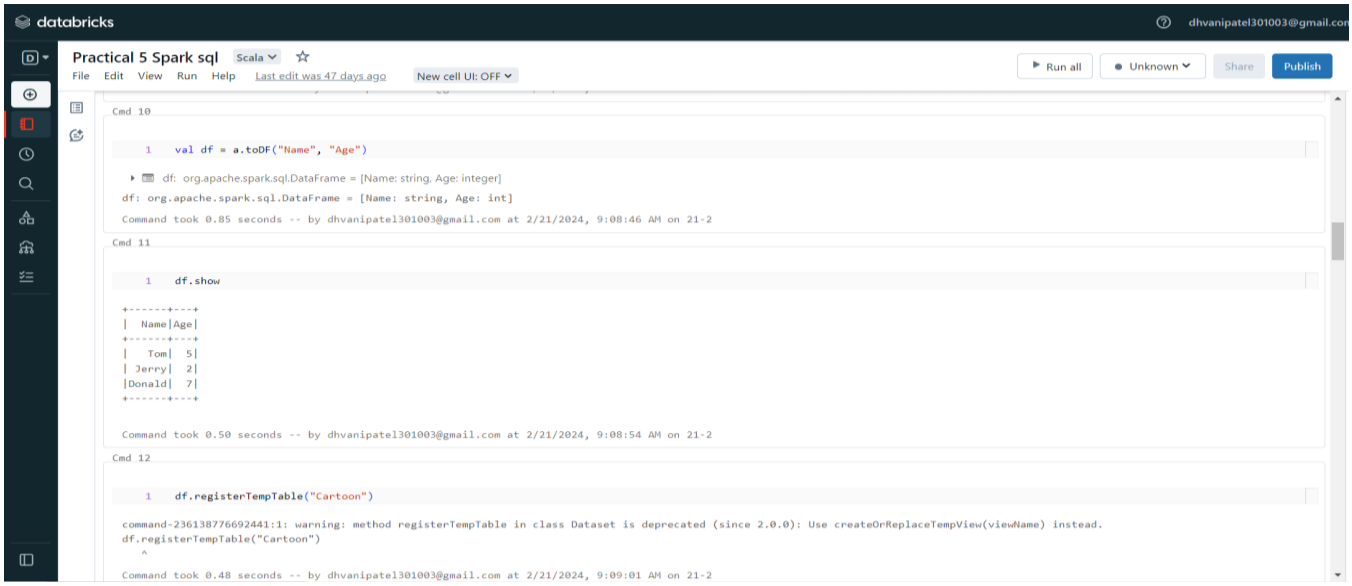
Roll no: 21BCP116

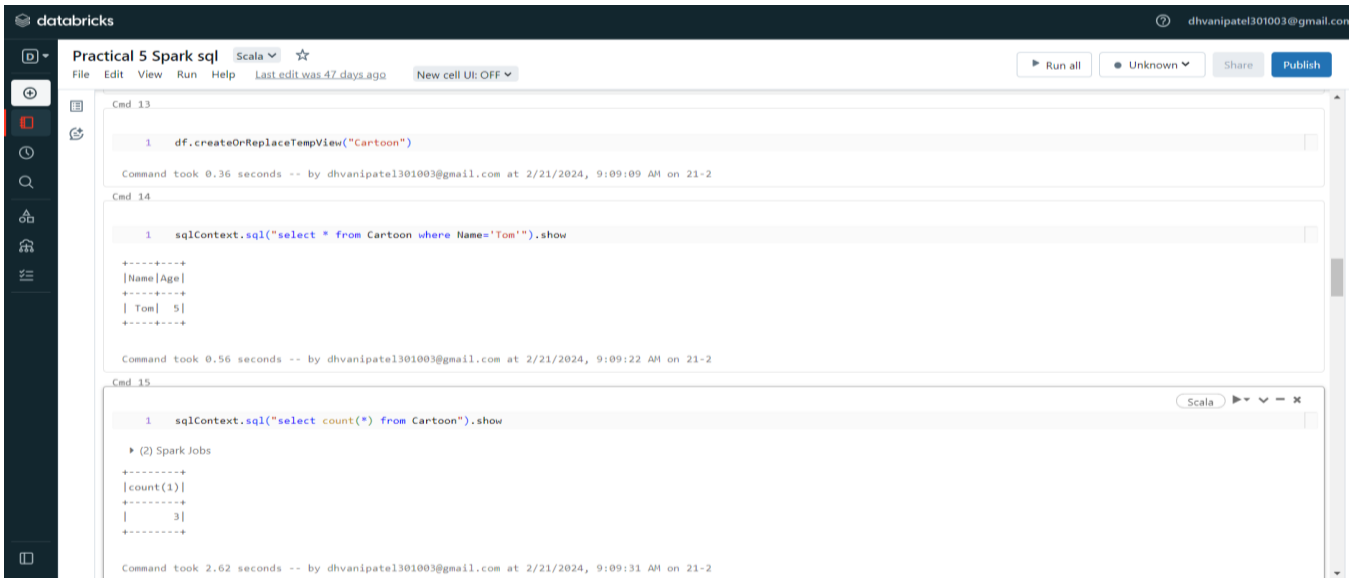
Big Data Analytics Lab

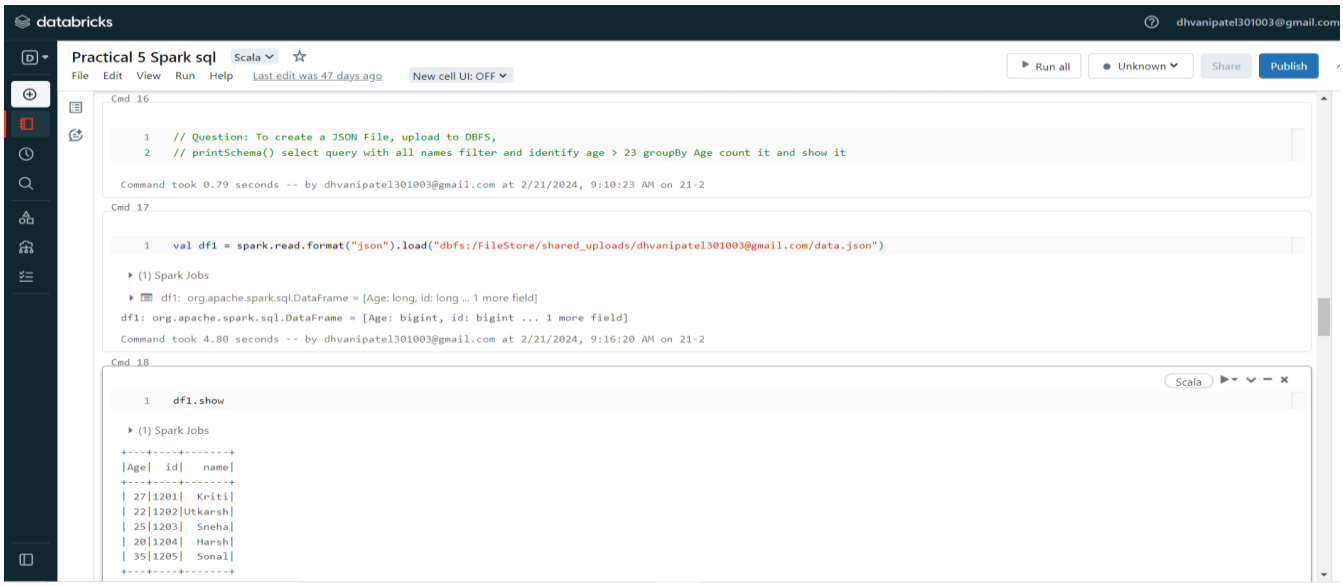
Practical 5: Spark SQL

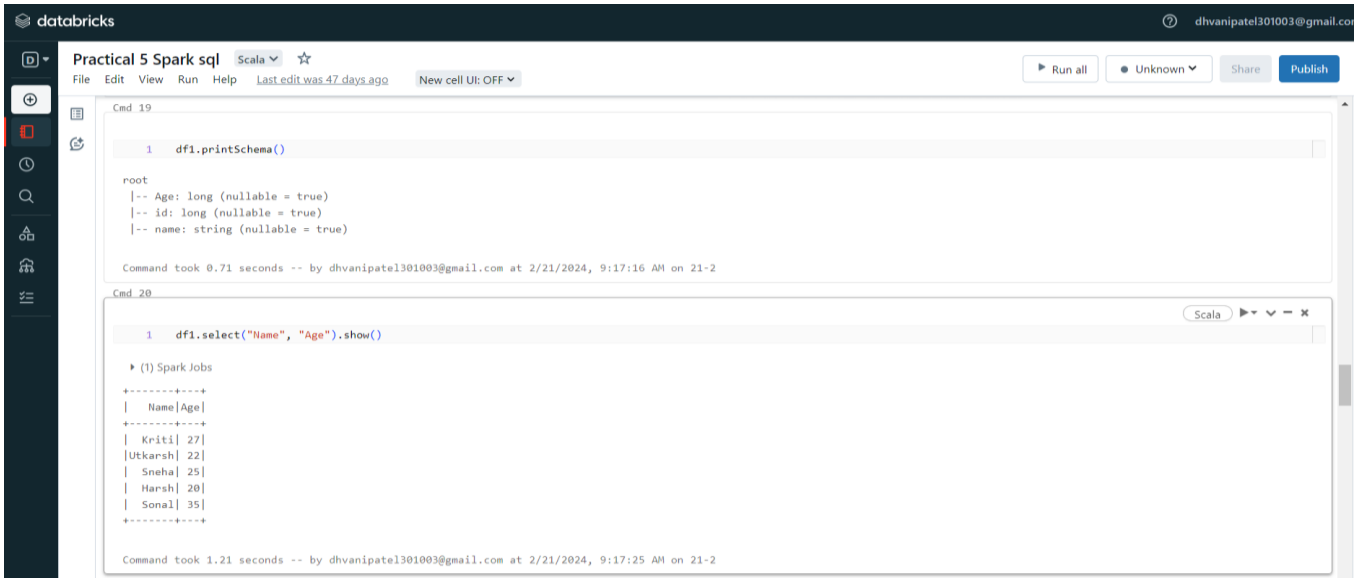


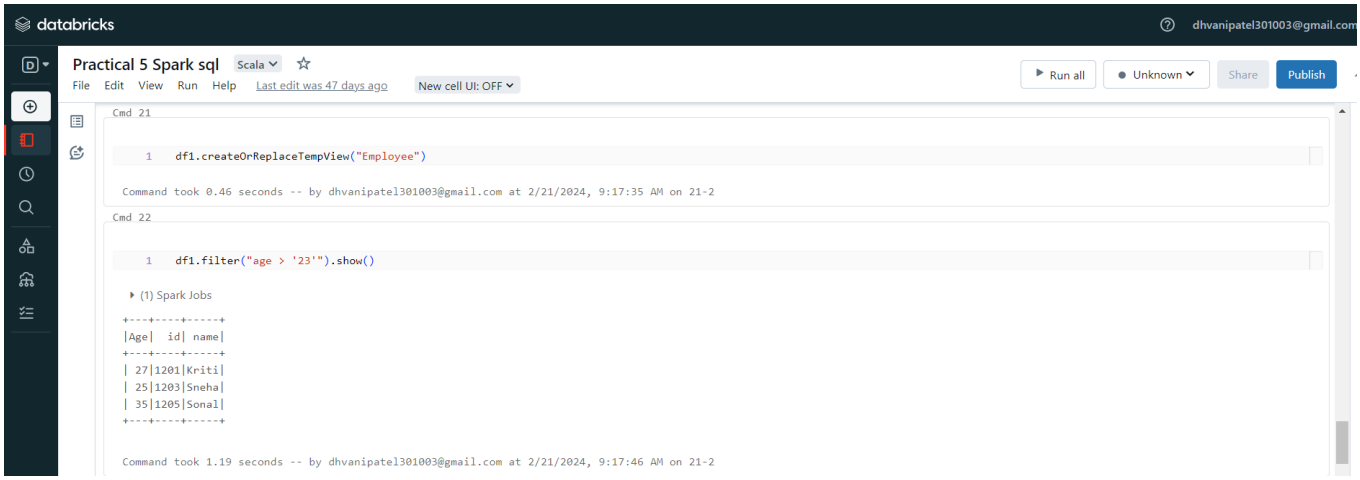


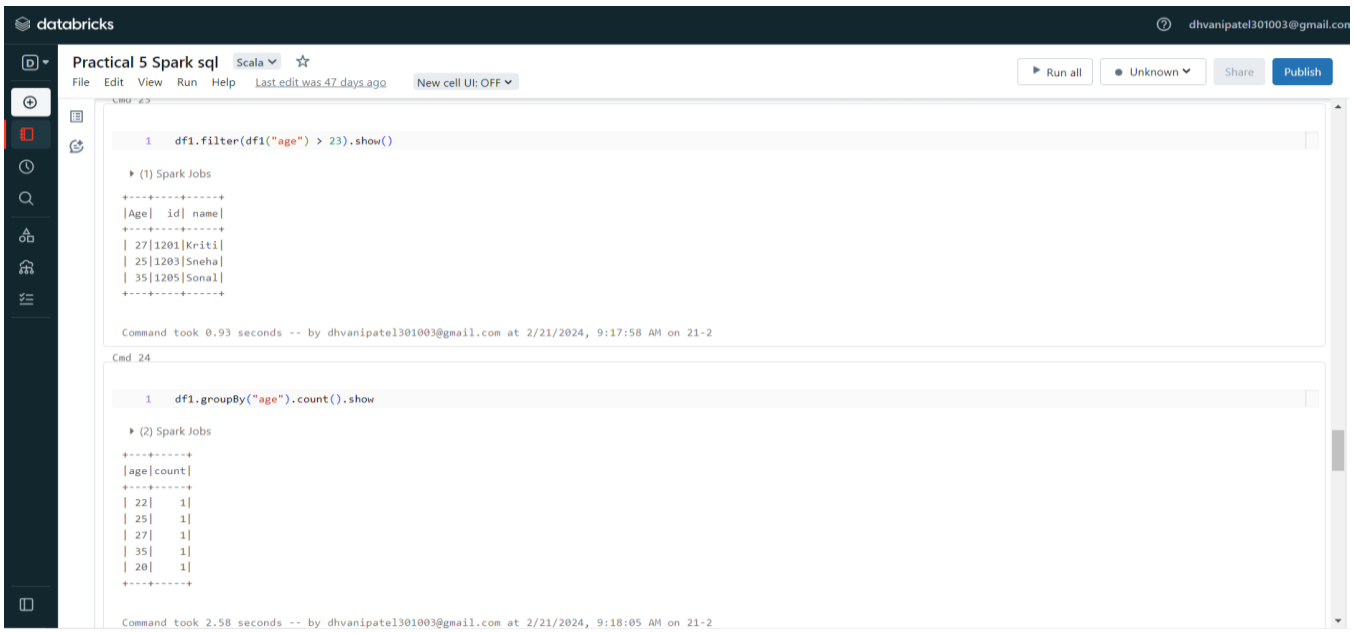

databricks
dhvanipatel301003@gmail.com
Practical 5 Spark sql
Scala
File Edit View Run Help
Last edit was 47 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 10
1 val df = a.toDF("Name", "Age")
df: org.apache.spark.sql.DataFrame = [Name: string, Age: integer]
df: org.apache.spark.sql.DataFrame = [Name: string, Age: int]
Command took 0.85 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:08:46 AM on 21-2
Cmd 11
1 df.show
+------+---+
| Name|Age|
+------+---+
| Tom| 5|
| Jerry| 2|
|Donald| 7|
+------+---+
Command took 0.50 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:08:54 AM on 21-2
Cmd 12
1 df.registerTempTable("Cartoon")
command-2361387766924l1:1: warning: method registerTempTable in class Dataset is deprecated (since 2.0.0): Use createOrReplaceTempView(viewName) instead.
df.registerTempTable("Cartoon")
^
Command took 0.48 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:09:01 AM on 21-2

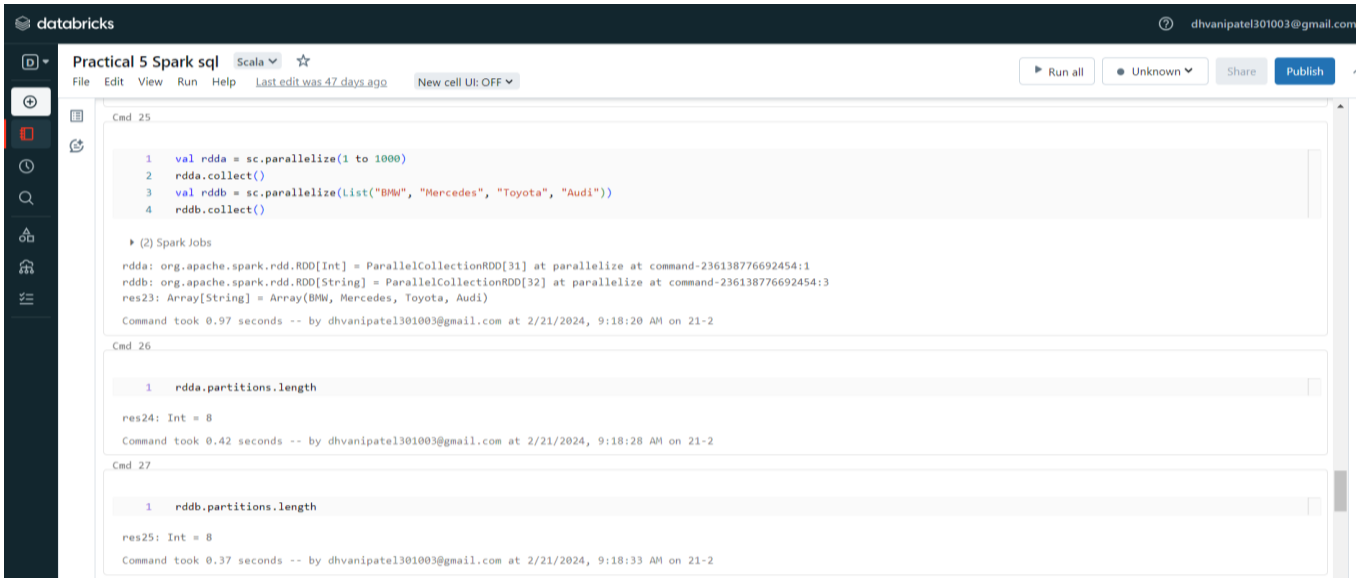

databricks
dhvanipatel301003@gmail.com
Practical 5 Spark sql
Scala
File Edit View Run Help
Last edit was 47 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 13
1 df.createOrReplaceTempView("Cartoon")
Command took 0.36 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:09:09 AM on 21-2
Cmd 14
1 sqlContext.sql("select * from Cartoon where Name='Tom'").show
+----+---+
|Name|Age|
+----+---+
| Tom| 5|
+----+---+
Command took 0.56 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:09:22 AM on 21-2
Cmd 15
Scala
1 sqlContext.sql("select count(*) from Cartoon").show
(2) Spark Jobs
+--------+
|count(1)|
+--------+
| 3|
+--------+
Command took 2.62 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:09:31 AM on 21-2

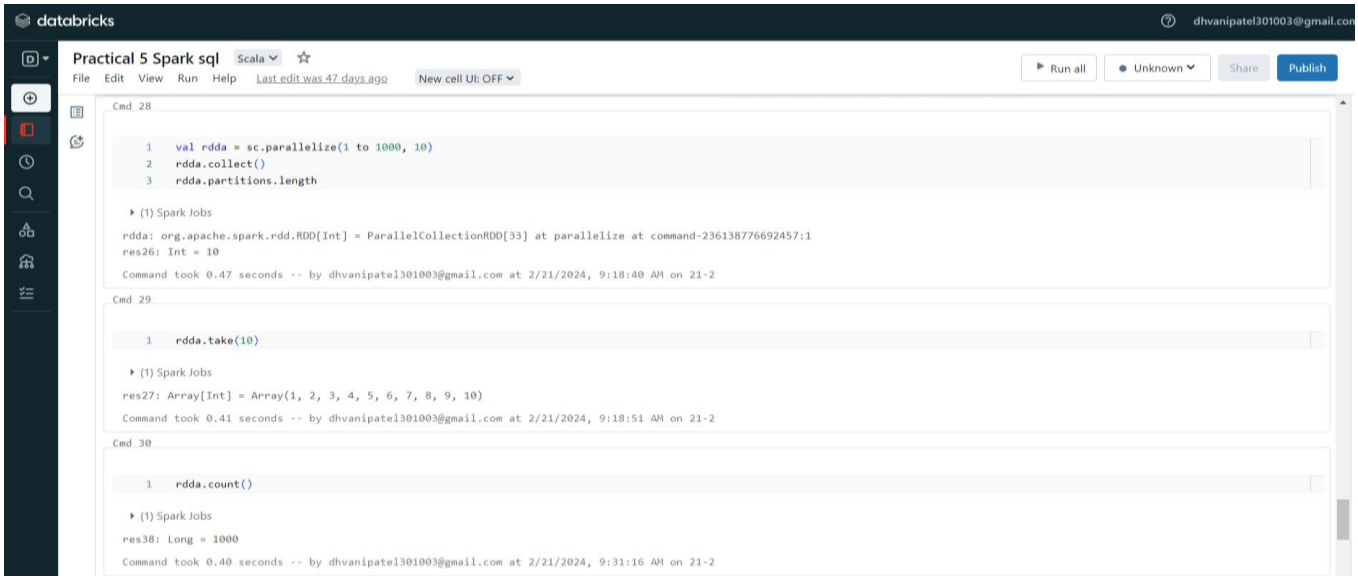

databricks
dhvanipatel301003@gmail.com
Practical 5 Spark sql
Scala
File Edit View Run Help
Last edit was 47 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 16
1 // Question: To create a JSON File, upload to DBFS,
2 // printSchema() select query with all names filter and identify age > 23 groupBy Age count it and show it
Command took 0.79 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:10:23 AM on 21-2
Cmd 17
1 val df1 = spark.read.format("json").load("dbfs:/FileStore/shared_uploads/dhvanipatel301003@gmail.com/data.json")
(1) Spark Jobs
df1: org.apache.spark.sql.DataFrame = [Age: long, id: long ... 1 more field]
df1: org.apache.spark.sql.DataFrame = [Age: bigint, id: bigint ... 1 more field]
Command took 4.80 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:16:20 AM on 21-2
Cmd 18
Scala
1 df1.show
(1) Spark Jobs
+---+----+-------+
|Age| id| name|
+---+----+-------+
27	1201	Kriti
22	1202	Utkarsh
25	1203	Sneha
20	1204	Harsh
35	1205	Sonal
+---+----+-------+

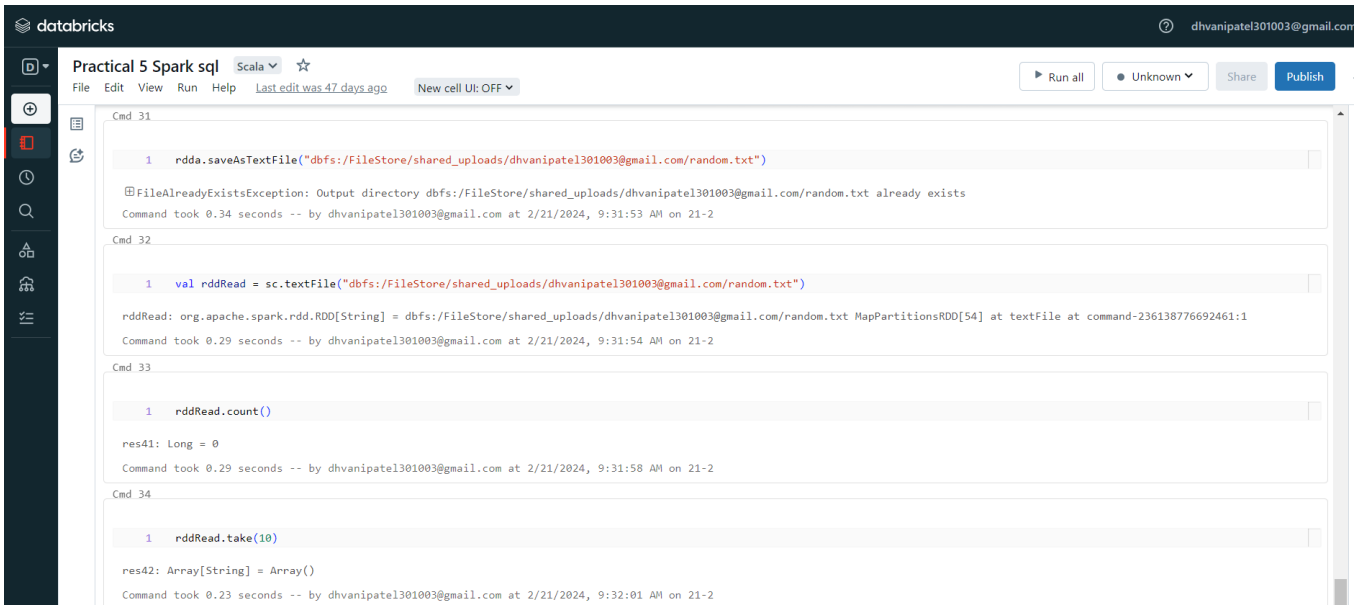

databricks
dhvanipatel301003@gmail.com
Practical 5 Spark sql
Scala
File Edit View Run Help Last edit was 47 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 19
1 df1.printSchema()
root
|-- Age: long (nullable = true)
|-- id: long (nullable = true)
|-- name: string (nullable = true)
Command took 0.71 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:17:16 AM on 21-2
Cmd 20
Scala
1 df1.select("Name", "Age").show()
(1) Spark Jobs
+-------+---+
| Name|Age|
+-------+---+
| Kriti| 27|
|Utkarsh| 22|
Sneha	25
Harsh	20
Sonal	35
+-------+---+
Command took 1.21 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:17:25 AM on 21-2

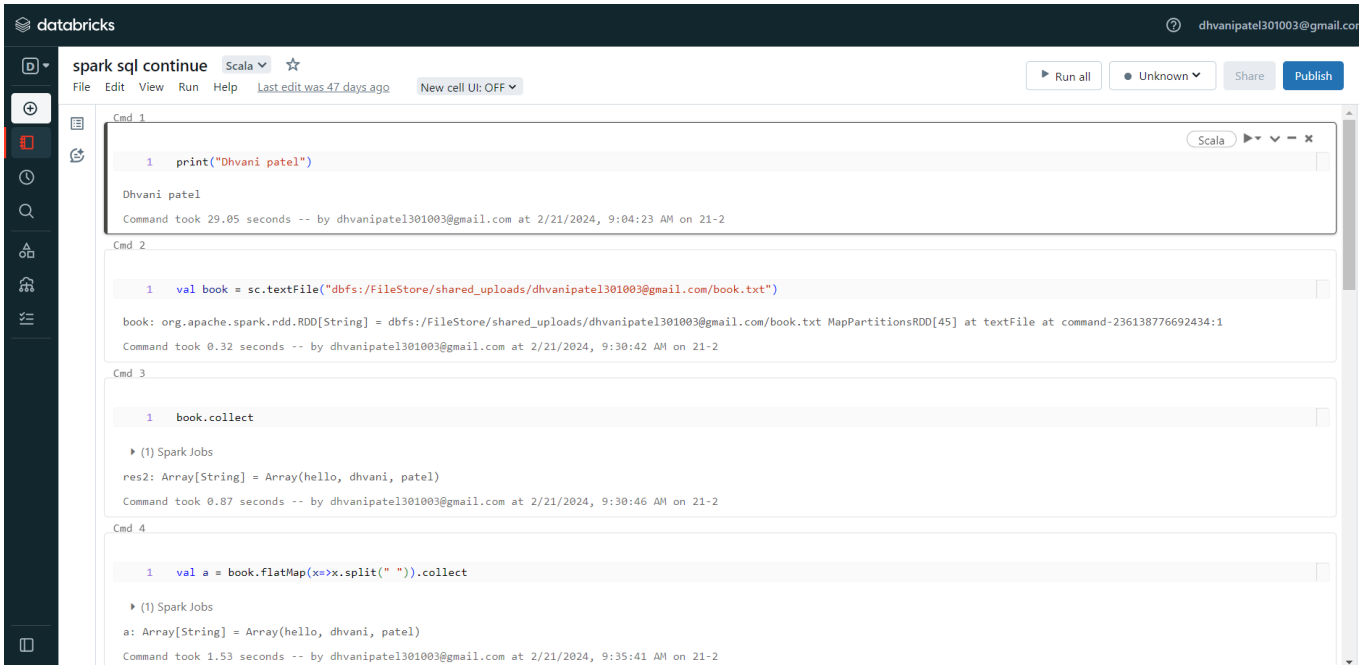

databricks
dhvanipatel301003@gmail.com
Practical 5 Spark sql
Scala
File Edit View Run Help Last edit was 47 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 21
1 df1.createOrReplaceTempView("Employee")
Command took 0.46 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:17:35 AM on 21-2
Cmd 22
1 df1.filter("age > '23'").show()
(1) Spark Jobs
+---+----+-----+
|Age| id| name|
+---+----+-----+
27	1201	Kriti
25	1203	Sneha
35	1205	Sonal
+---+----+-----+
Command took 1.19 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:17:46 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Practical 5 Spark sql
Scala
File Edit View Run Help Last edit was 47 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
1 df1.filter(df1("age") > 23).show()
(1) Spark Jobs
+---+----+-----+
|Age| id| name|
+---+----+-----+
27	1201	Kriti
25	1203	Sneha
35	1205	Sonal
+---+----+-----+
Command took 0.93 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:17:58 AM on 21-2
Cmd 24
1 df1.groupBy("age").count().show
(2) Spark Jobs
+---+-----+
|age|count|
+---+-----+
22	1
25	1
27	1
35	1
20	1
+---+-----+
Command took 2.58 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:18:05 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Practical 5 Spark sql
Scala
File Edit View Run Help Last edit was 47 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 25
1 val rdda = sc.parallelize(1 to 1000)
2 rdda.collect()
3 val rddb = sc.parallelize(List("BMW", "Mercedes", "Toyota", "Audi"))
4 rddb.collect()
(2) Spark Jobs
rdda: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[31] at parallelize at command-236138776692454:1
rddb: org.apache.spark.rdd.RDD[String] = ParallelCollectionRDD[32] at parallelize at command-236138776692454:3
res23: Array[String] = Array(BMW, Mercedes, Toyota, Audi)
Command took 0.97 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:18:20 AM on 21-2
Cmd 26
1 rdda.partitions.length
res24: Int = 8
Command took 0.42 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:18:28 AM on 21-2
Cmd 27
1 rddb.partitions.length
res25: Int = 8
Command took 0.37 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:18:33 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Practical 5 Spark sql
Scala
File Edit View Run Help Last edit was 47 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 28
1 val rdda = sc.parallelize(1 to 1000, 10)
2 rdda.collect()
3 rdda.partitions.length
(1) Spark Jobs
rdda: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[33] at parallelize at command-236138776692457:1
res26: Int = 10
Command took 0.47 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:18:40 AM on 21-2
Cmd 29
1 rdda.take(10)
(1) Spark Jobs
res27: Array[Int] = Array(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
Command took 0.41 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:18:51 AM on 21-2
Cmd 30
1 rdda.count()
(1) Spark Jobs
res38: Long = 1000
Command took 0.40 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:31:16 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Practical 5 Spark sql
Scala
File Edit View Run Help Last edit was 47 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 31
1 rdda.saveAsTextFile("dbfs:/FileStore/shared_uploads/dhvanipatel301003@gmail.com/random.txt")
FileAlreadyExistsException: Output directory dbfs:/FileStore/shared_uploads/dhvanipatel301003@gmail.com/random.txt already exists
Command took 0.34 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:31:53 AM on 21-2
Cmd 32
1 val rddRead = sc.textFile("dbfs:/FileStore/shared_uploads/dhvanipatel301003@gmail.com/random.txt")
rddRead: org.apache.spark.rdd.RDD[String] = dbfs:/FileStore/shared_uploads/dhvanipatel301003@gmail.com/random.txt MapPartitionsRDD[54] at textFile at command-236138776692461:1
Command took 0.29 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:31:54 AM on 21-2
Cmd 33
1 rddRead.count()
res41: Long = 0
Command took 0.29 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:31:58 AM on 21-2
Cmd 34
1 rddRead.take(10)
res42: Array[String] = Array()
Command took 0.23 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:32:01 AM on 21-2


databricks
dhvanipatel301003@gmail.com
spark sql continue
Scala
File Edit View Run Help Last edit was 47 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 1
print("Dhvani patel")
Dhvani patel
Command took 29.05 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:04:23 AM on 21-2
Cmd 2
val book = sc.textFile("dbfs:/FileStore/shared_uploads/dhvanipatel301003@gmail.com/book.txt")
book: org.apache.spark.rdd.RDD[String] = dbfs:/FileStore/shared_uploads/dhvanipatel301003@gmail.com/book.txt MapPartitionsRDD[45] at textFile at command-236138776692434:1
Command took 0.32 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:30:42 AM on 21-2
Cmd 3
book.collect
(1) Spark Jobs
res2: Array[String] = Array(hello, dhvani, patel)
Command took 0.87 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:30:46 AM on 21-2
Cmd 4
val a = book.flatMap(x=>x.split(" ")).collect
(1) Spark Jobs
a: Array[String] = Array(hello, dhvani, patel)
Command took 1.53 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:35:41 AM on 21-2


databricks
dhvanipatel301003@gmail.com
spark sql continue
Scala
File Edit View Run Help Last edit was 47 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 5
val b= a.map(y=>(y,1))
b: Array[(String, Int)] = Array((hello,1), (dhvani,1), (patel,1))
Command took 0.42 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:35:47 AM on 21-2
Cmd 6
val a= book.flatMap(x=>x.split(" ")).map(y=>(y,1)).reduceByKey((x,y) => (x+y)).collect
(1) Spark Jobs
a: Array[(String, Int)] = Array((patel,1), (dhvani,1), (hello,1))
Command took 1.99 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:37:22 AM on 21-2
Cmd 7
val a= book.flatMap(x=>x.split(" ")).map(y=>(y,1)).reduceByKey((x,y) => (x+y)).sortBy(_._1,false).collect
(2) Spark Jobs
a: Array[(String, Int)] = Array((patel,1), (hello,1), (dhvani,1))
Command took 1.25 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:38:22 AM on 21-2

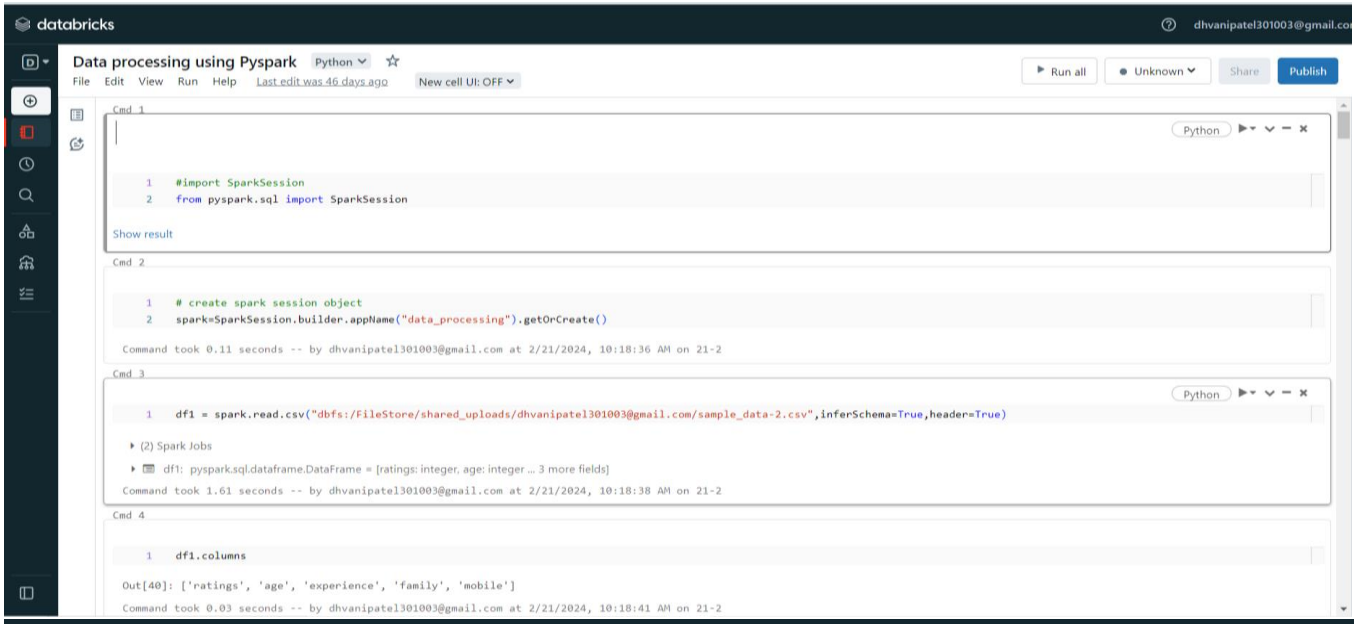

databricks
dhvanipatel301003@gmail.com
spark sql continue
Scala
File Edit View Run Help Last edit was 47 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 8
val a= book.flatMap(x=>x.split(" ")).map(y=>(y,1)).reduceByKey((x,y) => (x+y)).sortBy(_._2,false).take(1)
(3) Spark Jobs
a: Array[(String, Int)] = Array((patel,1))
Command took 1.61 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:40:56 AM on 21-2
Cmd 9
val a= book.flatMap(x=>x.split(" ")).map(y=>(y,1)).reduceByKey((x,y) => (x+y)).sortBy(_._2,false).filter(_._2 >1).collect
(2) Spark Jobs
a: Array[(String, Int)] = Array()
Command took 1.03 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 9:41:32 AM on 21-2

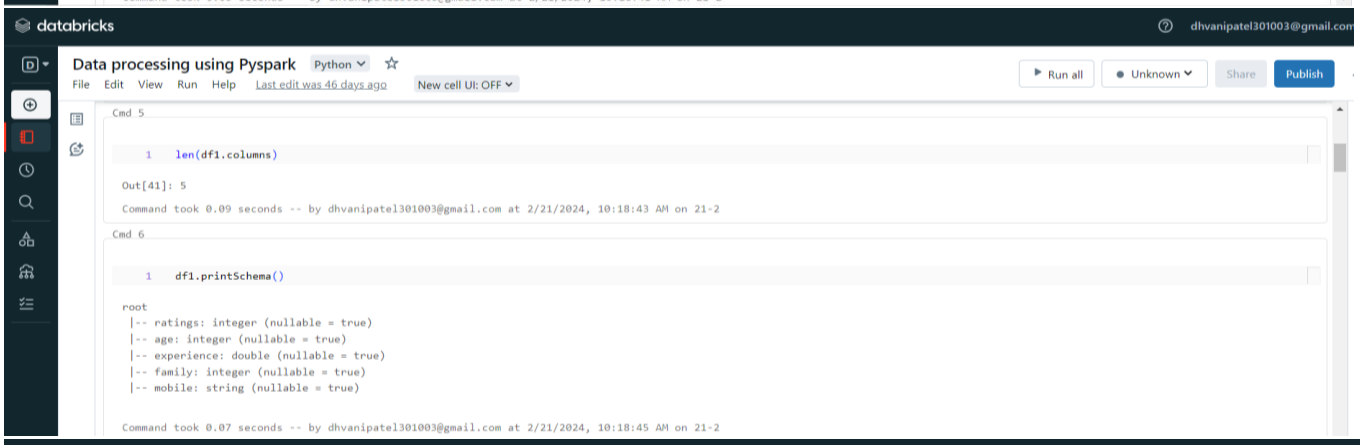
Name: Dhvani Patel

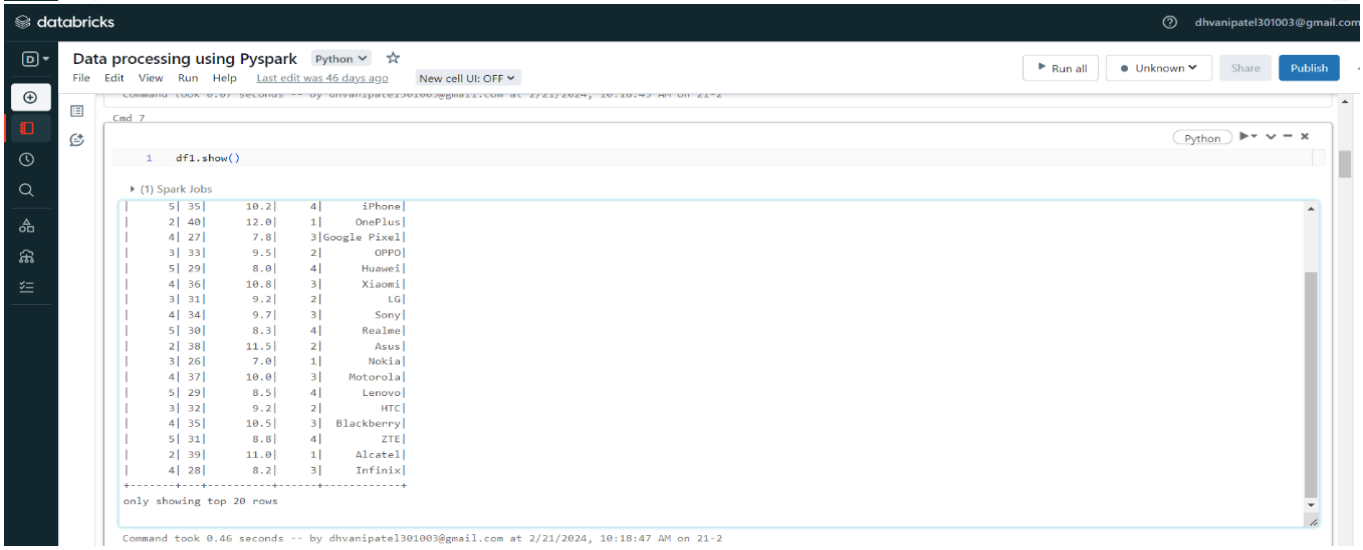
Roll no: 21BCP116

Big Data Analytics Lab

Practical 6 : Data Preprocessing using Pyspark







databricks

dhvanipatel301003@gmail.com

Data processing using Pyspark Python

File Edit View Run Help Last edit was 46 days ago New cell UI: OFF

Run all Unknown Share Publish

Cmd 8

```
1  df1.select('ratings','age','mobile').show(5)
```

(1) Spark Jobs

```
+-------+---+------------+
|ratings|age|      mobile|
+-------+---+------------+
|      3| 32|        Vivo|
|      4| 28|     Samsung|
|      5| 35|      iPhone|
|      2| 40|     OnePlus|
|      4| 27|Google Pixel|
+-------+---+------------+
only showing top 5 rows
```

Command took 0.38 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:18:49 AM on 21-2

databricks

dhvanipatel301003@gmail.com

Data processing using Pyspark Python

File Edit View Run Help Last edit was 46 days ago New cell UI: OFF

Run all Unknown Share Publish

Cmd 9

```
1  df1.describe().show()
```

(2) Spark Jobs

```
+-------+------------------+-----------------+-----------------+-----------------+-------+
|summary|           ratings|              age|       experience|           family| mobile|
+-------+------------------+-----------------+-----------------+-----------------+-------+
|  count|                47|               47|               47|               47|     47|
|   mean|3.6808510638297873|32.40425531914894|9.295744680851064|2.6808510638297873|   null|
| stddev| 1.0237693137472283|3.8485433408548855|1.17602082331618|1.0237693137472283|   null|
|    min|                 2|               26|              7.0|                1|Alcatel|
|    max|                 5|               40|             12.0|                4| iPhone|
+-------+------------------+-----------------+-----------------+-----------------+-------+
```

Command took 1.70 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:18:50 AM on 21-2

Cmd 10

Python

```
1  from pyspark.sql.types import StringType,DoubleType,IntegerType
```

Command took 0.06 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:18:54 AM on 21-2

databricks

dhvanipatel301003@gmail.com

Data processing using Pyspark Python

File Edit View Run Help Last edit was 46 days ago New cell UI: OFF

Run all Unknown Share Publish

Cmd 11

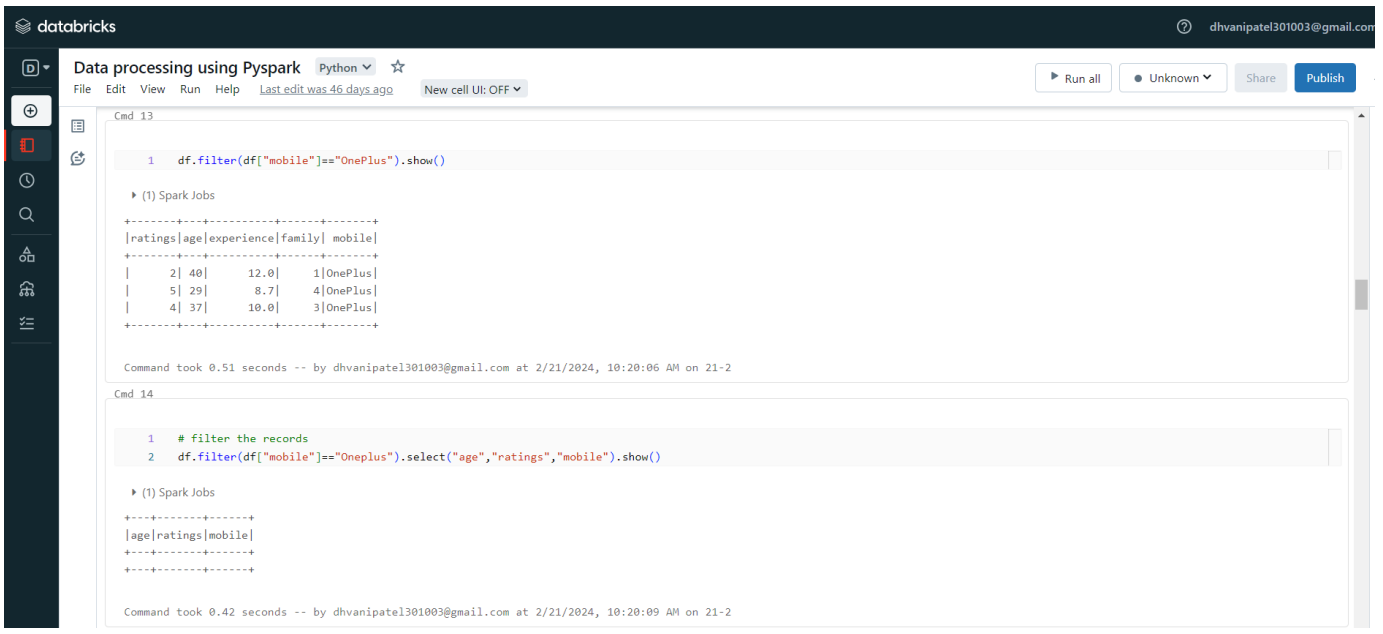
```
1  df1.withColumn("age_after_10_yrs", (df1["age"]+10)).show(10,False)
```

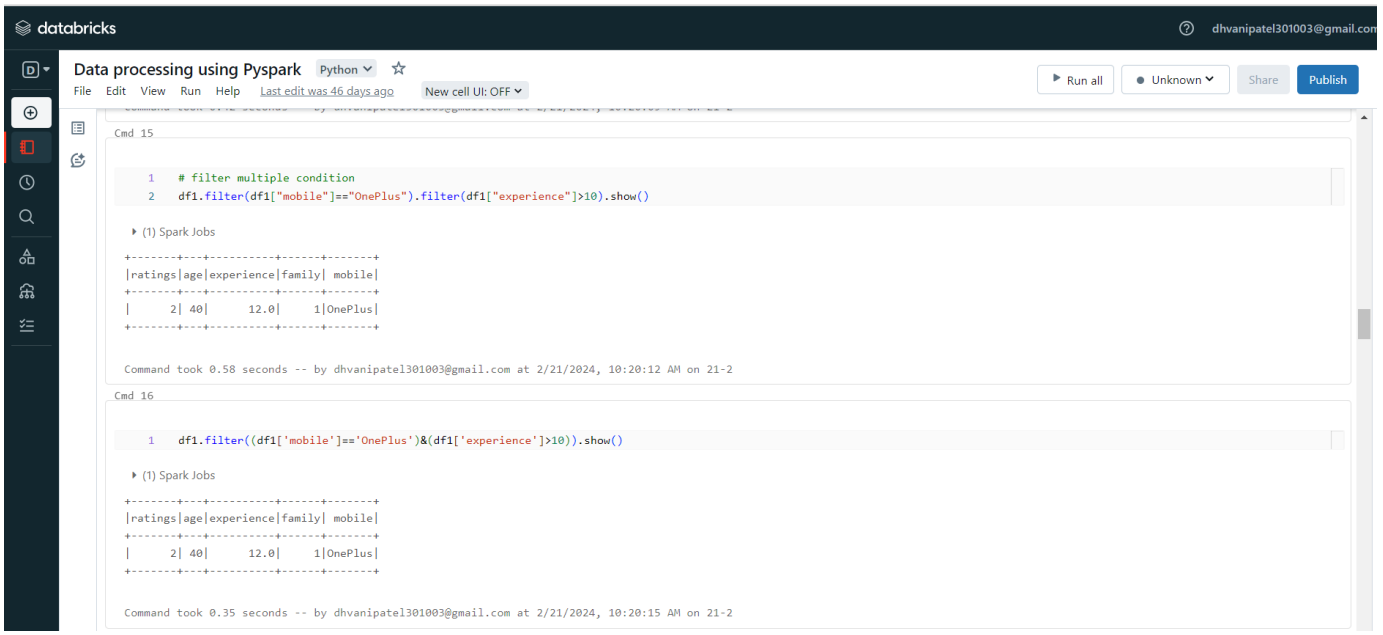
(1) Spark Jobs

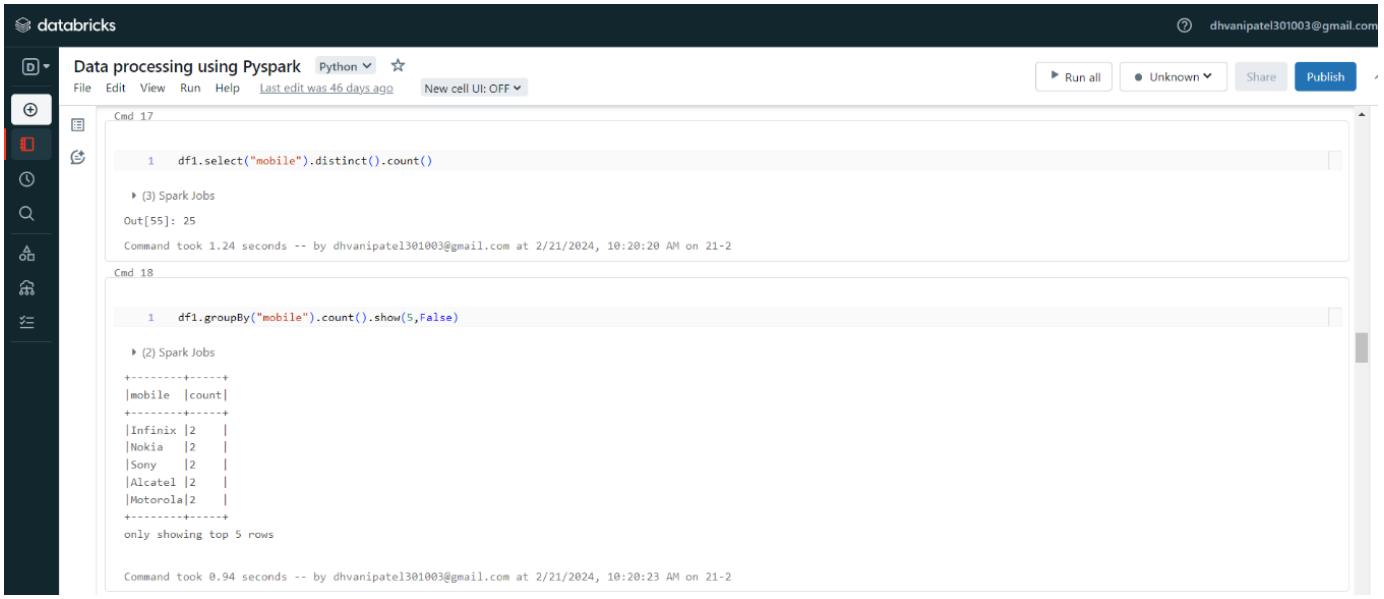
```
+-------+---+----------+------+------------+----------------+
|ratings|age|experience|family|mobile      |age_after_10_yrs|
+-------+---+----------+------+------------+----------------+
|3      |32 |9.0       |3     |Vivo        |42              |
|4      |28 |8.5       |2     |Samsung     |38              |
|5      |35 |10.2      |4     |iPhone      |45              |
|2      |40 |12.0      |1     |OnePlus     |50              |
|4      |27 |7.8       |3     |Google Pixel|37              |
|3      |33 |9.5       |2     |OPPO        |43              |
|5      |29 |8.0       |4     |Huawei      |39              |
|4      |36 |10.8      |3     |Xiaomi      |46              |
|3      |31 |9.2       |2     |LG          |41              |
|4      |34 |9.7       |3     |Sony        |44              |
+-------+---+----------+------+------------+----------------+
only showing top 10 rows
```

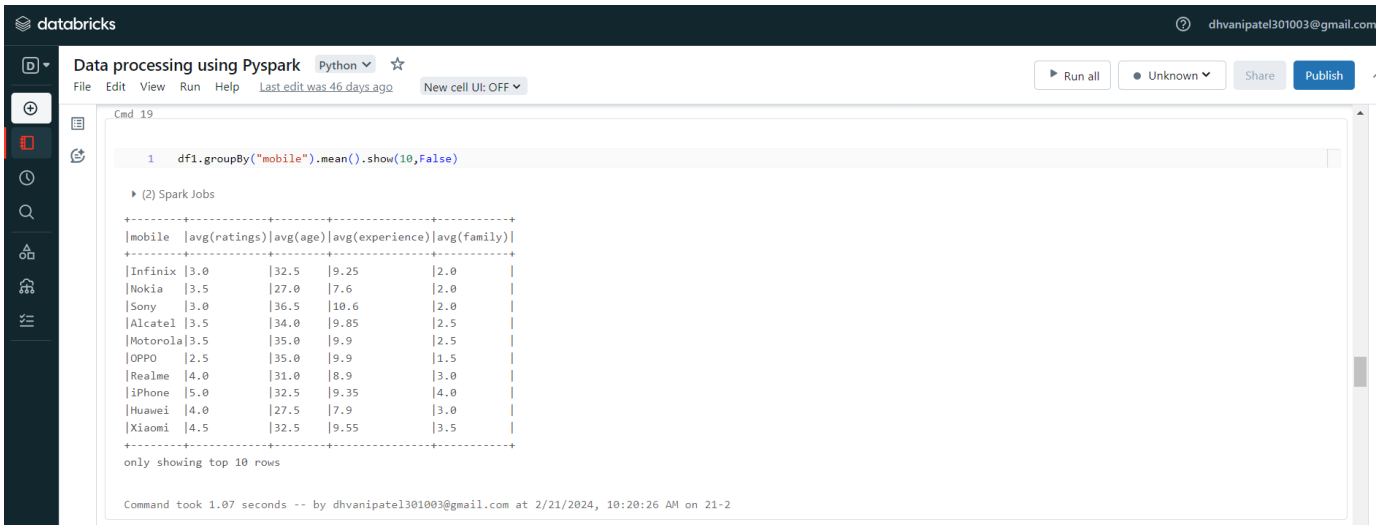
Command took 0.52 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:19:30 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 12
1 df1.withColumn("age_double", (df1["age"]).cast(DoubleType())).show(10,False)
(1) Spark Jobs
+-------+---+----------+------+------------+----------+
|ratings|age|experience|family|mobile |age_double|
+-------+---+----------+------+------------+----------+
3	32	9.0	3	Vivo	32.0
4	28	8.5	2	Samsung	28.0
5	35	10.2	4	iPhone	35.0
2	40	12.0	1	OnePlus	40.0
4	27	7.8	3	Google Pixel	27.0
3	33	9.5	2	OPPO	33.0
5	29	8.0	4	Huawei	29.0
4	36	10.8	3	Xiaomi	36.0
3	31	9.2	2	LG	31.0
4	34	9.7	3	Sony	34.0
+-------+---+----------+------+------------+----------+
only showing top 10 rows
Command took 0.55 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:20:04 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 13
1 df.filter(df["mobile"]=="OnePlus").show()
(1) Spark Jobs
+-------+---+----------+------+-------+
|ratings|age|experience|family| mobile|
+-------+---+----------+------+-------+
2	40	12.0	1	OnePlus
5	29	8.7	4	OnePlus
4	37	10.0	3	OnePlus
+-------+---+----------+------+-------+
Command took 0.51 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:20:06 AM on 21-2
Cmd 14
1 # filter the records
2 df.filter(df["mobile"]=="Oneplus").select("age","ratings","mobile").show()
(1) Spark Jobs
+---+-------+------+
|age|ratings|mobile|
+---+-------+------+
+---+-------+------+
Command took 0.42 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:20:09 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 15
1 # filter multiple condition
2 df1.filter(df1["mobile"]=="OnePlus").filter(df1["experience"]>10).show()
(1) Spark Jobs
+-------+---+----------+------+-------+
|ratings|age|experience|family| mobile|
+-------+---+----------+------+-------+
| 2| 40| 12.0| 1|OnePlus|
+-------+---+----------+------+-------+
Command took 0.58 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:20:12 AM on 21-2
Cmd 16
1 df1.filter((df1['mobile']=='OnePlus')&(df1['experience']>10)).show()
(1) Spark Jobs
+-------+---+----------+------+-------+
|ratings|age|experience|family| mobile|
+-------+---+----------+------+-------+
| 2| 40| 12.0| 1|OnePlus|
+-------+---+----------+------+-------+
Command took 0.35 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:20:15 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 17
1 df1.select("mobile").distinct().count()
(3) Spark Jobs
Out[55]: 25
Command took 1.24 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:20:20 AM on 21-2
Cmd 18
1 df1.groupBy("mobile").count().show(5,False)
(2) Spark Jobs
+--------+-----+
|mobile |count|
+--------+-----+
Infinix	2
Nokia	2
Sony	2
Alcatel	2
Motorola	2
+--------+-----+
only showing top 5 rows
Command took 0.94 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:20:23 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 19
1 df1.groupBy("mobile").mean().show(10,False)
(2) Spark Jobs
+--------+------------+--------+---------------+-----------+
|mobile |avg(ratings)|avg(age)|avg(experience)|avg(family)|
+--------+------------+--------+---------------+-----------+
Infinix	3.0	32.5	9.25	2.0
Nokia	3.5	27.0	7.6	2.0
Sony	3.0	36.5	10.6	2.0
Alcatel	3.5	34.0	9.85	2.5
Motorola	3.5	35.0	9.9	2.5
OPPO	2.5	35.0	9.9	1.5
Realme	4.0	31.0	8.9	3.0
iPhone	5.0	32.5	9.35	4.0
Huawei	4.0	27.5	7.9	3.0
Xiaomi	4.5	32.5	9.55	3.5
+--------+------------+--------+---------------+-----------+
only showing top 10 rows
Command took 1.07 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:20:26 AM on 21-2

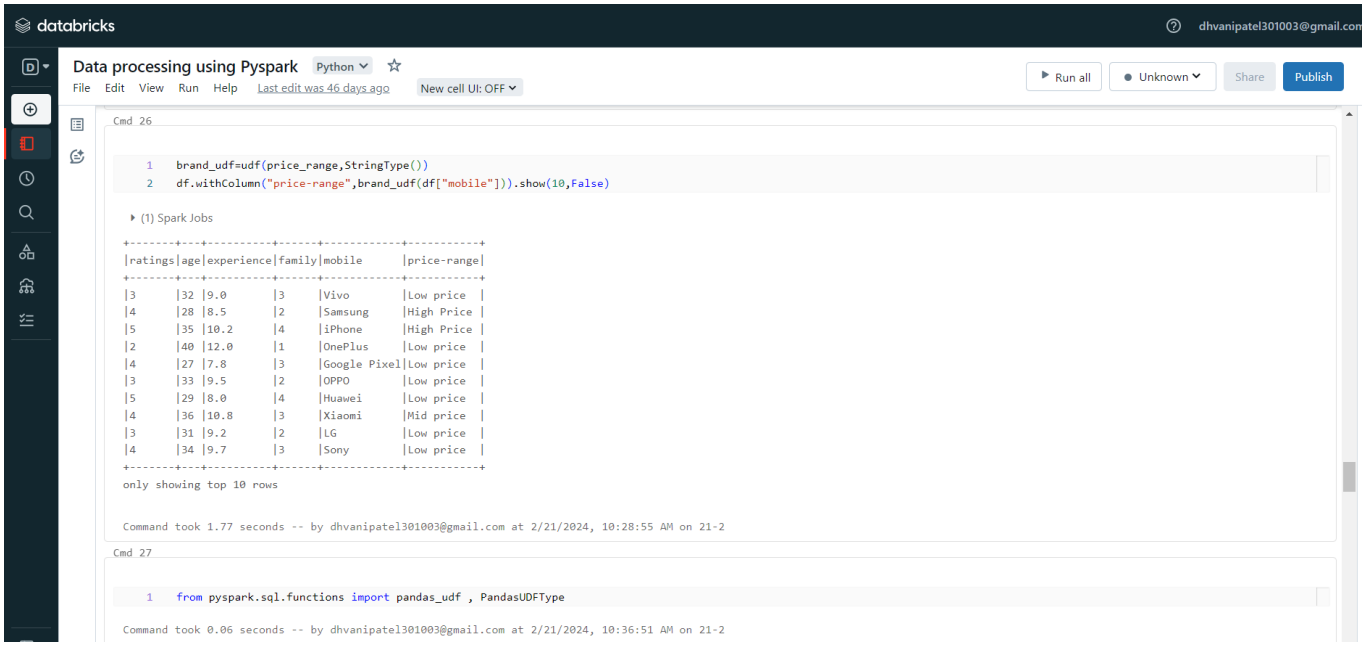

databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 20
1 df1.groupBy("mobile").sum().show(5,False)
(2) Spark Jobs
+--------+------------+--------+---------------+-----------+
|mobile |sum(ratings)|sum(age)|sum(experience)|sum(family)|
+--------+------------+--------+---------------+-----------+
Infinix	6	65	18.5	4
Nokia	7	54	15.2	4
Sony	6	73	21.2	4
Alcatel	7	68	19.7	5
Motorola	7	70	19.8	5
+--------+------------+--------+---------------+-----------+
only showing top 5 rows
Command took 0.99 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:20:31 AM on 21-2

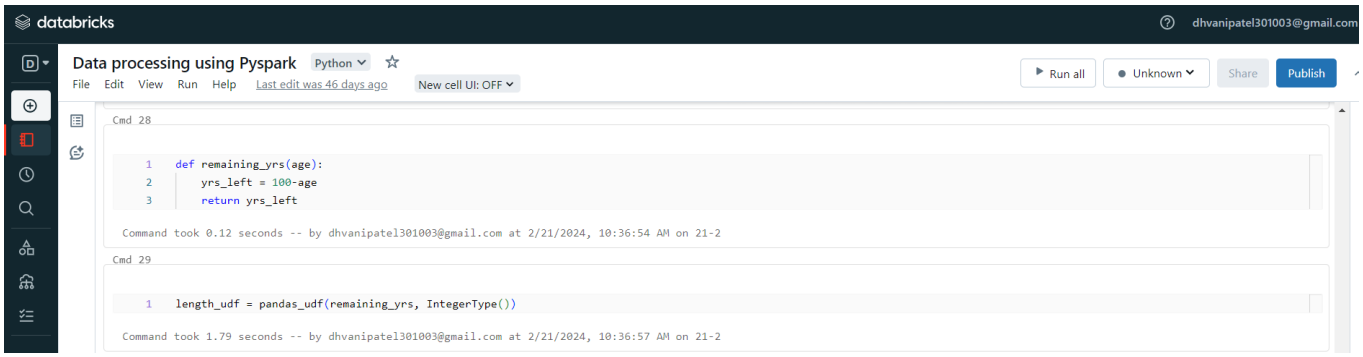

databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 21
1 df1.groupBy("mobile").min().show(5,False)
(2) Spark Jobs
+--------+------------+--------+--------------+-----------+
|mobile |min(ratings)|min(age)|min(experience)|min(family)|
+--------+------------+--------+--------------+-----------+
Infinix	2	28	8.2	1
Nokia	3	26	7.0	1
Sony	2	34	9.7	1
Alcatel	2	29	8.7	1
Motorola	3	33	9.8	2
+--------+------------+--------+--------------+-----------+
only showing top 5 rows
Command took 1.12 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:20:34 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 22
1 df1.groupBy("mobile").max().show(5,False)
(2) Spark Jobs
+--------+------------+--------+--------------+-----------+
|mobile |max(ratings)|max(age)|max(experience)|max(family)|
+--------+------------+--------+--------------+-----------+
Infinix	4	37	10.3	3
Nokia	4	28	8.2	3
Sony	4	39	11.5	3
Alcatel	5	39	11.0	4
Motorola	4	37	10.0	3
+--------+------------+--------+--------------+-----------+
only showing top 5 rows
Command took 0.97 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:20:37 AM on 21-2

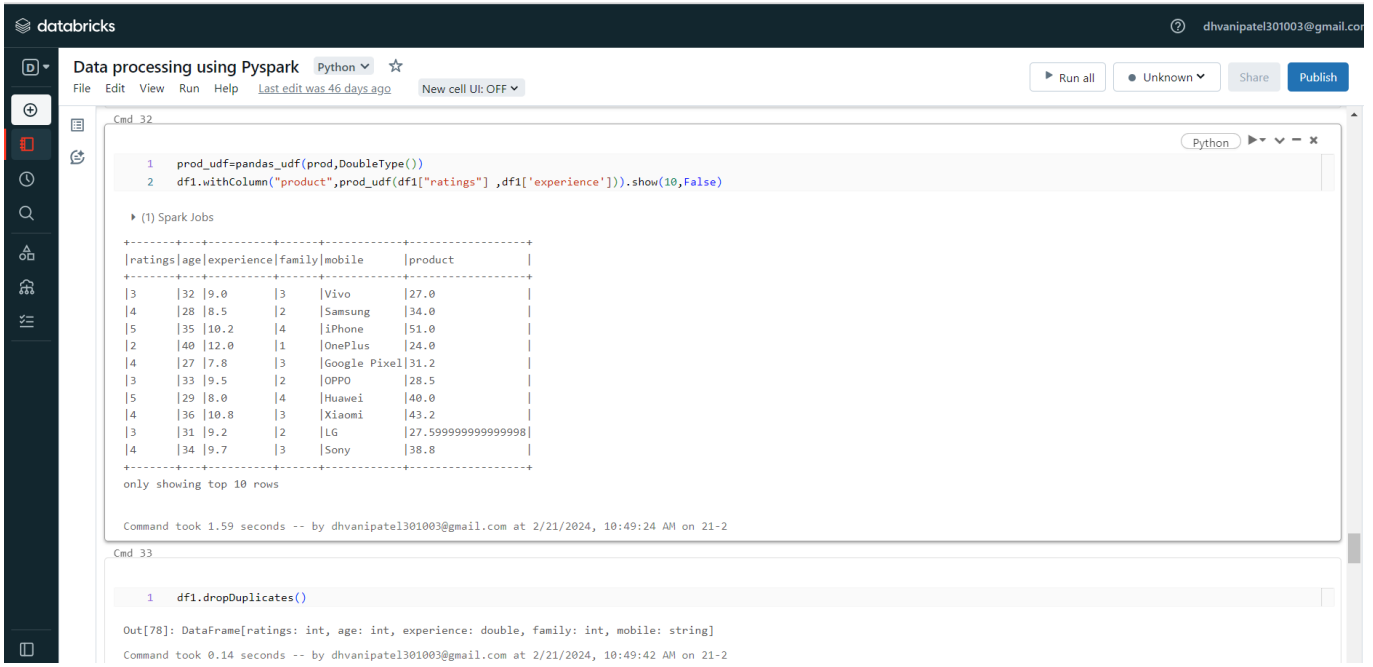

databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 23
1 df.groupBy("mobile").agg({"experience":"sum"}).show(5,False)
(2) Spark Jobs
+--------+---------------+
|mobile |sum(experience)|
+--------+---------------+
Infinix	18.5
Nokia	15.2
Sony	21.2
Alcatel	19.7
Motorola	19.8
+--------+---------------+
only showing top 5 rows
Command took 0.94 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:21:29 AM on 21-2
Cmd 24
1 from pyspark.sql.functions import udf
Command took 0.17 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:28:46 AM on 21-2

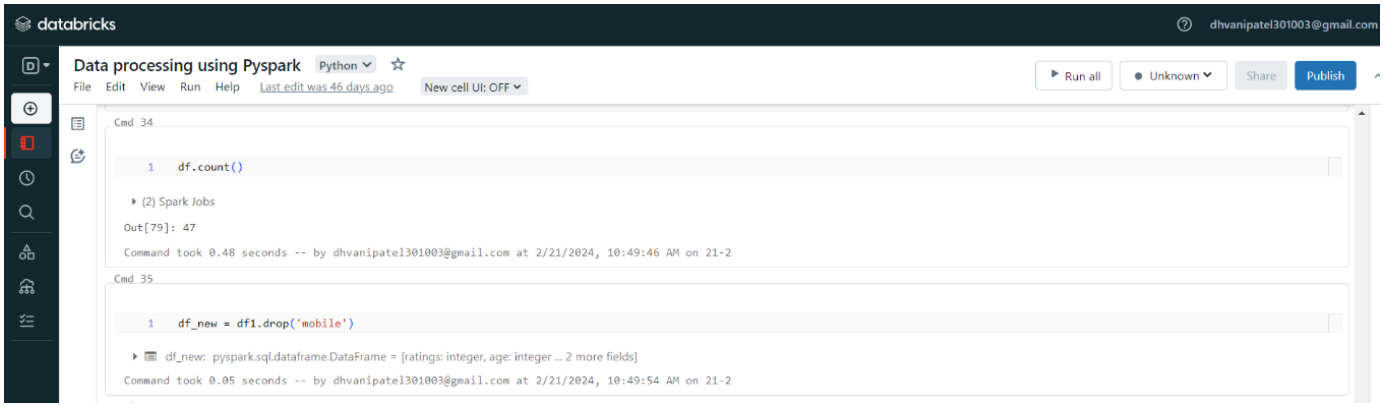

databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 25
1 def price_range(brand):
2 if brand in ["Samsung","iPhone"]:
3 return "High Price"
4 elif brand=="Xiaomi":
5 return "Mid price"
6 else:
7 return "Low price"
Command took 0.09 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:28:49 AM on 21-2

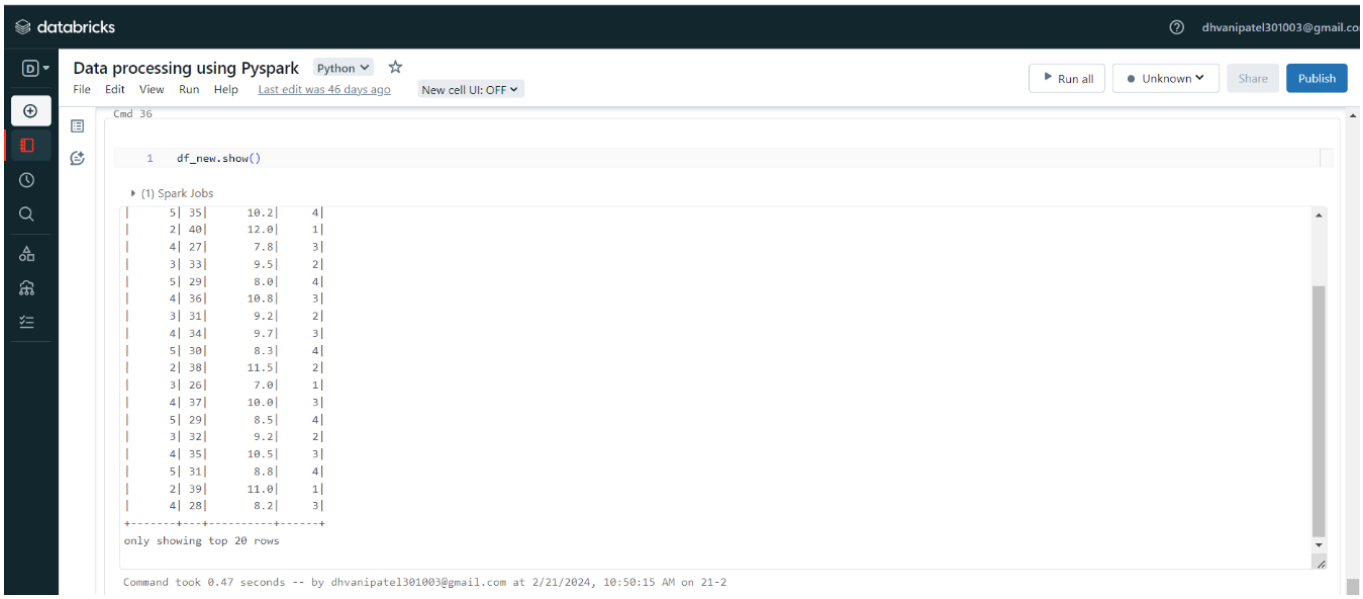

databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 26
1 brand_udf=udf(price_range,StringType())
2 df.withColumn("price-range",brand_udf(df["mobile"])).show(10,False)
(1) Spark Jobs
+-------+---+----------+------+------------+-----------+
|ratings|age|experience|family|mobile |price-range|
+-------+---+----------+------+------------+-----------+
3	32	9.0	3	Vivo	Low price
4	28	8.5	2	Samsung	High Price
5	35	10.2	4	iPhone	High Price
2	40	12.0	1	OnePlus	Low price
4	27	7.8	3	Google Pixel	Low price
3	33	9.5	2	OPPO	Low price
5	29	8.0	4	Huawei	Low price
4	36	10.8	3	Xiaomi	Mid price
3	31	9.2	2	LG	Low price
4	34	9.7	3	Sony	Low price
+-------+---+----------+------+------------+-----------+
only showing top 10 rows
Command took 1.77 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:28:55 AM on 21-2
Cmd 27
1 from pyspark.sql.functions import pandas_udf , PandasUDFType
Command took 0.06 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:36:51 AM on 21-2

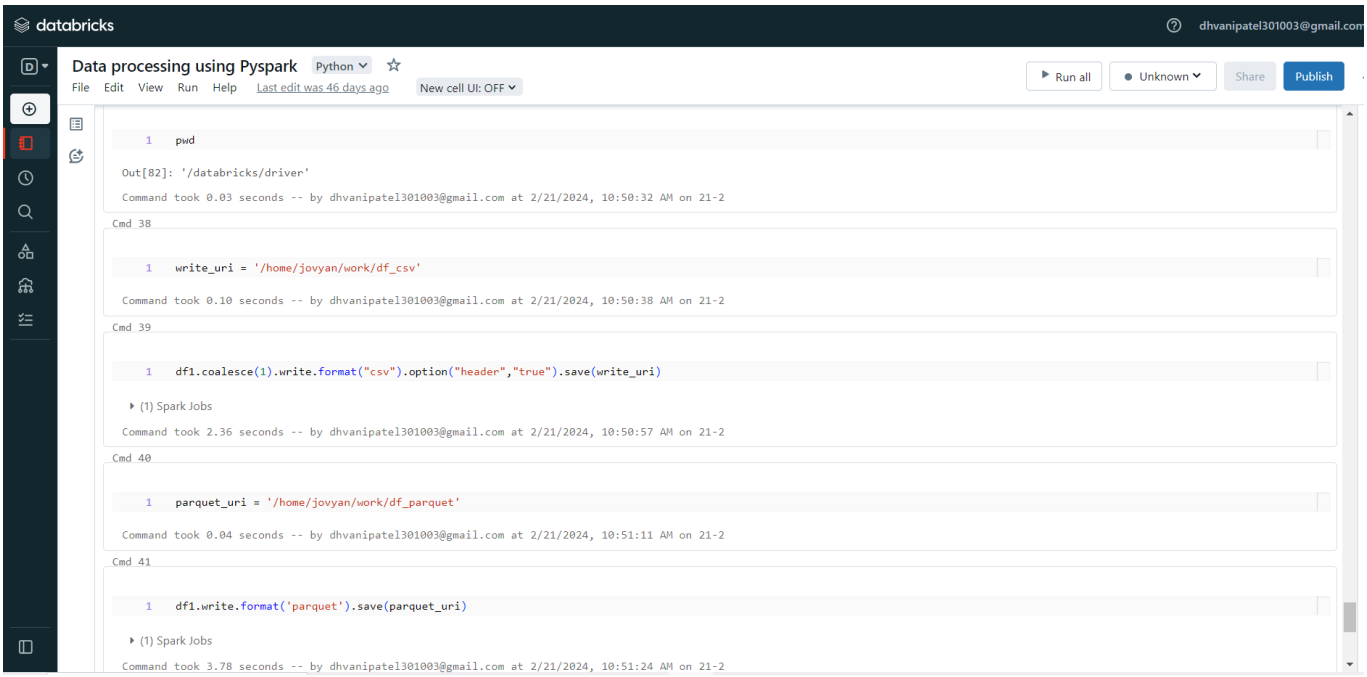

databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 28
1 def remaining_yrs(age):
2 yrs_left = 100-age
3 return yrs_left
Command took 0.12 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:36:54 AM on 21-2
Cmd 29
1 length_udf = pandas_udf(remaining_yrs, IntegerType())
Command took 1.79 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:36:57 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 30
1 df1.withColumn("yrs_left", length_udf(df1['age'])).show(10,False)
(1) Spark Jobs
+-------+---+----------+------+------------+--------+
|ratings|age|experience|family|mobile |yrs_left|
+-------+---+----------+------+------------+--------+
3	32	9.0	3	Vivo	68
4	28	8.5	2	Samsung	72
5	35	10.2	4	iPhone	65
2	40	12.0	1	OnePlus	60
4	27	7.8	3	Google Pixel	73
3	33	9.5	2	OPPO	67
5	29	8.0	4	Huawei	71
4	36	10.8	3	Xiaomi	64
3	31	9.2	2	LG	69
4	34	9.7	3	Sony	66
+-------+---+----------+------+------------+--------+
only showing top 10 rows
Command took 2.51 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:37:31 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 32
Python
1 prod_udf=pandas_udf(prod,DoubleType())
2 df1.withColumn("product",prod_udf(df1["ratings"] ,df1['experience'])).show(10,False)
(1) Spark Jobs
+-------+---+----------+------+------------+-----------------+
|ratings|age|experience|family|mobile |product |
+-------+---+----------+------+------------+-----------------+
3	32	9.0	3	Vivo	27.0
4	28	8.5	2	Samsung	34.0
5	35	10.2	4	iPhone	51.0
2	40	12.0	1	OnePlus	24.0
4	27	7.8	3	Google Pixel	31.2
3	33	9.5	2	OPPO	28.5
5	29	8.0	4	Huawei	40.0
4	36	10.8	3	Xiaomi	43.2
3	31	9.2	2	LG	27.599999999999998
4	34	9.7	3	Sony	38.8
+-------+---+----------+------+------------+-----------------+
only showing top 10 rows
Command took 1.59 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:49:24 AM on 21-2
Cmd 33
1 df1.dropDuplicates()
Out[78]: DataFrame[ratings: int, age: int, experience: double, family: int, mobile: string]
Command took 0.14 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:49:42 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 34
1 df.count()
(2) Spark Jobs
Out[79]: 47
Command took 0.48 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:49:46 AM on 21-2
Cmd 35
1 df_new = df1.drop('mobile')
df_new: pyspark.sql.dataframe.DataFrame = [ratings: integer, age: integer ... 2 more fields]
Command took 0.05 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:49:54 AM on 21-2


databricks
dhvanipatel301003@gmail.co
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
Cmd 36
1 df_new.show()
(1) Spark Jobs
5	35	10.2	4
2	40	12.0	1
4	27	7.8	3
3	33	9.5	2
5	29	8.0	4
4	36	10.8	3
3	31	9.2	2
4	34	9.7	3
5	30	8.3	4
2	38	11.5	2
3	26	7.0	1
4	37	10.0	3
5	29	8.5	4
3	32	9.2	2
4	35	10.5	3
5	31	8.8	4
2	39	11.0	1
4	28	8.2	3
only showing top 20 rows
Command took 0.47 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:50:15 AM on 21-2


databricks
dhvanipatel301003@gmail.com
Data processing using Pyspark
Python
File Edit View Run Help Last edit was 46 days ago
New cell UI: OFF
Run all
Unknown
Share
Publish
1 pwd
Out[82]: '/databricks/driver'
Command took 0.03 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:50:32 AM on 21-2
Cmd 38
1 write_uri = '/home/jovyan/work/df_csv'
Command took 0.10 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:50:38 AM on 21-2
Cmd 39
1 df1.coalesce(1).write.format("csv").option("header","true").save(write_uri)
(1) Spark Jobs
Command took 2.36 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:50:57 AM on 21-2
Cmd 40
1 parquet_uri = '/home/jovyan/work/df_parquet'
Command took 0.04 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:51:11 AM on 21-2
Cmd 41
1 df1.write.format('parquet').save(parquet_uri)
(1) Spark Jobs
Command took 3.78 seconds -- by dhvanipatel301003@gmail.com at 2/21/2024, 10:51:24 AM on 21-2

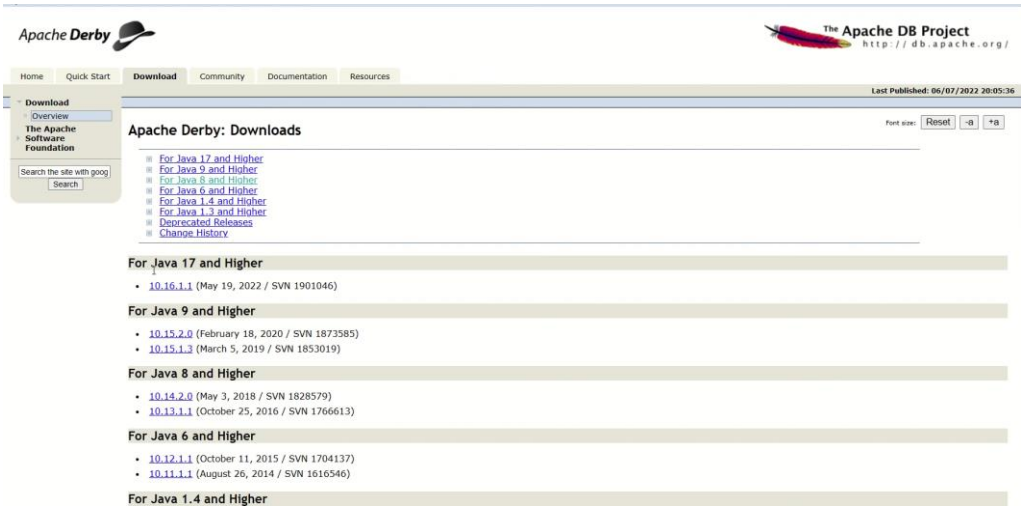
Name: Dhvani Patel

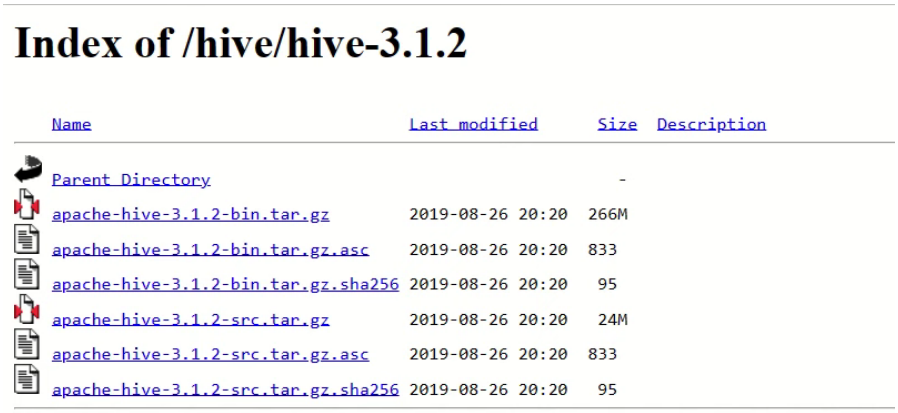
Rollno: 21BCP116

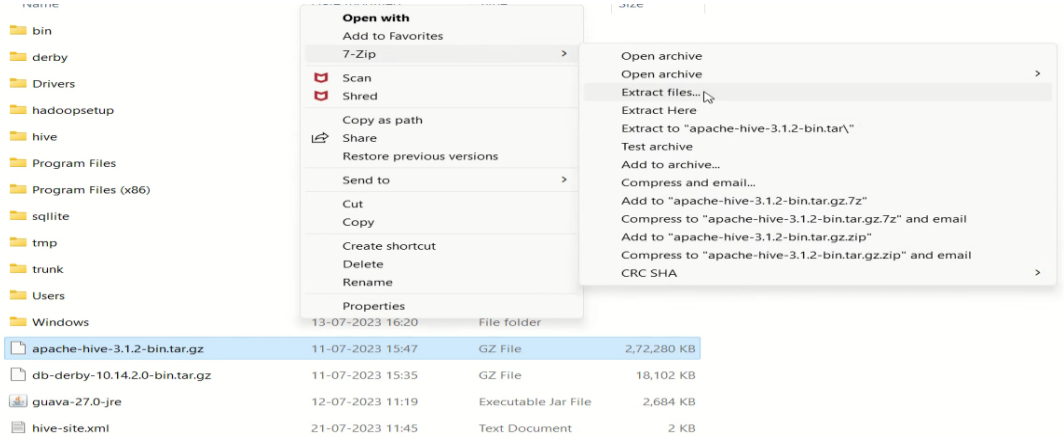
Big Data Analytics Lab

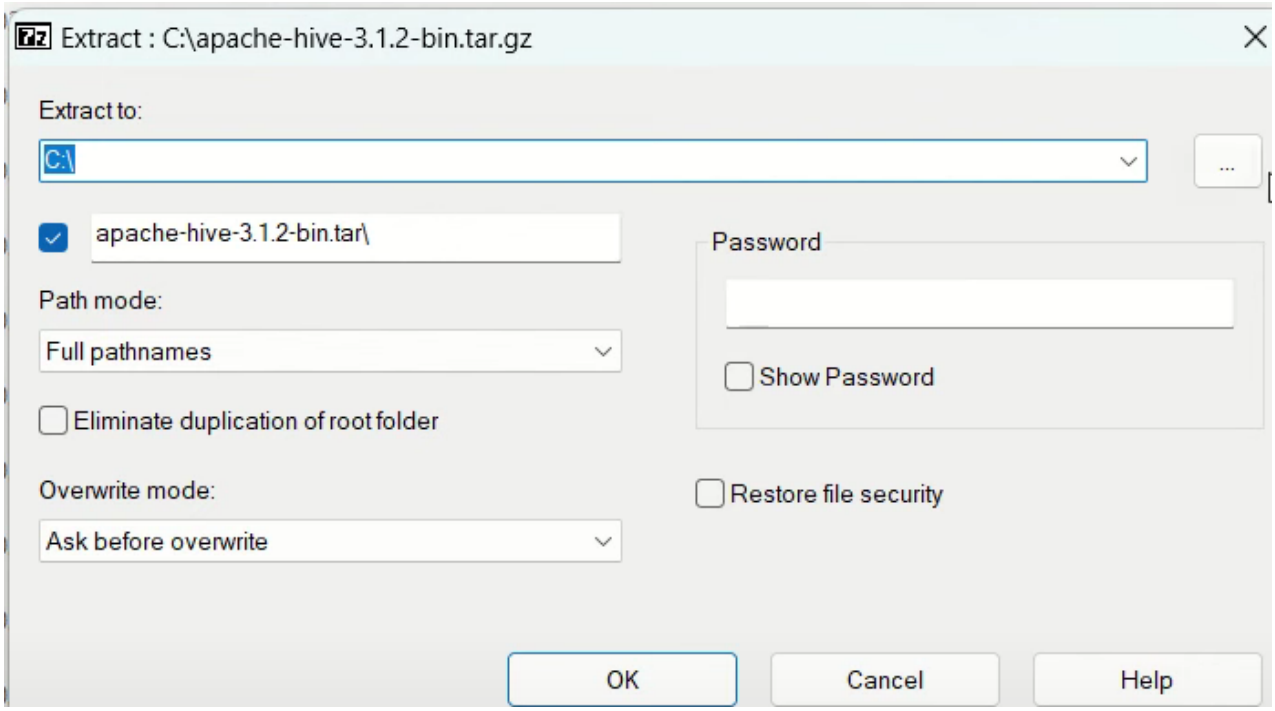
Practical 7 : Hadoop, Pig, Kafka, Spark and Hive installation

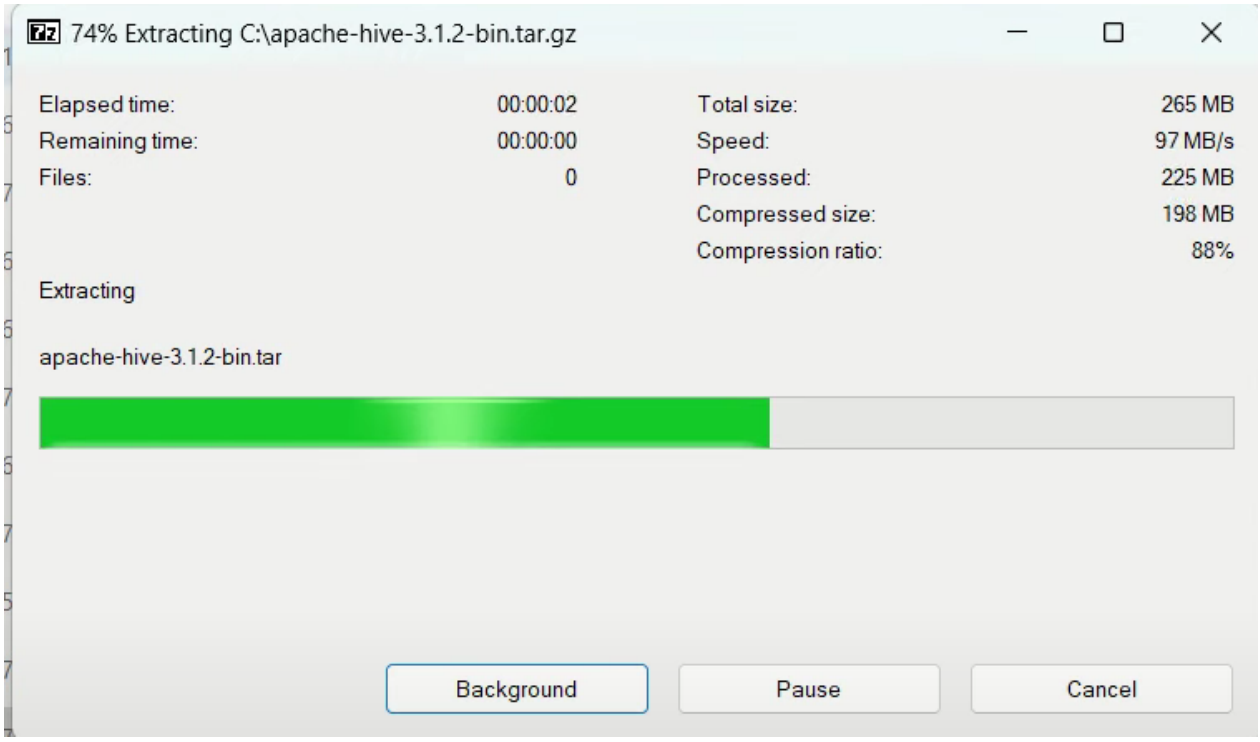
Apache Hive:





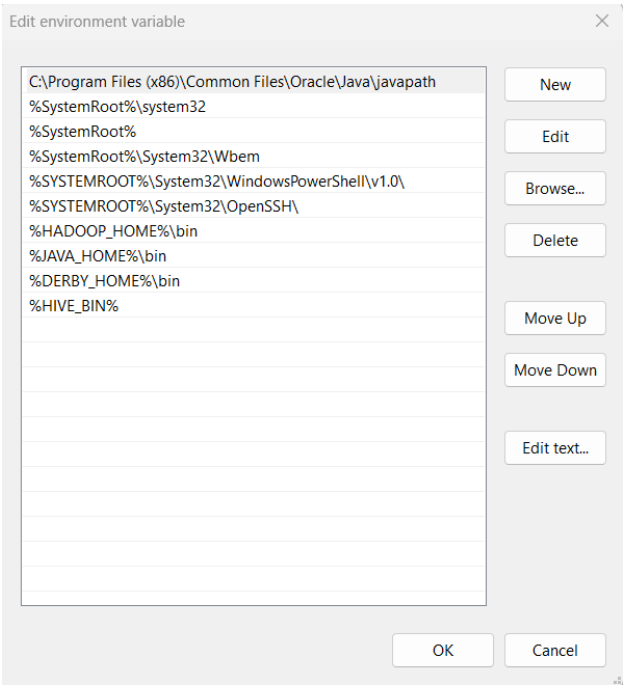




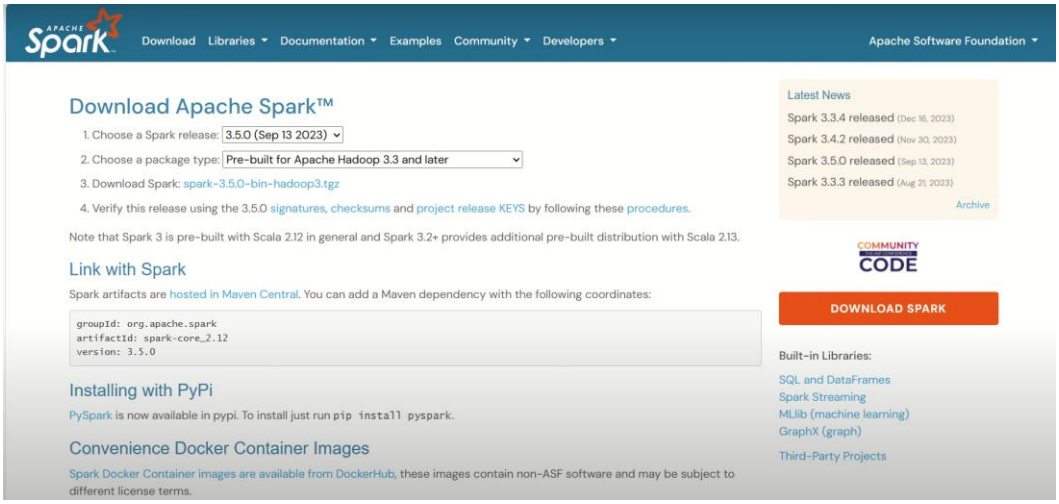


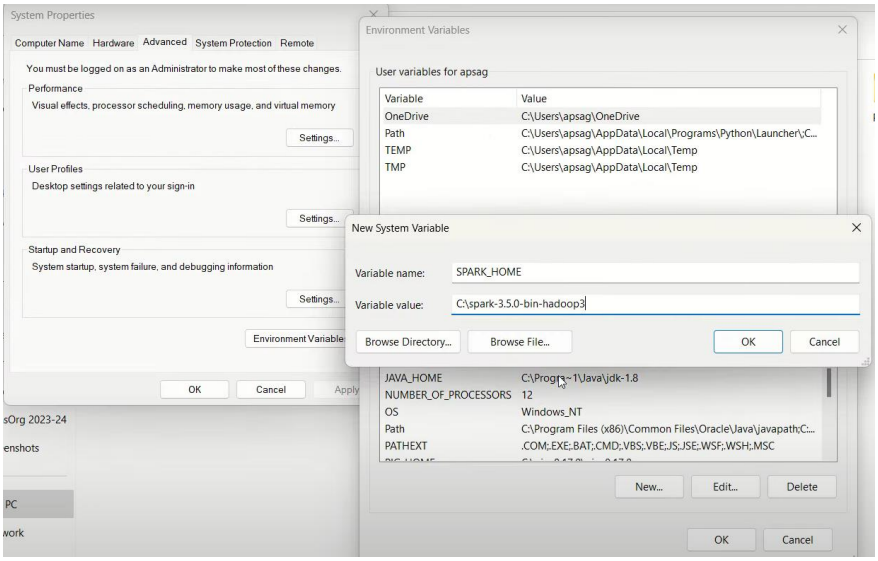
HIVE_BIN	%HIVE_HOME%\bin
HIVE_HOME	C:\hive\apache-hive-3.1.2-bin\apache-hive-3.1.2-bin
HIVE_LIB	%HIVE_HOME%\lib

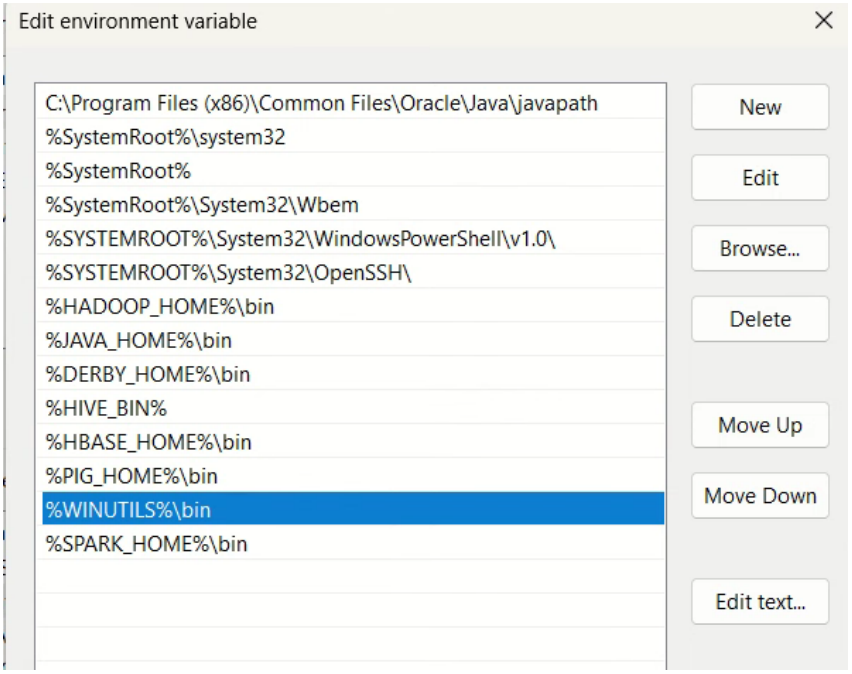
DERBY_HOME	C:\derby\db-derby-10.14.2.0-bin\db-derby-10.14.2.0-bin



Apache spark:



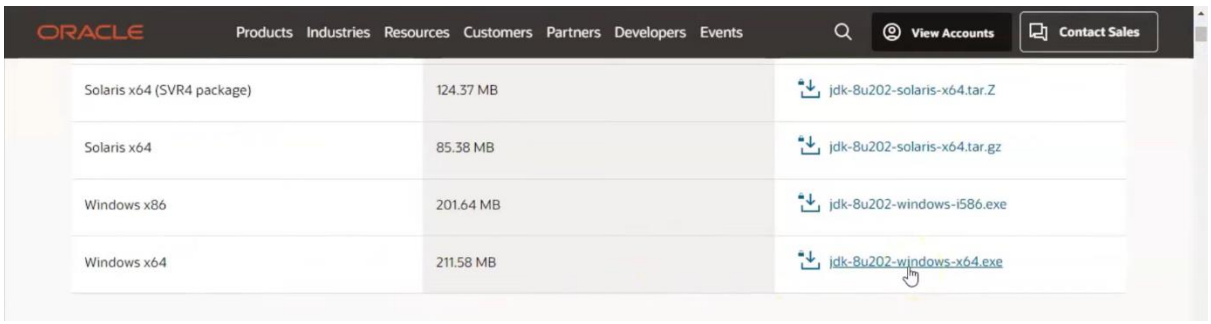




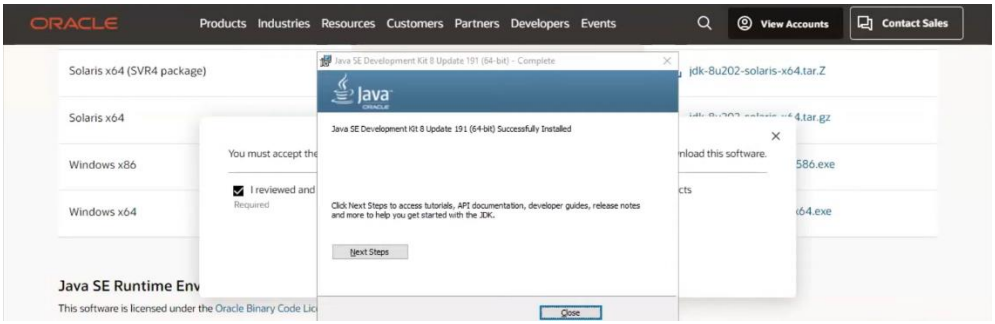

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Windows\system32> cd C:\spark-3.5.0-bin-hadoop3\bin\
PS C:\spark-3.5.0-bin-hadoop3\bin> spark-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://amrit:4040
Spark context available as 'sc' (master = local[*], app id = local-1703438479261).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.0
      /_/

Using Scala version 2.12.18 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_371)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

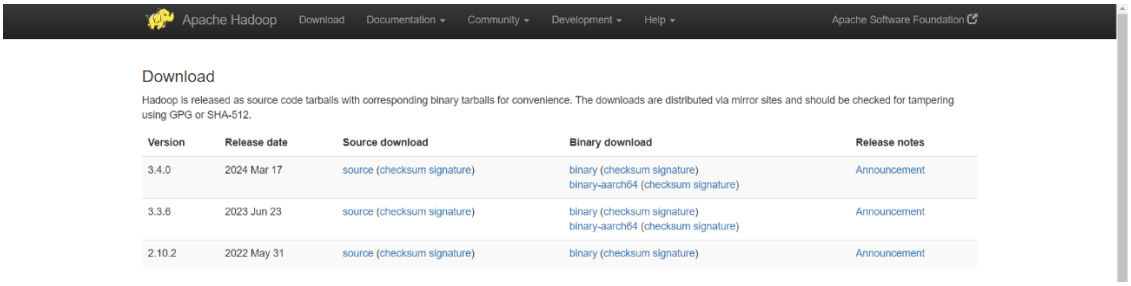
Hadoop installation

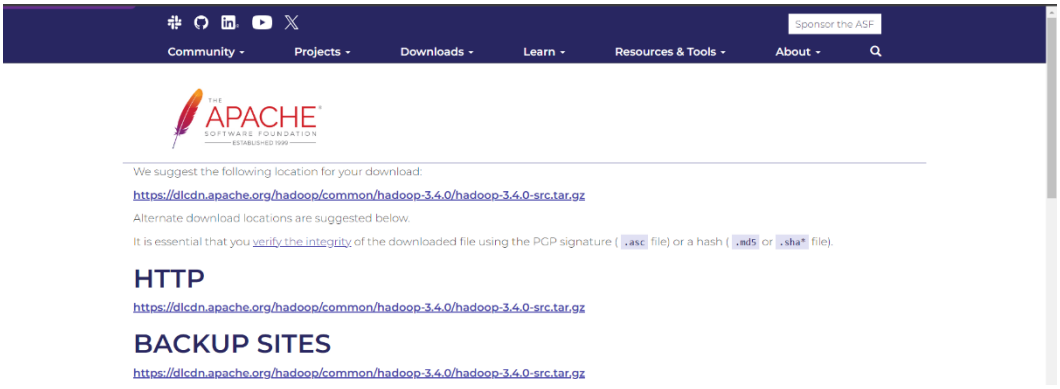
Java 8 installation:

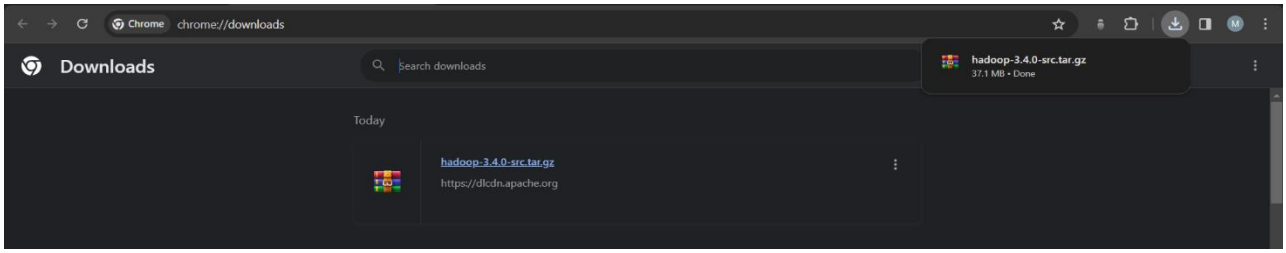


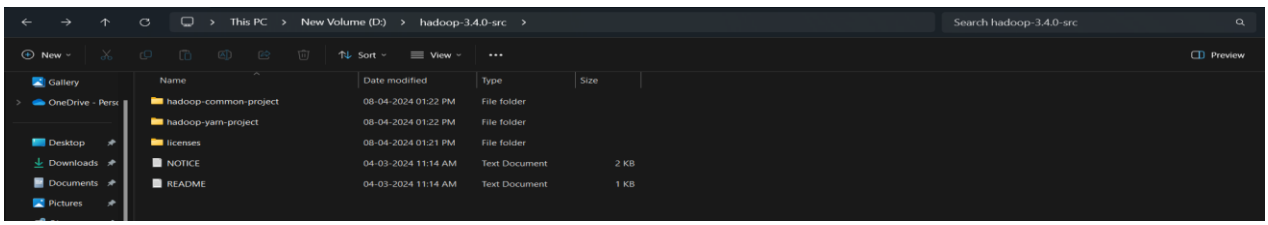

Solaris x64
Windows x86
Windows x64
You must accept the Oracle Binary Code License Agreement for the Java SE Platform Products to download this software.
I reviewed and accept the Oracle Binary Code License Agreement for the Java SE Platform Products
Required
Download jdk-8u202-windows-x64.exe
Java SE Runtime Env
This software is licensed under the Oracle Binary Code License Agreement for Java SE Platform Products

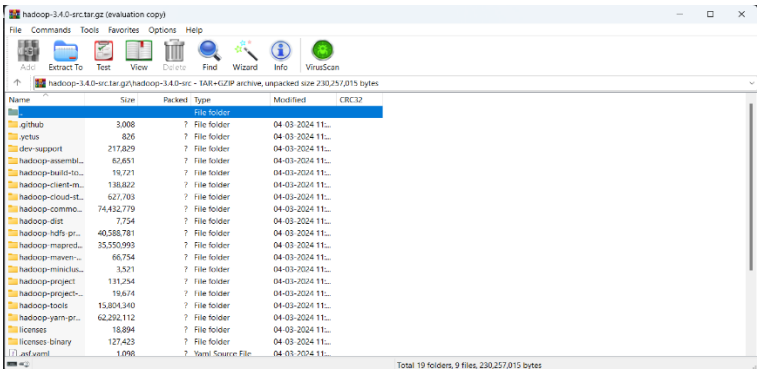

ORACLE
Products Industries Resources Customers Partners Developers Events
View Accounts
Contact Sales
Solaris x64 (SVR4 package)
jdk-8u202-solaris-x64.tar.Z
Solaris x64
Windows x86
Windows x64
Java SE Development Kit 8 Update 191 (64-bit) - Complete
Java
Java SE Development Kit 8 Update 191 (64-bit) Successfully Installed
Click Next Steps to access tutorials, API documentation, developer guides, release notes and more to help you get started with the JDK.
Next Steps
Close
Java SE Runtime Env
This software is licensed under the Oracle Binary Code Lic


Apache Hadoop
Download
Documentation
Community
Development
Help
Apache Software Foundation
Download
Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-512.
Version
Release date
Source download
Binary download
Release notes
3.4.0
2024 Mar 17
source (checksum signature)
binary (checksum signature)
binary-aarch64 (checksum signature)
Announcement
3.3.6
2023 Jun 23
source (checksum signature)
binary (checksum signature)
binary-aarch64 (checksum signature)
Announcement
2.10.2
2022 May 31
source (checksum signature)
binary (checksum signature)
Announcement

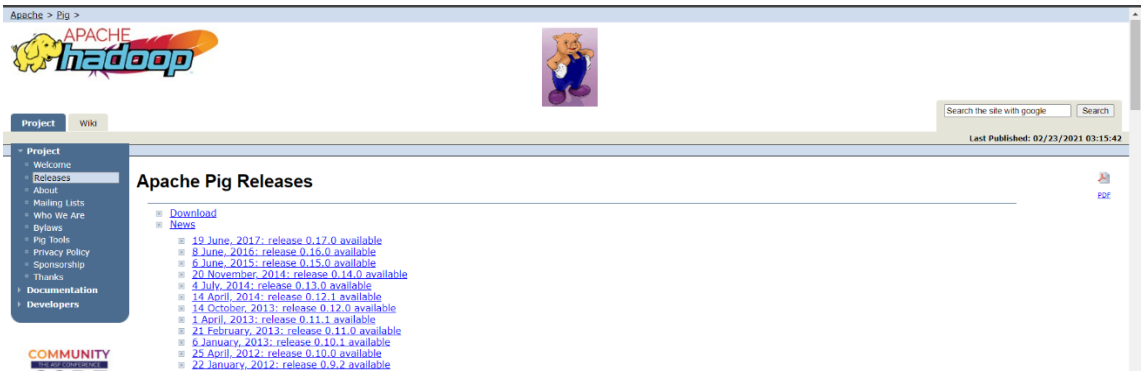

Sponsor the ASF
Community
Projects
Downloads
Learn
Resources & Tools
About
APACHE
SOFTWARE FOUNDATION
We suggest the following location for your download:
https://dlcdn.apache.org/hadoop/common/hadoop-3.4.0/hadoop-3.4.0-src.tar.gz
Alternate download locations are suggested below.
It is essential that you verify the integrity of the downloaded file using the PGP signature (.asc file) or a hash (.md5 or .sha* file).
HTTP
https://dlcdn.apache.org/hadoop/common/hadoop-3.4.0/hadoop-3.4.0-src.tar.gz
BACKUP SITES
https://dlcdn.apache.org/hadoop/common/hadoop-3.4.0/hadoop-3.4.0-src.tar.gz

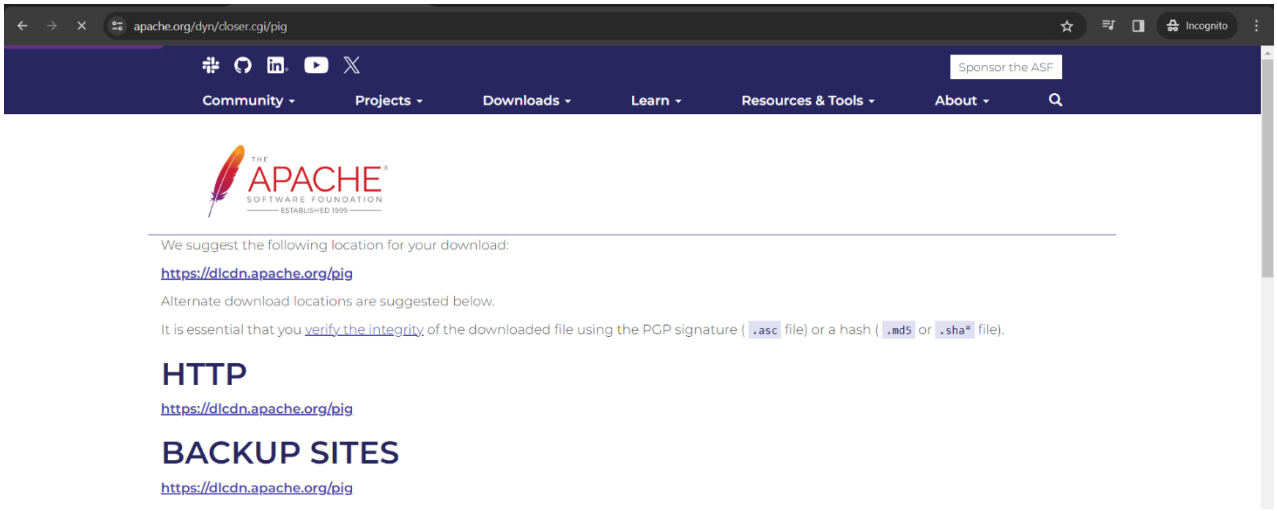

Chrome
chrome://downloads
Downloads
Search downloads
hadoop-3.4.0-src.tar.gz
37.1 MB • Done
Today
hadoop-3.4.0-src.tar.gz
https://dlcdn.apache.org

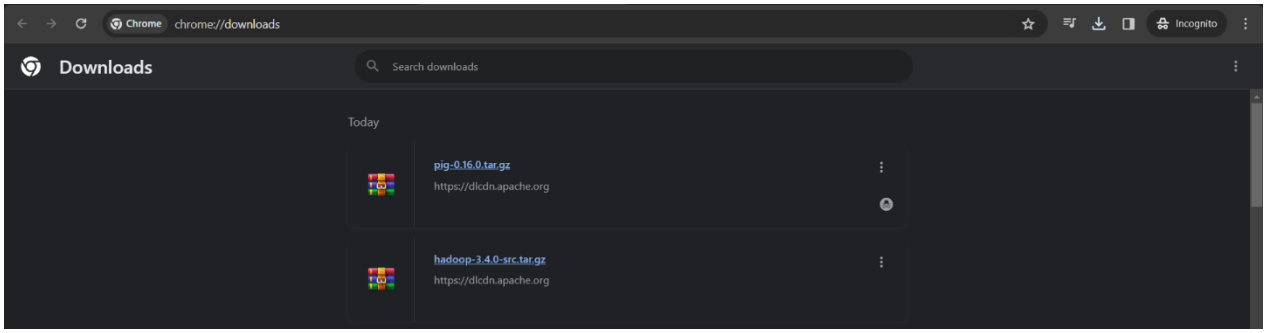

This PC
New Volume (D:)
hadoop-3.4.0-src
Search hadoop-3.4.0-src
New
Sort
View
Preview
Gallery
OneDrive - Pers
Desktop
Downloads
Documents
Pictures
Name
Date modified
Type
Size
hadoop-common-project
08-04-2024 01:22 PM
File folder
hadoop-yarn-project
08-04-2024 01:22 PM
File folder
licenses
08-04-2024 01:21 PM
File folder
NOTICE
04-03-2024 11:14 AM
Text Document
2 KB
README
04-03-2024 11:14 AM
Text Document
1 KB

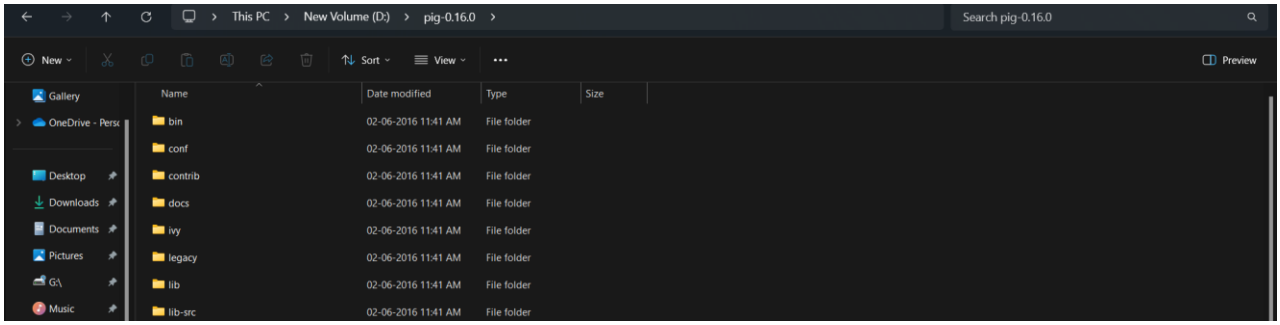


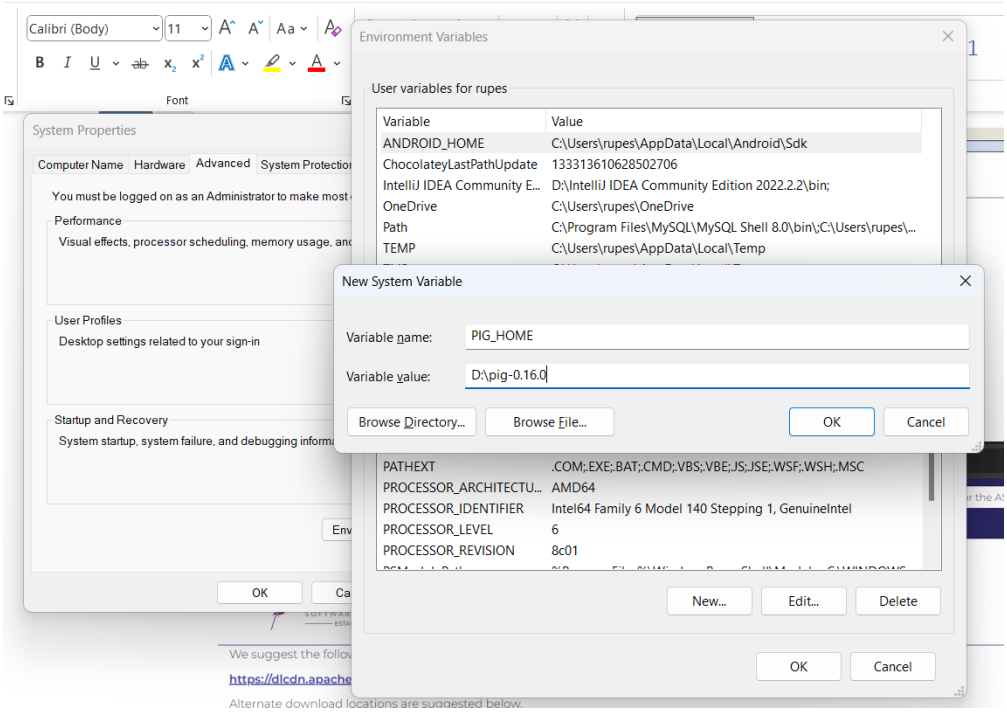
Pig Installation :



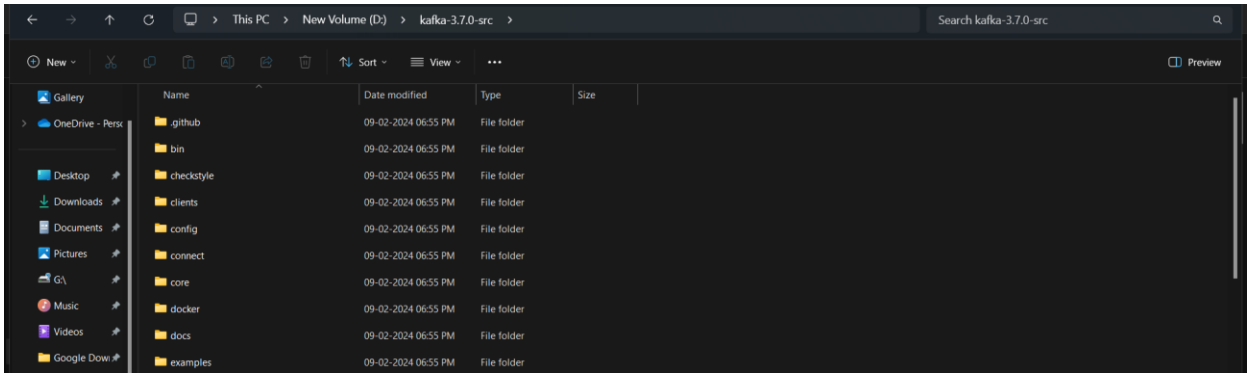


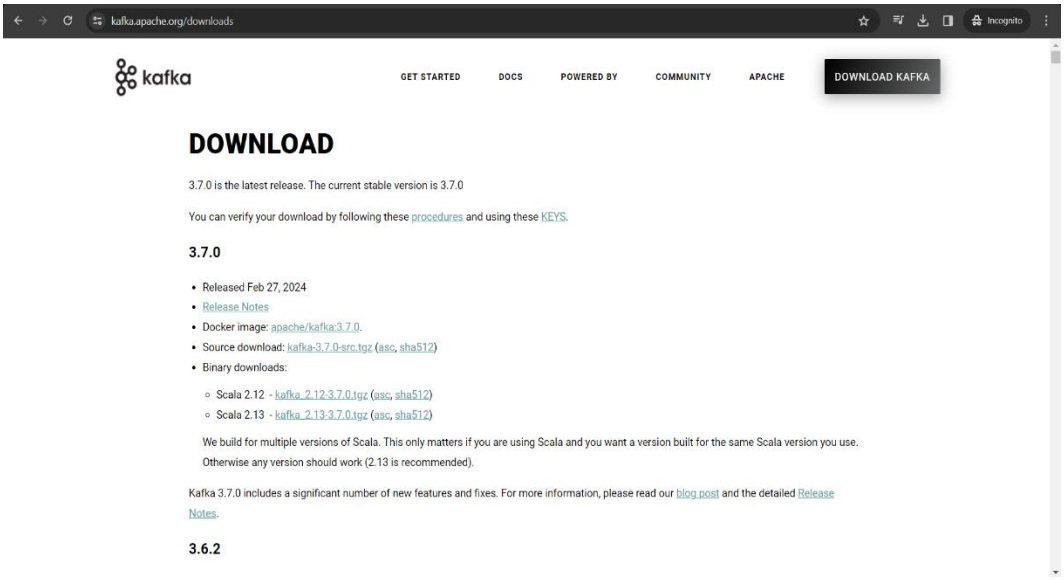


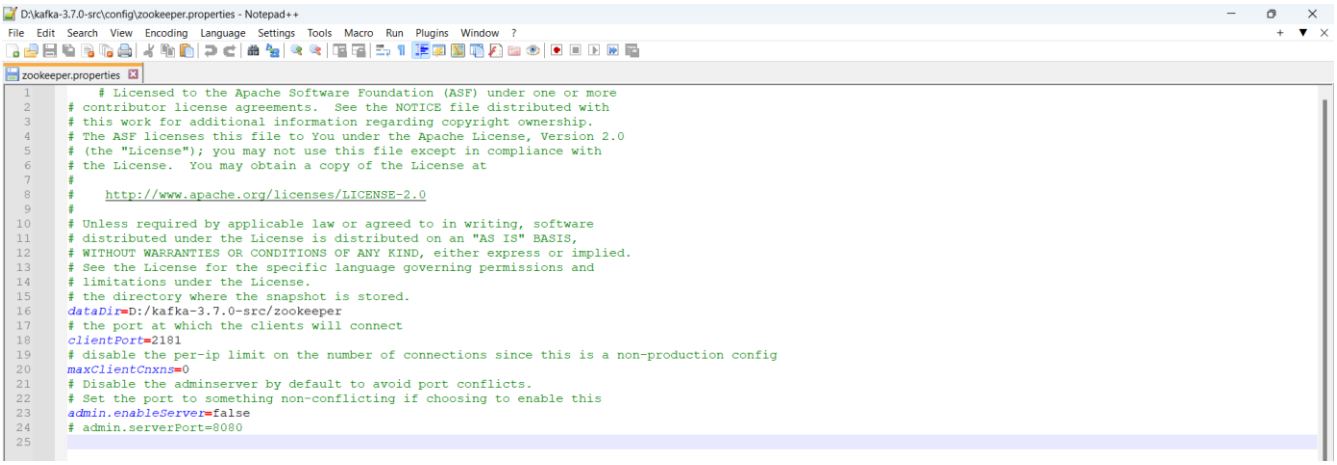


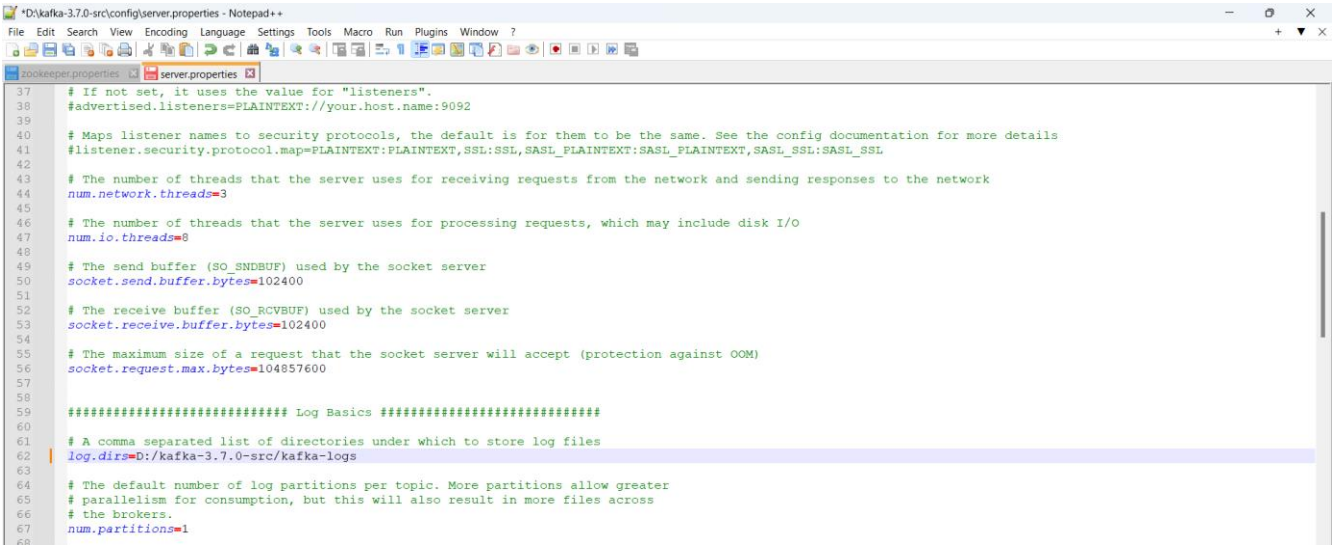


Kafka installation:




kafka.apache.org/downloads
Incognito
kafka
GET STARTED
DOCS
POWERED BY
COMMUNITY
APACHE
DOWNLOAD KAFKA
DOWNLOAD
3.7.0 is the latest release. The current stable version is 3.7.0
You can verify your download by following these procedures and using these KEYS.
3.7.0
Released Feb 27, 2024
Release Notes
Docker image: apache/kafka:3.7.0.
Source download: kafka-3.7.0-src.tgz (asc, sha512)
Binary downloads:
Scala 2.12 - kafka_2.12-3.7.0.tgz (asc, sha512)
Scala 2.13 - kafka_2.13-3.7.0.tgz (asc, sha512)
We build for multiple versions of Scala. This only matters if you are using Scala and you want a version built for the same Scala version you use.
Otherwise any version should work (2.13 is recommended).
Kafka 3.7.0 includes a significant number of new features and fixes. For more information, please read our blog post and the detailed Release Notes.
3.6.2


D:\kafka-3.7.0-src\config\zookeeper.properties - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
zookeeper.properties
Licensed to the Apache Software Foundation (ASF) under one or more
contributor license agreements. See the NOTICE file distributed with
this work for additional information regarding copyright ownership.
The ASF licenses this file to You under the Apache License, Version 2.0
(the "License"); you may not use this file except in compliance with
the License. You may obtain a copy of the License at
#
http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
# the directory where the snapshot is stored.
dataDir=D:/kafka-3.7.0-src/zookeeper
# the port at which the clients will connect
clientPort=2181
# disable the per-ip limit on the number of connections since this is a non-production config
maxClientCnxns=0
# Disable the adminserver by default to avoid port conflicts.
# Set the port to something non-conflicting if choosing to enable this
admin.enableServer=false
# admin.serverPort=8080


*D:\kafka-3.7.0-src\config\server.properties - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
zookeeper.properties
server.properties
If not set, it uses the value for "listeners".
#advertised.listeners=PLAINTEXT://your.host.name:9092
Maps listener names to security protocols, the default is for them to be the same. See the config documentation for more details
#listener.security.protocol.map=PLAINTEXT:PLAINTEXT,SSL:SSL,SASL_PLAINTEXT:SASL_PLAINTEXT,SASL_SSL:SASL_SSL
# The number of threads that the server uses for receiving requests from the network and sending responses to the network
num.network.threads=3
# The number of threads that the server uses for processing requests, which may include disk I/O
num.io.threads=8
# The send buffer (SO_SNDBUF) used by the socket server
socket.send.buffer.bytes=102400
# The receive buffer (SO_RCVBUF) used by the socket server
socket.receive.buffer.bytes=102400
# The maximum size of a request that the socket server will accept (protection against OOM)
socket.request.max.bytes=104857600
############################# Log Basics #############################
# A comma separated list of directories under which to store log files
log.dirs=D:/kafka-3.7.0-src/kafka-logs
# The default number of log partitions per topic. More partitions allow greater
# parallelism for consumption, but this will also result in more files across
# the brokers.
num.partitions=1

Name: Dhvani Patel

Roll no: 21BCP116

Big Data Analytics Lab

Practical 8: Implementation of graph data structure using networkx

jupyter bdalab_27_3_24 Last Checkpoint: 5 days ago

File Edit View Run Kernel Settings Help

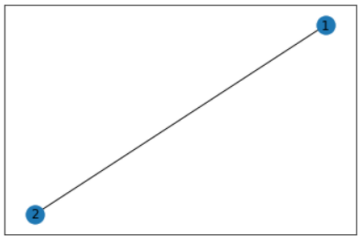
Trusted

Code

JupyterLab Python 3 (ipykernel)

[1]:
```
import networkx as nx
import matplotlib.pyplot as plt
```

[2]:
```
G = nx.Graph()

G.add_edge(1,2)

nx.draw_networkx(G)
plt.show()
```



jupyter bdalab_27_3_24 Last Checkpoint: 5 days ago

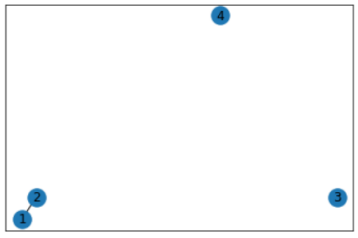
File Edit View Run Kernel Settings Help

Trusted

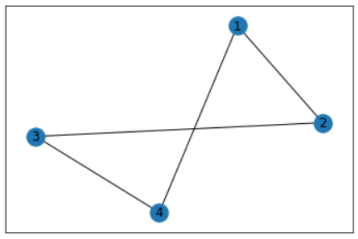
Code

JupyterLab Python 3 (ipykernel)

[3]:
```
G.add_nodes_from([3,4])
nx.draw_networkx(G)
plt.show()
```



[4]:
```
G.add_edge(3,4)
G.add_edges_from([(2,3),(4,1)])
nx.draw_networkx(G)
plt.show()
```



jupyter bdalab_27_3_24 Last Checkpoint: 5 days ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[5]:
```
G.nodes
```

[5]:
```
NodeView((1, 2, 3, 4))
```

[6]:
```
G.edges
```

[6]:
```
EdgeView([(1, 2), (1, 4), (2, 3), (3, 4)])
```

[7]:
```
list(nx.generate_adjlist(G))
```

[7]:
```
['1 2 4', '2 3', '3 4', '4']
```

[8]:
```
nx.to_dict_of_lists(G)
```

[8]:
```
{1: [2, 4], 2: [1, 3], 3: [4, 2], 4: [3, 1]}
```

[9]:
```
nx.to_edgelist(G)
```

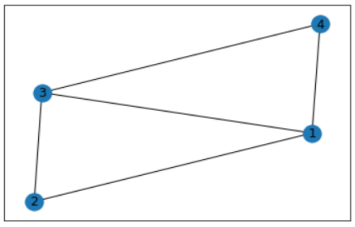
[9]:
```
EdgeDataView([(1, 2, {}), (1, 4, {}), (2, 3, {}), (3, 4, {})])
```

jupyter bdalab_27_3_24 Last Checkpoint: 5 days ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[10]:
```
G.add_edge(1,3)
nx.draw_networkx(G)
plt.show()
```



[11]:
```
G.degree
```

[11]:
```
DegreeView({1: 3, 2: 2, 3: 3, 4: 2})
```

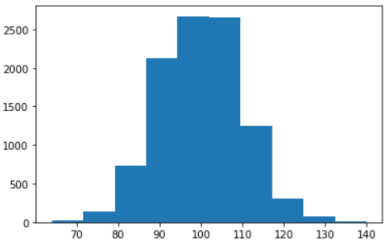
[12]:
```
kg=nx.fast_gnp_random_graph(10000,0.01)
k = kg.degree()
plt.hist(list(dict(k).values()))
plt.show()
```

jupyter bdalab_27_3_24 Last Checkpoint: 5 days ago

File Edit View Run Kernel Settings Help Trusted

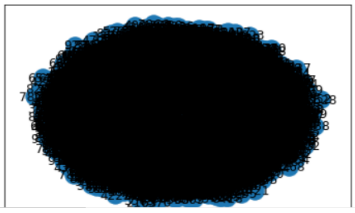
Code JupyterLab Python 3 (ipykernel)

[12]:
```
kg=nx.fast_gnp_random_graph(10000,0.01)
k = kg.degree()
plt.hist(list(dict(k).values()))
plt.show()
```



[13]:
```
nx.draw_networkx(kg)
plt.show
```

[13]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```

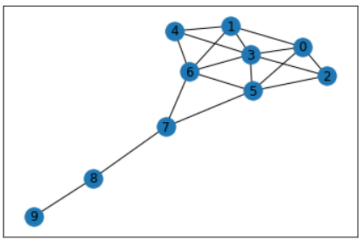


jupyter bdalab_27_3_24 Last Checkpoint: 5 days ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[14]:
```
G=nx.krackhardt_kite_graph()
nx.draw_networkx(G)
plt.show()
```



[15]:
```
print(nx.has_path(G,source=1, target=9))
print(nx.shortest_path(G,source=1, target=9))
print(nx.shortest_path_length(G,source=1, target=9))
paths=list(nx.all_pairs_shortest_path(G))
paths[5][1]
```

```
True
[1, 6, 7, 8, 9]
4
```

[15]:
```
{5: [5],
 0: [5, 0],
 2: [5, 2],
 3: [5, 3],
 6: [5, 6],
 7: [5, 7],
 1: [5, 0, 1],
 4: [5, 3, 4],
 8: [5, 7, 8],
 9: [5, 7, 8, 9]}
```

jupyter bdalab_27_3_24 Last Checkpoint: 5 days ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[16]:
```
nx.betweenness_centrality(G)
```

[16]:
```
{0: 0.023148148148148143,
 1: 0.023148148148148143,
 2: 0.0,
 3: 0.10185185185185183,
 4: 0.0,
 5: 0.2314814814814814,
 6: 0.2314814814814814,
 7: 0.38888888888888884,
 8: 0.2222222222222222,
 9: 0.0}
```

[17]:
```
nx.degree_centrality(G)
```

[17]:
```
{0: 0.4444444444444444,
 1: 0.4444444444444444,
 2: 0.3333333333333333,
 3: 0.6666666666666666,
 4: 0.3333333333333333,
 5: 0.5555555555555556,
 6: 0.5555555555555556,
 7: 0.3333333333333333,
 8: 0.2222222222222222,
 9: 0.1111111111111111}
```

[18]:
```
nx.closeness_centrality(G)
```

[18]:
```
{0: 0.5294117647058824,
 1: 0.5294117647058824,
 2: 0.5,
 3: 0.6,
 4: 0.5,
 5: 0.6428571428571429,
 6: 0.6428571428571429,
 7: 0.6,
 8: 0.42857142857142855,
 9: 0.3103448275862069}
```

jupyter bdalab_27_3_24 Last Checkpoint: 5 days ago

File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab Python 3 (ipykernel)

[19]:
```
nx.harmonic_centrality(G)
```

[19]:
```
{0: 6.083333333333333,
 1: 6.083333333333333,
 2: 5.583333333333333,
 3: 7.083333333333333,
 4: 5.583333333333333,
 5: 6.833333333333333,
 6: 6.833333333333333,
 7: 6.0,
 8: 4.666666666666666,
 9: 3.4166666666666665}
```

[20]:
```
nx.eigenvector_centrality(G)
```

[20]:
```
{0: 0.35220898139203594,
 1: 0.35220898139203594,
 2: 0.28583473531632414,
 3: 0.48102048812210046,
 4: 0.28583473531632414,
 5: 0.3976910106255469,
 6: 0.3976910106255469,
 7: 0.1958618517536038 2,
 8: 0.048074775014202945,
 9: 0.01116405857582424}
```

[21]:
```
nx.clustering(G)
```

[21]:
```
{0: 0.6666666666666666,
 1: 0.6666666666666666,
 2: 1.0,
 3: 0.5333333333333333,
 4: 1.0,
 5: 0.5,
 6: 0.5,
 7: 0.3333333333333333,
 8: 0,
 9: 0}
```

jupyter bdalab_27_3_24 Last Checkpoint: 5 days ago

File Edit View Run Kernel Settings Help

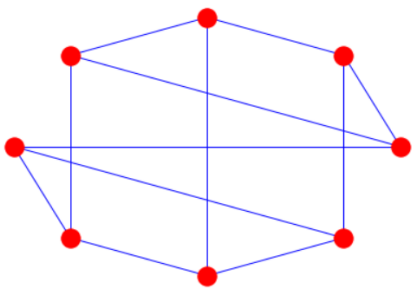
Trusted

Code

JupyterLab Python 3 (ipykernel)

```
 7: 0.3333333333333333,
 8: 0,
 9: 0}
```

[22]:
```
G=nx.cubical_graph(G)

nx.draw(G, pos=nx.circular_layout(G),node_color='r',edge_color='b')
```



[ ]: