05/12/2024

Versions:-
Py → 3.8 ; Java → Above 8
R → 3.5 ; Scala → 2.12/2.13

classmate
Date
Page

PySpark
- Apache spark library written in py to run python application
- We run application parallely on distributed cluster.
- Pyspark is python API which is an analytical process engine for large-scale data processing and ML applications

Apache Spark
- Open source, 100 times faster than traditional application like Hadoop
- ~~Open~~ Runs single and multi node clusters
- It creates to address the limitations of map-reduce by doing in-memory processing.

Features of py-spark
- Fault tolerence
- lazy-evaluation
- Inbulit-optimization when using Dataframes.
- Immutable
- In-memory computation.
- Cache and persistance
- Distributed processing using parallelize
- Used many clusters
- Supports ANSI SQL.

Advantages of Pyspark:-
- General purpose, in-memory, distributed processing engine allows data processing efficently in distributed fashion
- 100X faster than traditional
- Get great benefits from using Pyspark for data ingestion pipeli
- Using Pyspark we can process data from Hadoop
- process real time data using streaming & Kfka

Module and Package
Pyspark RDD (Pyspark.RDD) ; Pyspark streaming (Pyspark.streaming)
Pyspark Dataframe & SQL (Pyspark.SQL) , Pyspark Graph Frames (Graph fra

databricks

saisachin1616@gmail.com

Reading the csv file using spark.read.csv

Python

File Edit View Run Help Last edit was 1 minute ago New cell UI: OFF

Run all | My Cluster | Share | Publish

```
1  import pyspark
2  from pyspark.sql import SparkSession
3  spark = SparkSession.builder.appName("practice").getOrCreate()
4
5  spark
6  df=spark.read.csv("/FileStore/tables/Profile.csv")
7  df
8  df.show()
```

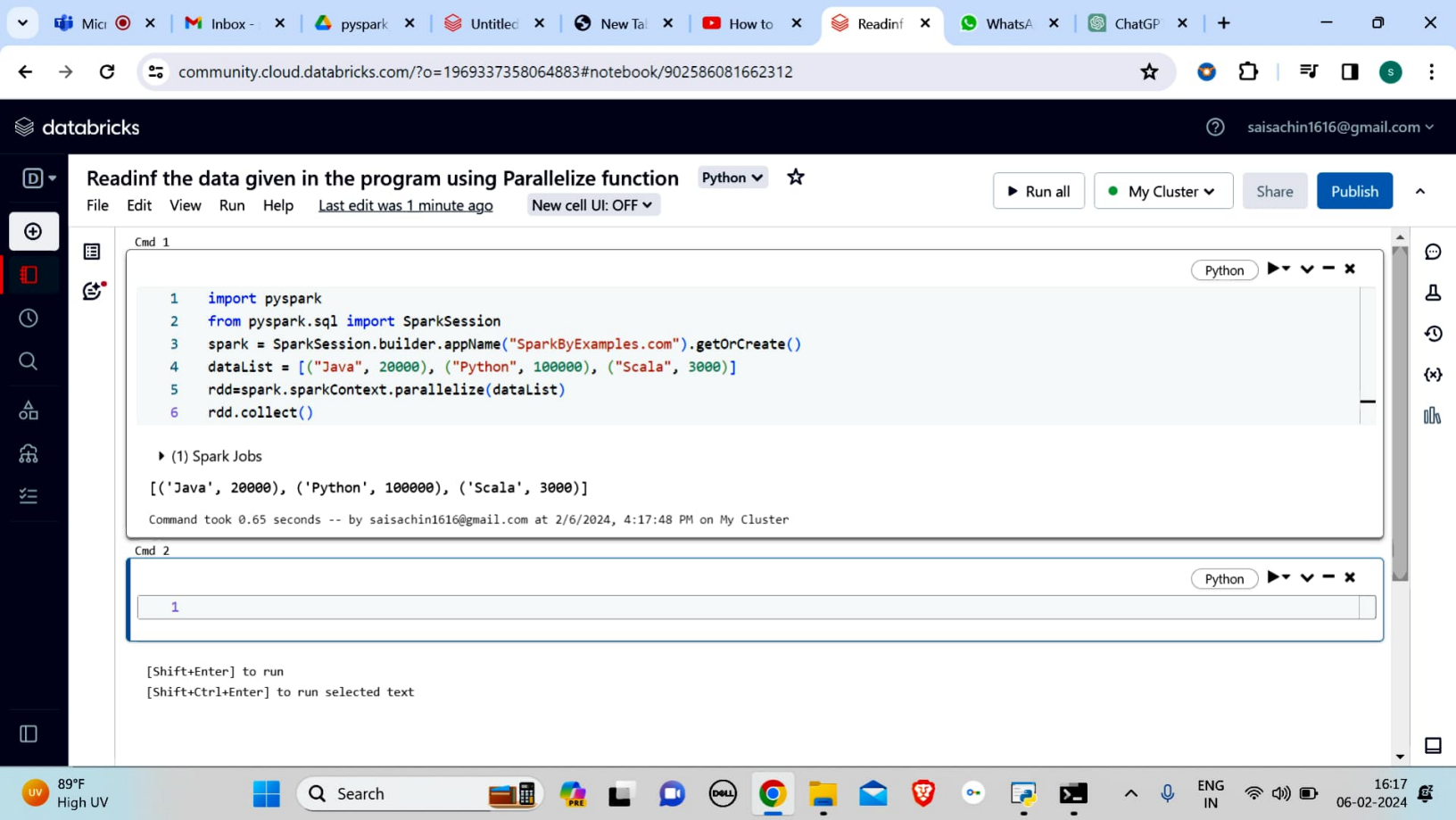
▸ (2) Spark Jobs

▸ df: pyspark.sql.dataframe.DataFrame = [_c0: string, _c1: string ... 1 more field]

```
+----+---+----------+
| _c0|_c1|       _c2|
+----+---+----------+
|Name|Age|Occupation|
|John| 25|  Engineer|
|Jane| 30|    Doctor|
| Bob| 22|   Student|
+----+---+----------+
```

Command took 2.63 seconds -- by saisachin1616@gmail.com at 2/6/2024, 4:14:33 PM on My Cluster

Cmd 2


Readinf
community.cloud.databricks.com/?o=1969337358064883#notebook/902586081662312
databricks
saisachin1616@gmail.com
Readinf the data given in the program using Parallelize function
Python
File Edit View Run Help Last edit was 1 minute ago
New cell UI: OFF
Run all
My Cluster
Share
Publish
Cmd 1
Python
1 import pyspark
2 from pyspark.sql import SparkSession
3 spark = SparkSession.builder.appName("SparkByExamples.com").getOrCreate()
4 dataList = [("Java", 20000), ("Python", 100000), ("Scala", 3000)]
5 rdd=spark.sparkContext.parallelize(dataList)
6 rdd.collect()
(1) Spark Jobs
[('Java', 20000), ('Python', 100000), ('Scala', 3000)]
Command took 0.65 seconds -- by saisachin1616@gmail.com at 2/6/2024, 4:17:48 PM on My Cluster
Cmd 2
Python
1
[Shift+Enter] to run
[Shift+Ctrl+Enter] to run selected text
89°F
High UV
Search
ENG
IN
16:17
06-02-2024