

## Content

I.	Project Details	1
II.	Databricks Workspace	2
	a. Create a Cluster in Databricks	2
	b. Create a Notebook under the Cluster	2
	c. Load the StudentData.csv file in Data	3
III.	Importing Pyspark, SparkSession and Dataset	4
	a. Starting Spark Session in PySpark	
	b. Reading "StudentData" dataset with respect to Spark	
IV.	Queries	
	1. Show the number of students in the file.	5
	2. Show the total marks achieved by Female and Male students.	5
	3. Show the total number of students that have passed and failed. 50+ marks are required to pass the course.	6
	4. Show the total number of students enrolled per course.	6
	5. Show the total marks that students have achieved per course	7
	6. Show the average marks that students have achieved per course.	7
	7. Show the minimum and maximum marks achieved per course	8
	8. Show the average age of male and female students.	9

## Project Details :

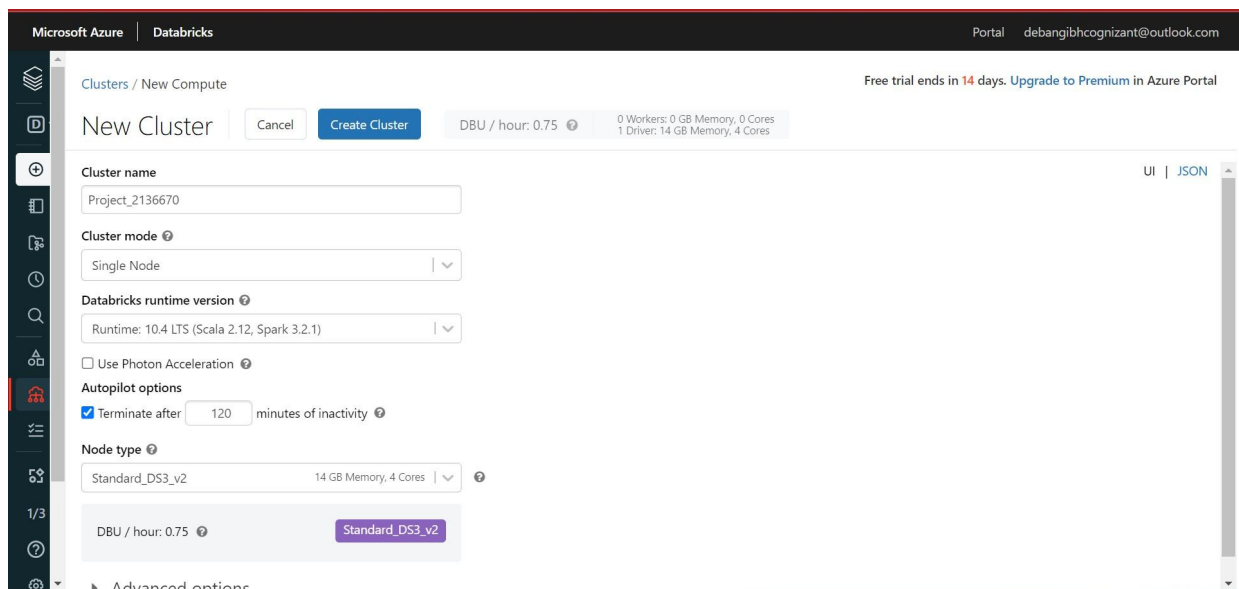
---

To execute the following tasks in **Azure Databricks** using **Spark** using the StudentData csv file.

1. Show the number of students in the file.
2. Show the total marks achieved by Female and Male students.
3. Show the total number of students that have passed and failed. 50+ marks are required to pass the course
4. Show the total number of students enrolled per course.
5. Show the total marks that students have achieved per course.
6. Show the average marks that students have achieved per course.
7. Show the minimum and maximum marks that students have achieved per course.
8. Show the average age of male and female students.

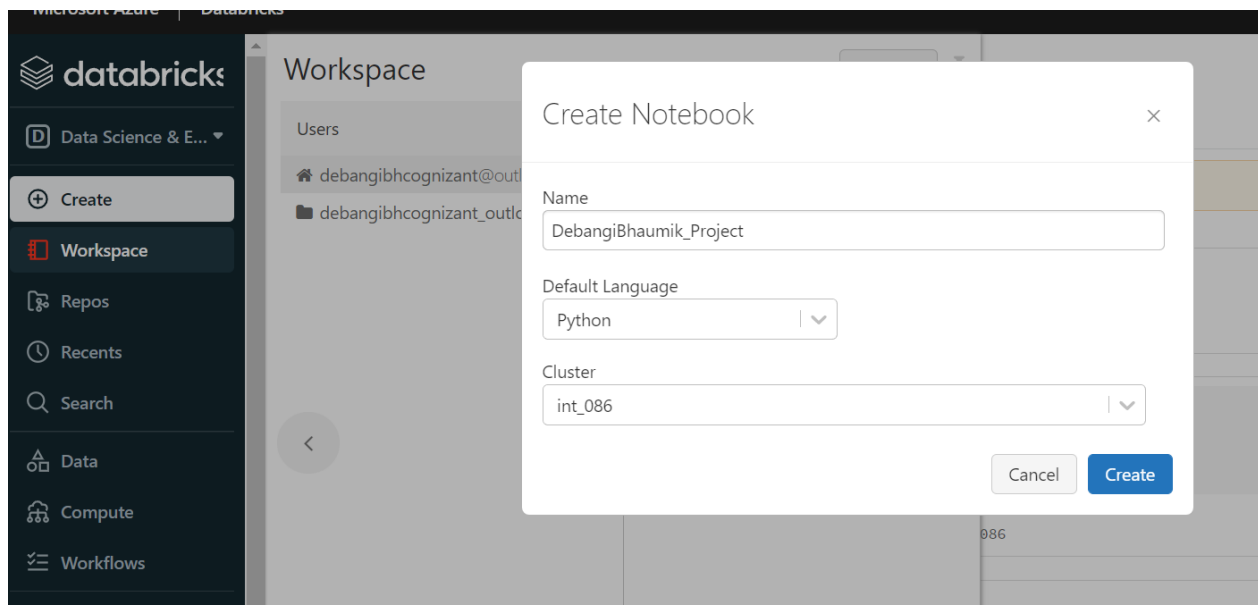
# Databricks Workspace

## 1. Create a Cluster in Databricks



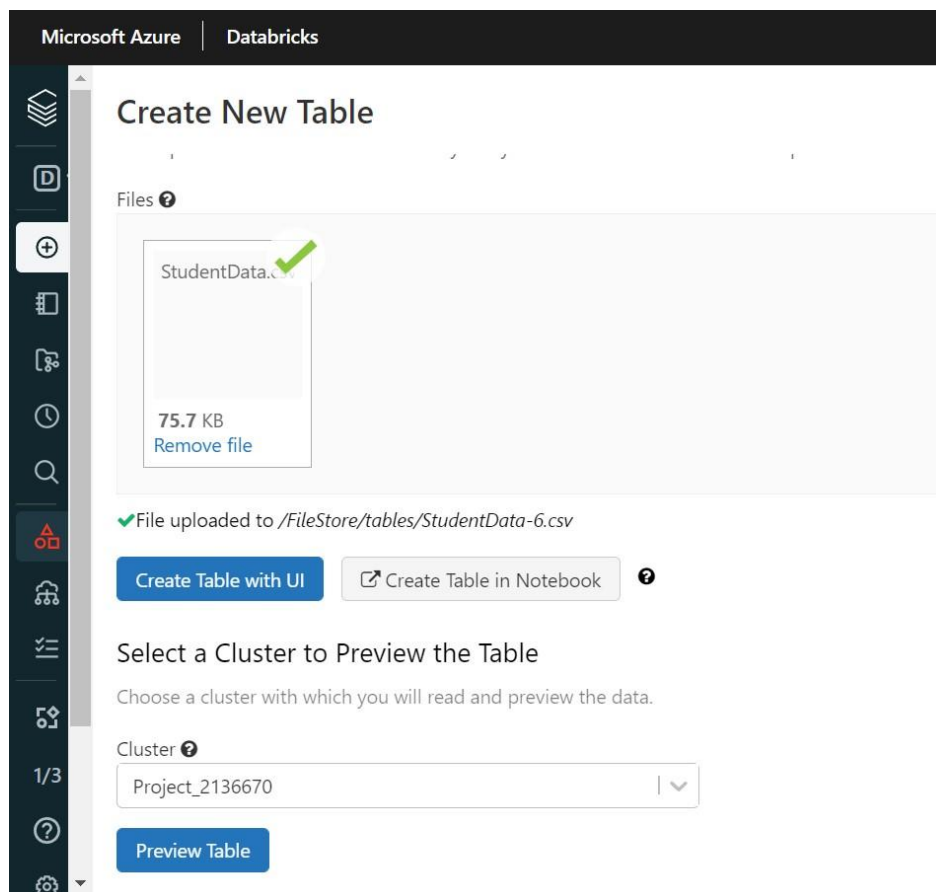
The screenshot shows the 'New Cluster' configuration page in the Databricks Azure portal. The page has a dark sidebar on the left with navigation icons. The main content area is titled 'New Cluster' and includes a 'Cancel' button and a 'Create Cluster' button. Below these, there are several configuration options: 'Cluster name' (text input with 'Project\_2136670'), 'Cluster mode' (dropdown menu with 'Single Node'), 'Databricks runtime version' (dropdown menu with 'Runtime: 10.4 LTS (Scala 2.12, Spark 3.2.1)'), 'Use Photon Acceleration' (checkbox), 'Autopilot options' (checkbox 'Terminate after' with a value of '120' minutes of inactivity), and 'Node type' (dropdown menu with 'Standard\_DS3\_v2'). A summary box at the bottom shows 'DBU / hour: 0.75' and 'Standard\_DS3\_v2'. The top right corner indicates 'Free trial ends in 14 days. Upgrade to Premium in Azure Portal'.

## 2. Create a Notebook under the Cluster



The screenshot shows the 'Create Notebook' dialog box in the Databricks workspace. The dialog box is titled 'Create Notebook' and has a close button (X) in the top right corner. It contains three fields: 'Name' (text input with 'DebangiBhaumik\_Project'), 'Default Language' (dropdown menu with 'Python'), and 'Cluster' (dropdown menu with 'int\_086'). At the bottom right, there are 'Cancel' and 'Create' buttons. The background shows the Databricks workspace interface with a sidebar on the left and a 'Workspace' view on the right.

### 3. Load the StudentData.csv file in Data



# Importing Pyspark, SparkSession and Dataset

---

- Starting Spark Session in PySpark

Cmd 1

## Starting Spark Session in PySpark

Cmd 2

```
1 import pyspark
2 from pyspark.sql import SparkSession
3 spark = SparkSession.builder.appName("Azure_Databricks_Project").getOrCreate
```

Command took 0.03 seconds -- by debangibhcnizant@outlook.com at 7/31/2022, 7:23:15 PM on int\_086

- Reading "StudentData" dataset with respect to Spark

Cmd 3



## Reading "StudentData" dataset with respect to Spark

Cmd 4

```
1 df = spark.read.options(header="True",inferSchema="True",delimiter=',').csv("/FileStore/tables/StudentData-8.csv")
```

▸ (2) Spark Jobs

▼ df: pyspark.sql.dataframe.DataFrame

```
age: integer
gender: string
name: string
course: string
roll: integer
marks: integer
email: string
```

Command took 0.60 seconds -- by debangibhcnizant@outlook.com at 7/31/2022, 7:08:50 PM on int\_086

## Queries :

---

### 1. Show the number of students in the file.

Cmd 8

#### Question 1 : Show the number of students in the file.

Cmd 9

```
1 print("Total number of Students : ", df.count())
```

► (2) Spark Jobs

Total number of Students : 1000

Command took 0.43 seconds -- by debangibhcnizant@outlook.com at 7/31/2022, 11:57:22 AM on Project\_2136670

### 2. Show the total marks achieved by Female and Male students.

Cmd 10

#### Question 2 : Show the total marks achieved by Female and Male students

Cmd 11

```
1 print("The total marks achieved by Female and Male Students : ")
2 df.groupBy("gender").sum("marks").show()
```

► (2) Spark Jobs

The total marks achieved by Female and Male Students :

```
+-----+-----+
|gender|sum(marks)|
+-----+-----+
|Female|      29636|
|  Male|      30461|
+-----+-----+
```

Command took 0.59 seconds -- by debangibhcnizant@outlook.com at 7/31/2022, 7:40:55 PM on int\_086

### 3. Show the total number of students that have passed and failed. 50+ marks are required to pass the course.

**Question 3 : Show the total number of students that have passed and failed. 50+ marks are required to pass the course.**

```
Cmd 13
```

```
1 from pyspark.sql.functions import when
2 df = df.withColumn("Result",when(df.marks>=50,"Pass").when(df.marks<50,"Fail").when(df.marks.isNull(),"0").otherwise(df.marks))
3 print("The total number of students that have passed and failed : ")
4 df.groupBy("Result").count().show()
```

Python ▶ ▾ - ✕

▶ (2) Spark Jobs

▶ df: pyspark.sql.dataframe.DataFrame = [age: integer, gender: string ... 6 more fields]

The total number of students that have passed and failed :

Result	count
Fail	356
Pass	644

Command took 0.57 seconds -- by debangibhcnizant@outlook.com at 7/31/2022, 7:41:28 PM on int\_086

### 4. Show the total number of students enrolled per course.

**Question 4 : Show the total number of students enrolled per course**

```
Cmd 15
```

```
1 print("The total number of students enrolled per course :")
2 df.groupBy("course").count().show()
```

▶ (2) Spark Jobs

The total number of students enrolled per course :

course	count
PF	166
DB	157
MVC	157
DSA	176
Cloud	192
OOP	152

Command took 1.91 seconds -- by debangibhcnizant@outlook.com at 7/31/2022, 6:50:30 PM on int\_086

## 5. Show the total marks that students have achieved per course

### Question 5 : Show the total marks that students have achieved per course

Cmd 17

```
1 print("The total marks that students have achieved per course.")
2 df.groupBy("Course").sum("marks").withColumnRenamed("sum(marks)", "Total_Marks").show()
```

► (2) Spark Jobs

The total marks that students have achieved per course.

```
+-----+-----+
| Course|Total_Marks|
+-----+-----+
|    PF|      9933|
|    DB|      9270|
|   MVC|      9585|
|   DSA|     10950|
| Cloud|     11443|
|   OOP|      8916|
+-----+-----+
```

Command took 0.34 seconds -- by debangibhcnizant@outlook.com at 7/31/2022, 6:56:09 PM on int\_086

## 6. Show the average marks that students have achieved per course.

### Question 6 : Show the average marks that students have achieved per course

Cmd 19

```
1 print("Average marks of students per course:")
2 df.groupBy("Course").avg("marks").withColumnRenamed("avg(marks)", "Average_Marks").show()
```

► (2) Spark Jobs

Average marks of students per course:

```
+-----+-----+
| Course|Average_Marks|
+-----+-----+
|    PF| 59.83734939759036|
|    DB| 59.044585987261144|
|   MVC| 61.05095541401274|
|   DSA| 62.21590909090909|
| Cloud| 59.598958333333336|
|   OOP| 58.6578947368421|
+-----+-----+
```

Command took 0.58 seconds -- by debangibhcnizant@outlook.com at 7/31/2022, 6:58:30 PM on int\_086



## 7. Show the minimum and maximum marks achieved per course

### Question 7 : Show the minimum and maximum marks achieved per course

Cmd 21

```
1 print("Maximum marks achieved per course :")
2 df.groupBy("course").agg({'marks':'max'}).show()
3
4
```

► (2) Spark Jobs

Maximum marks achieved per course :

```
+-----+-----+
|course|max(marks)|
+-----+-----+
|   PF   |      99   |
|   DB   |      98   |
|   MVC  |      99   |
|   DSA  |      99   |
| Cloud |      99   |
|   OOP  |      99   |
+-----+-----+
```

```
1 print("Minimum marks achieved per course :")
2 df.groupBy("course").agg({'marks':'min'}).show()
```

► (2) Spark Jobs

Minimum marks achieved per course :

```
+-----+-----+
|course|min(marks)|
+-----+-----+
|   PF   |      20   |
|   DB   |      20   |
|   MVC  |      22   |
|   DSA  |      20   |
| Cloud |      20   |
|   OOP  |      20   |
+-----+-----+
```

## 8. Show the average age of male and female students.

### Question 8 : Show the average age of male and female students

Cmd 24

```
1 df.groupBy('gender').avg('age').show()
```

► (2) Spark Jobs

```
+-----+-----+
|gender|      avg(age)|
+-----+-----+
|Female|28.489021956087825|
|  Male| 28.52304609218437|
+-----+-----+
```

Command took 0.51 seconds -- by debangibhcognizant@outlook.com at 7/31/2022, 7:06:42 PM on int\_086