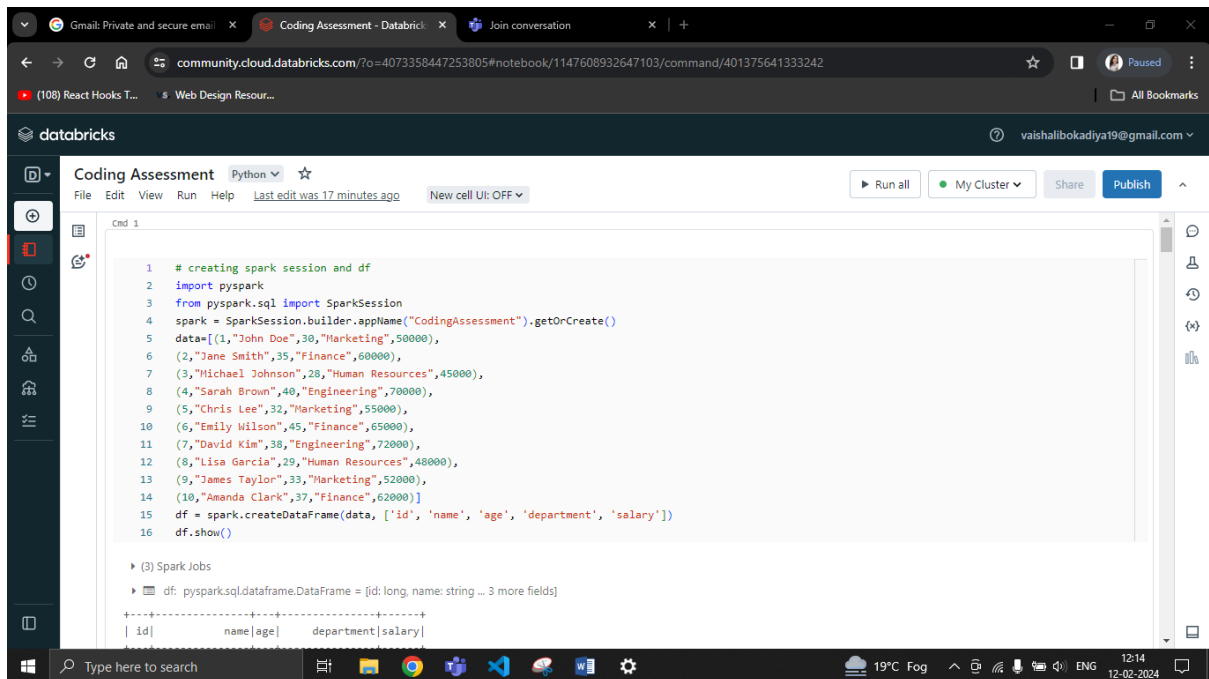


VAISHALI BOKADIYA
CODING ASSESSMENT
QUESTION 1

Execute Manipulating, Dropping, Sorting, Aggregations, Joining, GroupBy DataFrames:

Creating session and df:



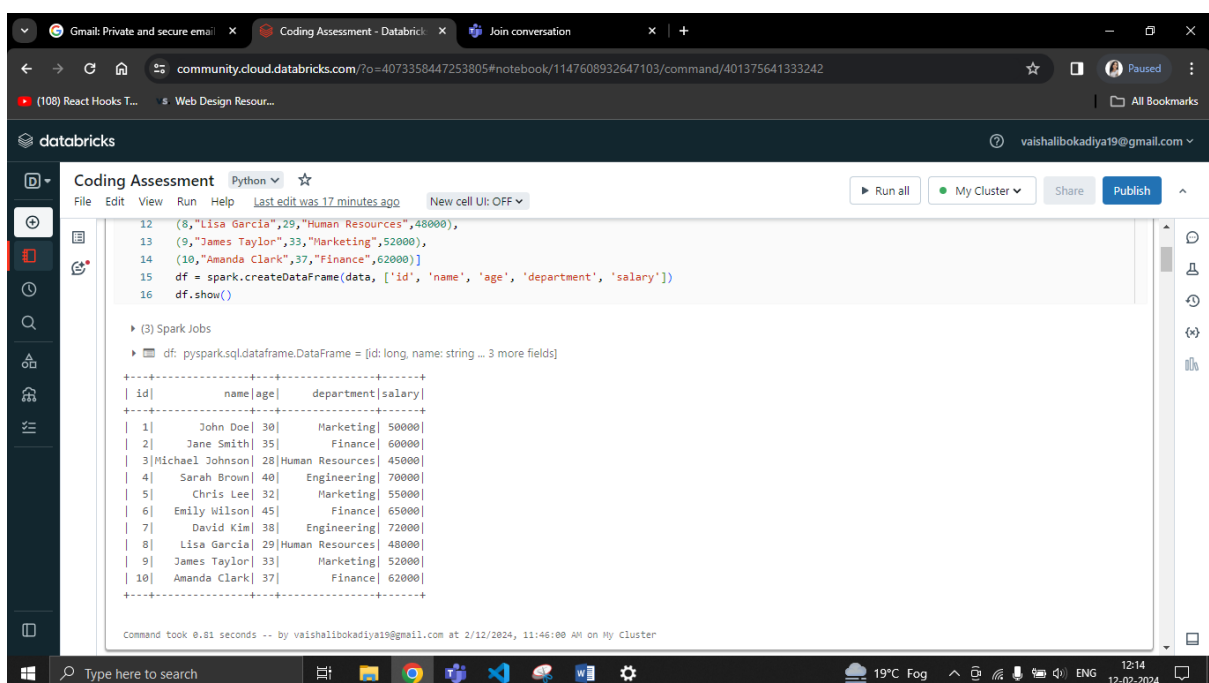
The screenshot shows a Databricks notebook titled "Coding Assessment" in Python. The code in the first cell creates a Spark session and a DataFrame with 10 rows of employee data. The output shows the Spark Jobs and the DataFrame structure.

```
1 # creating spark session and df
2 import pyspark
3 from pyspark.sql import SparkSession
4 spark = SparkSession.builder.appName("CodingAssessment").getOrCreate()
5 data=[(1,"John Doe",30,"Marketing",50000),
6 (2,"Jane Smith",35,"Finance",60000),
7 (3,"Michael Johnson",28,"Human Resources",45000),
8 (4,"Sarah Brown",40,"Engineering",70000),
9 (5,"Chris Lee",32,"Marketing",55000),
10 (6,"Emily Wilson",45,"Finance",65000),
11 (7,"David Kim",38,"Engineering",72000),
12 (8,"Lisa Garcia",29,"Human Resources",48000),
13 (9,"James Taylor",33,"Marketing",52000),
14 (10,"Amanda Clark",37,"Finance",62000)]
15 df = spark.createDataFrame(data, ['id', 'name', 'age', 'department', 'salary'])
16 df.show()
```

(3) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [id: long, name: string ... 3 more fields]

| id | name | age | department | salary |
|----|------|-----|------------|--------|
|----|------|-----|------------|--------|



The screenshot shows the same Databricks notebook, but the code is now partially obscured by the output. The output displays the DataFrame content as a table with 10 rows of employee data.

```
12 (8,"Lisa Garcia",29,"Human Resources",48000),
13 (9,"James Taylor",33,"Marketing",52000),
14 (10,"Amanda Clark",37,"Finance",62000)]
15 df = spark.createDataFrame(data, ['id', 'name', 'age', 'department', 'salary'])
16 df.show()
```

(3) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [id: long, name: string ... 3 more fields]

| id | name | age | department | salary |
|----|-----------------|-----|-----------------|--------|
| 1 | John Doe | 30 | Marketing | 50000 |
| 2 | Jane Smith | 35 | Finance | 60000 |
| 3 | Michael Johnson | 28 | Human Resources | 45000 |
| 4 | Sarah Brown | 40 | Engineering | 70000 |
| 5 | Chris Lee | 32 | Marketing | 55000 |
| 6 | Emily Wilson | 45 | Finance | 65000 |
| 7 | David Kim | 38 | Engineering | 72000 |
| 8 | Lisa Garcia | 29 | Human Resources | 48000 |
| 9 | James Taylor | 33 | Marketing | 52000 |
| 10 | Amanda Clark | 37 | Finance | 62000 |

Command took 0.81 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:46:00 AM on My Cluster

Manipulating:

The screenshot shows a Databricks notebook titled "Coding Assessment" in Python. The code in the cell is:

```
1 # manipulating df
2 # changing column name
3 df.withColumnRenamed("salary", "salary_amount").show()
```

The output shows the DataFrame after renaming the "salary" column to "salary_amount".

| id | name | age | department | salary_amount |
|----|-----------------|-----|-----------------|---------------|
| 1 | John Doe | 30 | Marketing | 50000 |
| 2 | Jane Smith | 35 | Finance | 60000 |
| 3 | Michael Johnson | 28 | Human Resources | 45000 |
| 4 | Sarah Brown | 40 | Engineering | 70000 |
| 5 | Chris Lee | 32 | Marketing | 55000 |
| 6 | Emily Wilson | 45 | Finance | 65000 |
| 7 | David Kim | 38 | Engineering | 72000 |
| 8 | Lisa Garcia | 29 | Human Resources | 48000 |
| 9 | James Taylor | 33 | Marketing | 52000 |
| 10 | Amanda Clark | 37 | Finance | 62000 |

The screenshot shows a Databricks notebook titled "Coding Assessment" in Python. The code in the cell is:

```
1 print(df.take(3))
```

The output shows the first three rows of the DataFrame:

```
[Row(id=1, name='John Doe', age=30, department='Marketing', salary=50000), Row(id=2, name='Jane Smith', age=35, department='Finance', salary=60000), Row(id=3, name='Michael Johnson', age=28, department='Human Resources', salary=45000)]
```

Below this, the code in the next cell is:

```
1 # pivoting
2 df.groupBy("department").pivot("name").sum("salary").show()
```

The output shows the DataFrame after pivoting the data by department and summing the salary for each employee.

| department | Amanda Clark | Chris Lee | David Kim | Emily Wilson | James Taylor | Jane Smith | John Doe | Lisa Garcia | Michael Johnson | Sarah Brown |
|-----------------|--------------|-----------|-----------|--------------|--------------|------------|----------|-------------|-----------------|-------------|
| Engineering | NULL | NULL | 72000 | NULL | NULL | NULL | NULL | NULL | NULL | 70000 |
| Finance | 62000 | NULL | NULL | 65000 | NULL | 60000 | NULL | NULL | NULL | NULL |
| Marketing | NULL | 55000 | NULL | NULL | 52000 | NULL | 50000 | NULL | NULL | NULL |
| Human Resources | NULL | NULL | NULL | NULL | NULL | NULL | 48000 | 45000 | NULL | NULL |

Dropping missing values:

The screenshot shows a Databricks notebook titled "Coding Assessment" in Python. The code in the first cell reads a CSV file from the FileStore. The output shows a DataFrame with 5 rows and 3 columns: Name, Department, and Salary. The data is as follows:

| Name | Department | Salary |
|----------|------------|--------|
| Vaibhav | IT | 50000 |
| NULL | HR | 45000 |
| Shubham | IT | NULL |
| NULL | HR | 56000 |
| Vaishali | IT | NULL |

The command took 1.29 seconds to execute.

The screenshot shows the same Databricks notebook after running two more commands. The first command drops rows with any missing values, resulting in a DataFrame with only one row:

| Name | Department | Salary |
|---------|------------|--------|
| Vaibhav | IT | 50000 |

The second command drops rows with missing values in the 'Salary' column, resulting in a DataFrame with only one row:

| Name | Department | Salary |
|---------|------------|--------|
| Vaibhav | IT | 50000 |

The command took 0.36 seconds to execute.

community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333216

vaishalibokadiya19@gmail.com

Coding Assessment

Python

File Edit View Run Help Last edit was 19 minutes ago New cell UI: OFF

Run all My Cluster Share Publish

```
1 df_missing.na.drop(how="any", subset=["name"]).show()
```

(1) Spark Jobs

| Name | Department | Salary |
|----------|------------|--------|
| Vaibhav | IT | 50000 |
| Shubham | IT | NULL |
| Vaishali | IT | NULL |
| Nisha | NULL | 65000 |

Command took 0.31 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:01:37 PM on My Cluster

```
1 # sorting
2 df.sort(df["salary"].desc()).show()
```

(1) Spark Jobs

| Name | Department | Salary |
|----------|------------|--------|
| Vaibhav | IT | 50000 |
| Shubham | IT | NULL |
| Vaishali | IT | NULL |
| Nisha | NULL | 65000 |

19°C Fog 12:16 12-02-2024

Sorting:

community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333216

vaishalibokadiya19@gmail.com

Coding Assessment

Python

File Edit View Run Help Last edit was 19 minutes ago New cell UI: OFF

Run all My Cluster Share Publish

```
1 # sorting
2 df.sort(df["salary"].desc()).show()
```

(1) Spark Jobs

| id | name | age | department | salary |
|----|-----------------|-----|-----------------|--------|
| 7 | David Kim | 38 | Engineering | 72000 |
| 4 | Sarah Brown | 40 | Engineering | 70000 |
| 6 | Emily Wilson | 45 | Finance | 65000 |
| 10 | Amanda Clark | 37 | Finance | 62000 |
| 2 | Jane Smith | 35 | Finance | 60000 |
| 5 | Chris Lee | 32 | Marketing | 55000 |
| 9 | James Taylor | 33 | Marketing | 52000 |
| 1 | John Doe | 30 | Marketing | 50000 |
| 8 | Lisa Garcia | 29 | Human Resources | 48000 |
| 3 | Michael Johnson | 28 | Human Resources | 45000 |

Command took 0.77 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:46:12 AM on My Cluster

19°C Fog 12:16 12-02-2024

community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333216

databricks vaishalibokadiya19@gmail.com

Coding Assessment Python ☆

File Edit View Run Help Last edit was 19 minutes ago New cell UI: OFF

Run all My Cluster Share Publish

Command took 0.77 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:46:12 AM on My Cluster

Cmd 10

```
1 df.sort("age","salary").show()
```

(1) Spark Jobs

| id | name | age | department | salary |
|----|-----------------|-----|-----------------|--------|
| 3 | Michael Johnson | 28 | Human Resources | 45000 |
| 8 | Lisa Garcia | 29 | Human Resources | 48000 |
| 1 | John Doe | 30 | Marketing | 50000 |
| 5 | Chris Lee | 32 | Marketing | 55000 |
| 9 | James Taylor | 33 | Marketing | 52000 |
| 2 | Jane Smith | 35 | Finance | 60000 |
| 10 | Amanda Clark | 37 | Finance | 62000 |
| 7 | David Kim | 38 | Engineering | 72000 |
| 4 | Sarah Brown | 40 | Engineering | 70000 |
| 6 | Emily Wilson | 45 | Finance | 65000 |

Command took 0.41 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:46:34 AM on My Cluster

Cmd 11

GroupBy

community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333216

databricks vaishalibokadiya19@gmail.com

Coding Assessment Python ☆

File Edit View Run Help Last edit was 20 minutes ago New cell UI: OFF

Run all My Cluster Share Publish

Command took 0.41 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:46:34 AM on My Cluster

Cmd 11

```
1 # groupby
2 df.groupby("department").sum("salary").alias("total_salary").show()
```

(2) Spark Jobs

| department | sum(salary) |
|-----------------|-------------|
| Marketing | 157000 |
| Finance | 187000 |
| Human Resources | 93000 |
| Engineering | 142000 |

Command took 0.82 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:49:46 AM on My Cluster

Cmd 12

```
1 # aggregation
```

community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333243

vaishalibokadiya19@gmail.com

Coding Assessment Python

File Edit View Run Help Last edit was now New cell UI: OFF

Run all My Cluster Share Publish

```
df.groupBy("department").count().alias("number_of_employees").show()
```

Command took 0.82 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:49:46 AM on My Cluster

Cmd 12

```
df.groupBy("department").count().alias("number_of_employees").show()
```

(2) Spark Jobs

| department | count |
|-----------------|-------|
| Marketing | 3 |
| Finance | 3 |
| Human Resources | 2 |
| Engineering | 2 |

Command took 1.49 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:28:05 PM on My Cluster

Cmd 13

```
# aggregation
```

Aggregation:

community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333216

vaishalibokadiya19@gmail.com

Coding Assessment Python

File Edit View Run Help Last edit was 20 minutes ago New cell UI: OFF

Run all My Cluster Share Publish

```
# aggregation
df.groupBy("department").agg({"salary": "sum"}).show()
```

Command took 0.82 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:49:46 AM on My Cluster

Cmd 12

```
# aggregation
df.groupBy("department").agg({"salary": "sum"}).show()
```

(2) Spark Jobs

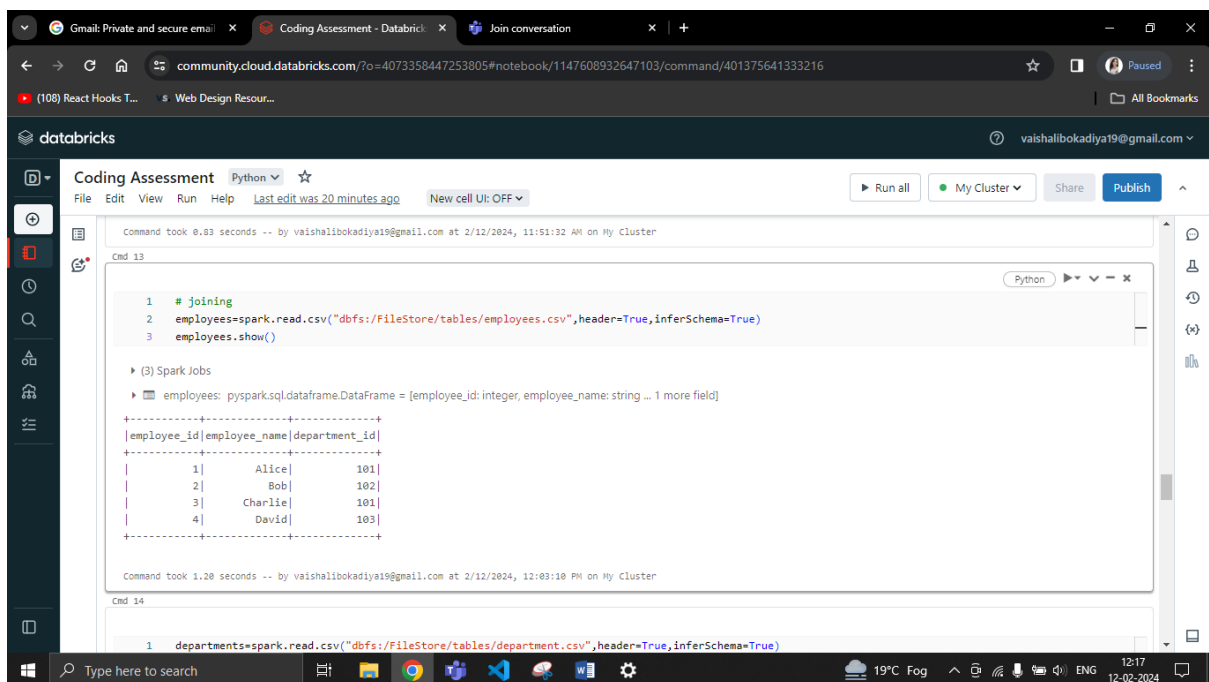
| department | sum(salary) |
|-----------------|-------------|
| Marketing | 157000 |
| Finance | 187000 |
| Human Resources | 93000 |
| Engineering | 142000 |

Command took 0.83 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:51:32 AM on My Cluster

Cmd 13

```
# joining
```

Joins



community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333216

databricks vaishalibokadiya19@gmail.com

Coding Assessment Python ☆

File Edit View Run Help Last edit was 20 minutes ago New cell UI: OFF

Command took 0.83 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 11:51:32 AM on My Cluster

Cmd 13

```
1 # joining
2 employees=spark.read.csv("dbfs:/FileStore/tables/employees.csv",header=True,inferSchema=True)
3 employees.show()
```

(3) Spark Jobs

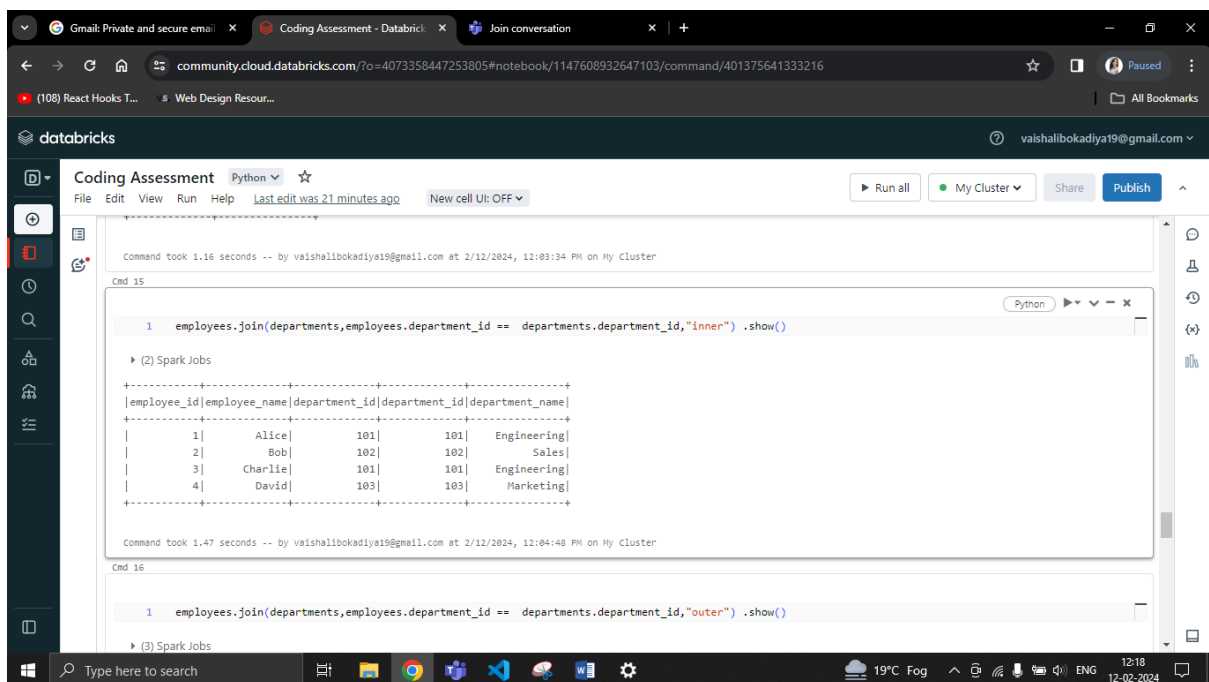
employees: pyspark.sql.dataframe.DataFrame = [employee_id: integer, employee_name: string ... 1 more field]

| employee_id | employee_name | department_id |
|-------------|---------------|---------------|
| 1 | Alice | 101 |
| 2 | Bob | 102 |
| 3 | Charlie | 101 |
| 4 | David | 103 |

Command took 1.20 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:03:10 PM on My Cluster

Cmd 14

```
1 departments=spark.read.csv("dbfs:/FileStore/tables/department.csv",header=True,inferSchema=True)
```



community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333216

databricks vaishalibokadiya19@gmail.com

Coding Assessment Python ☆

File Edit View Run Help Last edit was 21 minutes ago New cell UI: OFF

Command took 1.16 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:03:34 PM on My Cluster

Cmd 15

```
1 employees.join(departments,employees.department_id == departments.department_id,"inner").show()
```

(2) Spark Jobs

| employee_id | employee_name | department_id | department_id | department_name |
|-------------|---------------|---------------|---------------|-----------------|
| 1 | Alice | 101 | 101 | Engineering |
| 2 | Bob | 102 | 102 | Sales |
| 3 | Charlie | 101 | 101 | Engineering |
| 4 | David | 103 | 103 | Marketing |

Command took 1.47 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:04:48 PM on My Cluster

Cmd 16

```
1 employees.join(departments,employees.department_id == departments.department_id,"outer").show()
```

(3) Spark Jobs

community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333216

databricks vaishalibokadiya19@gmail.com

Coding Assessment Python ☆

File Edit View Run Help Last edit was 21 minutes ago New cell UI: OFF

Run all My Cluster Share Publish

Command took 1.47 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:04:48 PM on My Cluster

cmd 16

```
1 employees.join(departments,employees.department_id == departments.department_id,"outer").show()
```

▶ (3) Spark Jobs

| employee_id | employee_name | department_id | department_id | department_name |
|-------------|---------------|---------------|---------------|-----------------|
| 1 | Alice | 101 | 101 | Engineering |
| 3 | Charlie | 101 | 101 | Engineering |
| 2 | Bob | 102 | 102 | Sales |
| 4 | David | 103 | 103 | Marketing |
| NULL | NULL | NULL | 104 | HR |

Command took 1.12 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:04:54 PM on My Cluster

cmd 17

```
1 employees.join(departments,employees.department_id == departments.department_id,"left").show()
```

community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333216

databricks vaishalibokadiya19@gmail.com

Coding Assessment Python ☆

File Edit View Run Help Last edit was 21 minutes ago New cell UI: OFF

Run all My Cluster Share Publish

Command took 1.12 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:04:54 PM on My Cluster

cmd 17

```
1 employees.join(departments,employees.department_id == departments.department_id,"left").show()
```

▶ (2) Spark Jobs

| employee_id | employee_name | department_id | department_id | department_name |
|-------------|---------------|---------------|---------------|-----------------|
| 1 | Alice | 101 | 101 | Engineering |
| 2 | Bob | 102 | 102 | Sales |
| 3 | Charlie | 101 | 101 | Engineering |
| 4 | David | 103 | 103 | Marketing |

Command took 0.68 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:05:00 PM on My Cluster

cmd 18

```
1 employees.join(departments,employees.department_id == departments.department_id,"right").show()
```

▶ (2) Spark Jobs

community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333216

databricks vaishalibokadiya19@gmail.com

Coding Assessment

Python Last edit was 21 minutes ago New cell UI: OFF

Run all My Cluster Share Publish

```
1 employees.join(departments,employees.department_id == departments.department_id,"right").show()
```

Command took 0.68 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:05:00 PM on My Cluster

cmd 18

(2) Spark Jobs

| employee_id | employee_name | department_id | department_id | department_name |
|-------------|---------------|---------------|---------------|-----------------|
| 3 | Charlie | 101 | 101 | Engineering |
| 4 | David | 103 | 103 | Marketing |
| 1 | Alice | 101 | 101 | Engineering |
| 2 | Bob | 102 | 102 | Sales |
| 4 | David | 103 | 103 | Marketing |
| NULL | NULL | NULL | 104 | HR |

Command took 0.84 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:05:06 PM on My Cluster

cmd 19

community.cloud.databricks.com/?o=4073358447253805#notebook/1147608932647103/command/401375641333216

databricks vaishalibokadiya19@gmail.com

Coding Assessment

Python Last edit was 21 minutes ago New cell UI: OFF

Run all My Cluster Share Publish

```
1 employees.join(departments,employees.department_id == departments.department_id,"leftsemi").show()
```

Command took 0.73 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:05:11 PM on My Cluster

cmd 19

(2) Spark Jobs

| employee_id | employee_name | department_id |
|-------------|---------------|---------------|
| 1 | Alice | 101 |
| 2 | Bob | 102 |
| 3 | Charlie | 101 |
| 4 | David | 103 |

Command took 0.73 seconds -- by vaishalibokadiya19@gmail.com at 2/12/2024, 12:05:11 PM on My Cluster

cmd 20

```
1 employees.join(departments,employees.department_id == departments.department_id,"leftanti").show()
```

(2) Spark Jobs

The screenshot displays a Databricks workspace with a notebook titled "Coding Assessment". The notebook is written in Python and contains a table of employee data and a Spark SQL query. The table has columns for employee_id, employee_name, and department_id. The query filters for employees in the 'Sales' department. The interface shows the command history and the execution status of the code cells.

| employee_id | employee_name | department_id |
|-------------|---------------|---------------|
| 1 | Alice | 101 |
| 2 | Bob | 102 |
| 3 | Charlie | 101 |
| 4 | David | 103 |

```

employees.join(departments, employees.department_id == departments.department_id, "leftanti").show()
  
```

(2) Spark Jobs

| employee_id | employee_name | department_id |
|-------------|---------------|---------------|
| 1 | Alice | 101 |
| 2 | Bob | 102 |
| 3 | Charlie | 101 |
| 4 | David | 103 |