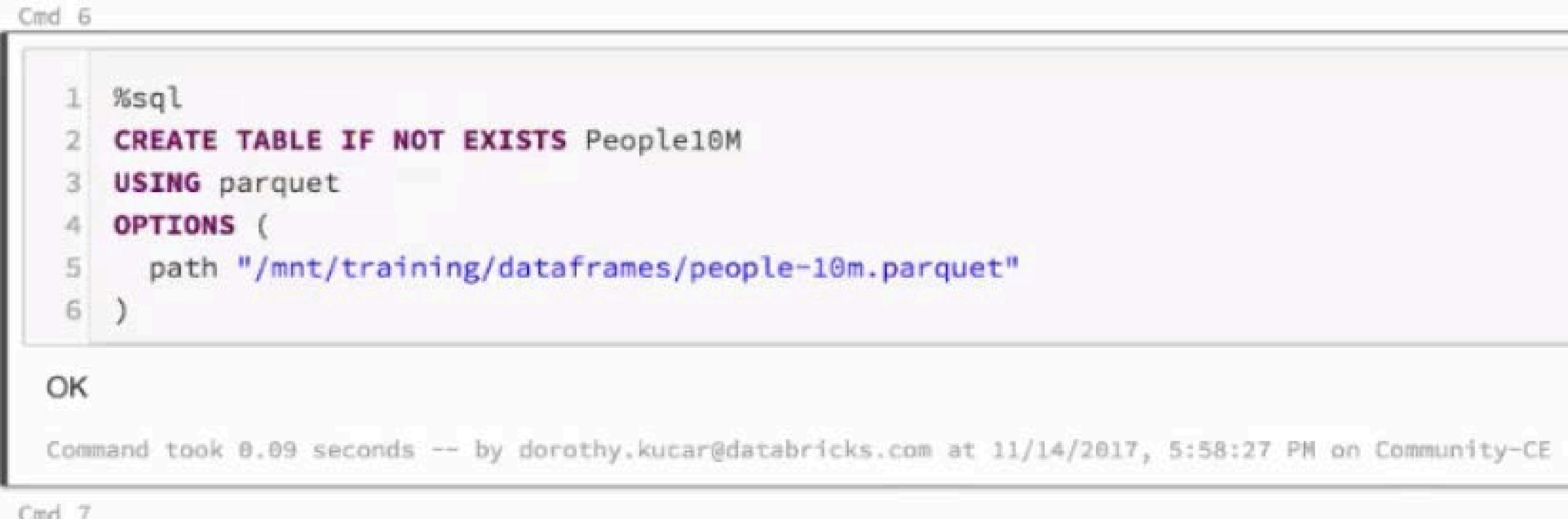
Querying Files with SQL

Apache Spark™ and Databricks® allow you to use SQL to query large data files.

In this lesson you:

- Query large files using Spark SQL
- Visualize query results using charts
- Create temporary views to simplify complex queries



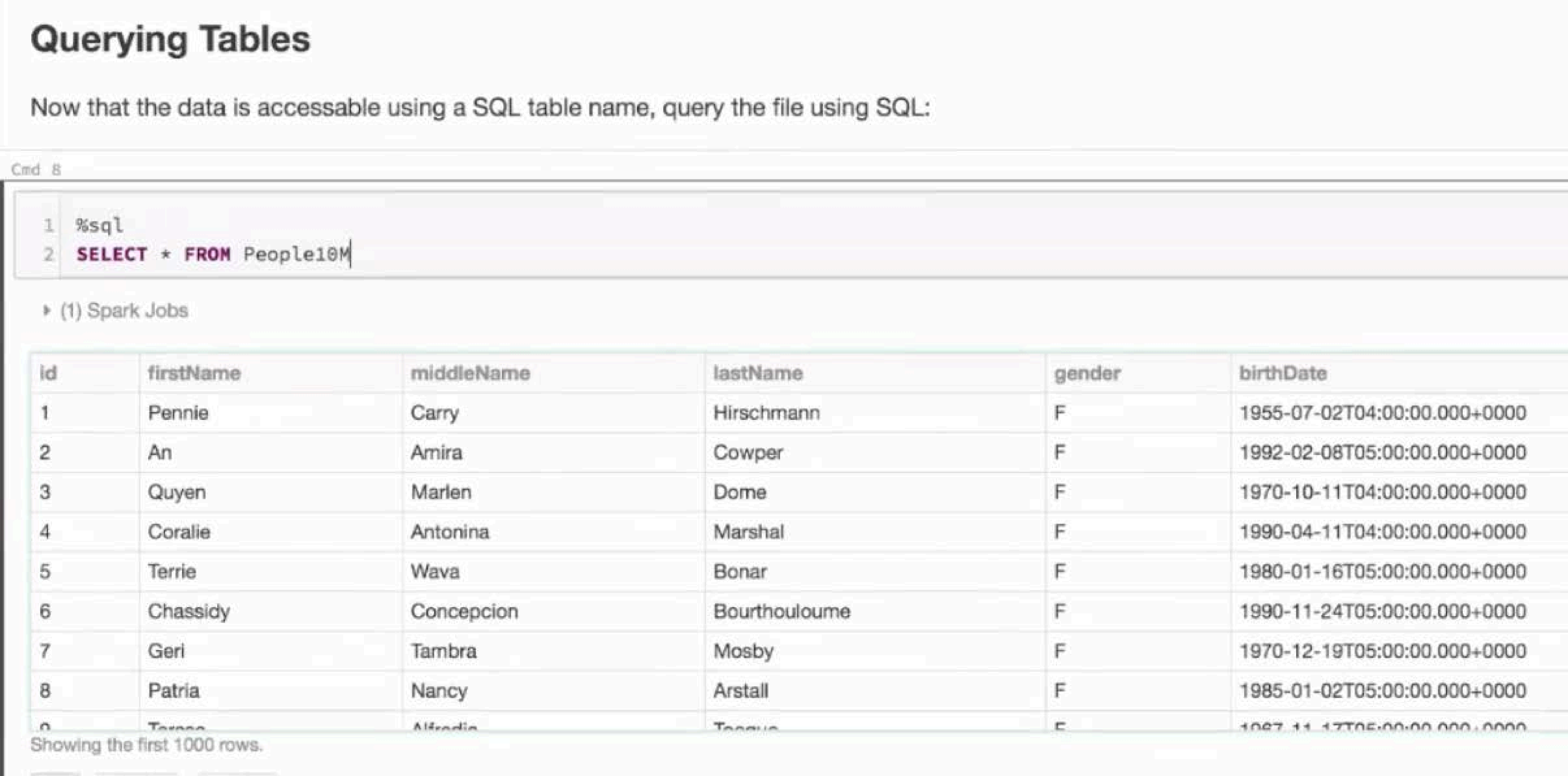
Cmd 6

```
%sql
CREATE TABLE IF NOT EXISTS People10M
USING parquet
OPTIONS (
  path "/mnt/training/dataframes/people-10m.parquet"
)
```

OK

Command took 0.09 seconds -- by dorothy.kucar@databricks.com at 11/14/2017, 5:58:27 PM on Community-CE

Cmd 7



Querying Tables

Now that the data is accessable using a SQL table name, query the file using SQL:

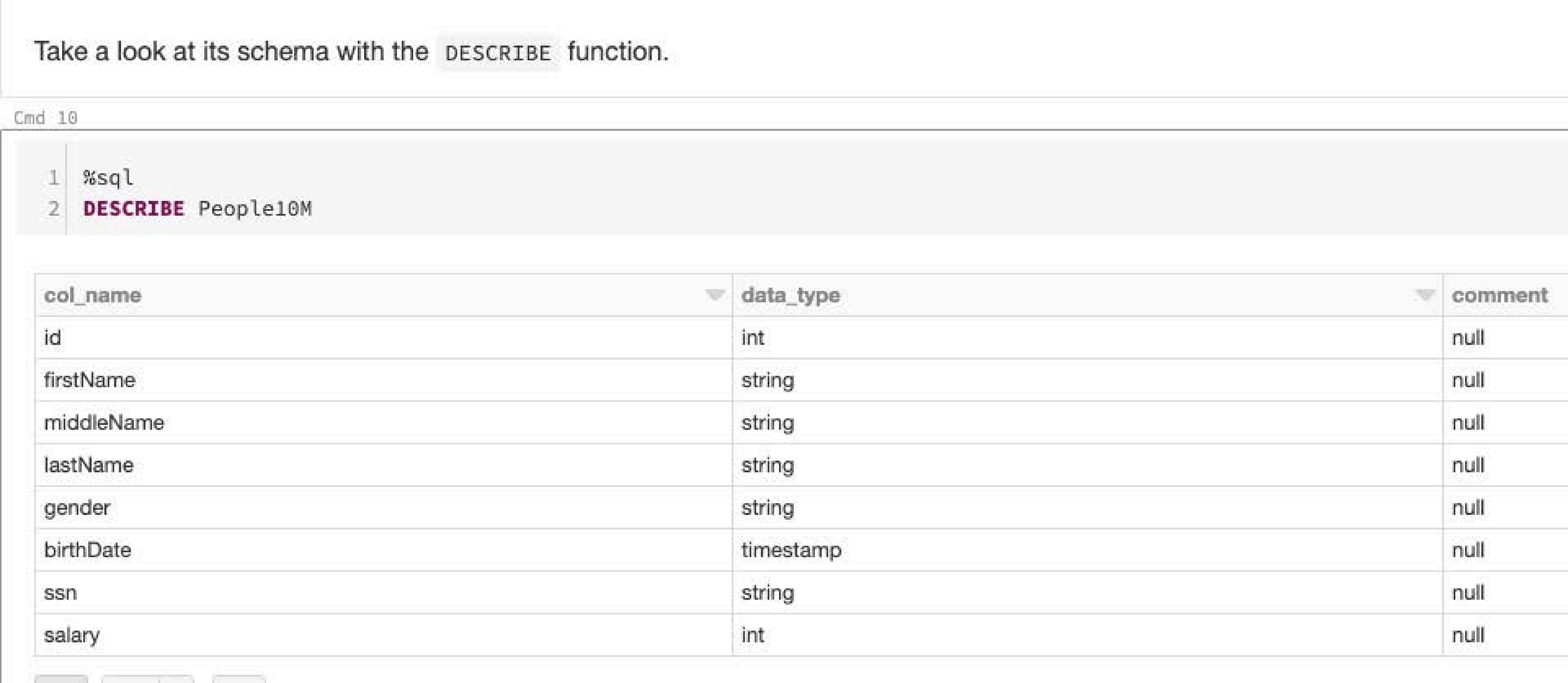
Cmd 8

```
%sql
SELECT * FROM People10M
```

▸ (1) Spark Jobs

id	firstName	middleName	lastName	gender	birthDate
1	Pennie	Carry	Hirschmann	F	1955-07-02T04:00:00.000+0000
2	An	Amira	Cowper	F	1992-02-08T05:00:00.000+0000
3	Quyen	Marlen	Dome	F	1970-10-11T04:00:00.000+0000
4	Coralie	Antonina	Marshal	F	1990-04-11T04:00:00.000+0000
5	Terrie	Wava	Bonar	F	1980-01-16T05:00:00.000+0000
6	Chassidy	Concepcion	Bourthouloume	F	1990-11-24T05:00:00.000+0000
7	Geri	Tambra	Mosby	F	1970-12-19T05:00:00.000+0000
8	Patria	Nancy	Arstall	F	1985-01-02T05:00:00.000+0000
9	Teresa	Alfredia	Teague	F	1967-11-17T05:00:00.000+0000

Showing the first 1000 rows.

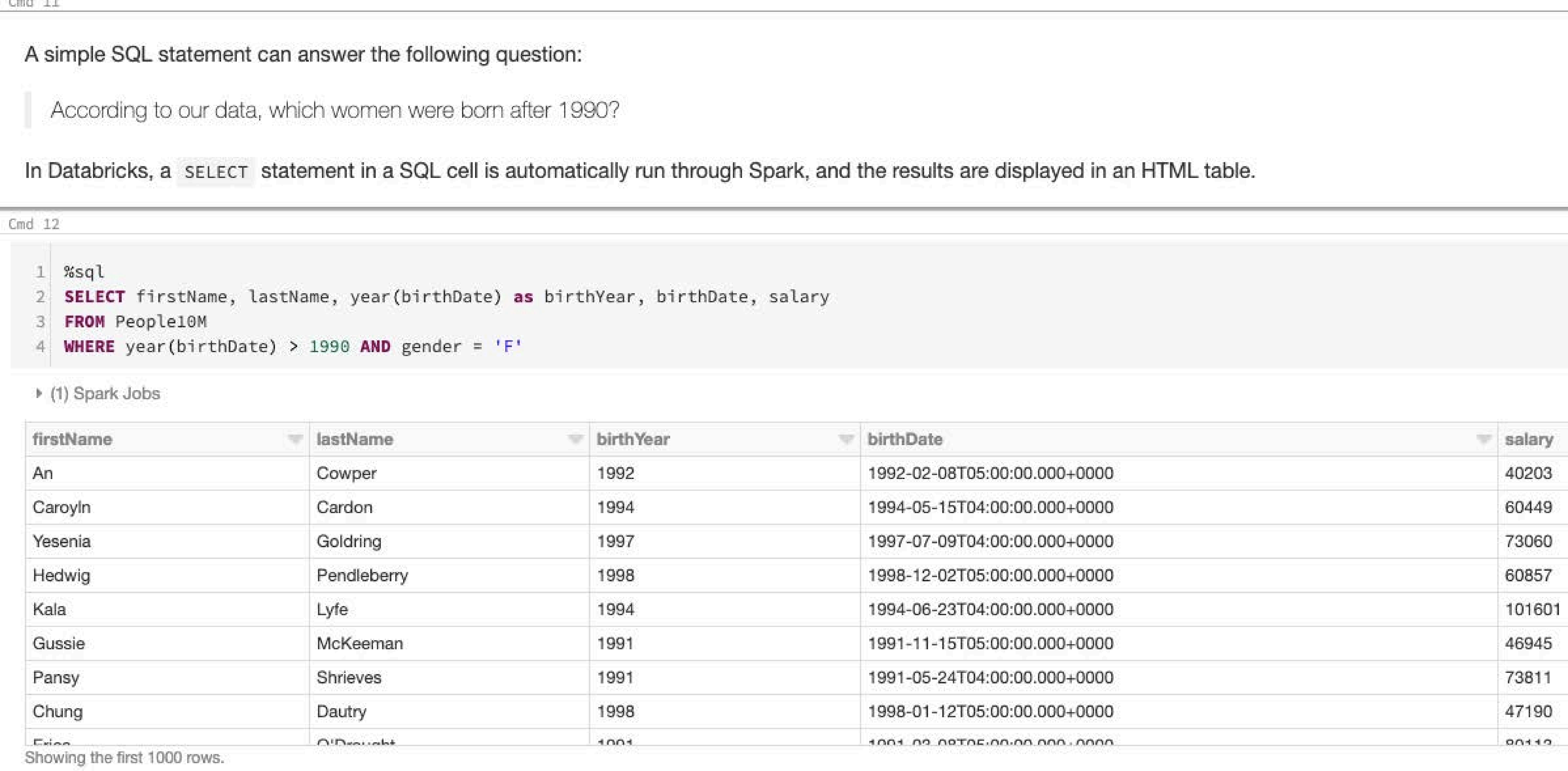


Take a look at its schema with the `DESCRIBE` function.

Cmd 10

```
%sql
DESCRIBE People10M
```

col_name	data_type	comment
id	int	null
firstName	string	null
middleName	string	null
lastName	string	null
gender	string	null
birthDate	timestamp	null
ssn	string	null
salary	int	null



Cmd 11

A simple SQL statement can answer the following question:

> According to our data, which women were born after 1990?

In Databricks, a `SELECT` statement in a SQL cell is automatically run through Spark, and the results are displayed in an HTML table.

Cmd 12

```
%sql
SELECT firstName, lastName, year(birthDate) as birthYear, birthDate, salary
FROM People10M
WHERE year(birthDate) > 1990 AND gender = 'F'
```

▸ (1) Spark Jobs

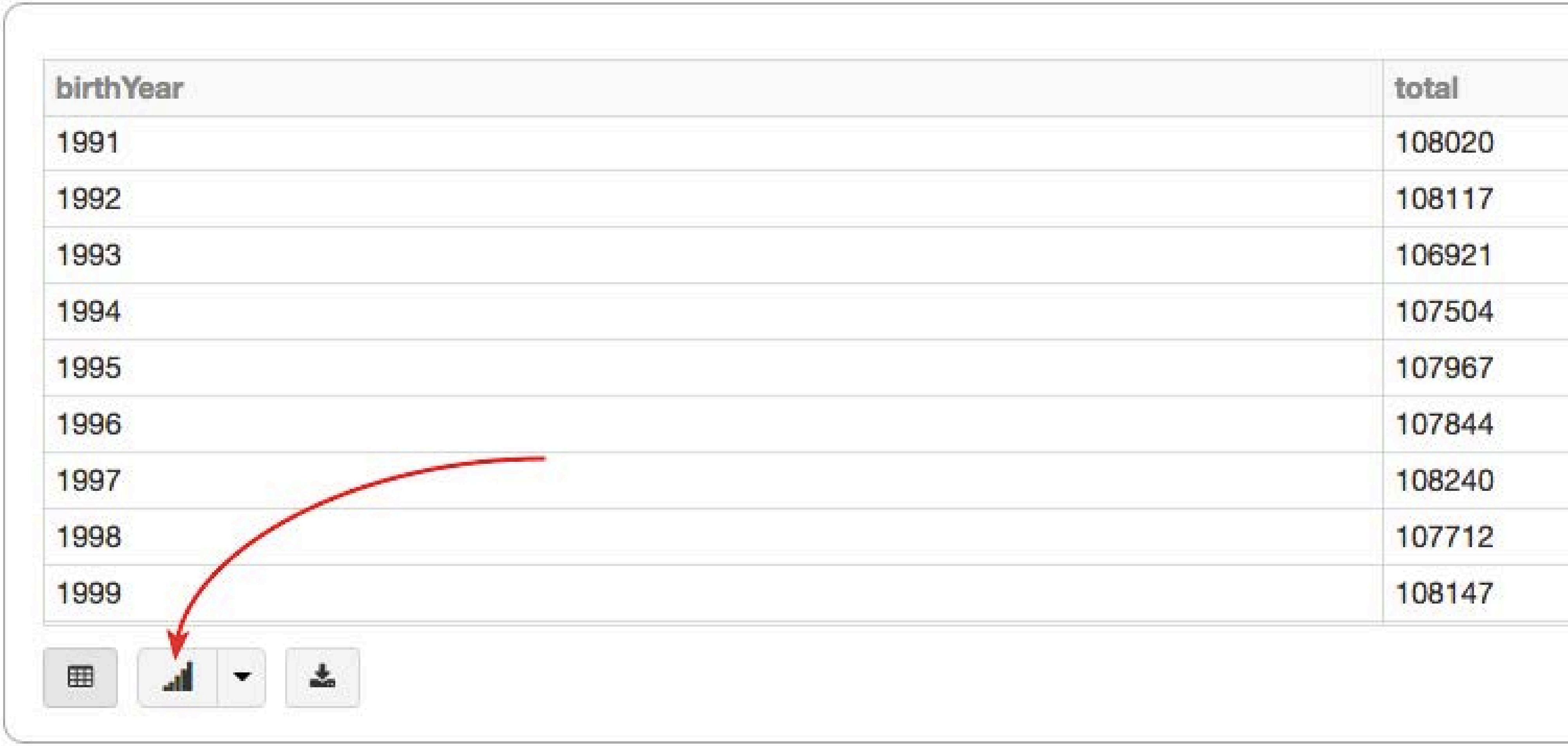
firstName	lastName	birthYear	birthDate	salary
An	Cowper	1992	1992-02-08T05:00:00.000+0000	40203
Caroyln	Cardon	1994	1994-05-15T04:00:00.000+0000	60449
Yesenia	Goldring	1997	1997-07-09T04:00:00.000+0000	73060
Hedwig	Pendleberry	1998	1998-12-02T05:00:00.000+0000	60857
Kala	Lyfe	1994	1994-06-23T04:00:00.000+0000	101601
Gussie	McKeeman	1991	1991-11-15T05:00:00.000+0000	46945
Pansy	Shrieves	1991	1991-05-24T04:00:00.000+0000	73811
Chung	Dautry	1998	1998-01-12T05:00:00.000+0000	47190

Showing the first 1000 rows.

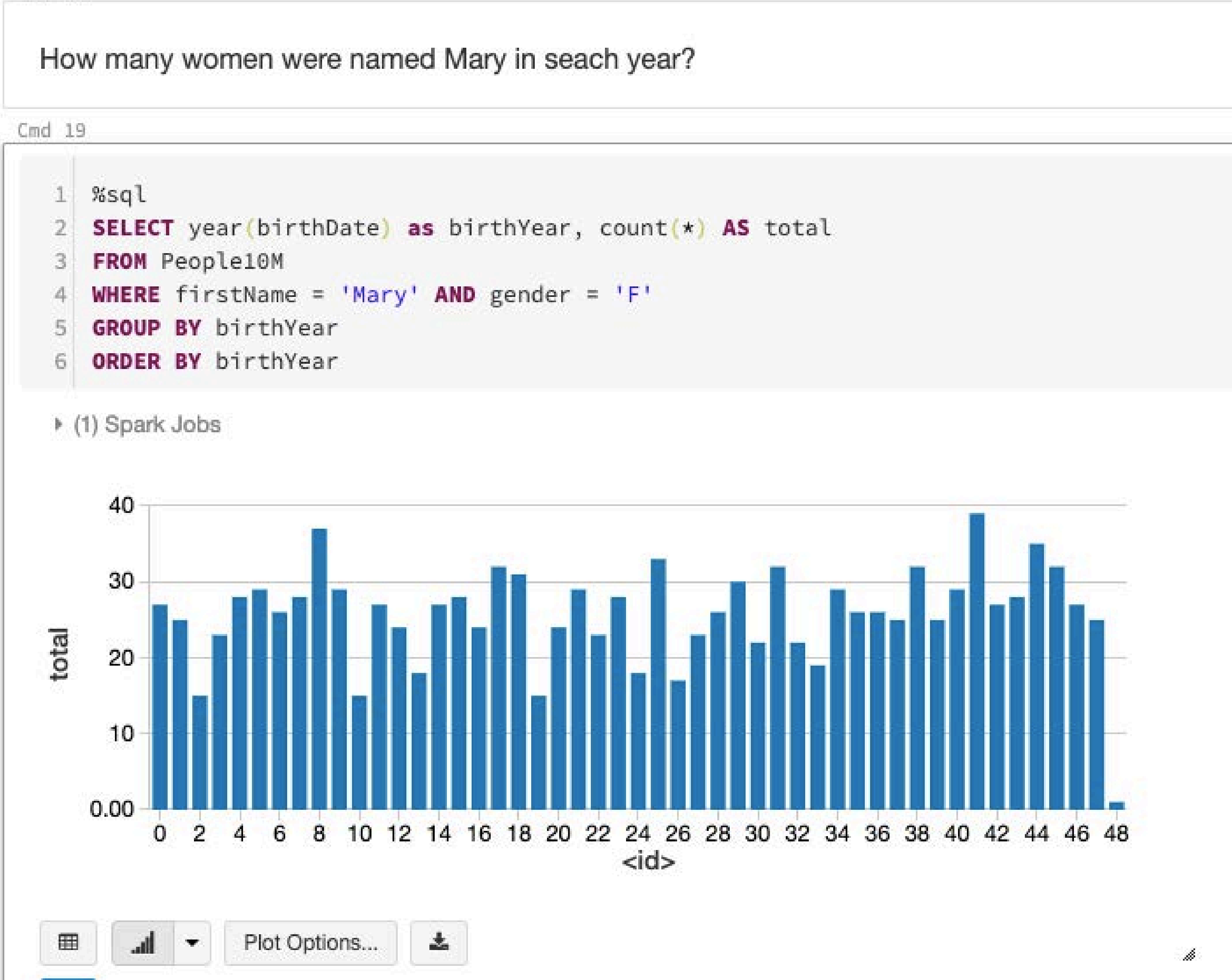
Visualization

Databricks provides built-in easy to use visualizations for your data.

Take the query below, and visualize it by selecting the bar graph icon once the table is displayed:



birthYear	total
1991	108020
1992	108117
1993	106921
1994	107504
1995	107967
1996	107844
1997	108240
1998	107712
1999	108147

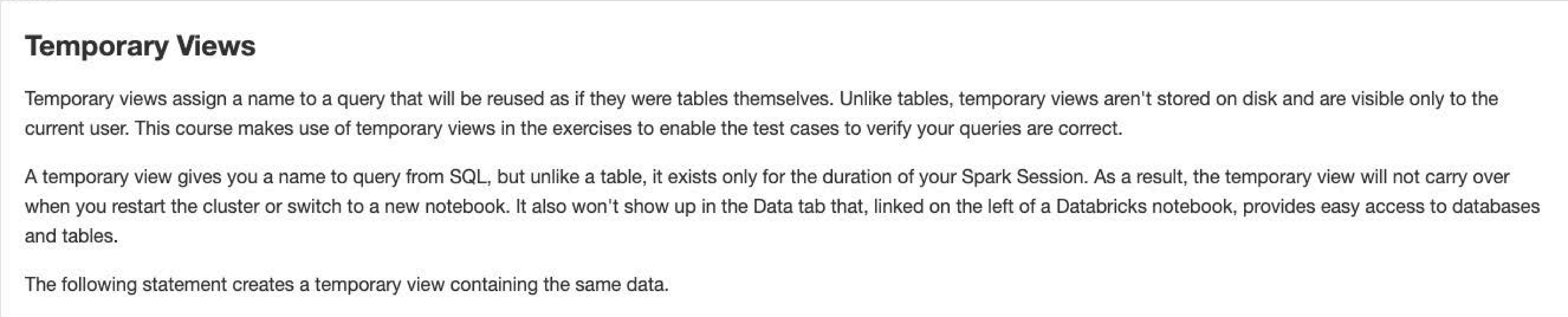


How many women were named Mary in seach year?

Cmd 19

```
%sql
SELECT year(birthDate) as birthYear, count(*) AS total
FROM People10M
WHERE firstName = 'Mary' AND gender = 'F'
GROUP BY birthYear
ORDER BY birthYear
```

▸ (1) Spark Jobs



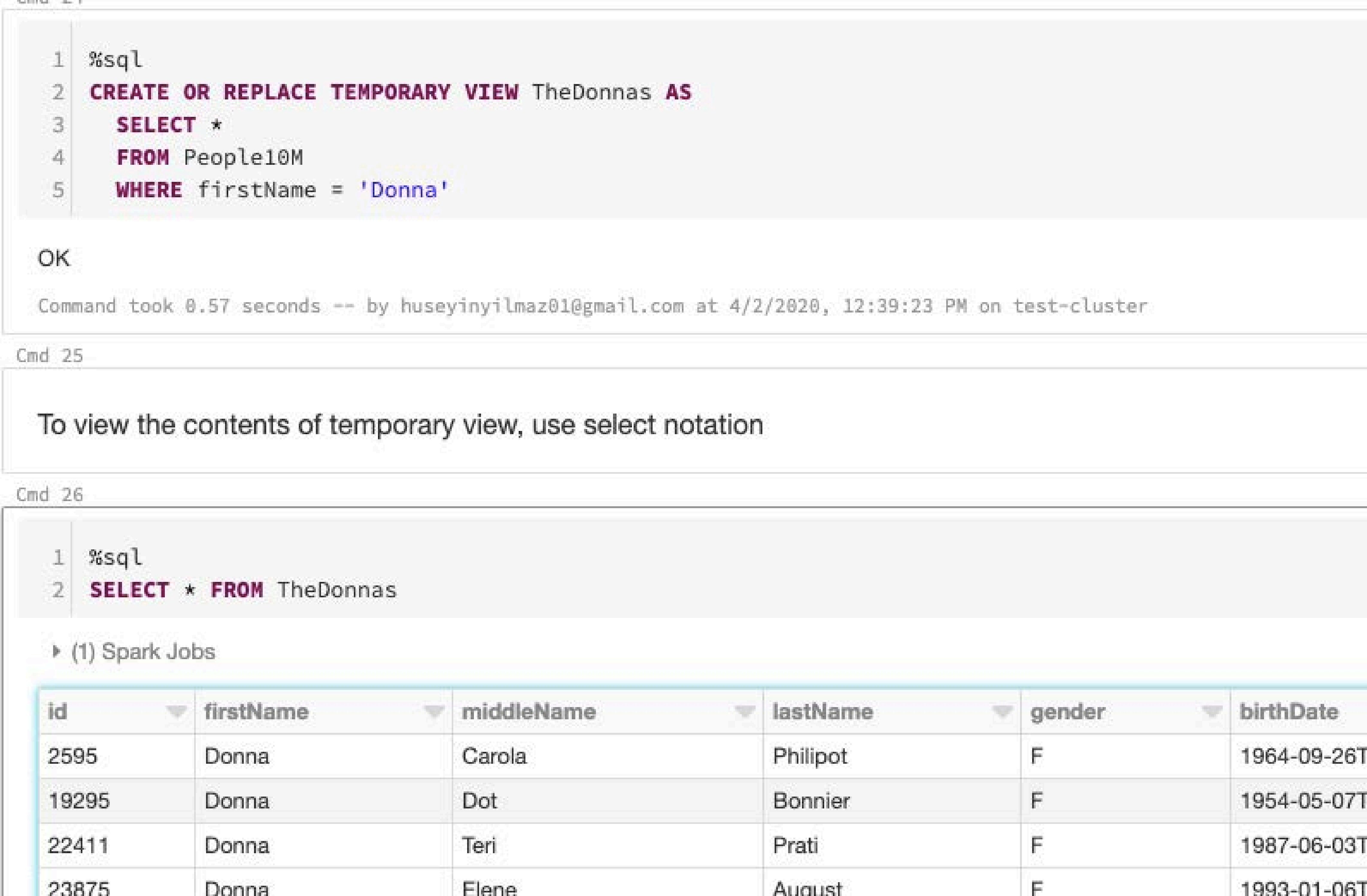
Cmd 22

Temporary Views

Temporary views assign a name to a query that will be reused as if they were tables themselves. Unlike tables, temporary views aren't stored on disk and are visible only to the current user. This course makes use of temporary views in the exercises to enable the test cases to verify your queries are correct.

A temporary view gives you a name to query from SQL, but unlike a table, it exists only for the duration of your Spark Session. As a result, the temporary view will not carry over when you restart the cluster or switch to a new notebook. It also won't show up in the Data tab that, linked on the left of a Databricks notebook, provides easy access to databases and tables.

The following statement creates a temporary view containing the same data.



Cmd 24

```
%sql
CREATE OR REPLACE TEMPORARY VIEW TheDonnas AS
  SELECT *
  FROM People10M
  WHERE firstName = 'Donna'
```

OK

Command took 0.57 seconds -- by huseyinyilmaz01@gmail.com at 4/2/2020, 12:39:23 PM on test-cluster

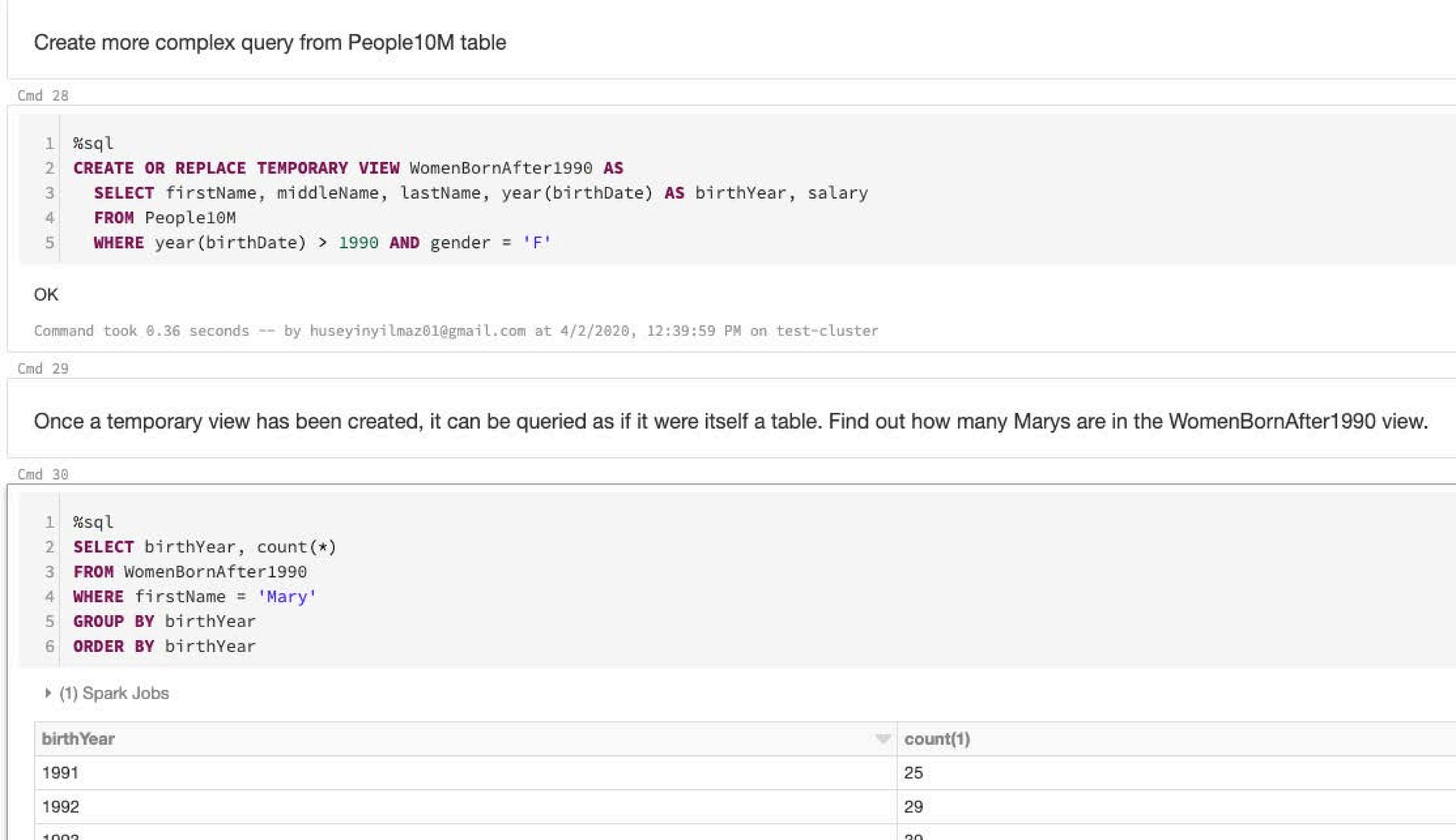
Cmd 25

To view the contents of temporary view, use select notation

Cmd 26

```
%sql
SELECT * FROM TheDonnas
```

id	firstName	middleName	lastName	gender	birthDate
2595	Donna	Carola	Philipot	F	1964-09-26T
19295	Donna	Dot	Bonnier	F	1954-05-07T
22411	Donna	Teri	Prati	F	1987-06-03T
23875	Donna	Elene	August	F	1993-01-06T



Create more complex query from People10M table

Cmd 28

```
%sql
CREATE OR REPLACE TEMPORARY VIEW WomenBornAfter1990 AS
  SELECT firstName, middleName, lastName, year(birthDate) AS birthYear, salary
  FROM People10M
  WHERE year(birthDate) > 1990 AND gender = 'F'
```

OK

Command took 0.36 seconds -- by huseyinyilmaz01@gmail.com at 4/2/2020, 12:39:59 PM on test-cluster

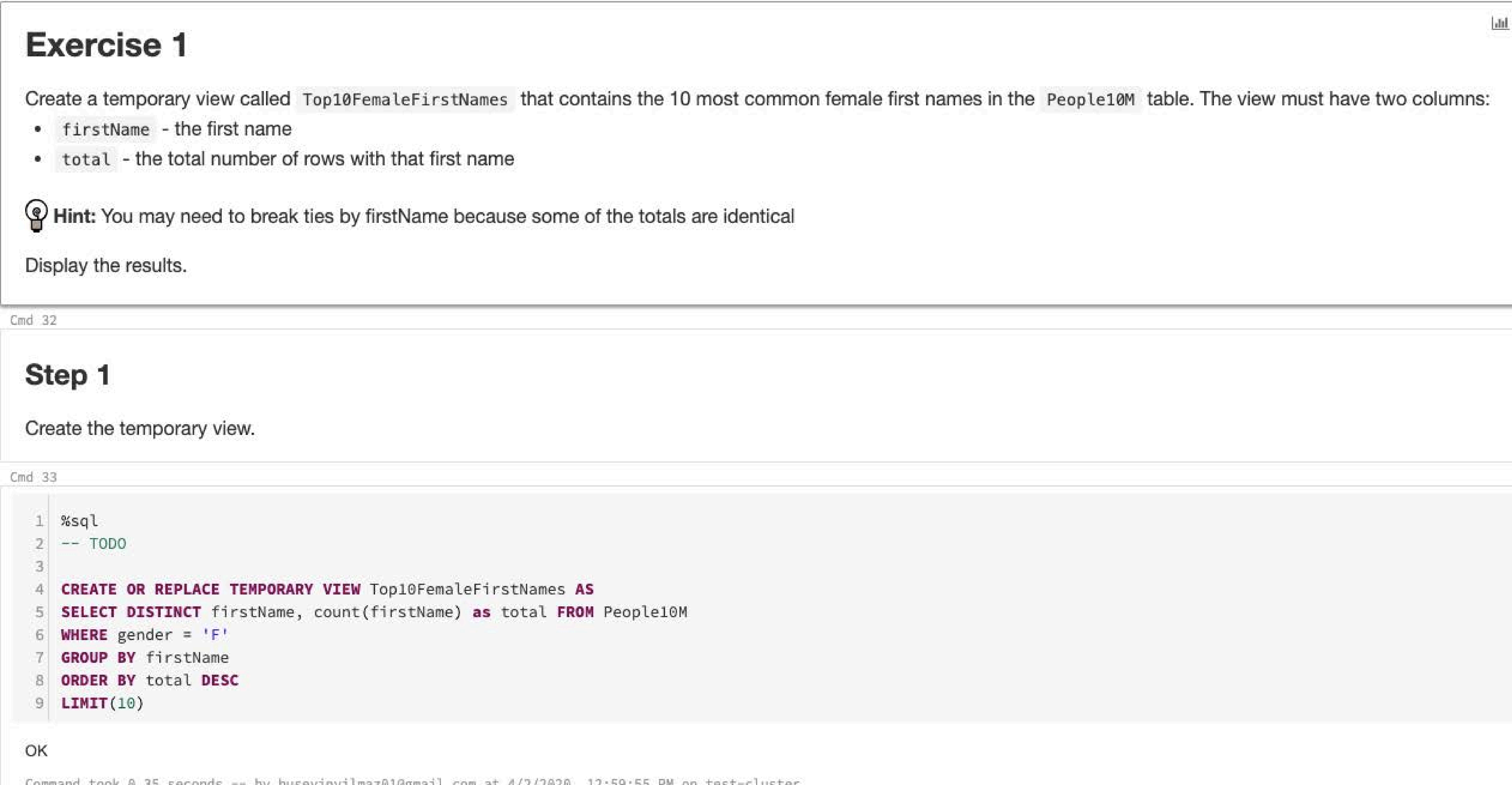
Cmd 29

Once a temporary view has been created, it can be queried as if it were itself a table. Find out how many Marys are in the WomenBornAfter1990 view.

Cmd 30

```
%sql
SELECT birthYear, count(*)
FROM WomenBornAfter1990
WHERE firstName = 'Mary'
GROUP BY birthYear
ORDER BY birthYear
```

birthYear	count(1)
1991	25
1992	29
1993	30



Exercise 1

Create a temporary view called `Top10FemaleFirstNames` that contains the 10 most common female first names in the `People10M` table. The view must have two columns:

- `firstName` - the first name
- `total` - the total number of rows with that first name

Hint: You may need to break ties by firstName because some of the totals are identical

Display the results.

Cmd 32

Step 1

Create the temporary view.

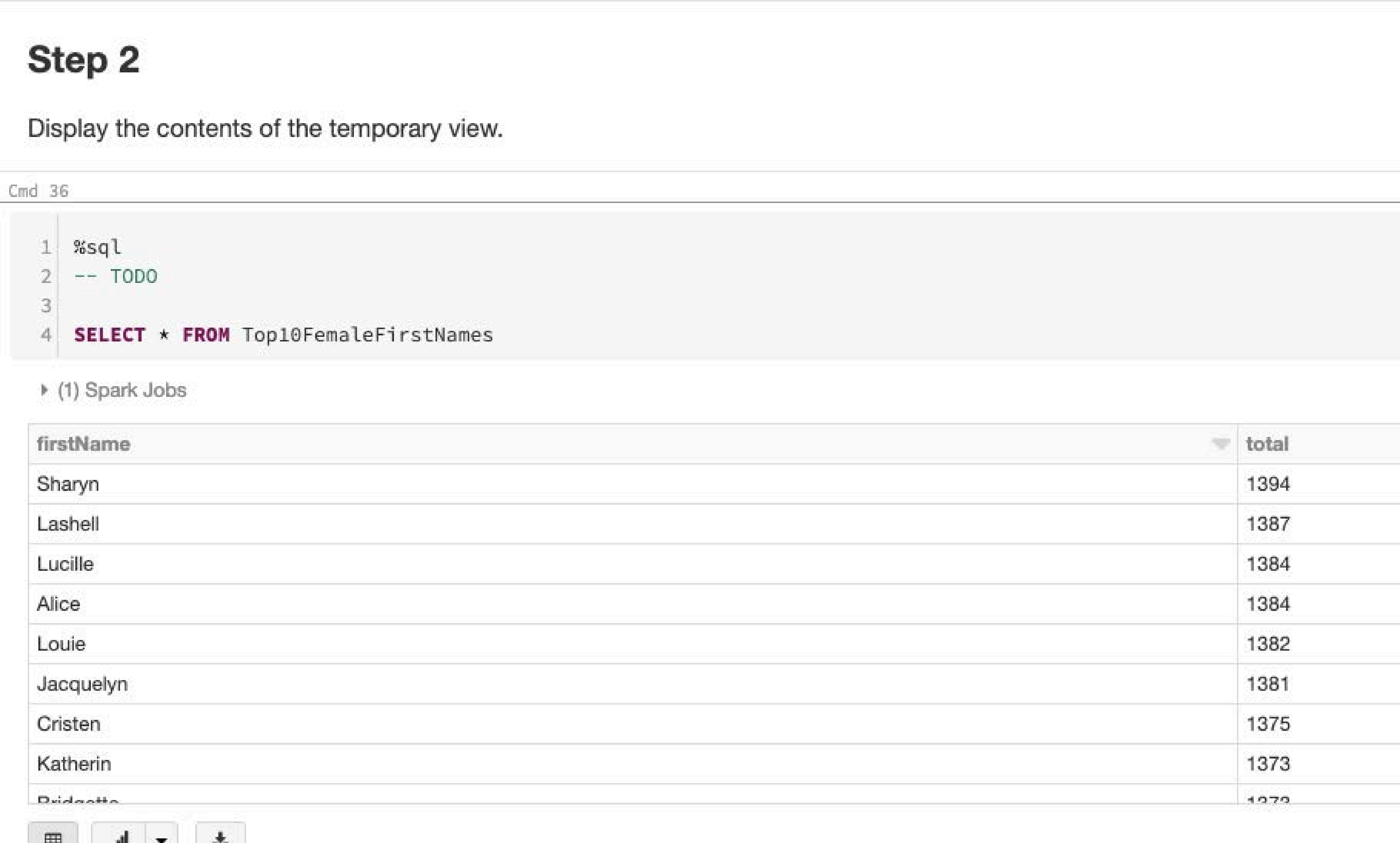
Cmd 33

```
%sql
-- TODO

CREATE OR REPLACE TEMPORARY VIEW Top10FemaleFirstNames AS
SELECT DISTINCT firstName, count(firstName) as total FROM People10M
WHERE gender = 'F'
GROUP BY firstName
ORDER BY total DESC
LIMIT(10)
```

OK

Command took 0.35 seconds -- by huseyinyilmaz01@gmail.com at 4/2/2020, 12:59:55 PM on test-cluster



Step 2

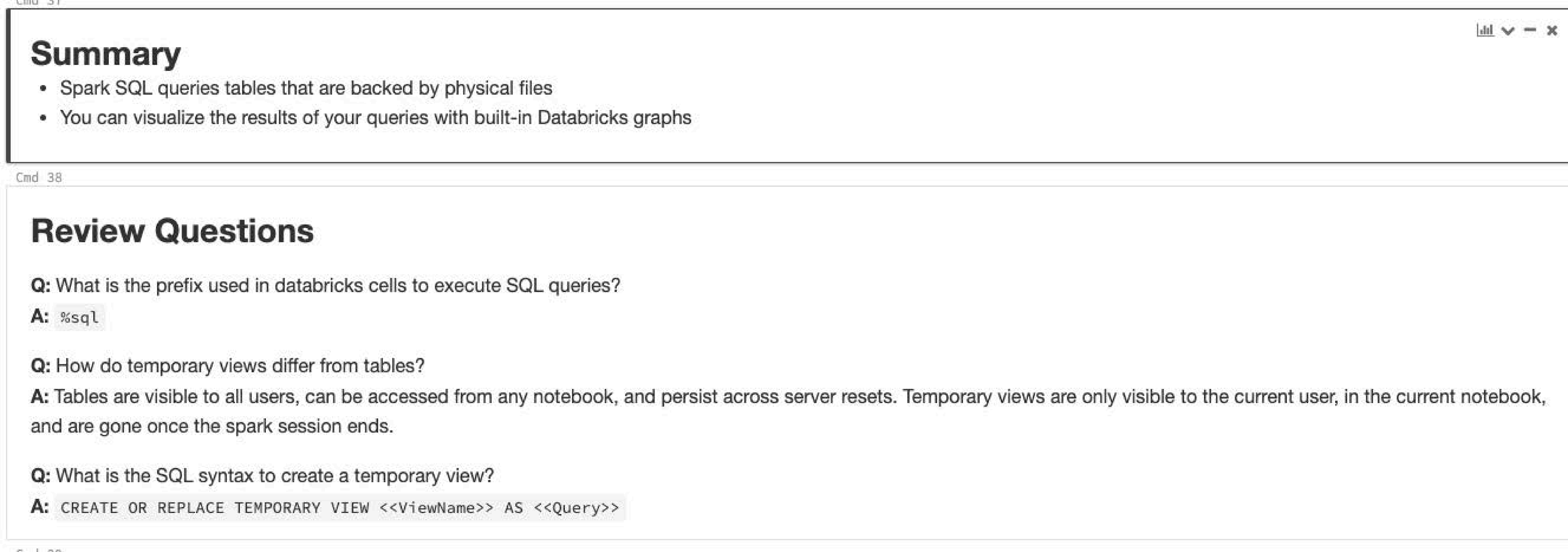
Display the contents of the temporary view.

Cmd 36

```
%sql
-- TODO

SELECT * FROM Top10FemaleFirstNames
```

firstName	total
Sharyn	1394
Lashell	1387
Lucille	1384
Alice	1384
Louie	1382
Jacquelyn	1381
Cristen	1375
Katherin	1373
Bridgette	1373



Summary

- Spark SQL queries tables that are backed by physical files
- You can visualize the results of your queries with built-in Databricks graphs

Cmd 38

Review Questions

Q: What is the prefix used in databricks cells to execute SQL queries?

A: `%sql`

Q: How do temporary views differ from tables?

A: Tables are visible to all users, can be accessed from any notebook, and persist across server resets. Temporary views are only visible to the current user, in the current notebook, and are gone once the spark session ends.

Q: What is the SQL syntax to create a temporary view?

A: `CREATE OR REPLACE TEMPORARY VIEW <<ViewName>> AS <<Query>>`