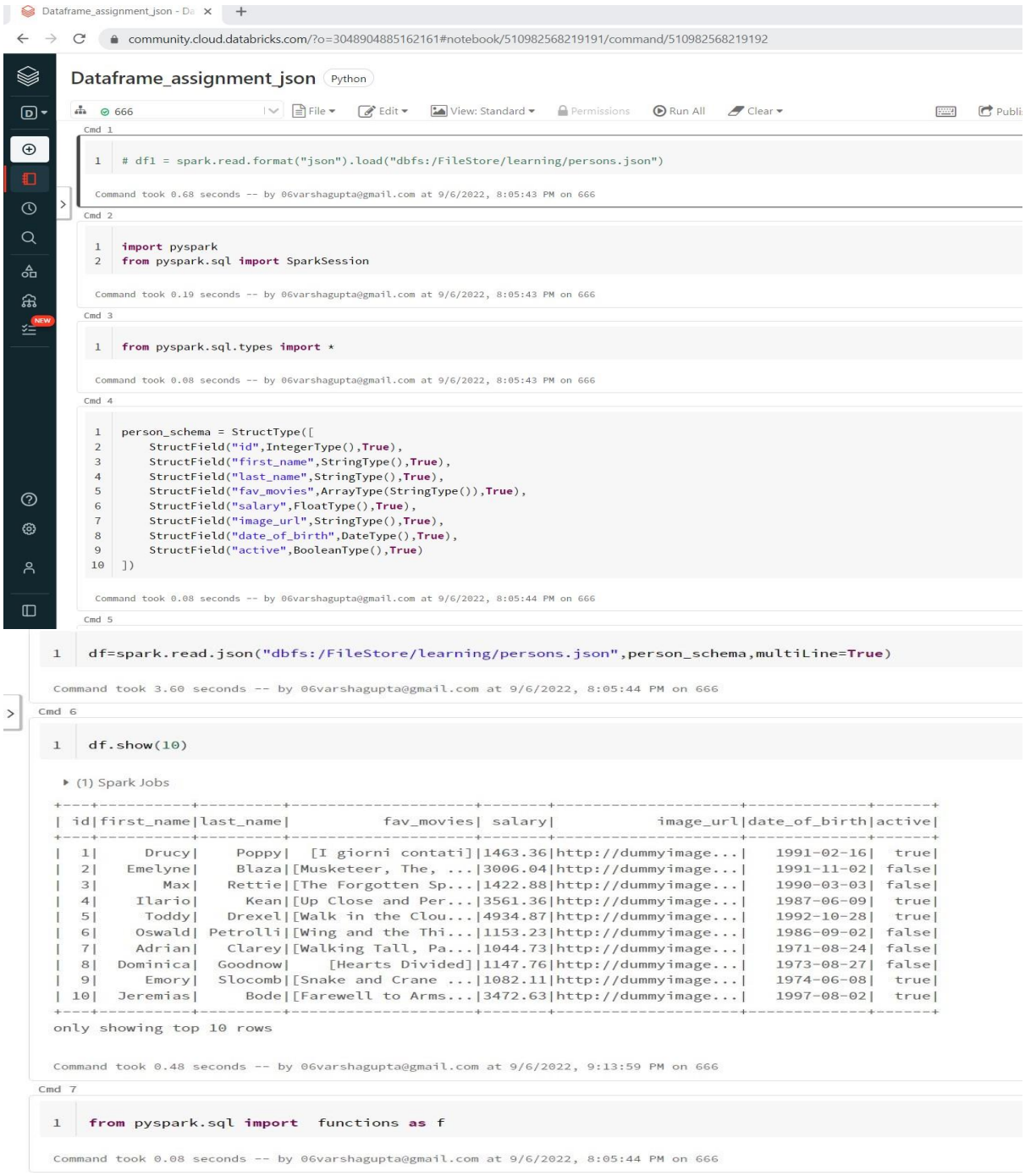
Task-20

Name: Varsha Gupta

Referral id: DIRSS3129

Q1. You need to use the person json file

- You need to create three column as year , month, quarter of the year
- All the three column should be extracted from the date_of_birth columns



```
1 from pyspark.sql.functions import col
```

Command took 0.07 seconds -- by 06varshagupta@gmail.com at 9/6/2022, 8:05:44 PM on 666

Cmd 9

Q1. You need to use the person json file

- You need to create three column as year , month, quarter of the year
- All the three column should be extracted from the date_of_birth columns

Cmd 10

```
1 df.select( col("date_of_birth"), f.year( col("date_of_birth") ), f.month( col("date_of_birth") ), f.quarter( col("date_of_birth") ) ).show(10)
```

▸ (1) Spark Jobs

```
+-------------+------------------+-------------------+----------------------+
|date_of_birth|year(date_of_birth)|month(date_of_birth)|quarter(date_of_birth)|
+-------------+------------------+-------------------+----------------------+
|   1991-02-16|              1991|                  2|                     1|
|   1991-11-02|              1991|                 11|                     4|
|   1990-03-03|              1990|                  3|                     1|
|   1987-06-09|              1987|                  6|                     2|
|   1992-10-28|              1992|                 10|                     4|
|   1986-09-02|              1986|                  9|                     3|
|   1971-08-24|              1971|                  8|                     3|
|   1973-08-27|              1973|                  8|                     3|
|   1974-06-08|              1974|                  6|                     2|
|   1997-08-02|              1997|                  8|                     3|
+-------------+------------------+-------------------+----------------------+
only showing top 10 rows
```

only showing top 10 rows

Command took 0.56 seconds -- by 06varshagupta@gmail.com at 9/6/2022, 9:14:51 PM on 666

Cmd 11

```
1 # Get year from date_of_birth in pyspark
2
3 from pyspark.sql.functions import year
4 from pyspark.sql.functions import to_date
5
6 df2 = df.withColumn('year', year(df.date_of_birth))
7 df2.show(10)
```

▸ (1) Spark Jobs

```
+---+----------+---------+--------------------+-------+--------------------+-------------+------+----+
| id|first_name|last_name|           fav_movies| salary|           image_url|date_of_birth|active|year|
+---+----------+---------+--------------------+-------+--------------------+-------------+------+----+
|  1|     Drucy|    Poppy|  [I giorni contati]|1463.36|http://dummyimage...|   1991-02-16|  true|1991|
|  2|   Emelyne|    Blaza|[Musketeer, The, ...|3006.04|http://dummyimage...|   1991-11-02| false|1991|
|  3|       Max|   Rettie|[The Forgotten Sp...|1422.88|http://dummyimage...|   1990-03-03| false|1990|
|  4|    Ilario|     Kean|[Up Close and Per...|3561.36|http://dummyimage...|   1987-06-09|  true|1987|
|  5|     Toddy|   Drexel|[Walk in the Clou...|4934.87|http://dummyimage...|   1992-10-28|  true|1992|
|  6|    Oswald| Petrolli|[Wing and the Thi...|1153.23|http://dummyimage...|   1986-09-02| false|1986|
|  7|    Adrian|    Clarey|[Walking Tall, Pa...|1044.73|http://dummyimage...|   1971-08-24| false|1971|
|  8|  Dominica|  Goodnow|    [Hearts Divided]|1147.76|http://dummyimage...|   1973-08-27| false|1973|
|  9|     Emory|  Slocomb|[Snake and Crane ...|1082.11|http://dummyimage...|   1974-06-08|  true|1974|
| 10|  Jeremias|     Bode|[Farewell to Arms...|3472.63|http://dummyimage...|   1997-08-02|  true|1997|
+---+----------+---------+--------------------+-------+--------------------+-------------+------+----+
only showing top 10 rows
```

only showing top 10 rows

Command took 1.66 seconds -- by 06varshagupta@gmail.com at 9/6/2022, 8:05:44 PM on 666

- check if the person is having the year of birth from 1990 to 1995 give 25% hike in salary
- if person year is more than 1995 give 50% hike in salary

Cmd 12

- check if the person is having the year of birth from 1990 to 1995 give 25% hike in salary

Cmd 13

```
y= df2.where(((col("year"))>1990) & ((col("year"))<=1995) )
y.show(10)
```

▸ (1) Spark Jobs

```
+---+----------+---------+--------------------+-------+--------------------+-------------+------+----+
| id|first_name|last_name|          fav_movies| salary|           image_url|date_of_birth|active|year|
+---+----------+---------+--------------------+-------+--------------------+-------------+------+----+
|  1|     Drucy|    Poppy|  [I giorni contati]|1463.36|http://dummyimage...|   1991-02-16|  true|1991|
|  2|   Emelyne|    Blaza|[Musketeer, The, ...|3006.04|http://dummyimage...|   1991-11-02| false|1991|
|  5|    Toddy|   Drexel|[Walk in the Clou...|4934.87|http://dummyimage...|   1992-10-28|  true|1992|
| 23|       Min|   Latter|[Kurt Cobain: Mon...|1909.87|http://dummyimage...|   1991-03-17| false|1991|
| 28|   Frankie|Copestick|[Computer Wore Te...|1186.68|http://dummyimage...|   1994-03-09| false|1994|
| 29|       Eli|Normabell|[Return to Peyton...|4917.48|http://dummyimage...|   1993-01-02|  true|1993|
| 35|     Fanni|    Dyson|     [Wild One, The]|3228.58|http://dummyimage...|   1995-01-09|  true|1995|
| 36|     Ozzie| Brownlie|     [Orange County]|3945.18|http://dummyimage...|   1992-01-25| false|1992|
| 38|    Camile|     Mace|[Family Guy Prese...|3559.93|http://dummyimage...|   1994-12-12| false|1994|
| 46|    Giffie|  Kemmish|[State Witness, T...|4543.29|http://dummyimage...|   1995-06-25| false|1995|
+---+----------+---------+--------------------+-------+--------------------+-------------+------+----+
only showing top 10 rows
```

Command took 0.34 seconds -- by 06varshagupta@gmail.com at 9/6/2022, 8:54:17 PM on 666

Cmd 14

```
(y.select(col("salary"),col("year"),
          col("salary")*.25
         )
).show(10)
```

▸ (1) Spark Jobs

```
+-------+----+------------------+
| salary|year|   (salary * 0.25)|
+-------+----+------------------+
|1463.36|1991| 365.8399963378906|
|3006.04|1991|  751.510009765625|
|4934.87|1992| 1233.717529296875|
|1909.87|1991| 477.4674987792969|
|1186.68|1994| 296.6700134277344|
|4917.48|1993|1229.3699951171875|
|3228.58|1995|   807.14501953125|
|3945.18|1992| 986.2949829101562|
|3559.93|1994| 889.9824829101562|
|4543.29|1995| 1135.822509765625|
+-------+----+------------------+
only showing top 10 rows
```

Command took 0.30 seconds -- by 06varshagupta@gmail.com at 9/6/2022, 8:54:33 PM on 666

Cmd 15

- if person year is more than 1995 give 50% hike in salary

Cmd 16

```
1  x = df2.where((col("year"))>1995)
2  x.show(10)
```

▸ (1) Spark Jobs

```
+---+----------+-----------+--------------------+-------+--------------------+-------------+------+----+
| id|first_name|  last_name|          fav_movies| salary|           image_url|date_of_birth|active|year|
+---+----------+-----------+--------------------+-------+--------------------+-------------+------+----+
| 10|  Jeremias|       Bode|[Farewell to Arms...|3472.63|http://dummyimage...|   1997-08-02|  true|1997|
| 13|  Claiborn|     Denham|[McCullin, Max Pa...|2623.33|http://dummyimage...|   1996-03-07| false|1996|
| 15|    Feodor|Nancekivell|   [Monsoon Wedding]|2218.46|http://dummyimage...|   2000-10-07| false|2000|
| 19|      Lura|     Follis|[My Life in Pink ...|3331.26|http://dummyimage...|   1998-11-03| false|1998|
| 21|      Dian|      Dancy|[Double, Double, ...| 3720.3|http://dummyimage...|   1998-12-01|  true|1998|
| 22|  Theodore|   Climance|[Story of the Wee...|3008.56|http://dummyimage...|   1999-01-30| false|1999|
| 24|  Camellia|   Jervoise| [Waiting to Exhale]|3431.17|http://dummyimage...|   1996-07-14| false|1996|
| 25|     Kelcy|     Wogdon|    [Iron Mask, The]|4512.51|http://dummyimage...|   2000-10-20|  true|2000|
| 32|      Redd|   Akenhead|[Century of the D...| 2470.9|http://dummyimage...|   2000-06-05| false|2000|
| 37|    Carlen|    Sharply|[Dr. Jekyll and M...|2051.85|http://dummyimage...|   2002-06-01|  true|2002|
+---+----------+-----------+--------------------+-------+--------------------+-------------+------+----+
only showing top 10 rows

Command took 0.61 seconds -- by 06varshagupta@gmail.com at 9/6/2022, 9:10:12 PM on 666
```

Cmd 17

```
1  (x.select(col("salary"),col("year"),
2            col("salary")*.5
3           )
4  ).show(10)
```

▸ (1) Spark Jobs

```
+-------+----+------------------+
| salary|year|    (salary * 0.5)|
+-------+----+------------------+
|3472.63|1997|  1736.31494140625|
|2623.33|1996|   1311.6650390625|
|2218.46|2000|  1109.22998046875|
|3331.26|1998|1665.6300048828125|
| 3720.3|1998|1860.1500244140625|
|3008.56|1999| 1504.280029296875|
|3431.17|1996|   1715.5849609375|
|4512.51|2000|   2256.2548828125|
| 2470.9|2000| 1235.449951171875|
|2051.85|2002| 1025.925048828125|
+-------+----+------------------+
only showing top 10 rows

Command took 0.56 seconds -- by 06varshagupta@gmail.com at 9/6/2022, 9:12:26 PM on 666
```

Q2. Get the details of the person in such a way
- it should give you only 1 person detail per year

(Hint : Duplicate remove function)

Cmd 18

Q2. Get the details of the person in such a way - it should give you only 1 person detail per year

Cmd 19

```
1  df2.dropDuplicates(["year"]).show(10)
```

▸ (2) Spark Jobs

```
+---+----------+----------+--------------------+-------+--------------------+-------------+------+----+
| id|first_name| last_name|          fav_movies| salary|           image_url|date_of_birth|active|year|
+---+----------+----------+--------------------+-------+--------------------+-------------+------+----+
|  7|    Adrian|    Clarey|[Walking Tall, Pa...|1044.73|http://dummyimage...|   1971-08-24| false|1971|
| 33|  Sherline|   Primett|   [Jungle Fighters]|2309.39|http://dummyimage...|   1972-07-23|  true|1972|
|  8|  Dominica|   Goodnow|    [Hearts Divided]|1147.76|http://dummyimage...|   1973-08-27| false|1973|
|  9|     Emory|   Slocomb|[Snake and Crane ...|1082.11|http://dummyimage...|   1974-06-08|  true|1974|
| 93|    Janean|      Pelz|              [Once]|4906.92|http://dummyimage...|   1975-09-23|  true|1975|
| 53| Franciska|   Gilford|[Cathedral, The (...|4406.27|http://dummyimage...|   1976-04-08| false|1976|
| 54|     Johny|Goodenough| [Maximum Overdrive]|3592.31|http://dummyimage...|   1977-08-18| false|1977|
| 43|   Guthrie|    Veeler|[Mr. Brooks, Seco...|4806.88|http://dummyimage...|   1978-07-31|  true|1978|
| 20|      Maxi|     Cluet|[All I Want for C...|4046.46|http://dummyimage...|   1979-05-06| false|1979|
| 30|    Carter|     Ferre|[In a World..., B...|1737.73|http://dummyimage...|   1980-10-10| false|1980|
+---+----------+----------+--------------------+-------+--------------------+-------------+------+----+
only showing top 10 rows
```

Command took 1.70 seconds -- by 06varshagupta@gmail.com at 9/6/2022, 8:13:30 PM on 666