Analyzing airline data with Spark SQL

In [5]:

```
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Analyzing airline data") \
    .getOrCreate()
```

Exploring SQL query options

In [1]:

```
from pyspark.sql.types import Row
from datetime import datetime
```

Creating a dataframe with different data types

In [23]:

```
record = sc.parallelize([Row(id = 1,
                             name = "Jill",
                             active = True,
                             clubs = ['chess', 'hockey'],
                             subjects = {"math": 80, 'english': 56},
                             enrolled = datetime(2014, 8, 1, 14, 1, 5)),
                         Row(id = 2,
                             name = "George",
                             active = False,
                             clubs = ['chess', 'soccer'],
                             subjects = {"math": 60, 'english': 96},
                             enrolled = datetime(2015, 3, 21, 8, 2, 5))
])
```

In [24]:

```
record_df = record.toDF()
record_df.show()
```

```
+------+---------------+-------------------+---+------+--------------------+
|active|          clubs|           enrolled| id|  name|            subjects|
+------+---------------+-------------------+---+------+--------------------+
|  true|[chess, hockey]|2014-08-01 14:01:05|  1|  Jill|[english -> 56, m...|
| false|[chess, soccer]|2015-03-21 08:02:05|  2|George|[english -> 96, m...|
+------+---------------+-------------------+---+------+--------------------+
```

Register the dataframe as a temporary view

- **The view is valid for one session**
- **This is required to run SQL commands on the dataframe**

In [25]:

```
record_df.createOrReplaceTempView("records")
```

In [26]:

```
all_records_df = sqlContext.sql('SELECT * FROM records')
```

```
all_records_df.show()
```

```
+------+---------------+-------------------+---+------+--------------------+
|active|          clubs|           enrolled| id|  name|            subjects|
+------+---------------+-------------------+---+------+--------------------+
|  true|[chess, hockey]|2014-08-01 14:01:05|  1|  Jill|[english -> 56, m...|
| false|[chess, soccer]|2015-03-21 08:02:05|  2|George|[english -> 96, m...|
+------+---------------+-------------------+---+------+--------------------+
```

In [27]:

```
sqlContext.sql('SELECT id, clubs[1], subjects["english"] FROM records').show()
```

```
+---+--------+-----------------+
| id|clubs[1]|subjects[english]|
+---+--------+-----------------+
|  1|  hockey|               56|
|  2|  soccer|               96|
+---+--------+-----------------+
```

In [28]:

```
sqlContext.sql('SELECT id, NOT active FROM records').show()
```

```
+---+------------+
| id|(NOT active)|
+---+------------+
|  1|       false|
|  2|        true|
+---+------------+
```

Conditional statements in SQL

In [29]:

```
sqlContext.sql('SELECT * FROM records where active').show()
```

```
+------+---------------+-------------------+---+----+--------------------+
|active|          clubs|           enrolled| id|name|            subjects|
+------+---------------+-------------------+---+----+--------------------+
|  true|[chess, hockey]|2014-08-01 14:01:05|  1|Jill|[english -> 56, m...|
+------+---------------+-------------------+---+----+--------------------+
```

In [30]:

```
sqlContext.sql('SELECT * FROM records where subjects["english"] > 90').show()
```

```
+------+---------------+-------------------+---+------+--------------------+
|active|          clubs|           enrolled| id|  name|            subjects|
+------+---------------+-------------------+---+------+--------------------+
| false|[chess, soccer]|2015-03-21 08:02:05|  2|George|[english -> 96, m...|
+------+---------------+-------------------+---+------+--------------------+
```

Global temporary view

- **Temporary view shared across multiple sessions**
- **Kept alive till the Spark application terminates**

In [32]:

```
record_df.createGlobalTempView("global_records")
```

In [35]:

```
sqlContext.sql('SELECT * FROM global_temp.global_records').show()
```

```
+------+---------------+-------------------+---+------+--------------------+
|active|          clubs|           enrolled| id|  name|            subjects|
+------+---------------+-------------------+---+------+--------------------+
|  true|[chess, hockey]|2014-08-01 14:01:05|  1|  Jill|[english -> 56, m...|
| false|[chess, soccer]|2015-03-21 08:02:05|  2|George|[english -> 96, m...|
+------+---------------+-------------------+---+------+--------------------+
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: