Exercise 06.02 – Simple Analytics Experience

In this exercise, you will accomplish the following:

- Launch the Analytics Spark SQL shell
- Run a simple query

Step 1: Set-Up the Cluster

To complete this exercise, we configured the cluster with all nodes running DSE with Analytics enabled (we will also enable Search and Graph, although these are not strictly necessary for this exercise).

Launch the Apache Spark™ SQL Shell

With the cluster running and all nodes enabled with analytics, start up the SQL shell and run some queries.

1. Open your DSE-node2 terminal.

2. Launch the Spark SQL shell by typing the following:

```
dse spark-sql
```

If the app does not launch, use the `sudo` command to execute.

```
ubuntu@DSE-node2:~$ sudo dse spark-sql
The log file is at /home/ubuntu/.spark-sql-shell.log
spark-sql>
```

3. Using the `spark-sql` command interface, identify the most common first name and count the number of times it is used in the *killrvideo* database. Paste in the following SQL query:

```
SELECT firstname, COUNT(*)
     FROM killrvideo.users
     GROUP BY firstname
     ORDER BY count(*) desc
     LIMIT 3;
```

Notice the query does not return immediately, but instead it take a few seconds to execute. Change the `LIMIT` number to identify other frequently used first names.

```
Randi       9
```

```
Kelcy       7
Appolonia   7
Time taken: 22.191 seconds, Fetched 3 row(s)
```

4. Attempt the same query again, but change the field options; increase the `LIMIT` size, `ORDER BY asc`, `GROUP BY` another value, etc. Change multiple fields in the same SQL query:

```
SELECT lastname, COUNT(*)
     FROM killrvideo.users
     GROUP BY lastname
     ORDER BY count(*) asc
     LIMIT 5;
```

What happened this time? Were there any issues? Why is the `spark-sql` command so powerful? What other possible uses can you see for this command line application?

5. Exit the Spark SQL shell:

```
spark-sql> exit;
```

END OF EXERCISE