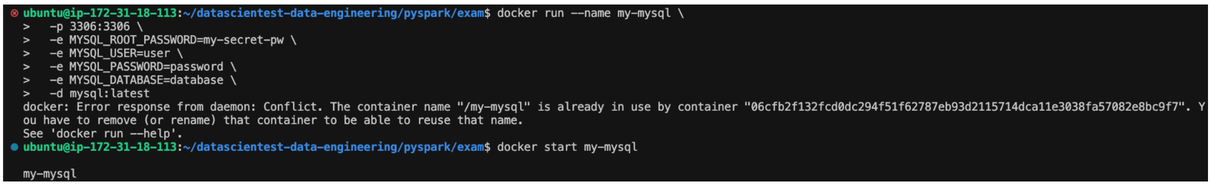
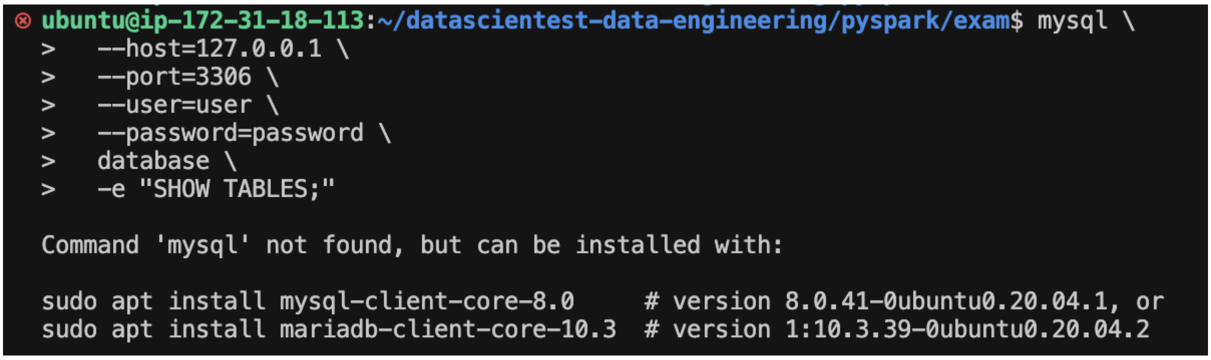
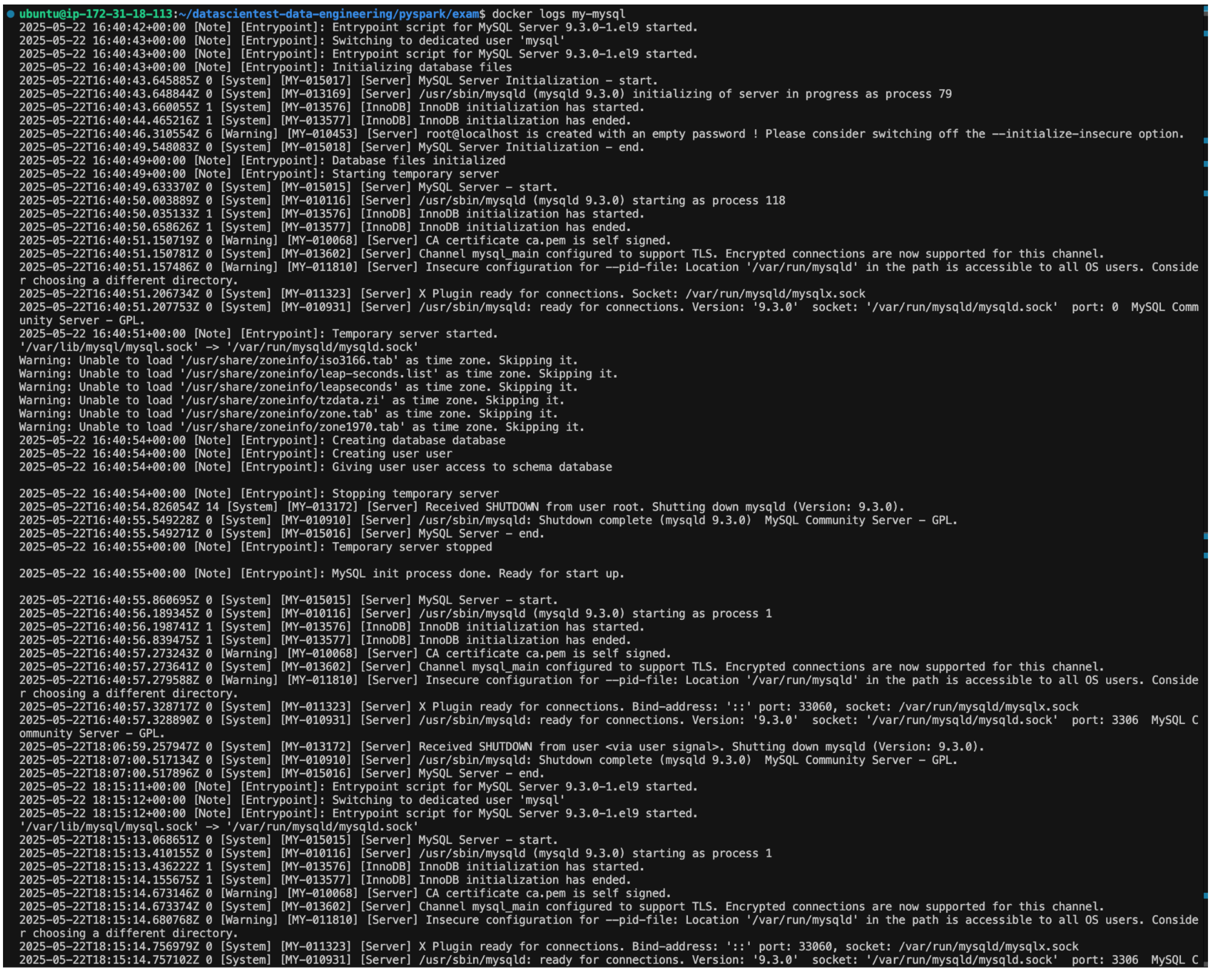
Verification Steps of PySpark Exam



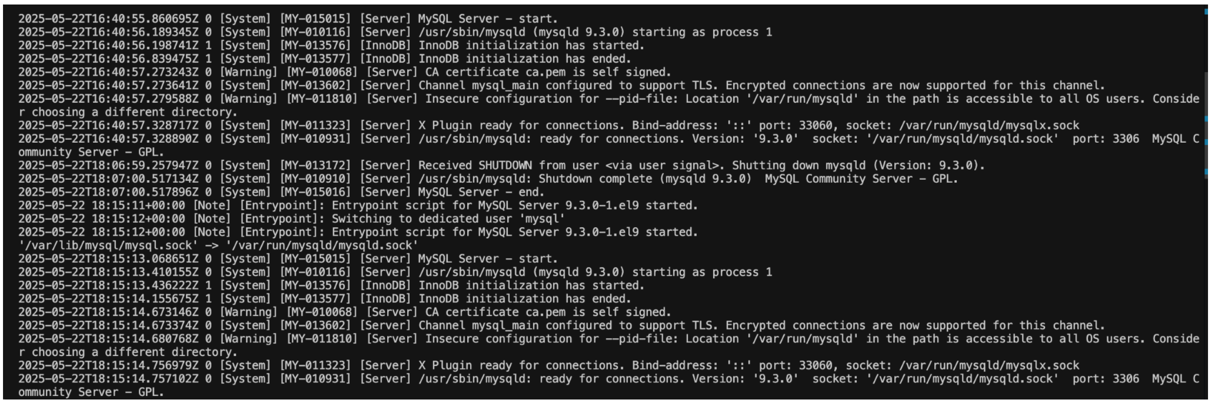
```
⊗ ubuntu@ip-172-31-18-113:~/datascientest-data-engineering/pyspark/exam$ docker run --name my-mysql \
  >   -p 3306:3306 \
  >   -e MYSQL_ROOT_PASSWORD=my-secret-pw \
  >   -e MYSQL_USER=user \
  >   -e MYSQL_PASSWORD=password \
  >   -e MYSQL_DATABASE=database \
  >   -d mysql:latest
  docker: Error response from daemon: Conflict. The container name "/my-mysql" is already in use by container "06cfb2f132fcd0dc294f51f62787eb93d2115714dca11e3038fa57082e8bc9f7". Y
  ou have to remove (or rename) that container to be able to reuse that name.
  See 'docker run --help'.
● ubuntu@ip-172-31-18-113:~/datascientest-data-engineering/pyspark/exam$ docker start my-mysql

  my-mysql
```



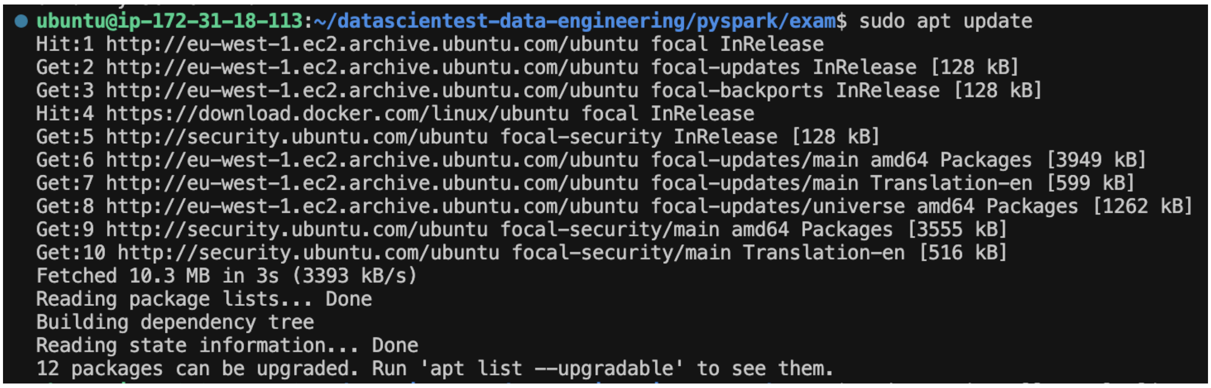
```
⊗ ubuntu@ip-172-31-18-113:~/datascientest-data-engineering/pyspark/exam$ mysql \
  >   --host=127.0.0.1 \
  >   --port=3306 \
  >   --user=user \
  >   --password=password \
  >   database \
  >   -e "SHOW TABLES;"

  Command 'mysql' not found, but can be installed with:

  sudo apt install mysql-client-core-8.0     # version 8.0.41-0ubuntu0.20.04.1, or
  sudo apt install mariadb-client-core-10.3  # version 1:10.3.39-0ubuntu0.20.04.2
```



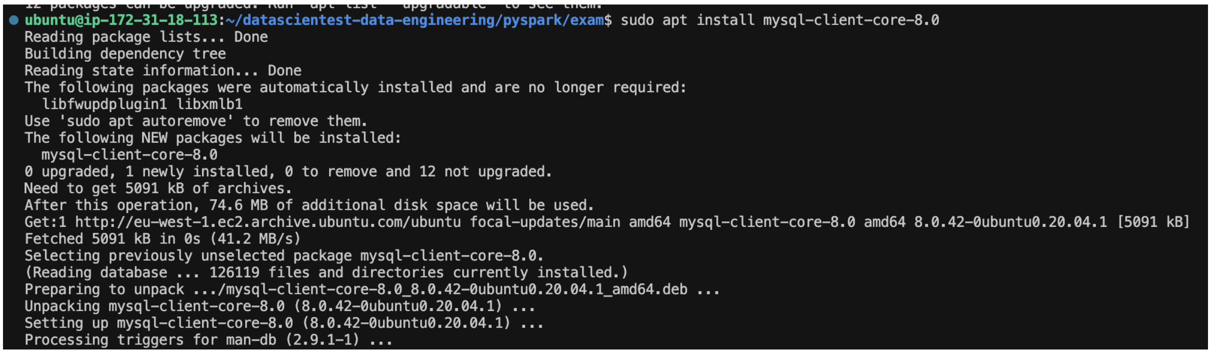
```
● ubuntu@ip-172-31-18-113:~/datascientest-data-engineering/pyspark/exam$ docker logs my-mysql
  2025-05-22 16:40:42+00:00 [Note] [Entrypoint]: Entrypoint script for MySQL Server 9.3.0-1.el9 started.
  2025-05-22 16:40:43+00:00 [Note] [Entrypoint]: Switching to dedicated user 'mysql'
  2025-05-22 16:40:43+00:00 [Note] [Entrypoint]: Entrypoint script for MySQL Server 9.3.0-1.el9 started.
  2025-05-22 16:40:43+00:00 [Note] [Entrypoint]: Initializing database files
  2025-05-22T16:40:43.645885Z 0 [System] [MY-015017] [Server] MySQL Server Initialization - start.
  2025-05-22T16:40:43.648844Z 0 [System] [MY-013169] [Server] /usr/sbin/mysqld (mysqld 9.3.0) initializing of server in progress as process 79
  2025-05-22T16:40:43.660055Z 1 [System] [MY-013576] [InnoDB] InnoDB initialization has started.
  2025-05-22T16:40:44.465216Z 1 [System] [MY-013577] [InnoDB] InnoDB initialization has ended.
  2025-05-22T16:40:46.310554Z 6 [Warning] [MY-010453] [Server] root@localhost is created with an empty password ! Please consider switching off the --initialize-insecure option.
  2025-05-22T16:40:49.548083Z 0 [System] [MY-015018] [Server] MySQL Server Initialization - end.
  2025-05-22 16:40:49+00:00 [Note] [Entrypoint]: Database files initialized
  2025-05-22 16:40:49+00:00 [Note] [Entrypoint]: Starting temporary server
  2025-05-22T16:40:49.633370Z 0 [System] [MY-015015] [Server] MySQL Server - start.
  2025-05-22T16:40:50.003889Z 0 [System] [MY-010116] [Server] /usr/sbin/mysqld (mysqld 9.3.0) starting as process 118
  2025-05-22T16:40:50.035133Z 1 [System] [MY-013576] [InnoDB] InnoDB initialization has started.
  2025-05-22T16:40:50.658626Z 1 [System] [MY-013577] [InnoDB] InnoDB initialization has ended.
  2025-05-22T16:40:51.150719Z 0 [Warning] [MY-010068] [Server] CA certificate ca.pem is self signed.
  2025-05-22T16:40:51.150781Z 0 [System] [MY-013602] [Server] Channel mysql_main configured to support TLS. Encrypted connections are now supported for this channel.
  2025-05-22T16:40:51.157486Z 0 [Warning] [MY-011810] [Server] Insecure configuration for --pid-file: Location '/var/run/mysqld' in the path is accessible to all OS users. Conside
  r choosing a different directory.
  2025-05-22T16:40:51.206734Z 0 [System] [MY-011323] [Server] X Plugin ready for connections. Socket: /var/run/mysqld/mysqlx.sock
  2025-05-22T16:40:51.207753Z 0 [System] [MY-010931] [Server] /usr/sbin/mysqld: ready for connections. Version: '9.3.0'  socket: '/var/run/mysqld/mysqld.sock'  port: 0  MySQL Comm
  unity Server - GPL.
  2025-05-22 16:40:51+00:00 [Note] [Entrypoint]: Temporary server started.
  '/var/lib/mysql/mysql.sock' -> '/var/run/mysqld/mysqld.sock'
  Warning: Unable to load '/usr/share/zoneinfo/iso3166.tab' as time zone. Skipping it.
  Warning: Unable to load '/usr/share/zoneinfo/leap-seconds.list' as time zone. Skipping it.
  Warning: Unable to load '/usr/share/zoneinfo/leapseconds' as time zone. Skipping it.
  Warning: Unable to load '/usr/share/zoneinfo/tzdata.zi' as time zone. Skipping it.
  Warning: Unable to load '/usr/share/zoneinfo/zone.tab' as time zone. Skipping it.
  Warning: Unable to load '/usr/share/zoneinfo/zone1970.tab' as time zone. Skipping it.
  2025-05-22 16:40:54+00:00 [Note] [Entrypoint]: Creating database database
  2025-05-22 16:40:54+00:00 [Note] [Entrypoint]: Creating user user
  2025-05-22 16:40:54+00:00 [Note] [Entrypoint]: Giving user user access to schema database

  2025-05-22 16:40:54+00:00 [Note] [Entrypoint]: Stopping temporary server
  2025-05-22T16:40:54.826054Z 14 [System] [MY-013172] [Server] Received SHUTDOWN from user root. Shutting down mysqld (Version: 9.3.0).
  2025-05-22T16:40:55.549228Z 0 [System] [MY-010910] [Server] /usr/sbin/mysqld: Shutdown complete (mysqld 9.3.0)  MySQL Community Server - GPL.
  2025-05-22T16:40:55.549271Z 0 [System] [MY-015016] [Server] MySQL Server - end.
  2025-05-22 16:40:55+00:00 [Note] [Entrypoint]: Temporary server stopped

  2025-05-22 16:40:55+00:00 [Note] [Entrypoint]: MySQL init process done. Ready for start up.

  2025-05-22T16:40:55.860695Z 0 [System] [MY-015015] [Server] MySQL Server - start.
  2025-05-22T16:40:56.189345Z 0 [System] [MY-010116] [Server] /usr/sbin/mysqld (mysqld 9.3.0) starting as process 1
  2025-05-22T16:40:56.198741Z 1 [System] [MY-013576] [InnoDB] InnoDB initialization has started.
  2025-05-22T16:40:56.839475Z 1 [System] [MY-013577] [InnoDB] InnoDB initialization has ended.
  2025-05-22T16:40:57.273243Z 0 [Warning] [MY-010068] [Server] CA certificate ca.pem is self signed.
  2025-05-22T16:40:57.273641Z 0 [System] [MY-013602] [Server] Channel mysql_main configured to support TLS. Encrypted connections are now supported for this channel.
  2025-05-22T16:40:57.279588Z 0 [Warning] [MY-011810] [Server] Insecure configuration for --pid-file: Location '/var/run/mysqld' in the path is accessible to all OS users. Conside
  r choosing a different directory.
  2025-05-22T16:40:57.328717Z 0 [System] [MY-011323] [Server] X Plugin ready for connections. Bind-address: '::' port: 33060, socket: /var/run/mysqld/mysqlx.sock
  2025-05-22T16:40:57.328890Z 0 [System] [MY-010931] [Server] /usr/sbin/mysqld: ready for connections. Version: '9.3.0'  socket: '/var/run/mysqld/mysqld.sock'  port: 3306  MySQL C
  ommunity Server - GPL.
  2025-05-22T18:06:59.257947Z 0 [System] [MY-013172] [Server] Received SHUTDOWN from user <via user signal>. Shutting down mysqld (Version: 9.3.0).
  2025-05-22T18:07:00.517134Z 0 [System] [MY-010910] [Server] /usr/sbin/mysqld: Shutdown complete (mysqld 9.3.0)  MySQL Community Server - GPL.
  2025-05-22T18:07:00.517896Z 0 [System] [MY-015016] [Server] MySQL Server - end.
  2025-05-22 18:15:11+00:00 [Note] [Entrypoint]: Entrypoint script for MySQL Server 9.3.0-1.el9 started.
  2025-05-22 18:15:12+00:00 [Note] [Entrypoint]: Switching to dedicated user 'mysql'
  2025-05-22 18:15:12+00:00 [Note] [Entrypoint]: Entrypoint script for MySQL Server 9.3.0-1.el9 started.
  '/var/lib/mysql/mysql.sock' -> '/var/run/mysqld/mysqld.sock'
  2025-05-22T18:15:13.068651Z 0 [System] [MY-015015] [Server] MySQL Server - start.
  2025-05-22T18:15:13.410155Z 0 [System] [MY-010116] [Server] /usr/sbin/mysqld (mysqld 9.3.0) starting as process 1
  2025-05-22T18:15:13.436222Z 1 [System] [MY-013576] [InnoDB] InnoDB initialization has started.
  2025-05-22T18:15:14.155675Z 1 [System] [MY-013577] [InnoDB] InnoDB initialization has ended.
  2025-05-22T18:15:14.673146Z 0 [Warning] [MY-010068] [Server] CA certificate ca.pem is self signed.
  2025-05-22T18:15:14.673374Z 0 [System] [MY-013602] [Server] Channel mysql_main configured to support TLS. Encrypted connections are now supported for this channel.
  2025-05-22T18:15:14.680768Z 0 [Warning] [MY-011810] [Server] Insecure configuration for --pid-file: Location '/var/run/mysqld' in the path is accessible to all OS users. Conside
  r choosing a different directory.
  2025-05-22T18:15:14.756979Z 0 [System] [MY-011323] [Server] X Plugin ready for connections. Bind-address: '::' port: 33060, socket: /var/run/mysqld/mysqlx.sock
  2025-05-22T18:15:14.757102Z 0 [System] [MY-010931] [Server] /usr/sbin/mysqld: ready for connections. Version: '9.3.0'  socket: '/var/run/mysqld/mysqld.sock'  port: 3306  MySQL C
```



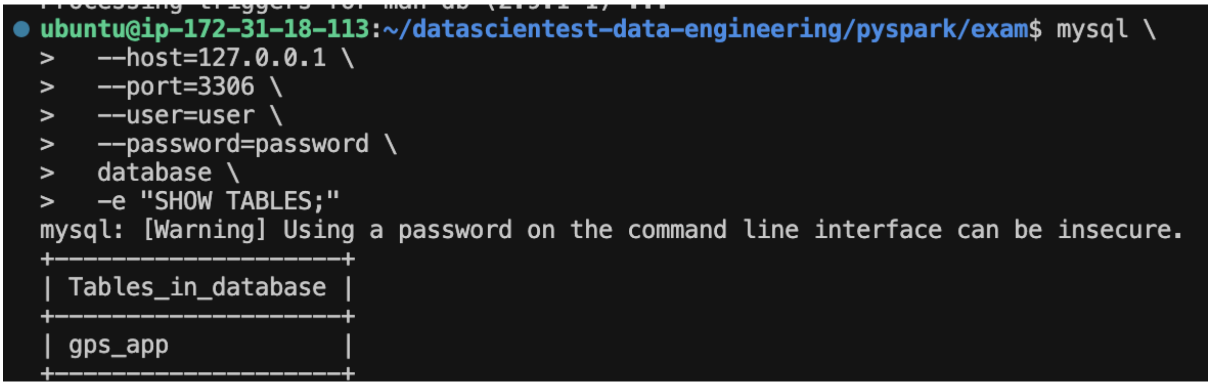
```
2025-05-22T16:40:55.860695Z 0 [System] [MY-015015] [Server] MySQL Server - start.
2025-05-22T16:40:56.189345Z 0 [System] [MY-010116] [Server] /usr/sbin/mysqld (mysqld 9.3.0) starting as process 1
2025-05-22T16:40:56.198741Z 1 [System] [MY-013576] [InnoDB] InnoDB initialization has started.
2025-05-22T16:40:56.839475Z 1 [System] [MY-013577] [InnoDB] InnoDB initialization has ended.
2025-05-22T16:40:57.273243Z 0 [Warning] [MY-010068] [Server] CA certificate ca.pem is self signed.
2025-05-22T16:40:57.273641Z 0 [System] [MY-013602] [Server] Channel mysql_main configured to support TLS. Encrypted connections are now supported for this channel.
2025-05-22T16:40:57.279588Z 0 [Warning] [MY-011810] [Server] Insecure configuration for --pid-file: Location '/var/run/mysqld' in the path is accessible to all OS users. Conside
r choosing a different directory.
2025-05-22T16:40:57.328717Z 0 [System] [MY-011323] [Server] X Plugin ready for connections. Bind-address: '::' port: 33060, socket: /var/run/mysqld/mysqlx.sock
2025-05-22T16:40:57.328890Z 0 [System] [MY-010931] [Server] /usr/sbin/mysqld: ready for connections. Version: '9.3.0'  socket: '/var/run/mysqld/mysqld.sock'  port: 3306  MySQL C
ommunity Server - GPL.
2025-05-22T18:06:59.257947Z 0 [System] [MY-013172] [Server] Received SHUTDOWN from user <via user signal>. Shutting down mysqld (Version: 9.3.0).
2025-05-22T18:07:00.517134Z 0 [System] [MY-010910] [Server] /usr/sbin/mysqld: Shutdown complete (mysqld 9.3.0)  MySQL Community Server - GPL.
2025-05-22T18:07:00.517896Z 0 [System] [MY-015016] [Server] MySQL Server - end.
2025-05-22 18:15:11+00:00 [Note] [Entrypoint]: Entrypoint script for MySQL Server 9.3.0-1.el9 started.
2025-05-22 18:15:12+00:00 [Note] [Entrypoint]: Switching to dedicated user 'mysql'
2025-05-22 18:15:12+00:00 [Note] [Entrypoint]: Entrypoint script for MySQL Server 9.3.0-1.el9 started.
'/var/lib/mysql/mysql.sock' -> '/var/run/mysqld/mysqld.sock'
2025-05-22T18:15:13.068651Z 0 [System] [MY-015015] [Server] MySQL Server - start.
2025-05-22T18:15:13.410155Z 0 [System] [MY-010116] [Server] /usr/sbin/mysqld (mysqld 9.3.0) starting as process 1
2025-05-22T18:15:13.436222Z 1 [System] [MY-013576] [InnoDB] InnoDB initialization has started.
2025-05-22T18:15:14.155675Z 1 [System] [MY-013577] [InnoDB] InnoDB initialization has ended.
2025-05-22T18:15:14.673146Z 0 [Warning] [MY-010068] [Server] CA certificate ca.pem is self signed.
2025-05-22T18:15:14.673374Z 0 [System] [MY-013602] [Server] Channel mysql_main configured to support TLS. Encrypted connections are now supported for this channel.
2025-05-22T18:15:14.680768Z 0 [Warning] [MY-011810] [Server] Insecure configuration for --pid-file: Location '/var/run/mysqld' in the path is accessible to all OS users. Conside
r choosing a different directory.
2025-05-22T18:15:14.756979Z 0 [System] [MY-011323] [Server] X Plugin ready for connections. Bind-address: '::' port: 33060, socket: /var/run/mysqld/mysqlx.sock
2025-05-22T18:15:14.757102Z 0 [System] [MY-010931] [Server] /usr/sbin/mysqld: ready for connections. Version: '9.3.0'  socket: '/var/run/mysqld/mysqld.sock'  port: 3306  MySQL C
ommunity Server - GPL.
```



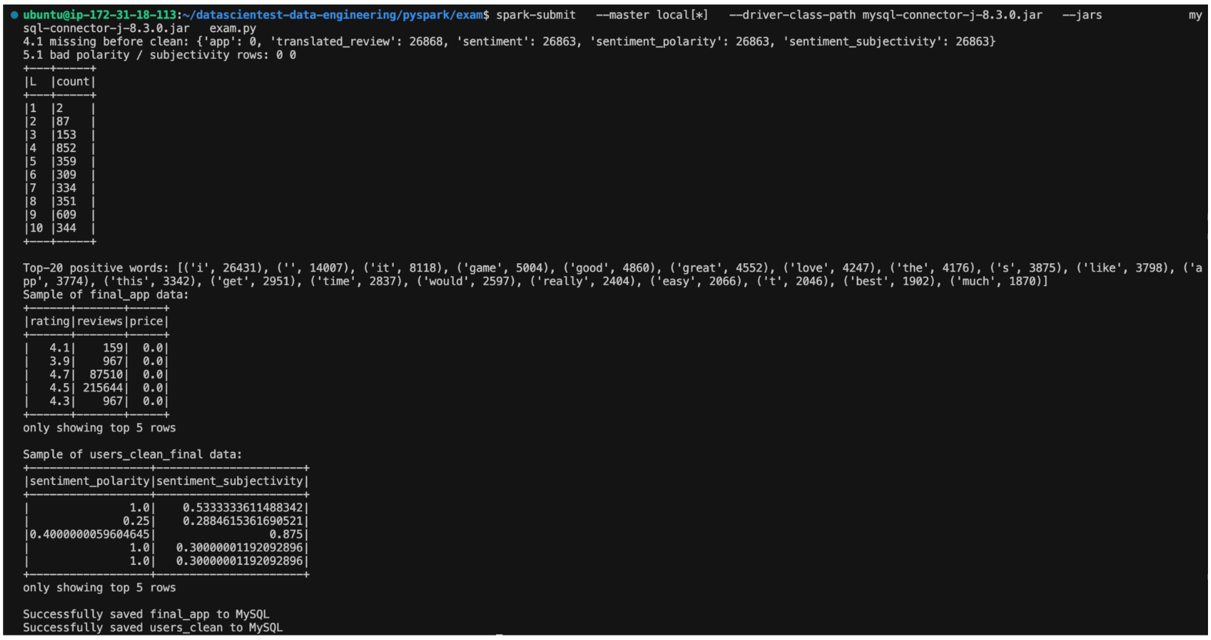
```
● ubuntu@ip-172-31-18-113:~/datascientest-data-engineering/pyspark/exam$ sudo apt update
  Hit:1 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu focal InRelease
  Get:2 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu focal-updates InRelease [128 kB]
  Get:3 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu focal-backports InRelease [128 kB]
  Hit:4 https://download.docker.com/linux/ubuntu focal InRelease
  Get:5 http://security.ubuntu.com/ubuntu focal-security InRelease [128 kB]
  Get:6 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu focal-updates/main amd64 Packages [3949 kB]
  Get:7 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu focal-updates/main Translation-en [599 kB]
  Get:8 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu focal-updates/universe amd64 Packages [1262 kB]
  Get:9 http://security.ubuntu.com/ubuntu focal-security/main amd64 Packages [3555 kB]
  Get:10 http://security.ubuntu.com/ubuntu focal-security/main Translation-en [516 kB]
  Fetched 10.3 MB in 3s (3393 kB/s)
  Reading package lists... Done
  Building dependency tree
  Reading state information... Done
  12 packages can be upgraded. Run 'apt list --upgradable' to see them.
```



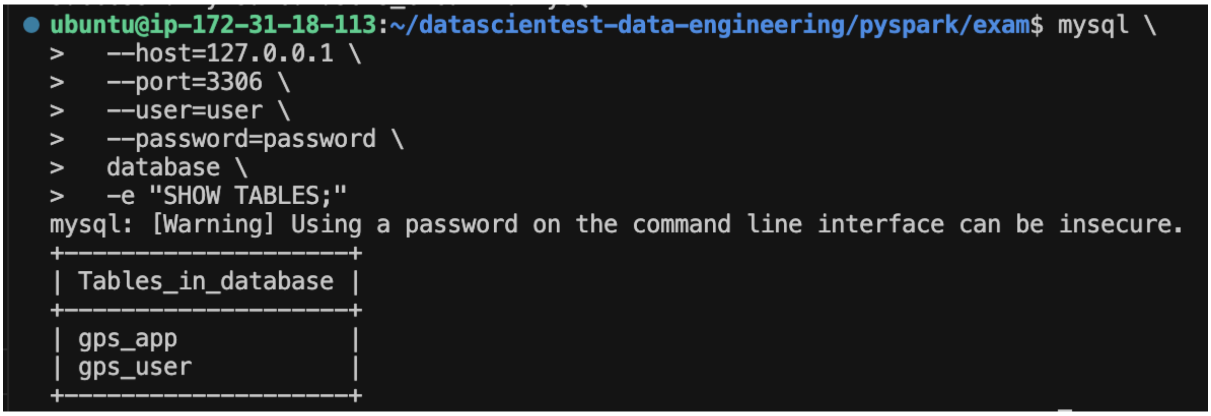
```
● ubuntu@ip-172-31-18-113:~/datascientest-data-engineering/pyspark/exam$ sudo apt install mysql-client-core-8.0
  Reading package lists... Done
  Building dependency tree
  Reading state information... Done
  The following packages were automatically installed and are no longer required:
    libfwupdplugin1 libxmlb1
  Use 'sudo apt autoremove' to remove them.
  The following NEW packages will be installed:
    mysql-client-core-8.0
  0 upgraded, 1 newly installed, 0 to remove and 12 not upgraded.
  Need to get 5091 kB of archives.
  After this operation, 74.6 MB of additional disk space will be used.
  Get:1 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu focal-updates/main amd64 mysql-client-core-8.0 amd64 8.0.42-0ubuntu0.20.04.1 [5091 kB]
  Fetched 5091 kB in 0s (41.2 MB/s)
  Selecting previously unselected package mysql-client-core-8.0.
  (Reading database ... 126119 files and directories currently installed.)
  Preparing to unpack .../mysql-client-core-8.0_8.0.42-0ubuntu0.20.04.1_amd64.deb ...
  Unpacking mysql-client-core-8.0 (8.0.42-0ubuntu0.20.04.1) ...
  Setting up mysql-client-core-8.0 (8.0.42-0ubuntu0.20.04.1) ...
  Processing triggers for man-db (2.9.1-1) ...
```



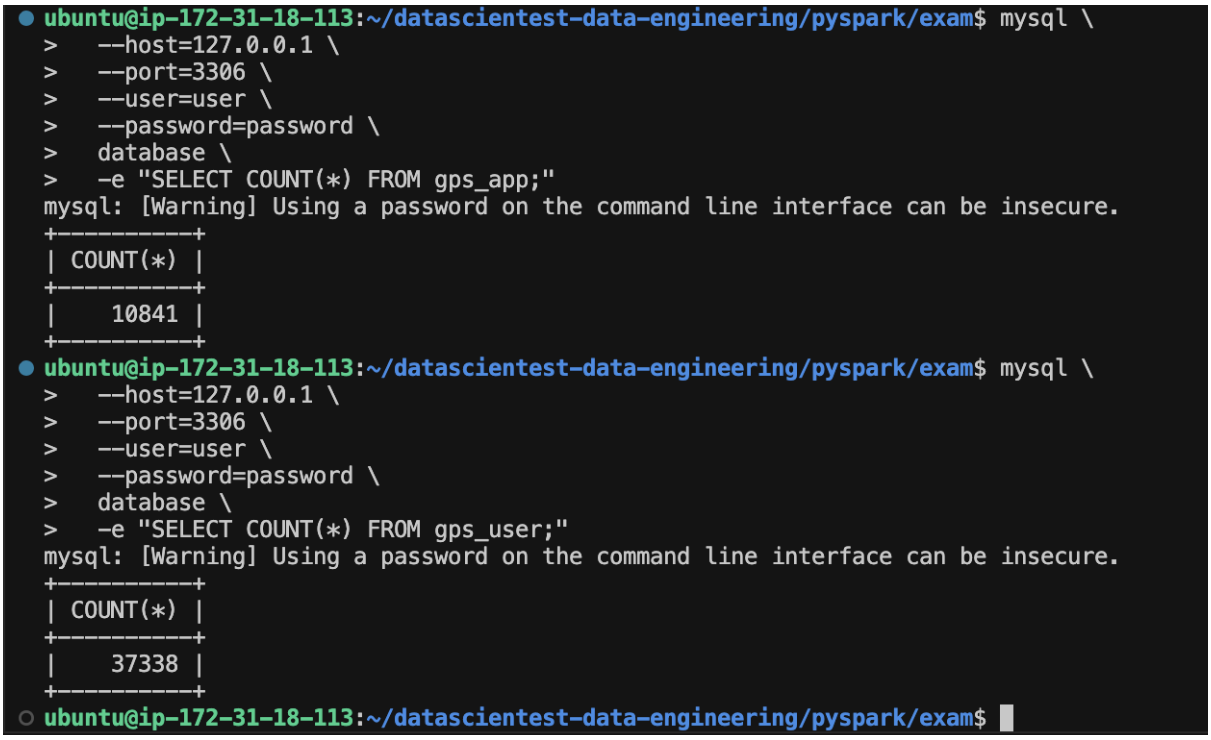
```
● ubuntu@ip-172-31-18-113:~/datascientest-data-engineering/pyspark/exam$ mysql \
  >   --host=127.0.0.1 \
  >   --port=3306 \
  >   --user=user \
  >   --password=password \
  >   database \
  >   -e "SHOW TABLES;"
  mysql: [Warning] Using a password on the command line interface can be insecure.
  +---------------------+
  | Tables_in_database |
  +---------------------+
  | gps_app             |
  +---------------------+
```



```
● ubuntu@ip-172-31-18-113:~/datascientest-data-engineering/pyspark/exam$ spark-submit   --master local[*]   --driver-class-path mysql-connector-j-8.3.0.jar   --jars          my
  sql-connector-j-8.3.0.jar   exam.py
  4.1 missing before clean: {'app': 0, 'translated_review': 26868, 'sentiment': 26863, 'sentiment_polarity': 26863, 'sentiment_subjectivity': 26863}
  5.1 bad polarity / subjectivity rows: 0 0
  +---+-----+
  |L  |count|
  +---+-----+
  |1  |2    |
  |2  |87   |
  |3  |153  |
  |4  |852  |
  |5  |359  |
  |6  |309  |
  |7  |334  |
  |8  |351  |
  |9  |609  |
  |10 |344  |
  +---+-----+

  Top-20 positive words: [('i', 26431), ('', 14007), ('it', 8118), ('game', 5004), ('good', 4860), ('great', 4552), ('love', 4247), ('the', 4176), ('s', 3875), ('like', 3798), ('a
  pp', 3774), ('this', 3342), ('get', 2951), ('time', 2837), ('would', 2597), ('really', 2404), ('easy', 2066), ('t', 2046), ('best', 1902), ('much', 1870)]
  Sample of final_app data:
  +------+-------+-----+
  |rating|reviews|price|
  +------+-------+-----+
  |   4.1|    159|  0.0|
  |   3.9|    967|  0.0|
  |   4.7|  87510|  0.0|
  |   4.5| 215644|  0.0|
  |   4.3|    967|  0.0|
  +------+-------+-----+
  only showing top 5 rows

  Sample of users_clean_final data:
  +------------------+----------------------+
  |sentiment_polarity|sentiment_subjectivity|
  +------------------+----------------------+
  |               1.0|    0.5333333611488342|
  |              0.25|    0.2884615361690521|
  |0.4000000059604645|                 0.875|
  |               1.0|   0.30000001192092896|
  |               1.0|   0.30000001192092896|
  +------------------+----------------------+
  only showing top 5 rows

  Successfully saved final_app to MySQL
  Successfully saved users_clean to MySQL
```



```
● ubuntu@ip-172-31-18-113:~/datascientest-data-engineering/pyspark/exam$ mysql \
  >   --host=127.0.0.1 \
  >   --port=3306 \
  >   --user=user \
  >   --password=password \
  >   database \
  >   -e "SHOW TABLES;"
  mysql: [Warning] Using a password on the command line interface can be insecure.
  +---------------------+
  | Tables_in_database |
  +---------------------+
  | gps_app             |
  | gps_user            |
  +---------------------+
```



```
ubuntu@ip-172-31-18-113:~/datascientest-data-engineering/pyspark/exam$ mysql \
>   --host=127.0.0.1 \
>   --port=3306 \
>   --user=user \
>   --password=password \
>   database \
>   -e "SELECT COUNT(*) FROM gps_app;"
mysql: [Warning] Using a password on the command line interface can be insecure.
+----------+
| COUNT(*) |
+----------+
|    10841 |
+----------+
ubuntu@ip-172-31-18-113:~/datascientest-data-engineering/pyspark/exam$ mysql \
>   --host=127.0.0.1 \
>   --port=3306 \
>   --user=user \
>   --password=password \
>   database \
>   -e "SELECT COUNT(*) FROM gps_user;"
mysql: [Warning] Using a password on the command line interface can be insecure.
+----------+
| COUNT(*) |
+----------+
|    37338 |
+----------+
ubuntu@ip-172-31-18-113:~/datascientest-data-engineering/pyspark/exam$ 
```