

SQOOP Pyspark with my sql

SQOOP

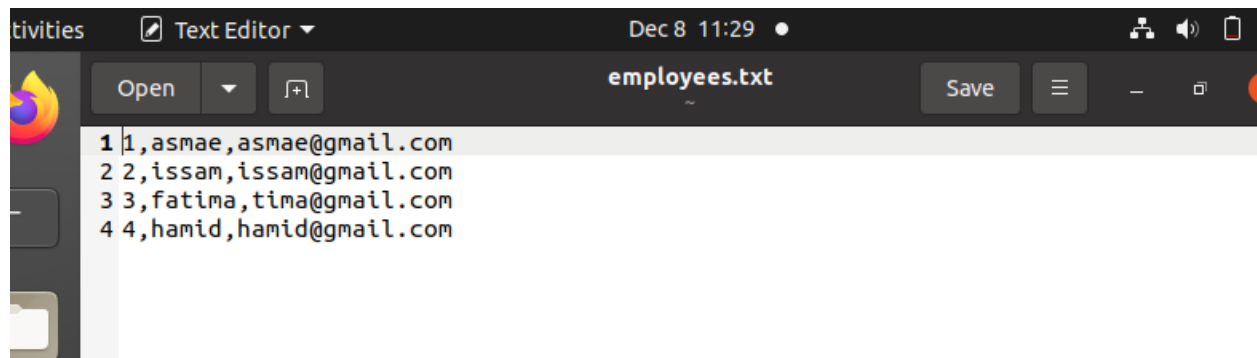
```
asmae@asmae-VirtualBox:/opt/lampp$ ls
apache2      htdocs      libexec      pear          sbin
bin           icons       licenses     php           share
build        img         logs         phpmyadmin    xampp
cgi-bin      include     man          proftpd       THIRDPARTY
ctlscrip.sh  info       manager-linux-x64.run  properties.ini  uninstall
docs         lampp      manual       README.md     uninstall.dat
error        lib        modules     README-wsrep  var
etc          lib64      mysql       RELEASENOTES  xampp

asmae@asmae-VirtualBox:/opt/lampp$ sudo manager-linux-x64.run
[sudo] password for asmae:
sudo: manager-linux-x64.run: command not found
asmae@asmae-VirtualBox:/opt/lampp$ sudo ./manager-linux-x64.run
asmae@asmae-VirtualBox:/opt/lampp$ start-dfs.sh
Starting namenodes on [localhost]
asmae@localhost's password:
localhost: starting namenode, logging to /home/asmae/hadoop-2.7.1/logs/hadoop-asmae-
namenode-asmae-VirtualBox.out
asmae@localhost's password:
localhost: starting datanode, logging to /home/asmae/hadoop-2.7.1/logs/hadoop-asmae-
datanode-asmae-VirtualBox.out
Starting secondary namenodes [0.0.0.0]
asmae@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /home/asmae/hadoop-2.7.1/logs/hadoo
p-asmae-secondarynamenode-asmae-VirtualBox.out
asmae@asmae-VirtualBox:/opt/lampp$ jps
28321 DataNode
28535 SecondaryNameNode
28648 Jps
4170 Master
28139 NameNode
asmae@asmae-VirtualBox:/opt/lampp$
```

Import (Prendre le contenu de la table employees est le mettre dans le dossier scoop)

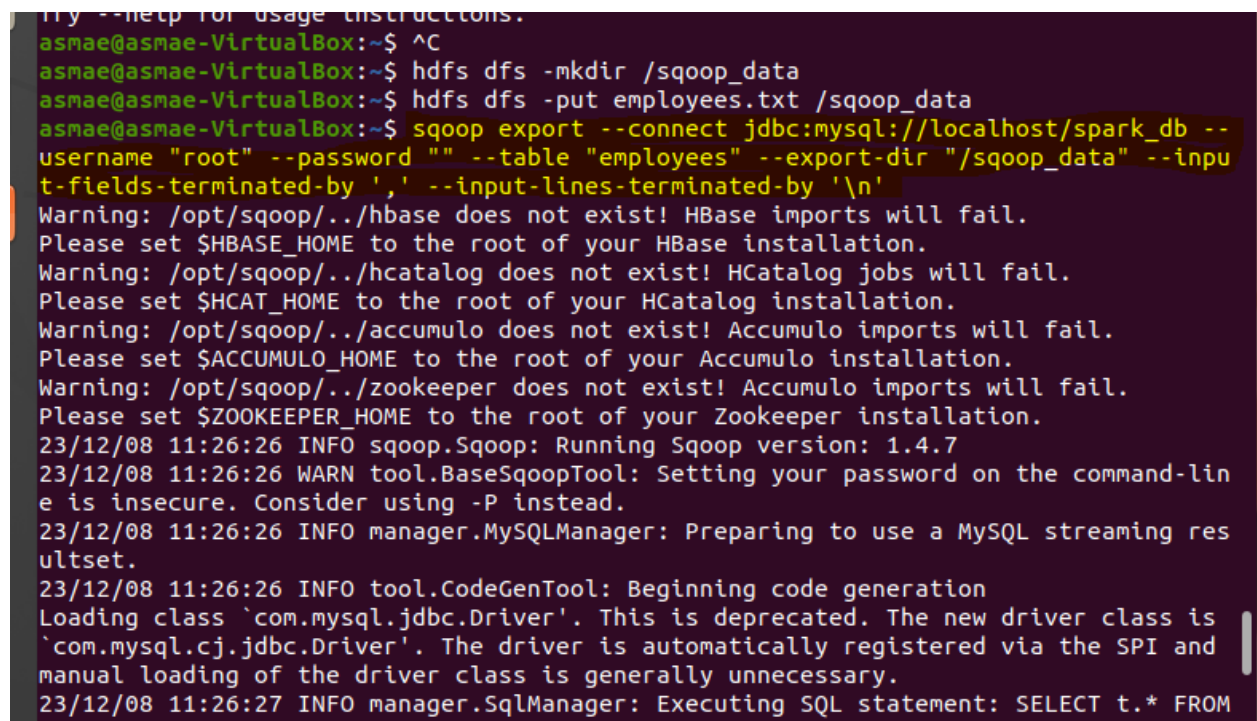
```
28139 NameNode
asmae@asmae-VirtualBox:/opt/lampp$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /home/asmae/hadoop-2.7.1/logs/yarn-asmae-resou
rcemanager-asmae-VirtualBox.out
asmae@localhost's password:
localhost: starting nodemanager, logging to /home/asmae/hadoop-2.7.1/logs/yarn-asma
e-nodemanager-asmae-VirtualBox.out
asmae@asmae-VirtualBox:/opt/lampp$ jps
28321 DataNode
28791 ResourceManager
28535 SecondaryNameNode
4170 Master
28139 NameNode
29180 Jps
29119 NodeManager
asmae@asmae-VirtualBox:/opt/lampp$ sqoop import --connect jdbc:mysql://localhost/sp
ark_db --username "root" --password "" --table employees --target-dir /sqoop
Warning: /opt/sqoop/./hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /opt/sqoop/./hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /opt/sqoop/./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /opt/sqoop/./zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
23/12/08 11:02:51 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
23/12/08 11:02:51 WARN tool.BaseSqoopTool: Setting your password on the command-lin
```

Export (Mettre le contenu de employees.txt dans la base de donnés)



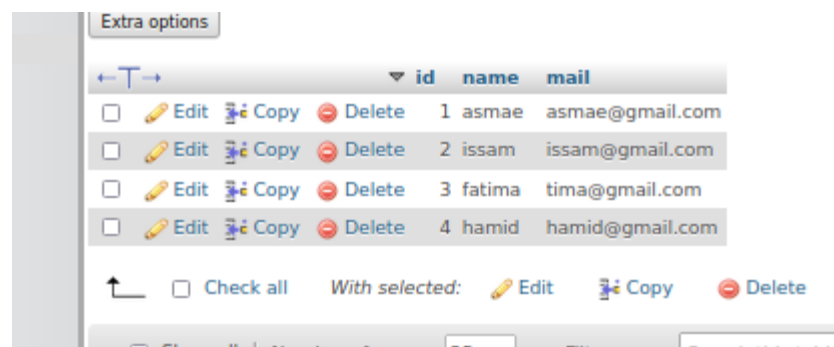
A screenshot of a text editor window titled 'employees.txt'. The editor shows four lines of text, each representing an employee record with an ID, name, and email address.

```
1 1,asmae,asmae@gmail.com
2 2,issam,issam@gmail.com
3 3,fatima,tima@gmail.com
4 4,hamid,hamid@gmail.com
```



A screenshot of a terminal window showing the execution of the 'sqoop export' command. The terminal output includes warnings about missing HBase, HCatalog, and Accumulo installations, followed by status messages from sqoop and the MySQL driver, and finally the execution of the SQL statement 'SELECT t.* FROM'.

```
asmae@asmae-VirtualBox:~$ ^C
asmae@asmae-VirtualBox:~$ hdfs dfs -mkdir /sqoop_data
asmae@asmae-VirtualBox:~$ hdfs dfs -put employees.txt /sqoop_data
asmae@asmae-VirtualBox:~$ sqoop export --connect jdbc:mysql://localhost/spark_db --
username "root" --password "" --table "employees" --export-dir "/sqoop_data" --input-
t-fields-terminated-by ',' --input-lines-terminated-by '\n'
Warning: /opt/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /opt/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /opt/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /opt/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
23/12/08 11:26:26 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
23/12/08 11:26:26 WARN tool.BaseSqoopTool: Setting your password on the command-lin
e is insecure. Consider using -P instead.
23/12/08 11:26:26 INFO manager.MySQLManager: Preparing to use a MySQL streaming res
ultset.
23/12/08 11:26:26 INFO tool.CodeGenTool: Beginning code generation
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is
'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and
manual loading of the driver class is generally unnecessary.
23/12/08 11:26:27 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM
```



A screenshot of a database management tool interface showing the exported data. The table has columns 'id', 'name', and 'mail'. There are four rows of data, each with a checkbox, an 'Edit' button, a 'Copy' button, and a 'Delete' button. Below the table, there are options to 'Check all', 'With selected:', 'Edit', 'Copy', and 'Delete'.

	id	name	mail
<input type="checkbox"/>	1	asmae	asmae@gmail.com
<input type="checkbox"/>	2	issam	issam@gmail.com
<input type="checkbox"/>	3	fatima	tima@gmail.com
<input type="checkbox"/>	4	hamid	hamid@gmail.com

Base de données

	ID_PROJECT	TITLE	DESCRIPTION	LIEU	DATE_DEBUT	DATE_FINE
<input type="checkbox"/> Edit Copy Delete	1	Project One	Description One	Location One	2023-01-01	2023-12-31
<input type="checkbox"/> Edit Copy Delete	2	Project Two	Description Two	Location Two	2023-02-01	2023-11-30
<input type="checkbox"/> Edit Copy Delete	3	Project Three	Description Three	Location Three	2023-03-01	2023-10-15
<input type="checkbox"/> Edit Copy Delete	4	Project Four	Description Four	Location Four	2023-04-15	2023-09-30
<input type="checkbox"/> Edit Copy Delete	5	Project Five	Description Five	Location Five	2023-05-01	2023-08-31

Extra options

	ID_TACHE	TITLE	DATE_DEBUT	DATE_FIN	TERMINE	ID_PROJET
<input type="checkbox"/> Edit Copy Delete	1	Task One	2023-01-01	2023-01-15	0	1
<input type="checkbox"/> Edit Copy Delete	2	Task Two	2023-02-01	2023-02-28	1	2
<input type="checkbox"/> Edit Copy Delete	3	Task Three	2023-03-01	2023-07-20	1	3
<input type="checkbox"/> Edit Copy Delete	4	Task Four	2023-04-01	2023-04-30	0	4
<input type="checkbox"/> Edit Copy Delete	5	Task Five	2023-05-01	2023-05-15	0	5
<input type="checkbox"/> Edit Copy Delete	6	Task Six	2023-06-01	2023-08-30	0	1
<input type="checkbox"/> Edit Copy Delete	7	Task Seven	2023-07-01	2023-09-30	1	2
<input type="checkbox"/> Edit Copy Delete	8	Task Eight	2023-08-01	2023-10-15	1	3
<input type="checkbox"/> Edit Copy Delete	9	Task One	2023-01-01	2023-01-15	1	1
<input type="checkbox"/> Edit Copy Delete	10	Task Two	2023-02-01	2023-06-28	1	2
<input type="checkbox"/> Edit Copy Delete	11	Task Three	2023-03-01	2023-03-20	0	1
<input type="checkbox"/> Edit Copy Delete	12	Task Four	2023-04-01	2023-05-02	0	2
<input type="checkbox"/> Edit Copy Delete	13	Task Five	2023-05-01	2023-05-15	0	1
<input type="checkbox"/> Edit Copy Delete	14	Task Six	2023-06-01	2023-06-30	0	2

Check all With selected: Edit Copy Delete Export

```
asmae@asmae-VirtualBox:~$ pyspark --jars "/home/asmae/mysql-connector-j-8.2.0.jar"
Python 3.8.10 (default, Nov 22 2023, 10:22:35)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license()" for more information.
23/12/08 23:22:25 WARN Utils: Your hostname, asmae-VirtualBox resolves to a loopback address: 127.0.0.1; using 10.0.2.15 instead (on interface enp0s3)
23/12/08 23:22:25 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
23/12/08 23:22:27 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform.. using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to

Spark version 3.0.1

Using Python version 3.8.10 (default, Nov 22 2023 10:22:35)
SparkSession available as 'spark'.
>>> from pyspark.sql import SparkSession
>>> spark=SparkSession.builder.appName("TP SQL").master("localhost[*]").getOrCreate();
```

```
asmae@asmae-VirtualBox: ~
File "<stdin>", line 1, in <module>
TypeError: 'Builder' object is not callable
>>> spark=SparkSession.builder.appName("TP SQL").master("localhost[*]").getOrCreate();
>>> dfProjets=spark.read.format("jdbc").option("driver","com.mysql.jdbc.Driver").option("url","jdbc:mysql://localhost:3306/spark_db").option("user","root").option("password","").load()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/opt/spark/python/pyspark/sql/readwriter.py", line 184, in load
    return self._jreader.load()
  File "/opt/spark/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py", line 1304, in __call__
  File "/opt/spark/python/pyspark/sql/utils.py", line 134, in deco
    raise_from(converted)
  File "<string>", line 3, in raise_from
pyspark.sql.utils.IllegalArgumentException: Option 'dbtable' or 'query' is required.
>>> dfProjets=spark.read.format("jdbc").option("driver","com.mysql.jdbc.Driver").option("url","jdbc:mysql://localhost:3306/spark_db").option("user","root").option("password","").option("dbtable","PROJETS").load()
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
>>> dfProjets.show()
+-----+-----+-----+-----+-----+-----+
|ID_PROJECT|TITLE|DESCRIPTION|LIEU|DATE_DEBUT|DATE_FINE|
+-----+-----+-----+-----+-----+-----+
|1|Project One|Description One|Location One|2023-01-01|2023-12-31|
|2|Project Two|Description Two|Location Two|2023-02-01|2023-11-30|
|3|Project Three|Description Three|Location Three|2023-03-01|2023-10-15|
|4|Project Four|Description Four|Location Four|2023-04-15|2023-09-30|
|5|Project Five|Description Five|Location Five|2023-05-01|2023-08-31|
+-----+-----+-----+-----+-----+-----+
```

1. Afficher les projets en cours de réalisation.

```
>>> spark.read.format("jdbc").option("driver","com.mysql.jdbc.Driver").option("url","jdbc:mysql://localhost:3306/spark_db").option("user","root").option("password","").option("query","SELECT * FROM `PROJETS` WHERE `DATE_FINE` >= CURDATE()").load().show()
+-----+-----+-----+-----+-----+-----+
|ID_PROJECT|    TITLE| DESCRIPTION|    LIEU|DATE_DEBUT| DATE_FINE|
+-----+-----+-----+-----+-----+-----+
|         1|Project One|Description One|Location One|2023-01-01|2023-12-31|
+-----+-----+-----+-----+-----+-----+
```

2. Afficher pour chaque projet, le nombre de tâches dont la durée dépasse un mois. Le format d'affichage est le suivant :

ID_PROJET | TITRE | NOMBRE

```
>>> spark.read.format("jdbc").option("driver","com.mysql.jdbc.Driver").option("url","jdbc:mysql://localhost:3306/spark_db").option("user","root").option("password","").option("query","SELECT p.ID_PROJECT, p.TITLE, COUNT(ID_TACHE) FROM PROJETS p JOIN TACHES t WHERE p.ID_PROJECT=t.ID_PROJECT and (t.DATE_FIN - t.DATE_DEBUT) > 60 GROUP BY p.ID_PROJECT").load().show()
[Stage 3:>                                (0 + 0) / 1
[Stage 3:>                                (0 + 1) / 1
+-----+-----+-----+
|ID_PROJECT|    TITLE|COUNT(ID_TACHE)|
+-----+-----+-----+
|         1|Project One|                1|
|         2|Project Two|                3|
|         3|Project Three|                2|
+-----+-----+-----+
>>> 
```

3. Afficher pour chaque projet les tâches en retard (avec la durée de retard).

```
>>> spark.read.format("jdbc").option("driver","com.mysql.jdbc.Driver").option("url","jdbc:mysql://localhost:3306/spark_db").option("user","root").option("password","").option("query","SELECT t.ID_TACHE , p.ID_PROJECT,p.TITLE,t.DATE_FIN,t.TERMINER from PROJETS p JOIN TACHES t WHERE p.ID_PROJECT=t.ID_PROJECT and t.TERMINER=0 AND t.DATE_FIN<CURRENT_DATE").load().show()
+-----+-----+-----+-----+-----+
|ID_TACHE|ID_PROJECT|    TITLE| DATE_FIN|TERMINER|
+-----+-----+-----+-----+-----+
|        1|         1|Project One|2023-01-15|   false|
|        6|         1|Project One|2023-08-30|   false|
|       11|         1|Project One|2023-03-20|   false|
|       13|         1|Project One|2023-05-15|   false|
|       12|         2|Project Two|2023-05-02|   false|
|       14|         2|Project Two|2023-06-30|   false|
|         4|         4|Project Four|2023-04-30|   false|
|         5|         5|Project Five|2023-05-15|   false|
+-----+-----+-----+-----+-----+
```