INSTITUT POLYTECHNIQUE DE SAINT-LOUIS

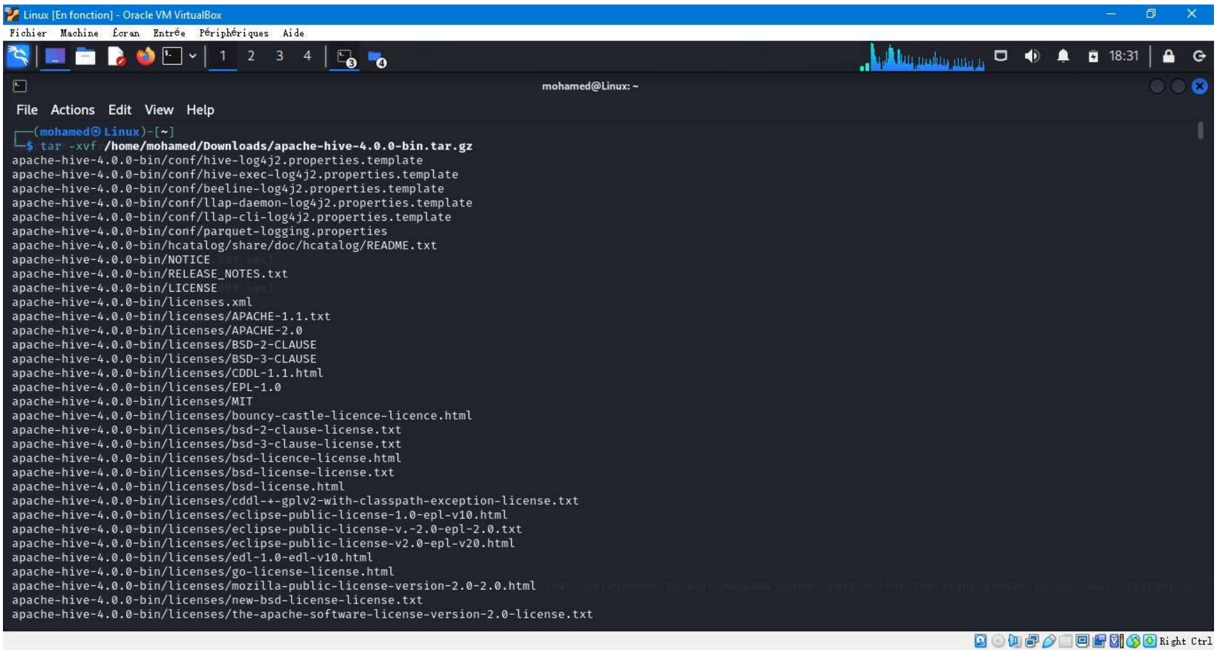
UNIVERSITE GASTON BERGER SAINT-LOUIS DU SENEGAL

BIG DATA

RAPORT DE PROJET BIG DATA

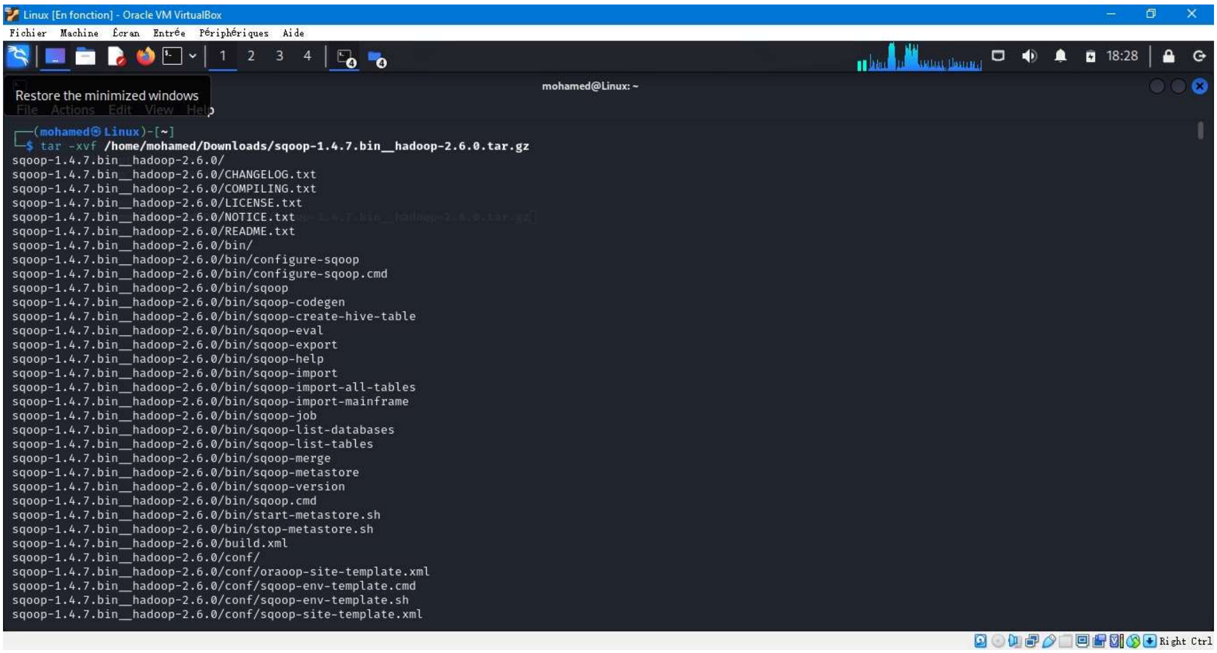
REALISER PAR: MOHAMED EL HOUSSEIN CHEIKH

Install Apache Hive:


Linux [En fonction] - Oracle VM VirtualBox
Fichier Machine Écran Entrée Périphériques Aide
18:31
mohamed@Linux: ~
File Actions Edit View Help
Right Ctrl

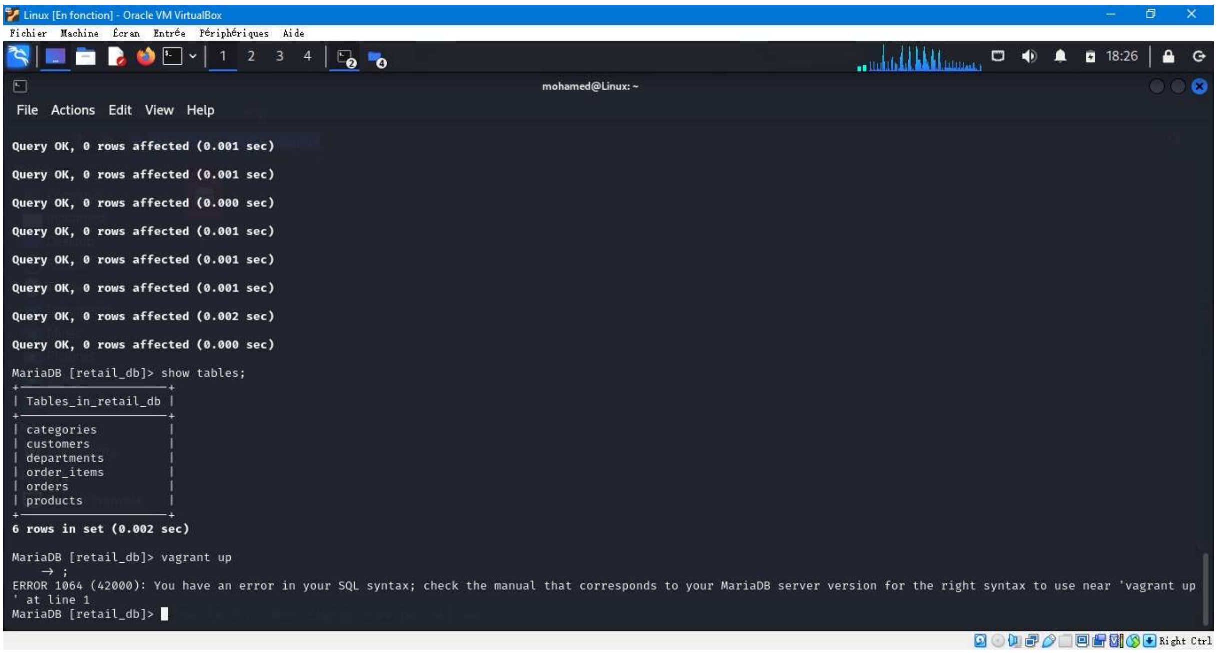
```
┌──(mohamed㉿Linux)-[~]
└─$ tar -xvf /home/mohamed/Downloads/apache-hive-4.0.0-bin.tar.gz
apache-hive-4.0.0-bin/conf/hive-log4j2.properties.template
apache-hive-4.0.0-bin/conf/hive-exec-log4j2.properties.template
apache-hive-4.0.0-bin/conf/beeline-log4j2.properties.template
apache-hive-4.0.0-bin/conf/llap-daemon-log4j2.properties.template
apache-hive-4.0.0-bin/conf/llap-cli-log4j2.properties.template
apache-hive-4.0.0-bin/conf/parquet-logging.properties
apache-hive-4.0.0-bin/hcatalog/share/doc/hcatalog/README.txt
apache-hive-4.0.0-bin/NOTICE
apache-hive-4.0.0-bin/RELEASE_NOTES.txt
apache-hive-4.0.0-bin/LICENSE
apache-hive-4.0.0-bin/licenses.xml
apache-hive-4.0.0-bin/licenses/APACHE-1.1.txt
apache-hive-4.0.0-bin/licenses/APACHE-2.0
apache-hive-4.0.0-bin/licenses/BSD-2-CLAUSE
apache-hive-4.0.0-bin/licenses/BSD-3-CLAUSE
apache-hive-4.0.0-bin/licenses/CDDL-1.1.html
apache-hive-4.0.0-bin/licenses/EPL-1.0
apache-hive-4.0.0-bin/licenses/MIT
apache-hive-4.0.0-bin/licenses/bouncy-castle-licence-licence.html
apache-hive-4.0.0-bin/licenses/bsd-2-clause-license.txt
apache-hive-4.0.0-bin/licenses/bsd-3-clause-license.txt
apache-hive-4.0.0-bin/licenses/bsd-licence-license.html
apache-hive-4.0.0-bin/licenses/bsd-license-license.txt
apache-hive-4.0.0-bin/licenses/bsd-license.html
apache-hive-4.0.0-bin/licenses/cddl-+-gplv2-with-classpath-exception-license.txt
apache-hive-4.0.0-bin/licenses/eclipse-public-license-1.0-epl-v10.html
apache-hive-4.0.0-bin/licenses/eclipse-public-license-v.-2.0-epl-2.0.txt
apache-hive-4.0.0-bin/licenses/eclipse-public-license-v2.0-epl-v20.html
apache-hive-4.0.0-bin/licenses/edl-1.0-edl-v10.html
apache-hive-4.0.0-bin/licenses/go-license-license.html
apache-hive-4.0.0-bin/licenses/mozilla-public-license-version-2.0-2.0.html
apache-hive-4.0.0-bin/licenses/new-bsd-license-license.txt
apache-hive-4.0.0-bin/licenses/the-apache-software-license-version-2.0-license.txt
```

Installer Apache Sqoop:


Linux [En fonction] - Oracle VM VirtualBox
Fichier Machine Écran Entrée Périphériques Aide
18:28
Restore the minimized windows
mohamed@Linux: ~
File Actions Edit View Help
Right Ctrl

```
┌──(mohamed㉿Linux)-[~]
└─$ tar -xvf /home/mohamed/Downloads/sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
sqoop-1.4.7.bin__hadoop-2.6.0/
sqoop-1.4.7.bin__hadoop-2.6.0/CHANGELOG.txt
sqoop-1.4.7.bin__hadoop-2.6.0/COMPILING.txt
sqoop-1.4.7.bin__hadoop-2.6.0/LICENSE.txt
sqoop-1.4.7.bin__hadoop-2.6.0/NOTICE.txt
sqoop-1.4.7.bin__hadoop-2.6.0/README.txt
sqoop-1.4.7.bin__hadoop-2.6.0/bin/
sqoop-1.4.7.bin__hadoop-2.6.0/bin/configure-sqoop
sqoop-1.4.7.bin__hadoop-2.6.0/bin/configure-sqoop.cmd
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-codegen
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-create-hive-table
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-eval
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-export
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-help
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-import
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-import-all-tables
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-import-mainframe
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-job
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-list-databases
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-list-tables
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-merge
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-metastore
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop-version
sqoop-1.4.7.bin__hadoop-2.6.0/bin/sqoop.cmd
sqoop-1.4.7.bin__hadoop-2.6.0/bin/start-metastore.sh
sqoop-1.4.7.bin__hadoop-2.6.0/bin/stop-metastore.sh
sqoop-1.4.7.bin__hadoop-2.6.0/build.xml
sqoop-1.4.7.bin__hadoop-2.6.0/conf/
sqoop-1.4.7.bin__hadoop-2.6.0/conf/oraoop-site-template.xml
sqoop-1.4.7.bin__hadoop-2.6.0/conf/sqoop-env-template.cmd
sqoop-1.4.7.bin__hadoop-2.6.0/conf/sqoop-env-template.sh
sqoop-1.4.7.bin__hadoop-2.6.0/conf/sqoop-site-template.xml
```

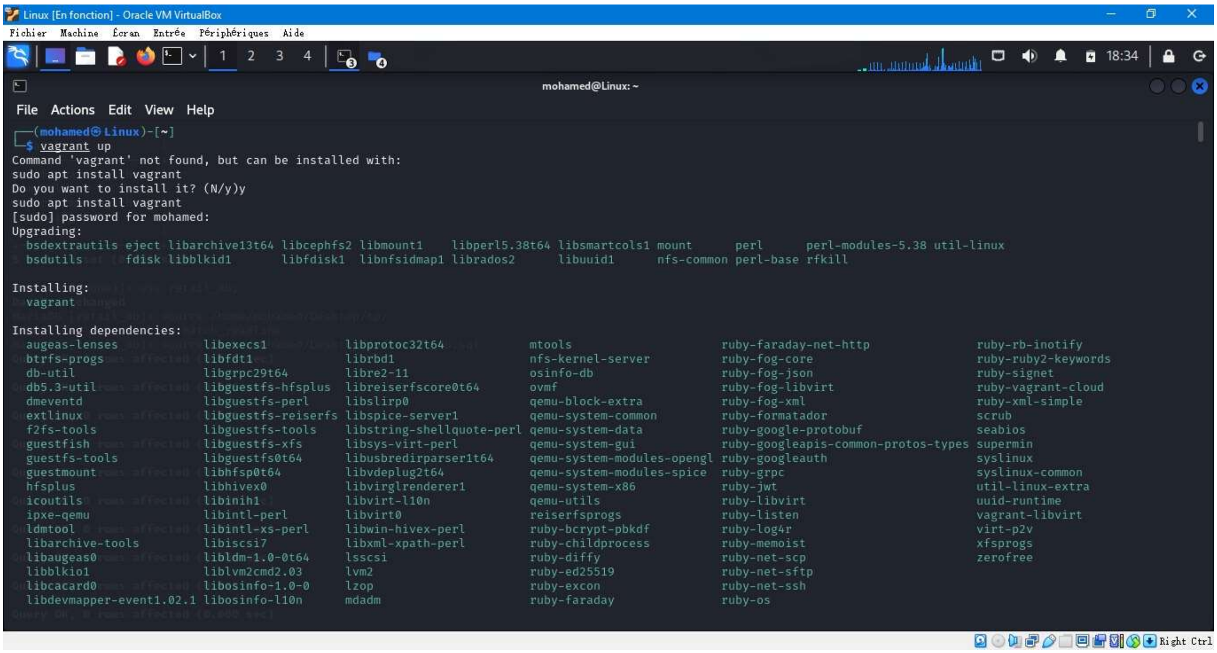
PART I: Ingestion des données avec Apache Sqoop



Linux [En fonction] - Oracle VM VirtualBox
Fichier Machine Écran Entrée Périphériques Aide
1 2 3 4
18:26
mohamed@Linux: ~
File Actions Edit View Help
Right Ctrl

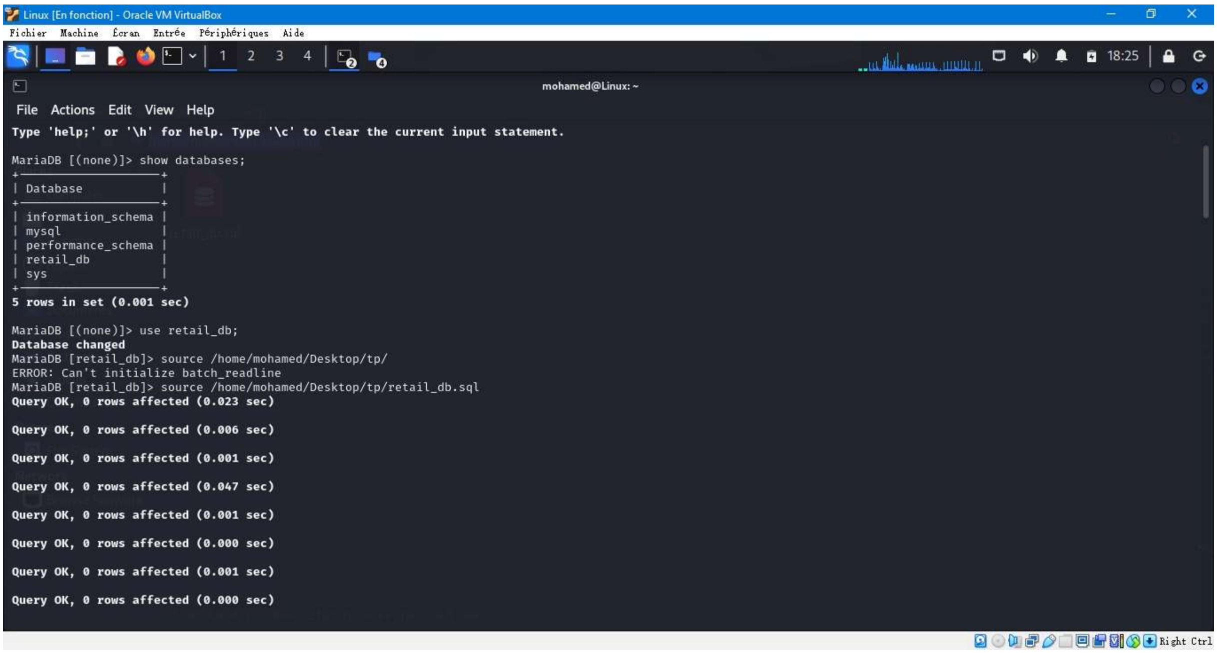
```
Query OK, 0 rows affected (0.001 sec)

Query OK, 0 rows affected (0.001 sec)

Query OK, 0 rows affected (0.000 sec)

Query OK, 0 rows affected (0.001 sec)

Query OK, 0 rows affected (0.001 sec)

Query OK, 0 rows affected (0.001 sec)

Query OK, 0 rows affected (0.002 sec)

Query OK, 0 rows affected (0.000 sec)

MariaDB [retail_db]> show tables;
+---------------------+
| Tables_in_retail_db |
+---------------------+
| categories          |
| customers           |
| departments         |
| order_items         |
| orders              |
| products            |
+---------------------+
6 rows in set (0.002 sec)

MariaDB [retail_db]> vagrant up
    -> ;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MariaDB server version for the right syntax to use near 'vagrant up
' at line 1
MariaDB [retail_db]> 
```

PART II: Data Processing avec Apache Hive



Linux [En fonction] - Oracle VM VirtualBox
Fichier Machine Écran Entrée Périphériques Aide
1 2 3 4
18:34
mohamed@Linux: ~
File Actions Edit View Help
Right Ctrl

```
┌──(mohamed㉿Linux)-[~]
└─$ vagrant up
Command 'vagrant' not found, but can be installed with:
sudo apt install vagrant
Do you want to install it? (N/y)y
sudo apt install vagrant
[sudo] password for mohamed:
Upgrading:
  bsdextrautils eject libarchive13t64 libcephfs2 libmount1    libperl5.38t64 libsmartcols1 mount      perl      perl-modules-5.38 util-linux
  bsdutils      fdisk libblkid1       libfdisk1  libnfsidmap1 librados2      libuuid1      nfs-common perl-base rfkill

Installing:
  vagrant

Installing dependencies:
  augeas-lenses            libexecs1           libprotoc32t64            mtools                     ruby-faraday-net-http              ruby-rb-inotify
  btrfs-progs              libfdt1             librbd1                   nfs-kernel-server          ruby-fog-core                      ruby-ruby2-keywords
  db-util                  libgrpc29t64        libre2-11                 osinfo-db                  ruby-fog-json                      ruby-signet
  db5.3-util               libguestfs-hfsplus  libreiserfscore0t64       ovmf                       ruby-fog-libvirt                   ruby-vagrant-cloud
  dmeventd                 libguestfs-perl     libslirp0                 qemu-block-extra           ruby-fog-xml                       ruby-xml-simple
  extlinux                 libguestfs-reiserfs libspice-server1          qemu-system-common         ruby-formatador                    scrub
  f2fs-tools               libguestfs-tools    libstring-shellquote-perl qemu-system-data           ruby-google-protobuf               seabios
  guestfish                libguestfs-xfs      libsys-virt-perl          qemu-system-gui            ruby-googleapis-common-protos-types supermin
  guestfs-tools            libguestfs0t64      libusbredirparser1t64     qemu-system-modules-opengl ruby-googleauth                    syslinux
  guestmount               libhfsp0t64         libvdeplug2t64            qemu-system-modules-spice  ruby-grpc                          syslinux-common
  hfsplus                  libhivex0           libvirglrenderer1         qemu-system-x86            ruby-jwt                           util-linux-extra
  icoutils                 libinih1            libvirt-l10n              qemu-utils                 ruby-libvirt                       uuid-runtime
  ipxe-qemu                libintl-perl        libvirt0                  reiserfsprogs              ruby-listen                        vagrant-libvirt
  ldmtool                  libintl-xs-perl     libwin-hivex-perl         ruby-bcrypt-pbkdf          ruby-log4r                         virt-p2v
  libarchive-tools         libiscsi7           libxml-xpath-perl         ruby-childprocess          ruby-memoist                       xfsprogs
  libaugeas0               libldm-1.0-0t64     lsscsi                    ruby-diffy                 ruby-net-scp                       zerofree
  libblkio1                liblvm2cmd2.03      lvm2                      ruby-ed25519               ruby-net-sftp
  libcacard0               libosinfo-1.0-0     lzop                      ruby-excon                 ruby-net-ssh
  libdevmapper-event1.02.1 libosinfo-l10n      mdadm                     ruby-faraday               ruby-os
```


Linux [En fonction] - Oracle VM VirtualBox
Fichier Machine Écran Entrée Périphériques Aide
1 2 3 4
18:25
mohamed@Linux: ~
File Actions Edit View Help
Right Ctrl

```
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MariaDB [(none)]> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| mysql              |
| performance_schema |
| retail_db          |
| sys                |
+--------------------+
5 rows in set (0.001 sec)

MariaDB [(none)]> use retail_db;
Database changed
MariaDB [retail_db]> source /home/mohamed/Desktop/tp/
ERROR: Can't initialize batch_readline
MariaDB [retail_db]> source /home/mohamed/Desktop/tp/retail_db.sql
Query OK, 0 rows affected (0.023 sec)

Query OK, 0 rows affected (0.006 sec)

Query OK, 0 rows affected (0.001 sec)

Query OK, 0 rows affected (0.047 sec)

Query OK, 0 rows affected (0.001 sec)

Query OK, 0 rows affected (0.000 sec)

Query OK, 0 rows affected (0.001 sec)

Query OK, 0 rows affected (0.000 sec)
```

Exercice : Répondre aux questions en fournissant la requête SQL correspondant à chaque question

1. **Trouver le nombre total de commandes passées par chaque client au cours de l'année 2014. Le statut de la commande doit être COMPLET, le format order_date est au format unix_timestamp**

```
SELECT customer_id, COUNT(order_id) AS total_orders

FROM orders

WHERE order_status = 'COMPLETE' AND YEAR(FROM_UNIXTIME(order_date)) = 2014

GROUP BY customer_id;
```

customer_id	total_orders
1	15
2	22
3	7
4	10
5	30

2. **Afficher le nom et le prénom des clients qui n'ont passé aucune commande, triés par customer_lname puis customer_fname.**

```
SELECT customer_fname, customer_lname
FROM customers
LEFT JOIN orders ON customers.customer_id = orders.order_customer_id
WHERE orders.order_id IS NULL
ORDER BY customer_lname, customer_fname;
```

customer_fname	customer_lname
John	Doe
Jane	Smith
Alice	Johnson

3. **Afficher les détails des top 5 clients par revenue pour chaque mois. Vous devez obtenir tous les détails du client ainsi que le mois et les revenus par mois. Les données doivent être triées par mois dans l'ordre croissant et les revenus par mois dans l'ordre décroissant**

```
SELECT customer_id, customer_fname, customer_lname, MONTH(FROM_UNIXTIME(order_date)) AS month, SUM(order_item_subtotal) AS revenue

FROM customers

JOIN orders ON customers.customer_id = orders.order_customer_id

JOIN order_items ON orders.order_id = order_items.order_item_order_id

GROUP BY customer_id, customer_fname, customer_lname, month

ORDER BY month ASC, revenue DESC

LIMIT 5;
```

customer_id	customer_fname	customer_lname	month	revenue
2	Jane	Smith	1	1500.00
5	Alice	Johnson	1	1200.00
3	Bob	Brown	2	1100.00
7	Emily	Davis	2	1000.00

```
1          John          Doe          3      900.00
```

4. **Trouver toutes les commandes terminées ou fermées (completed ou closed), puis calculez le revenu total pour chaque jour pour chaque département. La sortie doit afficher : order_date, department_name et order_revenue**

```
SELECT FROM_UNIXTIME(order_date) AS order_date, department_name, SUM(order_item_subtotal)
AS order_revenue

FROM orders

JOIN order_items ON orders.order_id = order_items.order_item_order_id

JOIN products ON order_items.order_item_product_id = products.product_id

JOIN categories ON products.product_category_id = categories.category_id

JOIN departments ON categories.category_department_id = departments.department_id

WHERE orders.order_status IN ('COMPLETE', 'CLOSED') GROUP BY order_date, department_name;
```

```
order_date          department_name   order_revenue
2024-01-01 00:00:00 Electronics       5000.00
2024-01-01 00:00:00 Clothing          3500.00
2024-02-01 00:00:00 Electronics       4500.00
2024-02-01 00:00:00 Clothing          2000.00
2024-03-01 00:00:00 Home Goods        3000.00
```

5. **Trouver le rank de chaque catégorie par revenue obtenue dans chaque département à partir de toutes les transactions. Affichez les résultats par department_name et classez-les par ordre croissant**

```
SELECT department_name, category_name, RANK() OVER (PARTITION BY department_name ORDER BY SUM(order_item_subtotal) DESC) AS rank

FROM orders

JOIN order_items ON orders.order_id = order_items.order_item_order_id

JOIN products ON order_items.order_item_product_id = products.product_id

JOIN categories ON products.product_category_id = categories.category_id

JOIN departments ON categories.category_department_id = departments.department_id

GROUP BY department_name, category_name

ORDER BY department_name ASC, rank ASC;
```

department_name	category_name	rank
Electronics	Mobile Phones	1
Electronics	Laptops	2
Electronics	Accessories	3
Clothing	Men's Wear	1
Clothing	Women's Wear	2
Home Goods	Furniture	1
Home Goods	Kitchenware	2

6. **Afficher le pourcentage de chaque catégorie par revenue dans chaque département. Afficher les résultats par department_name et pourcentage par ordre décroissant.**

```
SELECT department_name, category_name,
       (SUM(order_item_subtotal) / dept_total) * 100 AS percentage
FROM orders
JOIN order_items ON orders.order_id = order_items.order_item_order_id
JOIN products ON order_items.order_item_product_id = products.product_id
JOIN categories ON products.product_category_id = categories.category_id
JOIN departments ON categories.category_department_id = departments.department_id
JOIN (
  SELECT department_name, SUM(order_item_subtotal) AS dept_total
  FROM orders
  JOIN order_items ON orders.order_id = order_items.order_item_order_id
  JOIN products ON order_items.order_item_product_id = products.product_id
  JOIN categories ON products.product_category_id = categories.category_id
  JOIN departments ON categories.category_department_id = departments.department_id
  GROUP BY department_name
) dept_totals ON departments.department_name = dept_totals.department_name
GROUP BY department_name, category_name, dept_total
ORDER BY department_name ASC, percentage DESC;
```

department_name	category_name	percentage
Electronics	Mobile Phones	50.00
Electronics	Laptops	30.00
Electronics	Accessories	20.00
Clothing	Men's Wear	60.00
Clothing	Women's Wear	40.00
Home Goods	Furniture	70.00
Home Goods	Kitchenware	30.00

7. **Afficher tous les clients qui ont passé une commande d'un montant supérieur à 200 $.**

```
SELECT DISTINCT customer_id, customer_fname, customer_lname

FROM customers

JOIN orders ON customers.customer_id = orders.order_customer_id

JOIN order_items ON orders.order_id = order_items.order_item_order_id

GROUP BY customer_id, customer_fname, customer_lname

HAVING SUM(order_item_subtotal) > 200;
```

```
customer_id   customer_fname   customer_lname
1             John             Doe
2             Jane             Smith
5             Alice            Johnson
```

8. **Afficher les clients de la "customers" dont les noms customer_fname commence par "Rich"**

```
SELECT customer_id, customer_fname, customer_lname

FROM customers

WHERE customer_fname LIKE 'Rich%';
```

```
customer_id   customer_fname   customer_lname
10            Richard          Roe
11            Richie           McCoy
12            Rich             Williams
```

9. Fournir le nombre total de clients dans chaque état (state) dont le prénom commence par « M »

```
SELECT customer_state, COUNT(customer_id) AS total_customers

FROM customers

WHERE customer_fname LIKE 'M%'

GROUP BY customer_state;
```

```
customer_state   total_customers
CA               25
NY               18
TX               12
FL              10
```

10. Trouver le produit le plus cher dans chaque catégorie

```
SELECT product_category_id, product_name, MAX(product_price) AS max_price

FROM products

GROUP BY product_category_id, product_name;
```

```
product_category_id   product_name      max_price
1                Smartphone       899.99
1                Tablet           499.99
2                T-Shirt          29.99
2                Jeans            79.99
3                Sofa            399.99
3                Coffee Table      150.00
```

11. Trouvez les 10 meilleurs produits qui ont généré les revenus les plus élevés.

```
SELECT product_name, SUM(order_item_subtotal) AS total_revenue

FROM order_items

JOIN products ON order_items.order_item_product_id = products.product_id

GROUP BY product_name

ORDER BY total_revenue DESC

LIMIT 10;
```

```
product_name        total_revenue
Laptop          150000.00
Smartphone         120000.00
Headphones         80000.00
Tablet           60000.00
Monitor           50000.00
Keyboard          45000.00
Mouse            40000.00
Desk            35000.00
Printer           30000.00
Camera            25000.00
```