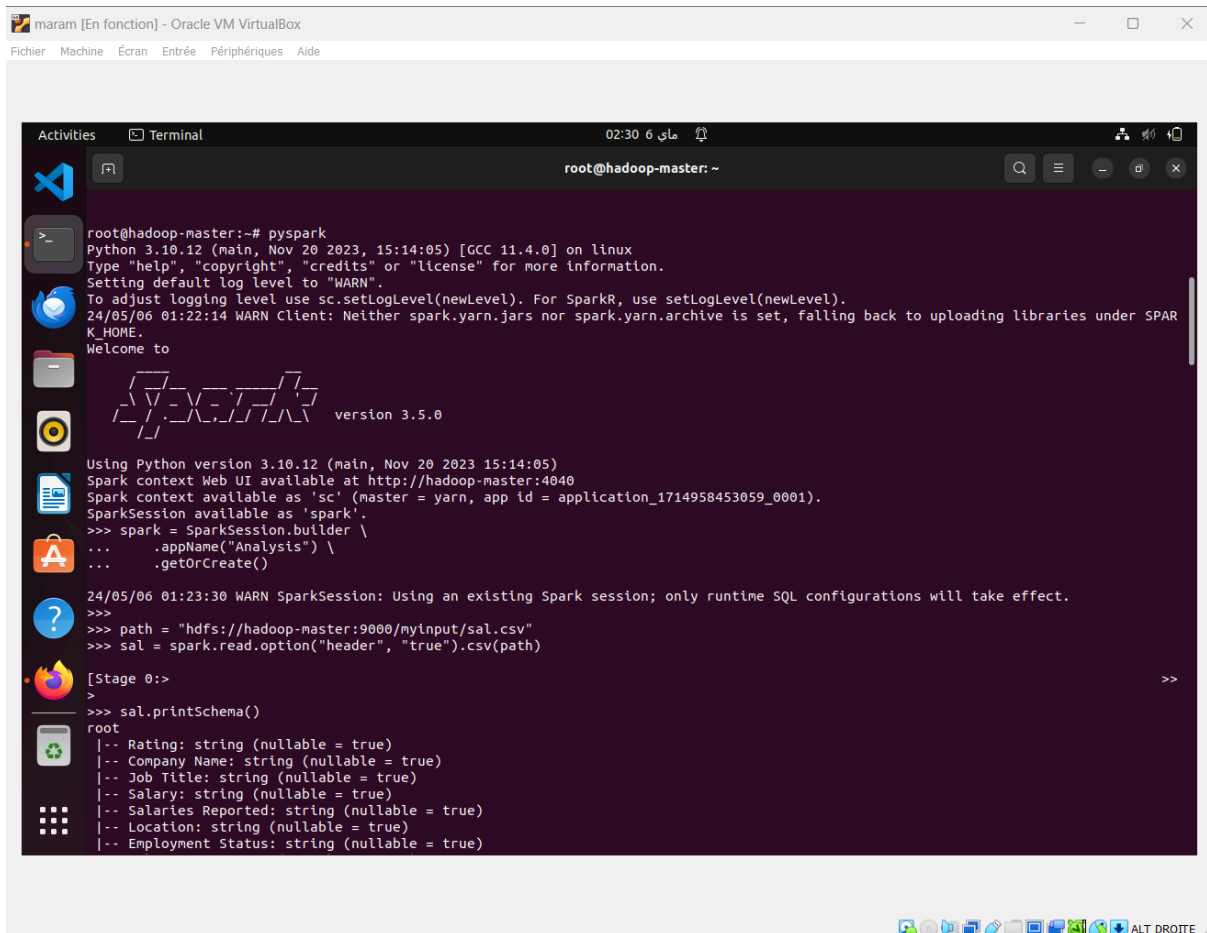


Code SPARK

1/ code scala:

1. Charger le fichier CSV et afficher le schéma des données :

```
val path = "hdfs://hadoop-master:9000/myinput/sal.csv"
val sal = spark.read.option("header", "true").csv(path)
sal.printSchema()
```



The screenshot shows a terminal window titled 'maram [En fonction] - Oracle VM VirtualBox'. The terminal is running a Spark shell (pyspark) on a machine named 'root@hadoop-master'. The output shows the Spark version (3.5.0) and the Python version (3.10.12). The user enters the following commands:

```
root@hadoop-master:~# pyspark
Python 3.10.12 (main, Nov 20 2023, 15:14:05) [GCC 11.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/05/06 01:22:14 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
Welcome to
      _/ _ \| | | | _/_/
     / _ \| |_| | | | |
    / ___ \| | | | | | |
   /_/   \_\_|_|_|_|_|_|
version 3.5.0

Using Python version 3.10.12 (main, Nov 20 2023 15:14:05)
Spark context Web UI available at http://hadoop-master:4040
Spark context available as 'sc' (master = yarn, app id = application_1714958453059_0001).
SparkSession available as 'spark'.
>>> spark = SparkSession.builder \
...     .appName("Analysis") \
...     .getOrCreate()
24/05/06 01:23:30 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
>>>
>>> path = "hdfs://hadoop-master:9000/myinput/sal.csv"
>>> sal = spark.read.option("header", "true").csv(path)

[Stage 0:>
>
>>> sal.printSchema()
root
 |-- Rating: string (nullable = true)
 |-- Company Name: string (nullable = true)
 |-- Job Title: string (nullable = true)
 |-- Salary: string (nullable = true)
 |-- Salaries Reported: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Employment Status: string (nullable = true)
```

description:Cette commande charge le fichier CSV depuis HDFS dans un DataFrame Spark appelé `sal` et affiche le schéma des données, montrant les noms des colonnes et leurs types.

2. Afficher les premières lignes du DataFrame :

```
sal.show()
```

description:Cela affiche les premières lignes du DataFrame `sal`.

```
maram [En fonction] - Oracle VM VirtualBox
Fichier Machine Écran Entrée Périphériques Aide

Activities Terminal 02:30 6 ماي
root@hadoop-master: ~

[Stage 0:~>
>
>>> sal.printSchema()
root
|-- Rating: string (nullable = true)
|-- Company Name: string (nullable = true)
|-- Job Title: string (nullable = true)
|-- Salary: string (nullable = true)
|-- Salaries Reported: string (nullable = true)
|-- Location: string (nullable = true)
|-- Employment Status: string (nullable = true)
|-- Job Roles: string (nullable = true)

>>> sal.show()
+-----+-----+-----+-----+-----+-----+-----+-----+
|Rating|      Company Name|      Job Title| Salary|Salaries Reported| Location|Employment Status|Job Roles|
+-----+-----+-----+-----+-----+-----+-----+-----+
| 3.8|      Saksen|      Android Developer| 400000|          3|Bangalore|      Full Time|      Android|
| 4.5|Advanced Millenni...|      Android Developer| 400000|          3|Bangalore|      Full Time|      Android|
| 4.0|      Unacademy|      Android Developer| 1000000|          3|Bangalore|      Full Time|      Android|
| 3.8|      SnapBizz Cloudtech|      Android Developer| 300000|          3|Bangalore|      Full Time|      Android|
| 4.4|Appoids Tech Solu...|      Android Developer| 600000|          3|Bangalore|      Full Time|      Android|
| 4.2|      Freelancer|      Android Developer| 100000|          3|Bangalore|      Full Time|      Android|
| 3.7|      SQUARE N CUBE|      Android Developer| 192000|          3|Bangalore|      Full Time|      Android|
| 3.1|Samsung R&D Insti...|      Android Developer| 400000|          3|Bangalore|      Full Time|      Android|
| 3.7|DXMinds Technologies|      Android Developer| 300000|          3|Bangalore|      Full Time|      Android|
| 3.6|Endeavour Softwar...|      Android Developer| 600000|          3|Bangalore|      Full Time|      Android|
| 3.6|      Craft Silicon|      Android Developer| 300000|          3|Bangalore|      Full Time|      Android|
| 3.9|Baronford & Assoc...|      Android Developer| 240000|          2|Bangalore|      Full Time|      Android|
| 3.7|      Wibmo|      Android Developer| 900000|          2|Bangalore|      Full Time|      Android|
| 4.8|      Retail Pulse|      Android Developer...| 24000|          2|Bangalore|      Intern|      Android|
| 3.9|      Booknyshow|      Android Developer| 600000|          2|Bangalore|      Full Time|      Android|
| 3.9|      Knowledge Flex|      Android Developer| 228000|          2|Bangalore|      Full Time|      Android|
| 3.6|      Novopay Solutions|      Android Developer| 600000|          2|Bangalore|      Full Time|      Android|
| 3.7|      WealthEngine|      Android Developer| 360000|          2|Bangalore|      Full Time|      Android|
| 4.0|      J.P. Morgan|      Android Developer| 1000000|          2|Bangalore|      Full Time|      Android|
| 3.6|      Acviss|      Android Developer| 500000|          2|Bangalore|      Full Time|      Android|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows
```

3. Compter le nombre total de lignes dans le DataFrame :

sal.count()

description:Cette commande compte le nombre total de lignes dans le DataFrame

sal.

```
maram [En fonction] - Oracle VM VirtualBox
Fichier Machine Écran Entrée Périphériques Aide

Activities Terminal 02:30 6 ماي
root@hadoop-master: ~

[Stage 0:~>
| 4.0|      J.P. Morgan|      Android Developer| 1000000|          2|Bangalore|      Full Time|      Android|
| 3.6|      Acviss|      Android Developer| 500000|          2|Bangalore|      Full Time|      Android|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows

>>> sal.count()
[Stage 2:~>
770

>>> sal.describe("Salary", "Salaries Reported").show()
[Stage 3:~>
[Stage 7:~>
+-----+-----+-----+
|summary|      Salary| Salaries Reported|
+-----+-----+-----+
| count|      22770|      22770|
| mean|695387.2112428634|1.8557751427310644|
| stddev|884399.0136761835| 6.823668180958269|
| min|      100000|      1|
| max|      990000|      98|
+-----+-----+-----+

>>> high_salaries = sal.filter(sal["Salary"] > 100000)
>>> high_salaries.show()
+-----+-----+-----+-----+-----+-----+-----+-----+
|Rating|      Company Name|      Job Title| Salary|Salaries Reported| Location|Employment Status|Job Roles|
+-----+-----+-----+-----+-----+-----+-----+-----+
| 3.8|      Saksen|      Android Developer| 400000|          3|Bangalore|      Full Time|      Android|
| 4.5|Advanced Millenni...|      Android Developer| 400000|          3|Bangalore|      Full Time|      Android|
| 4.0|      Unacademy|      Android Developer| 1000000|          3|Bangalore|      Full Time|      Android|
| 3.8|      SnapBizz Cloudtech|      Android Developer| 300000|          3|Bangalore|      Full Time|      Android|
| 4.4|Appoids Tech Solu...|      Android Developer| 600000|          3|Bangalore|      Full Time|      Android|
| 3.7|      SQUARE N CUBE|      Android Developer| 192000|          3|Bangalore|      Full Time|      Android|
| 3.1|Samsung R&D Insti...|      Android Developer| 400000|          3|Bangalore|      Full Time|      Android|
| 3.7|DXMinds Technologies|      Android Developer| 300000|          3|Bangalore|      Full Time|      Android|
| 3.6|Endeavour Softwar...|      Android Developer| 600000|          3|Bangalore|      Full Time|      Android|
| 3.6|      Craft Silicon|      Android Developer| 300000|          3|Bangalore|      Full Time|      Android|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows
```

4. Afficher des statistiques descriptives pour les colonnes "Salary" et "Salaries Reported" :

```
sal.describe("Salary", "Salaries Reported").show()
```

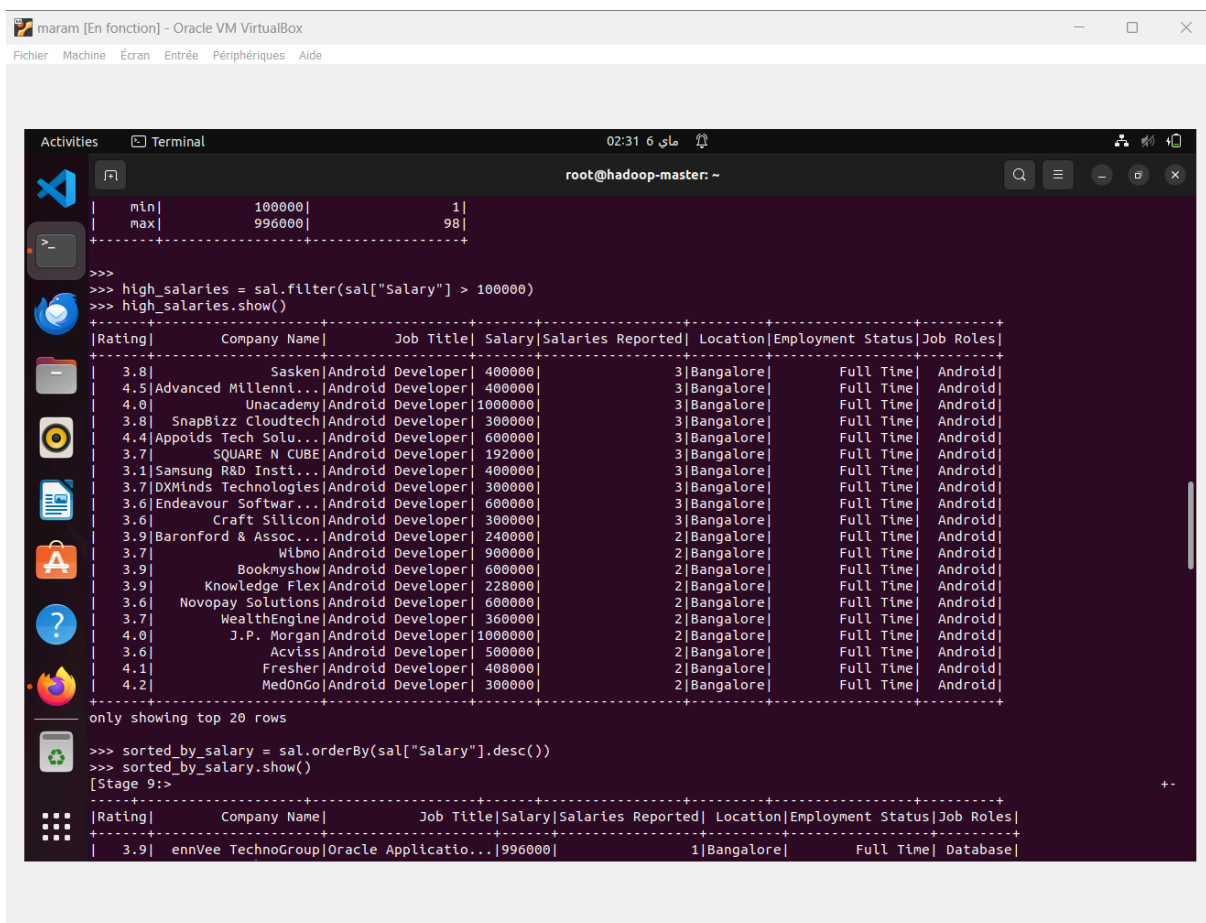
description:Cela affiche des statistiques descriptives telles que la moyenne, l'écart type, le minimum, le maximum, etc., pour les colonnes spécifiées.

5. Filtrer les lignes avec un salaire supérieur à 100 000 :

```
val highSalaries = sal.filter($"Salary" > 100000)
```

```
highSalaries.show()
```

description:Cela filtre les lignes du DataFrame où la valeur de la colonne "Salary" est supérieure à 100 000 et affiche le résultat.



```
maram [En fonction] - Oracle VM VirtualBox
Fichier Machine Écran Entrée Périphériques Aide

Activities Terminal 02:31 6 ماي
root@hadoop-master: ~

min|          100000|          1|
max|          996000|          98|
+-----+-----+-----+
|Rating| Company Name| Job Title|Salary|Salaries Reported| Location|Employment Status|Job Roles|
+-----+-----+-----+-----+-----+-----+-----+-----+
|3.8|   Sasken|Android Developer|400000|          400000|   3|Bangalore|   Full Time|Android|
|4.5|Advanced Millenni...|Android Developer|400000|          400000|   3|Bangalore|   Full Time|Android|
|4.0|   Unacademy|Android Developer|1000000|          1000000|   3|Bangalore|   Full Time|Android|
|3.8|   SnapBizz Cloudtech|Android Developer|300000|          300000|   3|Bangalore|   Full Time|Android|
|4.4|Appoids Tech Solu...|Android Developer|600000|          600000|   3|Bangalore|   Full Time|Android|
|3.7|   SQUARE N CUBE|Android Developer|192000|          192000|   3|Bangalore|   Full Time|Android|
|3.1|Samsung R&D Insti...|Android Developer|400000|          400000|   3|Bangalore|   Full Time|Android|
|3.7|DXMinds Technologies|Android Developer|300000|          300000|   3|Bangalore|   Full Time|Android|
|3.6|Endeavour Softwar...|Android Developer|600000|          600000|   3|Bangalore|   Full Time|Android|
|3.6|   Craft Silicon|Android Developer|300000|          300000|   3|Bangalore|   Full Time|Android|
|3.9|Baronford & Assoc...|Android Developer|240000|          240000|   2|Bangalore|   Full Time|Android|
|3.7|   Wibmo|Android Developer|900000|          900000|   2|Bangalore|   Full Time|Android|
|3.9|   Bookmyshow|Android Developer|600000|          600000|   2|Bangalore|   Full Time|Android|
|3.9|   Knowledge Flex|Android Developer|228000|          228000|   2|Bangalore|   Full Time|Android|
|3.6|   Novopay Solutions|Android Developer|600000|          600000|   2|Bangalore|   Full Time|Android|
|3.7|   HealthEngine|Android Developer|360000|          360000|   2|Bangalore|   Full Time|Android|
|4.0|   J.P. Morgan|Android Developer|1000000|          1000000|   2|Bangalore|   Full Time|Android|
|3.6|   Acviss|Android Developer|500000|          500000|   2|Bangalore|   Full Time|Android|
|4.1|   Fresher|Android Developer|400000|          400000|   2|Bangalore|   Full Time|Android|
|4.2|   MedOnGo|Android Developer|300000|          300000|   2|Bangalore|   Full Time|Android|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows

>>> sorted_by_salary = sal.orderBy(sal["Salary"].desc())
>>> sorted_by_salary.show()
[Stage 9:>]
+-----+-----+-----+-----+-----+-----+-----+-----+
|Rating| Company Name| Job Title|Salary|Salaries Reported| Location|Employment Status|Job Roles|
+-----+-----+-----+-----+-----+-----+-----+-----+
|3.9|   ennVee TechnoGroup|Oracle Applicatio...|996000|          996000|   1|Bangalore|   Full Time|Database|
```

6. Calculer le salaire moyen par entreprise :

```
val avgSalaryByCompany = sal.groupBy("Company Name").avg("Salary")
```

```
avgSalaryByCompany.show()
```

```
maram [En fonction] - Oracle VM VirtualBox
Fichier Machine Écran Entrée Périphériques Aide

Activities Terminal 02:10 6 مای
root@hadoop-master: ~

scala> avgSalaryByCompany.show()
+-----+
| Company Name | avg(Salary) |
+-----+
| Dcoder | 398000.0 |
| Launchers World S... | 192000.0 |
| Mobile Apps Company | 200000.0 |
| Aequalis | 516000.0 |
| Digital Minds (In... | 408000.0 |
| YunzyX | 600000.0 |
| Newfold Digital | 924888.8888888889 |
| Fyp | 120000.0 |
| Grey Coconut Designs | 300000.0 |
| Dr. Reddy's | 500000.0 |
| Intertec Systems | 588000.0 |
| Think201 | 120000.0 |
| AskGalore Digital | 120000.0 |
| IBI Group | 700000.0 |
| nagesh patil | 300000.0 |
| Appgram Technologies | 100000.0 |
| GoalsR | 216000.0 |
| Fresher Zones | 308615.3846153846 |
| Teachers College | 228000.0 |
| Blue Ribbon | 168000.0 |
+-----+
only showing top 20 rows

scala> val sortedBySalary = sal.orderBy($"Salary".desc)
sortedBySalary: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Rating: string, Company Name: string ... 6 more fields]

scala> sortedBySalary.show()
+-----+
| Rating | Company Name | Job Title | Salary | Salaries Reported | Location | Employment Status | Job Roles |
+-----+
| 3.9 | ennVee TechnoGroup | Oracle Applicatio... | 996000 | 1 | Bangalore | Full Time | Database |
| 4.2 | MakeMyTrip | Software Developm... | 996000 | 4 | Bangalore | Full Time | SDE |
| 4.3 | Esper | Android Engineer | 996000 | 1 | Bangalore | Full Time | Android |
| 4.3 | Iris Software | Senior Android De... | 996000 | 1 | New Delhi | Contractor | Android |
+-----+
```

description:Cela regroupe les données par entreprise et calcule la moyenne du salaire pour chaque entreprise, puis affiche le résultat.

7. Trier le DataFrame par salaire décroissant :

```
val sortedBySalary = sal.orderBy($"Salary".desc)
sortedBySalary.show()
```

description:Cela trie le DataFrame en fonction de la colonne "Salary" de manière décroissante et affiche le résultat.

```

scala> val sortedBySalary = sal.orderBy($"Salary".desc)
sortedBySalary: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Rating: string, Company Name: string ... 6 more fields]

scala> sortedBySalary.show()
+-----+-----+-----+-----+-----+-----+-----+
|Rating|Company Name|Job Title|Salary|Salaries Reported|Location|Employment Status|Job Roles|
+-----+-----+-----+-----+-----+-----+-----+
|3.9|ennVee TechnoGroup|Oracle Applicatio...|996000|1|Bangalore|Full Time|Database|
|4.2|MakeMyTrip|Software Developm...|996000|4|Bangalore|Full Time|SDE|
|4.3|Esper|Android Engineer|996000|1|Bangalore|Full Time|Android|
|4.3|Iris Software|Senior Android De...|996000|1|New Delhi|Contractor|Android|
|4.2|9Logic Technologi...|Android Developer|996000|1|Chennai|Full Time|Android|
|3.8|Ecom Express|Front End Developer|996000|1|New Delhi|Full Time|Frontend|
|3.9|Sociolla|Senior IOS App De...|996000|1|New Delhi|Full Time|IOS|
|5.0|Ramcides|Software Engineer...|996000|1|Hyderabad|Full Time|Java|
|5.0|AB Solutions Lab|Mobile App Developer|996000|1|Bangalore|Full Time|Mobile|
|4.0|Goldman Sachs|Software Developm...|996000|4|Bangalore|Intern|SDE|
|4.4|Microsoft|Software Developm...|996000|1|Chennai|Intern|SDE|
|3.8|Amazon|Software Developm...|996000|1|Hyderabad|Full Time|SDE|
|3.7|US Department of ...|Software Developm...|996000|1|Hyderabad|Full Time|SDE|
|2.7|Data Dimensions|Software Developm...|996000|1|Mumbai|Full Time|SDE|
|4.0|Sira Consulting|Software Developm...|996000|1|New Delhi|Intern|SDE|
|4.0|CrossChannel|Software Developm...|996000|1|Pune|Full Time|SDE|
|3.5|Diamanti|Software Developm...|996000|1|Pune|Full Time|SDE|
|4.4|ThinkBridge|Software Developm...|996000|1|Pune|Full Time|SDE|
|3.7|CDK Global|Test Engineer|996000|1|Hyderabad|Full Time|Testing|
|3.9|Tata Consultancy ...|Tester|996000|8|Hyderabad|Full Time|Testing|
+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows

scala> val jobCountByLocation = sal.groupBy("Location").count()
jobCountByLocation: org.apache.spark.sql.DataFrame = [Location: string, count: bigint]

scala> jobCountByLocation.show()
+-----+-----+
|Location|count|
+-----+-----+
|Bangalore|8264|
+-----+-----+

```

8. Compter le nombre d'emplois par emplacement :

```
val jobCountByLocation = sal.groupBy("Location").count()
jobCountByLocation.show()
```

description:Cela regroupe les données par emplacement et compte le nombre d'emplois pour chaque emplacement, puis affiche le résultat.

The screenshot shows a terminal window titled "root@hadoop-master: ~" with the following content:

```
only showing top 20 rows

scala> val jobCountByLocation = sal.groupBy("Location").count()
jobCountByLocation: org.apache.spark.sql.DataFrame = [Location: string, count: bigint]

scala> jobCountByLocation.show()
+-----+
| Location | count |
+-----+
| Bangalore | 8264 |
| Kerala | 108 |
| Madhya Pradesh | 155 |
| Chennai | 2458 |
| Mumbai | 749 |
| Kolkata | 178 |
| Pune | 2134 |
| New Delhi | 4176 |
| Hyderabad | 4467 |
| Jaipur | 81 |
+-----+

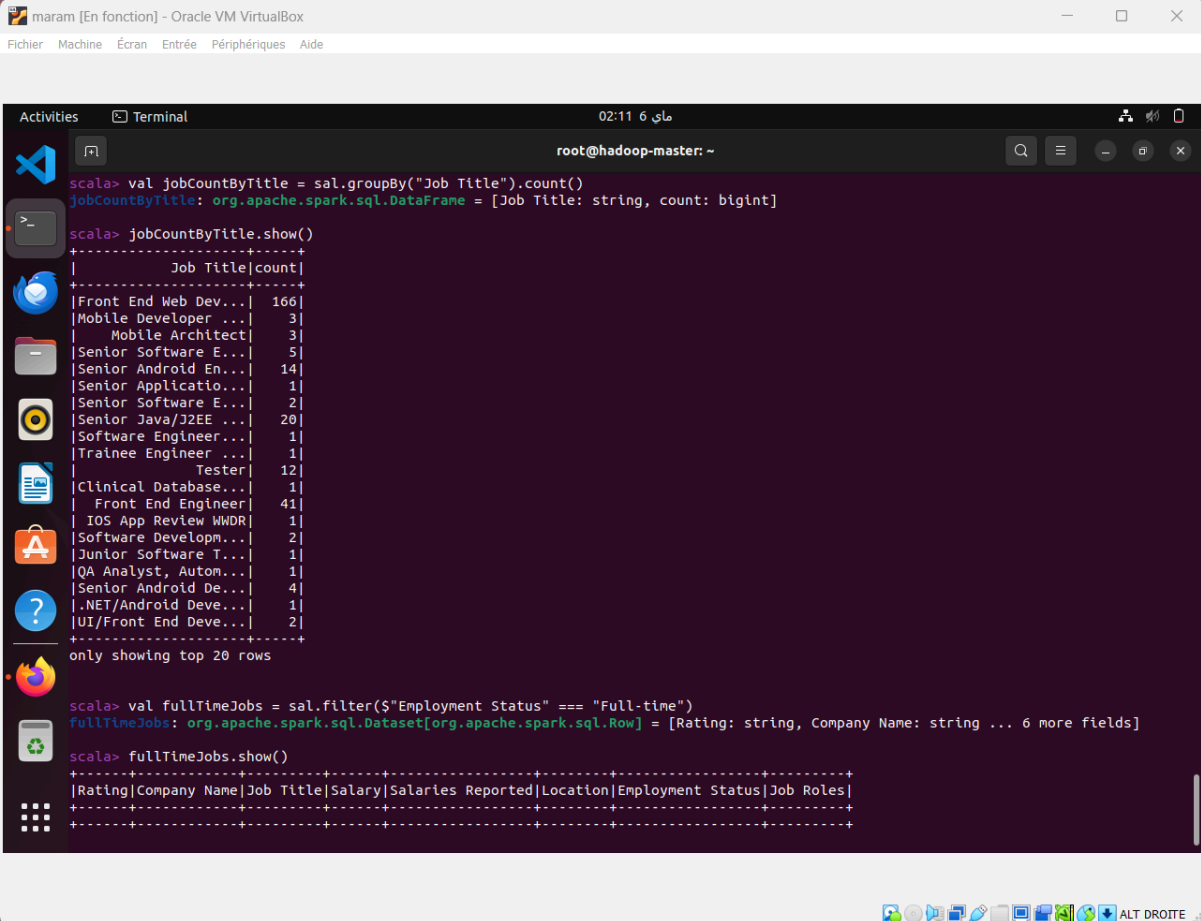
scala> val jobCountByTitle = sal.groupBy("Job Title").count()
jobCountByTitle: org.apache.spark.sql.DataFrame = [Job Title: string, count: bigint]

scala> jobCountByTitle.show()
+-----+
| Job Title | count |
+-----+
| Front End Web Dev... | 166 |
| Mobile Developer ... | 3 |
| Mobile Architect | 3 |
| Senior Software E... | 5 |
| Senior Android En... | 14 |
| Senior Applicatio... | 1 |
| Senior Software E... | 2 |
| Senior Java/J2EE ... | 20 |
| Software Engineer... | 1 |
+-----+
```

9. Compter le nombre d'emplois par titre de poste :

```
val jobCountByTitle = sal.groupBy("Job Title").count()
jobCountByTitle.show()
```

description:Cela regroupe les données par titre de poste et compte le nombre d'emplois pour chaque titre de poste, puis affiche le résultat.



```
maram [En fonction] - Oracle VM VirtualBox
Fichier Machine Écran Entrée Périphériques Aide

Activities Terminal 02:11 6 ماي
root@hadoop-master: ~

scala> val jobCountByTitle = sal.groupBy("Job Title").count()
jobCountByTitle: org.apache.spark.sql.DataFrame = [Job Title: string, count: bigint]

scala> jobCountByTitle.show()
+-----+-----+
| Job Title|count|
+-----+-----+
|Front End Web Dev...| 166|
|Mobile Developer ...| 3|
|Mobile Architect| 3|
|Senior Software E...| 5|
|Senior Android En...| 14|
|Senior Applicatio...| 1|
|Senior Software E...| 2|
|Senior Java/J2EE ...| 20|
|Software Engineer...| 1|
|Trainee Engineer...| 1|
|Tester| 12|
|Clinical Database...| 1|
|Front End Engineer| 41|
|IOS App Review WWDR| 1|
|Software Developm...| 2|
|Junior Software T...| 1|
|QA Analyst, Auton...| 1|
|Senior Android De...| 4|
|.NET/Android Deve...| 1|
|UI/Front End Deve...| 2|
+-----+-----+
only showing top 20 rows

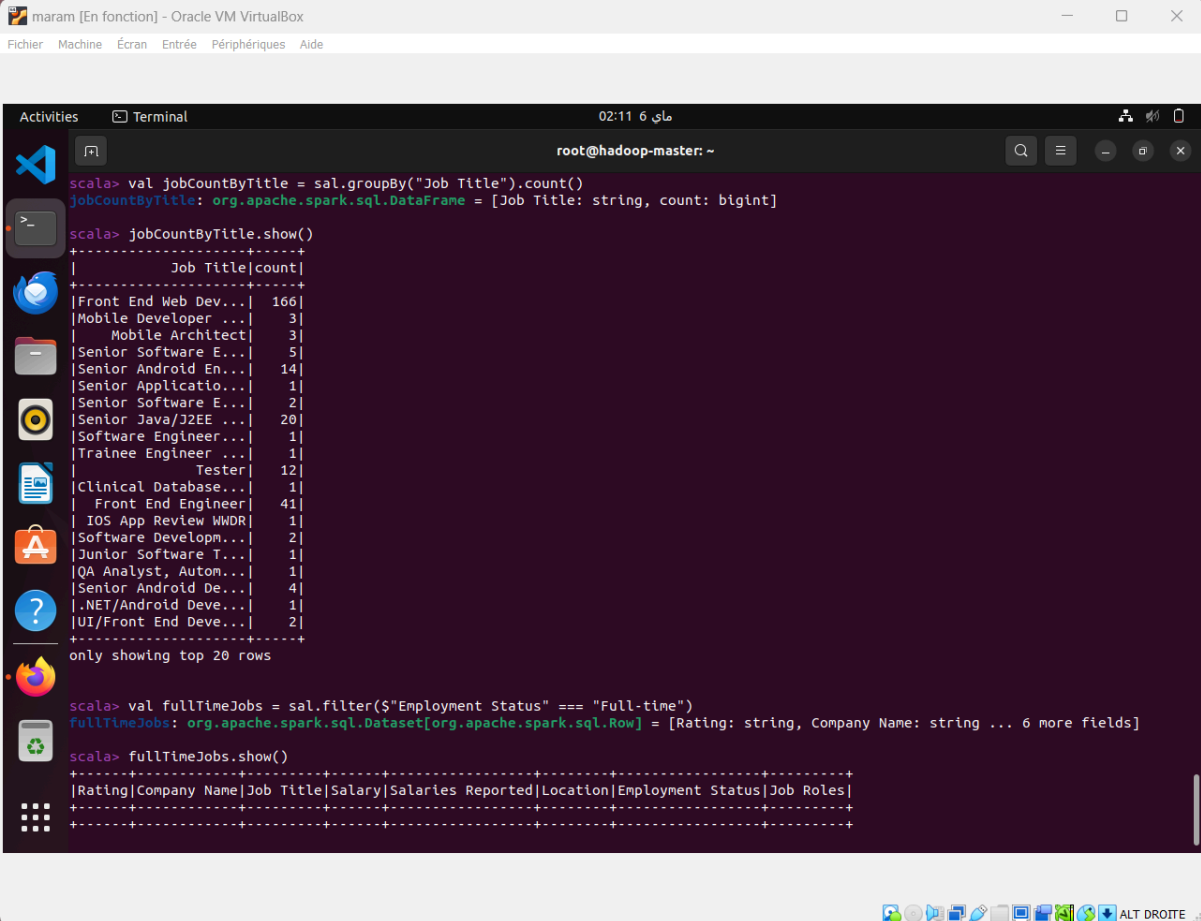
scala> val fullTimeJobs = sal.filter($"Employment Status" === "Full-time")
fullTimeJobs: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Rating: string, Company Name: string ... 6 more fields]

scala> fullTimeJobs.show()
+-----+-----+-----+-----+-----+-----+-----+-----+
|Rating|Company Name|Job Title|Salary|Salaries Reported|Location|Employment Status|Job Roles|
+-----+-----+-----+-----+-----+-----+-----+-----+
|
```

10. Filtrer les emplois à temps plein :

```
val fullTimeJobs = sal.filter($"Employment Status" === "Full-time")
fullTimeJobs.show()
```

description: Cela filtre les emplois à temps plein en sélectionnant les lignes où la valeur de la colonne "Employment Status" est "Full-time", puis affiche le résultat.



```
maram [En fonction] - Oracle VM VirtualBox
Fichier Machine Écran Entrée Périphériques Aide

Activities Terminal 02:11 6 ماي
root@hadoop-master: ~

scala> val jobCountByTitle = sal.groupBy("Job Title").count()
jobCountByTitle: org.apache.spark.sql.DataFrame = [Job Title: string, count: bigint]

scala> jobCountByTitle.show()
+-----+-----+
| Job Title|count|
+-----+-----+
|Front End Web Dev...| 166|
|Mobile Developer ...| 3|
|Mobile Architect...| 3|
|Senior Software E...| 5|
|Senior Android En...| 14|
|Senior Applicatio...| 1|
|Senior Software E...| 2|
|Senior Java/J2EE ...| 20|
|Software Engineer...| 1|
|Trainee Engineer...| 1|
|Tester| 12|
|Clinical Database...| 1|
|Front End Engineer| 41|
|IOS App Review WWRD| 1|
|Software Developm...| 2|
|Junior Software T...| 1|
|QA Analyst, Auton...| 1|
|Senior Android De...| 4|
|.NET/Android Deve...| 1|
|UI/Front End Deve...| 2|
+-----+-----+
only showing top 20 rows

scala> val fullTimeJobs = sal.filter($"Employment Status" === "Full-time")
fullTimeJobs: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Rating: string, Company Name: string ... 6 more fields]

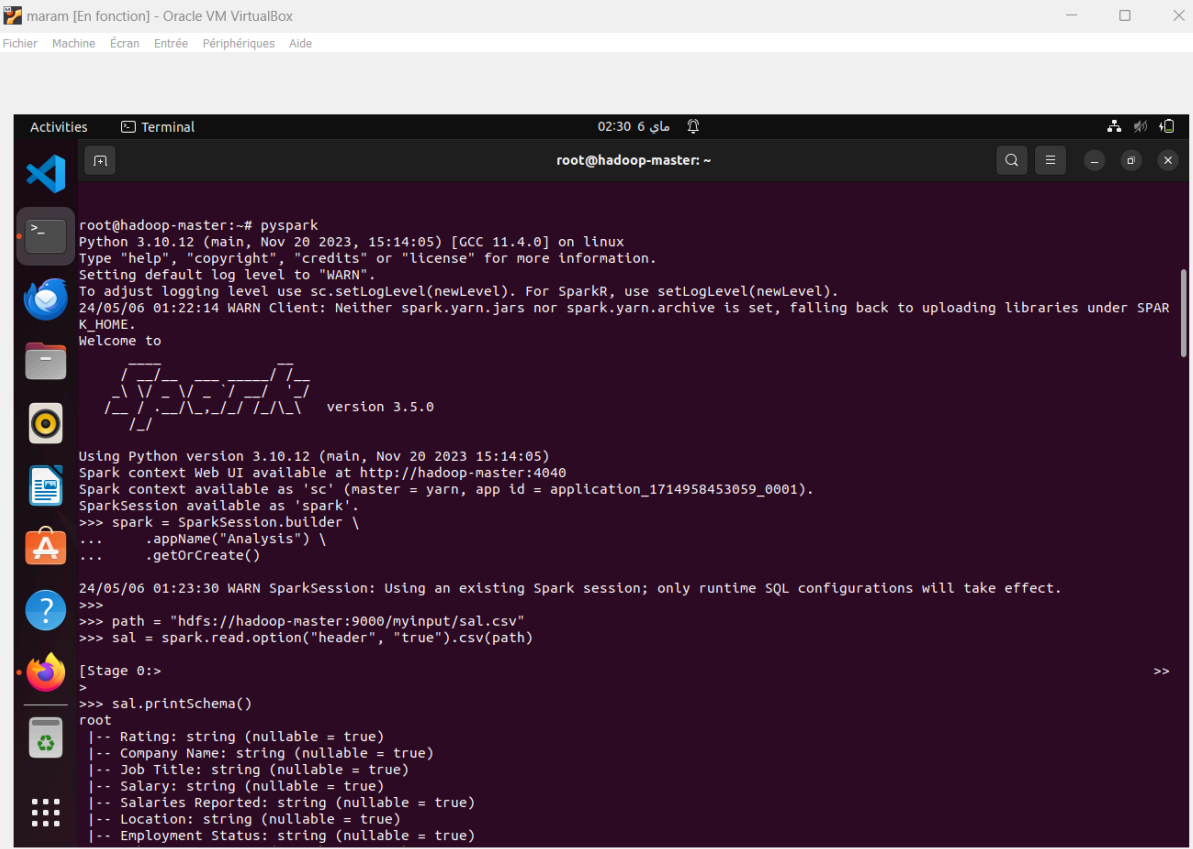
scala> fullTimeJobs.show()
+-----+-----+-----+-----+-----+-----+-----+-----+
|Rating|Company Name|Job Title|Salary|Salaries Reported|Location|Employment Status|Job Roles|
+-----+-----+-----+-----+-----+-----+-----+-----+
|
```

1/ code pyspark:

1. Créer une session Spark et charger le fichier CSV :

```
from pyspark.sql import SparkSession
spark = SparkSession.builder \
    .appName("Analysis") \
    .getOrCreate()
path = "hdfs://hadoop-master:9000/myinput/sal.csv"
sal = spark.read.option("header", "true").csv(path)
```


description : Cela crée une session Spark et charge le fichier CSV depuis HDFS dans un DataFrame Spark appelé `sal`.



```
root@hadoop-master:~# pyspark
Python 3.10.12 (main, Nov 20 2023, 15:14:05) [GCC 11.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/05/06 01:22:14 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
Welcome to

      ____              __
     / __ \__  _/___  /  /
    / /_/ //  /_  __/  /
   / ____//_/ |_/ /_  /
  /_/   /_/ |_/___/_/

version 3.5.0

Using Python version 3.10.12 (main, Nov 20 2023 15:14:05)
Spark context Web UI available at http://hadoop-master:4040
Spark context available as 'sc' (master = yarn, app id = application_1714958453059_0001).
SparkSession available as 'spark'.
>>> spark = SparkSession.builder \
...     .appName("Analysis") \
...     .getOrCreate()

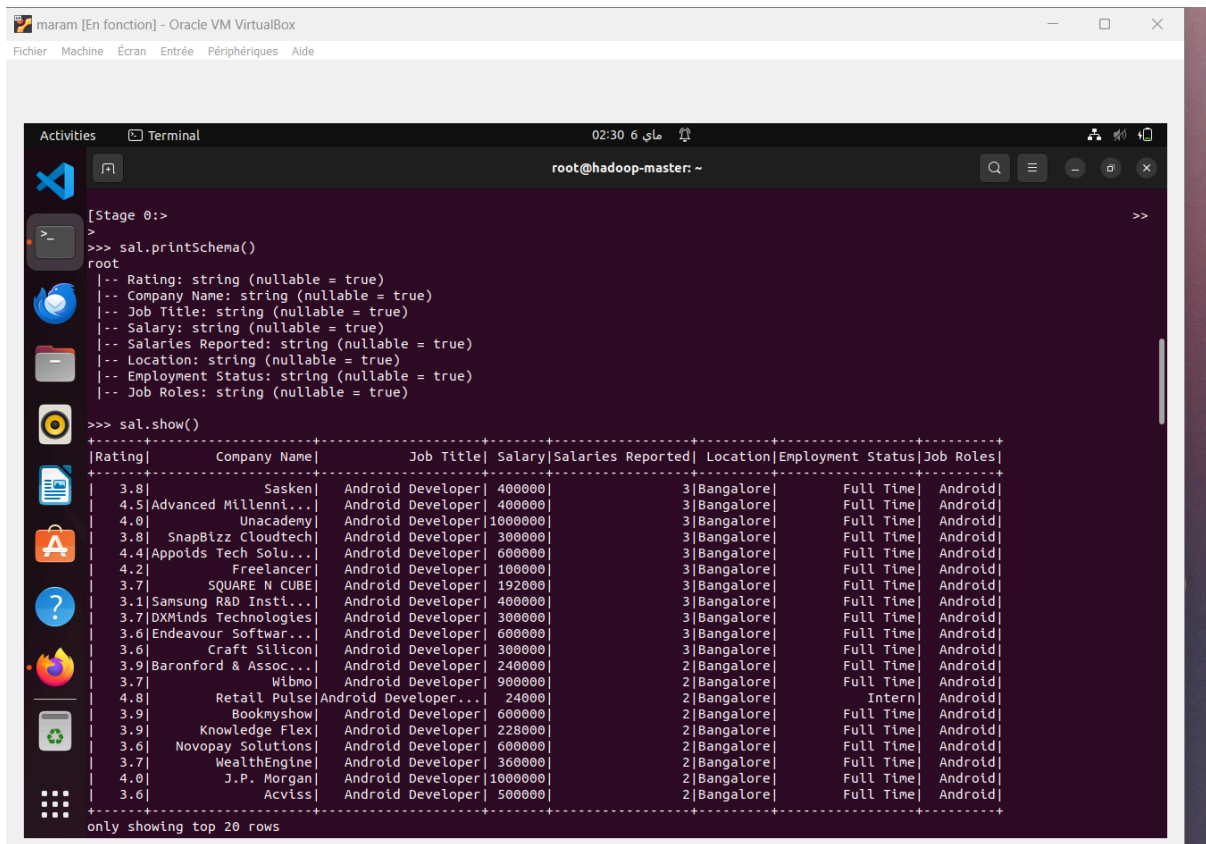
24/05/06 01:23:30 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
>>>
>>> path = "hdfs://hadoop-master:9000/myinput/sal.csv"
>>> sal = spark.read.option("header", "true").csv(path)

[Stage 0:>
>
>>> sal.printSchema()
root
 |-- Rating: string (nullable = true)
 |-- Company Name: string (nullable = true)
 |-- Job Title: string (nullable = true)
 |-- Salary: string (nullable = true)
 |-- Salaries Reported: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Employment Status: string (nullable = true)
```

2. Afficher le schéma des données :

```
sal.printSchema()
```

description : Cette commande affiche le schéma des données du DataFrame `sal`, montrant les noms des colonnes et leurs types.



The screenshot shows a terminal window titled "root@hadoop-master: ~" with the following commands and output:

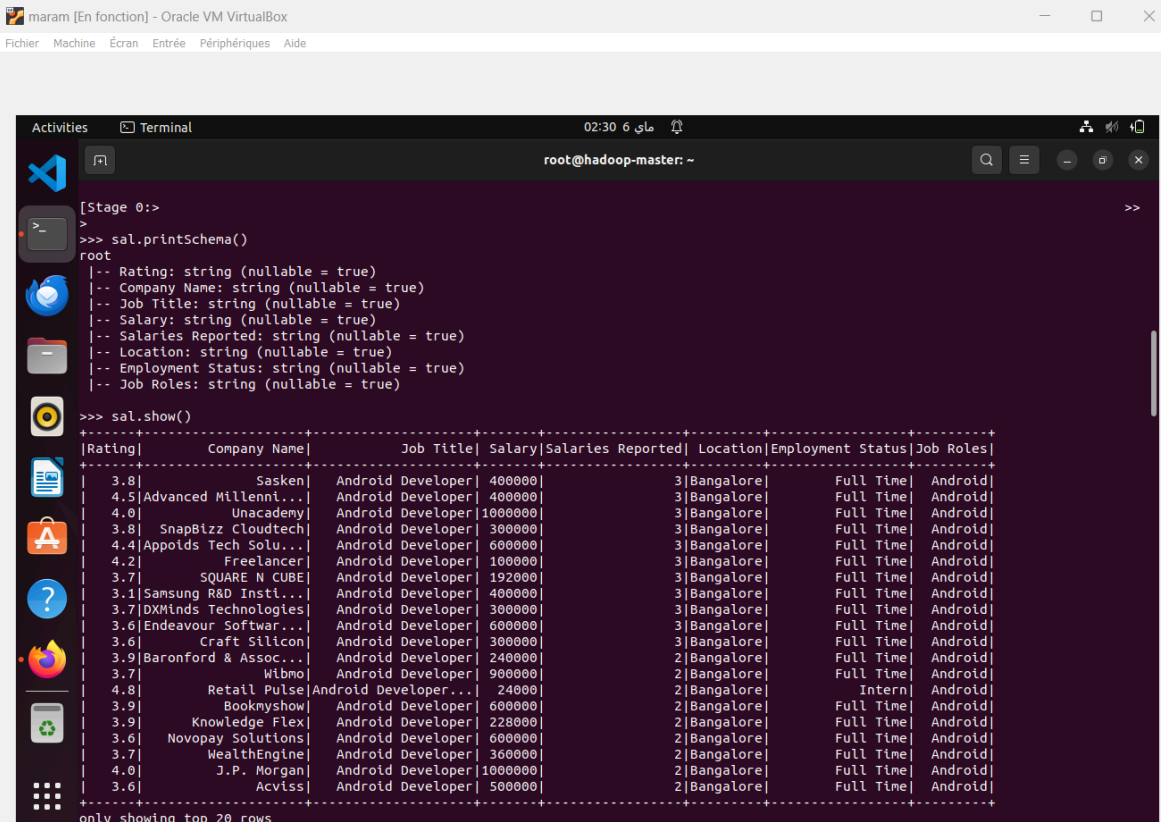
```
[Stage 0:>
>
>>> sal.printSchema()
root
 |-- Rating: string (nullable = true)
 |-- Company Name: string (nullable = true)
 |-- Job Title: string (nullable = true)
 |-- Salary: string (nullable = true)
 |-- Salaries Reported: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Employment Status: string (nullable = true)
 |-- Job Roles: string (nullable = true)

>>> sal.show()
+-----+-----+-----+-----+-----+-----+-----+
|Rating|Company Name|Job Title|Salary|Salaries Reported|Location|Employment Status|Job Roles|
+-----+-----+-----+-----+-----+-----+-----+
|3.8|Sasken|Android Developer|400000|3|Bangalore|Full Time|Android|
|4.5|Advanced Millenni...|Android Developer|400000|3|Bangalore|Full Time|Android|
|4.0|Unacademy|Android Developer|1000000|3|Bangalore|Full Time|Android|
|3.8|SnapBizz Cloudtech|Android Developer|300000|3|Bangalore|Full Time|Android|
|4.4|Appoids Tech Solu...|Android Developer|600000|3|Bangalore|Full Time|Android|
|4.2|Freelancer|Android Developer|100000|3|Bangalore|Full Time|Android|
|3.7|SQUARE N CUBE|Android Developer|192000|3|Bangalore|Full Time|Android|
|3.1|Samsung R&D Insti...|Android Developer|400000|3|Bangalore|Full Time|Android|
|3.7|DXMinds Technologies|Android Developer|300000|3|Bangalore|Full Time|Android|
|3.6|Endeavour Softwar...|Android Developer|600000|3|Bangalore|Full Time|Android|
|3.6|Craft Silicon|Android Developer|300000|3|Bangalore|Full Time|Android|
|3.9|Baronford & Assoc...|Android Developer|240000|2|Bangalore|Full Time|Android|
|3.7|Wibmo|Android Developer|900000|2|Bangalore|Full Time|Android|
|4.8|Retail Pulse|Android Developer...|24000|2|Bangalore|Intern|Android|
|3.9|Bookmyshow|Android Developer|600000|2|Bangalore|Full Time|Android|
|3.9|Knowledge Flex|Android Developer|228000|2|Bangalore|Full Time|Android|
|3.6|Novopay Solutions|Android Developer|600000|2|Bangalore|Full Time|Android|
|3.7|WealthEngine|Android Developer|360000|2|Bangalore|Full Time|Android|
|4.0|J.P. Morgan|Android Developer|1000000|2|Bangalore|Full Time|Android|
|3.6|Acviss|Android Developer|500000|2|Bangalore|Full Time|Android|
+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows
```

3. Afficher les premières lignes du DataFrame :

`sal.show()`

description : Cela affiche les premières lignes du DataFrame `sal`.



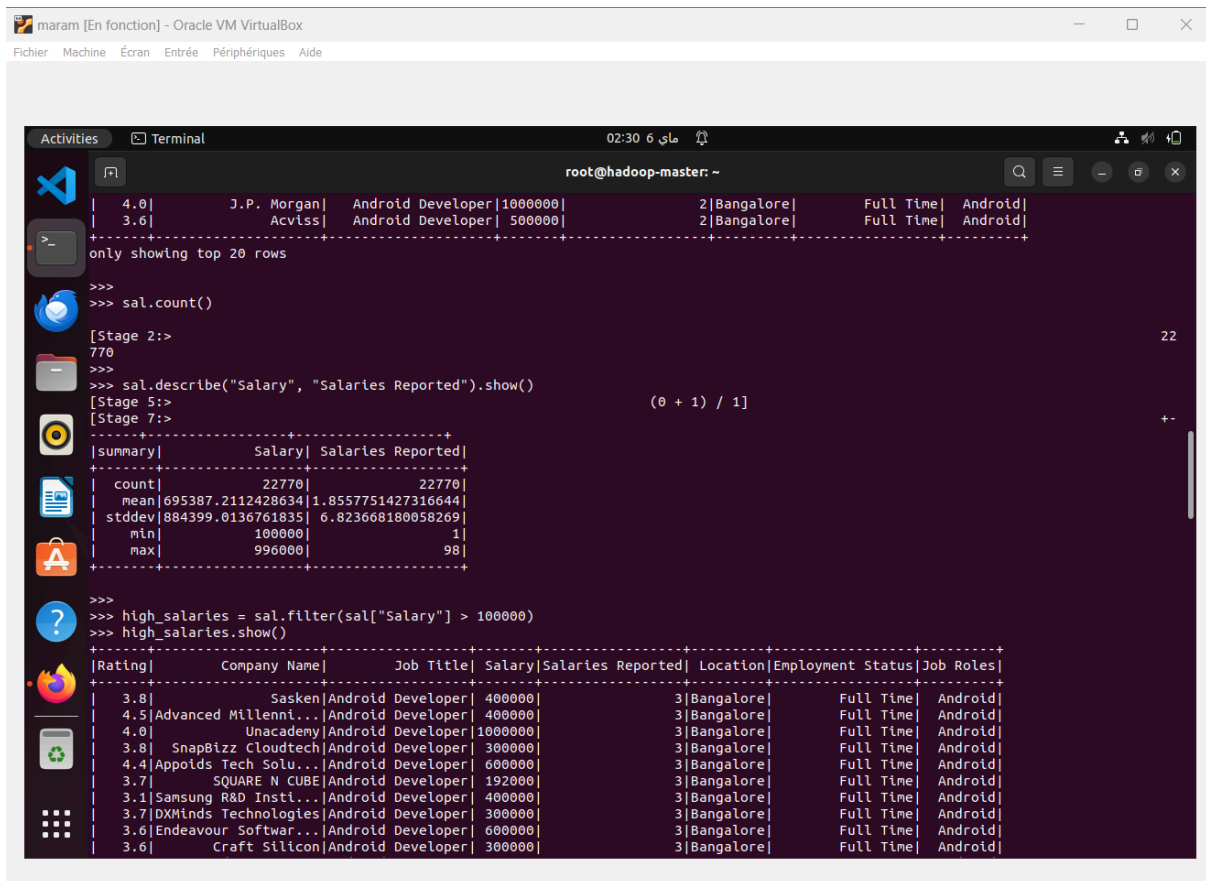
```
[Stage 0:>
>
>>> sal.printSchema()
root
 |-- Rating: string (nullable = true)
 |-- Company Name: string (nullable = true)
 |-- Job Title: string (nullable = true)
 |-- Salary: string (nullable = true)
 |-- Salaries Reported: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Employment Status: string (nullable = true)
 |-- Job Roles: string (nullable = true)

>>> sal.show()
+-----+-----+-----+-----+-----+-----+-----+-----+
|Rating|Company Name|Job Title|Salary|Salaries Reported|Location|Employment Status|Job Roles|
+-----+-----+-----+-----+-----+-----+-----+-----+
|3.8|Sasken|Android Developer|400000|400000|3|Bangalore|Full Time|Android|
|4.5|Advanced Millenni...|Android Developer|400000|400000|3|Bangalore|Full Time|Android|
|4.0|Unacademy|Android Developer|1000000|1000000|3|Bangalore|Full Time|Android|
|3.8|SnapBizz Cloudtech|Android Developer|300000|300000|3|Bangalore|Full Time|Android|
|4.4|Appoids Tech Solu...|Android Developer|600000|600000|3|Bangalore|Full Time|Android|
|4.2|Freelancer|Android Developer|100000|100000|3|Bangalore|Full Time|Android|
|3.7|SQUARE N CUBE|Android Developer|192000|192000|3|Bangalore|Full Time|Android|
|3.1|Samsung R&D Insti...|Android Developer|400000|400000|3|Bangalore|Full Time|Android|
|3.7|DXMinds Technologies|Android Developer|300000|300000|3|Bangalore|Full Time|Android|
|3.6|Endeavour Softwar...|Android Developer|600000|600000|3|Bangalore|Full Time|Android|
|3.6|Craft Silicon|Android Developer|300000|300000|3|Bangalore|Full Time|Android|
|3.9|Baronford & Assoc...|Android Developer|240000|240000|2|Bangalore|Full Time|Android|
|3.7|Wibno|Android Developer|900000|900000|2|Bangalore|Full Time|Android|
|4.8|Retail Pulse|Android Developer...|24000|24000|2|Bangalore|Intern|Android|
|3.9|Bookmyshow|Android Developer|600000|600000|2|Bangalore|Full Time|Android|
|3.9|Knowledge Flex|Android Developer|228000|228000|2|Bangalore|Full Time|Android|
|3.6|Novopay Solutions|Android Developer|600000|600000|2|Bangalore|Full Time|Android|
|3.7|HealthEngine|Android Developer|360000|360000|2|Bangalore|Full Time|Android|
|4.0|J.P. Morgan|Android Developer|1000000|1000000|2|Bangalore|Full Time|Android|
|3.6|Acviss|Android Developer|500000|500000|2|Bangalore|Full Time|Android|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows
```

4. Compter le nombre total de lignes dans le DataFrame :

`sal.count()`

description : Cette commande compte le nombre total de lignes dans le DataFrame `sal`.

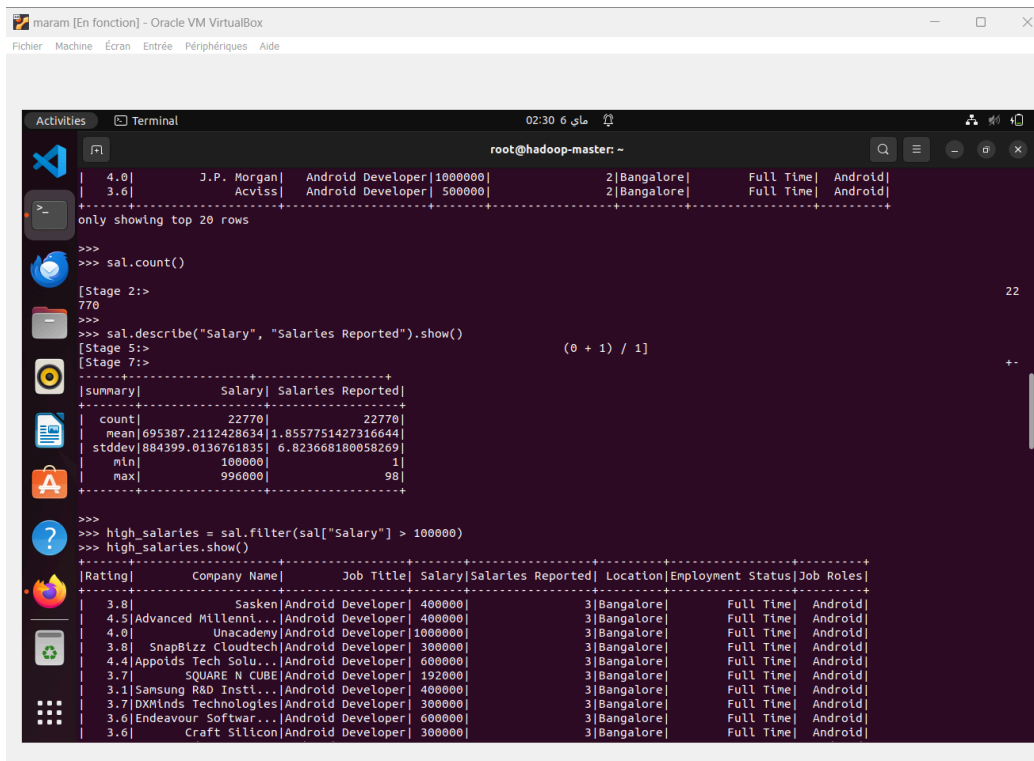


```
root@hadoop-master: ~  
| 4.0| J.P. Morgan| Android Developer|1000000| 2|Bangalore| Full Time| Android|  
| 3.6| Acviss| Android Developer| 500000| 2|Bangalore| Full Time| Android|  
+-----+-----+-----+-----+-----+-----+-----+  
only showing top 20 rows  
>>> sal.count()  
[Stage 2:;> 22  
770  
>>> sal.describe("Salary", "Salaries Reported").show()  
[Stage 5:;> (0 + 1) / 1]  
[Stage 7:;> +  
+-----+-----+-----+  
|summary| Salary| Salaries Reported|  
+-----+-----+-----+  
|count| 22770| 22770|  
|mean|695387.2112428634|1.8557751427316644|  
|stddev|884399.0136761835| 6.823668180058269|  
|min| 100000| 1|  
|max| 996000| 98|  
+-----+-----+-----+  
>>>  
>>> high_salaries = sal.filter(sal["Salary"] > 100000)  
>>> high_salaries.show()  
+-----+-----+-----+-----+-----+-----+-----+  
|Rating| Company Name| Job Title| Salary|Salaries Reported| Location|Employment Status|Job Roles|  
+-----+-----+-----+-----+-----+-----+-----+  
|3.8| Sasken|Android Developer| 400000| 3|Bangalore| Full Time| Android|  
|4.5|Advanced Millenni...|Android Developer| 400000| 3|Bangalore| Full Time| Android|  
|4.0| Unacademy|Android Developer|1000000| 3|Bangalore| Full Time| Android|  
|3.8| SnapBizz Cloudtech|Android Developer| 300000| 3|Bangalore| Full Time| Android|  
|4.4|Appoids Tech Solu...|Android Developer| 600000| 3|Bangalore| Full Time| Android|  
|3.7| SQUARE N CUBE|Android Developer| 192000| 3|Bangalore| Full Time| Android|  
|3.1|Samsung R&D Insti...|Android Developer| 400000| 3|Bangalore| Full Time| Android|  
|3.7|DXMinds Technologies|Android Developer| 300000| 3|Bangalore| Full Time| Android|  
|3.6|Endeavour Softwar...|Android Developer| 600000| 3|Bangalore| Full Time| Android|  
|3.6| Craft Silicon|Android Developer| 300000| 3|Bangalore| Full Time| Android|
```

5. Afficher des statistiques descriptives pour les colonnes "Salary" et "Salaries Reported" :

```
sal.describe("Salary", "Salaries Reported").show()
```

description: Cela affiche des statistiques descriptives telles que la moyenne, l'écart type, le minimum, le maximum, etc., pour les colonnes spécifiées.



```
root@hadoop-master: ~  
| 4.0| J.P. Morgan| Android Developer|1000000| 2|Bangalore| Full Time| Android|  
| 3.6| Acviss| Android Developer| 500000| 2|Bangalore| Full Time| Android|  
+-----+-----+-----+-----+-----+-----+-----+-----+  
only showing top 20 rows  
>>>  
>>> sal.count()  
[Stage 2:> 22  
770  
>>> sal.describe("Salary", "Salaries Reported").show()  
[Stage 5:>  
[Stage 7:> (0 + 1) / 1]  
+-----+-----+-----+-----+-----+-----+-----+-----+  
|summary| Salary| Salaries Reported|  
+-----+-----+-----+-----+-----+-----+-----+-----+  
| count| 22770| 22770|  
| mean|695387.2112428634|1.8557751427316644|  
| stddev|884399.0136761835| 6.823668180058269|  
| min| 100000| 1|  
| max| 996000| 98|  
+-----+-----+-----+-----+-----+-----+-----+-----+  
>>>  
>>> high_salaries = sal.filter(sal["Salary"] > 100000)  
>>> high_salaries.show()  
+-----+-----+-----+-----+-----+-----+-----+-----+  
|Rating| Company Name| Job Title| Salary|Salaries Reported| Location|Employment Status|Job Roles|  
+-----+-----+-----+-----+-----+-----+-----+-----+  
| 3.8| Sasken|Android Developer| 400000| 3|Bangalore| Full Time| Android|  
| 4.5|Advanced Millenni...|Android Developer| 400000| 3|Bangalore| Full Time| Android|  
| 4.0| Unacademy|Android Developer|1000000| 3|Bangalore| Full Time| Android|  
| 3.8| SnapBlizz Cloudtech|Android Developer| 300000| 3|Bangalore| Full Time| Android|  
| 4.4|Appoids Tech Solu...|Android Developer| 600000| 3|Bangalore| Full Time| Android|  
| 3.7| SQUARE N CUBE|Android Developer| 192000| 3|Bangalore| Full Time| Android|  
| 3.1|Samsung R&D Instl...|Android Developer| 400000| 3|Bangalore| Full Time| Android|  
| 3.7|DXMInds Technologies|Android Developer| 300000| 3|Bangalore| Full Time| Android|  
| 3.6|Endeavour Softwar...|Android Developer| 600000| 3|Bangalore| Full Time| Android|  
| 3.6| Craft Silicon|Android Developer| 300000| 3|Bangalore| Full Time| Android|
```

6. Filtrer les lignes avec un salaire supérieur à 100 000 :

```
high_salaries = sal.filter(sal["Salary"] > 100000)  
high_salaries.show()
```

description: Cela filtre les lignes du DataFrame où la valeur de la colonne "Salary" est supérieure à 100 000 et affiche le résultat.

```

maram [En fonction] - Oracle VM VirtualBox
Fichier Machine Écran Entrée Périphériques Aide

Activities Terminal 02:31 6 ماي
root@hadoop-master: ~

min|      100000|      1|
max|     996000|     98|

>>>
>>> high_salaries = sal.filter(sal["Salary"] > 100000)
>>> high_salaries.show()

+-----+-----+-----+-----+-----+-----+-----+-----+
|Rating| Company Name| Job Title| Salary|Salaries Reported| Location|Employment Status|Job Roles|
+-----+-----+-----+-----+-----+-----+-----+-----+
| 3.8|      Saksen|Android Developer| 400000|              3|Bangalore|      Full Time|  Android|
| 4.5|Advanced Millenni...|Android Developer| 400000|              3|Bangalore|      Full Time|  Android|
| 4.0|      Unacademy|Android Developer|1000000|              3|Bangalore|      Full Time|  Android|
| 3.8|   SnapBizz Cloudtech|Android Developer| 300000|              3|Bangalore|      Full Time|  Android|
| 4.4|Appoids Tech Solu...|Android Developer| 600000|              3|Bangalore|      Full Time|  Android|
| 3.7|    SQUARE N CUBE|Android Developer| 192000|              3|Bangalore|      Full Time|  Android|
| 3.1|Samsung R&D Instl...|Android Developer| 400000|              3|Bangalore|      Full Time|  Android|
| 3.7|DXMinds Technologies|Android Developer| 300000|              3|Bangalore|      Full Time|  Android|
| 3.6|Endeavour Softwar...|Android Developer| 600000|              3|Bangalore|      Full Time|  Android|
| 3.6|   Craft Silicon|Android Developer| 300000|              3|Bangalore|      Full Time|  Android|
| 3.9|Baronford & Assoc...|Android Developer| 240000|              2|Bangalore|      Full Time|  Android|
| 3.7|      Wibmo|Android Developer| 900000|              2|Bangalore|      Full Time|  Android|
| 3.9|   Bookmyshow|Android Developer| 600000|              2|Bangalore|      Full Time|  Android|
| 3.9| Knowledge Flex|Android Developer| 228000|              2|Bangalore|      Full Time|  Android|
| 3.6|  Novopay Solutions|Android Developer| 600000|              2|Bangalore|      Full Time|  Android|
| 3.7|   WealthEngine|Android Developer| 360000|              2|Bangalore|      Full Time|  Android|
| 4.0|   J.P. Morgan|Android Developer|1000000|              2|Bangalore|      Full Time|  Android|
| 3.6|   Acviss|Android Developer| 500000|              2|Bangalore|      Full Time|  Android|
| 4.1|   Fresher|Android Developer| 408000|              2|Bangalore|      Full Time|  Android|
| 4.2|   MedOnGo|Android Developer| 300000|              2|Bangalore|      Full Time|  Android|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows

>>> sorted_by_salary = sal.orderBy(sal["Salary"].desc())
>>> sorted_by_salary.show()
[Stage 9:]

+-----+-----+-----+-----+-----+-----+-----+-----+
|Rating| Company Name| Job Title| Salary|Salaries Reported| Location|Employment Status|Job Roles|
+-----+-----+-----+-----+-----+-----+-----+-----+
| 3.9|   ennVee TechnoGroup|Oracle Applicatio...|996000|              1|Bangalore|      Full Time| Database|

```

7. Calculer le salaire moyen par entreprise :

```

avg_salary_by_company = sal.groupBy("Company Name").avg("Salary")
avg_salary_by_company.show()

```

description: Cela regroupe les données par entreprise et calcule la moyenne du salaire pour chaque entreprise, puis affiche le résultat.

8. Trier le DataFrame par salaire décroissant :

```

sorted_by_salary = sal.orderBy(sal["Salary"].desc())
sorted_by_salary.show()

```

description: Cela trie le DataFrame en fonction de la colonne "Salary" de manière décroissante et affiche le résultat.

maram [En fonction] - Oracle VM VirtualBox

Fichier Machine Écran Entrée Périphériques Aide

Activities Terminal 02:31 مای ٦

root@hadoop-master: ~

```
4.1|      Fresher|Android Developer| 400000|      2|Bangalore|      Full Time|  Android|
4.2|      MedOnGo|Android Developer| 300000|      2|Bangalore|      Full Time|  Android|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows

>>> sorted_by_salary = sal.orderBy(sal["Salary"].desc())
>>> sorted_by_salary.show()
[Stage 9:]

+-----+-----+-----+-----+-----+-----+-----+-----+
|Rating|      Company Name|      Job Title|Salary|Salaries Reported| Location|Employment Status|Job Roles|
+-----+-----+-----+-----+-----+-----+-----+-----+
3.9|ennVee TechnoGroup|Oracle Applicatio...|996000|      1|Bangalore|      Full Time|  Database|
4.2|      MakeMyTrip|Software Developm...|996000|      4|Bangalore|      Full Time|  SDE|
4.3|      Esper|      Android Engineer|996000|      1|Bangalore|      Full Time|  Android|
4.3|      Iris Software|Senior Android De...|996000|      1|New Delhi|      Contractor|  Android|
4.2|9Logic Technologi...|      Android Developer|996000|      1|Chennai|      Full Time|  Android|
3.8|      Ecom Express|      Front End Developer|996000|      1|New Delhi|      Full Time|  Frontend|
3.9|      Sociolla|Senior IOS App De...|996000|      1|New Delhi|      Full Time|  IOS|
5.0|      Ramcides|Software Engineer...|996000|      1|Hyderabad|      Full Time|  Java|
5.0|      AB Solutions Lab|Mobile App Developer|996000|      1|Bangalore|      Full Time|  Mobile|
4.0|      Goldman Sachs|Software Developm...|996000|      4|Bangalore|      Intern|  SDE|
4.4|      Microsoft|Software Developm...|996000|      1|Chennai|      Intern|  SDE|
3.8|      Amazon|Software Developm...|996000|      1|Hyderabad|      Full Time|  SDE|
3.7|US Department of ...|Software Developm...|996000|      1|Hyderabad|      Full Time|  SDE|
2.7|      Data Dimensions|Software Developm...|996000|      1|Mumbai|      Full Time|  SDE|
4.0|      Sira Consulting|Software Developm...|996000|      1|New Delhi|      Intern|  SDE|
4.0|      CrossChannel|Software Developm...|996000|      1|Pune|      Full Time|  SDE|
3.5|      Diamantl|Software Developm...|996000|      1|Pune|      Full Time|  SDE|
4.4|      ThinkBridge|Software Developm...|996000|      1|Pune|      Full Time|  SDE|
3.7|      CDK Global|      Test Engineer|996000|      1|Hyderabad|      Full Time|  Testing|
3.9|Tata Consultancy ...|      Tester|996000|      8|Hyderabad|      Full Time|  Testing|
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows

>>> df.select('Company Name', 'Salary').show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'df' is not defined
>>> avg_salary_by_company = sal.groupBy("Company Name").avg("Salary")
Traceback (most recent call last):
```