

**COMPTE RENDU**

**DE**

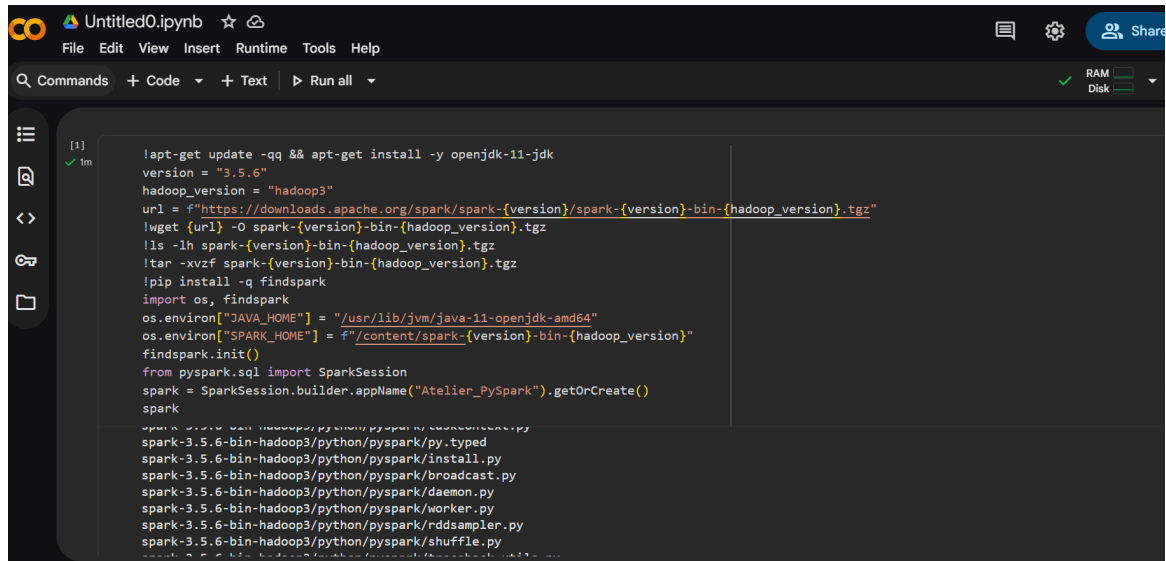
**Analyse des ventes avec Spark SQL**

Réalisé par :

***Hajar Rachid*** (G2)

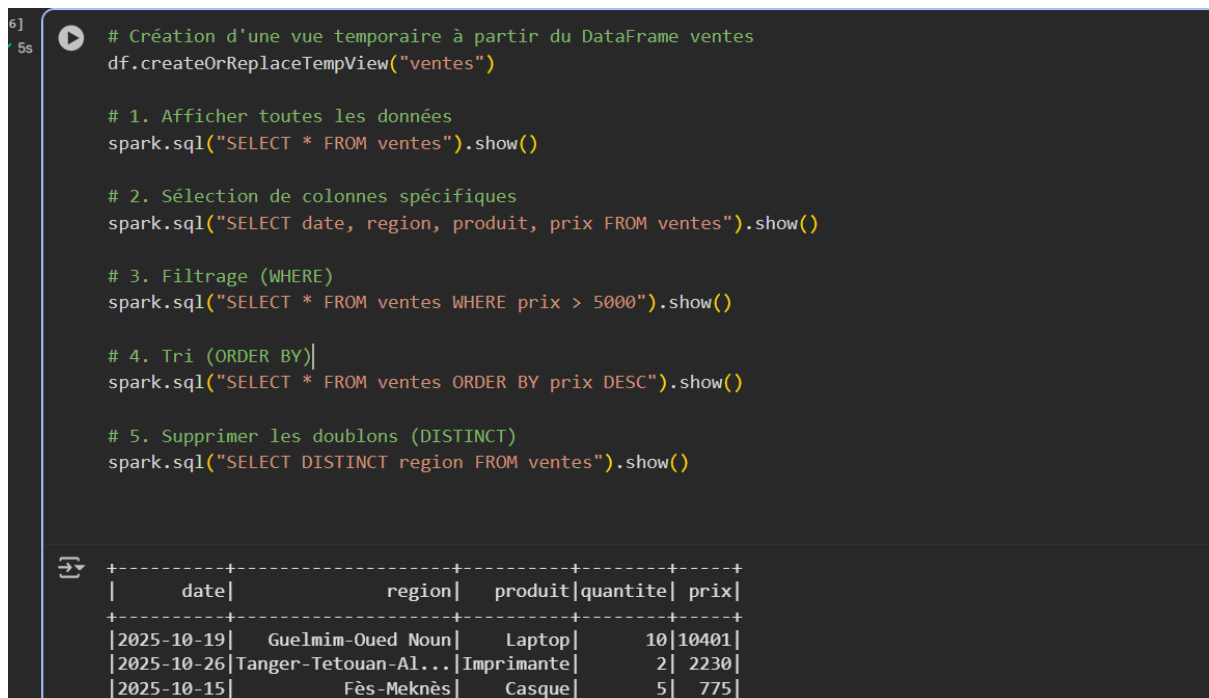
**SIR**

## II. Mise en place de l'environnement



```
[1] ✓ 1m
!apt-get update -qq && apt-get install -y openjdk-11-jdk
version = "3.5.6"
hadoop_version = "hadoop3"
url = f"https://downloads.apache.org/spark/spark-{version}/spark-{version}-bin-{hadoop_version}.tgz"
!wget {url} -O spark-{version}-bin-{hadoop_version}.tgz
!ls -lh spark-{version}-bin-{hadoop_version}.tgz
!tar -xvzf spark-{version}-bin-{hadoop_version}.tgz
!pip install -q findspark
import os, findspark
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["SPARK_HOME"] = f"/content/spark-{version}-bin-{hadoop_version}"
findspark.init()
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("Atelier_PySpark").getOrCreate()
spark
spark-3.5.6-bin-hadoop3/python/pyspark/submit.py
spark-3.5.6-bin-hadoop3/python/pyspark/py.typed
spark-3.5.6-bin-hadoop3/python/pyspark/install.py
spark-3.5.6-bin-hadoop3/python/pyspark/broadcast.py
spark-3.5.6-bin-hadoop3/python/pyspark/daemon.py
spark-3.5.6-bin-hadoop3/python/pyspark/worker.py
spark-3.5.6-bin-hadoop3/python/pyspark/rddsampler.py
spark-3.5.6-bin-hadoop3/python/pyspark/shuffle.py
```

## III. Rappel sur les requêtes SQL de base



```
6] 5s
# Création d'une vue temporaire à partir du DataFrame ventes
df.createOrReplaceTempView("ventes")

# 1. Afficher toutes les données
spark.sql("SELECT * FROM ventes").show()

# 2. Sélection de colonnes spécifiques
spark.sql("SELECT date, region, produit, prix FROM ventes").show()

# 3. Filtrage (WHERE)
spark.sql("SELECT * FROM ventes WHERE prix > 5000").show()

# 4. Tri (ORDER BY)
spark.sql("SELECT * FROM ventes ORDER BY prix DESC").show()

# 5. Supprimer les doublons (DISTINCT)
spark.sql("SELECT DISTINCT region FROM ventes").show()
```

date	region	produit	quantite	prix
2025-10-19	Guelmim-Oued Noun	Laptop	10	10401
2025-10-26	Tanger-Tetouan-Al...	Imprimante	2	2230
2025-10-15	Fès-Meknès	Casque	5	775

	date	region	produit	quantite	prix
	2025-10-19	Guelmim-Oued Noun	Laptop	10	10401
	2025-10-26	Tanger-Tetouan-Al...	Imprimante	2	2230
	2025-10-15	Fès-Meknès	Casque	5	775
	2025-10-02	Fès-Meknès	Imprimante	9	3129
	2025-10-02	Marrakech-Safi	Laptop	1	8514
	2025-10-30	Tanger-Tétouan-Al...	USB	9	85
	2025-10-19	Fès-Meknès	Laptop	4	9466
	2025-10-25	Marrakech-Safi	Imprimante	9	2216

	date	region	produit	prix
	2025-10-19	Guelmim-Oued Noun	Laptop	10401
	2025-10-26	Tanger-Tetouan-Al...	Imprimante	2230
	2025-10-15	Fès-Meknès	Casque	775
	2025-10-02	Fès-Meknès	Imprimante	3129
	2025-10-02	Marrakech-Safi	Laptop	8514
	2025-10-30	Tanger-Tétouan-Al...	USB	85
	2025-10-19	Fès-Meknès	Laptop	9466
	2025-10-25	Marrakech-Safi	Imprimante	2216

	date	region	produit	quantite	prix
	2025-10-19	Guelmim-Oued Noun	Laptop	10	10401

	date	region	produit	quantite	prix
	2025-10-19	Guelmim-Oued Noun	Laptop	10	10401
	2025-10-02	Marrakech-Safi	Laptop	1	8514
	2025-10-19	Fès-Meknès	Laptop	4	9466

	date	region	produit	quantite	prix
	2025-10-19	Guelmim-Oued Noun	Laptop	10	10401
	2025-10-19	Fès-Meknès	Laptop	4	9466
	2025-10-02	Marrakech-Safi	Laptop	1	8514
	2025-10-02	Fès-Meknès	Imprimante	9	3129
	2025-10-26	Tanger-Tetouan-Al...	Imprimante	2	2230
	2025-10-25	Marrakech-Safi	Imprimante	9	2216
	2025-10-15	Fès-Meknès	Casque	5	775
	2025-10-30	Tanger-Tétouan-Al...	USB	9	85

	region
	Fès-Meknès
	Guelmim-Oued Noun
	Marrakech-Safi
	Tanger-Tetouan-Al...
	Tanger-Tétouan-Al...

## IV. Calculs et agrégations

```
[11]
✓ 4s

# Somme totale par région
spark.sql("""
SELECT region, Sum(prix) AS total_prix
FROM ventes
GROUP BY region
""").show()

# Moyenne des prix par produit
spark.sql("""
SELECT produit, AVG(prix) AS moyenne_prix
FROM ventes
GROUP BY produit
""").show()

# Total par région et produit
spark.sql("""
SELECT region, produit, SUM(prix) AS total
FROM ventes
GROUP BY region, produit
""").show()

# Condition sur l'agrégation (HAVING)
spark.sql("""
SELECT region, SUM(prix) AS total
FROM ventes
GROUP BY region
HAVING total > 10000
""").show()
```

```
[11]
✓ 4s

+-----+-----+
|          region|total_prix|
+-----+-----+
|      Fès-Meknès|    13370|
| Guelmim-Oued Noun|    10401|
| Marrakech-Safi|    10730|
|Tanger-Tétouan-Al...|     2230|
|Tanger-Tétouan-Al...|       85|
+-----+-----+

+-----+-----+
| produit| moyenne_prix|
+-----+-----+
| Laptop|9460.333333333334|
| Imprimante|    2525.0|
| Casque|    775.0|
| USB|    85.0|
+-----+-----+

+-----+-----+-----+
|          region| produit|total|
+-----+-----+-----+
|Tanger-Tétouan-Al...|    USB|    85|
|      Fès-Meknès|Imprimante|  3129|
| Guelmim-Oued Noun|  Laptop|10401|
|      Fès-Meknès|  Casque|   775|
| Marrakech-Safi|  Laptop|  8514|
|      Fès-Meknès|  Laptop|  9466|
+-----+-----+-----+
```

```
+-----+-----+
+-----+-----+-----+
|          region| produit|total|
+-----+-----+-----+
|Tanger-Tétouan-Al...|    USB|    85|
|      Fès-Meknès|Imprimante|  3129|
| Guelmim-Oued Noun|  Laptop|10401|
|      Fès-Meknès|  Casque|   775|
| Marrakech-Safi|  Laptop|  8514|
|      Fès-Meknès|  Laptop|  9466|
|Tanger-Tetouan-Al...|Imprimante| 2230|
| Marrakech-Safi|Imprimante| 2216|
+-----+-----+-----+

+-----+-----+
|          region|total|
+-----+-----+
|      Fès-Meknès|13370|
| Guelmim-Oued Noun|10401|
| Marrakech-Safi|10730|
+-----+-----+
```

## V. Création de nouvelles vues

```
3] 0s # Création d'une nouvelles vues
spark.sql("""
CREATE OR REPLACE TEMP VIEW ventes_resumees AS
SELECT region, produit, SUM(quantite) AS qte_totale, SUM(prix) AS totale_prix
FROM ventes
GROUP BY region, produit
""")
spark.sql("SELECT * FROM ventes_resumees").show()
```

region	produit	qte_totale	totale_prix
Tanger-Tétouan-Al...	USB	9	85
Fès-Meknès	Imprimante	9	3129
Guelmim-Oued Noun	Laptop	10	10401
Fès-Meknès	Casque	5	775
Marrakech-Safi	Laptop	1	8514
Fès-Meknès	Laptop	4	9466
Tanger-Tetouan-Al...	Imprimante	2	2230
Marrakech-Safi	Imprimante	9	2216

## VI. Jointures entre vues

```
4] 3s # Jointures entre vues
data_cat = [
    ("Laptop", "Informatique"),
    ("Imprimante", "Bureau"),
    ("Casque", "Audio"),
    ("USB", "Accessoires")
]

df_cat = spark.createDataFrame(data_cat, ["produit", "categorie"])
df_cat.createOrReplaceTempView("categories")

# Jointure entre ventes et categories
spark.sql("""
SELECT v.region, v.produit, c.categorie, v.quantite, v.prix
FROM ventes v
JOIN categories c ON v.produit = c.produit
""").show()
```

region	produit	categorie	quantite	prix
Fès-Meknès	Laptop	Informatique	4	9466
Marrakech-Safi	Laptop	Informatique	1	8514
Guelmim-Oued Noun	Laptop	Informatique	10	10401
Marrakech-Safi	Imprimante	Bureau	9	2216
Fès-Meknès	Imprimante	Bureau	9	3129
Tanger-Tetouan-Al...	Imprimante	Bureau	2	2230
Fès-Meknès	Casque	Audio	5	775
Tanger-Tétouan-Al...	USB	Accessoires	9	85

## VII. Fonctions SQL avancées

```
(19)
✓ 1s
# Extraire l'année et le mois
spark.sql("""
SELECT
  date,
  SUBSTRING(date, 1, 4) AS annee,
  SUBSTRING(date, 6, 2) AS mois,
  produit,
  prix
FROM ventes
""").show()

# Ajouter des colonnes calculées
spark.sql("""
SELECT *,
  quantite * prix AS total_vente
FROM ventes
""").show()

#Utiliser des fonctions d'agrégation multiples
spark.sql("""
SELECT
  region,
  MAX(prix) AS prix_max,
  MIN(prix) AS prix_min,
  ROUND(AVG(prix),2) AS prix_moyen
FROM ventes
GROUP BY region
""").show()
```

date	annee	mois	produit	prix
2025-10-19	2025	10	Laptop	10401
2025-10-26	2025	10	Imprimante	2230
2025-10-15	2025	10	Casque	775
2025-10-02	2025	10	Imprimante	3129
2025-10-02	2025	10	Laptop	8514
2025-10-30	2025	10	USB	85
2025-10-19	2025	10	Laptop	9466
2025-10-25	2025	10	Imprimante	2216

date	region	produit	quantite	prix	total_vente
2025-10-19	Guelmim-Oued Noun	Laptop	10	10401	104010
2025-10-26	Tanger-Tetouan-Al...	Imprimante	2	2230	4460
2025-10-15	Fès-Meknès	Casque	5	775	3875
2025-10-02	Fès-Meknès	Imprimante	9	3129	28161
2025-10-02	Marrakech-Safi	Laptop	1	8514	8514
2025-10-30	Tanger-Tétouan-Al...	USB	9	85	765
2025-10-19	Fès-Meknès	Laptop	4	9466	37864
2025-10-25	Marrakech-Safi	Imprimante	9	2216	19944

region	prix_max	prix_min	prix_moyen
Fès-Meknès	9466	775	4456.67
Guelmim-Oued Noun	10401	10401	10401.0
Marrakech-Safi	8514	2216	5365.0
Tanger-Tetouan-Al...	2230	2230	2230.0
Tanger-Tétouan-Al...	85	85	85.0

## → Synthèse

- o Spark SQL permet de manipuler les données comme en SQL classique.
- o Les vues facilitent la réutilisation des requêtes complexes.
- o Les fonctions d'agrégation et de jointure sont essentielles pour l'analyse.
- o On peut combiner SQL + DataFrame API selon le besoin (flexibilité)

## Exercice pratique : Analyse des ventes

### Exercice pratique : Analyse des ventes

```
# 1. Création de la vue temporaire
df.createOrReplaceTempView("ventes_view")

# Vérification
spark.sql("SELECT * FROM ventes_view LIMIT 5").show()

# 2 Nombre total d'unités vendues par région
spark.sql("""
    SELECT region, SUM(quantite) AS total_quantite
    FROM ventes_view
    GROUP BY region
    ORDER BY total_quantite DESC
""").show()

# 3 Prix maximum par produit
spark.sql("""
    SELECT produit, MAX(prix) AS prix_max
    FROM ventes_view
    GROUP BY produit
""").show()
```

```
SELECT produit, MAX(prix) AS prix_max
FROM ventes_view
GROUP BY produit
""").show()

# Produit le plus cher
spark.sql("""
    SELECT produit, MAX(prix) AS prix_max
    FROM ventes_view
    GROUP BY produit
    ORDER BY prix_max DESC
    LIMIT 1
""").show()

# 4 Chiffre d'affaires total par produit
spark.sql("""
    SELECT produit,
        SUM(quantite * prix) AS chiffre_affaires
    FROM ventes_view
    GROUP BY produit
    ORDER BY chiffre_affaires DESC
""").show()

# 5 Analyse mensuelle des ventes
spark.sql("""
    SELECT
```

```
[ ]
        SUBSTRING(date, 6, 2) AS mois,
        SUM(quantite * prix) AS chiffre_affaires
    FROM ventes_view
    WHERE SUBSTRING(date, 1, 4) = '2025'
    GROUP BY annee, mois
    ORDER BY mois
    """).show()

# 6 Ventes dépassant 10 000
spark.sql("""
    SELECT *,
           (quantite * prix) AS total_vente
    FROM ventes_view
    WHERE (quantite * prix) > 10000
    """).show()

# 7 Produit le plus vendu par région
spark.sql("""
    SELECT region, produit, SUM(quantite) AS total_quantite
    FROM ventes_view
    GROUP BY region, produit
    """).createOrReplaceTempView("ventes_region")

spark.sql("""
    SELECT region, produit, total_quantite
    FROM (

```

```

    )
    WHERE rang = 1
    """).show()

# 8 Statistiques des prix par produit
spark.sql("""
    SELECT produit,
           MIN(prix) AS prix_min,
           MAX(prix) AS prix_max,
           ROUND(AVG(prix), 2) AS prix_moyen
    FROM ventes_view
    GROUP BY produit
    """).show()

# 9 Création d'une vue ventes_analyse
spark.sql("""
    CREATE OR REPLACE TEMP VIEW ventes_analyse AS
    SELECT region,
           produit,
           quantite,
           prix,
           (quantite * prix) AS total_vente,
           SUBSTRING(date, 1, 4) AS annee,
           SUBSTRING(date, 6, 2) AS mois
    FROM ventes_view
    """)

```



```

# 10 Top 3 des produits les plus rentables
spark.sql("""
    SELECT produit, SUM(total_vente) AS chiffre_affaires
    FROM ventes_analyse
    GROUP BY produit
    ORDER BY chiffre_affaires DESC
    LIMIT 3
""").show()

# 11 Chiffre d'affaires total par région
spark.sql("""
    SELECT region, SUM(total_vente) AS total_region
    FROM ventes_analyse
    GROUP BY region
    ORDER BY total_region DESC
""").show()

# Région avec le plus haut chiffre d'affaires
spark.sql("""
    SELECT region, SUM(total_vente) AS total_region
    FROM ventes_analyse
    GROUP BY region
    ORDER BY total_region DESC
    LIMIT 1
""").show()

```

```

+-----+-----+-----+-----+-----+-----+
|2025-10-19|Guelmim-Oued Noun|Laptop|10|10401|104010|
|2025-10-02|Fès-Meknès|Imprimante|9|3129|28161|
|2025-10-19|Fès-Meknès|Laptop|4|9466|37864|
|2025-10-25|Marrakech-Safi|Imprimante|9|2216|19944|
+-----+-----+-----+-----+-----+-----+

+-----+-----+-----+-----+
|      region|produit|total_quantite|
+-----+-----+-----+-----+
|Fès-Meknès|Imprimante|9|
|Guelmim-Oued Noun|Laptop|10|
|Marrakech-Safi|Imprimante|9|
|Tanger-Tetouan-Al...|Imprimante|2|
|Tanger-Tétouan-Al...|USB|9|
+-----+-----+-----+-----+

+-----+-----+-----+-----+
|produit|prix_min|prix_max|prix_moyen|
+-----+-----+-----+-----+
|Laptop|8514|10401|9460.33|
|Imprimante|2216|3129|2525.0|
|Casque|775|775|775.0|
|USB|85|85|85.0|
+-----+-----+-----+-----+

+-----+-----+-----+-----+-----+-----+-----+-----+
|      region|produit|quantite|prix|total_vente|annee|mois|
+-----+-----+-----+-----+-----+-----+-----+-----+
|Guelmim-Oued Noun|Laptop|10|10401|104010|2025|10|

```



region	produit	quantite	prix	total_vente	annee	mois
Guelmim-Oued Noun	Laptop	10	10401	104010	2025	10
Tanger-Tetouan-Al...	Imprimante	2	2230	4460	2025	10
Fès-Meknès	Casque	5	775	3875	2025	10
Fès-Meknès	Imprimante	9	3129	28161	2025	10
Marrakech-Safi	Laptop	1	8514	8514	2025	10

produit	chiffre_affaires
Laptop	150388
Imprimante	52565
Casque	3875

region	total_region
Guelmim-Oued Noun	104010
Fès-Meknès	69900
Marrakech-Safi	28458
Tanger-Tetouan-Al...	4460
Tanger-Tétouan-Al...	765