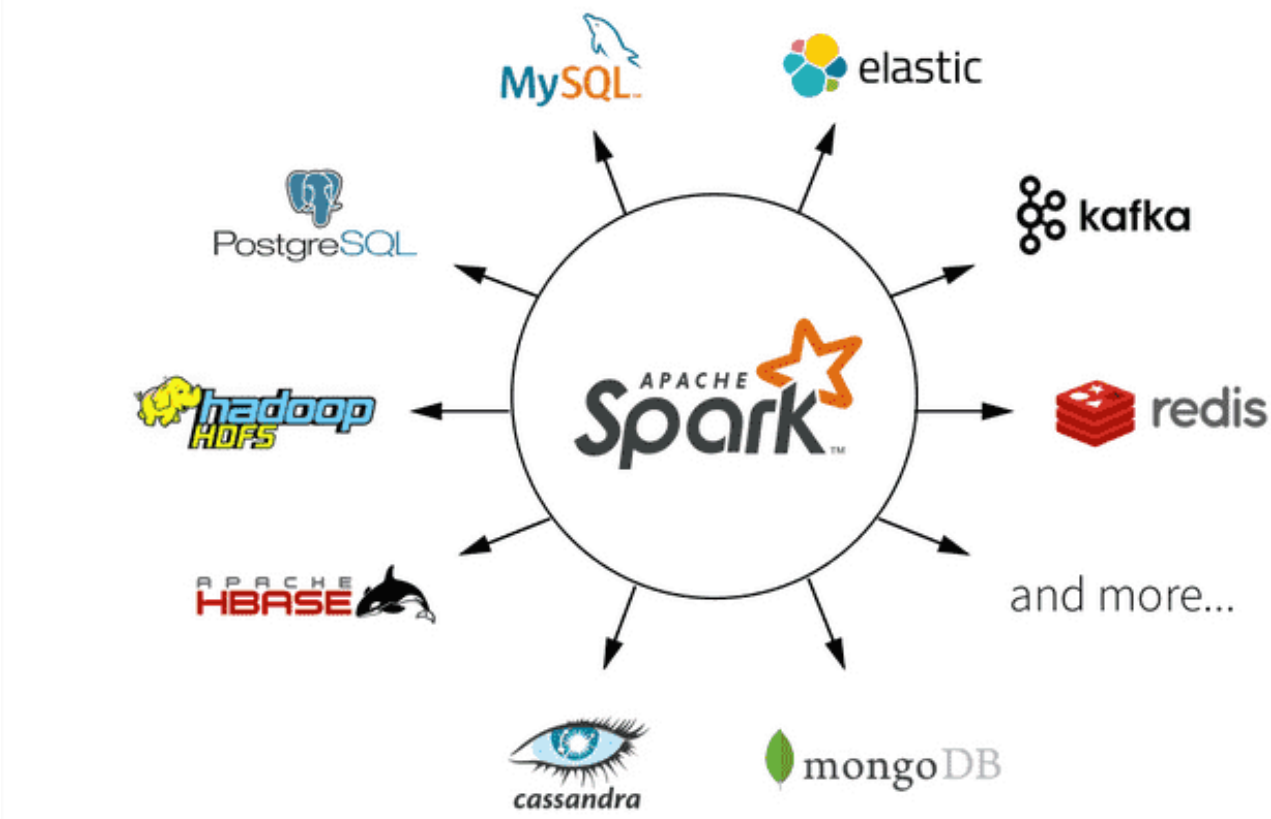
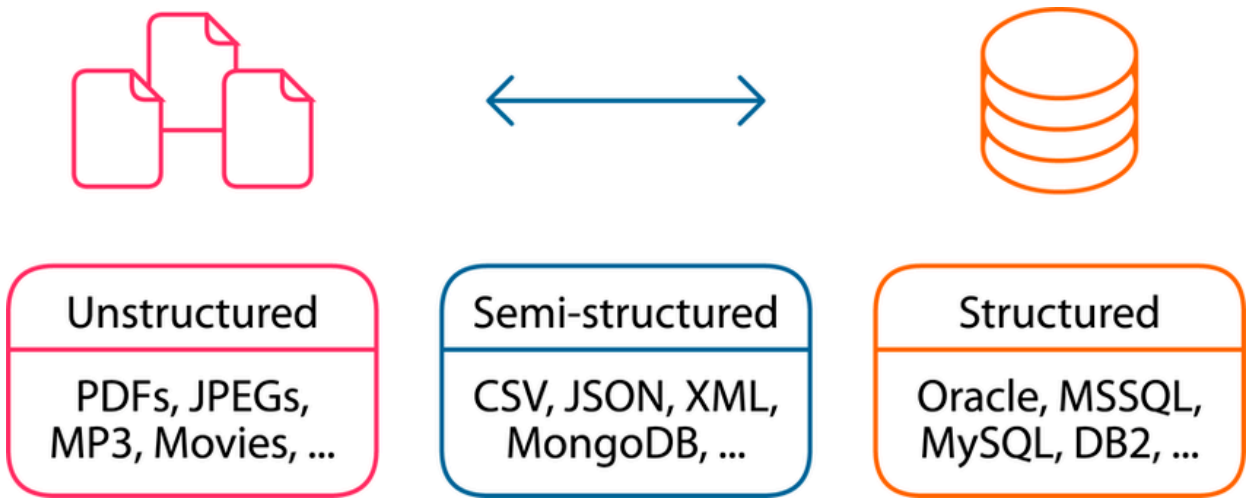
Read Data

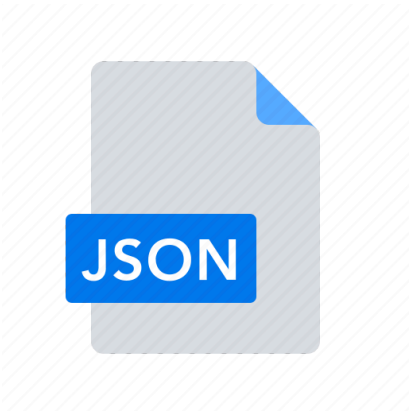


- spark = 3.2.1
- scala= 2.12.15

Data Types



Json



In [ ]:
```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("Load JSON files") \
    .getOrCreate()
```

In [ ]:
```
df = spark.read.json("sales_data.json")
```

In [ ]:
```
df = spark.read.load("sales_data.json",format="json")
```

In [ ]:
```
df = spark.read.format("json").load("sales_data.json")
```

In [ ]:
```
df.show()
```

```
+---------------+-----------+------+--------+----+--------+-------------------+---------+------------------+
|        article|payment_mode|pos_id|pos_name|prix|quantity|          sale_time|sale_type|             total|
+---------------+-----------+------+--------+----+--------+-------------------+---------+------------------+
|   Key Lime Tart|      online|     6|   tunis| 8.2|      16|2024-03-20 02:58:00|livraison|             131.2|
|       Croissant|      online|     6|   tunis| 1.5|       6|2024-03-20 02:58:01|livraison|               9.0|
|   Danish Pastry|        cash|     6|   tunis| 6.5|       6|2024-03-20 02:58:04|   direct|              39.0|
|         Palmier|      online|     6|   tunis| 3.0|       3|2024-03-20 02:58:05|livraison|               9.0|
|       Bear Claw|        card|     6|   tunis| 6.8|       5|2024-03-20 02:58:07|   direct|              34.0|
|Chocolate Eclair|        cash|     6|   tunis| 7.5|      20|2024-03-20 02:58:08|   direct|             150.0|
|  Apple Turnover|      online|     6|   tunis| 5.0|       9|2024-03-20 02:58:09|livraison|              45.0|
|   Danish Pastry|      online|     6|   tunis| 6.5|      10|2024-03-20 02:58:10|livraison|              65.0|
|         Strudel|        card|     6|   tunis| 8.0|      16|2024-03-20 02:58:14|   direct|             128.0|
|Blueberry Muffin|        card|     6|   tunis| 3.8|       4|2024-03-20 02:58:17|   direct|              15.2|
|        Baguette|      online|     6|   tunis| 2.0|       7|2024-03-20 02:58:18|livraison|              14.0|
|       Lemon Bar|      online|     6|   tunis| 6.0|      11|2024-03-20 02:58:22|livraison|              66.0|
|       Croissant|      online|     6|   tunis| 1.5|      10|2024-03-20 02:58:25|livraison|              15.0|
|           Scone|      online|     6|   tunis| 2.8|      12|2024-03-20 02:58:28|livraison|33.599999999999994|
|      Cherry Pie|      online|     6|   tunis| 9.5|      13|2024-03-20 02:58:32|livraison|             123.5|
|   Key Lime Tart|        cash|     6|   tunis| 8.2|       2|2024-03-20 02:58:35|   direct|              16.4|
|        Napoleon|      online|     6|   tunis| 7.9|       2|2024-03-20 02:58:37|livraison|              15.8|
|         Cupcake|        cash|     6|   tunis| 4.5|      20|2024-03-20 02:58:38|   direct|              90.0|
|       Lemon Bar|        cash|     6|   tunis| 6.0|       7|2024-03-20 02:58:40|   direct|              42.0|
|         Cupcake|      online|     6|   tunis| 4.5|       4|2024-03-20 02:58:41|livraison|              18.0|
+---------------+-----------+------+--------+----+--------+-------------------+---------+------------------+
only showing top 20 rows
```

CSV

Load Data With Defined StructType Schema

In [ ]:
```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("Load CSV files") \
    .getOrCreate()

df = spark.read.csv("sales_data.csv", header=True)

df.show()
```

```
+------+--------+---------------+--------+----------+-----+---------+------------+-------------------+
|POS ID|POS Name|        Article|Quantity|Unit Price|Total|Sale Type|Payment Mode|          Sale Time|
+------+--------+---------------+--------+----------+-----+---------+------------+-------------------+
|     4|    beja|  Danish Pastry|       2|       6.5| 13.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|Red Velvet Cake|       1|      12.0| 12.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|      Lemon Bar|      16|       6.0| 96.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|  Danish Pastry|      18|       6.5|117.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|  Danish Pastry|       8|       6.5| 52.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|      Croissant|      11|       1.5| 16.5|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|  Key Lime Tart|      16|       8.2|131.2|   direct|        card|2024-03-31 01:30:18|
|     4|    beja|Red Velvet Cake|      14|      12.0|168.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|  Cinnamon Roll|       3|       4.0| 12.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|  Cinnamon Roll|       6|       4.0| 24.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|Red Velvet Cake|       1|      12.0| 12.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|      Croissant|      20|       1.5| 30.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|  Danish Pastry|      16|       6.5|104.0|   direct|        card|2024-03-31 01:30:18|
|     4|    beja|      Pecan Pie|      10|      11.0|110.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|      Croissant|      11|       1.5| 16.5|   direct|        cash|2024-03-31 01:30:18|
|     4|    beja|      Pecan Pie|       6|      11.0| 66.0|   direct|        card|2024-03-31 01:30:18|
|     4|    beja|  Danish Pastry|       2|       6.5| 13.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|      Croissant|       7|       1.5| 10.5|   direct|        card|2024-03-31 01:30:18|
|     4|    beja|Red Velvet Cake|      10|      12.0|120.0|   direct|        cash|2024-03-31 01:30:18|
|     4|    beja|      Lemon Bar|      16|       6.0| 96.0|   direct|        card|2024-03-31 01:30:18|
+------+--------+---------------+--------+----------+-----+---------+------------+-------------------+
only showing top 20 rows
```

In [ ]:
```
df.printSchema()
```

```
root
 |-- POS ID: string (nullable = true)
 |-- POS Name: string (nullable = true)
 |-- Article: string (nullable = true)
 |-- Quantity: string (nullable = true)
 |-- Unit Price: string (nullable = true)
 |-- Total: string (nullable = true)
 |-- Sale Type: string (nullable = true)
 |-- Payment Mode: string (nullable = true)
 |-- Sale Time: string (nullable = true)
```

Load Data With Specific StructType Schema

Struct Type Example

In [ ]:
```
from pyspark.sql.types import StructType, StructField, StringType, IntegerType, FloatType, DoubleType, TimestampType

schema = StructType([
    StructField("POS ID", IntegerType()),
    StructField("POS Name", StringType()),
    StructField("Article", StringType()),
    StructField("Quantity", FloatType()),
    StructField("Unit Price", FloatType()),
    StructField("Total", DoubleType()),
    StructField("Sale Type", StringType()),
    StructField("Payment Mode", StringType()),
    StructField("Sale Time", TimestampType())
])
```

DDL (Data Definition Language) Example

In [ ]:
```
schema = "POS_ID INT, POS_Name STRING, Article STRING, Quantity FLOAT, Unit_Price FLOAT, Total DOUBLE, Sale_Type STRING, Paymer
```

In [ ]:
```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("Load CSV files") \
    .getOrCreate()

df = spark.read.csv("sales_data.csv", header=True, sep=",", schema=schema)

df.show()
```

```
+------+--------+--------------+--------+----------+-----+---------+------------+-------------------+
|POS ID|POS Name|       Article|Quantity|Unit Price|Total|Sale Type|Payment Mode|          Sale Time|
+------+--------+--------------+--------+----------+-----+---------+------------+-------------------+
|     4|    beja|  Danish Pastry|     2.0|       6.5| 13.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|Red Velvet Cake|     1.0|      12.0| 12.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|      Lemon Bar|    16.0|       6.0| 96.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|  Danish Pastry|    18.0|       6.5|117.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|  Danish Pastry|     8.0|       6.5| 52.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|      Croissant|    11.0|       1.5| 16.5|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|  Key Lime Tart|    16.0|       8.2|131.2|   direct|        card|2024-03-31 01:30:18|
|     4|    beja|Red Velvet Cake|    14.0|      12.0|168.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|  Cinnamon Roll|     3.0|       4.0| 12.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|  Cinnamon Roll|     6.0|       4.0| 24.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|Red Velvet Cake|     1.0|      12.0| 12.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|      Croissant|    20.0|       1.5| 30.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|  Danish Pastry|    16.0|       6.5|104.0|   direct|        card|2024-03-31 01:30:18|
|     4|    beja|      Pecan Pie|    10.0|      11.0|110.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|      Croissant|    11.0|       1.5| 16.5|   direct|        cash|2024-03-31 01:30:18|
|     4|    beja|      Pecan Pie|     6.0|      11.0| 66.0|   direct|        card|2024-03-31 01:30:18|
|     4|    beja|  Danish Pastry|     2.0|       6.5| 13.0|livraison|      online|2024-03-31 01:30:18|
|     4|    beja|      Croissant|     7.0|       1.5| 10.5|   direct|        card|2024-03-31 01:30:18|
|     4|    beja|Red Velvet Cake|    10.0|      12.0|120.0|   direct|        cash|2024-03-31 01:30:18|
|     4|    beja|      Lemon Bar|    16.0|       6.0| 96.0|   direct|        card|2024-03-31 01:30:18|
+------+--------+--------------+--------+----------+-----+---------+------------+-------------------+
only showing top 20 rows
```

In [ ]:
```
df.printSchema()
```

```
root
 |-- POS ID: integer (nullable = true)
 |-- POS Name: string (nullable = true)
 |-- Article: string (nullable = true)
 |-- Quantity: float (nullable = true)
 |-- Unit Price: float (nullable = true)
 |-- Total: double (nullable = true)
 |-- Sale Type: string (nullable = true)
 |-- Payment Mode: string (nullable = true)
 |-- Sale Time: timestamp (nullable = true)
```

Excel

Methode 1 With Config "spark.jars.packages"

In [ ]:
```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("Read Excel Data") \
    .config('spark.jars.packages', 'com.crealytics:spark-excel_2.12:3.2.1_0.17.1') \
    .getOrCreate()

excel_df = spark.read \
    .format("com.crealytics.spark.excel") \
    .option("header", "true") \
    .load("sales_data.xlsx")
```

```
excel_df.show()
```

```
+------+--------+----------------+--------+----------+-----+---------+------------+-------------------+
|POS ID|pos_name|         Article|Quantity|Unit Price|Total|Sale Type|Payment Mode|          Sale Time|
+------+--------+----------------+--------+----------+-----+---------+------------+-------------------+
|     5|  bizert|         Strudel|      19|         8|  152|livraison|      online|2024-03-20 03:01:39|
|     5|  bizert|Chocolate Eclair|      12|       7.5|   90|   direct|        card|2024-03-20 03:01:41|
|     5|  bizert|      Cheesecake|       6|      10.5|   63|livraison|      online|2024-03-20 03:01:44|
|     5|  bizert|          Muffin|      20|       3.5|   70|   direct|        card|2024-03-20 03:01:46|
|     5|  bizert|        Baguette|      15|         2|   30|livraison|      online|2024-03-20 03:01:48|
|     5|  bizert|         Cupcake|      19|       4.5| 85.5|   direct|        cash|2024-03-20 03:01:52|
|     5|  bizert|         Palmier|      14|         3|   42|livraison|      online|2024-03-20 03:01:56|
|     5|  bizert|      Cherry Pie|      12|       9.5|  114|   direct|        card|2024-03-20 03:02:00|
|     5|  bizert|  Apple Turnover|       6|         5|   30|livraison|      online|2024-03-20 03:02:03|
|     5|  bizert|   Cinnamon Roll|      12|         4|   48|   direct|        card|2024-03-20 03:02:07|
|     5|  bizert|      Cream Puff|      19|         9|  171|   direct|        cash|2024-03-20 03:02:08|
|     5|  bizert|       Lemon Bar|      15|         6|   90|livraison|      online|2024-03-20 03:02:12|
|     5|  bizert|      Fruit Tart|      17|       8.5|144.5|livraison|      online|2024-03-20 03:02:16|
|     5|  bizert| Red Velvet Cake|       5|        12|   60|   direct|        cash|2024-03-20 03:02:19|
|     5|  bizert|        Napoleon|      18|       7.9|142.2|   direct|        cash|2024-03-20 03:02:22|
|     5|  bizert|      Cheesecake|      15|      10.5|157.5|livraison|      online|2024-03-20 03:02:24|
|     5|  bizert|Blueberry Muffin|      10|       3.8|   38|livraison|      online|2024-03-20 03:02:26|
|     5|  bizert|Blueberry Muffin|      11|       3.8| 41.8|   direct|        card|2024-03-20 03:02:30|
|     5|  bizert|      Cream Puff|      16|         9|  144|livraison|      online|2024-03-20 03:02:32|
|     5|  bizert|        Baguette|      18|         2|   36|   direct|        cash|2024-03-20 03:02:33|
+------+--------+----------------+--------+----------+-----+---------+------------+-------------------+
only showing top 20 rows
```

Methode 2 With Config "spark.jars"

In [ ]:
```
import subprocess

jdbc_url = "https://repo1.maven.org/maven2/com/crealytics/spark-excel_2.12/3.2.1_0.17.1/spark-excel_2.12-3.2.1_0.17.1.jar"
destination_directory = "/home/jovyan"
subprocess.run(["wget", jdbc_url, "-P", destination_directory])
```

Out[ ]:
```
CompletedProcess(args=['wget', 'https://repo1.maven.org/maven2/com/crealytics/spark-excel_2.12/3.4.1_0.19.0/spark-excel_2.12-3.4.1_0.19.0.jar', '-P', '/home/jovyan'], returncode=0)
```

In [ ]:
```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("Read Excel Data") \
    .config("spark.jars", "/home/jovyan/spark-excel_2.12-3.2.1_0.17.1.jar") \
    .getOrCreate()

excel_df = spark.read \
    .format("com.crealytics.spark.excel") \
    .option("header", "true") \
    .load("sales_data.xlsx")

excel_df.show()
```

POS ID	pos_name	Article	Quantity	Unit Price	Total	Sale Type	Payment Mode	Sale Time
5	bizert	Strudel	19	8	152	livraison	online	2024-03-20 03:01:39
5	bizert	Chocolate Eclair	12	7.5	90	direct	card	2024-03-20 03:01:41
5	bizert	Cheesecake	6	10.5	63	livraison	online	2024-03-20 03:01:44
5	bizert	Muffin	20	3.5	70	direct	card	2024-03-20 03:01:46
5	bizert	Baguette	15	2	30	livraison	online	2024-03-20 03:01:48
5	bizert	Cupcake	19	4.5	85.5	direct	cash	2024-03-20 03:01:52
5	bizert	Palmier	14	3	42	livraison	online	2024-03-20 03:01:56
5	bizert	Cherry Pie	12	9.5	114	direct	card	2024-03-20 03:02:00
5	bizert	Apple Turnover	6	5	30	livraison	online	2024-03-20 03:02:03
5	bizert	Cinnamon Roll	12	4	48	direct	card	2024-03-20 03:02:07
5	bizert	Cream Puff	19	9	171	direct	cash	2024-03-20 03:02:08
5	bizert	Lemon Bar	15	6	90	livraison	online	2024-03-20 03:02:12
5	bizert	Fruit Tart	17	8.5	144.5	livraison	online	2024-03-20 03:02:16
5	bizert	Red Velvet Cake	5	12	60	direct	cash	2024-03-20 03:02:19
5	bizert	Napoleon	18	7.9	142.2	direct	cash	2024-03-20 03:02:22
5	bizert	Cheesecake	15	10.5	157.5	livraison	online	2024-03-20 03:02:24
5	bizert	Blueberry Muffin	10	3.8	38	livraison	online	2024-03-20 03:02:26
5	bizert	Blueberry Muffin	11	3.8	41.8	direct	card	2024-03-20 03:02:30
5	bizert	Cream Puff	16	9	144	livraison	online	2024-03-20 03:02:32
5	bizert	Baguette	18	2	36	direct	cash	2024-03-20 03:02:33

only showing top 20 rows

MySQL

- MySQL (Version: 8.0.32)

In [ ]:
```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("Read from MySQL") \
    .config('spark.jars.packages', 'mysql:mysql-connector-java:8.0.32') \
    .getOrCreate()

mysql_host = "jendouba_host"
mysql_port = "3306"
mysql_database = "jendouba_sales_db"
mysql_username = "root"
mysql_password = "secret"
mysql_table = "jendouba_sales"

jdbc_url = f"jdbc:mysql://{mysql_host}:{mysql_port}/{mysql_database}"

mysql_properties = {
    "user": mysql_username,
    "password": mysql_password,
    "driver": "com.mysql.cj.jdbc.Driver"
}

df = spark.read.jdbc(url=jdbc_url, table=mysql_table, properties=mysql_properties)

df.show()
```

```
+------+--------+---------------+--------+----+-----+---------+------------+-------------------+
|pos_id|pos_name|        article|quantity|prix|total|sale_type|payment_mode|          sale_time|
+------+--------+---------------+--------+----+-----+---------+------------+-------------------+
|     1|jendouba|Chocolate Eclair|    13.0| 7.5| 97.5|   direct|        cash|2024-03-20 02:53:39|
|     1|jendouba|   Cinnamon Roll|    17.0| 4.0| 68.0|livraison|      online|2024-03-20 02:53:43|
|     1|jendouba|       Lemon Bar|    20.0| 6.0|120.0|   direct|        card|2024-03-20 02:53:44|
|     1|jendouba|        Napoleon|     8.0| 7.9| 63.2|   direct|        cash|2024-03-20 02:53:45|
|     1|jendouba|   Danish Pastry|     2.0| 6.5| 13.0|   direct|        cash|2024-03-20 02:53:48|
|     1|jendouba|         Palmier|     6.0| 3.0| 18.0|livraison|      online|2024-03-20 02:53:49|
|     1|jendouba|      Cream Puff|     6.0| 9.0| 54.0|   direct|        cash|2024-03-20 02:53:51|
|     1|jendouba|      Cream Puff|     3.0| 9.0| 27.0|livraison|      online|2024-03-20 02:53:52|
|     1|jendouba|Chocolate Eclair|     1.0| 7.5|  7.5|livraison|      online|2024-03-20 02:53:56|
|     1|jendouba| Red Velvet Cake|     1.0|12.0| 12.0|livraison|      online|2024-03-20 02:53:59|
|     1|jendouba|       Pecan Pie|    13.0|11.0|143.0|livraison|      online|2024-03-20 02:54:00|
|     1|jendouba|       Croissant|    17.0| 1.5| 25.5|livraison|      online|2024-03-20 02:54:03|
|     1|jendouba|        Baguette|     2.0| 2.0|  4.0|livraison|      online|2024-03-20 02:54:07|
|     1|jendouba|           Scone|    17.0| 2.8| 47.6|   direct|        card|2024-03-20 02:54:08|
|     1|jendouba|       Pecan Pie|     7.0|11.0| 77.0|   direct|        card|2024-03-20 02:54:11|
|     1|jendouba|   Key Lime Tart|    20.0| 8.2|164.0|   direct|        card|2024-03-20 02:54:13|
|     1|jendouba|      Cream Puff|     9.0| 9.0| 81.0|   direct|        card|2024-03-20 02:54:17|
|     1|jendouba|  Apple Turnover|    18.0| 5.0| 90.0|livraison|      online|2024-03-20 02:54:19|
|     1|jendouba|       Croissant|    15.0| 1.5| 22.5|   direct|        cash|2024-03-20 02:54:21|
|     1|jendouba|      Fruit Tart|     6.0| 8.5| 51.0|livraison|      online|2024-03-20 02:54:23|
+------+--------+---------------+--------+----+-----+---------+------------+-------------------+
only showing top 20 rows
```

Postgresql

- PostgreSQL (Version: 42.5.4)

In [ ]:
```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("Read from PostgreSQL") \
    .config('spark.jars.packages', 'org.postgresql:postgresql:42.5.4') \
    .getOrCreate()

postgresql_host = "siliana_host"
postgresql_port = "5432"
postgresql_database = "siliana_sales_db"
postgresql_username = "postgres"
postgresql_password = "secret"
postgresql_table = "siliana_sales"

jdbc_url = f"jdbc:postgresql://{postgresql_host}:{postgresql_port}/{postgresql_database}"

postgresql_properties = {
    "user": postgresql_username,
    "password": postgresql_password,
    "driver": "org.postgresql.Driver"
}

df = spark.read.jdbc(url=jdbc_url, table=postgresql_table, properties=postgresql_properties)

df.show()
```

```
+------+--------+--------------+--------+----+------------------+---------+------------+-------------------+
|pos_id|pos_name|       article|quantity|prix|             total|sale_type|payment_mode|          sale_time|
+------+--------+--------------+--------+----+------------------+---------+------------+-------------------+
|     2| siliana|         Scone|     7.0| 2.8|19.599999999999998|livraison|      online|2024-03-20 02:53:21|
|     2| siliana|    Cherry Pie|     6.0| 9.5|              57.0|   direct|        cash|2024-03-20 02:53:23|
|     2| siliana|Red Velvet Cake|     4.0|12.0|              48.0|livraison|      online|2024-03-20 02:53:25|
|     2| siliana|  Danish Pastry|     1.0| 6.5|               6.5|   direct|        cash|2024-03-20 02:53:26|
|     2| siliana|         Scone|     1.0| 2.8|               2.8|livraison|      online|2024-03-20 02:53:29|
|     2| siliana|    Cherry Pie|    15.0| 9.5|             142.5|   direct|        card|2024-03-20 02:53:31|
|     2| siliana|  Key Lime Tart|    11.0| 8.2| 90.19999999999999|   direct|        cash|2024-03-20 02:53:34|
|     2| siliana| Apple Turnover|    14.0| 5.0|              70.0|   direct|        cash|2024-03-20 02:53:37|
|     2| siliana|    Cherry Pie|    17.0| 9.5|             161.5|   direct|        cash|2024-03-20 02:53:38|
|     2| siliana|     Croissant|     2.0| 1.5|               3.0|   direct|        card|2024-03-20 02:53:40|
|     2| siliana|    Cherry Pie|    12.0| 9.5|             114.0|livraison|      online|2024-03-20 02:53:42|
|     2| siliana|      Pecan Pie|    20.0|11.0|             220.0|livraison|      online|2024-03-20 02:53:44|
|     2| siliana|       Baguette|     4.0| 2.0|               8.0|   direct|        card|2024-03-20 02:53:47|
|     2| siliana|Red Velvet Cake|    13.0|12.0|             156.0|   direct|        card|2024-03-20 02:53:50|
|     2| siliana|        Palmier|     6.0| 3.0|              18.0|livraison|      online|2024-03-20 02:53:51|
|     2| siliana|        Strudel|     9.0| 8.0|              72.0|livraison|      online|2024-03-20 02:53:52|
|     2| siliana|      Pecan Pie|    16.0|11.0|             176.0|livraison|      online|2024-03-20 02:53:54|
|     2| siliana| Apple Turnover|    20.0| 5.0|             100.0|   direct|        cash|2024-03-20 02:53:57|
|     2| siliana|     Cheesecake|     8.0|10.5|              84.0|   direct|        card|2024-03-20 02:53:59|
|     2| siliana|     Fruit Tart|     2.0| 8.5|              17.0|livraison|      online|2024-03-20 02:54:02|
+------+--------+--------------+--------+----+------------------+---------+------------+-------------------+
only showing top 20 rows
```

MongoDB

- MongoDB (Version: 3.0.2)

In [ ]:
```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("Read from Mongo") \
    .config('spark.jars.packages', 'org.mongodb.spark:mongo-spark-connector_2.12:3.0.2') \
    .getOrCreate()

mongo_host = "kef_host"
mongo_port = "27017"
mongo_database = "kef_sales_db"
mongo_username = "mongo"
mongo_password = "mongo"
mongo_collection = "kef_sales"

df = spark.read.format("mongo") \
    .option("uri", f"mongodb://{mongo_username}:{mongo_password}@{mongo_host}:{mongo_port}/") \
    .option("database", mongo_database) \
    .option("collection", mongo_collection) \
    .load()

df.show()
```

```
+--------------------+---------------+------------+------+--------+----+--------+--------------------+---------+-----+
|                 _id|        article|payment_mode|pos_id|pos_name|prix|quantity|           sale_time|sale_type|total|
+--------------------+---------------+------------+------+--------+----+--------+--------------------+---------+-----+
|{65fa4f9b7c4ce0a1...|      Pecan Pie|        card|     1|     kef|11.0|       9|2024-03-20 02:53:...|   direct| 99.0|
|{65fa4faa7c4ce0a1...|  Cinnamon Roll|      online|     1|     kef| 4.0|      17|2024-03-20 02:53:...|livraison| 68.0|
|{65fa4fad7c4ce0a1...|     Cherry Pie|      online|     1|     kef| 9.5|      15|2024-03-20 02:53:...|livraison|142.5|
|{65fa4fb17c4ce0a1...|     Cheesecake|      online|     1|     kef|10.5|       7|2024-03-20 02:53:...|livraison| 73.5|
|{65fa4fb57c4ce0a1...|          Scone|      online|     1|     kef| 2.8|       5|2024-03-20 02:53:...|livraison| 14.0|
|{65fa4fb67c4ce0a1...|  Danish Pastry|      online|     1|     kef| 6.5|      16|2024-03-20 02:53:...|livraison|104.0|
|{65fa4fb97c4ce0a1...|Red Velvet Cake|      online|     1|     kef|12.0|       8|2024-03-20 02:53:...|livraison| 96.0|
|{65fa4fba7c4ce0a1...|     Cream Puff|        cash|     1|     kef| 9.0|      20|2024-03-20 02:53:...|   direct|180.0|
|{65fa4fbb7c4ce0a1...|       Baguette|        cash|     1|     kef| 2.0|      10|2024-03-20 02:53:...|   direct| 20.0|
|{65fa4fbf7c4ce0a1...|        Cupcake|        card|     1|     kef| 4.5|      10|2024-03-20 02:53:...|   direct| 45.0|
|{65fa4fc27c4ce0a1...|      Bear Claw|        card|     1|     kef| 6.8|      12|2024-03-20 02:53:...|   direct| 81.6|
|{65fa4fc67c4ce0a1...|     Cheesecake|      online|     1|     kef|10.5|       2|2024-03-20 02:53:...|livraison| 21.0|
|{65fa4fc97c4ce0a1...|       Baguette|      online|     1|     kef| 2.0|       8|2024-03-20 02:54:...|livraison| 16.0|
|{65fa4fcb7c4ce0a1...|      Lemon Bar|      online|     1|     kef| 6.0|      10|2024-03-20 02:54:...|livraison| 60.0|
|{65fa4fcf7c4ce0a1...|     Cherry Pie|        card|     1|     kef| 9.5|       5|2024-03-20 02:54:...|   direct| 47.5|
|{65fa4fd27c4ce0a1...|  Cinnamon Roll|      online|     1|     kef| 4.0|      18|2024-03-20 02:54:...|livraison| 72.0|
|{65fa4fd57c4ce0a1...|        Cupcake|      online|     1|     kef| 4.5|       5|2024-03-20 02:54:...|livraison| 22.5|
|{65fa4fd77c4ce0a1...|        Palmier|      online|     1|     kef| 3.0|       1|2024-03-20 02:54:...|livraison|  3.0|
|{65fa4fda7c4ce0a1...|      Bear Claw|        card|     1|     kef| 6.8|       1|2024-03-20 02:54:...|   direct|  6.8|
|{65fa4fdb7c4ce0a1...|      Croissant|        card|     1|     kef| 1.5|       4|2024-03-20 02:54:...|   direct|  6.0|
+--------------------+---------------+------------+------+--------+----+--------+--------------------+---------+-----+
only showing top 20 rows
```

Cassandra

- Cassandra (Version: 3.2.0)

In [ ]:
```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("Read from Cassandra") \
    .config("spark.jars.packages", "com.datastax.spark:spark-cassandra-connector_2.12:3.2.0") \
    .config("spark.cassandra.connection.host", "cassandra") \
    .config("spark.cassandra.connection.port", "9042") \
    .config("spark.cassandra.auth.username", "cassandra") \
    .config("spark.cassandra.auth.password", "cassandra") \
    .getOrCreate()

df = spark.read.format("org.apache.spark.sql.cassandra") \
    .options(table="ariana_sales", keyspace="ariana_sales_db") \
    .load()

df.show()
```

```
+--------------------+----------------+------------+------+--------+----+--------+-------------------+---------+-----+
|             sale_id|         article|payment_mode|pos_id|pos_name|prix|quantity|          sale_time|sale_type|total|
+--------------------+----------------+------------+------+--------+----+--------+-------------------+---------+-----+
|2ca9ce16-d9f9-4df...|   Danish Pastry|      online|     1|  ariana| 6.5|     6.0|2024-03-20 02:59:04|livraison| 39.0|
|f39cd47c-ab05-4ce...|   Danish Pastry|      online|     1|  ariana| 6.5|     7.0|2024-03-20 03:00:54|livraison| 45.5|
|64fa9ff5-d470-4e0...|         Strudel|        card|     1|  ariana| 8.0|     6.0|2024-03-20 03:00:00|   direct| 48.0|
|028f4c93-8a85-45d...|       Bear Claw|        cash|     1|  ariana| 6.8|    14.0|2024-03-20 03:03:48|   direct| 95.2|
|00e8799a-6265-45a...|         Cupcake|        card|     1|  ariana| 4.5|    16.0|2024-03-20 03:02:25|   direct| 72.0|
|e5d2f4b7-bc8a-475...|           Scone|        cash|     1|  ariana| 2.8|    20.0|2024-03-20 03:00:26|   direct| 56.0|
|2f89cc7d-d83f-4b6...|   Key Lime Tart|        card|     1|  ariana| 8.2|    20.0|2024-03-20 02:58:17|   direct|164.0|
|6a325d72-4b2f-416...|   Key Lime Tart|      online|     1|  ariana| 8.2|    12.0|2024-03-20 02:56:12|livraison| 98.4|
|52c29b14-fc9e-4b0...|   Danish Pastry|      online|     1|  ariana| 6.5|     8.0|2024-03-20 02:57:32|livraison| 52.0|
|3aabcb03-f25f-4dd...|       Lemon Bar|        cash|     1|  ariana| 6.0|     5.0|2024-03-20 02:59:44|   direct| 30.0|
|2617c334-891b-46d...|   Cinnamon Roll|      online|     1|  ariana| 4.0|     7.0|2024-03-20 03:01:09|livraison| 28.0|
|6b7e200c-9324-44e...|        Baguette|        card|     1|  ariana| 2.0|    18.0|2024-03-20 02:57:39|   direct| 36.0|
|d59b4fa9-c0ff-4fe...|           Scone|        card|     1|  ariana| 2.8|    18.0|2024-03-20 03:05:11|   direct| 50.4|
|66a89219-31fa-489...|       Lemon Bar|      online|     1|  ariana| 6.0|     9.0|2024-03-20 03:03:11|livraison| 54.0|
|5285ae9f-d12f-4a9...|   Cinnamon Roll|      online|     1|  ariana| 4.0|    13.0|2024-03-20 02:55:32|livraison| 52.0|
|99c63254-54dd-420...|       Croissant|        card|     1|  ariana| 1.5|    12.0|2024-03-20 03:00:28|   direct| 18.0|
|0e1b50d3-6c71-4f6...|Chocolate Eclair|        card|     1|  ariana| 7.5|    13.0|2024-03-20 03:02:18|   direct| 97.5|
|8a9cbc33-13ac-452...|      Cheesecake|      online|     1|  ariana|10.5|    17.0|2024-03-20 03:02:47|livraison|178.5|
|613a5f2a-dfe0-43f...|        Napoleon|      online|     1|  ariana| 7.9|     9.0|2024-03-20 02:58:58|livraison| 71.1|
|249cd58b-3379-4b6...|       Bear Claw|      online|     1|  ariana| 6.8|    14.0|2024-03-20 03:00:58|livraison| 95.2|
+--------------------+----------------+------------+------+--------+----+--------+-------------------+---------+-----+
only showing top 20 rows
```

Kafka

- Kafka (Version: 3.4.1)

In [ ]:
```
# version read batch data

from pyspark.sql import SparkSession
from pyspark.sql.functions import from_json, col
from pyspark.sql.types import StructType, StructField, StringType, IntegerType, FloatType

schema = StructType([
    StructField("sale_id", StringType()),
    StructField("pos_name", StringType()),
    StructField("pos_id", IntegerType()),
    StructField("article", StringType()),
    StructField("quantity", FloatType()),
    StructField("prix", FloatType()),
    StructField("total", FloatType()),
    StructField("sale_type", StringType()),
    StructField("payment_mode", StringType()),
    StructField("sale_time", StringType()),
])

spark = SparkSession.builder \
    .appName("KafkaConsumer") \
    .config("spark.jars.packages",
            "org.apache.spark:spark-sql-kafka-0-10_2.12:3.2.1," +
            "org.apache.kafka:kafka-clients:3.4.1") \
    .getOrCreate()

input_df = spark \
    .read \
    .format("kafka") \
    .option("kafka.bootstrap.servers", "kafka:9092") \
    .option("subscribe", "kairouan_sales") \
    .option("startingOffsets", "earliest") \
    .load()

expanded_df = input_df \
    .selectExpr("CAST(value AS STRING)") \
    .select(from_json(col("value"), schema).alias("kairouan_sales")) \
    .select("kairouan_sales.*")

expanded_df.show()
```

```
+--------------------+--------+------+--------------+--------+----+-----+---------+------------+-------------------+
|             sale_id|pos_name|pos_id|       article|quantity|prix|total|sale_type|payment_mode|          sale_time|
+--------------------+--------+------+--------------+--------+----+-----+---------+------------+-------------------+
|367cfce8-068f-472...|kairouan|     9|     Bear Claw|     1.0| 6.8|  6.8|   direct|        cash|2024-03-20 02:53:15|
|5dacb383-e2c2-44e...|kairouan|     9|Red Velvet Cake|     6.0|12.0| 72.0|livraison|      online|2024-03-20 02:53:29|
|a215243e-c2b9-4a7...|kairouan|     9|     Croissant|    19.0| 1.5| 28.5|   direct|        card|2024-03-20 02:53:30|
|6dfe0a3b-62a1-4ea...|kairouan|     9|    Cheesecake|     9.0|10.5| 94.5|   direct|        cash|2024-03-20 02:53:34|
|6fb0b47a-2c89-42a...|kairouan|     9|     Pecan Pie|     4.0|11.0| 44.0|   direct|        cash|2024-03-20 02:53:37|
|433d7041-25e1-43f...|kairouan|     9|      Napoleon|    16.0| 7.9|126.4|   direct|        card|2024-03-20 02:53:38|
|e30c3959-81ec-499...|kairouan|     9|      Napoleon|    15.0| 7.9|118.5|   direct|        card|2024-03-20 02:53:40|
|c50ae742-4415-485...|kairouan|     9| Danish Pastry|    15.0| 6.5| 97.5|livraison|      online|2024-03-20 02:53:42|
|f197d1d2-b862-493...|kairouan|     9| Danish Pastry|    18.0| 6.5|117.0|   direct|        cash|2024-03-20 02:53:46|
|255b8b1c-e7b7-428...|kairouan|     9|    Cherry Pie|     2.0| 9.5| 19.0|livraison|      online|2024-03-20 02:53:50|
|d6a24eb6-1265-441...|kairouan|     9|    Cheesecake|    14.0|10.5|147.0|livraison|      online|2024-03-20 02:53:53|
|ff03035c-e550-4fb...|kairouan|     9|    Cheesecake|     7.0|10.5| 73.5|   direct|        cash|2024-03-20 02:53:56|
|e9054906-b9f6-4ca...|kairouan|     9|     Bear Claw|    16.0| 6.8|108.8|livraison|      online|2024-03-20 02:54:00|
|b947b60d-ca42-452...|kairouan|     9|     Bear Claw|     4.0| 6.8| 27.2|livraison|      online|2024-03-20 02:54:04|
|3708b2d7-6c40-47c...|kairouan|     9|    Cherry Pie|     9.0| 9.5| 85.5|livraison|      online|2024-03-20 02:54:07|
|908a6f24-18e3-4ab...|kairouan|     9|     Lemon Bar|    17.0| 6.0|102.0|livraison|      online|2024-03-20 02:54:10|
|974d2317-108a-4cc...|kairouan|     9|    Cream Puff|    16.0| 9.0|144.0|   direct|        card|2024-03-20 02:54:12|
|01a3bfdc-72e6-42f...|kairouan|     9| Cinnamon Roll|    13.0| 4.0| 52.0|livraison|      online|2024-03-20 02:54:13|
|6d69414b-c567-4af...|kairouan|     9|     Lemon Bar|     4.0| 6.0| 24.0|livraison|      online|2024-03-20 02:54:17|
|9dee4ddb-1295-484...|kairouan|     9| Cinnamon Roll|     3.0| 4.0| 12.0|livraison|      online|2024-03-20 02:54:19|
+--------------------+--------+------+--------------+--------+----+-----+---------+------------+-------------------+
only showing top 20 rows
```

In [ ]:

```
# version stream data

from pyspark.sql import SparkSession
from pyspark.sql.functions import from_json, col
from pyspark.sql.types import StructType, StructField, StringType, IntegerType, FloatType

schema = StructType([
    StructField("sale_id", StringType()),
    StructField("pos_name", StringType()),
    StructField("pos_id", IntegerType()),
    StructField("article", StringType()),
    StructField("quantity", FloatType()),
    StructField("prix", FloatType()),
    StructField("total", FloatType()),
    StructField("sale_type", StringType()),
    StructField("payment_mode", StringType()),
    StructField("sale_time", StringType()),
])

spark = SparkSession.builder \
    .appName("KafkaConsumer") \
    .config("spark.jars.packages",
            "org.apache.spark:spark-sql-kafka-0-10_2.12:3.2.1," +
            "org.apache.kafka:kafka-clients:3.4.1") \
    .getOrCreate()

input_df = spark \
    .readStream \
    .format("kafka") \
    .option("kafka.bootstrap.servers", "kafka:9092") \
    .option("subscribe", "kairouan_sales") \
    .option("startingOffsets", "earliest") \
    .load()

expanded_df = input_df \
    .selectExpr("CAST(value AS STRING)") \
    .select(from_json(col("value"), schema).alias("kairouan_sales")) \
    .select("kairouan_sales.*") \

query1 = expanded_df \
    .writeStream \
    .outputMode("update") \
    .format("console") \
    .start()


query1.awaitTermination()
```