```
#check that java is installed
!java -version
```

```
openjdk version "11.0.27" 2025-04-15
OpenJDK Runtime Environment (build 11.0.27+6-post-Ubuntu-0ubuntu122.04)
OpenJDK 64-Bit Server VM (build 11.0.27+6-post-Ubuntu-0ubuntu122.04, mixed mode, sharing)
```

```
#install pyspark
!pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.11/dist-packages (3.5.1)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.11/dist-packages (from pyspark) (0.10.9.7)
```

```
import os
import pandas as pd
from pyspark.sql import SparkSession
from pyspark.sql.types import StructType, StructField, StringType, IntegerType, DoubleType, LongType, TimestampType
from pyspark.sql.functions import spark_partition_id
```

```
!curl -O https://raw.githubusercontent.com/Apoorva-888/Spark-Optimization/main/BigMart_Sales.csv
```

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  849k  100  849k    0     0  2019k      0 --:--:-- --:--:-- --:--:-- 2017k
```

```
spark = SparkSession.builder.appName('AutomotiveData').getOrCreate()

print(f'The Spark version is {spark.version}')
```

```
The Spark version is 3.5.1
```

```
spark.conf.set("spark.sql.adaptive.enabled","false")
spark.conf.get("spark.sql.adaptive.enabled")
```

```
'false'
```

```
spark.conf.get("spark.sql.adaptive.enabled")
```

```
'false'
```

```
df = spark.read.format("csv") \
    .option("inferSchema", True) \
    .option("header", True) \
    .load("BigMart_Sales.csv")

df.show(5)
```

```
+---------------+-----------+----------------+---------------+--------------------+--------+-----------------+-------------------------+
|Item_Identifier|Item_Weight|Item_Fat_Content|Item_Visibility|           Item_Type|Item_MRP|Outlet_Identifier|Outlet_Establishment_Year|
+---------------+-----------+----------------+---------------+--------------------+--------+-----------------+-------------------------+
|          FDA15|        9.3|         Low Fat|    0.016047301|               Dairy|249.8092|           OUT049|                     1999|
|          DRC01|       5.92|         Regular|    0.019278216|         Soft Drinks| 48.2692|           OUT018|                     2009|
|          FDN15|       17.5|         Low Fat|    0.016760075|                Meat| 141.618|           OUT049|                     1999|
|          FDX07|       19.2|         Regular|            0.0|Fruits and Vegeta...| 182.095|           OUT010|                     1998|
|          NCD19|       8.93|         Low Fat|            0.0|           Household| 53.8614|           OUT013|                     1987|
+---------------+-----------+----------------+---------------+--------------------+--------+-----------------+-------------------------+
only showing top 5 rows
```

```
df.printSchema()
```

```
root
 |-- Item_Identifier: string (nullable = true)
 |-- Item_Weight: double (nullable = true)
 |-- Item_Fat_Content: string (nullable = true)
 |-- Item_Visibility: double (nullable = true)
 |-- Item_Type: string (nullable = true)
 |-- Item_MRP: double (nullable = true)
 |-- Outlet_Identifier: string (nullable = true)
 |-- Outlet_Establishment_Year: integer (nullable = true)
 |-- Outlet_Size: string (nullable = true)
 |-- Outlet_Location_Type: string (nullable = true)
 |-- Outlet_Type: string (nullable = true)
 |-- Item_Outlet_Sales: double (nullable = true)
```

```
!ls -lh BigMart_Sales.csv
```

```
-rw-r--r-- 1 root root 850K Jul 17 14:21 BigMart_Sales.csv
```

```
spark.conf.get("spark.sql.files.maxPartitionBytes")
```

```
'134217728b'
```

```
 df.rdd.getNumPartitions()
```

```
1
```

**Changing default partition to 128kb**

```
Start coding or generate with AI.
```

```
spark.conf.set("spark.sql.files.maxPartitionBytes", 131072)
df = spark.read.format("csv").option("inferSchema", True).option("header", True).load("BigMart_Sales.csv")
df.rdd.getNumPartitions()
```

```
7
```

**Changing default partition to 128MB**

```
spark.conf.set("spark.sql.files.maxPartitionBytes", 1024 * 1024 * 1)
df = spark.read.format("csv").option("inferSchema", True).option("header", True).load("BigMart_Sales.csv")
df.rdd.getNumPartitions()
```

```
1
```

**Repartition**

```
df=df.repartition(10)
df.rdd.getNumPartitions()
```

```
10
```

```
#Write a DataFrame as Parquet in append mode in Google Colab.
df.write.format("parquet").mode("append").option("path", "/content/parquetWrite").save()
```

```
#After Saving: You can verify the files were written:
!ls -lh /content/parquetWrite
```

```
total 324K
-rw-r--r-- 1 root root 32K Jul 17 14:22 part-00000-383227c2-52d8-4466-afe5-868d48f6a786-c000.snappy.parquet
-rw-r--r-- 1 root root 31K Jul 17 14:22 part-00001-383227c2-52d8-4466-afe5-868d48f6a786-c000.snappy.parquet
-rw-r--r-- 1 root root 31K Jul 17 14:22 part-00002-383227c2-52d8-4466-afe5-868d48f6a786-c000.snappy.parquet
-rw-r--r-- 1 root root 32K Jul 17 14:22 part-00003-383227c2-52d8-4466-afe5-868d48f6a786-c000.snappy.parquet
-rw-r--r-- 1 root root 31K Jul 17 14:22 part-00004-383227c2-52d8-4466-afe5-868d48f6a786-c000.snappy.parquet
-rw-r--r-- 1 root root 33K Jul 17 14:22 part-00005-383227c2-52d8-4466-afe5-868d48f6a786-c000.snappy.parquet
-rw-r--r-- 1 root root 32K Jul 17 14:22 part-00006-383227c2-52d8-4466-afe5-868d48f6a786-c000.snappy.parquet
-rw-r--r-- 1 root root 32K Jul 17 14:22 part-00007-383227c2-52d8-4466-afe5-868d48f6a786-c000.snappy.parquet
-rw-r--r-- 1 root root 32K Jul 17 14:22 part-00008-383227c2-52d8-4466-afe5-868d48f6a786-c000.snappy.parquet
-rw-r--r-- 1 root root 31K Jul 17 14:22 part-00009-383227c2-52d8-4466-afe5-868d48f6a786-c000.snappy.parquet
-rw-r--r-- 1 root root   0 Jul 17 14:22 _SUCCESS
```

```
#Reading It Back:
df2 = spark.read.parquet("/content/parquetWrite")
df2.show()
```

```
+---------------+-----------+----------------+---------------+--------------------+--------+-----------------+-------------------------+
|Item_Identifier|Item_Weight|Item_Fat_Content|Item_Visibility|           Item_Type|Item_MRP|Outlet_Identifier|Outlet_Establishment_Year|
+---------------+-----------+----------------+---------------+--------------------+--------+-----------------+-------------------------+
|          FDU21|       NULL|         Regular|    0.134327613|         Snack Foods| 35.0558|           OUT019|                     1985|
|         FDN56|       5.46|         Regular|    0.106968096|Fruits and Vegeta...|142.6786|           OUT013|                     1987|
|         NCJ17|       7.68|         Low Fat|    0.153177941|  Health and Hygiene| 85.2224|           OUT018|                     2009|
|         FDG08|       NULL|         Regular|    0.289522833|Fruits and Vegeta...|172.0764|           OUT019|                     1985|
|         NCZ41|      19.85|         Low Fat|    0.064409056|  Health and Hygiene|126.7704|           OUT035|                     2004|
|         FDI60|       7.22|         Regular|      0.0383808|        Baking Goods|  62.351|           OUT049|                     1999|
|         FDR36|      6.715|             reg|    0.122274118|        Baking Goods| 40.3454|           OUT017|                     2007|
|         FDZ12|       9.17|         Low Fat|    0.102978817|        Baking Goods| 144.947|           OUT046|                     1997|
|         NCD07|        9.1|         Low Fat|    0.055382616|           Household|115.0518|           OUT013|                     1987|
|         DRG39|      14.15|         Low Fat|    0.042352822|               Dairy| 51.6982|           OUT018|                     2009|
|         FDR19|       13.5|         Regular|    0.160624116|Fruits and Vegeta...|147.0102|           OUT017|                     2007|
|         FDK56|      9.695|         Low Fat|    0.130261652|Fruits and Vegeta...|185.2898|           OUT045|                     2002|
|         NCM07|       NULL|         Low Fat|     0.03976832|              Others| 83.9908|           OUT027|                     1985|
|         NCI30|      20.25|         Low Fat|    0.059055045|           Household| 247.346|           OUT045|                     2002|
|         FDZ35|        9.6|         Regular|     0.02236923|              Breads| 104.799|           OUT018|                     2009|
|         FDB14|      20.25|         Regular|    0.103142373|              Canned|  94.612|           OUT018|                     2009|
|         NCN05|      8.235|         Low Fat|    0.014541462|  Health and Hygiene| 184.495|           OUT017|                     2007|
|         FDZ38|       17.6|              LF|    0.008001018|               Dairy|170.4422|           OUT046|                     1997|
|         FDM38|      5.885|         Regular|    0.092694107|              Canned| 53.6982|           OUT013|                     1987|
|         DRK13|       11.8|         Low Fat|    0.115346634|         Soft Drinks|200.2084|           OUT049|                     1999|
+--------------+-----------+----------------+---------------+--------------------+--------+-----------------+-------------------------+
only showing top 20 rows
```

```python
from pyspark.sql.functions import spark_partition_id
df_with_partition = df.withColumn("partition_id", spark_partition_id())
df_with_partition.show(truncate=False)  # Shows rows with partition info
df_with_partition.groupBy("partition_id").count().show()
```

```
+---------------+-----------+----------------+---------------+------------------+--------+-----------------+-------------------------+--
|Item_Identifier|Item_Weight|Item_Fat_Content|Item_Visibility|Item_Type         |Item_MRP|Outlet_Identifier|Outlet_Establishment_Year|Ou
+---------------+-----------+----------------+---------------+------------------+--------+-----------------+-------------------------+--
|DRG48          |5.78       |Low Fat         |0.014555066    |Soft Drinks       |145.2102|OUT046           |1997                     |Sm
|FDE24          |14.85      |Low Fat         |0.09344495     |Baking Goods      |141.0812|OUT035           |2004                     |Sm
|NCP18          |12.15      |Low Fat         |0.028714747    |Household         |151.9708|OUT018           |2009                     |Me
|FDX27          |20.7       |Regular         |0.114581955    |Dairy             |94.3436 |OUT018           |2009                     |Me
|FDU02          |13.35      |Low Fat         |0.102670882    |Dairy             |228.6352|OUT049           |1999                     |Me
|FDW57          |8.31       |Regular         |0.115911972    |Snack Foods       |177.3028|OUT045           |2002                     |NU
|FDZ01          |8.975      |Regular         |0.009057132    |Canned            |104.099 |OUT035           |2004                     |Sm
|FDL39          |16.1       |Regular         |0.063689583    |Dairy             |181.9318|OUT017           |2007                     |NU
|FDL43          |10.1       |Low Fat         |0.027064381    |Meat              |76.367  |OUT046           |1997                     |Sm
|FDP23          |NULL       |Low Fat         |0.035414528    |Breads            |218.2166|OUT027           |1985                     |Me
|FDL15          |17.85      |Low Fat         |0.046824729    |Meat              |153.6682|OUT018           |2009                     |Me
|FDB23          |19.2       |Regular         |0.0            |Starchy Foods     |223.8062|OUT045           |2002                     |NU
|FDY39          |NULL       |Regular         |0.08234117     |Meat              |185.7608|OUT019           |1985                     |Sm
|FDG05          |11.0       |Regular         |0.0            |Frozen Foods      |155.263 |OUT049           |1999                     |Me
|NCS29          |9.0        |Low Fat         |0.069653585    |Health and Hygiene|266.2884|OUT049           |1999                     |Me
|FDT01          |13.65      |Regular         |0.184454044    |Canned            |211.4902|OUT049           |1999                     |Me
|NCJ05          |18.7       |Low Fat         |0.046348967    |Health and Hygiene|153.6682|OUT017           |2007                     |NU
|NCX42          |6.36       |Low Fat         |0.00599072     |Household         |163.6526|OUT045           |2002                     |NU
|FDF10          |15.5       |reg             |0.156797787    |Snack Foods       |148.6418|OUT013           |1987                     |Hi
|FDZ15          |13.1       |Low Fat         |0.0            |Dairy             |117.8782|OUT035           |2004                     |Sm
+---------------+-----------+----------------+---------------+------------------+--------+-----------------+-------------------------+--
only showing top 20 rows

+------------+-----+
|partition_id|count|
+------------+-----+
|           1|  853|
|           6|  852|
|           3|  852|
|           5|  852|
|           9|  853|
|           4|  852|
|           8|  852|
|           7|  852|
|           2|  852|
|           0|  853|
+------------+-----+
```

## coalesce and repartition

```python
df2 = spark.read.parquet("/content/parquetWrite")
df2.show()
```

```
+---------------+-----------+----------------+---------------+--------------------+--------+-----------------+-------------------------+
|Item_Identifier|Item_Weight|Item_Fat_Content|Item_Visibility|           Item_Type|Item_MRP|Outlet_Identifier|Outlet_Establishment_Year|
+---------------+-----------+----------------+--------------+--------------------+--------+-----------------+-------------------------+
|          FDU21|       NULL|         Regular|   0.134327613|         Snack Foods| 35.0558|           OUT019|                     1985|
|          FDN56|       5.46|         Regular|   0.106968096|Fruits and Vegeta...|142.6786|           OUT013|                     1987|
|          NCJ17|       7.68|         Low Fat|   0.153177941|  Health and Hygiene| 85.2224|           OUT018|                     2009|
|          FDG08|       NULL|         Regular|   0.289522833|Fruits and Vegeta...|172.0764|           OUT019|                     1985|
|          NCZ41|      19.85|         Low Fat|   0.064409056|  Health and Hygiene|126.7704|           OUT035|                     2004|
|          FDI60|       7.22|         Regular|     0.0383808|        Baking Goods|  62.351|           OUT049|                     1999|
|          FDR36|      6.715|             reg|   0.122274118|        Baking Goods| 40.3454|           OUT017|                     2007|
|          FDZ12|       9.17|         Low Fat|   0.102978817|        Baking Goods| 144.947|           OUT046|                     1997|
|          NCD07|        9.1|         Low Fat|   0.055382616|           Household|115.0518|           OUT013|                     1987|
|          DRG39|      14.15|         Low Fat|   0.042352822|               Dairy| 51.6982|           OUT018|                     2009|
|          FDR19|       13.5|         Regular|   0.160624116|Fruits and Vegeta...|147.0102|           OUT017|                     2007|
|          FDK56|      9.695|         Low Fat|   0.130261652|Fruits and Vegeta...|185.2898|           OUT045|                     2002|
|          NCM07|       NULL|         Low Fat|    0.03976832|              Others| 83.9908|           OUT027|                     1985|
|          NCI30|      20.25|         Low Fat|   0.059055045|           Household| 247.346|           OUT045|                     2002|
|          FDZ35|        9.6|         Regular|    0.02236923|              Breads| 104.799|           OUT018|                     2009|
|          FDB14|      20.25|         Regular|   0.103142373|              Canned|  94.612|           OUT018|                     2009|
|          NCN05|      8.235|         Low Fat|   0.014541462|  Health and Hygiene| 184.495|           OUT017|                     2007|
|          FDZ38|       17.6|              LF|   0.008001018|               Dairy|170.4422|           OUT046|                     1997|
|          FDM38|      5.885|         Regular|   0.092694107|              Canned| 53.6982|           OUT013|                     1987|
|          DRK13|       11.8|         Low Fat|   0.115346634|         Soft Drinks|200.2084|           OUT049|                     1999|
+---------------+-----------+----------------+--------------+--------------------+--------+-----------------+-------------------------+
only showing top 20 rows
```

```python
from pyspark.sql.functions import *
df3=df2.filter(col("Outlet_Location_Type")=='Tier 1')
```

```python
from pyspark.sql.functions import *
df4=df2.filter(col("Outlet_Location_Type")=='Tier 2')
```

```python
df2_cache = spark.read.parquet("/content/parquetWrite").cache()
from pyspark.sql.functions import *
df42=df2_cache.filter(col("Outlet_Location_Type")=='Tier 2')
df42.show()
```

```
+---------------+-----------+----------------+--------------+--------------------+--------+-----------------+-------------------------+
|Item_Identifier|Item_Weight|Item_Fat_Content|Item_Visibility|           Item_Type|Item_MRP|Outlet_Identifier|Outlet_Establishment_Year|
+---------------+-----------+----------------+--------------+--------------------+--------+-----------------+-------------------------+
|          NCZ41|      19.85|         Low Fat|   0.064409056|  Health and Hygiene|126.7704|           OUT035|                     2004|
|          FDR36|      6.715|             reg|   0.122274118|        Baking Goods| 40.3454|           OUT017|                     2007|
|          FDR19|       13.5|         Regular|   0.160624116|Fruits and Vegeta...|147.0102|           OUT017|                     2007|
|          FDK56|      9.695|         Low Fat|   0.130261652|Fruits and Vegeta...|185.2898|           OUT045|                     2002|
|          NCI30|      20.25|         Low Fat|   0.059055045|           Household| 247.346|           OUT045|                     2002|
|          NCN05|      8.235|         Low Fat|   0.014541462|  Health and Hygiene| 184.495|           OUT017|                     2007|
|          FDU02|      13.35|         Low Fat|     0.1027194|               Dairy|228.8352|           OUT045|                     2002|
|          DRI13|      15.35|         Low Fat|   0.020441939|         Soft Drinks|216.4508|           OUT017|                     2007|
|          FDO08|       11.1|         Regular|   0.054079556|Fruits and Vegeta...|165.9526|           OUT017|                     2007|
|          FDJ09|       15.0|         low fat|           0.0|         Snack Foods| 47.2744|           OUT035|                     2004|
|          FDB26|       14.0|             reg|   0.031444357|              Canned|  53.764|           OUT017|                     2007|
|          FDB21|      7.475|         Low Fat|   0.149360698|Fruits and Vegeta...|243.4854|           OUT017|                     2007|
|          NCL41|      12.35|         Low Fat|   0.041973712|  Health and Hygiene| 35.7216|           OUT017|                     2007|
|          NCL31|       7.39|         Low Fat|   0.120524922|              Others| 142.247|           OUT045|                     2002|
|          FDT40|      5.985|         Low Fat|           0.0|        Frozen Foods|125.2678|           OUT035|                     2004|
|          NCN55|       14.6|         Low Fat|   0.059827007|              Others|239.2538|           OUT017|                     2007|
|          FDB51|       6.92|              LF|   0.038671588|               Dairy| 64.2852|           OUT017|                     2007|
|          DRG15|       6.13|         Low Fat|   0.076891526|               Dairy| 61.5536|           OUT045|                     2002|
|          FDF22|      6.865|         Low Fat|   0.056819936|         Snack Foods|212.6218|           OUT035|                     2004|
|          FDX50|       20.1|         Low Fat|   0.075049323|               Dairy|110.4228|           OUT017|                     2007|
+---------------+-----------+----------------+--------------+--------------------+--------+-----------------+-------------------------+
only showing top 20 rows
```

```python
df2_cache.unpersist()
```

```
DataFrame[Item_Identifier: string, Item_Weight: double, Item_Fat_Content: string, Item_Visibility: double, Item_Type: string, Item_MRP:
double, Outlet_Identifier: string, Outlet_Establishment_Year: int, Outlet_Size: string, Outlet_Location_Type: string, Outlet_Type:
string, Item_Outlet_Sales: double]
```

```python
from pyspark import StorageLevel
```

```python
df3.persist(StorageLevel.MEMORY_AND_DISK)
```

```
DataFrame[Item_Identifier: string, Item_Weight: double, Item_Fat_Content: string, Item_Visibility: double, Item_Type: string, Item_MRP:
double, Outlet_Identifier: string, Outlet_Establishment_Year: int, Outlet_Size: string, Outlet_Location_Type: string, Outlet_Type:
string, Item_Outlet_Sales: double]
```

```
# Default: MEMORY_AND_DISK (if memory full, spills to disk)
df3.persist(StorageLevel.MEMORY_AND_DISK)

# Other examples:
# df.persist(StorageLevel.MEMORY_ONLY)         # Stores only in memory, errors if not enough space
# df.persist(StorageLevel.DISK_ONLY)           # Skips memory, directly caches to disk
# df.persist(StorageLevel.MEMORY_AND_DISK_SER) # Serialized format, less memory, more CPU

df3.count()  # triggers cache
```

```
2388
```

```
Start coding or generate with AI.
```