spark-2-16-26

February 15, 2026

1 Some Spark Things

[1]:

```
import pyspark
pyspark.__version__
```

[1]:

```
'4.1.1'
```

[3]:

```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("MyDemo") \
    .getOrCreate()
```

```
WARNING: Using incubator modules: jdk.incubator.vector
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
26/02/15 15:55:04 WARN NativeCodeLoader: Unable to load native-hadoop library
for your platform… using builtin-java classes where applicable
26/02/15 15:55:06 WARN Utils: Service 'SparkUI' could not bind on port 4040.
Attempting port 4041.
26/02/15 15:55:06 WARN Utils: Service 'SparkUI' could not bind on port 4041.
Attempting port 4042.
```

[4]:

```
df = spark.read.parquet(
    "hdfs://pi1.knoxds.org:8020/datasets/yellow_tripdata_2025-01.parquet"
)
```

[9]:

```
df.columns
```

[9]:

```
['VendorID',
 'tpep_pickup_datetime',
 'tpep_dropoff_datetime',
 'passenger_count',
 'trip_distance',
 'RatecodeID',
 'store_and_fwd_flag',
 'PULocationID',
 'DOLocationID',
 'payment_type',
 'fare_amount',
 'extra',
 'mta_tax',
 'tip_amount',
 'tolls_amount',
 'improvement_surcharge',
 'total_amount',
 'congestion_surcharge',
 'Airport_fee',
 'cbd_congestion_fee']
```

[10]:

```
df.count()
```

[10]: 3475226

[12]:

```
df.filter(df.passenger_count > 4).count()
```

[12]: 29808

[14]:

```
mydf = df.filter(df.passenger_count > 4).select('passenger_count')
```

[15]:

```
mydf.count()
```

[15]: 29808

[16]:

```
mydf.columns
```

[16]: ['passenger_count']

[21]:

```
import pyspark.sql.functions as sf
mydf.select(sf.avg('passenger_count')).show()
```

```
+--------------------+
|avg(passenger_count)|
+--------------------+
|   5.404488727858293|
+--------------------+
```

[22]:

```
mydf.select(sf.avg('passenger_count'))
```

[22]: DataFrame[avg(passenger_count): double]

```
[ ]:
```