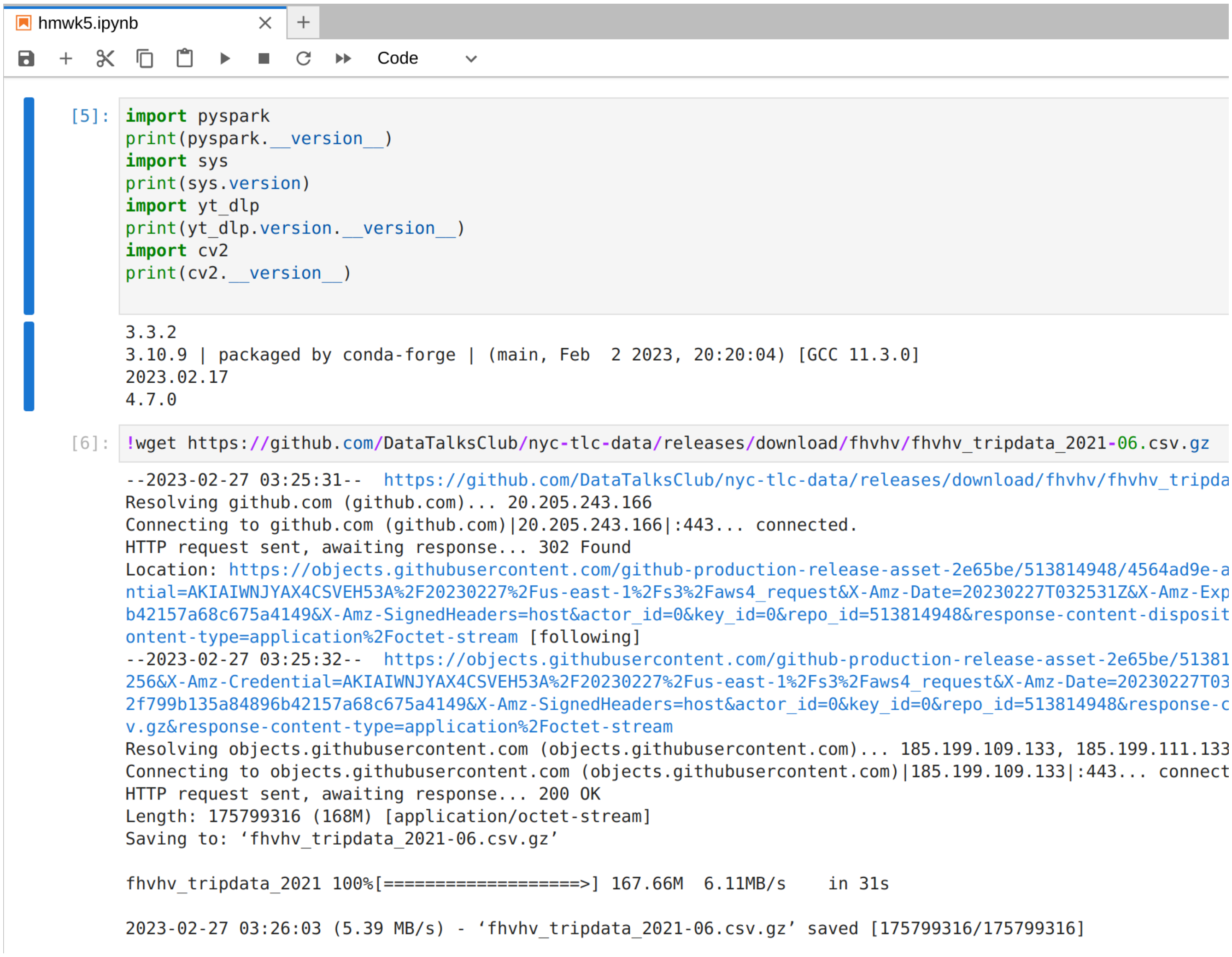


```
[5]: import pyspark
     print(pyspark.__version__)
     import sys
     print(sys.version)
     import yt_dlp
     print(yt_dlp.version.__version__)
     import cv2
     print(cv2.__version__)
```

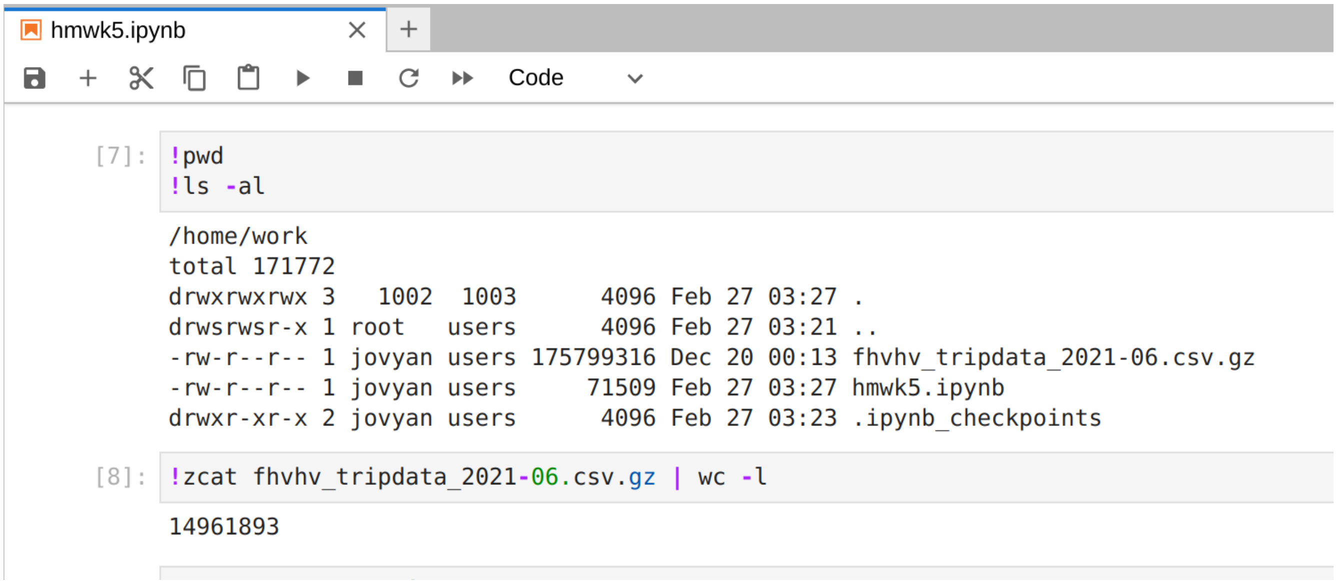
```
3.3.2
3.10.9 | packaged by conda-forge | (main, Feb  2 2023, 20:20:04) [GCC 11.3.0]
2023.02.17
4.7.0
```

```
[6]: !wget https://github.com/DataTalksClub/nyc-tlc-data/releases/download/fhvhv/fhvhv_tripdata_2021-06.csv.gz
```

```
--2023-02-27 03:25:31--  https://github.com/DataTalksClub/nyc-tlc-data/releases/download/fhvhv/fhvhv_tripda
Resolving github.com (github.com)... 20.205.243.166
Connecting to github.com (github.com)|20.205.243.166|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://objects.githubusercontent.com/github-production-release-asset-2e65be/513814948/4564ad9e-a
ntial=AKIAIWNJYAX4CSVEH53A%2F20230227%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20230227T032531Z&X-Amz-Exp
b42157a68c675a4149&X-Amz-SignedHeaders=host&actor_id=0&key_id=0&repo_id=513814948&response-content-disposit
ontent-type=application%2Foctet-stream [following]
--2023-02-27 03:25:32--  https://objects.githubusercontent.com/github-production-release-asset-2e65be/51381
256&X-Amz-Credential=AKIAIWNJYAX4CSVEH53A%2F20230227%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20230227T03
2f799b135a84896b42157a68c675a4149&X-Amz-SignedHeaders=host&actor_id=0&key_id=0&repo_id=513814948&response-c
v.gz&response-content-type=application%2Foctet-stream
Resolving objects.githubusercontent.com (objects.githubusercontent.com)... 185.199.109.133, 185.199.111.133
Connecting to objects.githubusercontent.com (objects.githubusercontent.com)|185.199.109.133|:443... connect
HTTP request sent, awaiting response... 200 OK
Length: 175799316 (168M) [application/octet-stream]
Saving to: ‘fhvhv_tripdata_2021-06.csv.gz’

fhvhv_tripdata_2021 100%[===================>] 167.66M  6.11MB/s    in 31s

2023-02-27 03:26:03 (5.39 MB/s) - ‘fhvhv_tripdata_2021-06.csv.gz’ saved [175799316/175799316]
```



```
[7]: !pwd
     !ls -al
```

```
/home/work
total 171772
drwxrwxrwx 3   1002  1003      4096 Feb 27 03:27 .
drwsrwsr-x 1 root   users      4096 Feb 27 03:21 ..
-rw-r--r-- 1 jovyan users 175799316 Dec 20 00:13 fhvhv_tripdata_2021-06.csv.gz
-rw-r--r-- 1 jovyan users     71509 Feb 27 03:27 hmwk5.ipynb
drwxr-xr-x 2 jovyan users      4096 Feb 27 03:23 .ipynb_checkpoints
```

```
[8]: !zcat fhvhv_tripdata_2021-06.csv.gz | wc -l
```

```
14961893
```

[5]:

```
df = spark.read \
        .option("header", "true") \
        .csv('/home/work/fhvhv_tripdata_2021-06.csv')

df.head(5)
```

[5]:

```
[Row(dispatching_base_num='B02764', pickup_datetime='2021-06-01 00:02:41', dropoff_datetime='2021-06-01 00:07:46',
base_number='B02764'),
 Row(dispatching_base_num='B02764', pickup_datetime='2021-06-01 00:16:16', dropoff_datetime='2021-06-01 00:21:14',
base_number='B02764'),
 Row(dispatching_base_num='B02764', pickup_datetime='2021-06-01 00:27:01', dropoff_datetime='2021-06-01 00:42:11',
_base_number='B02764'),
 Row(dispatching_base_num='B02764', pickup_datetime='2021-06-01 00:46:08', dropoff_datetime='2021-06-01 00:53:45',
_base_number='B02764'),
 Row(dispatching_base_num='B02510', pickup_datetime='2021-06-01 00:45:42', dropoff_datetime='2021-06-01 01:03:33',
_base_number=None)]
```

[11]:

```
!gunzip fhvhv_tripdata_2021-06.csv.gz
!head -n 1001 fhvhv_tripdata_2021-06.csv > head.csv
```

[12]:

```
import pandas as pd
df_pd = pd.read_csv('head.csv')
df_pd.dtypes
```

[12]:

```
dispatching_base_num      object
pickup_datetime           object
dropoff_datetime          object
PULocationID               int64
DOLocationID               int64
SR_Flag                   object
Affiliated_base_number    object
dtype: object
```

[6]:

```
from pyspark.sql import types

schema = types.StructType([
    types.StructField('dispatching_base_num', types.StringType(), True),
    types.StructField('pickup_datetime', types.TimestampType(), True),
    types.StructField('dropoff_datetime', types.TimestampType(), True),
    types.StructField('PULocationID', types.IntegerType(), True),
    types.StructField('DOLocationID', types.IntegerType(), True),
    types.StructField('SR_Flag', types.StringType(), True),
    types.StructField('Affiliated_base_number', types.StringType(), True)
])
```

[7]:

```
df = spark.read \
        .option("header", "true") \
        .schema(schema) \
        .csv('fhvhv_tripdata_2021-06.csv')

df.head(5)
```

[7]:

```
[Row(dispatching_base_num='B02764', pickup_datetime=datetime.datetime(2021, 6, 1, 0, 2, 41), d
nID=18, SR_Flag='N', Affiliated_base_number='B02764'),
 Row(dispatching_base_num='B02764', pickup_datetime=datetime.datetime(2021, 6, 1, 0, 16, 16),
onID=254, SR_Flag='N', Affiliated_base_number='B02764'),
 Row(dispatching_base_num='B02764', pickup_datetime=datetime.datetime(2021, 6, 1, 0, 27, 1), d
onID=127, SR_Flag='N', Affiliated_base_number='B02764'),
 Row(dispatching_base_num='B02764', pickup_datetime=datetime.datetime(2021, 6, 1, 0, 46, 8), d
onID=235, SR_Flag='N', Affiliated_base_number='B02764'),
 Row(dispatching_base_num='B02510', pickup_datetime=datetime.datetime(2021, 6, 1, 0, 45, 42),
onID=146, SR_Flag='N', Affiliated_base_number=None)]
```

[8]:

```
!mkdir -p /home/partitions/fhvhv/2021/06
```

[9]:

```
print(df.count())
rpdf = df.repartition(12)
rpdf.write.parquet('/home/partitions/fhvhv/2021/06/', mode='overwrite')
```

```
14961892
```

[10]:

```
!du /home/partitions/fhvhv/2021/06/
```

```
279328  /home/partitions/fhvhv/2021/06/
```

[11]:

```
dfp = spark.read.parquet('/home/partitions/fhvhv/2021/06/')
```

[12]:

```
from pyspark.sql import functions as F

dfp.select('dispatching_base_num', 'pickup_datetime', 'dropoff_datetime', 'PULocationID', 'DOL
    .filter((F.to_date(dfp.pickup_datetime) == '2021-06-15') &  (dfp.dispatching_base_num.isNo
    .count()
```

[12]:

```
452470
```

```
[13]: dfp = dfp.withColumn('trip_duration_hours', F.round((F.unix_timestamp('dropoff_datetime') - F.unix_timestamp('pickup_datetime'))/3600, 2))
      dfp.show()
```

```
+--------------------+-------------------+-------------------+------------+------------+-------+----------------------+-------------------+
|dispatching_base_num|    pickup_datetime|   dropoff_datetime|PULocationID|DOLocationID|SR_Flag|Affiliated_base_number|trip_duration_hours|
+--------------------+-------------------+-------------------+------------+------------+-------+----------------------+-------------------+
|              B02617|2021-06-04 16:50:34|2021-06-04 17:01:18|         118|         109|      N|                B02617|               0.18|
|              B02875|2021-06-02 22:28:45|2021-06-02 22:37:28|         163|          79|      N|                B02875|               0.15|
|              B02871|2021-06-03 11:47:48|2021-06-03 11:52:23|         231|          13|      N|                B02871|               0.08|
|              B02888|2021-06-03 08:45:25|2021-06-03 09:00:12|           9|          92|      N|                B02888|               0.25|
|              B02510|2021-06-05 09:50:43|2021-06-05 10:06:53|          14|         133|      N|                  null|               0.27|
|              B02764|2021-06-03 22:55:50|2021-06-03 23:21:24|         152|          74|      N|                B02764|               0.43|
|              B02889|2021-06-02 03:15:48|2021-06-02 03:26:20|         220|         235|      N|                B02889|               0.18|
|              B02872|2021-06-01 11:50:42|2021-06-01 12:00:36|         162|         161|      N|                B02872|               0.17|
|              B02510|2021-06-04 06:51:45|2021-06-04 06:56:26|         206|         206|      N|                  null|               0.08|
|              B02835|2021-06-02 01:21:31|2021-06-02 01:50:23|          49|         182|      N|                B02835|               0.48|
|              B02510|2021-06-02 17:11:31|2021-06-02 18:18:32|         236|          89|      N|                  null|               1.12|
|              B02869|2021-06-01 18:48:20|2021-06-01 18:54:59|         151|          43|      N|                B02869|               0.11|
|              B02510|2021-06-02 16:16:53|2021-06-02 16:39:03|         181|         189|      N|                  null|               0.37|
|              B02510|2021-06-03 21:23:35|2021-06-03 21:33:33|         181|          89|      N|                  null|               0.17|
|              B02764|2021-06-01 06:51:24|2021-06-01 06:58:09|          86|          86|      N|                B02764|               0.11|
|              B02867|2021-06-04 18:26:09|2021-06-04 18:41:43|         162|         263|      N|                B02867|               0.26|
|              B02510|2021-06-01 10:21:49|2021-06-01 11:08:24|         151|          17|      N|                  null|               0.78|
|              B02682|2021-06-04 01:45:17|2021-06-04 01:53:16|         125|         164|      N|                B02682|               0.13|
|              B02875|2021-06-01 14:06:42|2021-06-01 14:33:53|         192|           7|      N|                B02875|               0.45|
|              B02510|2021-06-03 21:15:15|2021-06-03 21:21:24|         171|          16|      N|                  null|                0.1|
+--------------------+-------------------+-------------------+------------+------------+-------+----------------------+-------------------+
only showing top 20 rows
```

```
[14]: max_trip_duration = dfp.agg(F.max('trip_duration_hours')).collect()[0][0]
      print(max_trip_duration)
```

```
[Stage 14:=============================>                            (2 + 2) / 4]
66.88
```

[15]:

```
!wget https://github.com/DataTalksClub/nyc-tlc-data/releases/download/misc/taxi_zone_lookup.csv
```

```
--2023-02-27 03:55:19--  https://github.com/DataTalksClub/nyc-tlc-data/releases/download/misc/taxi_zone_lookup.csv
Resolving github.com (github.com)... 20.205.243.166
Connecting to github.com (github.com)|20.205.243.166|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://objects.githubusercontent.com/github-production-release-asset-2e65be/513814948/5a2cc2f5-b4cd-4584-
ntial=AKIAIWNJYAX4CSVEH53A%2F20230227%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20230227T035519Z&X-Amz-Expires=300&
f90bc05f16236ab3a8&X-Amz-SignedHeaders=host&actor_id=0&key_id=0&repo_id=513814948&response-content-disposition=attac
pe=application%2Foctet-stream [following]
--2023-02-27 03:55:19--  https://objects.githubusercontent.com/github-production-release-asset-2e65be/513814948/5a2c
256&X-Amz-Credential=AKIAIWNJYAX4CSVEH53A%2F20230227%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20230227T035519Z&X-A
0d2c9c2290882fff90bc05f16236ab3a8&X-Amz-SignedHeaders=host&actor_id=0&key_id=0&repo_id=513814948&response-content-di
onse-content-type=application%2Foctet-stream
Resolving objects.githubusercontent.com (objects.githubusercontent.com)... 185.199.109.133, 185.199.108.133, 185.199
Connecting to objects.githubusercontent.com (objects.githubusercontent.com)|185.199.109.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 12322 (12K) [application/octet-stream]
Saving to: ‘taxi_zone_lookup.csv’

taxi_zone_lookup.cs 100%[===================>]  12.03K  --.-KB/s    in 0s

2023-02-27 03:55:20 (25.2 MB/s) - ‘taxi_zone_lookup.csv’ saved [12322/12322]
```

[20]:

```
zschema = types.StructType([
    types.StructField('LocationID', types.IntegerType(), True),
    types.StructField('Borough', types.StringType(), True),
    types.StructField('Zone', types.StringType(), True),
    types.StructField('service_zone', types.StringType(), True)
])


zdf = spark.read \
        .option("header", "true") \
        .schema(zschema) \
        .csv('/home/work/taxi_zone_lookup.csv')

zdf.head(5)
```

[20]:

```
[Row(LocationID=1, Borough='EWR', Zone='Newark Airport', service_zone='EWR'),
 Row(LocationID=2, Borough='Queens', Zone='Jamaica Bay', service_zone='Boro Zone'),
 Row(LocationID=3, Borough='Bronx', Zone='Allerton/Pelham Gardens', service_zone='Boro Zone'),
 Row(LocationID=4, Borough='Manhattan', Zone='Alphabet City', service_zone='Yellow Zone'),
 Row(LocationID=5, Borough='Staten Island', Zone='Arden Heights', service_zone='Boro Zone')]
```

[23]:

```python
from pyspark.sql.functions import col

joined_df = df.join(zdf, col("PULocationID") == col("LocationID"), "left")
joined_df.head(5)
```

[23]:

```
[Row(dispatching_base_num='B02764', pickup_datetime=datetime.datetime(2021, 6, 1, 0, 2, 41), dropoff_
nID=18, SR_Flag='N', Affiliated_base_number='B02764', LocationID=174, Borough='Bronx', Zone='Norwood'
 Row(dispatching_base_num='B02764', pickup_datetime=datetime.datetime(2021, 6, 1, 0, 16, 16), dropoff_
onID=254, SR_Flag='N', Affiliated_base_number='B02764', LocationID=32, Borough='Bronx', Zone='Bronxda
 Row(dispatching_base_num='B02764', pickup_datetime=datetime.datetime(2021, 6, 1, 0, 27, 1), dropoff_
onID=127, SR_Flag='N', Affiliated_base_number='B02764', LocationID=240, Borough='Bronx', Zone='Van Co
 Row(dispatching_base_num='B02764', pickup_datetime=datetime.datetime(2021, 6, 1, 0, 46, 8), dropoff_
onID=235, SR_Flag='N', Affiliated_base_number='B02764', LocationID=127, Borough='Manhattan', Zone='In
 Row(dispatching_base_num='B02510', pickup_datetime=datetime.datetime(2021, 6, 1, 0, 45, 42), dropoff_
onID=146, SR_Flag='N', Affiliated_base_number=None, LocationID=144, Borough='Manhattan', Zone='Little
```

[32]:

```python
from pyspark.sql.functions import count, max

zone_trips = joined_df.groupBy("Zone").agg(count("PULocationID").alias("NumTrips"))
max_count = zone_trips.agg(max(col("NumTrips")).alias("max_trips")).first()["max_trips"]
zone_trips.filter(col("NumTrips") == max_count).select("Zone", "NumTrips").show()
```

```
+-------------------+--------+
|               Zone|NumTrips|
+-------------------+--------+
|Crown Heights North|  231279|
+-------------------+--------+
```