DISTRIBUTED SYSTEMS ASSIGNMENT REPORT

ASSIGNMENT REPORT

Assignment ID: Assignment3 - Distributed Batch Processing Using Apache Spark

Student Name: 王谦益

Student ID: 12111003

DESIGN

Initialize

1. use `pip install pyspark` to build the environment
2. from `pyspark.sql` import `SparkSession` and other functions needed in package `functions`
3. initialize the spark session

   ```
   spark = SparkSession.builder.appName("ParkingDataAnalysis").getOrCreate()
   data = spark.read.csv('data/parking_data_sz.csv', header=True,
   inferSchema=True)
   ```

4. filter out invalid data in advance

   ```
   data = data.filter(col("out_time") > col("in_time"))
   ```

task1

SELECT COUNT(berthage) GROUP BY section

```
result1 = data.groupBy("section").agg(
    countDistinct("berthage").alias("count")
)
```

task2

SELECT DISTINCT(berthage, section)

```
result2 = data.select("berthage", "section").distinct()
```

task3

SELECT AVG(out_timg - in_time) GROUP BY section

```
result3 = data.withColumn("parking_time", (col("out_time") - col("in_time")))
result3 = result3.groupBy("section").agg(
    avg("parking_time").cast("int").alias("avg_parking_time")
)
```

task4

SELECT AVG(out_timg - in_time) GROUP BY berthage

```
result4 = data.withColumn("parking_time", (col("out_time") - col("in_time")))
result4 = result4.groupBy("berthage").agg(
    avg("parking_time").cast("int").alias("avg_parking_time")
)
```

task5

1. initialize the time limitation

    1. find the minimum and maximum time

        ```
        time_limitation = data.groupBy("section").agg(
            min("in_time").alias("min_in"),
            max("out_time").alias("max_out")
        )
        ```

    2. list the time sequence and form start_time & end_time

        ```
        time_limitation = time_limitation \
            .withColumn("hour_range", expr("sequence(min_in, max_out, interval 1 hour)").cast("array<timestamp>")) \
            .withColumn("start_time", explode("hour_range")) \
            .withColumn("end_time", expr("start_time + INTERVAL 1 HOUR"))
        time_limitation = time_limitation \
        ```

```
   .select("section", "start_time", "end_time") \
   .orderBy("section", "start_time")
```

2. prepare all the data needed

    ```
    all_data_needed = time_limitation.alias("time_limitation").join(
       data.alias("data"),
       (data.section == time_limitation.section) &
       (data.in_time < time_limitation.end_time) &
       (data.out_time > time_limitation.start_time),
       "left"
    )
    all_data_needed = all_data_needed \
       .select("time_limitation.section", "start_time", "end_time", "berthage") \
       .orderBy("section", "start_time")
    ```

3. calculate count & percentage

    1. select the berthage in_use and total

        ```
        in_use = all_data_needed.groupBy("section", "start_time", "end_time").agg(
           countDistinct("berthage").cast("long").alias("count")
        )
        total_count = all_data_needed.groupBy("section").agg(
           countDistinct("berthage").cast("long").alias("total_count")
        )
        ```

    2. form the result

        ```
           result5 = in_use \
              .join(total_count, "section") \
              .withColumn("percentage", round(col("count") / col("total_count") * 100, 1).cast("string")) \
              .select("section", "start_time", "end_time", "count", "percentage") \
              .orderBy("section", "start_time") \
              .withColumn("start_time", date_format("start_time", "yyyy-MM-dd HH:mm:ss")) \
              .withColumn("end_time", date_format("end_time", "yyyy-MM-dd HH:mm:ss")) \
              .withColumn("percentage", concat(col("percentage"), lit("%")))
        ```

subtask

1. import `plotly.express as px` and `pandas as pd`

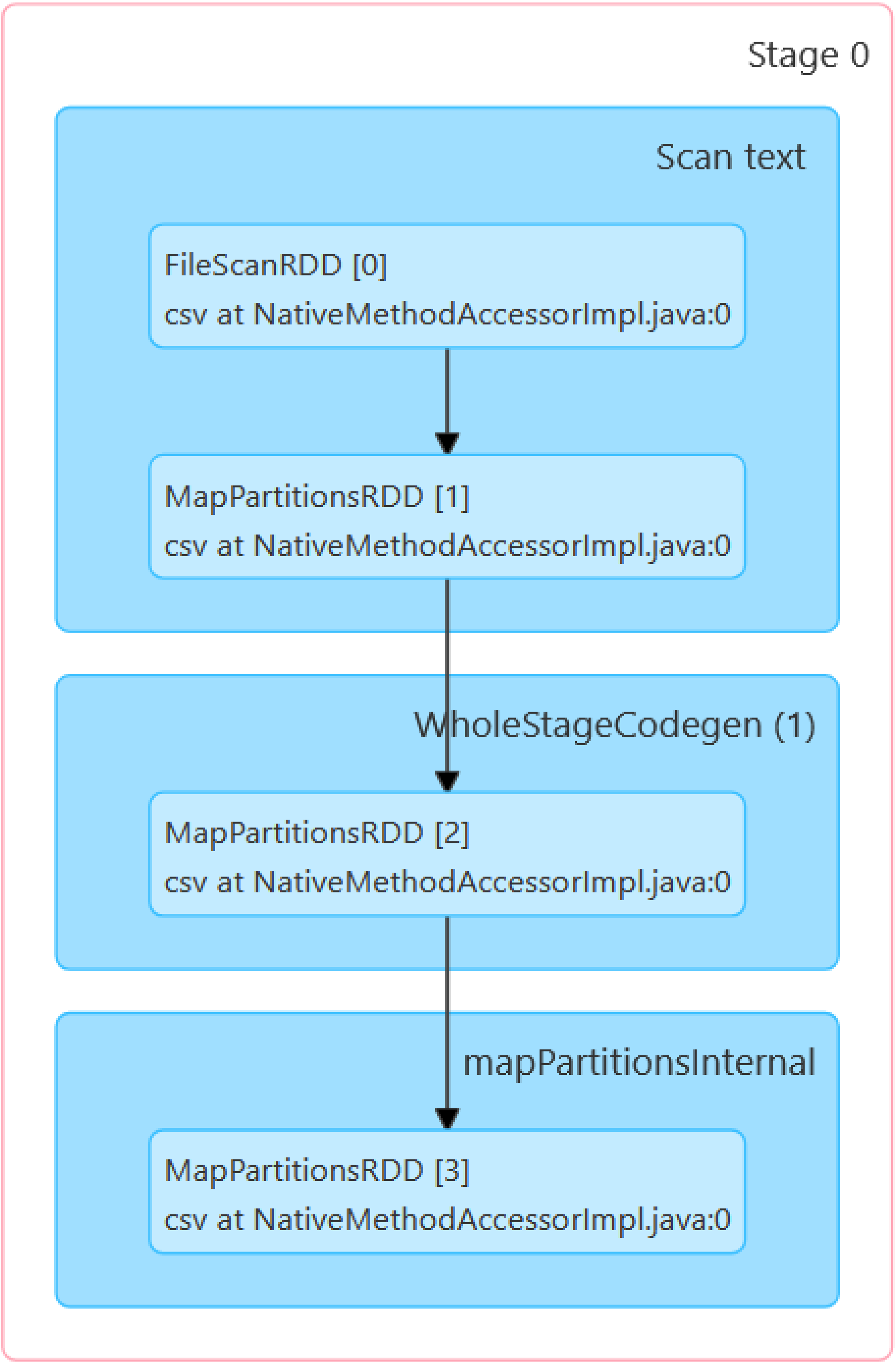
2. choose 3 sections in the data of task5

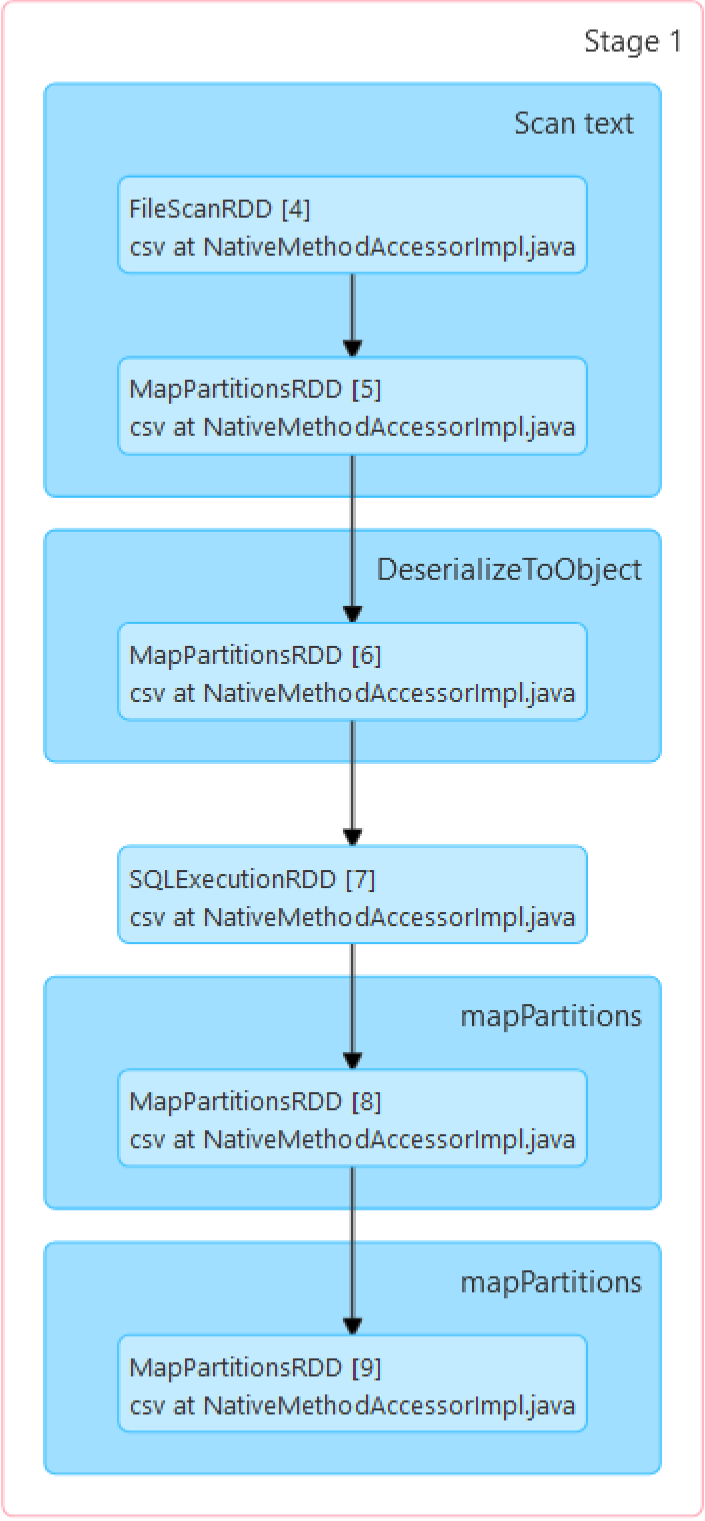
3. iterate the dataframe and plot the figure

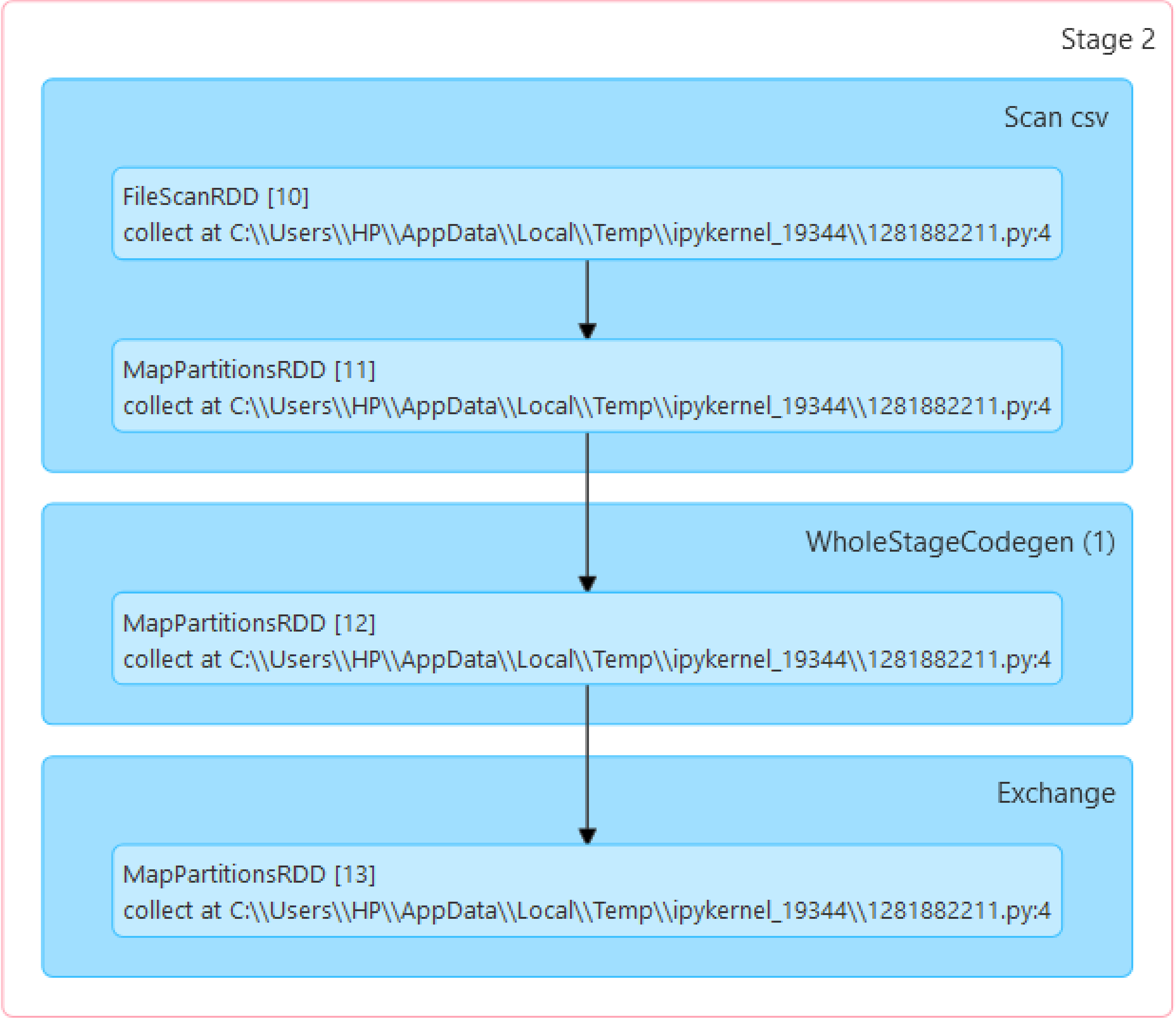
```
 for section in sections:
     section_name = section.section
     section_data = result5.filter(col("section") == section_name)
     x_array = [node[0] for node in
section_data.select("start_time").collect()]
     y_array = [float(node[0].split("%")[0]) for node in
section_data.select("percentage").collect()]
     data = pd.DataFrame({'Time': x_array, 'Percentage': y_array})
     fig = px.line(data, x='Time', y='Percentage', title="Percentage of
Berthages in Use Over Time")
     fig.show()
```

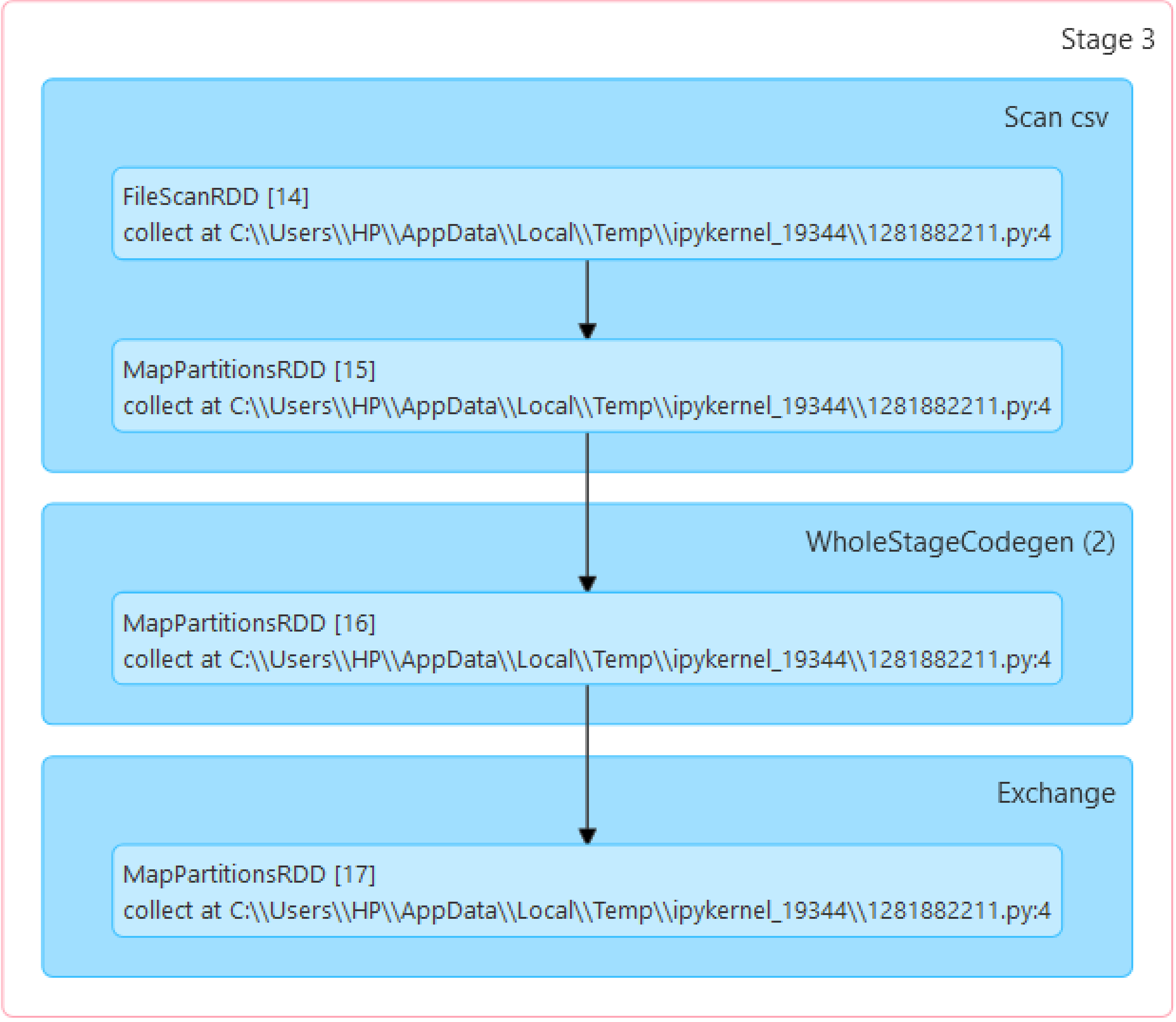
RUNNING RESULT

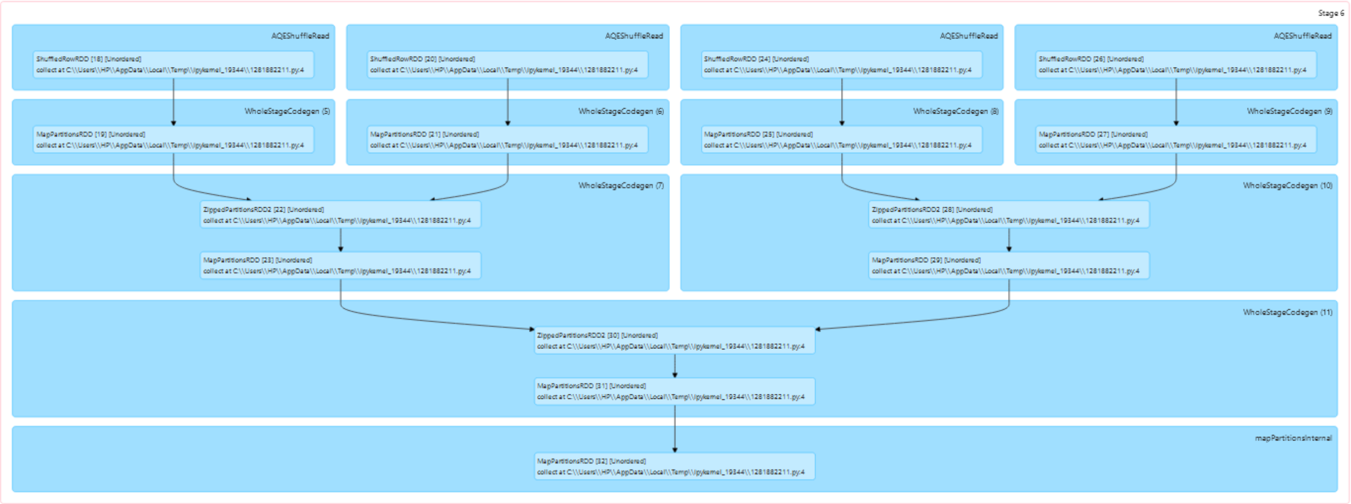
DAGs

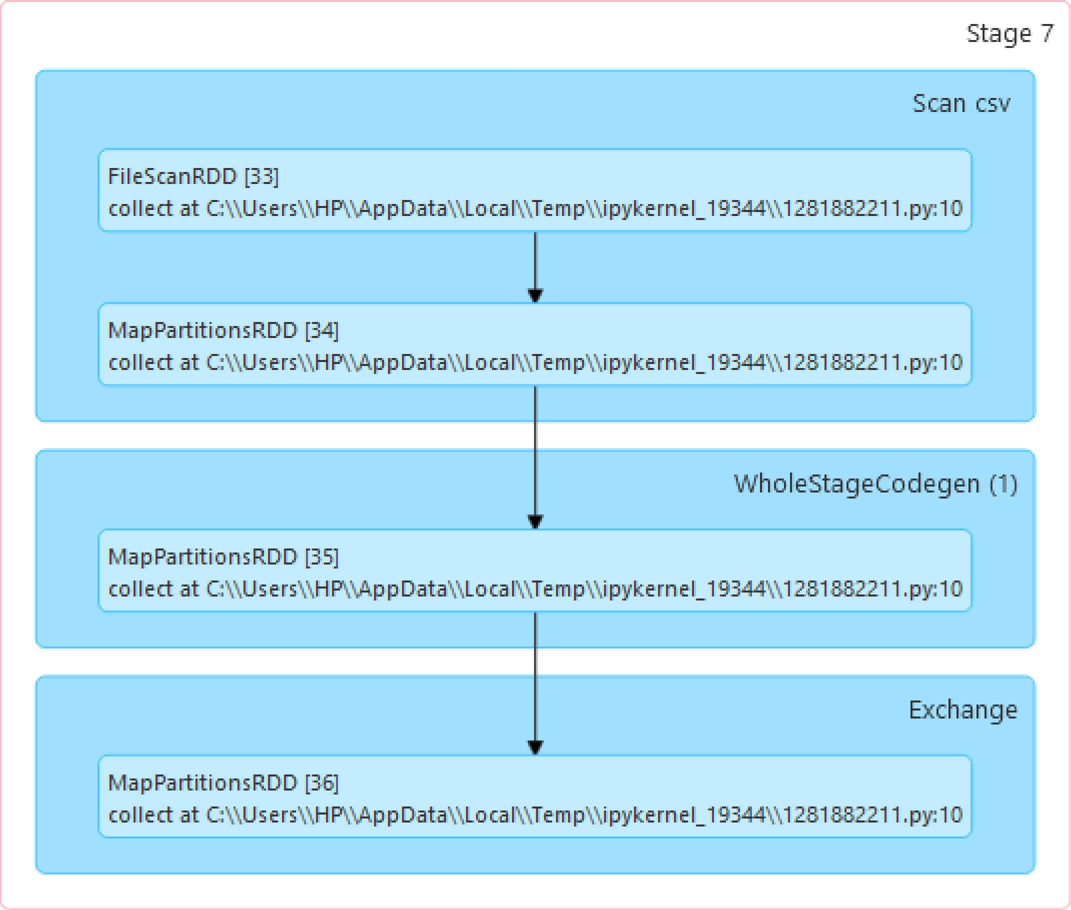

Stage 0
Scan text
FileScanRDD [0]
csv at NativeMethodAccessorImpl.java:0
MapPartitionsRDD [1]
csv at NativeMethodAccessorImpl.java:0
WholeStageCodegen (1)
MapPartitionsRDD [2]
csv at NativeMethodAccessorImpl.java:0
mapPartitionsInternal
MapPartitionsRDD [3]
csv at NativeMethodAccessorImpl.java:0


Stage 1
Scan text
FileScanRDD [4]
csv at NativeMethodAccessorImpl.java
MapPartitionsRDD [5]
csv at NativeMethodAccessorImpl.java
DeserializeToObject
MapPartitionsRDD [6]
csv at NativeMethodAccessorImpl.java
SQLExecutionRDD [7]
csv at NativeMethodAccessorImpl.java
mapPartitions
MapPartitionsRDD [8]
csv at NativeMethodAccessorImpl.java
mapPartitions
MapPartitionsRDD [9]
csv at NativeMethodAccessorImpl.java


Stage 2
Scan csv
FileScanRDD [10]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
MapPartitionsRDD [11]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
WholeStageCodegen (1)
MapPartitionsRDD [12]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
Exchange
MapPartitionsRDD [13]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4


Stage 3
Scan csv
FileScanRDD [14]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
MapPartitionsRDD [15]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
WholeStageCodegen (2)
MapPartitionsRDD [16]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
Exchange
MapPartitionsRDD [17]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4


Stage 6
AQEShuffleRead
ShuffledRowRDD [18] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
AQEShuffleRead
ShuffledRowRDD [20] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
AQEShuffleRead
ShuffledRowRDD [24] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
AQEShuffleRead
ShuffledRowRDD [26] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
WholeStageCodegen (5)
MapPartitionsRDD [19] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
WholeStageCodegen (6)
MapPartitionsRDD [21] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
WholeStageCodegen (8)
MapPartitionsRDD [25] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
WholeStageCodegen (9)
MapPartitionsRDD [27] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
WholeStageCodegen (7)
ZippedPartitionsRDD2 [22] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
MapPartitionsRDD [23] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
WholeStageCodegen (10)
ZippedPartitionsRDD2 [28] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
MapPartitionsRDD [29] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
WholeStageCodegen (11)
ZippedPartitionsRDD2 [30] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
MapPartitionsRDD [31] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4
mapPartitionsInternal
MapPartitionsRDD [32] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:4



Stage 7

Scan csv

FileScanRDD [33]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10

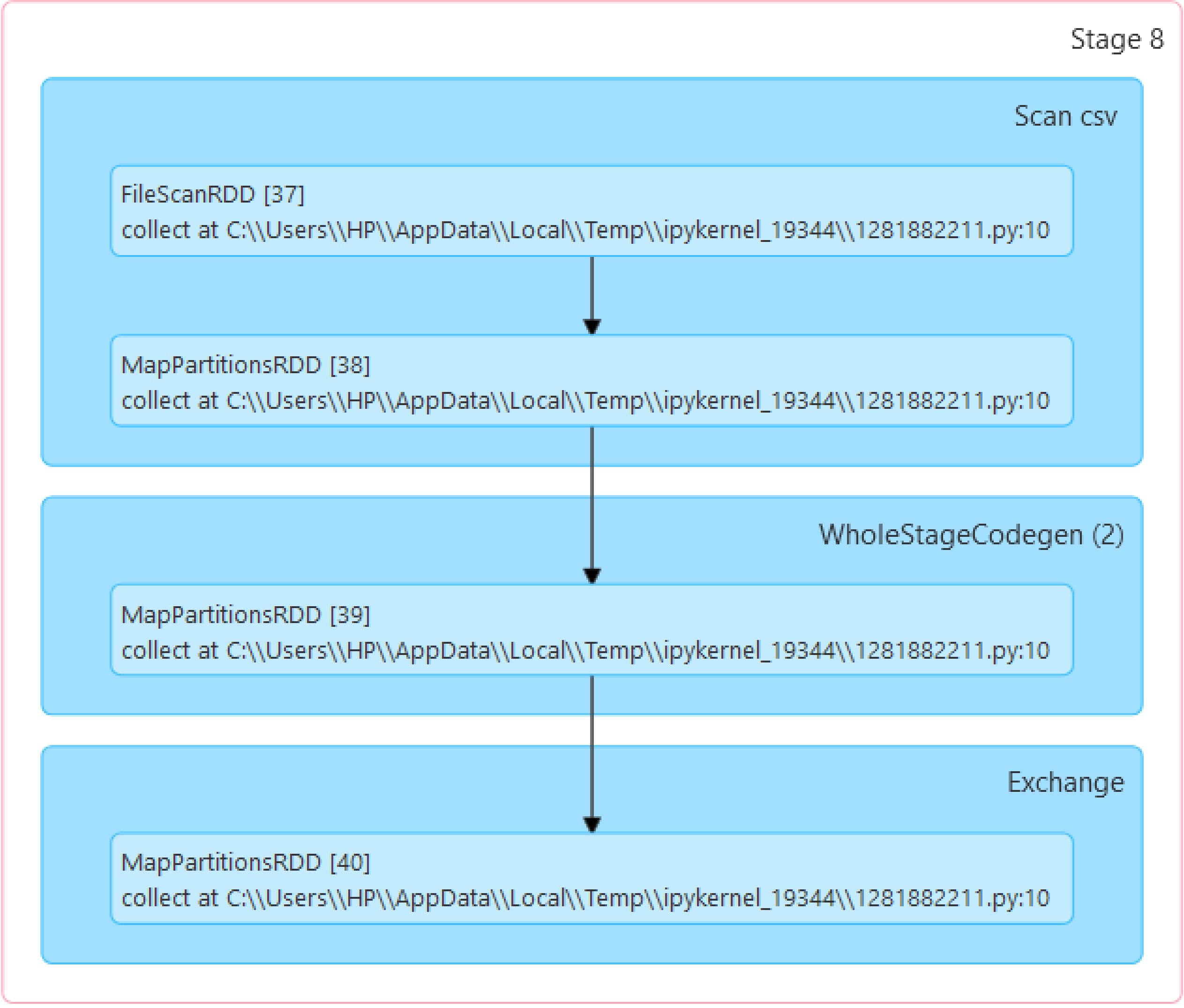
MapPartitionsRDD [34]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10

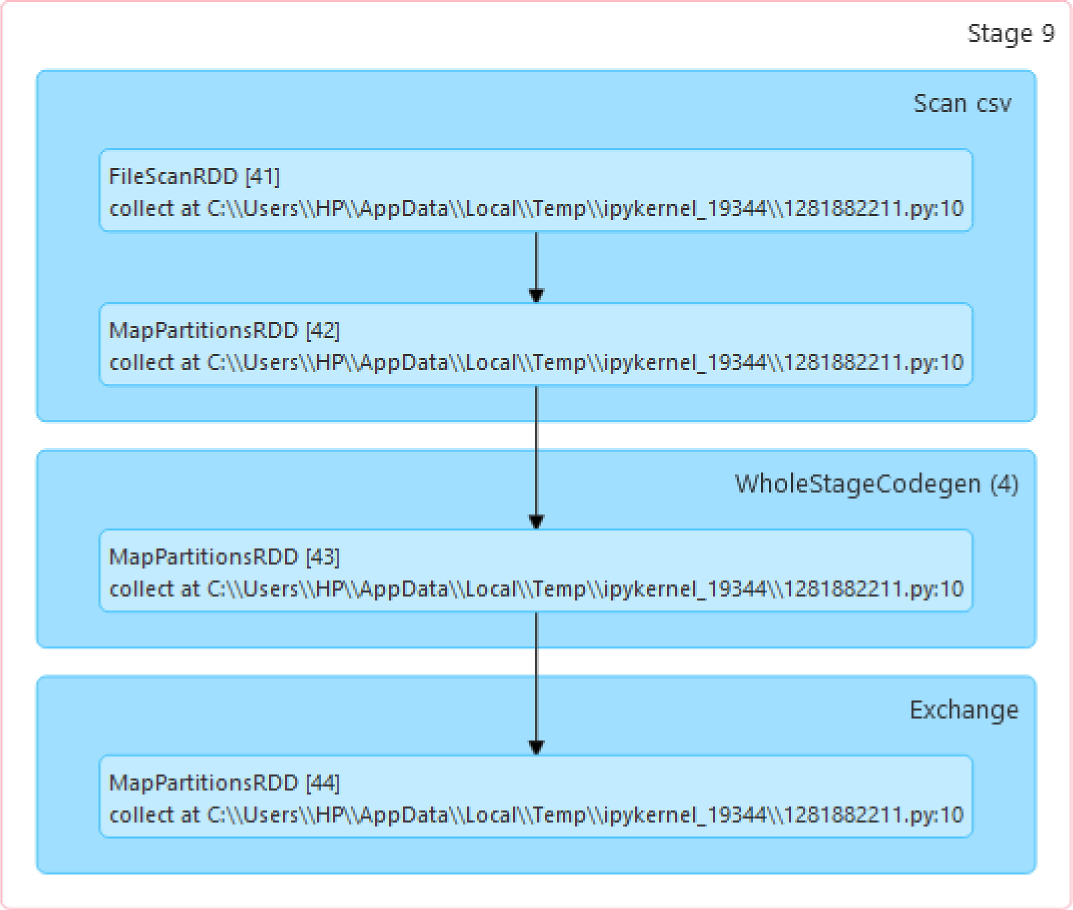
WholeStageCodegen (1)

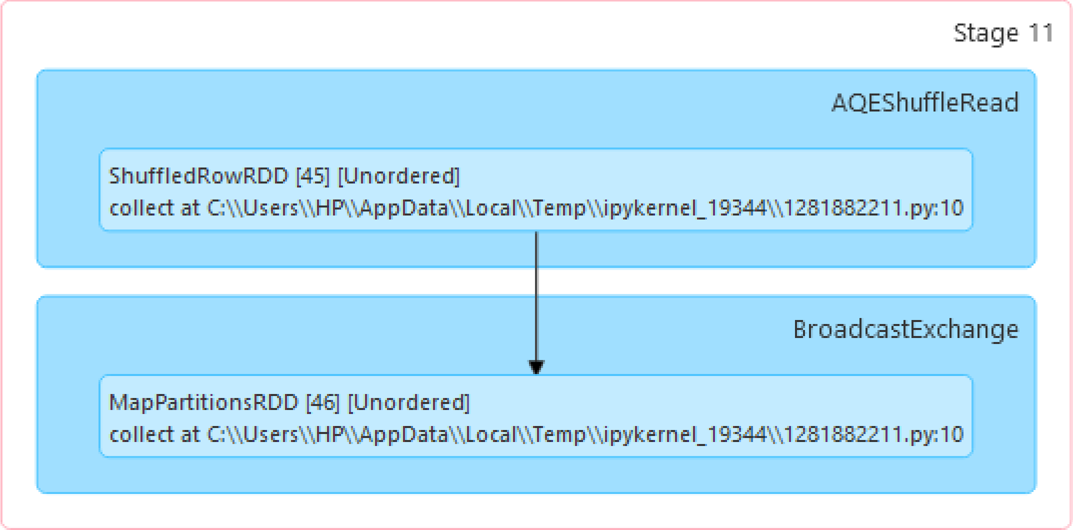
MapPartitionsRDD [35]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10

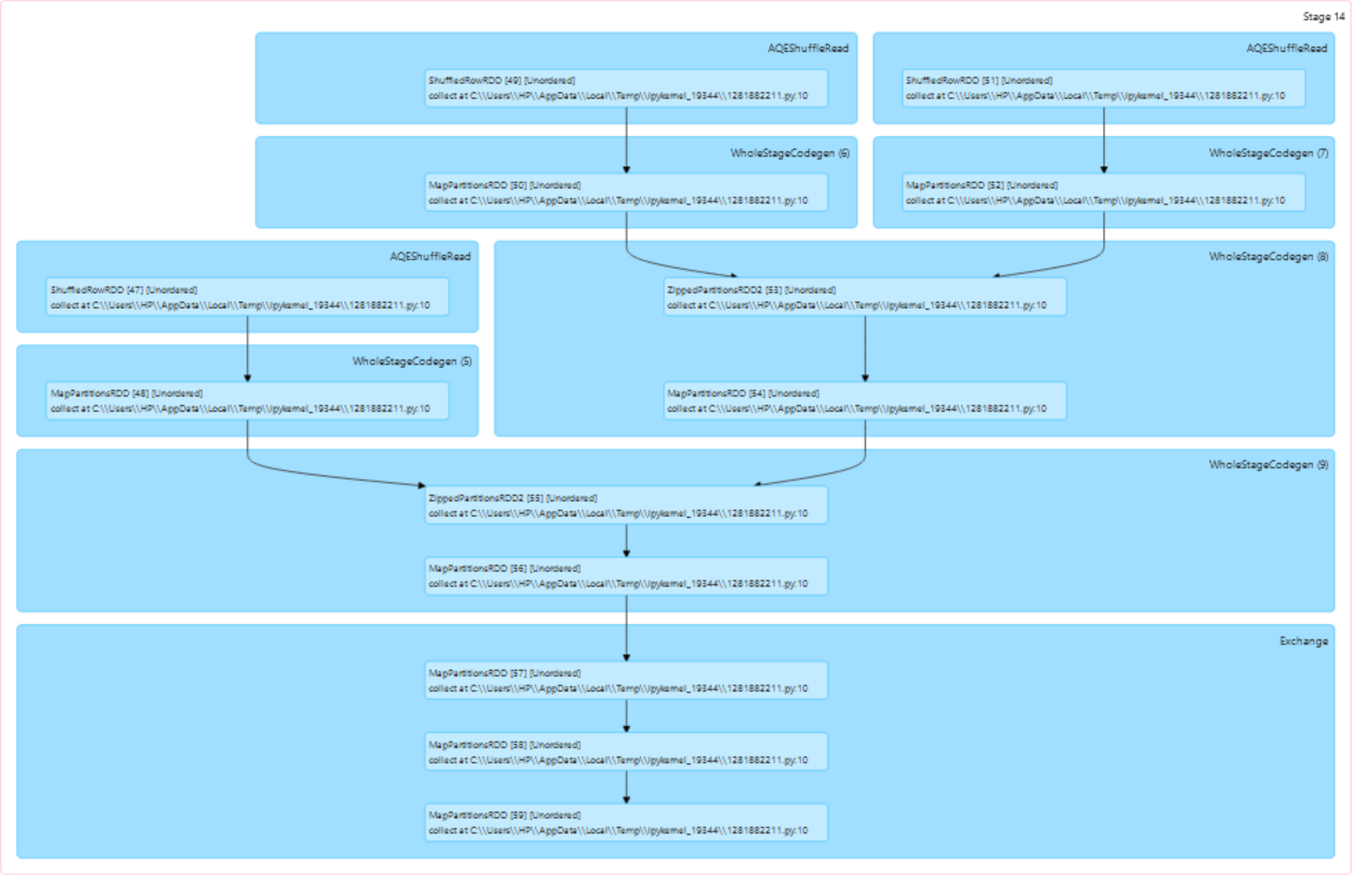
Exchange

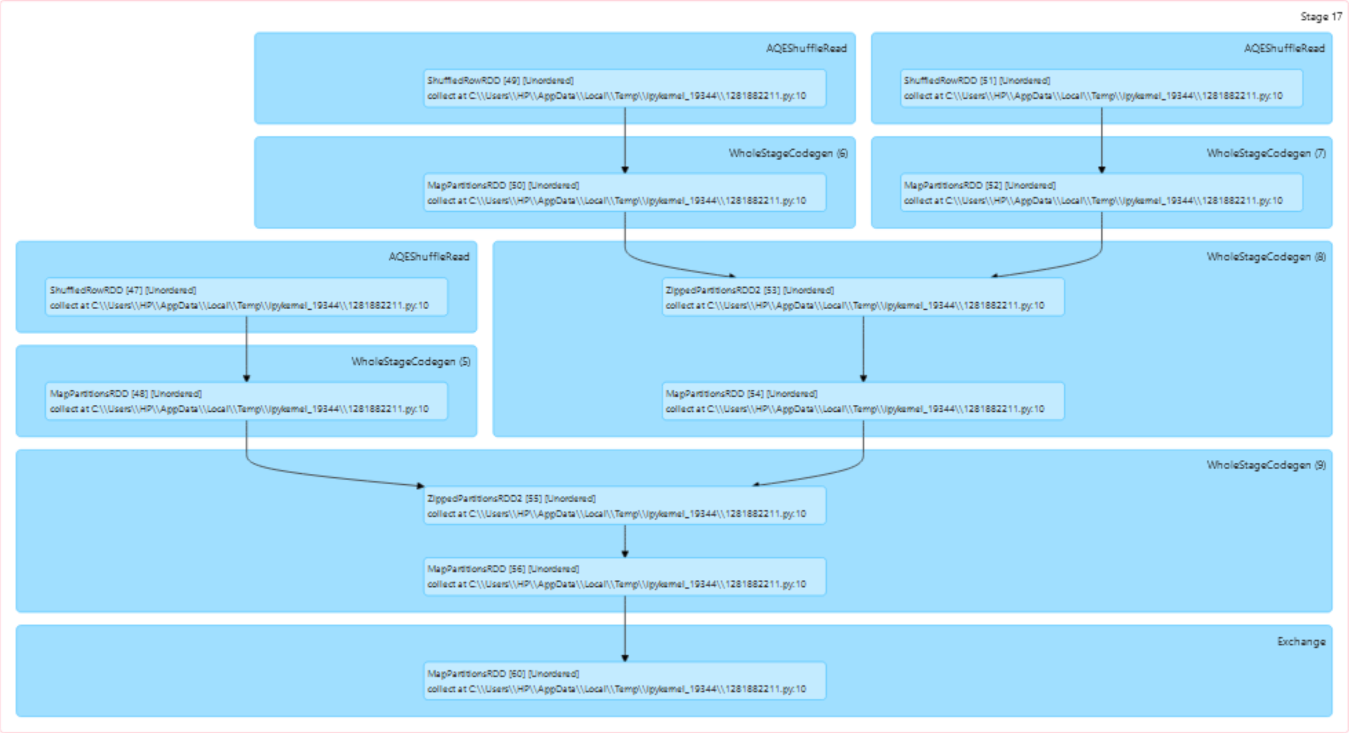
MapPartitionsRDD [36]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10

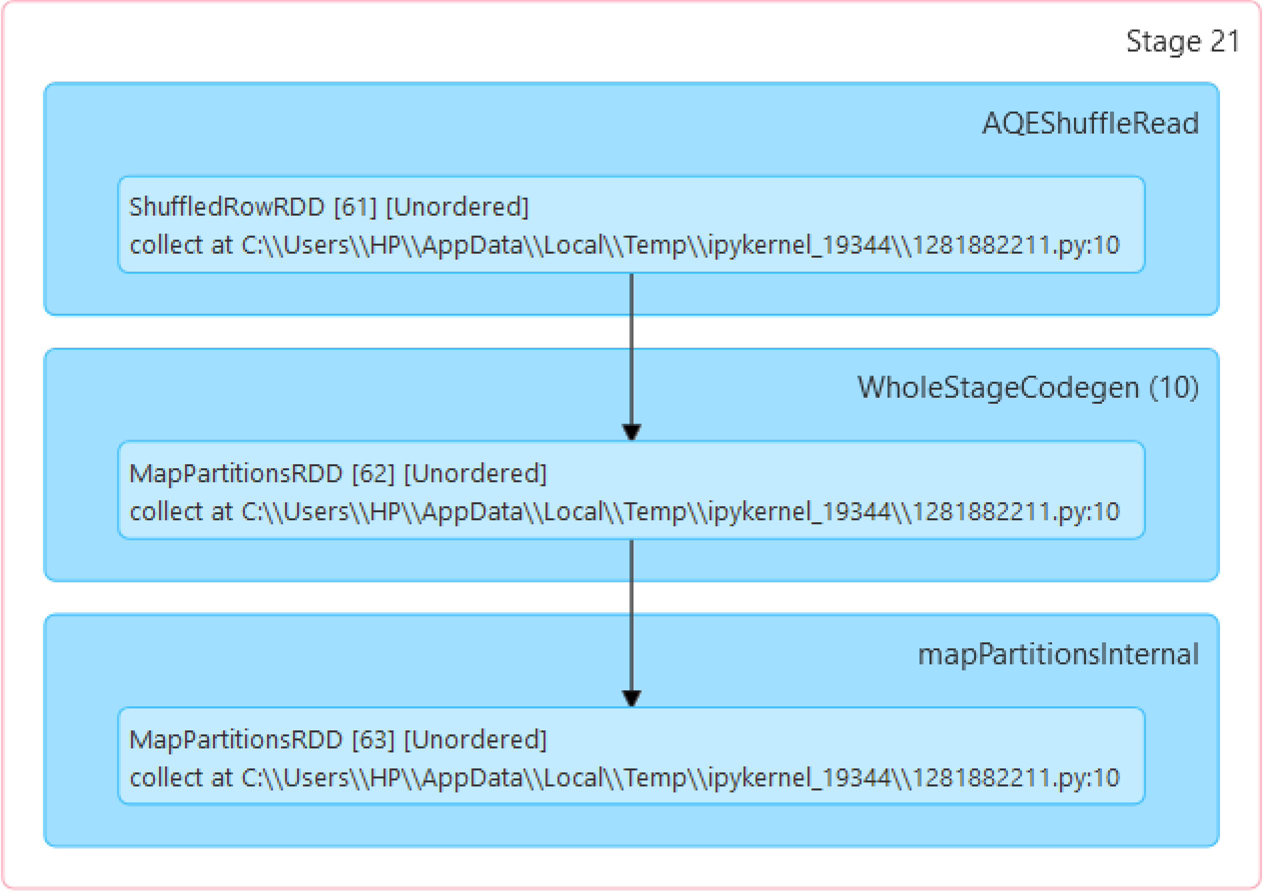

Stage 8
Scan csv
FileScanRDD [37]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
MapPartitionsRDD [38]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
WholeStageCodegen (2)
MapPartitionsRDD [39]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
Exchange
MapPartitionsRDD [40]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10

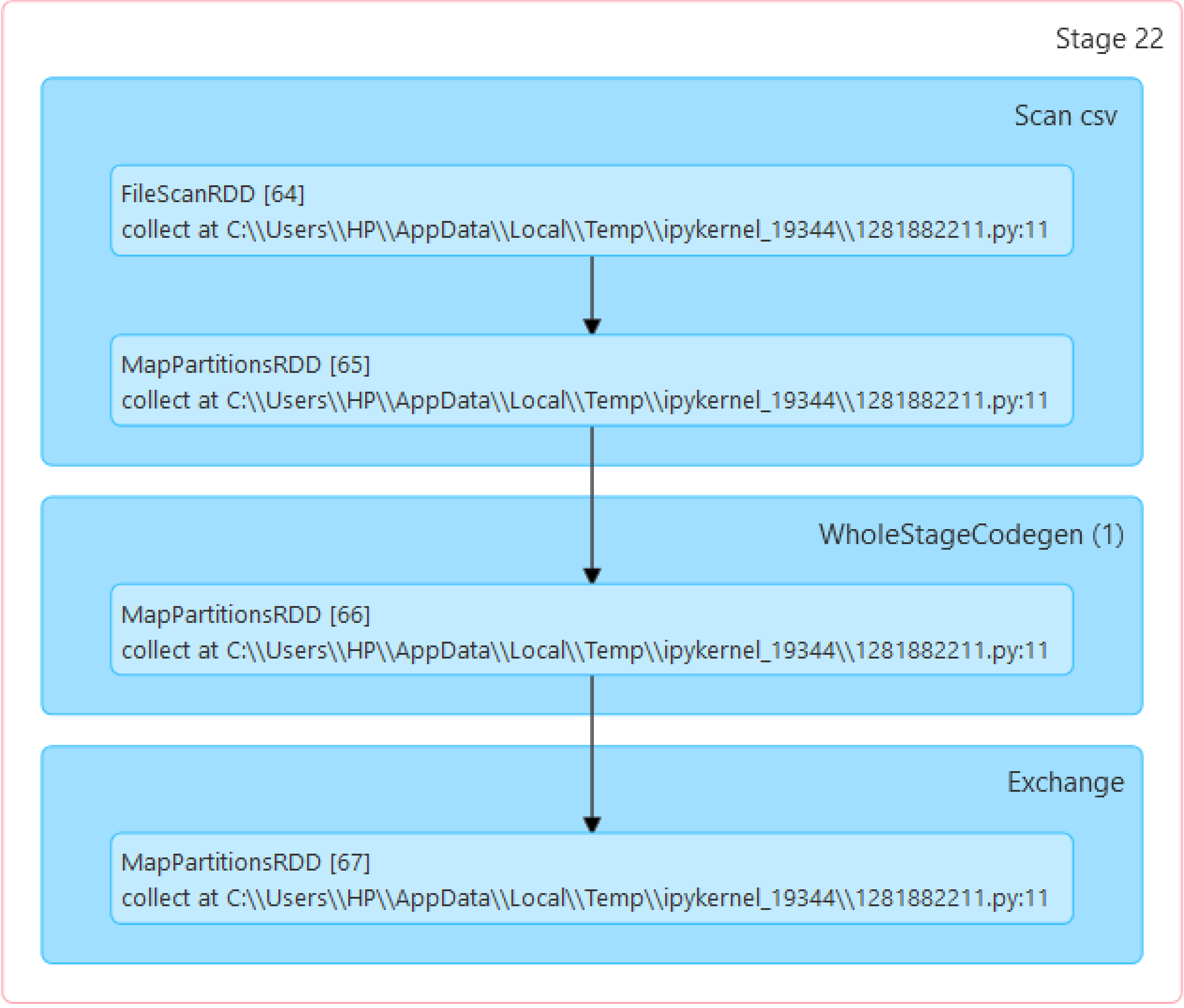

Stage 9
Scan csv
FileScanRDD [41]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
MapPartitionsRDD [42]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
WholeStageCodegen (4)
MapPartitionsRDD [43]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
Exchange
MapPartitionsRDD [44]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10

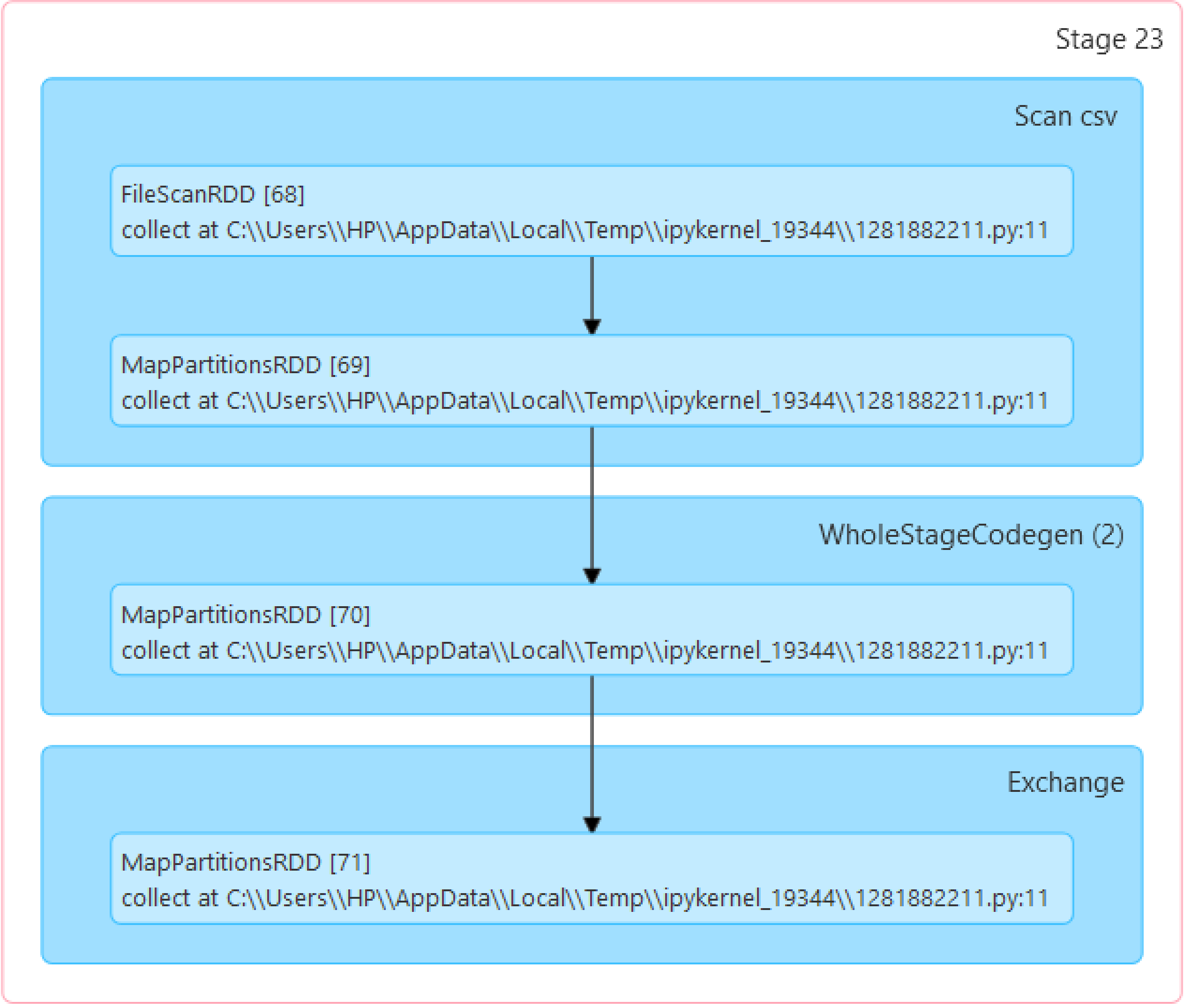

Stage 11
AQEShuffleRead
ShuffledRowRDD [45] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
BroadcastExchange
MapPartitionsRDD [46] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10


Stage 14
AQEShuffleRead
ShuffledRowRDD [49] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
AQEShuffleRead
ShuffledRowRDD [51] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
WholeStageCodegen (6)
MapPartitionsRDD [50] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
WholeStageCodegen (7)
MapPartitionsRDD [52] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
AQEShuffleRead
ShuffledRowRDD [47] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
WholeStageCodegen (8)
ZippedPartitionsRDD2 [53] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
WholeStageCodegen (5)
MapPartitionsRDD [48] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
MapPartitionsRDD [54] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
WholeStageCodegen (9)
ZippedPartitionsRDD2 [55] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
MapPartitionsRDD [56] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
Exchange
MapPartitionsRDD [57] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
MapPartitionsRDD [58] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
MapPartitionsRDD [59] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10

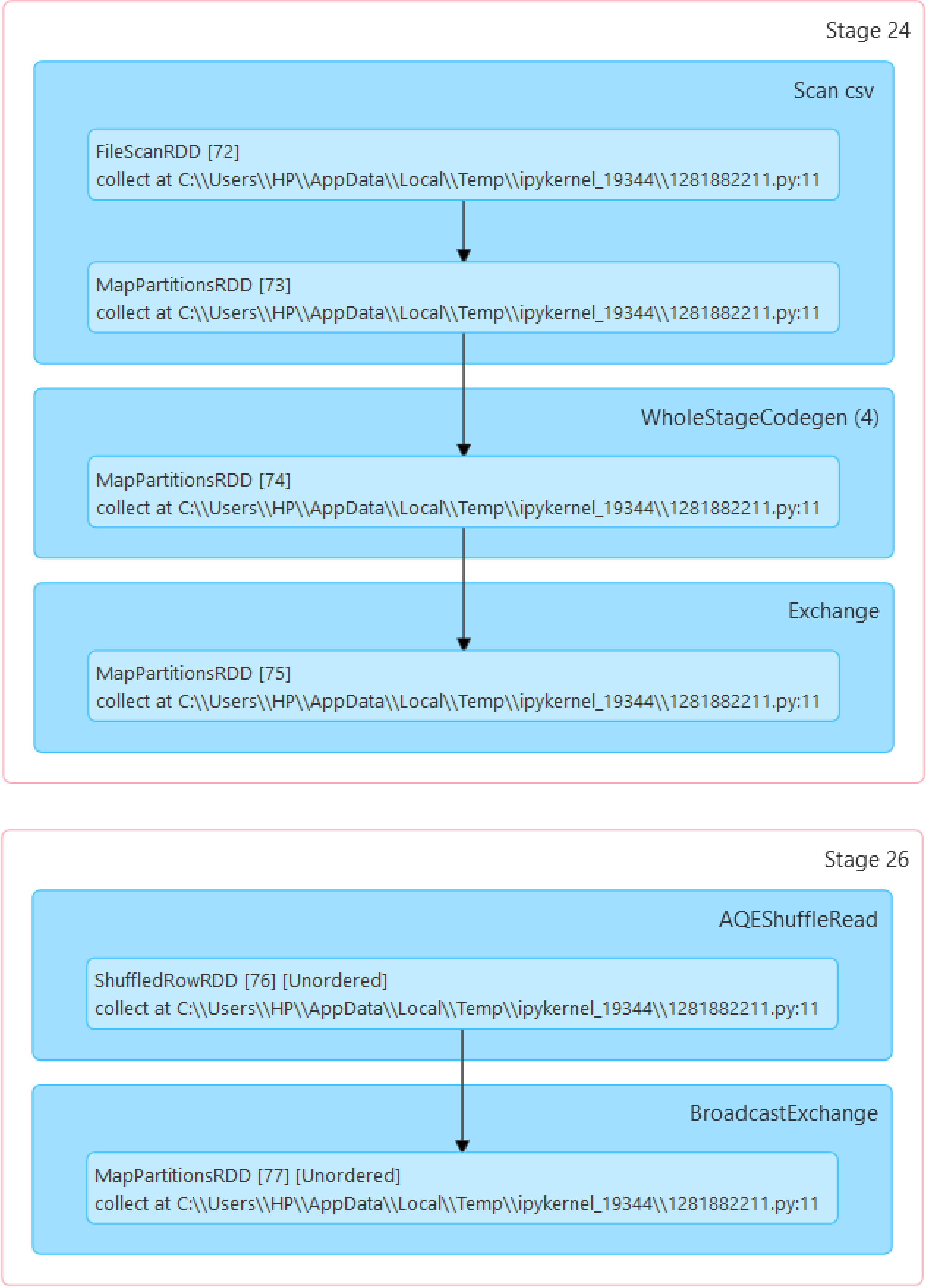

Stage 17
AQEShuffleRead
ShuffledRowRDD [49] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
AQEShuffleRead
ShuffledRowRDD [51] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
WholeStageCodegen (6)
MapPartitionsRDD [50] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
WholeStageCodegen (7)
MapPartitionsRDD [52] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
AQEShuffleRead
ShuffledRowRDD [47] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
WholeStageCodegen (8)
ZippedPartitionsRDD2 [53] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
WholeStageCodegen (5)
MapPartitionsRDD [48] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
MapPartitionsRDD [54] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
WholeStageCodegen (9)
ZippedPartitionsRDD2 [55] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
MapPartitionsRDD [56] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
Exchange
MapPartitionsRDD [60] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10

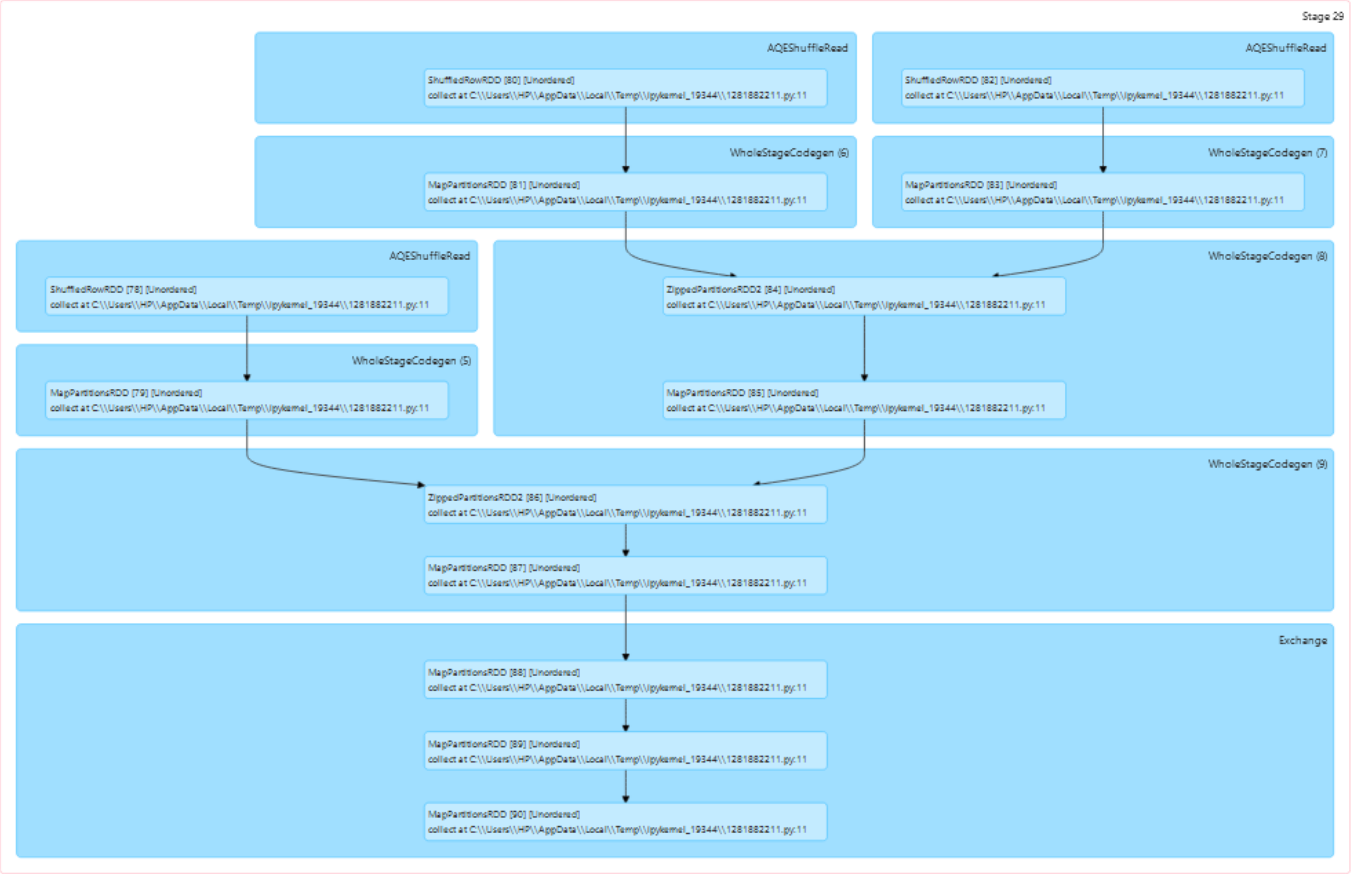

Stage 21
AQEShuffleRead
ShuffledRowRDD [61] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
WholeStageCodegen (10)
MapPartitionsRDD [62] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10
mapPartitionsInternal
MapPartitionsRDD [63] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:10

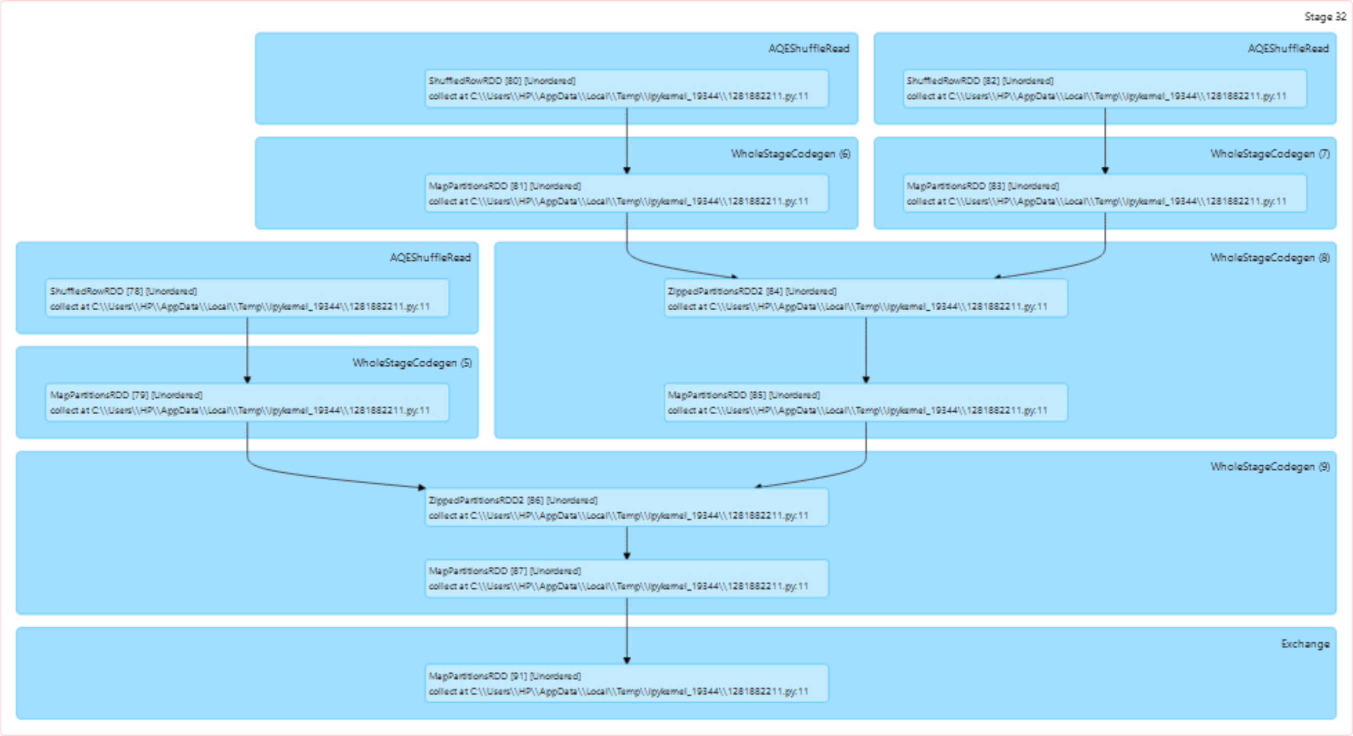

Stage 22
Scan csv
FileScanRDD [64]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
MapPartitionsRDD [65]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
WholeStageCodegen (1)
MapPartitionsRDD [66]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
Exchange
MapPartitionsRDD [67]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11

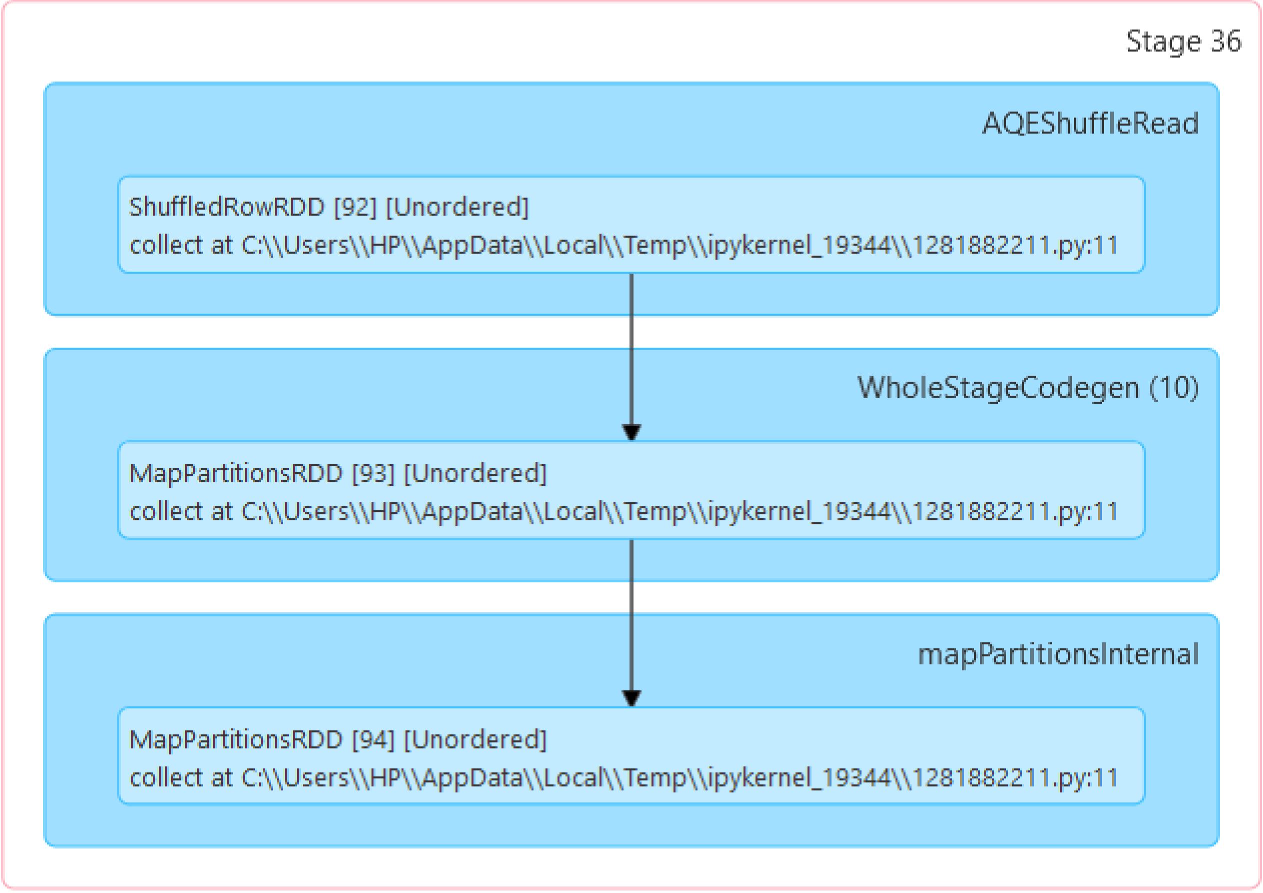


Stage 23
Scan csv
FileScanRDD [68]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
MapPartitionsRDD [69]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
WholeStageCodegen (2)
MapPartitionsRDD [70]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
Exchange
MapPartitionsRDD [71]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11


Stage 24
Scan csv
FileScanRDD [72]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
MapPartitionsRDD [73]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
WholeStageCodegen (4)
MapPartitionsRDD [74]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
Exchange
MapPartitionsRDD [75]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
Stage 26
AQEShuffleRead
ShuffledRowRDD [76] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
BroadcastExchange
MapPartitionsRDD [77] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11


Stage 29
AQEShuffleRead
ShuffledRowRDD [80] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
AQEShuffleRead
ShuffledRowRDD [82] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
WholeStageCodegen (6)
MapPartitionsRDD [81] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
WholeStageCodegen (7)
MapPartitionsRDD [83] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
AQEShuffleRead
ShuffledRowRDD [78] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
WholeStageCodegen (8)
ZippedPartitionsRDD2 [84] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
WholeStageCodegen (5)
MapPartitionsRDD [79] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
MapPartitionsRDD [85] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
WholeStageCodegen (9)
ZippedPartitionsRDD2 [86] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
MapPartitionsRDD [87] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
Exchange
MapPartitionsRDD [88] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
MapPartitionsRDD [89] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
MapPartitionsRDD [90] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11


Stage 32
AQEShuffleRead
ShuffledRowRDD [80] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
AQEShuffleRead
ShuffledRowRDD [82] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
WholeStageCodegen (6)
MapPartitionsRDD [81] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
WholeStageCodegen (7)
MapPartitionsRDD [83] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
AQEShuffleRead
ShuffledRowRDD [78] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
WholeStageCodegen (8)
ZippedPartitionsRDD2 [84] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
WholeStageCodegen (5)
MapPartitionsRDD [79] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
MapPartitionsRDD [85] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
WholeStageCodegen (9)
ZippedPartitionsRDD2 [86] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
MapPartitionsRDD [87] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11
Exchange
MapPartitionsRDD [91] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11



Stage 36

AQEShuffleRead

ShuffledRowRDD [92] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11

WholeStageCodegen (10)

MapPartitionsRDD [93] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11

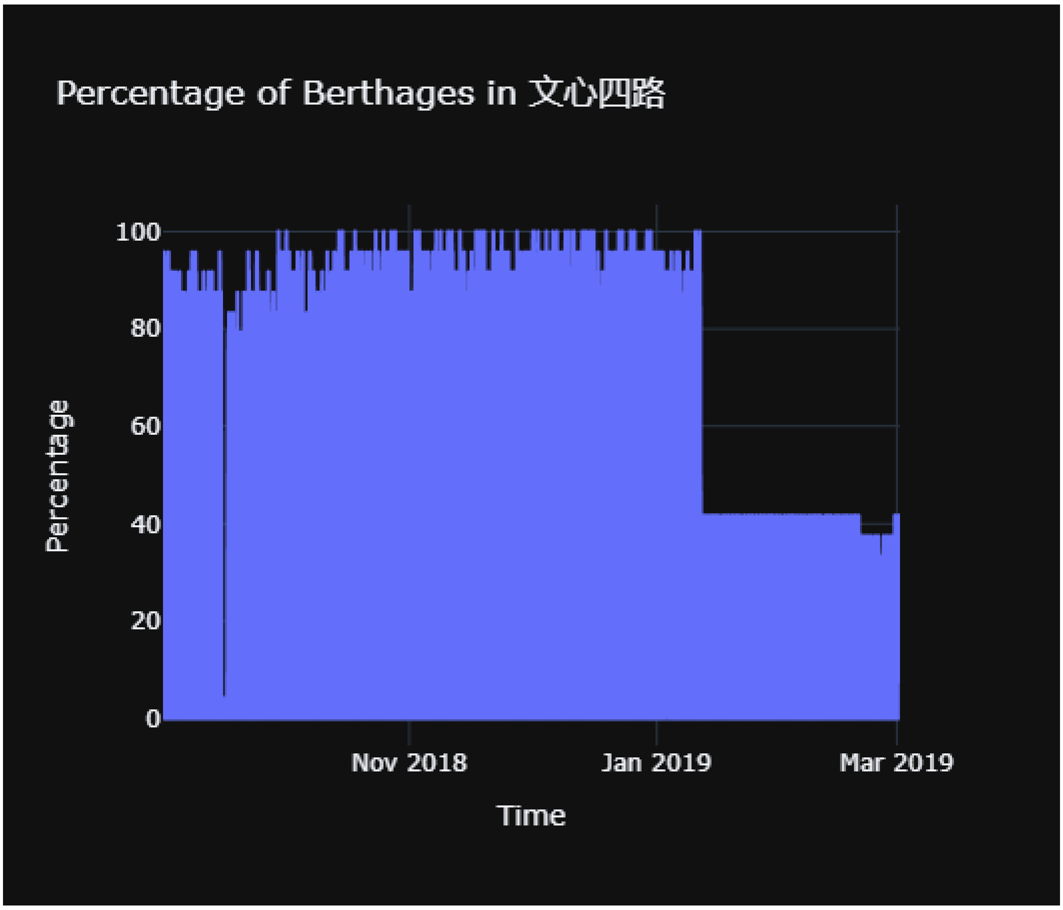
mapPartitionsInternal

MapPartitionsRDD [94] [Unordered]
collect at C:\\Users\\HP\\AppData\\Local\\Temp\\ipykernel_19344\\1281882211.py:11

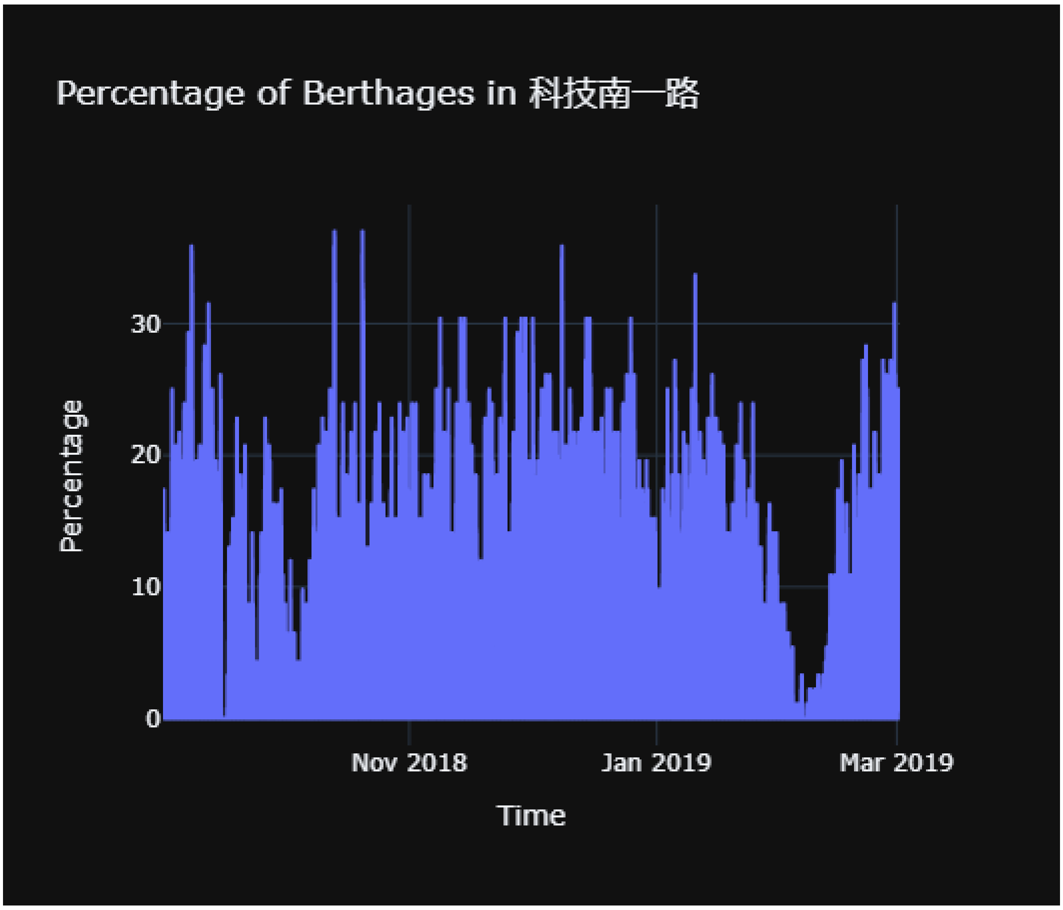
plots

(you can click the section name to visit the html file)

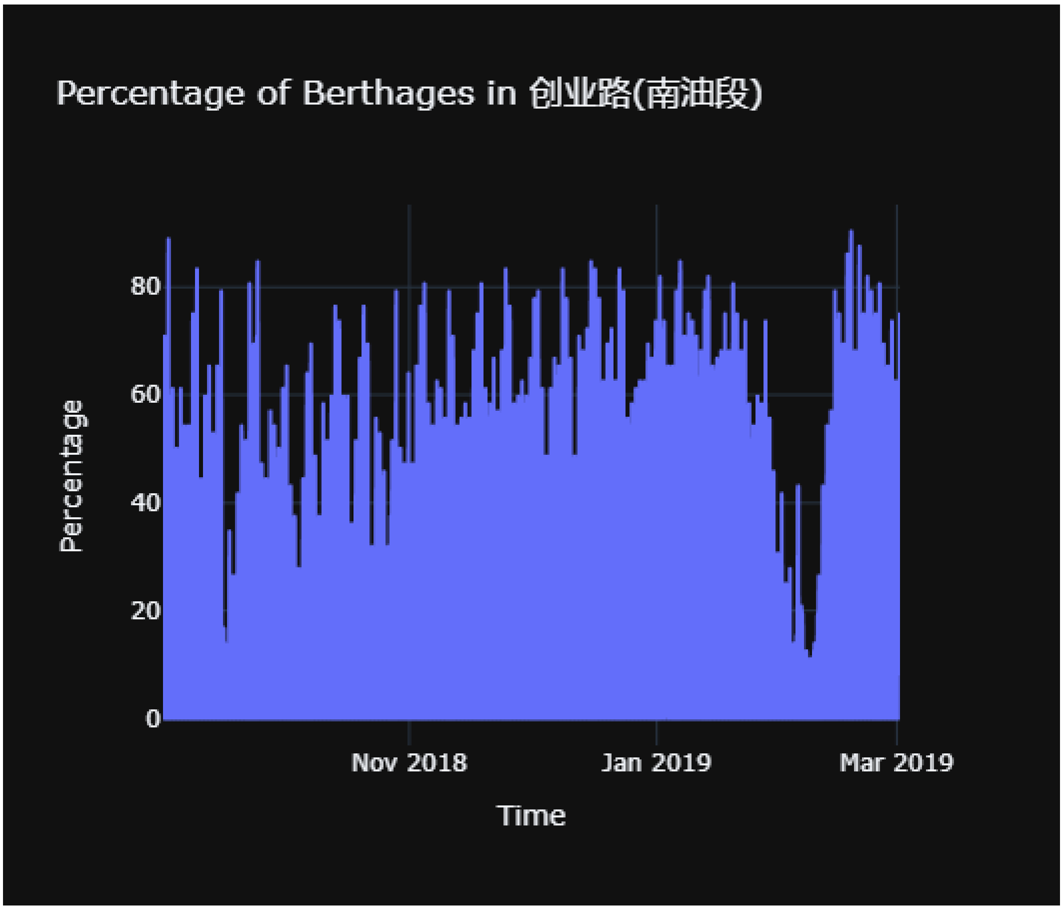
1. 文心四路 the in use berthage almost reach 100% before Jan 12 and only reach 40% after then



2. 科技南一路 Sep 16 reach 0% and never reach 40%



3. 创业路(南油段) increase and decrease repeatedly while highest almost reach 90%



PROBLEMS

1. file storage problem

in the beginning, I used the `pyspark` to store the data `result.write.csv(output_path, header=True)` but it is useless in Windows system and need Hadoop system to run in order to solve this problem, I used `pandas` to store the dataframe `result_pd.to_csv(output_path, index=False, header=True)`