MSIN0166 Data Engineering Individual Assignment

Due: 26th April 2022

Table of Contents

1.0 Introduction

"The National Basketball Association (NBA) is a professional basketball league in North America." (Wikipedia, 2022) NBA was founded in 1946. In terms of basic game information about the NBA, there are 30 teams playing in NBA games, each team would play 82 games every game season and 8 teams from each conference area would be selected for playoff games (Wikipedia, 2022).

During the NBA game season, it has always been a hot topic all around the world either on social media or in businesses related to betting the games. Although NBA is carried out in North America, it has an influence more than just in America but all around the world. There are many live streamers who would commentate each NBA game using different languages and their own styles.

NBA games would produce many static and dynamic data for each player in every team. How these data can bring more value for both the team and the investment companies would be a good topic for data analysts and big data management to explore. In this project, the players in each team's roster will be analysed with their current season and career per game result and their corresponding personal information. The data will be gathered from https://www.basketball-reference.com/ and they will be stored in a relational database (PostgreSQL). The aim of the machine learning model built in this project is to predict whether a player is in the playoff game.

2.0 Workflow

This project is designed in a structured sequence of processes and they are processed separately. The first process is to get the team related information then from each team gets the roster players' basic information. Once the player's basic information is gathered, the current season and career results are gathered. When all the data are ready, machine learning classifiers are used for predicting whether a player will be in a playoff or not.

2.1 Version Control

Version control is very important for project management for either individual or group projects. Version control is a tool to manage the changes made in the source code for the project over time. By using the version control tools, the team can work more efficiently and more effectively. (Atlassian, 2021) In this project, git is used for tracking the code changes. The code changes are committed and pushed after finishing each day's work.

Github: https://github.com/Haiyun-Zou/MSIN0166_Data_Engineering_individual

2.2 Data Lineage

"At its most basic it's the history of your data."(Jeffries, 2020) Data version control is necessary for data science related projects because data version control can help to understand the previous version of the data which could reduce the experiments required for training and reproduction(Jeffries, 2020). In this project, Data Version Control(DVC) is used for tracking the data version. The pipeline of each stage is written in the *dvc.yaml file*. Where there are five stages: **team, players, ML, data_transformtion, write_into_database**. There is also a file called *params.yaml* where all the parameters used in the ML stage are stored. This is useful because the parameters can be manipulated separately without changing the original machine learning code. DVC is also implemented in the auto script file.

In [1]:
```
from IPython.display import Image
```

In [2]:
```
Image("/project/MSIN0166_Data_Engineering_individual/graphs/dvc_files.png", width = 900)
```

Out[2]:

dvc.yaml

MSIN0166_Data_Engineering_individual > dvc.yaml

```yaml
stages:
  team:
    cmd: python python_files/teams_data.py
  players:
    cmd: python python_files/players_data.py
  ML:
    cmd: python python_files/ml.py
    params:
    - ML.seed
    - ML.split
    - ML.shuffle
    - ML.n_components
    metrics:
    - /project/MSIN0166_Data_Engineering_individual/scores.json
  data_transformation:
    cmd: python python_files/data_transformation.py
  write_into_database:
    cmd: python python_files/write_into_database.py
```

params.yaml

MSIN0166_Data_Engineering_individual > params.yaml

```yaml
ML:
  split: 0.20
  seed: 42
  shuffle: True
  n_components: 10
```

In [3]:
```
Image("/project/MSIN0166_Data_Engineering_individual/graphs/dvc_dag.png", width = 600)
```

Out[3]:
```
(Python3) /project/MSIN0166_Data_Engineering_individual(master)$ dvc dag
+------+
| team |
+------+
+---------+
| players |
+---------+
+----+
| ML |
+----+
+--------------------+
| data_transformation |
+--------------------+
+--------------------+
| write_into_database |
+--------------------+
```

In [4]:
```
Image("/project/MSIN0166_Data_Engineering_individual/graphs/auto_script.png", width = 400)
```

Out[4]:

auto_script.sh

MSIN0166_Data_Engineering_individual > auto_script.sh

```
bash install_spark.sh
pip install dvc

dvc exp run -f team
dvc exp run -f players
dvc exp run -f ML
dvc exp run -f data_transformation
dvc exp run -f write_into_database

dvc metrics show

dvc exp show
```

In [5]:
```
Image("/project/MSIN0166_Data_Engineering_individual/graphs/dvc_show.png", width = 900)
```

Out[5]:
```
Path         Accuract    F1_Score    Precision    Recall
scores.json  0.52941     0.5102      0.4902       0.53191
```

Experiment	Created	Accuract	Precision	Recall	F1_Score	ML.split	ML.seed	ML.shuffle	ML.n_components
workspace	**-**	**0.52941**	**0.4902**	**0.53191**	**0.5102**	**0.2**	**42**	**True**	**10**
master	**11:40 AM**	**0.42**	**0.40625**	**0.25**	**0.30952**	**0.2**	**42**	**True**	**10**
├── 5aa14f6 [**exp-01c3a**]	05:44 PM	0.52941	0.4902	0.53191	0.5102	0.2	42	True	10
├── 45c38b0 [**exp-3dc00**]	05:43 PM	0.52941	0.4902	0.53191	0.5102	0.2	42	True	10
├── ec969f9 [**exp-2c366**]	05:42 PM	0.52941	0.4902	0.53191	0.5102	0.2	42	True	10
├── f22f632 [**exp-dfbf8**]	05:39 PM	0.52941	0.4902	0.53191	0.5102	0.2	42	True	10
└── 946e02e [**exp-eec2a**]	05:35 PM	0.52941	0.4902	0.53191	0.5102	0.2	42	True	10

2.3 Terraform

Terraform automation has been setup for the GitHub repository. The file terraform.yml includes the information and configuration for the pull request in Github repository. A new plan will be generated by this terraform automated action when there is a PR created. All the commit and push actions to the master branch are also captured by this automation.

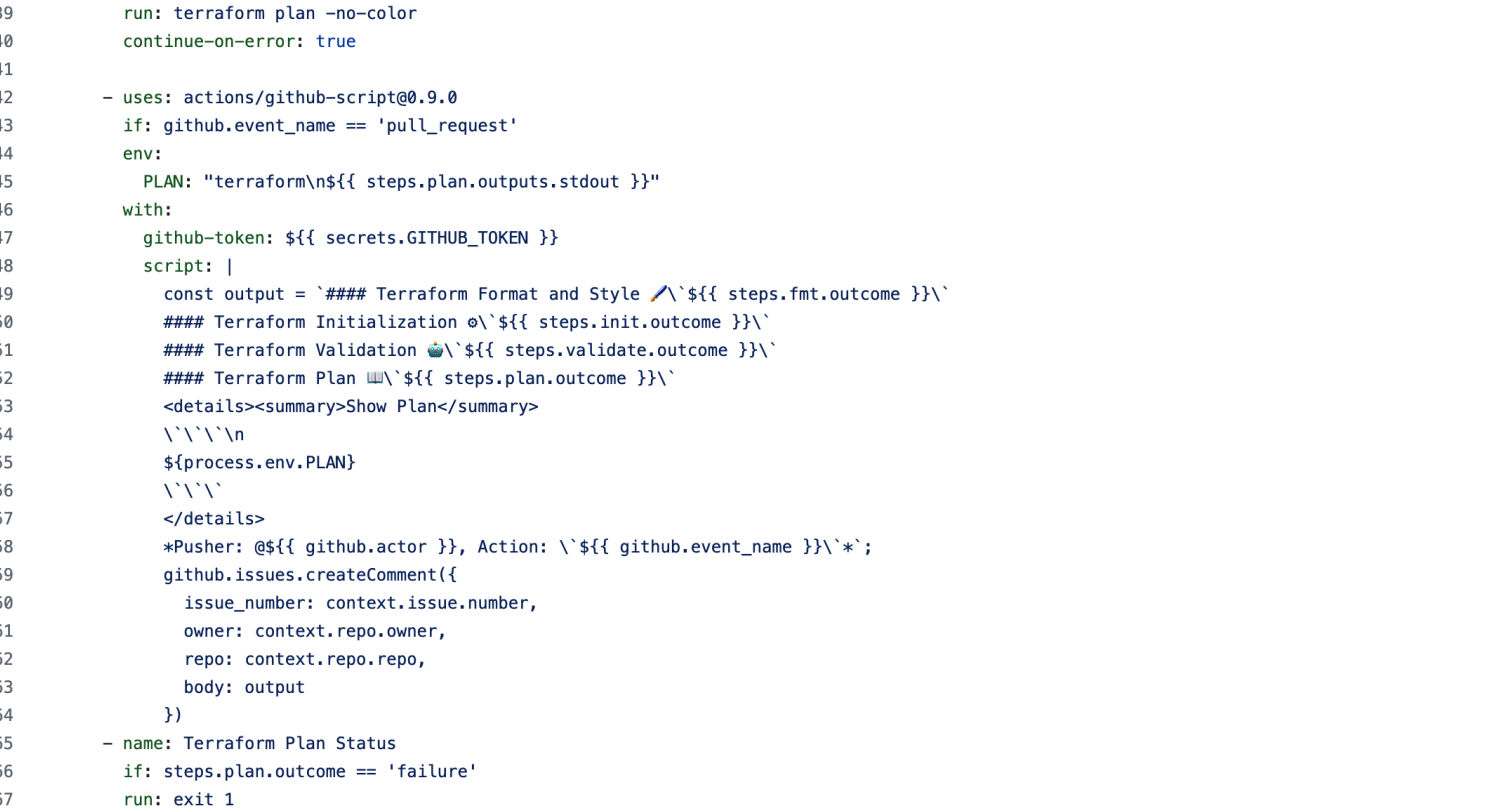
In [6]:
```
Image("/project/MSIN0166_Data_Engineering_individual/graphs/terraform.png", width = 500)
```

Out[6]:

```yaml
name: "Terraform"

on:
  push:
    branches:
      - master
  pull_request:

jobs:
  terraform:
    name: "Terraform"
    runs-on: ubuntu-latest
    environment: myenvironment
    steps:
      - name: Checkout
        uses: actions/checkout@v2

      - name: Setup Terraform
        uses: hashicorp/setup-terraform@v1
        with:
          # terraform_version: 0.13.0:
          cli_config_credentials_token: ${{ secrets.TERRAFORM_API_TOKEN }}

      - name: Terraform Format
        id: fmt
        run: terraform fmt -check

      - name: Terraform Init
        id: init
        run: terraform init

      - name: Terraform Validate
        id: validate
        run: terraform validate -no-color

      - name: Terraform Plan
        id: plan
        if: github.event_name == 'pull_request'
        run: terraform plan -no-color
        continue-on-error: true

      - uses: actions/github-script@0.9.0
        if: github.event_name == 'pull_request'
        env:
          PLAN: "terraform\n${{ steps.plan.outputs.stdout }}"
        with:
          github-token: ${{ secrets.GITHUB_TOKEN }}
          script: |
            const output = `#### Terraform Format and Style 🖌\`${{ steps.fmt.outcome }}\`
            #### Terraform Initialization ⚙\`${{ steps.init.outcome }}\`
            #### Terraform Validation 🤖\`${{ steps.validate.outcome }}\`
            #### Terraform Plan 📖\`${{ steps.plan.outcome }}\`
            <details><summary>Show Plan</summary>
            \`\`\`\n
            ${process.env.PLAN}
            \`\`\`
            </details>
            *Pusher: @${{ github.actor }}, Action: \`${{ github.event_name }}\`*;
            github.issues.createComment({
              issue_number: context.issue.number,
              owner: context.repo.owner,
              repo: context.repo.repo,
              body: output
            })
      - name: Terraform Plan Status
        if: steps.plan.outcome == 'failure'
        run: exit 1

      - name: Terraform Apply
        if: github.ref == 'refs/heads/main' && github.event_name == 'push'
        run: terraform apply -auto-approve
```

In [7]:
```
Image("/project/MSIN0166_Data_Engineering_individual/graphs/commit_automation.png", width = 800)
```

Out[7]:

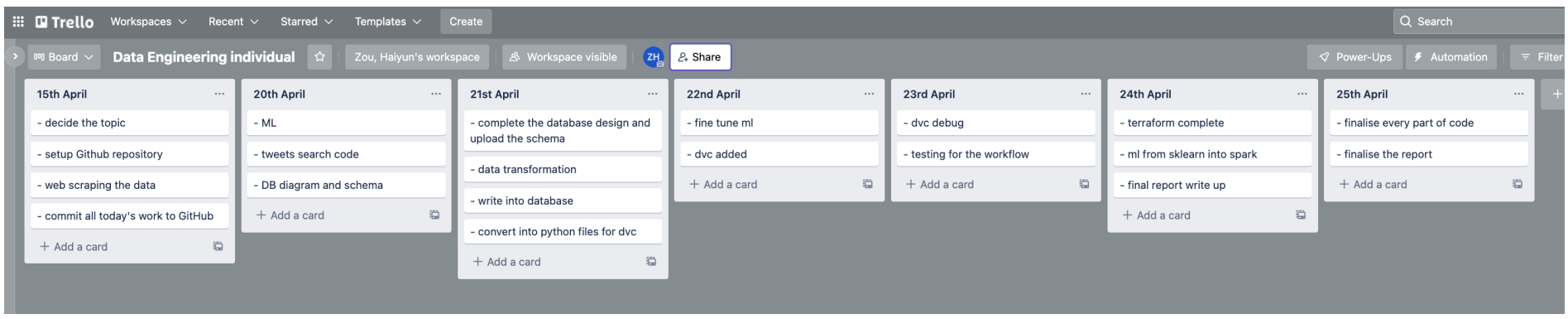


2.4 Project Management

Trello is used to track everyday progression. The details are shown below.

In [8]:
```
Image("/project/MSIN0166_Data_Engineering_individual/graphs/trello.png", width = 1500)
```

Out[8]:



In [9]:
```
# import the packages
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
import re
import os
import requests
from datetime import datetime
import matplotlib.pyplot as plt
```

In [10]:
```
np.random.seed(42)
```

In [11]:
```
# set up the display option
pd.set_option('display.max_columns', None)
```

In [12]:
```
# install the spark
!bash install_spark.sh
```

```
      .:::.      .:::.
    ....yy:      .yy.
    :.  .yy.      y.
         :y:     .:
         .yy     .:
          yy..:
          :y:..
          .y.
         .:..
    .....:
    :::.
```

• Project files and data should be stored in /project. This is shared among everyone in the project.
• Personal files and configuration should be stored in /home/faculty.
• Files outside /project and /home/faculty will be lost when this server is terminated.
• Create custom environments to setup your servers reproducibly.

```
Hit:1 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu bionic InRelease
Get:2 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu bionic-updates InRelease [88.7 kB]
Get:3 http://eu-west-1.ec2.archive.ubuntu.com/ubuntu bionic-backports InRelease [74.6 kB]
Get:4 http://security.ubuntu.com/ubuntu bionic-security InRelease [88.7 kB]
Hit:5 https://packages.cloud.google.com/apt cloud-sdk InRelease
Fetched 252 kB in 0s (622 kB/s)
Reading package lists... Done
Reading package lists... Done
Building dependency tree
Reading state information... Done
Calculating upgrade... Done
The following package was automatically installed and is no longer required:
  python3-crcmod
Use 'sudo apt autoremove' to remove it.
0 upgraded, 0 newly installed, 0 to remove and 0 not upgraded.
Reading package lists... Done
Building dependency tree
Reading state information... Done
openjdk-8-jdk-headless is already the newest version (8u312-b07-0ubuntu1~18.04).
The following package was automatically installed and is no longer required:
  python3-crcmod
Use 'sudo apt autoremove' to remove it.
0 upgraded, 0 newly installed, 0 to remove and 0 not upgraded.
/project /project/MSIN0166_Data_Engineering_individual
spark-3.2.1-bin-hadoop3.2/
spark-3.2.1-bin-hadoop3.2/LICENSE
spark-3.2.1-bin-hadoop3.2/NOTICE
spark-3.2.1-bin-hadoop3.2/R/
spark-3.2.1-bin-hadoop3.2/R/lib/
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/DESCRIPTION
```

spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/INDEX
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/Meta/
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/Meta/Rd.rds
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/Meta/features.rds
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/Meta/hsearch.rds
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/Meta/links.rds
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/Meta/nsInfo.rds
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/Meta/package.rds
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/Meta/vignette.rds
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/NAMESPACE
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/R/
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/R/SparkR
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/R/SparkR.rdb
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/R/SparkR.rdx
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/doc/
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/doc/index.html
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/doc/sparkr-vignettes.R
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/doc/sparkr-vignettes.Rmd
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/doc/sparkr-vignettes.html
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/help/
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/help/AnIndex
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/help/SparkR.rdb
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/help/SparkR.rdx
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/help/aliases.rds
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/help/paths.rds
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/html/
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/html/00Index.html
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/html/R.css
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/profile/
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/profile/general.R
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/profile/shell.R
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/tests/
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/tests/testthat/
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/tests/testthat/test_basic.R
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/worker/
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/worker/daemon.R
spark-3.2.1-bin-hadoop3.2/R/lib/SparkR/worker/worker.R
spark-3.2.1-bin-hadoop3.2/R/lib/sparkr.zip
spark-3.2.1-bin-hadoop3.2/README.md
spark-3.2.1-bin-hadoop3.2/RELEASE
spark-3.2.1-bin-hadoop3.2/bin/
spark-3.2.1-bin-hadoop3.2/bin/beeline
spark-3.2.1-bin-hadoop3.2/bin/beeline.cmd
spark-3.2.1-bin-hadoop3.2/bin/docker-image-tool.sh
spark-3.2.1-bin-hadoop3.2/bin/find-spark-home
spark-3.2.1-bin-hadoop3.2/bin/find-spark-home.cmd
spark-3.2.1-bin-hadoop3.2/bin/load-spark-env.cmd
spark-3.2.1-bin-hadoop3.2/bin/load-spark-env.sh
spark-3.2.1-bin-hadoop3.2/bin/pyspark
spark-3.2.1-bin-hadoop3.2/bin/pyspark.cmd
spark-3.2.1-bin-hadoop3.2/bin/pyspark2.cmd
spark-3.2.1-bin-hadoop3.2/bin/run-example
spark-3.2.1-bin-hadoop3.2/bin/run-example.cmd
spark-3.2.1-bin-hadoop3.2/bin/spark-class
spark-3.2.1-bin-hadoop3.2/bin/spark-class.cmd
spark-3.2.1-bin-hadoop3.2/bin/spark-class2.cmd
spark-3.2.1-bin-hadoop3.2/bin/spark-shell
spark-3.2.1-bin-hadoop3.2/bin/spark-shell.cmd
spark-3.2.1-bin-hadoop3.2/bin/spark-shell2.cmd
spark-3.2.1-bin-hadoop3.2/bin/spark-sql
spark-3.2.1-bin-hadoop3.2/bin/spark-sql.cmd
spark-3.2.1-bin-hadoop3.2/bin/spark-sql2.cmd
spark-3.2.1-bin-hadoop3.2/bin/spark-submit
spark-3.2.1-bin-hadoop3.2/bin/spark-submit.cmd
spark-3.2.1-bin-hadoop3.2/bin/spark-submit2.cmd
spark-3.2.1-bin-hadoop3.2/bin/sparkR
spark-3.2.1-bin-hadoop3.2/bin/sparkR.cmd
spark-3.2.1-bin-hadoop3.2/bin/sparkR2.cmd
spark-3.2.1-bin-hadoop3.2/conf/
spark-3.2.1-bin-hadoop3.2/conf/fairscheduler.xml.template
spark-3.2.1-bin-hadoop3.2/conf/log4j.properties.template
spark-3.2.1-bin-hadoop3.2/conf/metrics.properties.template
spark-3.2.1-bin-hadoop3.2/conf/spark-defaults.conf.template
spark-3.2.1-bin-hadoop3.2/conf/spark-env.sh.template
spark-3.2.1-bin-hadoop3.2/conf/workers.template
spark-3.2.1-bin-hadoop3.2/data/
spark-3.2.1-bin-hadoop3.2/data/graphx/
spark-3.2.1-bin-hadoop3.2/data/graphx/followers.txt
spark-3.2.1-bin-hadoop3.2/data/graphx/users.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/
spark-3.2.1-bin-hadoop3.2/data/mllib/als/
spark-3.2.1-bin-hadoop3.2/data/mllib/als/sample_movielens_ratings.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/als/test.data
spark-3.2.1-bin-hadoop3.2/data/mllib/gmm_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/images/
spark-3.2.1-bin-hadoop3.2/data/mllib/images/license.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/images/origin/
spark-3.2.1-bin-hadoop3.2/data/mllib/images/origin/kittens/
spark-3.2.1-bin-hadoop3.2/data/mllib/images/origin/kittens/29.5.a_b_EGDP022204.jpg

spark-3.2.1-bin-hadoop3.2/data/mllib/images/origin/kittens/54893.jpg
spark-3.2.1-bin-hadoop3.2/data/mllib/images/origin/kittens/DP153539.jpg
spark-3.2.1-bin-hadoop3.2/data/mllib/images/origin/kittens/DP802813.jpg
spark-3.2.1-bin-hadoop3.2/data/mllib/images/origin/kittens/not-image.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/images/origin/license.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/images/origin/multi-channel/
spark-3.2.1-bin-hadoop3.2/data/mllib/images/origin/multi-channel/BGRA.png
spark-3.2.1-bin-hadoop3.2/data/mllib/images/origin/multi-channel/BGRA_alpha_60.png
spark-3.2.1-bin-hadoop3.2/data/mllib/images/origin/multi-channel/chr30.4.184.jpg
spark-3.2.1-bin-hadoop3.2/data/mllib/images/origin/multi-channel/grayscale.jpg
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=kittens/
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=kittens/date=2018-01/
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=kittens/date=2018-01/29.5.a_b_EGDP022204.jpg
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=kittens/date=2018-01/not-image.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=kittens/date=2018-02/
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=kittens/date=2018-02/54893.jpg
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=kittens/date=2018-02/DP153539.jpg
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=kittens/date=2018-02/DP802813.jpg
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=multichannel/
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=multichannel/date=2018-01/
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=multichannel/date=2018-01/BGRA.png
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=multichannel/date=2018-01/BGRA_alpha_60.png
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=multichannel/date=2018-02/
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=multichannel/date=2018-02/chr30.4.184.jpg
spark-3.2.1-bin-hadoop3.2/data/mllib/images/partitioned/cls=multichannel/date=2018-02/grayscale.jpg
spark-3.2.1-bin-hadoop3.2/data/mllib/iris_libsvm.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/kmeans_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/pagerank_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/pic_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/ridge-data/
spark-3.2.1-bin-hadoop3.2/data/mllib/ridge-data/lpsa.data
spark-3.2.1-bin-hadoop3.2/data/mllib/sample_binary_classification_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/sample_fpgrowth.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/sample_isotonic_regression_libsvm_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/sample_kmeans_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/sample_lda_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/sample_lda_libsvm_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/sample_libsvm_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/sample_linear_regression_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/sample_movielens_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/sample_multiclass_classification_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/sample_svm_data.txt
spark-3.2.1-bin-hadoop3.2/data/mllib/streaming_kmeans_data_test.txt
spark-3.2.1-bin-hadoop3.2/data/streaming/
spark-3.2.1-bin-hadoop3.2/data/streaming/AFINN-111.txt
spark-3.2.1-bin-hadoop3.2/examples/
spark-3.2.1-bin-hadoop3.2/examples/jars/
spark-3.2.1-bin-hadoop3.2/examples/jars/scopt_2.12-3.7.1.jar
spark-3.2.1-bin-hadoop3.2/examples/jars/spark-examples_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/examples/src/
spark-3.2.1-bin-hadoop3.2/examples/src/main/
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/JavaHdfsLR.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/JavaLogQuery.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/JavaPageRank.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/JavaSparkPi.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/JavaStatusTrackerDemo.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/JavaTC.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/JavaWordCount.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaAFTSurvivalRegressionExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaALSExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaBinarizerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaBisectingKMeansExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaBucketedRandomProjectionLSHExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaBucketizerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaChiSqSelectorExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaChiSquareTestExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaCorrelationExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaCountVectorizerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaDCTExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaDecisionTreeClassificationExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaDecisionTreeRegressionExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaDocument.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaElementwiseProductExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaEstimatorTransformerParamExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaFMClassifierExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaFMRegressorExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaFPGrowthExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaFeatureHasherExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaGaussianMixtureExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaGeneralizedLinearRegressionExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaGradientBoostedTreeClassifierExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaGradientBoostedTreeRegressorExample.java

spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaImputerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaIndexToStringExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaInteractionExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaIsotonicRegressionExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaKMeansExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaLDAExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaLabeledDocument.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaLinearRegressionWithElasticNetExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaLinearSVCExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaLogisticRegressionSummaryExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaLogisticRegressionWithElasticNetExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaMaxAbsScalerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaMinHashLSHExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaMinMaxScalerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaModelSelectionViaCrossValidationExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaModelSelectionViaTrainValidationSplitExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaMulticlassLogisticRegressionWithElasticNetExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaMultilayerPerceptronClassifierExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaNGramExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaNaiveBayesExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaNormalizerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaOneHotEncoderExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaOneVsRestExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaPCAExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaPipelineExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaPolynomialExpansionExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaPowerIterationClusteringExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaPrefixSpanExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaQuantileDiscretizerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaRFormulaExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaRandomForestClassifierExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaRandomForestRegressorExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaRobustScalerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaSQLTransformerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaStandardScalerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaStopWordsRemoverExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaStringIndexerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaSummarizerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaTfIdfExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaTokenizerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaUnivariateFeatureSelectorExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaVarianceThresholdSelectorExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaVectorAssemblerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaVectorIndexerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaVectorSizeHintExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaVectorSlicerExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/ml/JavaWord2VecExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaALS.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaAssociationRulesExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaBinaryClassificationMetricsExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaBisectingKMeansExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaChiSqSelectorExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaCorrelationsExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaDecisionTreeClassificationExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaDecisionTreeRegressionExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaElementwiseProductExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaGaussianMixtureExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaGradientBoostingClassificationExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaGradientBoostingRegressionExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaHypothesisTestingExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaHypothesisTestingKolmogorovSmirnovTestExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaIsotonicRegressionExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaKMeansExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaKernelDensityEstimationExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaLBFGSExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaLatentDirichletAllocationExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaLogisticRegressionWithLBFGSExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaMultiLabelClassificationMetricsExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaMulticlassClassificationMetricsExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaNaiveBayesExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaPCAExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaPowerIterationClusteringExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaPrefixSpanExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaRandomForestClassificationExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaRandomForestRegressionExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaRankingMetricsExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaRecommendationExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaSVDExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaSVMWithSGDExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaSimpleFPGrowth.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaStratifiedSamplingExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaStreamingTestExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/mllib/JavaSummaryStatisticsExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/JavaSQLDataSourceExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/JavaSparkSQLExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/JavaUserDefinedScalar.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/JavaUserDefinedTypedAggregation.java

spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/JavaUserDefinedUntypedAggregation.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/hive/
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/hive/JavaSparkHiveExample.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/streaming/
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/streaming/JavaStructuredComplexSessionization.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/streaming/JavaStructuredKafkaWordCount.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/streaming/JavaStructuredKerberizedKafkaWordCount.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/streaming/JavaStructuredNetworkWordCount.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/streaming/JavaStructuredNetworkWordCountWindowed.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/sql/streaming/JavaStructuredSessionization.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/streaming/
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/streaming/JavaCustomReceiver.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/streaming/JavaDirectKafkaWordCount.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/streaming/JavaDirectKerberizedKafkaWordCount.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/streaming/JavaNetworkWordCount.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/streaming/JavaQueueStream.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/streaming/JavaRecord.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/streaming/JavaRecoverableNetworkWordCount.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/streaming/JavaSqlNetworkWordCount.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/java/org/apache/spark/examples/streaming/JavaStatefulNetworkWordCount.java
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/als.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/avro_inputformat.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/kmeans.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/logistic_regression.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/aft_survival_regression.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/als_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/binarizer_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/bisecting_k_means_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/bucketed_random_projection_lsh_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/bucketizer_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/chi_square_test_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/chisq_selector_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/correlation_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/count_vectorizer_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/cross_validator.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/dataframe_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/dct_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/decision_tree_classification_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/decision_tree_regression_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/elementwise_product_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/estimator_transformer_param_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/feature_hasher_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/fm_classifier_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/fm_regressor_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/fpgrowth_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/gaussian_mixture_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/generalized_linear_regression_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/gradient_boosted_tree_classifier_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/gradient_boosted_tree_regressor_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/imputer_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/index_to_string_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/interaction_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/isotonic_regression_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/kmeans_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/lda_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/linear_regression_with_elastic_net.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/linearsvc.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/logistic_regression_summary_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/logistic_regression_with_elastic_net.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/max_abs_scaler_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/min_hash_lsh_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/min_max_scaler_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/multiclass_logistic_regression_with_elastic_net.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/multilayer_perceptron_classification.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/n_gram_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/naive_bayes_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/normalizer_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/one_vs_rest_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/onehot_encoder_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/pca_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/pipeline_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/polynomial_expansion_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/power_iteration_clustering_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/prefixspan_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/quantile_discretizer_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/random_forest_classifier_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/random_forest_regressor_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/rformula_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/robust_scaler_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/sql_transformer.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/standard_scaler_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/stopwords_remover_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/string_indexer_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/summarizer_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/tf_idf_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/tokenizer_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/train_validation_split.py

spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/univariate_feature_selector_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/variance_threshold_selector_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/vector_assembler_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/vector_indexer_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/vector_size_hint_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/vector_slicer_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/ml/word2vec_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/binary_classification_metrics_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/bisecting_k_means_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/correlations.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/correlations_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/decision_tree_classification_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/decision_tree_regression_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/elementwise_product_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/fpgrowth_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/gaussian_mixture_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/gaussian_mixture_model.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/gradient_boosting_classification_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/gradient_boosting_regression_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/hypothesis_testing_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/hypothesis_testing_kolmogorov_smirnov_test_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/isotonic_regression_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/k_means_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/kernel_density_estimation_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/kmeans.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/latent_dirichlet_allocation_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/linear_regression_with_sgd_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/logistic_regression.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/logistic_regression_with_lbfgs_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/multi_class_metrics_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/multi_label_metrics_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/naive_bayes_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/normalizer_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/pca_rowmatrix_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/power_iteration_clustering_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/random_forest_classification_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/random_forest_regression_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/random_rdd_generation.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/ranking_metrics_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/recommendation_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/regression_metrics_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/sampled_rdds.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/standard_scaler_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/stratified_sampling_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/streaming_k_means_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/streaming_linear_regression_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/summary_statistics_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/svd_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/svm_with_sgd_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/tf_idf_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/word2vec.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/mllib/word2vec_example.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/pagerank.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/parquet_inputformat.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/pi.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/sort.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/sql/
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/sql/arrow.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/sql/basic.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/sql/datasource.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/sql/hive.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/sql/streaming/
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/sql/streaming/structured_kafka_wordcount.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/sql/streaming/structured_network_wordcount.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/sql/streaming/structured_network_wordcount_windowed.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/sql/streaming/structured_sessionization.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/status_api_demo.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/streaming/
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/streaming/hdfs_wordcount.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/streaming/network_wordcount.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/streaming/network_wordjoinsentiments.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/streaming/queue_stream.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/streaming/recoverable_network_wordcount.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/streaming/sql_network_wordcount.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/streaming/stateful_network_wordcount.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/transitive_closure.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/python/wordcount.py
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/RSparkSQLExample.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/data-manipulation.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/dataframe.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/als.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/bisectingKmeans.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/decisionTree.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/fmClassifier.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/fmRegressor.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/fpm.R

spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/gaussianMixture.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/gbt.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/glm.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/isoreg.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/kmeans.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/kstest.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/lda.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/lm_with_elastic_net.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/logit.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/ml.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/mlp.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/naiveBayes.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/powerIterationClustering.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/prefixSpan.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/randomForest.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/survreg.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/ml/svmLinear.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/streaming/
spark-3.2.1-bin-hadoop3.2/examples/src/main/r/streaming/structured_network_wordcount.R
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/META-INF/
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/META-INF/services/
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/META-INF/services/org.apache.spark.sql.SparkSessionExtensionsProvider
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/META-INF/services/org.apache.spark.sql.jdbc.JdbcConnectionProvider
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/dir1/
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/dir1/dir2/
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/dir1/dir2/file2.parquet
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/dir1/file1.parquet
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/dir1/file3.json
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/employees.json
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/full_user.avsc
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/kv1.txt
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/people.csv
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/people.json
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/people.txt
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/user.avsc
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/users.avro
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/users.orc
spark-3.2.1-bin-hadoop3.2/examples/src/main/resources/users.parquet
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/AccumulatorMetricsTest.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/BroadcastTest.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/DFSReadWriteTest.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/DriverSubmissionTest.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ExceptionHandlingTest.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/GroupByTest.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/HdfsTest.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/LocalALS.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/LocalFileLR.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/LocalKMeans.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/LocalLR.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/LocalPi.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/LogQuery.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/MultiBroadcastTest.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/SimpleSkewedGroupByTest.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/SkewedGroupByTest.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/SparkALS.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/SparkHdfsLR.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/SparkKMeans.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/SparkLR.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/SparkPageRank.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/SparkPi.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/SparkRemoteFileTest.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/SparkTC.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/extensions/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/extensions/AgeExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/extensions/SessionExtensionsWithLoader.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/extensions/SessionExtensionsWithoutLoader.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/extensions/SparkSessionExtensionsTest.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/graphx/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/graphx/AggregateMessagesExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/graphx/Analytics.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/graphx/ComprehensiveExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/graphx/ConnectedComponentsExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/graphx/LiveJournalPageRank.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/graphx/PageRankExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/graphx/SSSPExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/graphx/SynthBenchmark.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/graphx/TriangleCountingExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/AFTSurvivalRegressionExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/ALSExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/BinarizerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/BisectingKMeansExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/BucketedRandomProjectionLSHExample.scala

spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/BucketizerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/ChiSqSelectorExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/ChiSquareTestExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/CorrelationExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/CountVectorizerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/DCTExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/DataFrameExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/DecisionTreeClassificationExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/DecisionTreeExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/DecisionTreeRegressionExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/DeveloperApiExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/ElementwiseProductExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/EstimatorTransformerParamExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/FMClassifierExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/FMRegressorExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/FPGrowthExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/FeatureHasherExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/GBTExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/GaussianMixtureExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/GeneralizedLinearRegressionExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/GradientBoostedTreeClassifierExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/GradientBoostedTreeRegressorExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/ImputerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/IndexToStringExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/InteractionExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/IsotonicRegressionExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/KMeansExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/LDAExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/LinearRegressionExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/LinearRegressionWithElasticNetExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/LinearSVCExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/LogisticRegressionExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/LogisticRegressionSummaryExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/LogisticRegressionWithElasticNetExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/MaxAbsScalerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/MinHashLSHExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/MinMaxScalerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/ModelSelectionViaCrossValidationExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/ModelSelectionViaTrainValidationSplitExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/MulticlassLogisticRegressionWithElasticNetExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/MultilayerPerceptronClassifierExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/NGramExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/NaiveBayesExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/NormalizerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/OneHotEncoderExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/OneVsRestExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/PCAExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/PipelineExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/PolynomialExpansionExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/PowerIterationClusteringExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/PrefixSpanExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/QuantileDiscretizerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/RFormulaExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/RandomForestClassifierExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/RandomForestExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/RandomForestRegressorExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/RobustScalerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/SQLTransformerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/StandardScalerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/StopWordsRemoverExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/StringIndexerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/SummarizerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/TfIdfExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/TokenizerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/UnaryTransformerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/UnivariateFeatureSelectorExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/VarianceThresholdSelectorExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/VectorAssemblerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/VectorIndexerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/VectorSizeHintExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/VectorSlicerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/ml/Word2VecExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/AbstractParams.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/AssociationRulesExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/BinaryClassification.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/BinaryClassificationMetricsExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/BisectingKMeansExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/ChiSqSelectorExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/Correlations.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/CorrelationsExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/CosineSimilarity.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/DecisionTreeClassificationExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/DecisionTreeRegressionExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/DecisionTreeRunner.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/DenseKMeans.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/ElementwiseProductExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/FPGrowthExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/GaussianMixtureExample.scala

spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/GradientBoostedTreesRunner.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/GradientBoostingClassificationExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/GradientBoostingRegressionExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/HypothesisTestingExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/HypothesisTestingKolmogorovSmirnovTestExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/IsotonicRegressionExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/KMeansExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/KernelDensityEstimationExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/LBFGSExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/LDAExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/LatentDirichletAllocationExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/LogisticRegressionWithLBFGSExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/MovieLensALS.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/MultiLabelMetricsExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/MulticlassMetricsExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/MultivariateSummarizer.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/NaiveBayesExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/NormalizerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/PCAOnRowMatrixExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/PCAOnSourceVectorExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/PMMLModelExportExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/PowerIterationClusteringExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/PrefixSpanExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/RandomForestClassificationExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/RandomForestRegressionExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/RandomRDDGeneration.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/RankingMetricsExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/RecommendationExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/SVDExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/SVMWithSGDExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/SampledRDDs.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/SimpleFPGrowth.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/SparseNaiveBayes.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/StandardScalerExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/StratifiedSamplingExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/StreamingKMeansExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/StreamingLinearRegressionExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/StreamingLogisticRegression.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/StreamingTestExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/SummaryStatisticsExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/TFIDFExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/TallSkinnyPCA.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/TallSkinnySVD.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/mllib/Word2VecExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/pythonconverters/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/pythonconverters/AvroConverters.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/RDDRelation.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/SQLDataSourceExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/SimpleTypedAggregator.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/SparkSQLExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/UserDefinedScalar.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/UserDefinedTypedAggregation.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/UserDefinedUntypedAggregation.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/hive/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/hive/SparkHiveExample.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/jdbc/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/jdbc/ExampleJdbcConnectionProvider.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/streaming/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/streaming/StructuredComplexSessionization.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/streaming/StructuredKafkaWordCount.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/streaming/StructuredKerberizedKafkaWordCount.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/streaming/StructuredNetworkWordCount.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/streaming/StructuredNetworkWordCountWindowed.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/sql/streaming/StructuredSessionization.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/CustomReceiver.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/DirectKafkaWordCount.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/DirectKerberizedKafkaWordCount.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/HdfsWordCount.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/NetworkWordCount.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/QueueStream.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/RawNetworkGrep.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/RecoverableNetworkWordCount.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/SqlNetworkWordCount.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/StatefulNetworkWordCount.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/StreamingExamples.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/clickstream/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/clickstream/PageViewGenerator.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scala/org/apache/spark/examples/streaming/clickstream/PageViewStream.scala
spark-3.2.1-bin-hadoop3.2/examples/src/main/scripts/
spark-3.2.1-bin-hadoop3.2/examples/src/main/scripts/getGpusResources.sh
spark-3.2.1-bin-hadoop3.2/jars/
spark-3.2.1-bin-hadoop3.2/jars/HikariCP-2.5.1.jar
spark-3.2.1-bin-hadoop3.2/jars/JLargeArrays-1.5.jar
spark-3.2.1-bin-hadoop3.2/jars/JTransforms-3.1.jar
spark-3.2.1-bin-hadoop3.2/jars/RoaringBitmap-0.9.0.jar
spark-3.2.1-bin-hadoop3.2/jars/ST4-4.0.4.jar
spark-3.2.1-bin-hadoop3.2/jars/activation-1.1.1.jar

spark-3.2.1-bin-hadoop3.2/jars/aircompressor-0.21.jar
spark-3.2.1-bin-hadoop3.2/jars/algebra_2.12-2.0.1.jar
spark-3.2.1-bin-hadoop3.2/jars/annotations-17.0.0.jar
spark-3.2.1-bin-hadoop3.2/jars/antlr-runtime-3.5.2.jar
spark-3.2.1-bin-hadoop3.2/jars/antlr4-runtime-4.8.jar
spark-3.2.1-bin-hadoop3.2/jars/aopalliance-repackaged-2.6.1.jar
spark-3.2.1-bin-hadoop3.2/jars/arpack-2.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/arpack_combined_all-0.1.jar
spark-3.2.1-bin-hadoop3.2/jars/arrow-format-2.0.0.jar
spark-3.2.1-bin-hadoop3.2/jars/arrow-memory-core-2.0.0.jar
spark-3.2.1-bin-hadoop3.2/jars/arrow-memory-netty-2.0.0.jar
spark-3.2.1-bin-hadoop3.2/jars/arrow-vector-2.0.0.jar
spark-3.2.1-bin-hadoop3.2/jars/audience-annotations-0.5.0.jar
spark-3.2.1-bin-hadoop3.2/jars/automaton-1.11-8.jar
spark-3.2.1-bin-hadoop3.2/jars/avro-1.10.2.jar
spark-3.2.1-bin-hadoop3.2/jars/avro-ipc-1.10.2.jar
spark-3.2.1-bin-hadoop3.2/jars/avro-mapred-1.10.2.jar
spark-3.2.1-bin-hadoop3.2/jars/blas-2.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/bonecp-0.8.0.RELEASE.jar
spark-3.2.1-bin-hadoop3.2/jars/breeze-macros_2.12-1.2.jar
spark-3.2.1-bin-hadoop3.2/jars/breeze_2.12-1.2.jar
spark-3.2.1-bin-hadoop3.2/jars/cats-kernel_2.12-2.1.1.jar
spark-3.2.1-bin-hadoop3.2/jars/chill-java-0.10.0.jar
spark-3.2.1-bin-hadoop3.2/jars/chill_2.12-0.10.0.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-cli-1.2.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-codec-1.15.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-collections-3.2.2.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-compiler-3.0.16.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-compress-1.21.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-crypto-1.1.0.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-dbcp-1.4.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-io-2.8.0.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-lang-2.6.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-lang3-3.12.0.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-logging-1.1.3.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-math3-3.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-net-3.1.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-pool-1.5.4.jar
spark-3.2.1-bin-hadoop3.2/jars/commons-text-1.6.jar
spark-3.2.1-bin-hadoop3.2/jars/compress-lzf-1.0.3.jar
spark-3.2.1-bin-hadoop3.2/jars/core-1.1.2.jar
spark-3.2.1-bin-hadoop3.2/jars/curator-client-2.13.0.jar
spark-3.2.1-bin-hadoop3.2/jars/curator-framework-2.13.0.jar
spark-3.2.1-bin-hadoop3.2/jars/curator-recipes-2.13.0.jar
spark-3.2.1-bin-hadoop3.2/jars/datanucleus-api-jdo-4.2.4.jar
spark-3.2.1-bin-hadoop3.2/jars/datanucleus-core-4.1.17.jar
spark-3.2.1-bin-hadoop3.2/jars/datanucleus-rdbms-4.1.19.jar
spark-3.2.1-bin-hadoop3.2/jars/derby-10.14.2.0.jar
spark-3.2.1-bin-hadoop3.2/jars/dropwizard-metrics-hadoop-metrics2-reporter-0.1.2.jar
spark-3.2.1-bin-hadoop3.2/jars/flatbuffers-java-1.9.0.jar
spark-3.2.1-bin-hadoop3.2/jars/generex-1.0.2.jar
spark-3.2.1-bin-hadoop3.2/jars/gson-2.2.4.jar
spark-3.2.1-bin-hadoop3.2/jars/guava-14.0.1.jar
spark-3.2.1-bin-hadoop3.2/jars/hadoop-client-api-3.3.1.jar
spark-3.2.1-bin-hadoop3.2/jars/hadoop-client-runtime-3.3.1.jar
spark-3.2.1-bin-hadoop3.2/jars/hadoop-shaded-guava-1.1.1.jar
spark-3.2.1-bin-hadoop3.2/jars/hadoop-yarn-server-web-proxy-3.3.1.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-beeline-2.3.9.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-cli-2.3.9.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-common-2.3.9.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-exec-2.3.9-core.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-jdbc-2.3.9.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-llap-common-2.3.9.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-metastore-2.3.9.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-serde-2.3.9.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-service-rpc-3.1.2.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-shims-0.23-2.3.9.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-shims-2.3.9.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-shims-common-2.3.9.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-shims-scheduler-2.3.9.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-storage-api-2.7.2.jar
spark-3.2.1-bin-hadoop3.2/jars/hive-vector-code-gen-2.3.9.jar
spark-3.2.1-bin-hadoop3.2/jars/hk2-api-2.6.1.jar
spark-3.2.1-bin-hadoop3.2/jars/hk2-locator-2.6.1.jar
spark-3.2.1-bin-hadoop3.2/jars/hk2-utils-2.6.1.jar
spark-3.2.1-bin-hadoop3.2/jars/htrace-core4-4.1.0-incubating.jar
spark-3.2.1-bin-hadoop3.2/jars/httpclient-4.5.13.jar
spark-3.2.1-bin-hadoop3.2/jars/httpcore-4.4.14.jar
spark-3.2.1-bin-hadoop3.2/jars/istack-commons-runtime-3.0.8.jar
spark-3.2.1-bin-hadoop3.2/jars/ivy-2.5.0.jar
spark-3.2.1-bin-hadoop3.2/jars/jackson-annotations-2.12.3.jar
spark-3.2.1-bin-hadoop3.2/jars/jackson-core-2.12.3.jar
spark-3.2.1-bin-hadoop3.2/jars/jackson-core-asl-1.9.13.jar
spark-3.2.1-bin-hadoop3.2/jars/jackson-databind-2.12.3.jar
spark-3.2.1-bin-hadoop3.2/jars/jackson-dataformat-yaml-2.12.3.jar
spark-3.2.1-bin-hadoop3.2/jars/jackson-datatype-jsr310-2.11.2.jar
spark-3.2.1-bin-hadoop3.2/jars/jackson-mapper-asl-1.9.13.jar
spark-3.2.1-bin-hadoop3.2/jars/jackson-module-scala_2.12-2.12.3.jar
spark-3.2.1-bin-hadoop3.2/jars/jakarta.annotation-api-1.3.5.jar

spark-3.2.1-bin-hadoop3.2/jars/jakarta.inject-2.6.1.jar
spark-3.2.1-bin-hadoop3.2/jars/jakarta.servlet-api-4.0.3.jar
spark-3.2.1-bin-hadoop3.2/jars/jakarta.validation-api-2.0.2.jar
spark-3.2.1-bin-hadoop3.2/jars/jakarta.ws.rs-api-2.1.6.jar
spark-3.2.1-bin-hadoop3.2/jars/jakarta.xml.bind-api-2.3.2.jar
spark-3.2.1-bin-hadoop3.2/jars/janino-3.0.16.jar
spark-3.2.1-bin-hadoop3.2/jars/javassist-3.25.0-GA.jar
spark-3.2.1-bin-hadoop3.2/jars/javax.jdo-3.2.0-m3.jar
spark-3.2.1-bin-hadoop3.2/jars/javolution-5.5.1.jar
spark-3.2.1-bin-hadoop3.2/jars/jaxb-api-2.2.11.jar
spark-3.2.1-bin-hadoop3.2/jars/jaxb-runtime-2.3.2.jar
spark-3.2.1-bin-hadoop3.2/jars/jcl-over-slf4j-1.7.30.jar
spark-3.2.1-bin-hadoop3.2/jars/jdo-api-3.0.1.jar
spark-3.2.1-bin-hadoop3.2/jars/jersey-client-2.34.jar
spark-3.2.1-bin-hadoop3.2/jars/jersey-common-2.34.jar
spark-3.2.1-bin-hadoop3.2/jars/jersey-container-servlet-2.34.jar
spark-3.2.1-bin-hadoop3.2/jars/jersey-container-servlet-core-2.34.jar
spark-3.2.1-bin-hadoop3.2/jars/jersey-hk2-2.34.jar
spark-3.2.1-bin-hadoop3.2/jars/jersey-server-2.34.jar
spark-3.2.1-bin-hadoop3.2/jars/jline-2.14.6.jar
spark-3.2.1-bin-hadoop3.2/jars/joda-time-2.10.10.jar
spark-3.2.1-bin-hadoop3.2/jars/jodd-core-3.5.2.jar
spark-3.2.1-bin-hadoop3.2/jars/jpam-1.1.jar
spark-3.2.1-bin-hadoop3.2/jars/json-1.8.jar
spark-3.2.1-bin-hadoop3.2/jars/json4s-ast_2.12-3.7.0-M11.jar
spark-3.2.1-bin-hadoop3.2/jars/json4s-core_2.12-3.7.0-M11.jar
spark-3.2.1-bin-hadoop3.2/jars/json4s-jackson_2.12-3.7.0-M11.jar
spark-3.2.1-bin-hadoop3.2/jars/json4s-scalap_2.12-3.7.0-M11.jar
spark-3.2.1-bin-hadoop3.2/jars/jsr305-3.0.0.jar
spark-3.2.1-bin-hadoop3.2/jars/jta-1.1.jar
spark-3.2.1-bin-hadoop3.2/jars/jul-to-slf4j-1.7.30.jar
spark-3.2.1-bin-hadoop3.2/jars/kryo-shaded-4.0.2.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-client-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-admissionregistration-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-apiextensions-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-apps-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-autoscaling-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-batch-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-certificates-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-common-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-coordination-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-core-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-discovery-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-events-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-extensions-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-flowcontrol-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-metrics-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-networking-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-node-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-policy-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-rbac-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-scheduling-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/kubernetes-model-storageclass-5.4.1.jar
spark-3.2.1-bin-hadoop3.2/jars/lapack-2.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/leveldbjni-all-1.8.jar
spark-3.2.1-bin-hadoop3.2/jars/libfb303-0.9.3.jar
spark-3.2.1-bin-hadoop3.2/jars/libthrift-0.12.0.jar
spark-3.2.1-bin-hadoop3.2/jars/log4j-1.2.17.jar
spark-3.2.1-bin-hadoop3.2/jars/logging-interceptor-3.12.12.jar
spark-3.2.1-bin-hadoop3.2/jars/lz4-java-1.7.1.jar
spark-3.2.1-bin-hadoop3.2/jars/macro-compat_2.12-1.1.1.jar
spark-3.2.1-bin-hadoop3.2/jars/mesos-1.4.0-shaded-protobuf.jar
spark-3.2.1-bin-hadoop3.2/jars/metrics-core-4.2.0.jar
spark-3.2.1-bin-hadoop3.2/jars/metrics-graphite-4.2.0.jar
spark-3.2.1-bin-hadoop3.2/jars/metrics-jmx-4.2.0.jar
spark-3.2.1-bin-hadoop3.2/jars/metrics-json-4.2.0.jar
spark-3.2.1-bin-hadoop3.2/jars/metrics-jvm-4.2.0.jar
spark-3.2.1-bin-hadoop3.2/jars/minlog-1.3.0.jar
spark-3.2.1-bin-hadoop3.2/jars/netty-all-4.1.68.Final.jar
spark-3.2.1-bin-hadoop3.2/jars/objenesis-2.6.jar
spark-3.2.1-bin-hadoop3.2/jars/okhttp-3.12.12.jar
spark-3.2.1-bin-hadoop3.2/jars/okio-1.14.0.jar
spark-3.2.1-bin-hadoop3.2/jars/opencsv-2.3.jar
spark-3.2.1-bin-hadoop3.2/jars/orc-core-1.6.12.jar
spark-3.2.1-bin-hadoop3.2/jars/orc-mapreduce-1.6.12.jar
spark-3.2.1-bin-hadoop3.2/jars/orc-shims-1.6.12.jar
spark-3.2.1-bin-hadoop3.2/jars/oro-2.0.8.jar
spark-3.2.1-bin-hadoop3.2/jars/osgi-resource-locator-1.0.3.jar
spark-3.2.1-bin-hadoop3.2/jars/paranamer-2.8.jar
spark-3.2.1-bin-hadoop3.2/jars/parquet-column-1.12.2.jar
spark-3.2.1-bin-hadoop3.2/jars/parquet-common-1.12.2.jar
spark-3.2.1-bin-hadoop3.2/jars/parquet-encoding-1.12.2.jar
spark-3.2.1-bin-hadoop3.2/jars/parquet-format-structures-1.12.2.jar
spark-3.2.1-bin-hadoop3.2/jars/parquet-hadoop-1.12.2.jar
spark-3.2.1-bin-hadoop3.2/jars/parquet-jackson-1.12.2.jar
spark-3.2.1-bin-hadoop3.2/jars/protobuf-java-2.5.0.jar
spark-3.2.1-bin-hadoop3.2/jars/py4j-0.10.9.3.jar
spark-3.2.1-bin-hadoop3.2/jars/pyrolite-4.30.jar
spark-3.2.1-bin-hadoop3.2/jars/rocksdbjni-6.20.3.jar

spark-3.2.1-bin-hadoop3.2/jars/scala-collection-compat_2.12-2.1.1.jar
spark-3.2.1-bin-hadoop3.2/jars/scala-compiler-2.12.15.jar
spark-3.2.1-bin-hadoop3.2/jars/scala-library-2.12.15.jar
spark-3.2.1-bin-hadoop3.2/jars/scala-parser-combinators_2.12-1.1.2.jar
spark-3.2.1-bin-hadoop3.2/jars/scala-reflect-2.12.15.jar
spark-3.2.1-bin-hadoop3.2/jars/scala-xml_2.12-1.2.0.jar
spark-3.2.1-bin-hadoop3.2/jars/shapeless_2.12-2.3.3.jar
spark-3.2.1-bin-hadoop3.2/jars/shims-0.9.0.jar
spark-3.2.1-bin-hadoop3.2/jars/slf4j-api-1.7.30.jar
spark-3.2.1-bin-hadoop3.2/jars/slf4j-log4j12-1.7.30.jar
spark-3.2.1-bin-hadoop3.2/jars/snakeyaml-1.27.jar
spark-3.2.1-bin-hadoop3.2/jars/snappy-java-1.1.8.4.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-catalyst_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-core_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-graphx_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-hive-thriftserver_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-hive_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-kubernetes_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-kvstore_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-launcher_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-mesos_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-mllib-local_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-mllib_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-network-common_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-network-shuffle_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-repl_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-sketch_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-sql_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-streaming_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-tags_2.12-3.2.1-tests.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-tags_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-unsafe_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spark-yarn_2.12-3.2.1.jar
spark-3.2.1-bin-hadoop3.2/jars/spire-macros_2.12-0.17.0.jar
spark-3.2.1-bin-hadoop3.2/jars/spire-platform_2.12-0.17.0.jar
spark-3.2.1-bin-hadoop3.2/jars/spire-util_2.12-0.17.0.jar
spark-3.2.1-bin-hadoop3.2/jars/spire_2.12-0.17.0.jar
spark-3.2.1-bin-hadoop3.2/jars/stax-api-1.0.1.jar
spark-3.2.1-bin-hadoop3.2/jars/stream-2.9.6.jar
spark-3.2.1-bin-hadoop3.2/jars/super-csv-2.2.0.jar
spark-3.2.1-bin-hadoop3.2/jars/threeten-extra-1.5.0.jar
spark-3.2.1-bin-hadoop3.2/jars/tink-1.6.0.jar
spark-3.2.1-bin-hadoop3.2/jars/transaction-api-1.1.jar
spark-3.2.1-bin-hadoop3.2/jars/univocity-parsers-2.9.1.jar
spark-3.2.1-bin-hadoop3.2/jars/velocity-1.5.jar
spark-3.2.1-bin-hadoop3.2/jars/xbean-asm9-shaded-4.20.jar
spark-3.2.1-bin-hadoop3.2/jars/xz-1.8.jar
spark-3.2.1-bin-hadoop3.2/jars/zjsonpatch-0.3.0.jar
spark-3.2.1-bin-hadoop3.2/jars/zookeeper-3.6.2.jar
spark-3.2.1-bin-hadoop3.2/jars/zookeeper-jute-3.6.2.jar
spark-3.2.1-bin-hadoop3.2/jars/zstd-jni-1.5.0-4.jar
spark-3.2.1-bin-hadoop3.2/kubernetes/
spark-3.2.1-bin-hadoop3.2/kubernetes/dockerfiles/
spark-3.2.1-bin-hadoop3.2/kubernetes/dockerfiles/spark/
spark-3.2.1-bin-hadoop3.2/kubernetes/dockerfiles/spark/Dockerfile
spark-3.2.1-bin-hadoop3.2/kubernetes/dockerfiles/spark/bindings/
spark-3.2.1-bin-hadoop3.2/kubernetes/dockerfiles/spark/bindings/R/
spark-3.2.1-bin-hadoop3.2/kubernetes/dockerfiles/spark/bindings/R/Dockerfile
spark-3.2.1-bin-hadoop3.2/kubernetes/dockerfiles/spark/bindings/python/
spark-3.2.1-bin-hadoop3.2/kubernetes/dockerfiles/spark/bindings/python/Dockerfile
spark-3.2.1-bin-hadoop3.2/kubernetes/dockerfiles/spark/decom.sh
spark-3.2.1-bin-hadoop3.2/kubernetes/dockerfiles/spark/entrypoint.sh
spark-3.2.1-bin-hadoop3.2/kubernetes/tests/
spark-3.2.1-bin-hadoop3.2/kubernetes/tests/autoscale.py
spark-3.2.1-bin-hadoop3.2/kubernetes/tests/decommissioning.py
spark-3.2.1-bin-hadoop3.2/kubernetes/tests/decommissioning_cleanup.py
spark-3.2.1-bin-hadoop3.2/kubernetes/tests/py_container_checks.py
spark-3.2.1-bin-hadoop3.2/kubernetes/tests/pyfiles.py
spark-3.2.1-bin-hadoop3.2/kubernetes/tests/python_executable_check.py
spark-3.2.1-bin-hadoop3.2/kubernetes/tests/worker_memory_check.py
spark-3.2.1-bin-hadoop3.2/licenses/
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-AnchorJS.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-CC0.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-JLargeArrays.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-JTransforms.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-antlr.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-arpack.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-automaton.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-blas.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-bootstrap.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-cloudpickle.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-d3.min.js.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-dagre-d3.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-datatables.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-dnsjava.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-f2j.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-graphlib-dot.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-istack-commons-runtime.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-jakarta-annotation-api

spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-jakarta-ws-rs-api
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-jakarta.activation-api.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-jakarta.xml.bind-api.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-janino.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-javassist.html
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-javax-transaction-transaction-api.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-javolution.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-jaxb-runtime.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-jline.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-jodd.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-join.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-jquery.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-json-formatter.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-jsp-api.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-kryo.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-leveldbjni.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-machinist.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-matchMedia-polyfill.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-minlog.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-modernizr.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-mustache.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-netlib.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-paranamer.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-pmml-model.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-protobuf.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-py4j.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-pyrolite.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-re2j.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-reflectasm.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-respond.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-sbt-launch-lib.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-scala.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-scopt.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-slf4j.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-sorttable.js.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-spire.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-vis-timeline.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-xmlenc.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-zstd-jni.txt
spark-3.2.1-bin-hadoop3.2/licenses/LICENSE-zstd.txt
spark-3.2.1-bin-hadoop3.2/python/
spark-3.2.1-bin-hadoop3.2/python/.coveragerc
spark-3.2.1-bin-hadoop3.2/python/.gitignore
spark-3.2.1-bin-hadoop3.2/python/MANIFEST.in
spark-3.2.1-bin-hadoop3.2/python/README.md
spark-3.2.1-bin-hadoop3.2/python/dist/
spark-3.2.1-bin-hadoop3.2/python/docs/
spark-3.2.1-bin-hadoop3.2/python/docs/Makefile
spark-3.2.1-bin-hadoop3.2/python/docs/make.bat
spark-3.2.1-bin-hadoop3.2/python/docs/make2.bat
spark-3.2.1-bin-hadoop3.2/python/docs/source/
spark-3.2.1-bin-hadoop3.2/python/docs/source/_static/
spark-3.2.1-bin-hadoop3.2/python/docs/source/_static/copybutton.js
spark-3.2.1-bin-hadoop3.2/python/docs/source/_static/css/
spark-3.2.1-bin-hadoop3.2/python/docs/source/_static/css/pyspark.css
spark-3.2.1-bin-hadoop3.2/python/docs/source/_templates/
spark-3.2.1-bin-hadoop3.2/python/docs/source/_templates/autosummary/
spark-3.2.1-bin-hadoop3.2/python/docs/source/_templates/autosummary/class.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/_templates/autosummary/class_with_docs.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/conf.py
spark-3.2.1-bin-hadoop3.2/python/docs/source/development/
spark-3.2.1-bin-hadoop3.2/python/docs/source/development/contributing.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/development/debugging.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/development/index.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/development/setting_ide.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/development/testing.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/getting_started/
spark-3.2.1-bin-hadoop3.2/python/docs/source/getting_started/index.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/getting_started/install.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/getting_started/quickstart_df.ipynb
spark-3.2.1-bin-hadoop3.2/python/docs/source/getting_started/quickstart_ps.ipynb
spark-3.2.1-bin-hadoop3.2/python/docs/source/index.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/migration_guide/
spark-3.2.1-bin-hadoop3.2/python/docs/source/migration_guide/index.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/migration_guide/koalas_to_pyspark.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/migration_guide/pyspark_1.0_1.2_to_1.3.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/migration_guide/pyspark_1.4_to_1.5.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/migration_guide/pyspark_2.2_to_2.3.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/migration_guide/pyspark_2.3.0_to_2.3.1_above.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/migration_guide/pyspark_2.3_to_2.4.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/migration_guide/pyspark_2.4_to_3.0.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/migration_guide/pyspark_3.1_to_3.2.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/index.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.ml.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.mllib.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.pandas/
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.pandas/extensions.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.pandas/frame.rst

spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.pandas/general_functions.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.pandas/groupby.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.pandas/index.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.pandas/indexing.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.pandas/io.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.pandas/ml.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.pandas/series.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.pandas/window.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.resource.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.sql.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.ss.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/reference/pyspark.streaming.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/arrow_pandas.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/index.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/pandas_on_spark/
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/pandas_on_spark/best_practices.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/pandas_on_spark/faq.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/pandas_on_spark/from_to_dbms.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/pandas_on_spark/index.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/pandas_on_spark/options.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/pandas_on_spark/pandas_pyspark.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/pandas_on_spark/transform_apply.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/pandas_on_spark/typehints.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/pandas_on_spark/types.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/python_packaging.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/sql/
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/sql/arrow_pandas.rst
spark-3.2.1-bin-hadoop3.2/python/docs/source/user_guide/sql/index.rst
spark-3.2.1-bin-hadoop3.2/python/lib/
spark-3.2.1-bin-hadoop3.2/python/lib/PY4J_LICENSE.txt
spark-3.2.1-bin-hadoop3.2/python/lib/py4j-0.10.9.3-src.zip
spark-3.2.1-bin-hadoop3.2/python/lib/pyspark.zip
spark-3.2.1-bin-hadoop3.2/python/mypy.ini
spark-3.2.1-bin-hadoop3.2/python/pylintrc
spark-3.2.1-bin-hadoop3.2/python/pyspark/
spark-3.2.1-bin-hadoop3.2/python/pyspark/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/__init__.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/__pycache__/
spark-3.2.1-bin-hadoop3.2/python/pyspark/__pycache__/install.cpython-38.pyc
spark-3.2.1-bin-hadoop3.2/python/pyspark/_globals.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/_typing.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/accumulators.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/accumulators.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/broadcast.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/broadcast.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/cloudpickle/
spark-3.2.1-bin-hadoop3.2/python/pyspark/cloudpickle/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/cloudpickle/cloudpickle.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/cloudpickle/cloudpickle_fast.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/cloudpickle/compat.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/conf.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/conf.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/context.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/context.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/daemon.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/files.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/files.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/find_spark_home.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/install.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/java_gateway.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/join.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/_typing.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/base.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/base.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/classification.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/classification.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/clustering.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/clustering.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/common.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/common.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/evaluation.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/evaluation.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/feature.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/feature.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/fpm.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/fpm.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/functions.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/functions.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/image.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/image.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/linalg/
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/linalg/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/linalg/__init__.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/param/
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/param/__init__.py

spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/param/__init__.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/param/_shared_params_code_gen.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/param/_shared_params_code_gen.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/param/shared.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/param/shared.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/pipeline.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/pipeline.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/recommendation.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/recommendation.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/regression.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/regression.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/stat.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/stat.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_algorithms.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_base.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_evaluation.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_feature.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_image.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_linalg.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_param.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_persistence.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_pipeline.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_stat.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_training_summary.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_tuning.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_util.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tests/test_wrapper.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tree.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tree.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tuning.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/tuning.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/util.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/util.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/wrapper.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/ml/wrapper.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/_typing.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/classification.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/classification.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/clustering.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/clustering.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/common.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/common.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/evaluation.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/evaluation.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/feature.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/feature.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/fpm.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/fpm.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/linalg/
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/linalg/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/linalg/__init__.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/linalg/distributed.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/linalg/distributed.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/random.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/random.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/recommendation.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/recommendation.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/regression.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/regression.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/stat/
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/stat/KernelDensity.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/stat/KernelDensity.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/stat/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/stat/__init__.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/stat/_statistics.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/stat/_statistics.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/stat/distribution.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/stat/distribution.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/stat/test.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/stat/test.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/tests/
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/tests/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/tests/test_algorithms.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/tests/test_feature.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/tests/test_linalg.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/tests/test_stat.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/tests/test_streaming_algorithms.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/tests/test_util.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/tree.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/tree.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/util.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/mllib/util.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/_typing.py

spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/accessors.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/base.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/categorical.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/config.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/data_type_ops/
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/data_type_ops/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/data_type_ops/base.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/data_type_ops/binary_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/data_type_ops/boolean_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/data_type_ops/categorical_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/data_type_ops/complex_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/data_type_ops/date_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/data_type_ops/datetime_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/data_type_ops/null_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/data_type_ops/num_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/data_type_ops/string_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/data_type_ops/udt_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/datetimes.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/exceptions.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/extensions.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/frame.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/generic.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/groupby.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/indexes/
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/indexes/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/indexes/base.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/indexes/category.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/indexes/datetimes.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/indexes/multi.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/indexes/numeric.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/indexing.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/internal.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/missing/
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/missing/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/missing/common.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/missing/frame.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/missing/groupby.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/missing/indexes.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/missing/series.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/missing/window.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/ml.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/mlflow.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/namespace.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/numpy_compat.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/plot/
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/plot/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/plot/core.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/plot/matplotlib.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/plot/plotly.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/series.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/spark/
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/spark/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/spark/accessors.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/spark/functions.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/spark/utils.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/sql_processor.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/strings.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/test_base.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/test_binary_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/test_boolean_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/test_categorical_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/test_complex_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/test_date_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/test_datetime_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/test_null_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/test_num_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/test_string_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/test_udt_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/data_type_ops/testing_utils.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/indexes/
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/indexes/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/indexes/test_base.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/indexes/test_category.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/indexes/test_datetime.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/plot/
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/plot/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/plot/test_frame_plot.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/plot/test_frame_plot_matplotlib.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/plot/test_frame_plot_plotly.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/plot/test_series_plot.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/plot/test_series_plot_matplotlib.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/plot/test_series_plot_plotly.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_categorical.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_config.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_csv.py

spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_dataframe.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_dataframe_conversion.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_dataframe_spark_io.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_default_index.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_expanding.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_extension.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_frame_spark.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_groupby.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_indexing.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_indexops_spark.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_internal.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_namespace.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_numpy_compat.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_ops_on_diff_frames.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_ops_on_diff_frames_groupby.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_ops_on_diff_frames_groupby_expanding.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_ops_on_diff_frames_groupby_rolling.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_repr.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_reshape.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_rolling.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_series.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_series_conversion.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_series_datetime.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_series_string.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_spark_functions.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_sql.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_stats.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_typedef.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_utils.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/tests/test_window.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/typedef/
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/typedef/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/typedef/string_typehints.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/typedef/typehints.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/usage_logging/
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/usage_logging/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/usage_logging/usage_logger.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/utils.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/pandas/window.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/profiler.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/profiler.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/py.typed
spark-3.2.1-bin-hadoop3.2/python/pyspark/python/
spark-3.2.1-bin-hadoop3.2/python/pyspark/python/pyspark/
spark-3.2.1-bin-hadoop3.2/python/pyspark/python/pyspark/shell.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/rdd.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/rdd.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/rddsampler.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/resource/
spark-3.2.1-bin-hadoop3.2/python/pyspark/resource/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/resource/information.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/resource/information.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/resource/profile.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/resource/profile.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/resource/requests.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/resource/requests.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/resource/tests/
spark-3.2.1-bin-hadoop3.2/python/pyspark/resource/tests/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/resource/tests/test_resources.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/resultiterable.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/resultiterable.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/serializers.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/shell.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/shuffle.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/__init__.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/_typing.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/avro/
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/avro/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/avro/functions.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/avro/functions.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/catalog.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/catalog.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/column.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/column.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/conf.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/conf.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/context.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/context.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/dataframe.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/dataframe.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/functions.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/functions.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/group.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/group.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/_typing/

spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/_typing/__init__.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/_typing/protocols/
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/_typing/protocols/__init__.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/_typing/protocols/frame.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/_typing/protocols/series.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/conversion.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/conversion.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/functions.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/functions.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/group_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/group_ops.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/map_ops.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/map_ops.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/serializers.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/typehints.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/types.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/pandas/utils.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/readwriter.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/readwriter.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/session.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/session.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/streaming.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/streaming.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_arrow.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_catalog.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_column.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_conf.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_context.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_dataframe.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_datasources.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_functions.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_group.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_pandas_cogrouped_map.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_pandas_grouped_map.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_pandas_map.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_pandas_udf.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_pandas_udf_grouped_agg.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_pandas_udf_scalar.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_pandas_udf_typehints.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_pandas_udf_window.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_readwriter.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_serde.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_session.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_streaming.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_types.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_udf.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/tests/test_utils.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/types.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/types.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/udf.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/udf.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/utils.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/window.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/sql/window.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/statcounter.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/statcounter.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/status.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/status.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/storagelevel.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/storagelevel.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/context.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/context.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/dstream.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/dstream.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/kinesis.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/kinesis.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/listener.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/listener.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/tests/
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/tests/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/tests/test_context.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/tests/test_dstream.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/tests/test_kinesis.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/tests/test_listener.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/streaming/util.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/taskcontext.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/taskcontext.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/testing/
spark-3.2.1-bin-hadoop3.2/python/pyspark/testing/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/testing/mllibutils.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/testing/mlutils.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/testing/pandasutils.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/testing/sqlutils.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/testing/streamingutils.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/testing/utils.py

spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/__init__.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_appsubmit.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_broadcast.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_conf.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_context.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_daemon.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_install_spark.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_join.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_pin_thread.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_profiler.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_rdd.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_rddbarrier.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_readwrite.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_serializers.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_shuffle.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_taskcontext.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_util.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/tests/test_worker.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/traceback_utils.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/util.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/util.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/version.py
spark-3.2.1-bin-hadoop3.2/python/pyspark/version.pyi
spark-3.2.1-bin-hadoop3.2/python/pyspark/worker.py
spark-3.2.1-bin-hadoop3.2/python/pyspark.egg-info/
spark-3.2.1-bin-hadoop3.2/python/pyspark.egg-info/PKG-INFO
spark-3.2.1-bin-hadoop3.2/python/pyspark.egg-info/SOURCES.txt
spark-3.2.1-bin-hadoop3.2/python/pyspark.egg-info/dependency_links.txt
spark-3.2.1-bin-hadoop3.2/python/pyspark.egg-info/requires.txt
spark-3.2.1-bin-hadoop3.2/python/pyspark.egg-info/top_level.txt
spark-3.2.1-bin-hadoop3.2/python/run-tests
spark-3.2.1-bin-hadoop3.2/python/run-tests-with-coverage
spark-3.2.1-bin-hadoop3.2/python/run-tests.py
spark-3.2.1-bin-hadoop3.2/python/setup.cfg
spark-3.2.1-bin-hadoop3.2/python/setup.py
spark-3.2.1-bin-hadoop3.2/python/test_coverage/
spark-3.2.1-bin-hadoop3.2/python/test_coverage/conf/
spark-3.2.1-bin-hadoop3.2/python/test_coverage/conf/spark-defaults.conf
spark-3.2.1-bin-hadoop3.2/python/test_coverage/coverage_daemon.py
spark-3.2.1-bin-hadoop3.2/python/test_coverage/sitecustomize.py
spark-3.2.1-bin-hadoop3.2/python/test_support/
spark-3.2.1-bin-hadoop3.2/python/test_support/SimpleHTTPServer.py
spark-3.2.1-bin-hadoop3.2/python/test_support/hello/
spark-3.2.1-bin-hadoop3.2/python/test_support/hello/hello.txt
spark-3.2.1-bin-hadoop3.2/python/test_support/hello/sub_hello/
spark-3.2.1-bin-hadoop3.2/python/test_support/hello/sub_hello/sub_hello.txt
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/ages.csv
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/ages_newlines.csv
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/orc_partitioned/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/orc_partitioned/_SUCCESS
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/orc_partitioned/b=0/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/orc_partitioned/b=0/c=0/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/orc_partitioned/b=0/c=0/.part-r-00000-829af031-b970-49d6-ad39-30460a0be2c8.orc.crc
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/orc_partitioned/b=0/c=0/part-r-00000-829af031-b970-49d6-ad39-30460a0be2c8.orc
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/orc_partitioned/b=1/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/orc_partitioned/b=1/c=1/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/orc_partitioned/b=1/c=1/.part-r-00000-829af031-b970-49d6-ad39-30460a0be2c8.orc.crc
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/orc_partitioned/b=1/c=1/part-r-00000-829af031-b970-49d6-ad39-30460a0be2c8.orc
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/_SUCCESS
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/_common_metadata
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/_metadata
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2014/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2014/month=9/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2014/month=9/day=1/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2014/month=9/day=1/.part-r-00008.gz.parquet.crc
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2014/month=9/day=1/part-r-00008.gz.parquet
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/month=10/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/month=10/day=25/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/month=10/day=25/.part-r-00002.gz.parquet.crc
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/month=10/day=25/.part-r-00004.gz.parquet.crc
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/month=10/day=25/part-r-00002.gz.parquet
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/month=10/day=25/part-r-00004.gz.parquet
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/month=10/day=26/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/month=10/day=26/.part-r-00005.gz.parquet.crc
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/month=10/day=26/part-r-00005.gz.parquet
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/month=9/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/month=9/day=1/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/month=9/day=1/.part-r-00007.gz.parquet.crc
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/parquet_partitioned/year=2015/month=9/day=1/part-r-00007.gz.parquet
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/people.json
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/people1.json
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/people_array.json
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/people_array_utf16le.json
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/streaming/
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/streaming/text-test.txt

```
spark-3.2.1-bin-hadoop3.2/python/test_support/sql/text-test.txt
spark-3.2.1-bin-hadoop3.2/python/test_support/userlib-0.1.zip
spark-3.2.1-bin-hadoop3.2/python/test_support/userlibrary.py
spark-3.2.1-bin-hadoop3.2/sbin/
spark-3.2.1-bin-hadoop3.2/sbin/decommission-slave.sh
spark-3.2.1-bin-hadoop3.2/sbin/decommission-worker.sh
spark-3.2.1-bin-hadoop3.2/sbin/slaves.sh
spark-3.2.1-bin-hadoop3.2/sbin/spark-config.sh
spark-3.2.1-bin-hadoop3.2/sbin/spark-daemon.sh
spark-3.2.1-bin-hadoop3.2/sbin/spark-daemons.sh
spark-3.2.1-bin-hadoop3.2/sbin/start-all.sh
spark-3.2.1-bin-hadoop3.2/sbin/start-history-server.sh
spark-3.2.1-bin-hadoop3.2/sbin/start-master.sh
spark-3.2.1-bin-hadoop3.2/sbin/start-mesos-dispatcher.sh
spark-3.2.1-bin-hadoop3.2/sbin/start-mesos-shuffle-service.sh
spark-3.2.1-bin-hadoop3.2/sbin/start-slave.sh
spark-3.2.1-bin-hadoop3.2/sbin/start-slaves.sh
spark-3.2.1-bin-hadoop3.2/sbin/start-thriftserver.sh
spark-3.2.1-bin-hadoop3.2/sbin/start-worker.sh
spark-3.2.1-bin-hadoop3.2/sbin/start-workers.sh
spark-3.2.1-bin-hadoop3.2/sbin/stop-all.sh
spark-3.2.1-bin-hadoop3.2/sbin/stop-history-server.sh
spark-3.2.1-bin-hadoop3.2/sbin/stop-master.sh
spark-3.2.1-bin-hadoop3.2/sbin/stop-mesos-dispatcher.sh
spark-3.2.1-bin-hadoop3.2/sbin/stop-mesos-shuffle-service.sh
spark-3.2.1-bin-hadoop3.2/sbin/stop-slave.sh
spark-3.2.1-bin-hadoop3.2/sbin/stop-slaves.sh
spark-3.2.1-bin-hadoop3.2/sbin/stop-thriftserver.sh
spark-3.2.1-bin-hadoop3.2/sbin/stop-worker.sh
spark-3.2.1-bin-hadoop3.2/sbin/stop-workers.sh
spark-3.2.1-bin-hadoop3.2/sbin/workers.sh
spark-3.2.1-bin-hadoop3.2/yarn/
spark-3.2.1-bin-hadoop3.2/yarn/spark-3.2.1-yarn-shuffle.jar
/project/MSIN0166_Data_Engineering_individual
Requirement already satisfied: pyspark in /opt/anaconda/envs/Python3/lib/python3.8/site-packages (3.2.1)
Requirement already satisfied: py4j==0.10.9.3 in /opt/anaconda/envs/Python3/lib/python3.8/site-packages (from pyspark) (0.10.9.3)
```

In [13]:
```
# set up the spark envrionment
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/project/spark-3.2.1-bin-hadoop3.2"
```

In [14]:
```
# import spark
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("PySpark App").config("spark.jars", "postgresql-42.3.2.jar").getOrCreate()
spark.conf.set("spark.sql.parquet.enableVectorizedReader","false")
```

3.0 Data Mining

3.1 Team Data

Firstly, the team's basic information is gathered by using BeautifulSoup from https://www.basketball-reference.com/ . The information on winning and losing is in the same line with the team link and team full name for each team. The line of information is further processed to get the corresponding information and form it in a dictionary.

Secondly, the playoff team information is gathered from https://www.basketball-reference.com/playoffs/NBA_2022.html . Where only the teams in the playoff are listed.

Finally, form the team's basic information into a data frame and merge the playoff data for each team into the data frame. The data frame will be converted into a spark data frame and the schema is printed to check the data type for each column. Finally, the spark data frame is output as a parquet file.

In [15]:
```
# using BeautifulSoup to get the web page information
URL ="https://www.basketball-reference.com/"
page = requests.get(URL)
soup = BeautifulSoup(page.content, "html.parser")
soup_body = str(soup.body)
```

In [16]:
```
# get all the team related data
list_info = re.findall(r'data-stat="payroll_text">(.*)</td></tr>', soup_body)
```

In [17]:
```
list_info
```

Out[17]:
```
['<a href="/contracts/MIA.html" title="Miami Heat Team Payroll">$</a></td><td class="right" data-stat="wins">53</td><td class="right" data-stat="losses">29',
 '<a href="/contracts/BOS.html" title="Boston Celtics Team Payroll">$</a></td><td class="right" data-stat="wins">51</td><td class="right" data-stat="losses">31',
 '<a href="/contracts/MIL.html" title="Milwaukee Bucks Team Payroll">$</a></td><td class="right" data-stat="wins">51</td><td class="right" data-stat="losses">31',
 '<a href="/contracts/PHI.html" title="Philadelphia 76ers Team Payroll">$</a></td><td class="right" data-stat="wins">51</td><td class="right" data-stat="losses">31',
```

'$</td><td class="right" data-stat="wins">48</td><td class="right" data-stat="losses">34',
'$</td><td class="right" data-stat="wins">46</td><td class="right" data-stat="losses">36',
'$</td><td class="right" data-stat="wins">44</td><td class="right" data-stat="losses">38',
'$</td><td class="right" data-stat="wins">44</td><td class="right" data-stat="losses">38',
'$</td><td class="right" data-stat="wins">43</td><td class="right" data-stat="losses">39',
'$</td><td class="right" data-stat="wins">43</td><td class="right" data-stat="losses">39',
'$</td><td class="right" data-stat="wins">37</td><td class="right" data-stat="losses">45',
'$</td><td class="right" data-stat="wins">35</td><td class="right" data-stat="losses">47',
'$</td><td class="right" data-stat="wins">25</td><td class="right" data-stat="losses">57',
'$</td><td class="right" data-stat="wins">23</td><td class="right" data-stat="losses">59',
'$</td><td class="right" data-stat="wins">22</td><td class="right" data-stat="losses">60',
'$</td><td class="right" data-stat="wins">64</td><td class="right" data-stat="losses">18',
'$</td><td class="right" data-stat="wins">56</td><td class="right" data-stat="losses">26',
'$</td><td class="right" data-stat="wins">53</td><td class="right" data-stat="losses">29',
'$</td><td class="right" data-stat="wins">52</td><td class="right" data-stat="losses">30',
'$</td><td class="right" data-stat="wins">49</td><td class="right" data-stat="losses">33',
'$</td><td class="right" data-stat="wins">48</td><td class="right" data-stat="losses">34',
'$</td><td class="right" data-stat="wins">46</td><td class="right" data-stat="losses">36',
'$</td><td class="right" data-stat="wins">42</td><td class="right" data-stat="losses">40',
'$</td><td class="right" data-stat="wins">36</td><td class="right" data-stat="losses">46',
'$</td><td class="right" data-stat="wins">34</td><td class="right" data-stat="losses">48',
'$</td><td class="right" data-stat="wins">33</td><td class="right" data-stat="losses">49',
'$</td><td class="right" data-stat="wins">30</td><td class="right" data-stat="losses">52',
'$</td><td class="right" data-stat="wins">27</td><td class="right" data-stat="losses">55',
'$</td><td class="right" data-stat="wins">24</td><td class="right" data-stat="losses">58',
'$</td><td class="right" data-stat="wins">20</td><td class="right" data-stat="losses">62']

In [18]:
```
# set the team code and their full name into a dictionary
dict_info = {i[20:23]:[re.findall(r'title="(.*) Team Payroll', i)[0], int(re.findall(r'data-stat="wins">(.*)</', i)[0]), int(re.findall(r'data-stat="losses">(.*)', i)[0])]
```

In [19]:
```
# check the playoff data
URL ="https://www.basketball-reference.com/playoffs/NBA_2022.html"
page = requests.get(URL)
soup = BeautifulSoup(page.content, "html.parser")
soup_body = str(soup.body)
```

In [20]:
```
# get the playoff team name
list_playoff = list(set([i.split("html'>")[1] for i in re.findall(r'data-stat="team" >(.*?)</a></td><td', soup_body)]))
```

In [21]:
```
list_playoff
```

Out[21]:
```
['Dallas Mavericks',
 'Minnesota Timberwolves',
 'Atlanta Hawks',
 'Boston Celtics',
 'New Orleans Pelicans',
 'Utah Jazz',
 'Toronto Raptors',
 'Brooklyn Nets',
 'Phoenix Suns',
 'Chicago Bulls',
 'Philadelphia 76ers',
 'Miami Heat',
 'Denver Nuggets',
 'Milwaukee Bucks',
 'Memphis Grizzlies',
 'Golden State Warriors']
```

In [22]:
```
# convert the team data into a data frame
team_df = pd.DataFrame({'team': dict_info.keys(), 'name': [i[0] for i in dict_info.values()], 'win': [int(i[1]) for i in dict_info.values()], 'loss': [int(i[2]) for i in d
```

In [23]:
```
# match all the playoff information with the data frame
list_in_playoff = [1 if i in list_playoff else 0 for i in team_df.name]
```

In [24]:
```
# add the column of  in_playoff into the data frame for team information
team_df['in_playoff'] = list_in_playoff
```

In [25]:
```
# convert the team information into spark data frame
df_team = spark.createDataFrame(team_df)
```

In [26]:
```
# show the schema of the data frame
df_team.printSchema()
```

```
root
 |-- team: string (nullable = true)
 |-- name: string (nullable = true)
 |-- win: long (nullable = true)
 |-- loss: long (nullable = true)
 |-- in_playoff: long (nullable = true)
```

In [27]:
```
# convert the data frame into parquet format
df_team.write.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/team.parquet", mode = 'overwrite')
```

3.2 Players Data

Firstly, load the team information data frame to get the initials for each team.

Secondly, the new data frame for each players is created with the columns of 'team', 'playoff', 'name', 'link', 'No.', 'pos', 'height', 'weight', 'birth', 'age', 'exp', 'is_eastern'. Looping through each team page to get the basic information for each player.

Column Name	Description
team	The team the player belongs to
playoff	Whether this player is in a playoff team
name	The name for each player
link	The unique link for each player
No.	The players number
pos	The position the player is at in each team
height	The height for each player
weight	The weight for each player
birth	The year of birth for each player
age	The age of each player
exp	The year of experience for each player
is_eastern	Whether this player is in a team in the eastern conference

Thirdly, after all the player's links are taken from the team page, the complete version of their current season and career result are gathered using regular expressions. The team's contract link is used for getting the guaranteed data for each player. All the NaN are filled after this section of the process.

In [28]:
```
# read the parquet file for team df
team_df = spark.read.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/team.parquet").toPandas()
```

In [29]:
```
# create a data frame to store each player's information
players = pd.DataFrame(columns = ['team', 'playoff', 'name', 'link', 'No.', 'pos', 'height', 'weight', 'birth', 'age', 'exp', 'is_eastern'])
```

In [30]:
```
players
```

Out[30]:

team	playoff	name	link	No.	pos	height	weight	birth	age	exp	is_eastern

In [31]:

```
# loop through each team page and get all the players in each team and their related information
for t in team_df.team:
    URL =f"https://www.basketball-reference.com/teams/{t}/2022.html"
    page = requests.get(URL)
    soup = BeautifulSoup(page.content, "html.parser")
    soup_body = str(soup.body)

    no = [i for i in re.findall(r'data-stat="number" scope="row">(.*)</th>', soup_body)]
    link_name = [i.split('</a>')[0].split('">') for i in re.findall(r'"><a href="(.*?)data-stat="pos"', soup_body)]
    pos = re.findall(r'data-stat="pos">(.*?)</td>', soup_body)
    height = [float(i.replace('-', '.')) for i in re.findall(r'data-stat="height">(.*?)</td>', soup_body)]
    weight = [int(i) for i in re.findall(r'data-stat="weight">(.*?)</td>', soup_body)]
    birth = re.findall(r'</td><td class="left" csk="(.*)" data-stat="birth_date">', soup_body)
    exp = re.findall(r'data-stat="years_experience">(.*)</td><td class', soup_body)
    list_team = [t for i in range(len(no))]
    playoff = team_df[team_df['team'] == t]['in_playoff'].values[0]
    list_playoff_t = [int(playoff) for i in range(len(no))]
    is_eastern = [1 if 'Eastern' in re.findall(r'>NBA</a> \n(.*)\n', soup_body)[0] else 0 for i in range(len(no))]

    players = players.append(pd.DataFrame({'team': list_team,'playoff': list_playoff_t, 'name': [i[1] for i in link_name], 'link': [i[0] for i in link_name], 'No.': r
```

In [32]:
```
# convert the experience information with 'R' to 0.5 and convert the data type into float64
players['exp'] = players['exp'].replace('R', 0.5)
players['exp'] = players['exp'].astype(np.float64)
```

In [33]:
```
# temporary lists for all the players' information
temp_in_season = []
temp_list_2122 = []
temp_list_career = []
# loop through each player's page and get their game information
for count,l in enumerate(players.link):
    URL =f"https://www.basketball-reference.com{l}"
    page = requests.get(URL)

    soup = BeautifulSoup(page.content, "html.parser")
    soup_body = str(soup.body)


    temp_in_season.append('gamelog/2022' in soup_body)
    temp_list_2122.append(re.findall(fr'{l[1:-5]}/gamelog/2022">(.*?)</td></tr> ', soup_body))
    temp_list_career.append(re.findall(r'</tbody><tfoot><tr><th class="left" data-stat="season" scope="row">Career(.*?)</td></tr>', soup_body))
# add the columns into the players data frame
players['in_2021_22_season'] = temp_in_season
players['2021_2022_season'] = temp_list_2122
players['career'] = temp_list_career
```

In [34]:
```
temp_dict = {}
# loop through the teams and get each player's guarantee
for t in team_df.team:

    URL =f"https://www.basketball-reference.com/contracts/{t}.html"
    page = requests.get(URL)
    soup = BeautifulSoup(page.content, "html.parser")
    soup_body = str(soup.body)
    for i in re.findall(r'data-stat="player"(.*)</td></tr>', soup_body):
        if '.html">' in i and 'scope="row"><em>' not in i:
            temp_dict[re.findall(r'html">(.*)</a>', i)[0]] = re.findall(r'data-stat="remain_gtd">\$(.*)', i)
players = players.merge(pd.DataFrame({'name': temp_dict.keys(), 'guaranteed': temp_dict.values()}), how = 'left', on = 'name')
```

In [35]:
```
# fill all na
players = players.replace((np.inf, -np.inf, np.nan), '0').reset_index(drop = True)
```

In [36]:
```
# convert the list into int
players['guaranteed'] = players['guaranteed'].apply(lambda x: int(x[0].replace(',','')) if len(x)>0 else 0)
```

In [37]:
```
# get the age of each player by 2022-birth year
players['age'] = players['birth'].apply(lambda x: 2022 - int(x))
```

In [38]:
```
# fill the empty information
players['2021_2022_season'] = players['2021_2022_season'].apply(lambda x: x[0] if len(x) > 0 else '0')
players['career'] = players['career'].apply(lambda x: x[0] if len(x) > 0 else '0')
```

In [39]:
```
players
```

Out[39]:

	team	playoff	name	link	No.	pos	height	weight	birth	age	exp	is_eastern	in_2021_22_season	2021_2022_
0	MIN	1	Malik Beasley	/players/b/beaslma01.html	5	SG	6.40	187	1996	26	5.0	0	True	2021-22< <td class= data-

	team	playoff	name	link	No.	pos	height	weight	birth	age	exp	is_eastern	in_2021_22_season	2021_2022_
1	MIN	1	Naz Reid	/players/r/reidna01.html	11	C	6.90	264	1999	23	2.0	0	True	2021-22< <td class= data-
2	MIN	1	Karl-Anthony Towns	/players/t/townska01.html	32	C	6.11	248	1995	27	6.0	0	True	2021-22</ class="s
3	MIN	1	Jarred Vanderbilt	/players/v/vandeja01.html	8	PF	6.90	214	1999	23	3.0	0	True	2021-22< <td class= data-
4	MIN	1	Anthony Edwards	/players/e/edwaran01.html	1	SG	6.40	225	2001	21	1.0	0	True	2021-22< <td class= data-
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
503	DEN	1	DeMarcus Cousins	/players/c/couside01.html	4	C	6.10	270	1990	32	10.0	0	True	2021-22< <td class= data-
504	DEN	1	Markus Howard	/players/h/howarma02.html	00	SG	5.10	175	1999	23	1.0	0	True	2021-22< <td class= data-
505	DEN	1	Vlatko Čančar	/players/c/cancavl01.html	31	PF	6.80	236	1997	25	2.0	0	True	2021-22< <td class= data-
506	DEN	1	Michael Porter Jr.	/players/p/portemi01.html	1	SF	6.10	218	1998	24	2.0	0	True	2021-22< <td class= data-
507	DEN	1	Jamal Murray	/players/m/murraja01.html		PG	6.30	215	1997	25	5.0	0	False	

508 rows × 16 columns

Finally, the function *get_info* is written to return the specific score with its own pattern, therefore, it can be called straight away instead of writing repeated code several times.

There are 16 columns of data selected for both the current season and the career scores from the original data source.

Score type	Description
G	Games
GS	Game Started
MP	Minutes Played Per Game
FG%	Field Goals Percent
3P%	3 Points Percent
2P%	2 Points Percent
eFG%	Effective Field Goal Percent
FT%	Free Through Percent
ORB	Offensive Rebounds Per Game
DRB	Defensive Rebounds Per Game
AST	Assists Per Game
STL	Steals Per Game

Score type	Description
BLK	Blocks Per Game
TOV	Turnovers Per Game
PF	Personal Fouls Per Game
PTS	Points Per Game

Each of the scores in the above table is generated by calling the get_info function for both the 2021-2022 season and the player's career results. The columns with the original information are dropped and the final data frame is converted into a spark data frame for checking the data type and it is converted into parquet files for further uses.

In [40]:
```
def get_info(pattern, list_info):
    '''
    function to get information for different attributes
    '''
    temp_list = []
    for count, i in enumerate(list_info):
        if i!='0':
            result = re.findall(pattern, i)[0]
            if 'strong' in result:
                result = result.replace('strong','').replace('/','').replace('<','').replace('>','')
            if result == '':
                result = 0
            temp_list.append(result)
        else:
            temp_list.append(0)
    return temp_list
```

In [41]:
```
# create new columns and get the corresponding information and convert the data type
players['G_2122'] = get_info((r'data-stat="g">(.*?)</td'), players['2021_2022_season'])
players['G_2122'] = players['G_2122'].astype(np.int64)
players['GS_2122'] = get_info((r'data-stat="gs">(.*?)</td'), players['2021_2022_season'])
players['GS_2122'] = players['GS_2122'].astype(np.int64)
players['MP_2122'] = get_info((r'data-stat="mp_per_g">(.*?)</td'), players['2021_2022_season'])
players['MP_2122'] = players['MP_2122'].astype(np.float64)
players['FG%_2122'] = get_info((r'data-stat="fg_pct">(.*?)</td'), players['2021_2022_season'])
players['FG%_2122'] = players['FG%_2122'].astype(np.float64)
players['3P%_2122'] = get_info((r'data-stat="fg3_pct">(.*?)</td'), players['2021_2022_season'])
players['3P%_2122'] = players['3P%_2122'].astype(np.float64)
players['2P%_2122'] = get_info((r'data-stat="fg2_pct">(.*?)</td'), players['2021_2022_season'])
players['2P%_2122'] = players['2P%_2122'].astype(np.float64)
players['eFG%_2122'] = get_info((r'data-stat="efg_pct">(.*?)</td'), players['2021_2022_season'])
players['eFG%_2122'] = players['eFG%_2122'].astype(np.float64)
players['FT%_2122'] = get_info((r'data-stat="ft_pct">(.*?)</td'), players['2021_2022_season'])
players['FT%_2122'] = players['FT%_2122'].astype(np.float64)
players['ORB_2122'] = get_info((r'data-stat="orb_per_g">(.*?)</td'), players['2021_2022_season'])
players['ORB_2122'] = players['ORB_2122'].astype(np.float64)
players['DRB_2122'] = get_info((r'data-stat="drb_per_g">(.*?)</td'), players['2021_2022_season'])
players['DRB_2122'] = players['DRB_2122'].astype(np.float64)
players['AST_2122'] = get_info((r'data-stat="ast_per_g">(.*?)</td'), players['2021_2022_season'])
players['AST_2122'] = players['AST_2122'].astype(np.float64)
players['STL_2122'] = get_info((r'data-stat="stl_per_g">(.*?)</td'), players['2021_2022_season'])
players['STL_2122'] = players['STL_2122'].astype(np.float64)
players['BLK_2122'] = get_info((r'data-stat="blk_per_g">(.*?)</td'), players['2021_2022_season'])
players['BLK_2122'] = players['BLK_2122'].astype(np.float64)
players['TOV_2122'] = get_info((r'data-stat="tov_per_g">(.*?)</td'), players['2021_2022_season'])
players['TOV_2122'] = players['TOV_2122'].astype(np.float64)
players['PF_2122'] = get_info((r'data-stat="pf_per_g">(.*?)</td'), players['2021_2022_season'])
players['PF_2122'] = players['PF_2122'].astype(np.float64)
players['PTS_2122'] = get_info((r'data-stat="pts_per_g">(.*)'), players['2021_2022_season'])
players['PTS_2122'] = players['PTS_2122'].astype(np.float64)
```

In [42]:
```
# create new columns and get the corresponding information and convert the data type
players['G_career'] = get_info((r'data-stat="g">(.*?)</td'), players['career'])
players['G_career'] = players['G_career'].astype(np.int64)
players['GS_career'] = get_info((r'data-stat="gs">(.*?)</td'), players['career'])
players['GS_career'] = players['GS_career'].astype(np.int64)
players['MP_career'] = get_info((r'data-stat="mp_per_g">(.*?)</td'), players['career'])
players['MP_career'] = players['MP_career'].astype(np.float64)
players['FG%_career'] = get_info((r'data-stat="fg_pct">(.*?)</td'), players['career'])
players['FG%_career'] = players['FG%_career'].astype(np.float64)
players['3P%_career'] = get_info((r'data-stat="fg3_pct">(.*?)</td'), players['career'])
players['3P%_career'] = players['3P%_career'].astype(np.float64)
players['2P%_career'] = get_info((r'data-stat="fg2_pct">(.*?)</td'), players['career'])
players['2P%_career'] = players['2P%_career'].astype(np.float64)
players['eFG%_career'] = get_info((r'data-stat="efg_pct">(.*?)</td'), players['career'])
players['eFG%_career'] = players['eFG%_career'].astype(np.float64)
players['FT%_career'] = get_info((r'data-stat="ft_pct">(.*?)</td'), players['career'])
players['FT%_career'] = players['FT%_career'].astype(np.float64)
players['ORB_career'] = get_info((r'data-stat="orb_per_g">(.*?)</td'), players['career'])
players['ORB_career'] = players['ORB_career'].astype(np.float64)
players['DRB_career'] = get_info((r'data-stat="drb_per_g">(.*?)</td'), players['career'])
```

```
players['DRB_career'] = players['DRB_career'].astype(np.float64)
players['AST_career'] = get_info((r'data-stat="ast_per_g">(.*?)</td'), players['career'])
players['AST_career'] = players['AST_career'].astype(np.float64)
players['STL_career'] = get_info((r'data-stat="stl_per_g">(.*?)</td'), players['career'])
players['STL_career'] = players['STL_career'].astype(np.float64)
players['BLK_career'] = get_info((r'data-stat="blk_per_g">(.*?)</td'), players['career'])
players['BLK_career'] = players['BLK_career'].astype(np.float64)
players['TOV_career'] = get_info((r'data-stat="tov_per_g">(.*?)</td'), players['career'])
players['TOV_career'] = players['TOV_career'].astype(np.float64)
players['PF_career'] = get_info((r'data-stat="pf_per_g">(.*?)</td'), players['career'])
players['PF_career'] = players['PF_career'].astype(np.float64)
players['PTS_career'] = get_info((r'data-stat="pts_per_g">(.*)'), players['career'])
players['PTS_career'] = players['PTS_career'].astype(np.float64)
```

In [43]:
```
players = players.drop(['2021_2022_season', 'career'], axis = 1)
```

In [44]:
```
# convert the players data into spark data frame
players_df = spark.createDataFrame(players)
```

In [45]:
```
# show the schema of the spark data frame
players_df.printSchema()
```

```
root
 |-- team: string (nullable = true)
 |-- playoff: long (nullable = true)
 |-- name: string (nullable = true)
 |-- link: string (nullable = true)
 |-- No.: string (nullable = true)
 |-- pos: string (nullable = true)
 |-- height: double (nullable = true)
 |-- weight: long (nullable = true)
 |-- birth: long (nullable = true)
 |-- age: long (nullable = true)
 |-- exp: double (nullable = true)
 |-- is_eastern: long (nullable = true)
 |-- in_2021_22_season: boolean (nullable = true)
 |-- guaranteed: long (nullable = true)
 |-- G_2122: long (nullable = true)
 |-- GS_2122: long (nullable = true)
 |-- MP_2122: double (nullable = true)
 |-- FG%_2122: double (nullable = true)
 |-- 3P%_2122: double (nullable = true)
 |-- 2P%_2122: double (nullable = true)
 |-- eFG%_2122: double (nullable = true)
 |-- FT%_2122: double (nullable = true)
 |-- ORB_2122: double (nullable = true)
 |-- DRB_2122: double (nullable = true)
 |-- AST_2122: double (nullable = true)
 |-- STL_2122: double (nullable = true)
 |-- BLK_2122: double (nullable = true)
 |-- TOV_2122: double (nullable = true)
 |-- PF_2122: double (nullable = true)
 |-- PTS_2122: double (nullable = true)
 |-- G_career: long (nullable = true)
 |-- GS_career: long (nullable = true)
 |-- MP_career: double (nullable = true)
 |-- FG%_career: double (nullable = true)
 |-- 3P%_career: double (nullable = true)
 |-- 2P%_career: double (nullable = true)
 |-- eFG%_career: double (nullable = true)
 |-- FT%_career: double (nullable = true)
 |-- ORB_career: double (nullable = true)
 |-- DRB_career: double (nullable = true)
 |-- AST_career: double (nullable = true)
 |-- STL_career: double (nullable = true)
 |-- BLK_career: double (nullable = true)
 |-- TOV_career: double (nullable = true)
 |-- PF_career: double (nullable = true)
 |-- PTS_career: double (nullable = true)
```

In [46]:
```
# convert the data frame into parquet format
players_df.write.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/players.parquet", mode = 'overwrite')
```

4.0 Machine Learning

As mentioned in the introduction, there will be 8 teams in each area for each game season selected for playoff games. What if only looking into the players and predicting whether they will be in the playoff games?

4.1 Prepare the data

The players' data are loaded from its parquet file. The first step is to convert their position information from categorical data into dummy

variables. The method of get_dummies() is used and the new columns will be named with a prefix of 'is'. Then the column of 'in_2021_22_season' has been cleaned and replaced with 1/0 for True and False. The unused columns for building the model are dropped. Since the parameters are set in the params.yaml file, this information is loaded in the formate of the dictionary.

The Sklearn train_test_split package is used for splitting the training and testing data. The specific parameters are filled from the dictionary loaded. The ratio for training and testing is 0.8:0.2 and the random seed is set to 42. The training and testing set is split from the beginning due to further feature engineering will be taken place. Therefore splitting from the beginning is important to make sure the testing set is not polluted and mixed with the training set data.

After splitting the training and testing data frame is converted into spark data frames and the columns with score data are selected as new data frames for PCA.

In [47]:
```
# load the data from parquet file
players = spark.read.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/players.parquet").toPandas()
```

In [48]:
```
# convert the categorecal data into dummies
dummies_df = pd.get_dummies(players['pos'], prefix='is')
```

In [49]:
```
# add the columns into the data frame
for c in dummies_df.columns:
    players[c] = list(dummies_df[c])
```

In [50]:
```
# convert from True/False to 1/0
players['in_2021_22_season'] = players['in_2021_22_season'].replace({True:1, False:0})
```

In [51]:
```
players = players.sort_values('name').reset_index(drop = True)
```

In [52]:
```
players
```

Out[52]:

	team	playoff	name	link	No.	pos	height	weight	birth	age	exp	is_eastern	in_2021_22_season	guaranteed
0	DEN	1	Aaron Gordon	/players/g/gordoaa01.html	50	PF	6.80	235	1995	27	7.0	0	1	80207637
1	PHO	1	Aaron Holiday	/players/h/holidaa01.html	4	PG	6.00	185	1996	26	3.0	0	1	3980551
2	BOS	1	Aaron Nesmith	/players/n/nesmiaa01.html	26	SF	6.50	215	1999	23	1.0	1	1	7435560
3	OKC	0	Aaron Wiggins	/players/w/wiggiaa01.html	21	SG	6.60	200	1999	23	0.5	0	1	1000000
4	ORL	0	Admiral Schofield	/players/s/schofad01.html	25	SF	6.50	241	1997	25	1.0	1	1	169706
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
503	SAS	0	Zach Collins	/players/c/colliza01.html	23	PF	6.11	250	1997	25	3.0	0	1	22000000
504	CHI	1	Zach LaVine	/players/l/lavinza01.html	8	SF	6.50	200	1995	27	7.0	1	1	19500000
505	DEN	1	Zeke Nnaji	/players/n/nnajize01.html	22	PF	6.90	240	2001	21	1.0	0	1	5116560
506	MEM	1	Ziaire Williams	/players/w/willizi02.html	8	SF	6.80	215	2001	21	0.5	0	1	8965080
507	NOP	1	Zion Williamson	/players/w/willizi01.html		PF	6.60	284	2000	22	2.0	0	0	10733400

508 rows × 51 columns

In [53]:
```
# drop the columns that are not used for machine learning
players = players.drop(['team', 'name', 'link', 'pos', 'No.'], axis = 1)
```

In [54]:
```
import yaml
# load the yaml document
params = yaml.safe_load(open("/project/MSIN0166_Data_Engineering_individual/params.yaml"))["ML"]
```

In [55]:
```
params
```

Out[55]: {'split': 0.2, 'seed': 42, 'shuffle': True, 'n_components': 10}

In [56]:
```
from sklearn.model_selection import train_test_split
# split the train and test sets
players_train, players_test = train_test_split(players, test_size = params['split'], shuffle = params['shuffle'], random_state = params['seed'])
```

In [57]:
```
# create spark df
players_train  = spark.createDataFrame(players_train)
players_test  = spark.createDataFrame(players_test)
```

In [58]:
```
# select columns for pca
players_train_trans = players_train.select(players_train.columns[9:-5])
players_test_trans = players_test.select(players_test.columns[9:-5])
```

In [59]:
```
players_train_trans.toPandas().shape
```

Out[59]: (406, 32)

4.2 Feature Engineering

Principal Component Analysis (PCA) is used in the project. PCA is used to reduce the dimensionality of the data set (Jaadi, 2021). In this data set, since there are columns for both the current season and their career game result, and it is hard to make decisions about using/ not using variables, PCA is carried out to reduce the variables in the data set.

Spark is used to carry out the process for PCA. Firstly all the columns are converted into vectors and the StandardScaler is applied to each column. Then only the scaled columns will be used for the PCA and these columns are converted into vectors again. The number of components selected is from the dictionary of all the parameters.

The sum of the PCA explained variance is 91% which means, all the components after PCA can explain 91% of all the columns of scores. Finally, the PCA components columns are combined with the other basic information columns.

In [60]:
```
from pyspark.ml.feature import StandardScaler
from pyspark.ml import Pipeline
from pyspark.ml.feature import VectorAssembler

# assemble the training data first
assemblers = [VectorAssembler(inputCols=[col], outputCol=col + "_vec") for col in players_train_trans.columns]
# standerdise the data
scaler = [StandardScaler(inputCol=col + "_vec", outputCol=col + "_scaled") for col in players_train_trans.columns]
pipeline = Pipeline(stages= assemblers + scaler)
X_train = pipeline.fit(players_train_trans)
X_train = X_train.transform(players_train_trans)


# assemble the testing data
assemblers = [VectorAssembler(inputCols=[col], outputCol=col + "_vec") for col in players_test_trans.columns]
# standerdise the data
scaler = [StandardScaler(inputCol=col + "_vec", outputCol=col + "_scaled") for col in players_test_trans.columns]
pipeline = Pipeline(stages= assemblers + scaler)
X_test = pipeline.fit(players_test_trans)
X_test = X_test.transform(players_test_trans)
```

In [61]:
```
from pyspark.ml.feature import PCA
# assemble the columns into features column
assemblers = VectorAssembler(inputCols= [i for i in  X_train.columns if  '_scaled' in i], outputCol="features")
```

In [62]:
```
# transform the training data and return the PCA result in the column of PCA_features
X_train_v = assemblers.transform(X_train)
PCA_train = PCA(k = params['n_components'], inputCol="features", outputCol = 'PCA_features')


X_train_model = PCA_train.fit(X_train_v)
X_train = X_train_model.transform(X_train_v)
```

In [63]:
```
# transform the testing data and return the PCA result in the column of PCA_features
X_test_v = assemblers.transform(X_test)
PCA_test = PCA(k = params['n_components'] , inputCol="features", outputCol = 'PCA_features')

X_test_model = PCA_test.fit(X_test_v)
X_test = X_test_model.transform(X_test_v)
```

In [64]:
```
X_train.toPandas().shape
```

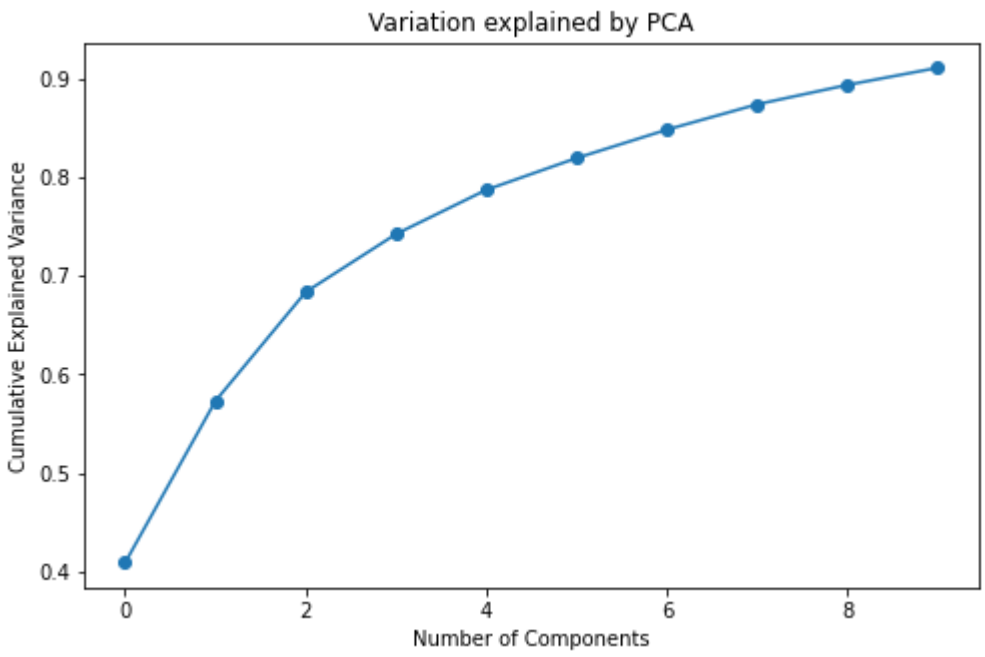
Out[64]: (406, 98)

In [65]:
```
sum(X_train_model.explainedVariance)
```

Out[65]: 0.9112038798252212

In [66]:
```
# plot the PCA variation
plt.figure(figsize = (8,5))
plt.plot(np.cumsum(X_train_model.explainedVariance),linestyle='solid', marker='o')
plt.xlabel('Number of Components')
plt.ylabel('Cumulative Explained Variance')
```

```
plt.title('Variation explained by PCA')

plt.show()
```



In [67]:
```
# get the nonPCA columns
players_train_first = players_train.select(list(set(players_train.columns[:9]) | set(players_train.columns[-5:])))
players_test_first = players_test.select(list(set(players_test.columns[:9]) | set(players_test.columns[-5:])))
```

In [68]:
```
# convert the PCA result column into data frame
temp_train = X_train.select('PCA_features').rdd.map(lambda x: [float(i) for i in x['PCA_features']]).toDF(['PCA' + str(i+1) for i in range(params['n_compor
temp_test = X_test.select('PCA_features').rdd.map(lambda x: [float(i) for i in x['PCA_features']]).toDF(['PCA' + str(i+1) for i in range(params['n_compone
```

In [69]:
```
# convert back to pandas
players_train_first_pd = players_train_first.toPandas()
players_test_first_pd = players_test_first.toPandas()
temp_train_pd = temp_train.toPandas()
temp_test_pd = temp_test.toPandas()
```

In [70]:
```
# merge the columns together
for c in temp_train_pd.columns:
    players_train_first_pd[c] = temp_train_pd[c]
for c in temp_test_pd.columns:
    players_test_first_pd[c] = temp_test_pd[c]
```

In [71]:
```
# convert back to spark df
players_train_final = spark.createDataFrame(players_train_first_pd)
players_test_final = spark.createDataFrame(players_test_first_pd)
```

4.3 Training the Model

DecisionTreeClassifier is selected for this project. The X variables are all converted into features and the playoff column is renamed as label for the model.

In [72]:
```
from pyspark.ml.classification import DecisionTreeClassifier
```

In [73]:
```
# assemble all the X columns into features
assembler = VectorAssembler(inputCols =[i for i in players_train_final.columns if i != 'playoff'] , outputCol='features')
```

In [74]:
```
# transform the train and test data
players_train_final = assembler.transform(players_train_final)
players_test_final= assembler.transform(players_test_final)
```

In [75]:
```
# rename the column playoff to label
players_train_final = players_train_final.withColumnRenamed('playoff', 'label')
players_test_final= players_test_final.withColumnRenamed('playoff', 'label')
```

In [76]:
```
players_train_final.toPandas().shape
```

Out[76]: (406, 25)

In [77]:
```
# build the model and make the prediction
clf_dt= DecisionTreeClassifier(featuresCol="features", labelCol="label")
clf_dt = clf_dt.fit(players_train_final)
pred_clf_dt = clf_dt.transform(players_test_final)
```

4.4 Evaluation of the model

The model is evaluated in both feature importance and confusion matrics. The data frame of the features importance shows that PCA3 and PCA4 have the highest importance. And the overall accuracy is 0.53. The other measurements are written into a JSON file for DVC to store.

In [78]:
```
# convert the importance into a data frame
importance = pd.DataFrame({'feature': [i for i in players_train_final.columns[:-1] if i != 'label'], 'importance':clf_dt.featureImportances.toArray()})
```

In [79]:
```
# order the data frame by its importance in descending order and print the head 10
print(importance.sort_values('importance', ascending = False).head(10))
```

```
   feature  importance
15    PCA3    0.213716
2      age    0.167800
16    PCA4    0.154172
7   weight    0.113176
21    PCA9    0.079174
17    PCA5    0.068354
18    PCA6    0.056910
9   height    0.056822
0     is_C    0.048477
14    PCA2    0.041399
```

In [80]:
```
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
# evaluate the model
evaluator = MulticlassClassificationEvaluator(predictionCol="prediction")
accuracy = evaluator.evaluate(pred_clf_dt)
print(accuracy)
```

```
0.5301365829350608
```

In [81]:
```
from sklearn.metrics import confusion_matrix
```

In [82]:
```
y_p=pred_clf_dt.select("prediction").collect()
y =pred_clf_dt.select("label").collect()

cm = confusion_matrix(y, y_p)
tn, fp, fn, tp = cm.ravel()
print("Confusion Matrix:")
print(cm)
```

```
Confusion Matrix:
[[29 26]
 [22 25]]
```

In [83]:
```
acc = (tn+tp)/(tp+tn+fp+fn)
precision = tp/(tp+fp)
recall = tp/(tp+fn)
```

In [84]:
```
import json
# write the metrics into the scores.json file
with open('/project/MSIN0166_Data_Engineering_individual/scores.json', "w") as fd:
    json.dump({"Accuract":acc , 'Precision':precision, 'Recall': recall, 'F1_Score':2*((precision * recall)/(precision + recall))}, fd, indent=4)
```

4.4.1 PCA component perfomence compare

5 and 10 components were used and compared. The results show that the model with 10 component performs better, therefore, the 10 component is used for building the model.

In [85]:
```
Image("/project/MSIN0166_Data_Engineering_individual/graphs/pca_compare.png", width = 900)
```

Out[85]:

```
Path          Accuract    F1_Score    Precision    Recall
scores.json   0.52941     0.5102      0.4902       0.53191
```

Experiment	Created	Accuract	Precision	Recall	F1_Score	ML.split	ML.seed	ML.shuffle	ML.n_components
workspace	**-**	**0.52941**	**0.4902**	**0.53191**	**0.5102**	**0.2**	**42**	**True**	**10**
master	**06:51 PM**	**0.42**	**0.40625**	**0.25**	**0.30952**	**0.2**	**42**	**True**	**5**
├── eaf35d9 **[exp-01c3a]**	07:06 PM	0.52941	0.4902	0.53191	0.5102	0.2	42	True	10
├── 3557382 **[exp-3dc00]**	07:05 PM	0.52941	0.4902	0.53191	0.5102	0.2	42	True	10
├── cc9efd6 **[exp-eec2a]**	07:05 PM	0.52941	0.4902	0.53191	0.5102	0.2	42	True	10
├── 0b9302d **[exp-2c366]**	07:04 PM	0.53922	0.5	0.06383	0.11321	0.2	42	True	10
├── eb03faa **[exp-dfbf8]**	07:02 PM	0.53922	0.5	0.06383	0.11321	0.2	42	True	10
└── c2b40bb **[exp-59038]**	06:58 PM	0.53922	0.5	0.06383	0.11321	0.2	42	True	5

5.0 Data Transformation

Data is transformed and separated into different data frames to write into the database.

The database is designed with four tables. The teams table contains the team code, team name, wins, loss, playoff and is eastern. This table only held the information in terms of the teams' views. The table players have all the personal information for each player. The other two tables

are the current season and career per game results for each player.

The new tables are formed by selecting the corresponding columns.

In [86]:
```
Image("/project/MSIN0166_Data_Engineering_individual/graphs/NBA_schema.png", width = 600)
```

Out[86]:

In [87]:
```
# load the data
teams = spark.read.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/team.parquet").toPandas()
players = spark.read.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/players.parquet").toPandas()
```

In [88]:
```
players
```

Out[88]:

	team	playoff	name	link	No.	pos	height	weight	birth	age	exp	is_eastern	in_2021_22_season	guaranteed
0	BRK	1	Nic Claxton	/players/c/claxtni01.html	33	C	6.11	215	1999	23	2.0	1	True	178262
1	BRK	1	Day'Ron Sharpe	/players/s/sharpda01.html	20	PF	6.11	265	2001	21	0.5	1	True	411840
2	BRK	1	Kyrie Irving	/players/i/irvinky01.html	11	PG	6.20	195	1992	30	10.0	1	True	34916200
3	BRK	1	Andre Drummond	/players/d/drumman01.html	0, 4	C	6.10	279	1993	29	9.0	1	True	2401537
4	BRK	1	David Duke Jr.	/players/d/dukeda01.html	6	PG	6.50	205	1999	23	0.5	1	True	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
503	BRK	1	Cam Thomas	/players/t/thomaca02.html	24	SG	6.40	210	2001	21	0.5	1	True	417444
504	BRK	1	Blake Griffin	/players/g/griffbl01.html	2	PF	6.90	250	1989	33	11.0	1	True	264169
505	BRK	1	Kevin Durant	/players/d/duranke01.html	7	PF	6.10	240	1988	34	13.0	1	True	22999722
506	BRK	1	Kessler Edwards	/players/e/edwarke02.html	14	SF	6.80	215	2000	22	0.5	1	True	5318
507	BRK	1	LaMarcus Aldridge	/players/a/aldrila01.html	21	C	6.11	250	1985	37	15.0	1	True	264169

508 rows × 46 columns

In [89]:
```
# get all the team_code and thier conference area
dict_team = {i:j for i,j in zip(players['team'], players['is_eastern'])}
```

In [90]:
```
# merge the area information with the team data frame
teams_t = teams.merge(pd.DataFrame({'team':dict_team.keys(), 'is_eastern': dict_team.values()}))
```

In [91]:
```
# convert the columns name
teams_t.columns = ['team_code', 'team_name', 'wins', 'loss', 'playoff', 'is_eastern']
```

In [92]:
```
teams_t
```

Out[92]:

	team_code	team_name	wins	loss	playoff	is_eastern
0	MIN	Minnesota Timberwolves	46	36	1	0
1	LAC	Los Angeles Clippers	42	40	0	0
2	NOP	New Orleans Pelicans	36	46	1	0
3	SAS	San Antonio Spurs	34	48	0	0
4	LAL	Los Angeles Lakers	33	49	0	0
5	SAC	Sacramento Kings	30	52	0	0
6	POR	Portland Trail Blazers	27	55	0	0
7	OKC	Oklahoma City Thunder	24	58	0	0
8	HOU	Houston Rockets	20	62	0	0
9	CLE	Cleveland Cavaliers	44	38	0	1
10	ATL	Atlanta Hawks	43	39	1	1
11	CHO	Charlotte Hornets	43	39	0	1
12	NYK	New York Knicks	37	45	0	1
13	WAS	Washington Wizards	35	47	0	1
14	IND	Indiana Pacers	25	57	0	1
15	DET	Detroit Pistons	23	59	0	1
16	MIA	Miami Heat	53	29	1	1
17	BOS	Boston Celtics	51	31	1	1
18	MIL	Milwaukee Bucks	51	31	1	1
19	PHI	Philadelphia 76ers	51	31	1	1
20	TOR	Toronto Raptors	48	34	1	1
21	CHI	Chicago Bulls	46	36	1	1
22	BRK	Brooklyn Nets	44	38	1	1
23	ORL	Orlando Magic	22	60	0	1
24	PHO	Phoenix Suns	64	18	1	0
25	MEM	Memphis Grizzlies	56	26	1	0
26	GSW	Golden State Warriors	53	29	1	0
27	DAL	Dallas Mavericks	52	30	1	0
28	UTA	Utah Jazz	49	33	1	0
29	DEN	Denver Nuggets	48	34	1	0

In [93]:
```
# extract the information for players table in the database
players_t = players[['link', 'name', 'team', 'birth', 'age', 'No.', 'pos', 'guaranteed', 'height', 'weight', 'exp']]
```

In [94]:
```
# convert the columns name
players_t.columns = ['id', 'name', 'team_code', 'birth_year', 'age', 'number', 'possition', 'guaranteed', 'height', 'weight', 'exp']
```

In [95]:
```
players_t
```

Out[95]:

	id	name	team_code	birth_year	age	number	possition	guaranteed	height	weight	exp
0	/players/c/claxtni01.html	Nic Claxton	BRK	1999	23	33	C	1782621	6.11	215	2.0
1	/players/s/sharpda01.html	Day'Ron Sharpe	BRK	2001	21	20	PF	4118400	6.11	265	0.5
2	/players/i/irvinky01.html	Kyrie Irving	BRK	1992	30	11	PG	34916200	6.20	195	10.0
3	/players/d/drumman01.html	Andre Drummond	BRK	1993	29	0, 4	C	2401537	6.10	279	9.0
4	/players/d/dukeda01.html	David Duke Jr.	BRK	1999	23	6	PG	0	6.50	205	0.5
...	...	...	...	...	...	...	...	...	...	...	...

	id	name	team_code	birth_year	age	number	possition	guaranteed	height	weight	exp
503	/players/t/thomaca02.html	Cam Thomas	BRK	2001	21	24	SG	4174440	6.40	210	0.5
504	/players/g/griffbl01.html	Blake Griffin	BRK	1989	33	2	PF	2641691	6.90	250	11.0
505	/players/d/duranke01.html	Kevin Durant	BRK	1988	34	7	PF	229997220	6.10	240	13.0
506	/players/e/edwarke02.html	Kessler Edwards	BRK	2000	22	14	SF	5318	6.80	215	0.5
507	/players/a/aldrila01.html	LaMarcus Aldridge	BRK	1985	37	21	C	2641691	6.11	250	15.0

508 rows × 11 columns

In [96]:
```
# extract the current season and career results from players
current_season_result = players[['link', 'G_2122', 'GS_2122','MP_2122', 'FG%_2122', '3P%_2122', '2P%_2122', 'eFG%_2122', 'FT%_2122', 'ORB_2122', 'DRE
current_career_result = players[['link', 'G_career', 'GS_career','MP_career', 'FG%_career', '3P%_career', '2P%_career', 'eFG%_career', 'FT%_career', 'ORB_ca
```

In [97]:
```
# convert the columns name
current_season_result.columns = [i.lower() for i in ['id', 'G', 'GS','MP', 'FGperc', '3Pperc', '2Pperc', 'eFGperc', 'FTperc', 'ORB', 'DRB', 'AST', 'STL', 'BLK', 'TOV
current_career_result.columns = [i.lower() for i in ['id', 'G', 'GS','MP', 'FGperc', '3Pperc', '2Pperc', 'eFGperc', 'FTperc', 'ORB', 'DRB', 'AST', 'STL', 'BLK', 'TOV'
```

In [98]:
```
current_season_result
```

Out[98]:

	id	g	gs	mp	fgperc	3pperc	2pperc	efgperc	ftperc	orb	drb	ast	stl	blk	tov	pf	pts
0	/players/c/claxtni01.html	47	19	20.7	0.674	0.000	0.674	0.674	0.581	1.9	3.7	0.9	0.5	1.1	0.8	2.3	8.7
1	/players/s/sharpda01.html	32	8	12.2	0.577	0.286	0.592	0.584	0.585	2.5	2.5	0.5	0.3	0.5	0.9	1.9	6.2
2	/players/i/irvinky01.html	29	29	37.6	0.469	0.418	0.501	0.550	0.915	0.6	3.8	5.8	1.4	0.6	2.5	2.8	27.4
3	/players/d/drumman01.html	73	36	19.7	0.570	0.000	0.574	0.570	0.524	3.1	6.2	1.8	1.1	0.9	1.6	2.6	7.9
4	/players/d/dukeda01.html	22	7	15.5	0.361	0.243	0.423	0.403	0.810	1.4	1.7	0.8	0.6	0.3	0.4	1.6	4.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
503	/players/t/thomaca02.html	67	2	17.6	0.433	0.270	0.520	0.480	0.829	0.2	2.2	1.2	0.5	0.1	0.8	1.0	8.5
504	/players/g/griffbl01.html	56	24	17.1	0.425	0.262	0.565	0.486	0.724	1.1	3.0	1.9	0.5	0.3	0.6	1.7	6.4
505	/players/d/duranke01.html	55	55	37.2	0.518	0.383	0.568	0.570	0.910	0.5	6.9	6.4	0.9	0.9	3.5	2.1	29.9
506	/players/e/edwarke02.html	48	23	20.6	0.412	0.353	0.473	0.502	0.842	0.9	2.7	0.6	0.6	0.5	0.9	1.8	5.9
507	/players/a/aldrila01.html	47	12	22.3	0.550	0.304	0.578	0.566	0.873	1.6	3.9	0.9	0.3	1.0	0.9	1.7	12.9

508 rows × 17 columns

In [99]:
```
current_career_result
```

Out[99]:

	id	g	gs	mp	fgperc	3pperc	2pperc	efgperc	ftperc	orb	drb	ast	stl	blk	tov	pf	pts
0	/players/c/claxtni01.html	94	20	18.7	0.646	0.167	0.658	0.648	0.539	1.6	3.4	0.9	0.5	1.1	0.7	2.0	7.3
1	/players/s/sharpda01.html	32	8	12.2	0.577	0.286	0.592	0.584	0.585	2.5	2.5	0.5	0.3	0.5	0.9	1.9	6.2
2	/players/i/irvinky01.html	611	611	34.0	0.470	0.393	0.505	0.532	0.882	0.8	3.1	5.7	1.3	0.4	2.6	2.3	23.1
3	/players/d/drumman01.html	718	630	29.6	0.540	0.132	0.546	0.541	0.473	4.5	8.7	1.4	1.4	1.5	2.0	3.1	13.8
4	/players/d/dukeda01.html	22	7	15.5	0.361	0.243	0.423	0.403	0.810	1.4	1.7	0.8	0.6	0.3	0.4	1.6	4.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
503	/players/t/thomaca02.html	67	2	17.6	0.433	0.270	0.520	0.480	0.829	0.2	2.2	1.2	0.5	0.1	0.8	1.0	8.5
504	/players/g/griffbl01.html	724	676	32.9	0.493	0.327	0.521	0.517	0.696	2.0	6.2	4.1	0.8	0.5	2.4	2.7	19.8
505	/players/d/duranke01.html	939	936	36.8	0.496	0.384	0.536	0.546	0.884	0.7	6.4	4.3	1.1	1.1	3.2	1.9	27.2
506	/players/e/edwarke02.html	48	23	20.6	0.412	0.353	0.473	0.502	0.842	0.9	2.7	0.6	0.6	0.5	0.9	1.8	5.9
507	/players/a/aldrila01.html	1076	997	33.7	0.493	0.320	0.500	0.499	0.813	2.6	5.5	1.9	0.7	1.1	1.5	2.4	19.1

508 rows × 17 columns

In [100]:
```
# convert all the data frame into spark data frame
teams_t_df = spark.createDataFrame(teams_t)
players_t_df = spark.createDataFrame(players_t)
current_season_result_df = spark.createDataFrame(current_season_result)
current_career_result_df = spark.createDataFrame(current_career_result)
```

In [101]:
```
# show the schema of the data frame
teams_t_df.printSchema()
players_t_df.printSchema()
current_season_result_df.printSchema()
current_career_result_df.printSchema()
```

```
root
```

```
 |-- team_code: string (nullable = true)
 |-- team_name: string (nullable = true)
 |-- wins: long (nullable = true)
 |-- loss: long (nullable = true)
 |-- playoff: long (nullable = true)
 |-- is_eastern: long (nullable = true)

root
 |-- id: string (nullable = true)
 |-- name: string (nullable = true)
 |-- team_code: string (nullable = true)
 |-- birth_year: long (nullable = true)
 |-- age: long (nullable = true)
 |-- number: string (nullable = true)
 |-- possition: string (nullable = true)
 |-- guaranteed: long (nullable = true)
 |-- height: double (nullable = true)
 |-- weight: long (nullable = true)
 |-- exp: double (nullable = true)

root
 |-- id: string (nullable = true)
 |-- g: long (nullable = true)
 |-- gs: long (nullable = true)
 |-- mp: double (nullable = true)
 |-- fgperc: double (nullable = true)
 |-- 3pperc: double (nullable = true)
 |-- 2pperc: double (nullable = true)
 |-- efgperc: double (nullable = true)
 |-- ftperc: double (nullable = true)
 |-- orb: double (nullable = true)
 |-- drb: double (nullable = true)
 |-- ast: double (nullable = true)
 |-- stl: double (nullable = true)
 |-- blk: double (nullable = true)
 |-- tov: double (nullable = true)
 |-- pf: double (nullable = true)
 |-- pts: double (nullable = true)

root
 |-- id: string (nullable = true)
 |-- g: long (nullable = true)
 |-- gs: long (nullable = true)
 |-- mp: double (nullable = true)
 |-- fgperc: double (nullable = true)
 |-- 3pperc: double (nullable = true)
 |-- 2pperc: double (nullable = true)
 |-- efgperc: double (nullable = true)
 |-- ftperc: double (nullable = true)
 |-- orb: double (nullable = true)
 |-- drb: double (nullable = true)
 |-- ast: double (nullable = true)
 |-- stl: double (nullable = true)
 |-- blk: double (nullable = true)
 |-- tov: double (nullable = true)
 |-- pf: double (nullable = true)
 |-- pts: double (nullable = true)
```

In [102]:

```
# convert the data frame into parquet format
teams_t_df.write.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/teams_t.parquet", mode = 'overwrite')
players_t_df.write.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/players_t.parquet", mode = 'overwrite')
current_season_result_df.write.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/current_season_result.parquet", mode = 'overw
current_career_result_df.write.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/current_career_result_df.parquet", mode = 'over
```

6.0 Write into the Database

The schema is built into the PostgreSQL database using the NBA.sql file. All four parquet files generated from the data transformation step are loaded and written into the database using spark.

In [103]:

```
!PGPASSWORD=qwerty123 psql -h depgdb.crhso94tou3n.eu-west-2.rds.amazonaws.com -d haiyunzou21 -U haiyunzou21 -c '\i /project/MSIN0166_Da
```

```
  .:::.     .::.
....yy:    .yy.
:.  .yy.    y.
     :y:   .:
     .yy  .:
      yy..:
      :y:.
      .y.
     .:.
.....:.
::::.
```

- Project files and data should be stored in /project. This is shared among everyone

in the project.
• Personal files and configuration should be stored in /home/faculty.
• Files outside /project and /home/faculty will be lost when this server is terminated.
• Create custom environments to setup your servers reproducibly.

```
psql:/project/MSIN0166_Data_Engineering_individual/NBA.sql:1: NOTICE:  drop cascades to 4 other objects
DETAIL:  drop cascades to table nba.teams
drop cascades to table nba.players
drop cascades to table nba.current_season_result
drop cascades to table nba.current_career_result
DROP SCHEMA
CREATE SCHEMA
psql:/project/MSIN0166_Data_Engineering_individual/NBA.sql:5: NOTICE:  table "teams" does not exist, skipping
DROP TABLE
psql:/project/MSIN0166_Data_Engineering_individual/NBA.sql:6: NOTICE:  table "players" does not exist, skipping
DROP TABLE
psql:/project/MSIN0166_Data_Engineering_individual/NBA.sql:7: NOTICE:  table "current_season_result" does not exist, skipping
DROP TABLE
psql:/project/MSIN0166_Data_Engineering_individual/NBA.sql:8: NOTICE:  table "current_career_result" does not exist, skipping
DROP TABLE
CREATE TABLE
CREATE TABLE
CREATE TABLE
CREATE TABLE
```

In [104]:
```
# load the data
teams_t_df = spark.read.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/teams_t.parquet")
players_t_df = spark.read.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/players_t.parquet")
current_season_result_df = spark.read.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/current_season_result.parquet")
current_career_result_df = spark.read.parquet("/project/MSIN0166_Data_Engineering_individual/parquet_files/current_career_result_df.parquet")
```

In [105]:
```
# information for log into postgresql
postgres_uri = "jdbc:postgresql://depgdb.crhso94tou3n.eu-west-2.rds.amazonaws.com:5432/haiyunzou21"
user = "haiyunzou21"
password = "qwerty123"
```

In [106]:
```
# write the data into the database
teams_t_df.write.jdbc(url=postgres_uri, table="NBA.teams", mode="append", properties={"user":user, "password": password, "driver": "org.postgresql.
players_t_df.write.jdbc(url=postgres_uri, table="NBA.players", mode="append", properties={"user":user, "password": password, "driver": "org.postgres
current_season_result_df.write.jdbc(url=postgres_uri, table="NBA.current_season_result", mode="append", properties={"user":user, "password": passw
current_career_result_df.write.jdbc(url=postgres_uri, table="NBA.current_career_result", mode="append", properties={"user":user, "password": passwo
```

6.1 SQL Query

The SQL query is to determine the players that have the highest guaranteed and their number of games played, experience and age.

The query results show that Stephen Curry and Kevin Durant are the two players has the highest guarantee.

This query can also be used for teams to reach out to the players with high potential and the needed position in the team.

In [107]:
```
sql1 = '''
SELECT name, possition, g, exp, age, guaranteed
FROM NBA.players as p
JOIN NBA.current_career_result as c
ON p.id = c.id
GROUP BY possition, guaranteed, g, name, exp, age
ORDER By guaranteed DESC
'''
```

In [108]:
```
sql1_df = spark.read \
    .format("jdbc") \
    .option("url", postgres_uri) \
    .option("query", sql1) \
    .option("user", user) \
    .option("password", password) \
    .option("driver", "org.postgresql.Driver") \
    .load()

sql1_df.printSchema()
```

```
root
 |-- name: string (nullable = true)
 |-- possition: string (nullable = true)
 |-- g: double (nullable = true)
 |-- exp: double (nullable = true)
 |-- age: integer (nullable = true)
 |-- guaranteed: integer (nullable = true)
```

In [109]:
```
sql1_df = sql1_df.toPandas()
```

In [110]:
```
sql1_df.head(10)
```

Out[110]:

	name	possition	g	exp	age	guaranteed
0	Stephen Curry	PG	826.0	12.0	34	261134628
1	Kevin Durant	PF	939.0	13.0	34	229997220
2	Luka Dončić	PG	264.0	3.0	23	217234391
3	Joel Embiid	C	328.0	5.0	28	206889460
4	Trae Young	PG	280.0	3.0	24	180876471
5	Shai Gilgeous-Alexander	PG	243.0	3.0	24	178045532
6	Giannis Antetokounmpo	PF	656.0	8.0	28	176265466
7	Damian Lillard	PG	711.0	9.0	32	176265152
8	Jimmy Butler	SF	690.0	10.0	33	167652137
9	Bam Adebayo	C	343.0	4.0	25	163000590

7.0 Conclusion and limitation

In conclusion, this project was carried out with the process of data mining, data cleaning, data preparation, feature engineering, machine learning model, data transformation, building schema and database interaction. Meanwhile, the code version control and data version control and automation are carried out using GitHub, DVC, and Terraform respectively.

In terms of the ML part, the final accuracy for the Decision Tree is 0.53, this result is not ideal but a possible way of fixing or improving this accuracy is to add more in the feature engineering part, where there could have a passion score for players at different positions. Since the position played in NBA games may affect their data in different aspects.

In future models, the actively moving data is also a choice of additional supporting data for example, on which part of the basketball court the players at the most.

8.0 Reference

Atlassian, (2021). What is version control? Available from: https://www.atlassian.com/git/tutorials/what-is-version-control (Accessed: 25 April 2022)

Jaadi, Zakaria, (2021). A Step-by-Step Explanation of Principal Component Analysis (PCA). Available from: https://builtin.com/data-science/step-step-explanation-principal-component-analysis (Accessed: 25 April 2022)

Jeffries, Dan, (2020). What Exactly Is Data Lineage. Available from: https://www.pachyderm.com/blog/what-exactly-is-data-lineage/ (Accessed: 25 April 2022)

Wikipedia, (2022). List of National Basketball Association seasons. Available from: https://en.wikipedia.org/wiki/List_of_National_Basketball_Association_seasons#:~:text=Each%20team%20plays%2082%20games,conferences%2((Accessed: 25 April 2022)

Wikipedia, (2022). National Basketball Association Available from: https://en.wikipedia.org/wiki/National_Basketball_Association (Accessed: 25 April 2022)