HOA 1.1: Setting up your Big Data Environment using PySpark

Procedure

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_
```

```
from pyspark.sql import SparkSession
my_spark = SparkSession.builder.appName("my_spark").getOrCreate()
print(my_spark)
```

```
<pyspark.sql.session.SparkSession object at 0x79707c59c790>
```

```
path = "/content/username.csv"

username_df = my_spark.read.csv(path,
                                header = True,
                                inferSchema = True)
username_df.show()
```

```
+---------------+----------+----------+---------+
|       Username|Identifier|First Name|Last Name|
+---------------+----------+----------+---------+
|  graylooker123|      1002|    Robert|  Jenkins|
|    ironlord982|      1003|     Linux| Cromwell|
|  skyrimlord023|      1004| Cassandra|    Ramos|
|wannabefortnite|      1005|    Joseph|   Pandas|
|    smithSummer|      1006|    Tensor|     Flow|
+---------------+----------+----------+---------+
```

```
username_df.count()
```

```
5
```

```
df_filtered = username_df.filter(username_df["Identifier"] > 1002)
df_filtered.show()
```

```
+---------------+----------+----------+---------+
|       Username|Identifier|First Name|Last Name|
+---------------+----------+----------+---------+
|    ironlord982|      1003|     Linux| Cromwell|
|  skyrimlord023|      1004| Cassandra|    Ramos|
|wannabefortnite|      1005|    Joseph|   Pandas|
|    smithSummer|      1006|    Tensor|     Flow|
+---------------+----------+----------+---------+
```

```
from pyspark.sql.functions import avg
username_df.groupBy("Username").agg(avg("Identifier")).show()
```

```
+---------------+---------------+
|       Username|avg(Identifier)|
+---------------+---------------+
|wannabefortnite|         1005.0|
|  skyrimlord023|         1004.0|
|    ironlord982|         1003.0|
|  graylooker123|         1002.0|
|    smithSummer|         1006.0|
+---------------+---------------+
```

Supplementary Activity

1. Create a DataFrame

```
from datetime import datetime, date
import pandas as pd
from pyspark.sql import Row

starting_df = my_spark.createDataFrame([
    (1, 2., 'string1', date(2025, 1, 1), datetime(2004, 1, 1, 12, 0)),
    (2, 3., 'string2', date(2026, 2, 1), datetime(2005, 1, 2, 9, 0)),
    (3, 4., 'string3', date(2027, 3, 1), datetime(2006, 1, 3, 8, 0))
], schema='a long, b double, c string, d date, e timestamp')

starting_df.show()
```

```
+---+---+-------+----------+-------------------+
|  a|  b|      c|         d|                  e|
+---+---+-------+----------+-------------------+
|  1|2.0|string1|2025-01-01|2004-01-01 12:00:00|
|  2|3.0|string2|2026-02-01|2005-01-02 09:00:00|
|  3|4.0|string3|2027-03-01|2006-01-03 08:00:00|
+---+---+-------+----------+-------------------+
```

3. Load the Employee_salaries.csv

```
path = "/content/Employee_Salaries.csv"

employees_df = my_spark.read.csv(path,
                                header = True,
                                inferSchema = True)
employees_df.show(25)
```

```
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+
|Department|     Department_Name|            Division|Gender|Base_Salary|Overtime_Pay|Longevity_Pay|Grade|
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   175873.0|         0.0|          0.0|   M2|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|  145613.36|         0.0|          0.0|   M3|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|   136970.0|         0.0|          0.0|   M3|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|  89432.694|         0.0|       2490.0|   21|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|    78947.0|      456.68|       6257.7|   16|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|    98228.0|       518.8|       998.28|   21|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F| 82405.3864|       549.2|          0.0|   18|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|    93986.0|     1187.06|      2452.94|  N20|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|  149464.15|         0.0|      9021.82|   18|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   117424.0|         0.0|          0.0|  N25|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|   82808.28|    11870.82|          0.0|   21|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M| 65961.8438|      2092.7|          0.0|   13|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   59288.86|     1013.01|          0.0|   13|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|  139407.15|         0.0|          0.0|   M3|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|   128531.0|         0.0|          0.0|  N27|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|  152632.07|         0.0|          0.0|   M3|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F| 73955.2951|     3509.43|          0.0|   16|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   117424.0|         0.0|          0.0|  N25|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   117424.0|         0.0|          0.0|  N25|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F| 110572.155|         0.0|          0.0|  N26|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M|    61240.0|     5294.68|       995.18|   10|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M|    61240.0|       92.35|       4827.4|   10|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M|    86103.0|    23917.44|      3665.64|  N18|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M|    61240.0|     9580.78|        459.3|   10|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M|    86103.0|     3165.46|      3665.65|  N18|
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+
only showing top 25 rows
```

4. Filter the Employees based on their gender

```
from pyspark.sql.functions import count

employees_df.groupBy("Gender").agg(count("Gender").alias("count")).show()
```

```
+------+-----+
|Gender|count|
+------+-----+
|     F| 4362|
|     M| 5929|
+------+-----+
```

```
# For Gender == F
employees_F = employees_df.filter(employees_df.Gender == "F")
employees_F.show()
```

```
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+
|Department|     Department_Name|            Division|Gender|Base_Salary|Overtime_Pay|Longevity_Pay|Grade|
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|   136970.0|         0.0|          0.0|   M3|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|  89432.694|         0.0|       2490.0|   21|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|    78947.0|      456.68|       6257.7|   16|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|    98228.0|       518.8|       998.28|   21|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F| 82405.3864|       549.2|          0.0|   18|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|  149464.15|         0.0|      9021.82|   18|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|   82808.28|    11870.82|          0.0|   21|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|   128531.0|         0.0|          0.0|  N27|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F| 73955.2951|     3509.43|          0.0|   16|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     F| 110572.155|         0.0|          0.0|  N26|
|       ABS|Alcohol Beverage ...| ABS 85 Beer Loading|     F|    76668.0|     3226.55|          0.0|   12|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Wareh...|     F|  56922.265|     6156.95|          0.0|   13|
|       ABS|Alcohol Beverage ...|ABS 85 Licensure,...|     F|  166140.03|         0.0|          0.0|   M2|
|       ABS|Alcohol Beverage ...|ABS 85 Licensure,...|     F| 99335.0745|     1540.34|          0.0|   22|
|       ABS|Alcohol Beverage ...|ABS 85 Licensure,...|     F|  94334.247|         0.0|          0.0|  N25|
|       ABS|Alcohol Beverage ...|ABS 85 Licensure,...|     F|    98349.0|         0.0|          0.0|   22|
|       ABS|Alcohol Beverage ...|ABS 85 Licensure,...|     F|    64633.0|         0.0|          0.0|   16|
|       ABS|Alcohol Beverage ...|ABS 85 Licensure,...|     F|    54583.0|         0.0|          0.0|   18|
|       ABS|Alcohol Beverage ...| ABS 85 Burtonsville|     F| 47447.9101|     1302.22|          0.0|   12|
|       ABS|Alcohol Beverage ...| ABS 85 Burtonsville|     F| 58239.9985|      536.86|          0.0|   14|
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+
only showing top 20 rows
```

```
# For Gender == M
employees_M = employees_df.filter(employees_df.Gender == "M")
employees_M.show()
```

```
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+
|Department|     Department_Name|            Division|Gender|Base_Salary|Overtime_Pay|Longevity_Pay|Grade|
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   175873.0|         0.0|          0.0|   M2|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|  145613.36|         0.0|          0.0|   M3|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|    93986.0|     1187.06|      2452.94|  N20|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   117424.0|         0.0|          0.0|  N25|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M| 65961.8438|      2092.7|          0.0|   13|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   59288.86|     1013.01|          0.0|   13|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|  139407.15|         0.0|          0.0|   M3|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|  152632.07|         0.0|          0.0|   M3|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   117424.0|         0.0|          0.0|  N25|
|       ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   117424.0|         0.0|          0.0|  N25|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M|    61240.0|     5294.68|       995.18|   10|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M|    61240.0|       92.35|       4827.4|   10|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M|    86103.0|    23917.44|      3665.64|  N18|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M|    61240.0|     9580.78|        459.3|   10|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M|    86103.0|     3165.46|      3665.65|  N18|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M| 75086.4937|     6857.37|          0.0|   15|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M| 81931.2453|    24672.86|          0.0|  N18|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M|    75621.0|     2065.26|      3471.32|   15|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M|  75621.002|     4453.91|      3572.85|   15|
|       ABS|Alcohol Beverage ...|ABS 85 Beer Deliv...|     M|    75621.0|    10989.68|      3436.83|   15|
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+
only showing top 20 rows
```

5. Group the Employees based on their Gender and Average their Salaries

```
from pyspark.sql.functions import avg
employees_df.groupBy("Gender").agg(avg("Base_Salary")).show()
```

```
+------+-----------------+
|Gender| avg(Base_Salary)|
+------+-----------------+
|     F|87497.50279041701|
|     M|92382.92975236966|
+------+-----------------+
```

6. Compute annual salary for each employee.

```
with_annual_salaries = employees_df.withColumn("Annual_Salary", employees_df["Base_Salary"] * 12)
with_annual_salaries.show()
```

```
---------+--------------------+--------------------+------+-----------+------------+-------------+-----+-----------
epartment|     Department_Name|            Division|Gender|Base_Salary|Overtime_Pay|Longevity_Pay|Grade|     Annual
---------+--------------------+--------------------+------+-----------+------------+-------------+-----+-----------
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   175873.0|         0.0|          0.0|   M2|         21
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|  145613.36|         0.0|          0.0|   M3|1747360.319
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|   136970.0|         0.0|          0.0|   M3|         16
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|  89432.694|         0.0|       2490.0|   21|       1073
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|    78947.0|      456.68|       6257.7|   16|          9
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|    98228.0|       518.8|       998.28|   21|         11
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     F| 82405.3864|       549.2|          0.0|   18|       9888
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|    93986.0|     1187.06|      2452.94|  N20|         11
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|  149464.15|         0.0|      9021.82|   18|1793569.799
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   117424.0|         0.0|          0.0|  N25|         14
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|   82808.28|    11870.82|          0.0|   21|         99
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     M| 65961.8438|      2092.7|          0.0|   13|       7915
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   59288.86|     1013.01|          0.0|   13| 711466.326
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|  139407.15|         0.0|          0.0|   M3|1672885.799
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     F|   128531.0|         0.0|          0.0|  N27|         15
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|  152632.07|         0.0|          0.0|   M3|        183
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     F| 73955.2951|     3509.43|          0.0|   16| 887463.541
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   117424.0|         0.0|          0.0|  N25|         14
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     M|   117424.0|         0.0|          0.0|  N25|         14
      ABS|Alcohol Beverage ...|ABS 85 Administra...|     F| 110572.155|         0.0|          0.0|  N26|1326865.859
---------+--------------------+--------------------+------+-----------+------------+-------------+-----+-----------
ly showing top 20 rows
```

7. Sort the result and display the highest average.

```
with_annual_salaries.sort("Annual_Salary", ascending=False).show(1)
```

```
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+--------
|Department|     Department_Name|            Division|Gender|Base_Salary|Overtime_Pay|Longevity_Pay|Grade|Annual_S
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+--------
|       CEX|Offices of the Co...|CEX 15 Chief Admi...|     M|   292000.0|         0.0|          0.0|  EX0|    3504
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+--------
only showing top 1 row
```

```
with_annual_salaries.sort("Annual_Salary", ascending=False).show(10)
```

```
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+--------
|Department|     Department_Name|            Division|Gender|Base_Salary|Overtime_Pay|Longevity_Pay|Grade|Annual_S
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+--------
|       CEX|Offices of the Co...|CEX 15 Chief Admi...|     M|   292000.0|         0.0|          0.0|  EX0|    3504
|       CAT|County Attorney's...|CAT 30 County Att...|     M|   258000.0|         0.0|          0.0|  EX1|    3096
|       POL|Department of Police|POL 47 HQ Police ...|     M|   258000.0|         0.0|          0.0|  EX1|    3096
|       CCL|      County Council|CCL 01 Council Ce...|     F|  246162.47|         0.0|          0.0| NULL|   29539
|       DGS|Department of Gen...|     DGS 36 Director|     M|   246000.0|         0.0|          0.0|  EX1|    2952
|       DOT|Department of Tra...|     DOT 50 Director|     M|   244000.0|         0.0|          0.0|  EX1|    2928
|       HHS|Department of Hea...|HHS 60 Director's...|     M|   240000.0|         0.0|          0.0|  EX1|    2880
|       ABS|Alcohol Beverage ...|     ABS 85 Director|     F|   236000.0|         0.0|          0.0|  EX1|    2832
|       FIN|Department of Fin...|     FIN 32 Director|     M|   236000.0|         0.0|          0.0|  EX1|    2832
|       OMB|Office of Managem...|OMB 31 Office of ...|     F|   236000.0|         0.0|          0.0|  EX1|    2832
+----------+--------------------+--------------------+------+-----------+------------+-------------+-----+--------
only showing top 10 rows
```

```
from pyspark.sql.functions import avg
with_annual_salaries.groupBy("Gender").agg(avg("Annual_Salary")).show()
```

```
+------+------------------+
|Gender|avg(Annual_Salary)|
+------+------------------+
|     F|1049970.0334850072|
|     M|1108595.1570284364|
+------+------------------+
```

Double-click (or enter) to edit