lab1

January 15, 2024

In [10]:
```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col
spark = SparkSession.builder.appName("SquareIntegers").getOrCreate()
integers = [4, 6, 7, 8, 9]
df = spark.createDataFrame([(i,) for i in integers], ["numbers"])
squared_df = df.withColumn("squared", col("numbers") ** 2)
squared_df.show()
spark.stop()
```

```
/home/lplab/anaconda3/lib/python3.7/site-packages/pyspark/context.py:317: FutureWarning: Pytho
  warnings.warn("Python 3.7 support is deprecated in Spark 3.4.", FutureWarning)
```

```
+-------+-------+
|numbers|squared|
+-------+-------+
|      4|   16.0|
|      6|   36.0|
|      7|   49.0|
|      8|   64.0|
|      9|   81.0|
+-------+-------+
```

In [11]:
```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, max as spark_max
spark = SparkSession.builder.appName("MaxOfNumbers").getOrCreate()
numbers = [2, 4, 8, 67, 32, 79]
df = spark.createDataFrame([(i,) for i in numbers], ["numbers"])
max_number = df.agg(spark_max(col("numbers"))).collect()[0][0]
print("The maximum number is:", max_number)
spark.stop()
```

```
/home/lplab/anaconda3/lib/python3.7/site-packages/pyspark/context.py:317: FutureWarning: Pytho
  warnings.warn("Python 3.7 support is deprecated in Spark 3.4.", FutureWarning)
```

```
The maximum number is: 79
```

In [12]:
```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, avg

# Create a Spark session
spark = SparkSession.builder.appName("AverageNumbers").getOrCreate()

# List of numbers
numbers = [4, 6, 7, 8, 9]

# Create a DataFrame
df = spark.createDataFrame([(i,) for i in numbers], ["numbers"])

# Calculate the average of the numbers
average_df = df.select(avg(col("numbers")).alias("average"))

# Show the result
average_df.show()

# Stop the Spark session
spark.stop()
```

```
/home/lplab/anaconda3/lib/python3.7/site-packages/pyspark/context.py:317: FutureWarning: Pytho
  warnings.warn("Python 3.7 support is deprecated in Spark 3.4.", FutureWarning)


+-------+
|average|
+-------+
|    6.8|
+-------+
```

In [13]:
```
from pyspark.sql import SparkSession

# Create a Spark session
spark = SparkSession.builder.appName("ReadCSV").getOrCreate()

# Specify the CSV file path
csv_file_path = "lab1.csv"

# Read the CSV file into a DataFrame
df = spark.read.csv(csv_file_path, header=True, inferSchema=True)

# Show the contents of the DataFrame
df.show()

# Stop the Spark session
spark.stop()
```

```
/home/lplab/anaconda3/lib/python3.7/site-packages/pyspark/context.py:317: FutureWarning: Pytho
  warnings.warn("Python 3.7 support is deprecated in Spark 3.4.", FutureWarning)


+-------+---+----+
|   Name|sec|cgpa|
+-------+---+----+
|karthik|  h|  10|
|  rahul|  a|   9|
|   soma|  d|   8|
|abhiram|  f|   7|
|  vamsi|  g|   6|
+-------+---+----+
```

```
In [15]: from pyspark.sql import SparkSession

         # Create a Spark session
         spark = SparkSession.builder.appName("ShowDataFrame").getOrCreate()

         # Specify the CSV file path (replace with your actual file path)
         csv_file_path = "lab1.csv"

         # Read the CSV file into a DataFrame
         df = spark.read.csv(csv_file_path, header=True, inferSchema=True)

         # Show the first few rows of the DataFrame
         print("First few rows:")
         df.show(2)  # Display the first 5 rows

         # Display the schema of the DataFrame
         print("\nDataFrame Schema:")
         df.printSchema()

         # Stop the Spark session
         spark.stop()
```

```
/home/lplab/anaconda3/lib/python3.7/site-packages/pyspark/context.py:317: FutureWarning: Pytho
  warnings.warn("Python 3.7 support is deprecated in Spark 3.4.", FutureWarning)


First few rows:
+-------+---+----+
|   Name|sec|cgpa|
+-------+---+----+
|karthik|  h|  10|
|  rahul|  a|   9|
+-------+---+----+
```

```
only showing top 2 rows


DataFrame Schema:
root
 |-- Name: string (nullable = true)
 |-- sec: string (nullable = true)
 |-- cgpa: integer (nullable = true)
```

In [20]:
```
from pyspark.sql import SparkSession

        # Create a Spark session
        spark = SparkSession.builder.appName("SummaryStatistics").getOrCreate()

        # Specify the CSV file path (replace with your actual file path)
        csv_file_path = "lab1.csv"

        # Read the CSV file into a DataFrame
        df = spark.read.csv(csv_file_path, header=True, inferSchema=True)

        # Specify the column for which you want to calculate summary statistics
        selected_column = "cgpa"  # Replace with the actual column name

        # Calculate summary statistics for the specified column
        summary_statistics = df.select(selected_column).describe()

        # Show the summary statistics
        summary_statistics.show()

        # Stop the Spark session
        spark.stop()
```

```
+-------+------------------+
|summary|              cgpa|
+-------+------------------+
|  count|                 5|
|   mean|               8.0|
| stddev|1.5811388300841898|
|    min|                 6|
|    max|                10|
+-------+------------------+
```

In [ ]: