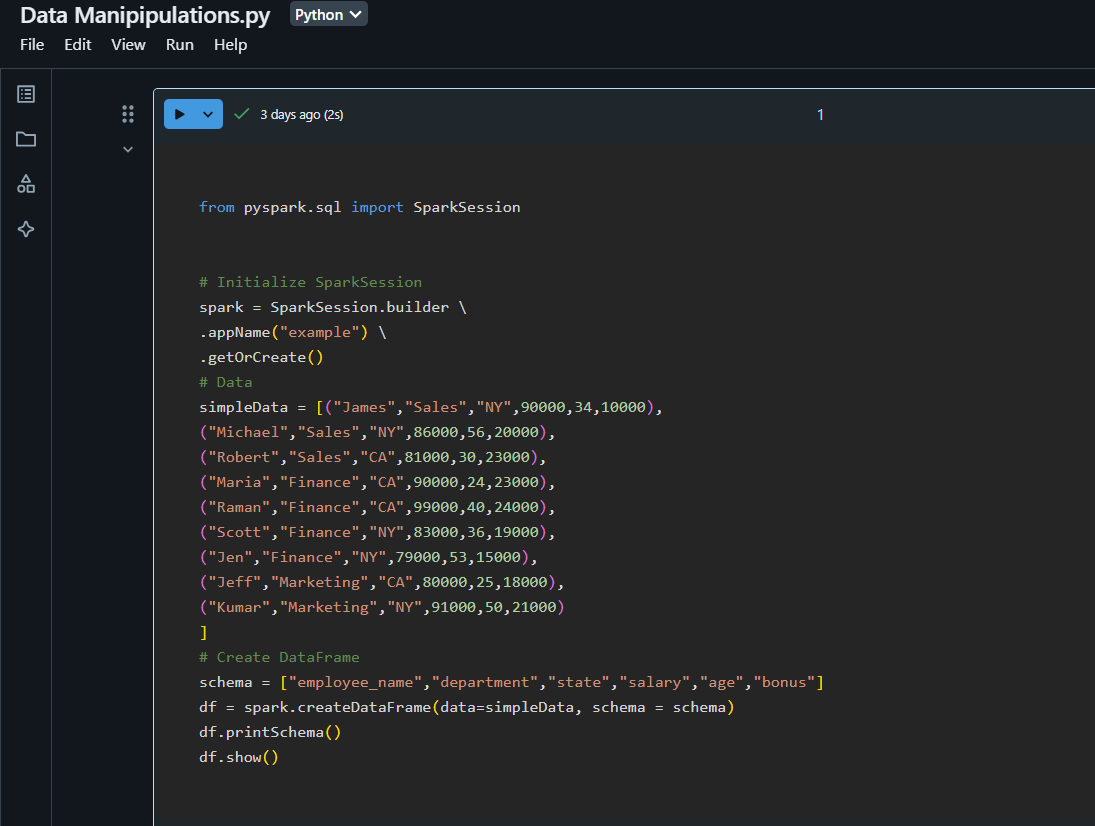
**Week 3 day 4**

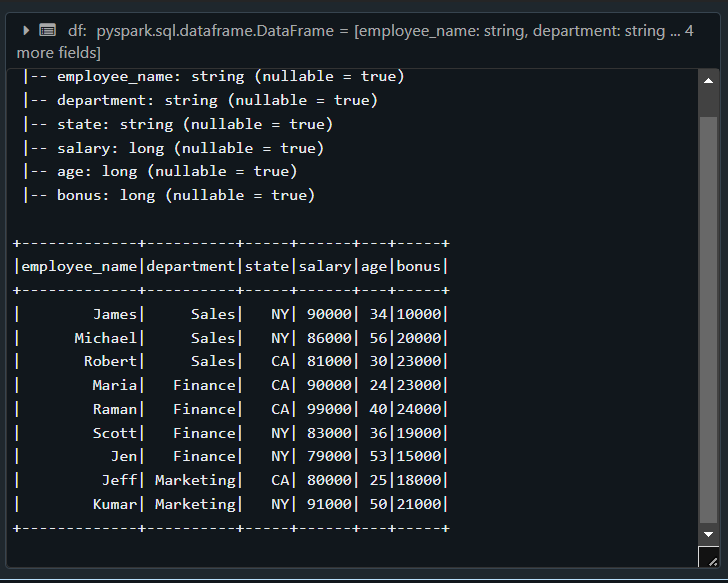
**(21-11-2024)**

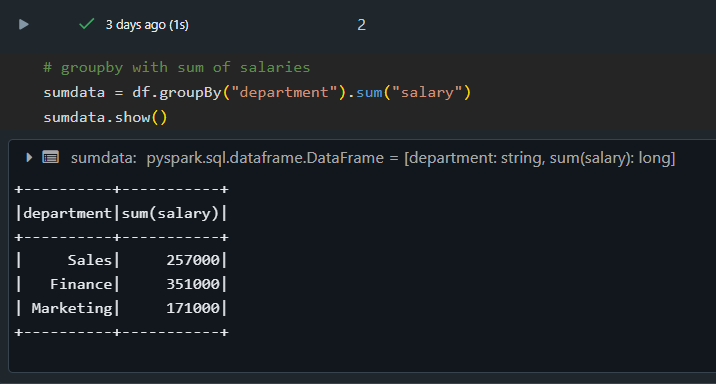
**Topics covered and a summary:**

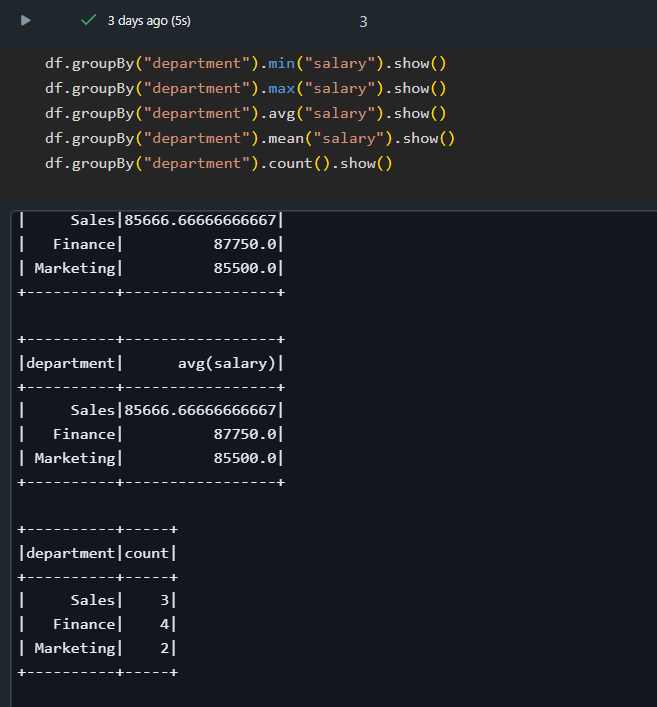
1. **Read Text File into PySpark**:
   * Use spark.read.text("path/to/textfile") to load text files into a PySpark DataFrame.
   * Alternatively, for CSV files, use spark.read.csv("path/to/csv", header=True, inferSchema=True) to load and infer schema.
2. **Manipulating**:
   * Perform transformations like withColumn (to add/modify a column), filter (to filter rows), and select (to select specific columns).
3. **Dropping**:
   * Use drop("column\_name") to remove a specific column from the DataFrame.
4. **Sorting**:
   * Use orderBy("column\_name", ascending=True/False) to sort the DataFrame by a column.
5. **Aggregations**:
   * Perform operations like groupBy("column\_name").agg() for aggregations such as sum, avg, count, etc.
6. **Joins**:
   * Use df1.join(df2, "common\_column", "join\_type") for joining two DataFrames (join\_type can be inner, left, right, outer).

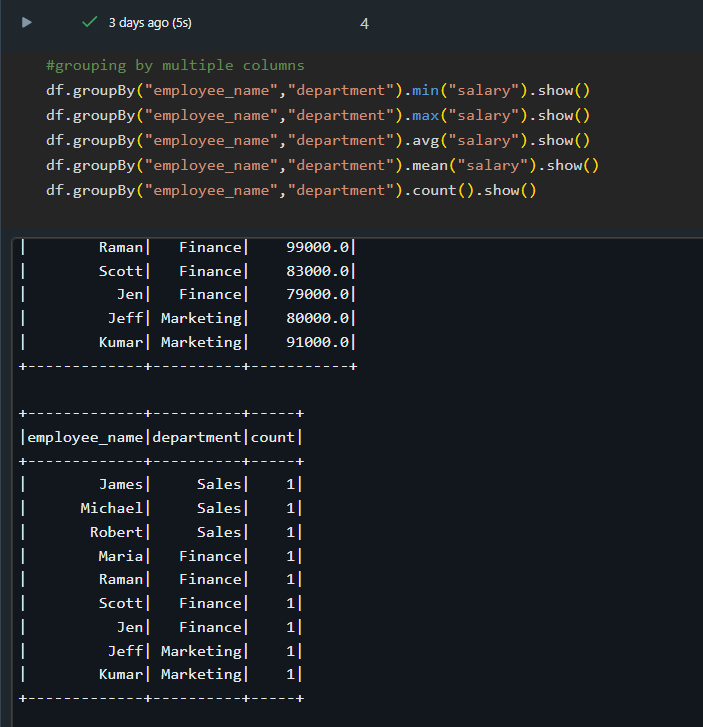
* This workflow allows for scalable and efficient handling of large datasets in PySpark.

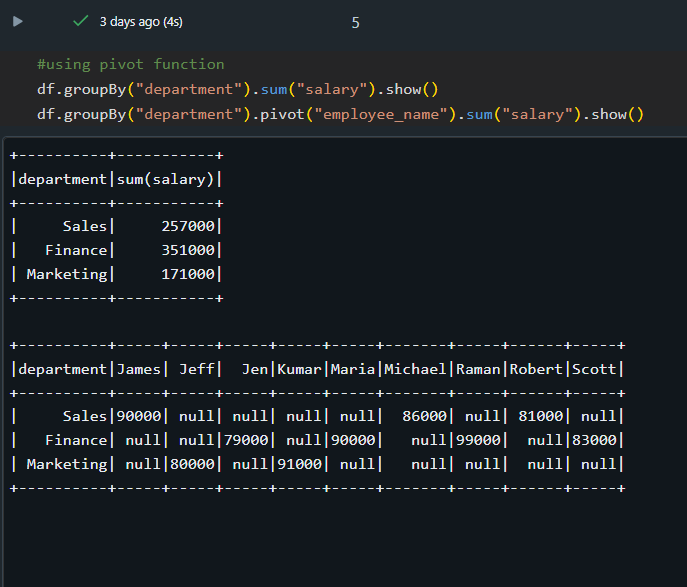
****

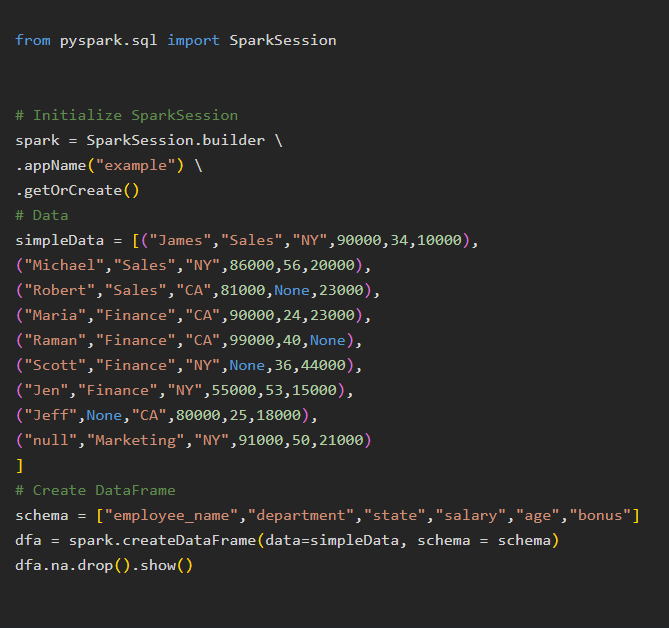
****

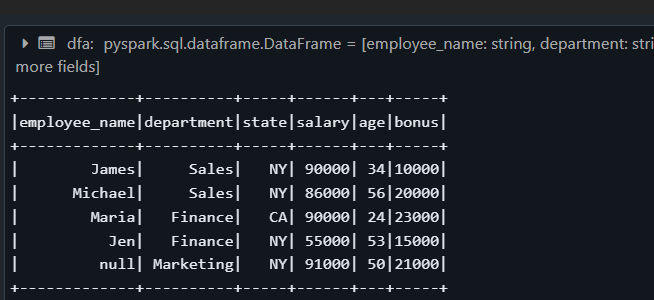
****

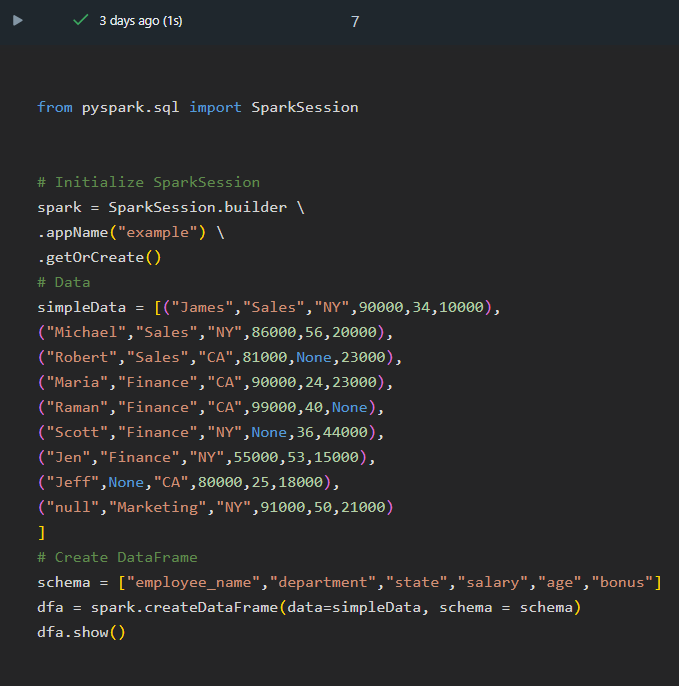
****

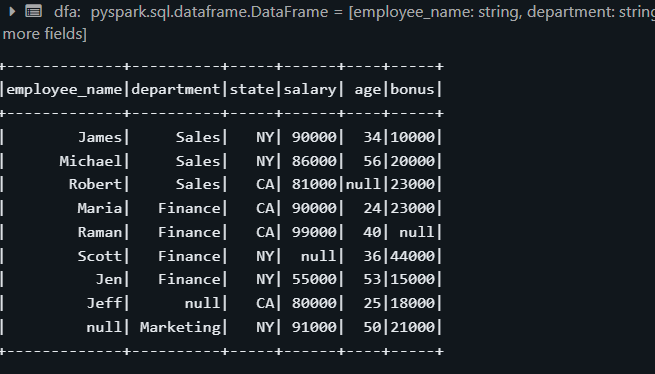
****

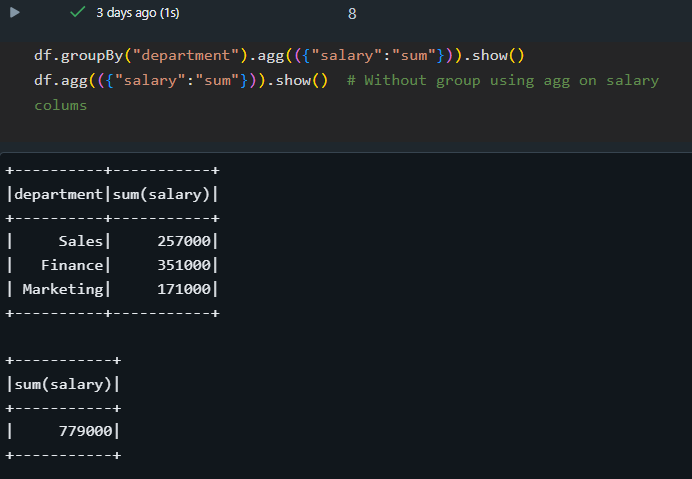
****

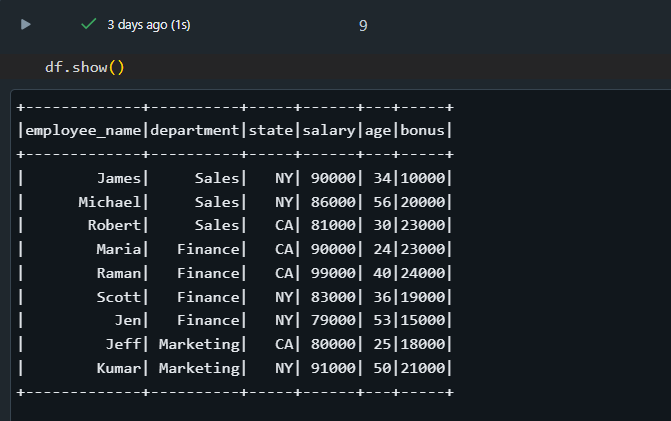
****

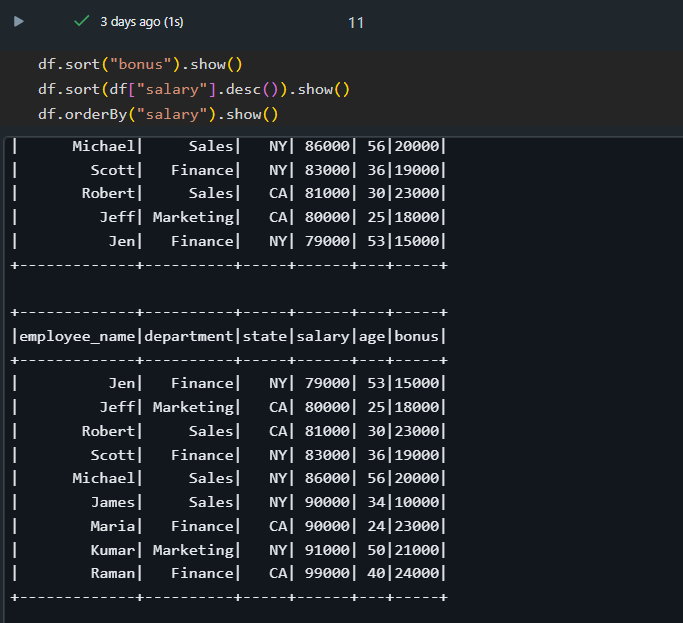
****

****

****

****

****

****