**Database Project (Fall 2023)**

**HW3 (100pts, Due date: Nov 3)**

**Student ID**: 2020315798

**Student Name**: Choi Jin Woo

**Instruction:** In this homework, we provide you with a jupyter notebook file (DBP\_HW3.ipynb). You should follow the instructions in these documents carefully.

**Submit two files as follows:**

- DBP\_HW3\_StudentID.zip

- DBP\_HW3\_StudentID.ipynb

- DBP\_HW3\_StudentID.pdf

**[Spark SQL]**

**1. [20pts]** The following data is related to the manufacturing process of computer manufacturers.

***Data:***

**- Already in the manufacturing process 🡪 used for problem (a):**[Row(phase='packing', model='book\_pro', serial=‘book\_pro1’),  
Row(phase='packing', model='book\_pro', serial=‘book\_pro2’),  
Row(phase='packing', model='plus', serial=‘plus1’),  
Row(phase='packing', model='book\_pro', serial=‘book\_pro3’),  
Row(phase='packing', model='plus', serial=‘plus2’),  
Row(phase='inspection', model='book\_pro', serial=‘book\_pro4’),  
Row(phase='inspection', model='plus', serial=‘plus3’),  
Row(phase='inspection', model='book\_pro', serial=‘book\_pro5’),  
Row(phase='inspection', model='book\_pro', serial=‘book\_pro6’)]

**- The manufacturing process to add 🡪 used for problem (b):**  
[Row(phase=‘assembly’, model=‘book\_pro’, serial=‘book\_pro7’),  
Row(phase=‘assembly’, model=‘plus’, serial=‘plus4’)]

1. Create a DataFrame with the given data and display the generated DataFrame.

**[Answer]**

Enter your code and result here. You must show your result (captured image).

|  |
| --- |
|  |

1. After adding ***two laptops to the manufacturing process***, find the number of products for each model.

**[Answer]**

Enter your code and result here. You must show your result (captured image).

|  |
| --- |
|  |

1. Group the data in the joined DataFrame by ‘phases’ and count the number of data for each phase.

**[Answer]** Enter your code and result here. You have to show your snapshot result.

|  |
| --- |
|  |

**2. [20pts]** The following data are *manufacturing process* and *customer information* for computer manufacturers.

***- Data 1(manufacturing process):***  
[Row(phase='packing', model='book\_pro', serial='book\_pro1'),  
Row(phase='packing', model='book\_pro', serial='book\_pro2'),  
Row(phase='packing', model='plus', serial='plus1'),  
Row(phase='packing', model='book\_pro', serial='book\_pro3'),  
Row(phase='packing', model='plus', serial='plus2'),  
Row(phase='inspection', model='book\_pro', serial='book\_pro4'),  
Row(phase='inspection', model='plus', serial='plus3'),  
Row(phase='inspection', model='book\_pro', serial='book\_pro5'),  
Row(phase='inspection', model='book\_pro', serial='book\_pro6')]

***- Data 2(customer information):***  
[Row(Customer='Lion', serial='book\_pro1', due\_date='2023-10-25'),  
Row(Customer='Rabbit', serial='book\_pro2', due\_date='2023-10-27'),  
Row(Customer='Rathor panda', serial='plus1', due\_date='2023-10-22'),  
Row(Customer='Fubao', serial='book\_pro3', due\_date='2023-10-28'),  
Row(Customer='Chinchilla', serial='plus2', due\_date='2023-10-22'),  
Row(Customer='Sloth', serial='book\_pro4', due\_date='2023-10-26')]

1. Create a DataFrame for the two given data and join Data 1 with Data 2 using an inner join based on the ‘serial’ column. (left side: Data 2, right side: Data 1)

**[Answer]** Enter your code and result here. You have to show your snapshot result.

|  |
| --- |
|  |

1. Use an SQL query to select the data from the joined DataFrame where the **‘due\_date’** is on or after **‘2023-10-25’**. And briefly explain the method you used.

**[Answer]** Enter your code and result here. You have to show your snapshot result.

|  |
| --- |
| [createOrReplaceTempView]  To save Dataframe into temporary SQL table named “joined\_df”  [sql\_query]  SQL query that filters where “due\_date” is on or after “2023-10-25”  [spark.sql()]  Returns new Dataframe with query result |

**[Spark ML]**

**3. [60pts]** We provide you with a ***Fashion-MNIST*** dataset.

***Dataset Description:***

**Training set:** 60,000 examples

**Test set:** 10,000 examples

Each example is a 28x28 grayscale image associated with a label from 10 classes.

|  |  |
| --- | --- |
| **Label** | **Description** |
| 0 | T-shirt/top |
| 1 | Trouser |
| 2 | Pullover |
| 3 | Dress |
| 4 | Coat |
| 5 | Sandal |
| 6 | Shirt |
| 7 | Sneaker |
| 8 | Bag |
| 9 | Ankle boot |

For more information, visit this website: <https://github.com/zalandoresearch/fashion-mnist>

1. Load the provided dataset, convert it into a DataFrame, and show it. You should follow the following instructions.

***-Instructions 1:*** *Assemble the features into a vector column and name the column “features.”*

***-Instructions 2:*** *Rename the target column to “label.”*

**[Answer]** Enter your code and result here. You have to show your snapshot result.

|  |
| --- |
|  |

1. Train models to classify the classes of the Fashion MNIST dataset and **report the results for test data**. The models used are **Logistic Regression,** **Decision Tree, and Random Forest.**

For detailed explanations of the usage of each model, please refer to the official documentation below.

- Logistic Regression: [[Link]](https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.classification.LogisticRegression.html)

- Decision Tree: [[Link]](https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.classification.DecisionTreeClassifier.html?highlight=decisiontreeclassifier)

- Random Forest: [[Link]](https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.classification.RandomForestClassifier.html)

**[Answer]** Fill in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Logistic Regression** | **Decision Tree** | **Random Forest** |
| **Test accuracy** | 0.7561 | 0.7441 | 0.7724 |

**[Answer]** Enter your code and result here. You have to show your snapshot result.

|  |
| --- |
|  |