CSCE 5300 Introduction to Big data and Data Science

ICE 7

Lesson Title: Spark
Lesson Description: Spark with RDDs (transformation and actions) and Spark with data frames and SQL Lesson Overview:

Spark is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.

In Class Exercise:

1.a. Create spark RDD from external dataset (word_list.txt).

- Display the line in Fiction.txt (which is in Word_list.txt dataset) And Count Who many time that "467" is repeated in Fiction.txt.

b. Create spark RDD from external dataset(shakespeare.txt).

- Execute transformation and action using Scala or Python and change all words to lowercase and show the last 10 lines. [Hint: use sort ()]

c. Create Spark dataframe from hotel_booking data and execute this query.

Load data from the hotel_booking.csv.

Count who need parking space(required_car_parking_space)in 2015 July.

Query who many are Reserved for Transited- party in Customer_type On august 2015.

2.a. Explain in detail of SparkSql, Data frames & Resilient Distributed Datasets?

- State the difference between resilient distributed dataset, dataframe and dataset. Explain when to use each one of these with proper real-life examples.

b. Assume you have two data frames, df1 and df2, with the following columns in each:

df1 ==> StudentId, StudentFirstName, and StudentLastName

df2 == > StudentId, StreetName, City, ZipCode

When selecting student Id, first name, last name, street name, city, zip code after merging both data frames based on the key, i.e., id, you receive an error ambiguous column id. What solution would you propose?

3. From the Hotel Bookings dataset:

a. Count the distinct number of hotels.

b. Count the number of adults with children who reserved in City Hotel in August 2015.

c. Query how many canceled reservations in year 2017.

d. Query how many booked the Type c room of Reserved_room_type category.

e. Count the Number of assigned _room_type for A.

f. Query how many are booked Direct in market_segment.

ICE Submission Guidelines

1. ICE Submission is individual.
2. ICE code must be properly commented.
3. The documentation should include the screenshots of your code/queries and results.
4. Provide the explanation of the exercise for each question as per your understanding.
5. The similarity score for your document should be less than 15%.
6. Submit the source code (if any) properly commented and documentation (.pdf/.doc) with explanation and screenshot of source code/queries having input logic and output results.
7. Submission after the deadline is considered as late submission.

References: https://spark.apache.org/docs/latest/rdd-programming-guide.html#rdd-operations https://spark.apache.org/docs/2.2.0/sql-programming-guide.html https://sparkbyexamples.com/pyspark-rdd https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/