Course Code: 1501413
Course Title: Big Data Management and Analytics

Assignment

General Instructions:

- ***a)*** *Students who Copy and Paste from Internet sites will receive ZERO marks.*
- ***b)*** *This assignment is an individual work. You are not permitted to work with anyone else on this assignment. All work submitted must be yours and yours alone.*
- ***c)*** *Academic Integrity: You will get an automatic ZERO for the course if you violate the academic integrity policy (No copying allowed).*
- ***d)*** *The last date to submit via LMS is:* ***April 11, 2025***
- ***e)*** *Attach the code & output as a screenshot, if it is a programming question.*
- ***f)*** *Your submission should be a single PDF report.*
- ***g)*** *No late submissions will be accepted.*
- ***h)*** *Maximum Marks : 10 (2 Marks for each question)*

1. **Module-2 Hadoop:**
   To prepare your development environment for this question you must first install and setup Java and Hadoop (refer to Lab tutorials.) Design a Hadoop Mapreduce program to count the number of distinct words in a given file. Use the attached data file as input for this question.

2. **Module-4 NB:**
   You are given the task of predicting whether a given SMS is Spam or not using Naïve Bayesian Classifier Model. For this, download the "SMS Spam Collection" dataset from UCI portal (https://archive.ics.uci.edu/dataset/228/sms+spam+collection.) Implement the classifier on PySpark MLib. Test your model with 20% of the dataset for testing and 80% for training. Display the accuracy.

3. **Module-6 APriori:**
   Consider the following transaction table depicting Crime data:

Crime ID	Crime Details
1	Gun, Murder
2	Knife, Robbery, Aged
3	Murder, Night, Knife
4	Theft, Youth, Bike
5	Youth, Robbery, Day, Outskirts
6	Day, Murder
7	Robbery, Jewels
8	Theft, Bike, Youth, Night

Using APriori algorithm find the frequent itemsets and the association rules. Assume Min. Support: 2 and Confidence: 50%. Show all the steps. What happens if you increase the min. support to 4?

4. **Module-7 Apache Kafka:**
   Design a complete solution with all the steps to produce and consume event details and updates in real time for the MFU activities. You can assume a specific event such as ADT OpenSpace 2025. Use Apache Kafka as your main tool to implement this task.

5. **Module-9 Data Warehouse:**
   Build a Data Warehouse for the following scenario:
   Assume a data warehouse called “HealthConnect” which integrates data from millions of patient records across all their facilities. With "HealthConnect", it's possible to achieve the following:
   - Reduce hospital stays by identifying at-risk patients earlier
   - Improve chronic disease management through better tracking and intervention
   - Enhance research capabilities, leading to improved treatment protocols
   - Streamline operations and reduce administrative costs

   You can use Star schema for the Dimensional modelling. Implement the "HealthConnect" with the help of MS-SQL Server and SSMS.