Assignment 1 - IN5040

General Information

This assignment is the first of two compulsory assignments. Both assignments for IN5040 are to be completed individually. All students are required to have passed both assignments as a prerequisite for taking the exam.

The due dates for each of the assignment are:

- Assignment 1: 6 October 2022. 11:59 AM
- Assignment 2: 27 October 2022. 11:59 AM

Delivery instructions - Assignment 1:

Where:	Devilry
Who:	Alone
Due:	Hard Deadline: 6 October 2022 – 11:59 AM
Deliverable:	Pdf containing: 1. Query and results for questions:1-3-4-6-7-8. You may include any assumptions you have made. 2. Query and results for question 2 Short written answer in response to the two questions. 3. Short written answer for question 5
Questions:	Espen Volnes (espenvol@ifi.uio.no)

Overview

The following assignment simulates a meteorological institute which receives summarized daily weather recordings from two weather stations over a period of 5 years. You will be using Esper, a software for Complex event processing (CEP) and Stream processing, to gain an understanding of some of the basic techniques used for analyzing stream data.

Please note that the provided Java code transmits the two streams into Esper at a high rate, otherwise the query results would take five years to produce. The rate at which the streams are generated is controlled by a Thread.sleep instruction. This instruction is executed by each thread in between tuple transmission. This instruction is especially important in the last query!

Getting started

Download the assignment from the course page and unzip it. The following files are discussed in detail:

- Makefile - A simple makefile with three commands:
  - $ make all
    - Cleans the java class files
    - Compiles the java files «WeatherTuple» and «Assignment1»

- $ make run‘X’
  - Runs the java program, where ‘X’ is the query number
- $ make clean
  - Cleans the java class files

- *san_francisco.csv* and *jfk.csv*: Csv files containing the daily weather recordings from JFK and San Francisco airports over the period 1 January 2014 - 31 December 2018.
- Assignment1.java: Code for generating weather tuples and reading queries.
- WeatherTuple.java: Represents the structure of a weather tuple and contains the following attributes:
  - timestamp
  - weatherStation
  - stationName
  - averageTemperature
  - minimumTemperature
  - maximumTemperature
  - averageWindSpeed
  - precipitation
  - weather
- query_1.epl - query_8.epl: The files in which you are to write your queries.

Query Structure:

SELECT * FROM jfk

The above query selects all tuples from the jfk.csv file. Notice that queries do not end with a semicolon as is standard in many SQL languages.

Similarly if we were to select all tuples from San Francisco, we would use the following query:

SELECT *

FROM san_francisco

Questions

Part 1 - Warm-up

Question 1.

We are interested in finding the dates for which the average temperature at JFK is above 85 degrees Fahrenheit. We would also like to know what the minimum and maximum temperatures for those days are.

Write a query which will output the date, average temperature, minimum and maximum temperatures for days which are above 85 degrees Fahrenheit.

Answer: 3 tuples

Part 2 - Window Queries

Question 2.

This question requires that you write two separate queries, one using a tuple-based sliding window and the other using a tuple-based tumbling window.

Write a query which returns any seven-day windows for which the average temperature is above 82 degrees fahrenheit.

Your results should contain the following attributes:

- Start_date
- End_date
- Temp

Which query produced the highest average temperature? Explain why?

Question 3.

We would like to identify any weeks in which three or more inches of precipitation are recorded at San Francisco airport. We require that you make use of the external timestamp for this question (the timestamp attribute in each tuple).

A week is defined as a seven day period, where week 1 contains the first 7 tuples in a stream (1 - 7), week 2 the next 7 tuples (8 -14)...

Create a query which identifies and returns any weeks which fulfill the criteria described above.

Your results should contain the following attributes:

- Week_start
- Week_end
- Precipitation

Answer: 5 tuples

Question 4

We wish to identify any three-day periods in San Francisco, which include both rain and an average wind speed exceeding 19 miles per hour.
The query does not require that it rains every day within the three-day period. It only requires that rainfall is recorded at some point during the three days.

Write a query which outputs any three days in a row which match the requirements described above.

Your results should contain the following attributes:

- Start_date
- End_date
- Wind
- Precipitation

Answer: 7 tuples

Part 3 - CEP

Question 5

Let us consider the following stream sequence:
A1 C1 A2 B1 D1 A3 B2 C2 B3 C3 C4 A4 B4 List the events which match on the following patterns:

- EveryA->B
- A->EveryB
- Every A -> Every B
- Every(A -> B)

Question 6

We would like to be able to identify an event in which the difference in average temperature between two days is 40 degrees or more. Furthermore, require we that the days are not more than seven days apart.

Write a query which identifies the described pattern for JFK.

Your results should contain the following attributes:

- First_date
- Second_date
- Temp1
- Temp2
- Temp_difference

Answer: 1 tuple

Question 7

We are interested in identifying a pattern for three consecutive days in a row. The pattern we wish to identify is as follows:

- From day 1 to day 2 we observe -> a 5 miles per hour increase in wind speed, and an increase in precipitation.
- From day 2 to day 3 we observe -> a 5 miles per hour increase in wind speed, and an increase in precipitation.

Your results should contain the following attributes:

- Start_date
- Stop_date
- Wind_1
- Wind_2
- Wind_3

Answer: 1 tuple

Part 4 - Joins

Question 8

We are interested in identifying any weeks in which the average temperature at San Francisco airport is more than 35 degrees warmer than at JFK. In addition to this we require that it has rained at JFK and that there has been no rain at San Francisco airport.

A week is defined as a seven day period, where the week 1 contains the first 7 tuples in a stream (1 - 7), week 2 the next 7 tuples (8 -14)...

It is important that you do not remove the Thread.sleep() in the java code while running this query.

Your results should contain the following attributes:

- Week_start
- Week_end
- Temperature_difference

Good luck :)