Streaming Data Analytics

Preview of the part of the exam about Streaming Data Engineering

Emanuele Della Valle
Politecnico di Milano

POLITECNICO
MILANO 1863

Exam content

- **Questions** on all lectures about Streaming Data Engineering **to test**
  - **The breadth** of your knowledge
  - **The depth** of your knowledge

(8 points)

- **Exercises**
  - Given a sequence of events and an **EPL pattern** using the followed-by operator (**->**) and pattern **guards**, tell **which** events the pattern matches, which it does not, and **why**
  - Given a realistic **Streaming Data Engineering problem**, **define** the **schema** of the streams, **provide** a minimal example of the **events** in those streams, **propose continuous queries** and **show** the **results** you expect from them**.** Formulate the solution **in EPL** and **in a language of your choice between** KSQL and Spark Structured Streaming. **Explain** your code in detail.

(7 points)

Questions to test the breadth of your knowledge

- Which are the three approaches to tame velocity? List them all before comparing and contrasting two of them
- Which are the three time-models we introduced? List them all before comparing and contrasting two of them
- Which are the types of windows? List them all and describe the behavior of at least two of them
- Which are the types of operations in spark? Illustrate them with a run-time example.
- Illustrate the programming model of spark structured streaming at the logical and physical level
- Describe Kafka at a conceptual, logical, system, and physical level
- How does ksqlDB work?
- Which types of joins are supported between two streams? Why?
- How does Spark Structured Stream treat late arrivals?

Questions to test the depth of your knowledge

- What's the difference between the stream-only and the absolute time models? Explain it, proposing examples of queries that can be answered in both models and queries that can be answered only with the absolute time model.
- What's the difference between the absolute and the interval-based time models? Explain it, proposing examples of queries that can be answered in both models and queries that can be answered only with the interval-based time model.
- What are streams and events? How do they relate to each other? Give an example.
- What is a sliding logical window? Give an example at the conceptual level and show that you know both the EPL syntax and that of another language of your choice.

Questions to test the depth of your knowledge

- What is a tumbling logical window? Give an example at the conceptual level and show that you know both the EPL syntax and that of another language of your choice.
- What is a tumbling tubling window? Give an example at the conceptual level and show that you know both the EPL syntax and that of another language of your choice.
- What is the role of the output clause in EPL? Give an example that supports your explanation.
- What's the difference between transformations and actions in spark?
- What's the difference between narrow and wide transformations in spark?
- What's the role of watermarking in Spark Structured Streaming? Support your claims with an example.

Questions to test the depth of your knowledge

- What’s the role of a topic in Kafka?
- What’s the role of a broker in Kafka?
- What’s the role of a partition in Kafka?
- What’s the relationship between a consumer, a topic, a broker, a partition, a consumer and a consumer group in Kafka? Explain it using an example.
- What’s the difference among a Kstream, a Ktable and a topic? Give an example.
- What is a session window? Give an example at the conceptual level and show that you know the ksqlDB syntax
- What’s the difference between a pull and a push query in KSQLDB?

Exercises on EPL patterns

- Suppose you receive the following stream of events:

```
A1@0,C1@1,B1@2,B2@3,A2@4,B3@5,A3@6,B4@10.
```

- Note that `A3@6` denotes an event of type A identified by the number 3 that is received at time 6.
- Given the patter:

```
every  A  -> (B and not C where timer:within(3 sec))
```

- Translate the pattern into an English sentence
- Which are the events that trigger the matching? Why?
- Which are the events that may trigger the matching but are excluded by the semantics of the `every` and the `where timer:within` clauses? Why?

Exercises on a Streaming Data Engineering problem

- In the git repo of the course you find a complete example about a Robotic Arm solved in
  - EPL
    - https://github.com/emanueledellavalle/streaming-data-analytics/tree/main/codes/epl_robotic-arm
  - Spark Structured Streaming
    - https://github.com/emanueledellavalle/streaming-data-analytics/tree/main/codes/sss_robotic-arm
  - ksqlDB
    - https://github.com/emanueledellavalle/streaming-data-analytics/tree/main/codes/ksql_robotic-arm
- There is also another example completely solved in EPL about drones picking tomatoes
  - https://github.com/emanueledellavalle/streaming-data-analytics/tree/main/codes/epl_tomatopick

Streaming Data Analytics

Preview of the part of the exam about Streaming Data Engineering

Emanuele Della Valle
Politecnico di Milano

POLITECNICO
MILANO 1863