Spark & Kafka

Summary Exercise

Contents

Exercise

Dimensions Creation

Car Models

EX 1

PySpark → S3 (car_models)   Cars Models

Object Name: ModelCreation

Output location: s3a://spark/data/dims/car_models

Data:

model_id	car_brand	car_model
1	Mazda	3
2	Mazda	6
3	Toyota	Corolla
4	Hyundai	i20
5	Kia	Sportage
6	Kia	Rio
7	Kia	Picanto

Car colors

EX 2

PySpark → S3 (car_colors)   Cars Colors

Object Name: ModelCreation

Output location: s3a://spark/data/dims/car_colors

Data:

color_id	color_name
1	Black
2	Red
3	Gray
4	White
5	Green
6	Blue
7	Pink

Cars

Object Name: CarsGenerator

Output location: s3a://spark/data/dims/cars

Logic:

- Generate 20 cars based on data logic
- The car_id must be unique

Data:

- car_id: Random number with 7 digits
- driver_id: Random number with 9 digits
- model_id: Random number from 1 to 7
- color_id: random number from 1 to 7

Data Generator

Object Name: DataGenerator

Input location: s3a://spark/data/dims/cars

Output KAFKA topic: sensors-sample

- Generate a row for each car every second
- The data should be in JSON format
- Event data:
  - event_id: unique ID
  - event_time: current timestamp
  - car_id: the id of the car
  - speed: random value from 0 to 200
  - rpm: random value from 0 to 8000
  - gear: random value from 1 to 7

- Please use While True, and Sleep 1 second.

Data Enrichment

EX 5
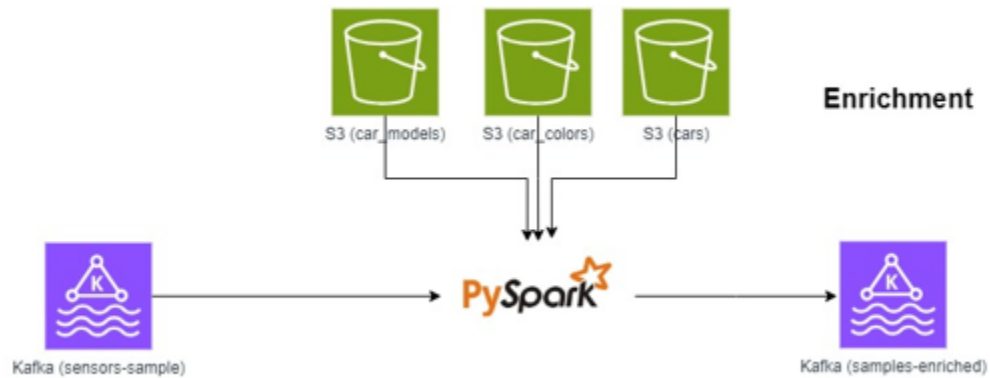
Object Name: DataEnrichment

Input KAFKA topic: sensors-sample

Output KAFKA topic: samples-enriched

Logic:

- Add to each event the following data:
    - driver_id: The ID of the driver
    - brand_name: The name of car brand
    - model_name: The model name of the car
    - color_name: The color name of the car
    - expected_gear: calculation: round(speed/30)

Alerting Detection

Alerts detection

Kafka (samples-enriched) → PySpark → Kafka (alert-data)

Object Name: AlertDetection

Input KAFKA topic: samples-enriched

Output KAFKA topic: alert-data

Logic:

- Filter rows to keep only alerting rows
- Alert conditions:
  - Speed is greater than 120
  - Expected gear is not equals to actual gear
  - RPM is greater than 6000

EX 7

Alerting Counter

Kafka (alert-data)   PySpark   Console

Object Name: AlertCounter

Input KAFKA topic: anomaly-alerts, alert-data

Output: console

Logic:

- Print the following aggregation across all data in last 15 minutes
- Aggregations:
  - num_of_rows: Total number of rows
  - num_of_black: Count rows with color == black
  - num_of_white: Count rows with color == white
  - num_of_silver: Count rows with color == silver
  - maximum_speed: The maximum speed
  - maximum_gear: The maximum gear
  - maximum_rpm: The maximum rpm