Big Data Technologies 874

Post Block Assignment 1: Streaming

Overview

Part 1

- To apply some of the concepts from the lectures and tutorials.
- To consider implementations and architectures in the streaming arena.

Part 2

- Perform hands-on processing tasks using streaming technologies.

Submission

- A form will be provided on SUNLearn for completion of this assignment.
- Any code submissions must be made in the form of working notebooks.

Part 1: Streaming and Big Data Processing

Question 1: Infrastructure [3] Consider the image in Figure 1.

1. What role is Kafka playing in this infrastructure. [1]
2. Suppose that the latest data ingested to the HADOOP cluster were completely destroyed. How would you recover those data? Describe it in detail. [2]

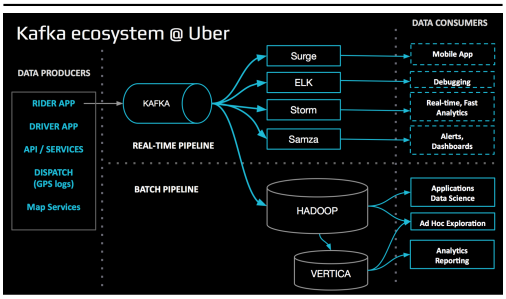

Figure 1: Uber Infrastructure

Question 2: Out of Order/Late Arriving [3] Consider a use case where you may not tolerate data loss (e.g. billing internet usage at an ISP). Packets/events may arrive out-of-order, but with a time difference between event time and processing time of at most *60s*.

1. What windowing options should you choose in Beam. Briefly explain your answer? [3]

Question 3: Anomaly Detection on Streams [16] Your organisation (an internet service provider) is embracing artificial intelligence for IT operations (AIOps). You have been tasked with assisting in the monitoring of operational data streams that are critical to business operations. One such stream is bytes sent and received per user. A large reduction in any of these values (or aggregates thereof) may indicate outages. This stream has been identified as the first use case to prove the viability of an online anomaly detection approach to business.

Details: These data are currently arriving via a single topic/stream; Demand varies throughout the day; These data are collected and attributed to four zones or areas; The stream is a high volume and high velocity stream (2 terabytes per day). Your organisation implements their data science solutions in Python. A prototype is required at first. All data science must be performed on a stream or a number of streams.

An example of two records is provided below.

user_id	zone	url	timestamp	bytes
01f4f1c2	A	www.google.com/colab	2020-09-15 12:36:19	212
01f4f1c2	B	www.sunlearn.co.za	2020-09-16 12:36:59	416

1. Propose an anomaly detection technique, identify a package/code that you will use, and motivate why this choice of technique is sufficient (please provide references/links to demonstrate your research). [5]
2. Explain how your technique may be integrated with Apache Beam, specifically with reference to the Beam API. [3]
3. Suppose the IT team placed the stream in Google PubSub so as to implement a decoupled architecture. How do you access these data and will there be messages in the queue if there are others already reading from it? [2]
4. Furthermore, IT has discussed creating derived streams for each *zone* as they have unique reporting needs. As your use case may be the first to use this functionality, you have been asked to provide an approximate cost estimate of additional streams. [2]
5. Alternatively, suppose Kafka was the chosen streaming technology, how would you:
    1. Use partitions to prevent the need to subscribe and publish to a new stream for each zone (thus maintaining only the initial stream)? [1]
    2. How would you ensure that messages are available for your detection algorithms, and are read once only. [1]
6. Describe how/if event ordering may affect your approach, and how you would remedy this. [2]

Part 2: Processing with Beam

Question 4: Merge PCollections [12] Consider the data from the streaming tutorial. The second file received (*orders.csv*) contains a data excerpt from the retailer's orders database. Combine the 2 input files (the other being *users.csv*) in order to provide insights into the average ordering profiles of their customers.

Expand your pipeline from the tutorial to do the following:

1. Merge the 2 input files. [3]
2. Perform a transformation that determines the average number of orders for female and male customers, respectively. [3]
3. Perform a transformation that groups users into age groups [16-26), [26-36), [36-46), [46-56), and determine the total number of orders placed by customers in each age group. [3]
4. Determine the total number of times that spinach was purchased within the [16-26), [26-36), [36-46), [46-56) age groups. [3]

Question 5: Stream Analytics [20] Consider the premise in the Question 3: *Anomaly Detection on Streams*, namely that you need to create an approach for monitoring a data stream that is reporting on internet usage per user, per zone (area).

You are required to implement a working prototype which will process the elements of the stream and determine if any outages occur (an outage is defined as a scenario in which many users experience limited

or no connectivity for some duration of time). As this is a first order solution, batch data may be used to develop your solution (synthetic data of internet usage is provided here.

Once a prototype has been created, your organisation will test it on live/production streams. To this end, it must be developed in Beam.

1. Perform initial exploratory data analysis to establish if outages have occurred in the provided data. Provide a working notebook to demonstrate your analyses and conclusion [3].
2. Implement a solution on a stream which identifies these outages [15]
3. Will your solution scale to distributed/parallelized streams? If not, identify the shortcomings. If so, what allows it to do so? [2]

Hints:

- It may be necessary to correctly timestamp data so that Beam is able to reason about it in the stream.
- Consider using a Python with pandas dataframes to test your approach before attempting it on a stream itself.