AIN429 Data Mining Laboratory

Assignment 2: Apache Kafka
Date Issued: 08.11.2023
Date Due: 15.11.2023

Aim of the Experiment

Apache Kafka is a widely used open-source stream processing platform for real-time data streaming and data processing. It is designed to handle high throughput, fault tolerance, and scalability, making it an essential tool in modern data engineering. In this assignment, you will explore various aspects of Apache Kafka, its architecture, and its applications in computer engineering. At the end of this assignment, you will gain practical knowledge and skills in working with Apache Kafka.

Assignment Tasks:

Task 1: Understanding Apache Kafka Basics (15 marks)

1. Provide an overview of Apache Kafka, explaining its purpose and the problems it solves in the context of data engineering.
2. Describe the key components of Apache Kafka architecture, including producers, consumers, topics, brokers, and partitions.
3. Explain the role of Zookeeper in Apache Kafka and how it ensures fault tolerance and distributed coordination.
4. Discuss the concepts of message retention and compaction in Kafka.

Task 2: Implementing Real-time Data Processing with Kafka (25 marks)

1. Demonstrate how to integrate Kafka with a sample application for real-time data processing.
2. Provide code examples (using Java or any preferred programming language) for creating Kafka producers and consumers.
3. Implement a simple data processing task using Kafka streams or Kafka Connect.
4. Discuss the challenges and solutions related to handling late-arriving data and out-of-order messages in real-time processing scenarios.

Task 3: Exploring Advanced Kafka Features (20 marks)

1. Discuss advanced features of Apache Kafka, such as exactly-once semantics, transactions, and security mechanisms (SSL/TLS, SASL).
2. Explore Kafka Connect and Kafka Streams API in detail, explaining their use cases and benefits.

3. Investigate the use of Kafka in microservices architectures and event-driven applications.
4. Analyze a case study or research paper that showcases the innovative use of Apache Kafka in a real-world scenario.

Task 4: Performance Optimization and Monitoring (20 marks)

1. Discuss strategies for optimizing Kafka cluster performance, including partitioning, replication, and hardware considerations.
2. Explain how to monitor Kafka cluster using tools like Kafka Manager, Confluent Control Center, or custom monitoring scripts.
3. Discuss common performance bottlenecks in Kafka and methods to mitigate them.
4. Provide insights into capacity planning and scaling Kafka clusters based on workload and usage patterns.

Grading

- Task 1 (%20)
- Task 2 (%30)
- Task 3 (%25)
- Task 4 (%25)

REMARKS:

- Submission format:
  - < studentID_name_surname_hw2.zip>
    - report.pdf
    - code files

- Your submission should be matched with the format above**. 10 point** penalty will be applied on mismatched submissions.
- You will use an online submission system to submit your experiments.
- https://submit.cs.hacettepe.edu.tr/ Deadline is 23:59. No other submission method (such as; CD or email) will be accepted.
- Do not submit any file via e-mail related to this assignment.
- The assignment must be original, INDIVIDUAL work. Duplicate or very similar assignments are both going to be punished. General discussion of the problem is allowed, but DO NOT SHARE answers, algorithms, or source codes.
- You can ask your questions through the course’s Piazza group and you are supposed to be aware of everything discussed in the group.