Towards a Data Lake Model using Real-Time ingestion and Querying

Steven Wen, Asher Song
Supervisor: Dr. Farhana Zulkernine, Student Mentor: Ahmed Harby

Queen's UNIVERSITY
COMPUTING

Overview

Problem Description

The challenge of efficiently and effectively processing and analyzing large amounts of real-time data. Real-time data analysis is currently expensive and complex, and traditional batch analytic approaches can result in outdated data. The goal of this project is to find a less costly way to ingest the data and build a pipeline for real-time processing. This requires the selection and implementation of appropriate data ingestion tools such as Nifi with Kafka, Apache Spark with Nifi and data management systems such as lakeFS.

Objective

- To develop a pipeline for real-time ingestion of large amounts of data using Apache Kafka and Apache Nifi.
- To store real-time data in a database (MongoDB) for immediate use.
- To store real-time data in a data lake (lakeFS) for post-processing.

Implementation

Dataset:

- MobiAct v2.0 human activity recognition sensor dataset.
  It contains sensor data from the accelerometers and gyroscopes of mobile devices worn by human subjects while performing various activities such as walking, running, sitting, standing, and others.

Software:

- Apache Nifi: Open-source data integration tool used for processing and moving data in real-time. It provides a web-based interface for designing data flows
- lakeFS: Open-source data version control designed for data lakes turning object stores into Git-like repositories

Data Ingestion

Figure 1. Kafka as Producer

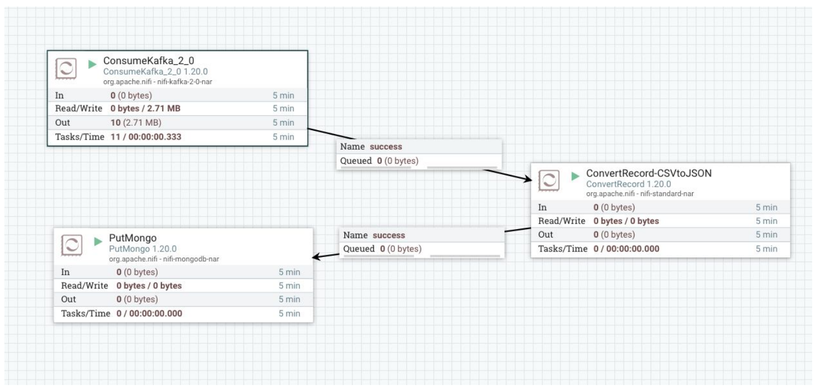


Figure 2. Kafka as Consumer

Components

- GetFile: Getting file from a local folder - it can be substitute to GetMongo to receive file from a database.
- PublishKafka:Sending data to an Apache Kafka topic. Apache Kafka is a distributed messaging system that is commonly used for building real-time data pipelines and streaming applications.
- ConsumeKafka: Consuming data from an Apache Kafka topic. Uses for moving data from a Kafka topic to another system.
- ConvertRecord: Convert CSV file to Json in order to ingest into MongoDB.
- PutMongo: Inserting or updating data in a MongoDB database. MongoDB is a popular NoSQL database used for storing and processing unstructured or semi-structured data.

Querying

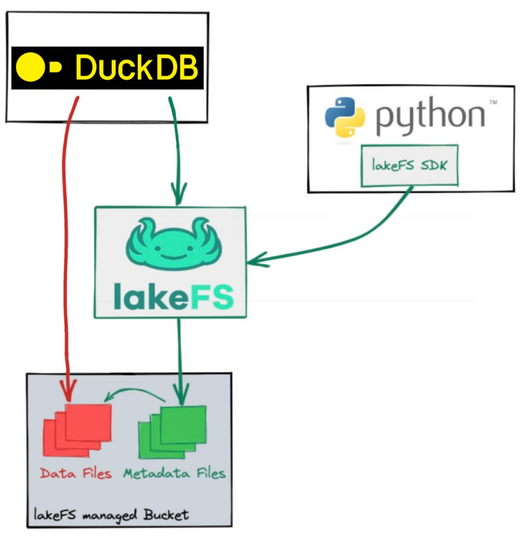
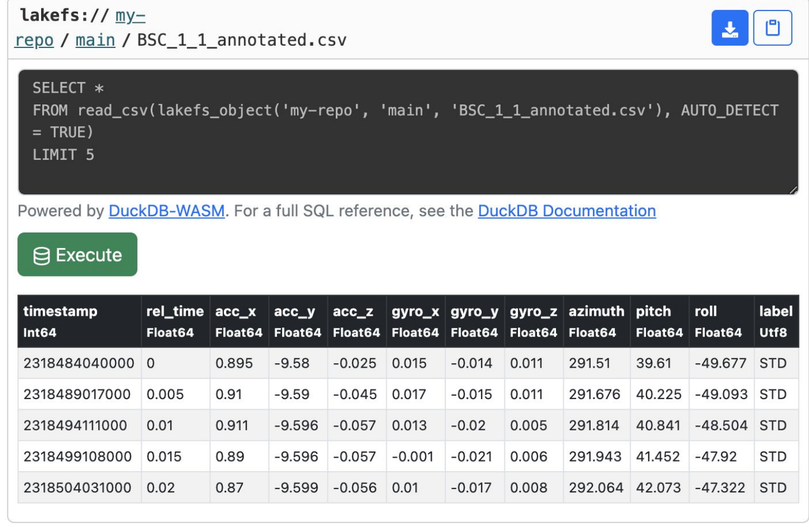


Figure 3. lakeFS interactions

lakeFS was used for their Git-like version control system to manage the ingested data.



timestamp Int64	rel_time Float64	acc_x Float64	acc_y Float64	acc_z Float64	gyro_x Float64	gyro_y Float64	gyro_z Float64	azimuth Float64	pitch Float64	roll Float64	label Utf8
2318484040000	0	0.895	-9.58	-0.025	0.015	-0.014	0.011	291.51	39.61	-49.677	STD
2318489017000	0.005	0.91	-9.59	-0.045	0.017	-0.015	0.011	291.676	40.225	-49.093	STD
2318494111000	0.01	0.911	-9.596	-0.057	0.013	-0.02	0.005	291.814	40.841	-48.504	STD
2318499108000	0.015	0.89	-9.596	-0.057	-0.001	-0.021	0.006	291.943	41.452	-47.92	STD
2318504031000	0.02	0.87	-9.599	-0.056	0.01	-0.017	0.008	292.064	42.073	-47.322	STD

Figure 4. SQL query using DuckDB on lakeFS web-UI

Interface

- Connect to and interface with lakeFS using their Python SDK
- Create a repository and branches to manage data
- Upload objects to then be queried
- lakeFS allows for SQL queries powered by DuckDB on their web-UI

Future Work

- Add path for the ingestion from a database than local file
- Connect the ingestion with the querying
- Perform queries externally i.e. through Python scripts using PySpark

Reference

1. Ahmet, A., & Abdullah, T. (2020, December). Real-time social media analytics with deep transformer language models: a big data approach. In 2020 IEEE 14th International Conference on Big Data Science and Engineering (BigDataSE) (pp. 41-48). IEEE.
2. Data ingestion: The first step to a sound data strategy. Stitch. (n.d.). Retrieved November 6, 2022, from https://www.stitchdata.com/resources/data-ingestion/
3. T. Hlupić, D. Oreščanin, D. Ružak and M. Baranović, "An Overview of Current Data Lake Architecture Models," 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2022, pp. 1082-1087, doi: 10.23919/MIPRO55190.2022.9803717.