Week 4 Labs

Due Date: Monday, 25th November 2024

Lab Exercise 1

Dataset

The Wikipedia Pageviews dataset records user interactions with Wikipedia pages, including date, time, language, title, and view counts. It offers insights into web traffic, user behavior, and content trends. For a project, students can analyze traffic patterns, evaluate topics, and practice big data techniques to derive actionable insights from large-scale datasets.

Prerequisites: Create a Dataproc cluster with Jupyter & Component Gateway on GCP.

Set Up Apache Spark and Jupyter Notebooks on Dataproc

Objective: By the end of this lab, you will be able to use Spark DataFrames and SQL to retrieve and manipulate Wikipedia page views data, write the data to BigQuery (a data warehouse on GCP) and query the data for insights.

Tasks:

Follow the instructions on the page via this link to perform the tasks (Repo Instructions Page)

- Read the Bigquery table into Spark DataFrame
- Filter for English version of Wikipedia for both desktop and mobile versions (‘en’ and ‘en.m’) with more than 100 views
- Group by title and order by page views to see the top pages
- Write the spark Dataframe to a BigQuery table
- Write a query to retrieve the top 10 most-viewed pages where the title contains the word "United".
- Repeat the same steps but perform the transformations using Spark SQL **(Steps to Use Spark SQL)**
- Visualize the total views across datehour using Pandas plotting (Steps to Pandas Plotting)

Lab Exercise 2

Overview

In this provisioned lab environment, you create a streaming data pipeline with Kafka providing you a hands-on look at the Kafka Streams API. You will run a Java application that uses the Kafka Streams library by showcasing a simple end-to-end data pipeline powered by Apache Kafka.

Prerequisites: An account on Cloud Skills Boost

Objective: By the end of this lab, you will be able to start a Kafka cluster on a Compute Engine single machine, write example input data to a Kafka topic using the console producer included in Kafka, process the input data and inspect the output data using console consumer.

Setup and requirements

Before you click the Start Lab button, read and follow the instructions. The lab is timed, and you cannot pause it.

Tasks: Follow the instructions on the page via this link to perform the tasks

Link to Lab

- Set up Kafka
- Prepare the topics and the input data
- Process the input data with Kafka streams
- Inspect the output data
- Stop the Kafka cluster