Introduction to Big Data
Graded Assignment 3

Name - Avijeet Palit
Roll - 21f1005675
Date -20/10/2024

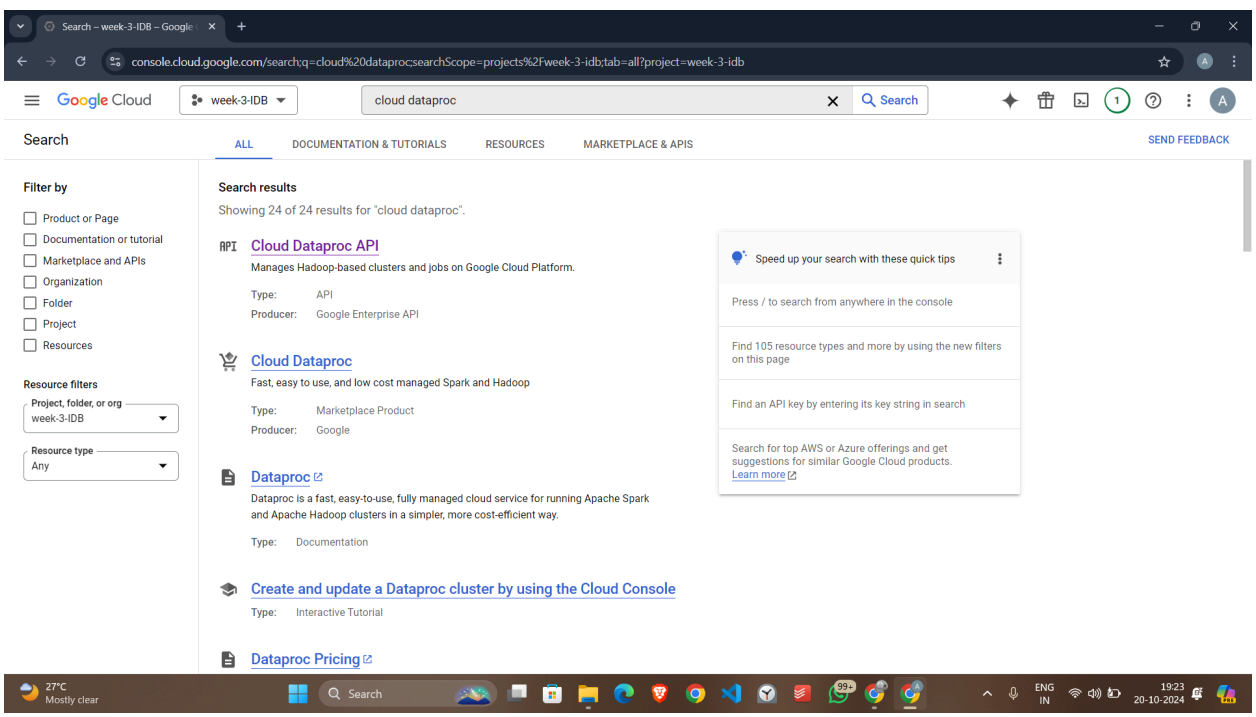
Problem Statement:

Implement a Spark code to analyze user clicks based on the hashing example discussed in the lecture video. The task involves creating a text file as shown in the video and determining the number of user clicks within specified time intervals: 0-6, 6-12, 12-18, and 18-24.

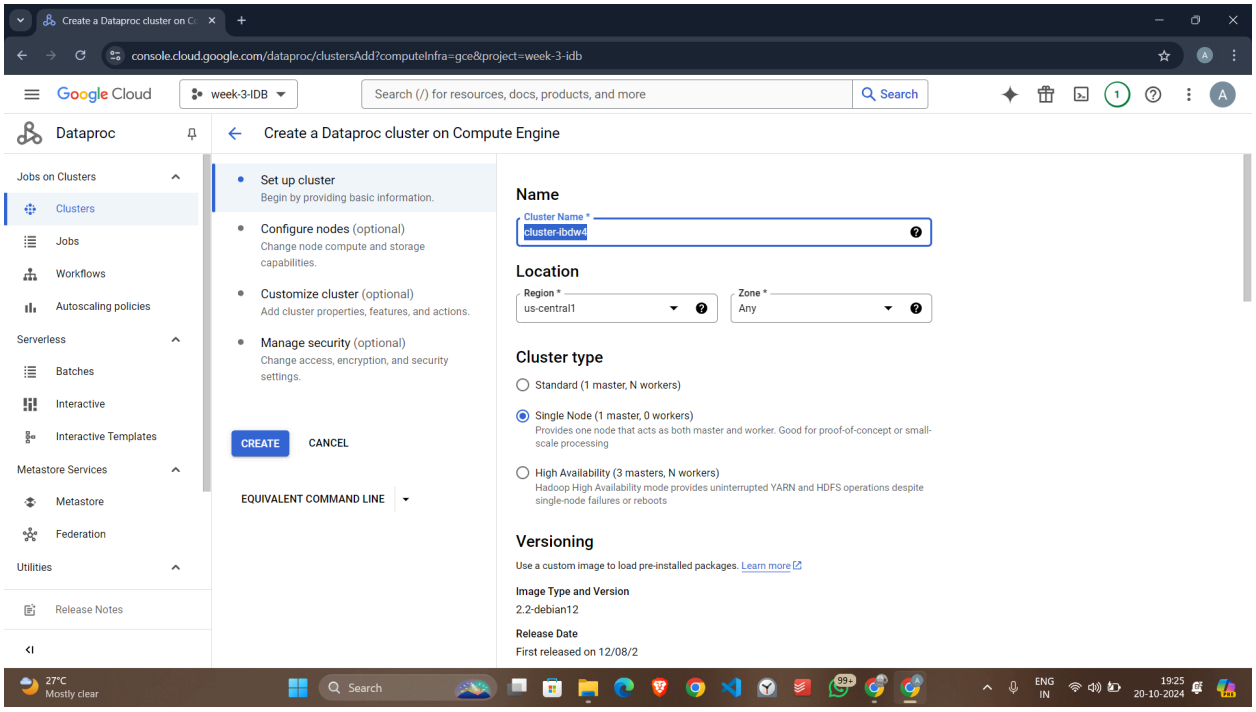
Approach:

To address the requirements outlined in the problem statement, I developed a Python program to execute Spark code within a Google Cloud Dataproc Cluster. Below are the detailed steps taken:

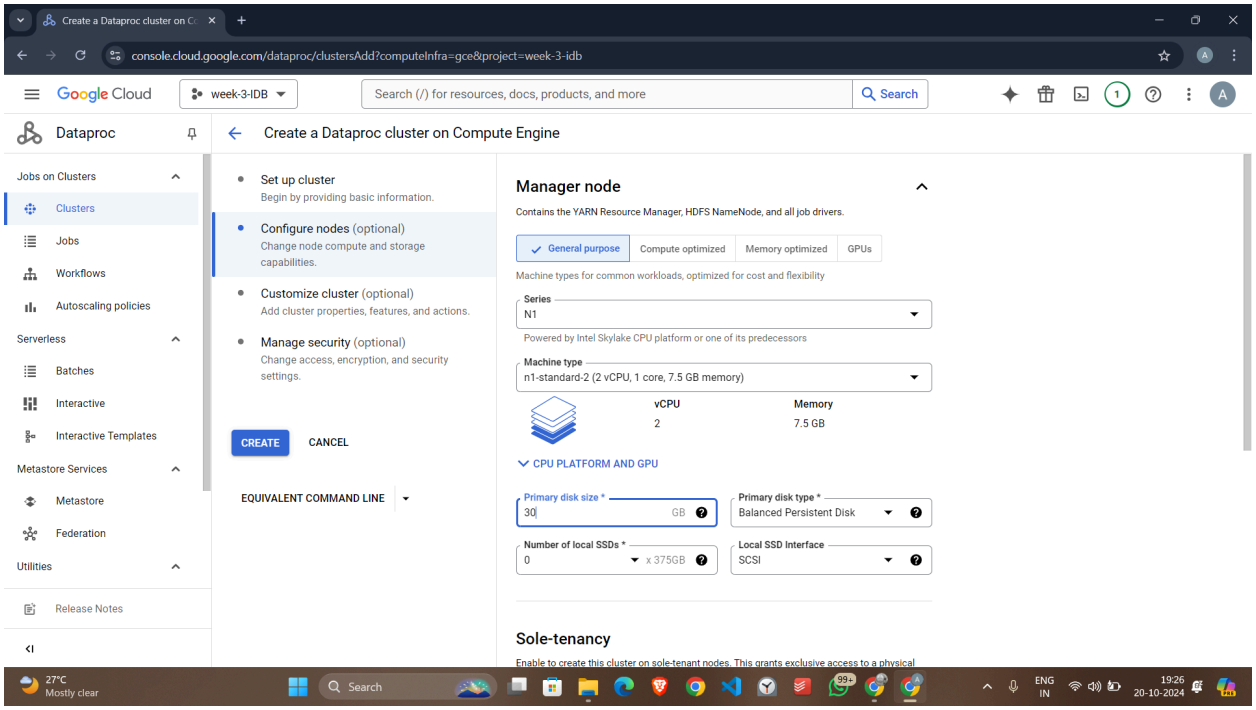
1. **Setting Up the Google Cloud Dataproc Cluster:**
    - Enabled the necessary APIs.

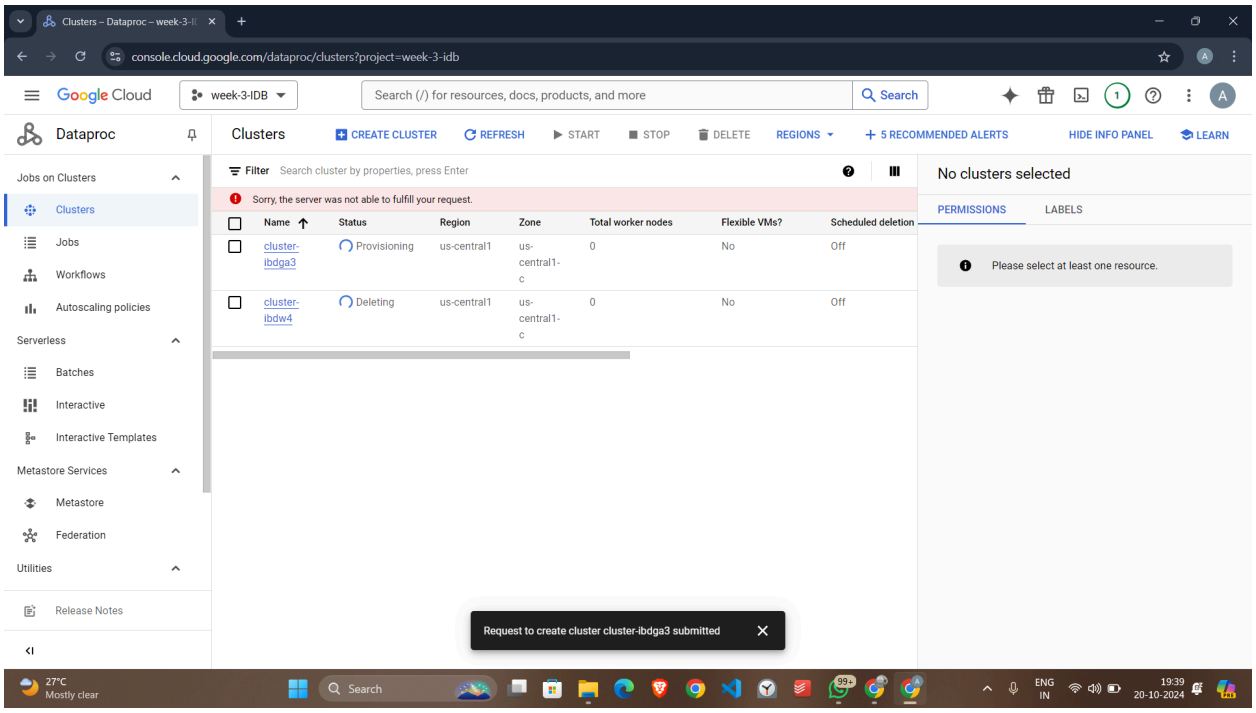


- Created a new Google Cloud Dataproc Cluster by navigating to "Dataproc" under the "Cluster" section in the "Analytics" sidebar.
- Named the cluster `cluster-ibdga3`, selected `us-central1` as the region, and chose "Single Node" for the cluster type.



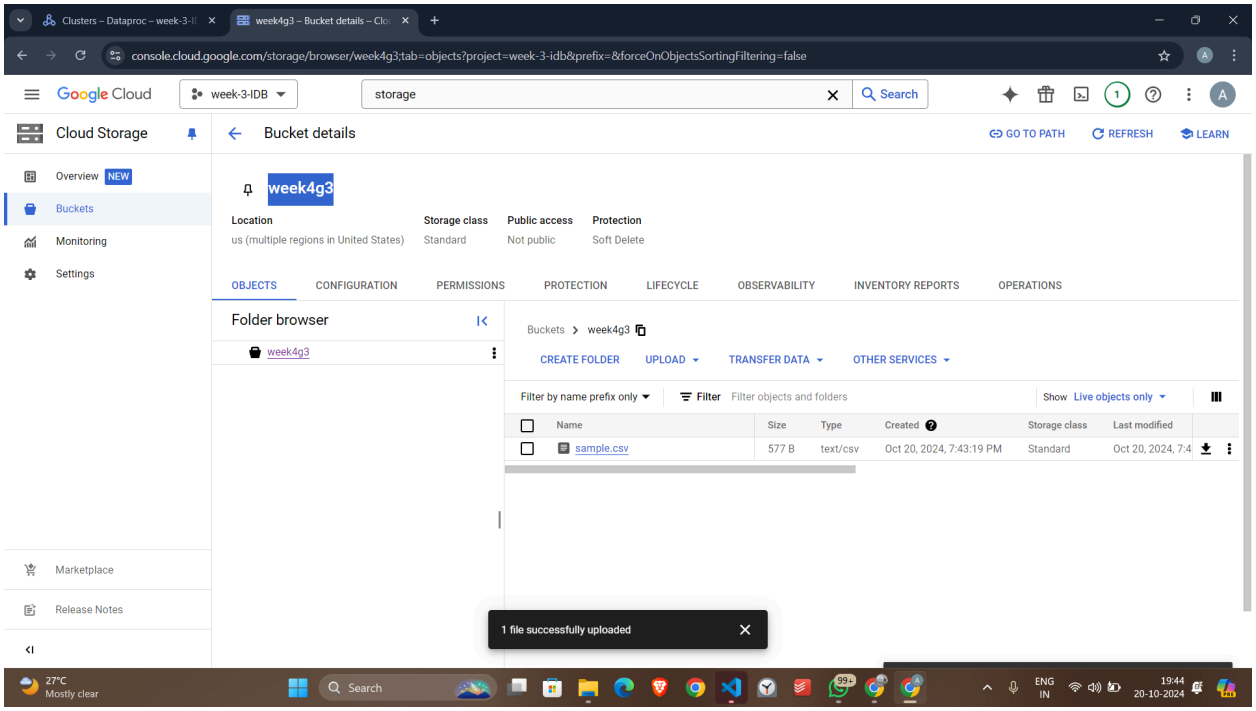
- Configured the Manager Node series as N1 along with the required settings.



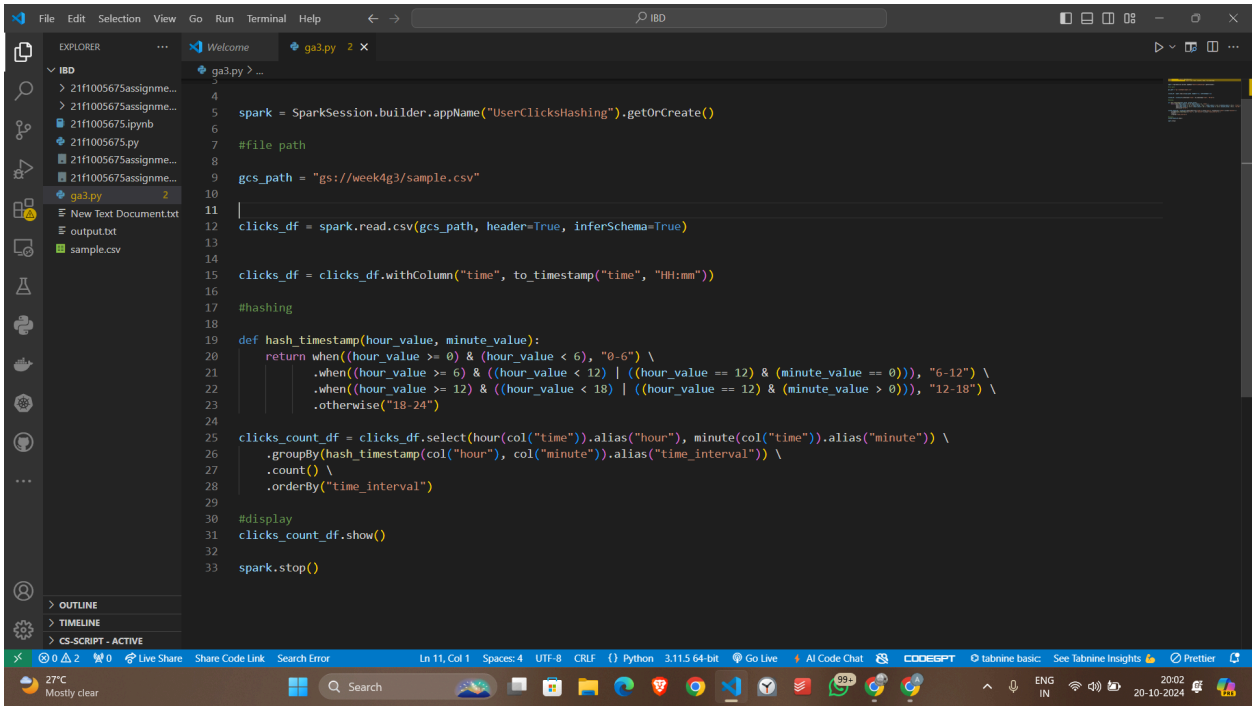


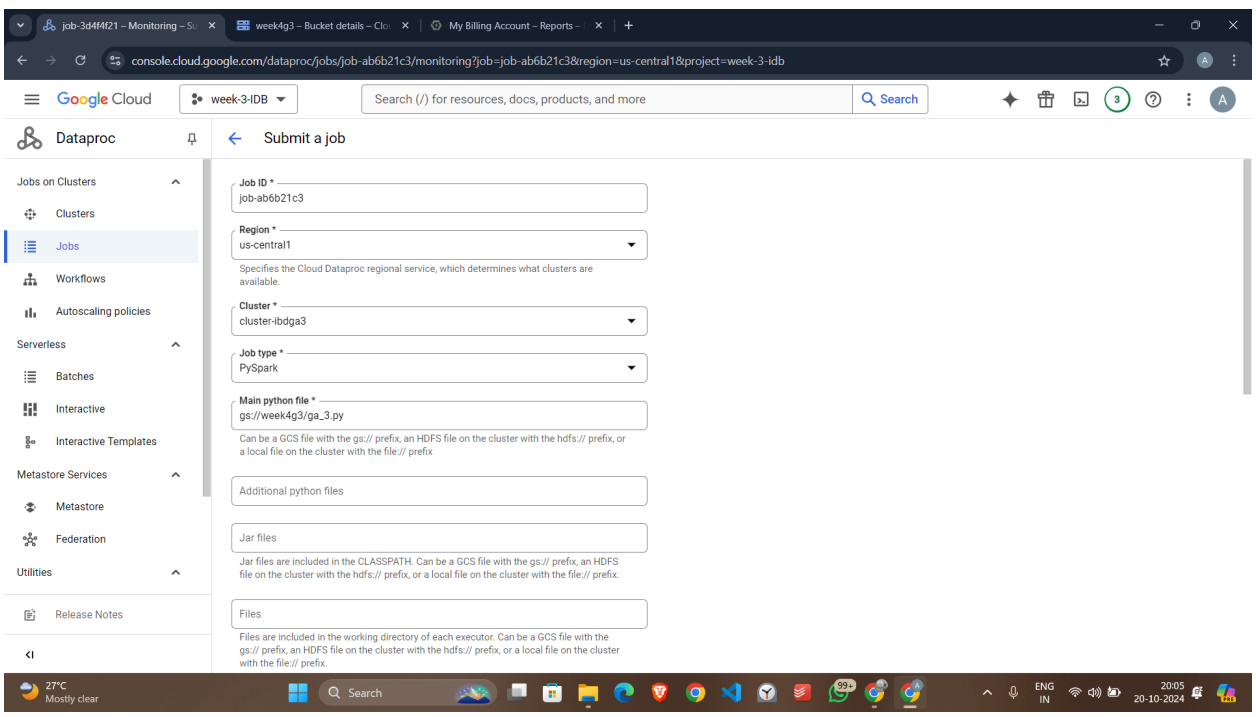
2. **Creating and Using a Cloud Storage Bucket:**
   - Created a bucket named week4g3 on Google Cloud Storage for data management and organization.

- Uploaded the Excel file `sample.csv` to the bucket.



3. **Writing and Submitting the PySpark Code:**
   - Developed the PySpark script `ga.py` using VS Code and saved the file in the bucket.
   - Set up a job with dataprop





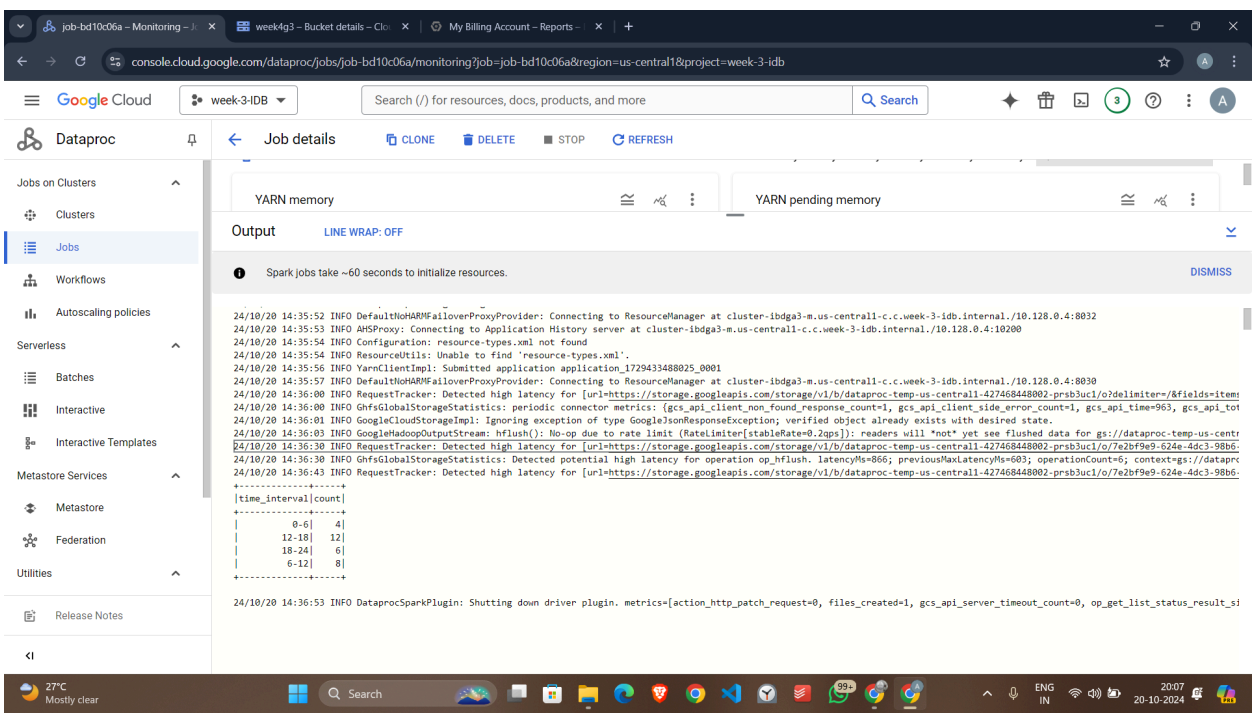
Explanation of the Code:

1. **Importing Libraries:**
   - Imported `SparkSession` and necessary functions from `pyspark.sql.functions` and `pyspark.sql`.

2. **Initializing Spark Session:**
    - Created a Spark session named "UserClicksHashing".
3. **Reading Data:**
    - Loaded the `sample.csv` file from the Google Cloud Storage bucket (`gs://week4g3/sample.csv`) into a DataFrame called `clicks_df`.
4. **Parsing the Time Column and Creating a Hash Function:**
    - Parsed the time column from the `sample.csv` file into a `time` timestamp and created a `hash_timestamp` function to hash the time values.
5. **Counting Clicks:**
    - Applied the hashing function to the DataFrame.
    - Grouped the data by specified time intervals and counted the number of clicks within each interval.

Results:

The output displayed the number of user clicks for each specified time interval, confirming the correct implementation of the logic. The results were as follows:



The counts accurately reflect the number of clicks that occurred within each defined interval.