AfterWork

Data Filtering with Python Project

Project Deliverable

- Your deliverable will be a Python notebook with your solution.

Instructions

Background Information

Welcome to the digital frontier of "FictionaTech," a visionary technology company committed to pioneering cutting-edge web services and digital solutions. At FictionaTech, we are dedicated to providing innovative web-based services that enrich the digital landscape and transform the way users interact with online platforms.

Meet Jane Anderson, a dedicated data analyst at FictionaTech. Jane's role is crucial in optimizing the user experience, enhancing content delivery, and ensuring the high performance of FictionaTech's web services. As part of her responsibilities, she is tasked with analyzing web server logs to uncover insights about user behavior and page access patterns.

Research Question

Jane embarks on a mission to tackle the following critical research questions:

1. Retrieve the first 20 log entries to explore the dataset.
2. What are the log records with the HTTP status code '404 Not Found'?
3. Identify log entries with server response times exceeding 100 milliseconds?
4. Retrieve data for the specific web page of interest: '/products.html'
5. Filter records for user agents matching 'Safari 15.0'.
6. Filter records with HTTP status codes '200 OK' and '404 Not Found'.
7. Select log entries for a specific date, e.g., '2023-10-29'.
8. Select log entries with a status code of '200 OK' and user agents matching 'Safari 15.0'.
9. Find log entries with response times between 50 and 100 milliseconds.
10. Filter log entries for bytes transferred between 1024 and 4096 bytes.

Methodology

1. **Business understanding**
2. **Data loading**
    - Load the provided web server log dataset, which includes timestamp, IP address, HTTP method, HTTP status code, URL, user agent, referrer, bytes transferred, and server response time.
3. **Data exploration**
    - Explore the dataset to understand its structure and characteristics.
    - Calculates summary statistics for numerical columns, such as bytes transferred and server response time.
4. **Data cleaning**
    - Addresses missing values, identifying gaps in the "Referrer" and "Bytes Transferred" columns.
    - Decides on the best approach to handling these missing values, either through imputation or removal.
    - Duplicate records in the dataset are identified and resolved.
    - Check for any inconsistencies or anomalies in the data and rectify them.
5. **Data analysis**
    - Perform analysis by counting the frequency of each web page accessed based on the "URL" column.
    - Rank the web pages according to their access frequency, ultimately identifying the most frequently accessed pages.
6. **Recommendations**
    - Provide recommendations based on the analysis results.

You can use the following guiding notebook (link) to get started.

Dataset Overview

This dataset (https://bit.ly/3FF0VA8) contains web server log records generated by FictionaTech's website. The log records provide information about user interactions with the website, including the timestamp of each access, IP addresses, HTTP methods, HTTP status codes, URLs of accessed pages, user agents, referrers, bytes transferred, and server response times.

Dataset Glossary

- **Timestamp:** The date and time of the access request, formatted as "YYYY-MM-DD HH:MM:SS"
- **IP Address:** The IP address of the client making the access request.
- **HTTP Method:** The HTTP method used to access the web page, such as GET or POST.

- **HTTP Status Code**: The HTTP response code indicating the outcome of the request (e.g., 200 for a successful request, 404 for not found).
- **URL:** The URL or URI of the web page accessed by the user.
- **User Agent:** A string providing information about the client's browser, operating system, and device.
- **Referrer:** The URL from which the user navigated to the current page. This column may contain missing values.
- **Bytes Transferred:** The size of the data transferred during the request. This column may contain missing values.
- **Server Response Time:** The time taken by the server to respond to the request, measured in milliseconds.