BITS Pilani

BIG DATA SYSTEMS – ASSIGNMENT 2

Web Server Log Analysis using Apache Spark

Overview & background:

A U.S.-based firm intends to analyze its web server logs to study application access patterns and determine if changes are needed in its deployment strategy. Two log files in compressed (.gz) format, each corresponding to a different month, have been made available for analysis. These logs follow the Common Log Format (CLF) of Apache and contain valuable information, such as request timestamps, the requesting host, requested resources, response codes, and more. The file names are:

- Access_log_Jul1995.gz
- Access_log_Aug1995.gz

Since the logs are large and contain lot of information, they need to be processed and analyzed efficiently using Big Data tools. For this assignment, you will leverage Apache Spark with PySpark and SparkSQL to handle and analyze the logs. You can choose to work with either Google Colab or a local Spark setup for this task.

Description:

1. METADATA

The log files follow Apache Common Log Format Logging in W3C httpd and it is given below:

remotehost rfc931 authuser [date] "request" status bytes

Each field in this format are explained below:

Field	Meaning
remotehost	*Remote hostname (or IP number if DNS hostname is not available, or if DNSLookup is Off).*
rfc931	*The remote logname of the user.* *A hyphen (-) indicates unavailable information.*

authuser	*The username as which the user has authenticated himself/herself.* *A hyphen (-) indicates unavailable information.*
[date]	*Date and time of the request. Format:* ***[day/month/year:hour:minute:second zone]***
"request"	*The request line exactly as it came from the client.* *eg:* ***"GET /history/skylab/skylab.html HTTP/1.0"*** *=> method used by the client is GET, requested resource (request endpoint) is /history/skylab/skylab.html, and the client used the protocol HTTP/1.0*
status	*The HTTP status code returned to the client.*
bytes	*The content-length of the document transferred.*

Notes:

1. Log files may contain missing data or malformed fields.
2. For simplicity, Zone can be ignored while processing Dates.

2. Data Wrangling

Before performing any analysis, you may perform appropriate data wrangling steps prepare the log data for analysis. These steps may include:

a. Parsing the log files correctly according to the Common Log Format.

b. Handling missing or malformed data.

c. Structuring the data into a format suitable for analysis and presentation.

Document each of the issues encountered during data wrangling and provide a justification for the actions taken to address them.

3. Analytics:

Analyze the web server logs to extract the following insights:

i. Count of total log records
   - Determine the total number of log entries in the dataset.

ii. Count of unique hosts:
   - Determine the number of unique hosts in the log data.

iii. Date wise unique host counts:
   - For each date, identify and count the number of unique hosts accessing the web server. List the result in the date order. Print the date in dd-MMM-yyyy format.

iv. Average Requests per Host per Day:
   - Calculate the average number of requests made per host for each day (day of the month). List the results in the day order.

v. Number of 404 Response Codes:
   - Identify and count the number of instances where the server returned a 404 (Not Found) response code.

vi. Top 15 Endpoints with 404 Responses:
   - Identify the endpoints with the most 404 errors and list the top fifteen.

vii. Top 15 Hosts with 404 Responses:
   - Identify the hosts generating the most 404 errors and list the top fifteen.

Submission Requirements:

Deliverables:

1. Jupyter Notebook:
   - Develop using either Google Colab or a local Jupyter Notebook. If using the latter, ensure Spark is installed and properly configured on the local machine.
   - Ensure high code quality with inline documentation.
   - Submit the notebook file (.ipynb) containing outputs for all cells.

2. Report (PDF/Word):
   - Document the following:
     a. Problem-solving approach.

b. Details of the development environment and setup. If Spark is set up locally, specify the Spark mode and configuration details.
c. Data loading scripts.
d. Data wrangling steps with encountered issues and actions taken to address each of them.
e. Code/query and results of each analytical query
f. Justifications for key decisions

3. Video (mp4):
   - Record the data loading and wrangling steps and execution of queries and their outcomes in real-time as a .mp4 file.

Submission File Format:

- Submit the three files specified in the 'Deliverables' section viz. the Jupyter notebook(.ipynb) file, the report (.docx/.pdf) file and the video file
  - Notebook: **Asgn2_Grp_<your_group_no>_code.ipynb**
  - Report: **Asgn2_Grp_<your_group_no>_report.pdf** OR **Asgn2_Grp_<your_group_no>_report.docx**
  - Video: **Asgn2_Grp_<your_group_no>_video.mp4**

General Notes:

☐ The report document should have full name and the BITS Registration number of each group member.

☐ Each group consists of 3 students. For the assignment, one of the members may act as the group leader.
   - The Group leader will be providing a contribution weightage of all group members and marks will be awarded accordingly.
   - The group leader should submit / upload a single set of files after collecting the individual contributions from group members.

☐ Evaluation will emphasize:
   - Data wrangling and analysis steps.
   - Code quality.
   - Result accuracy.
   - Result readability.

- ☐ Instances of plagiarism will result in mark deductions for all groups involved.

Academic Integrity

Honesty is primarily the responsibility of each student. The institute considers cheating to be a voluntary act for which there may be a reason, but for which there is no acceptable excuse. It is important to understand that collaborative learning is considered cheating unless specifically allowed for by the professor. The term cheating includes but is not limited to: plagiarism, receiving or knowingly supplying unauthorized information, using unauthorized material or sources, changing an answer after work has been graded and presenting it as improperly graded, illegally accessing confidential information through a computer, doing the assignment for another student or having another student do the assignment for you.