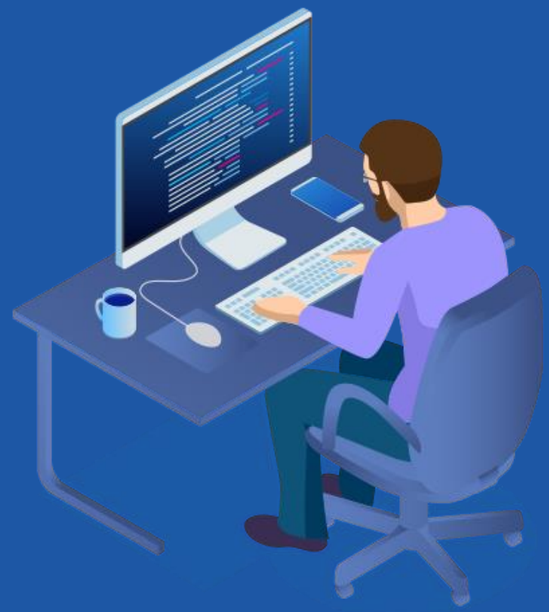
Big Data Hadoop and Spark Developer

Apache Server Log Analysis



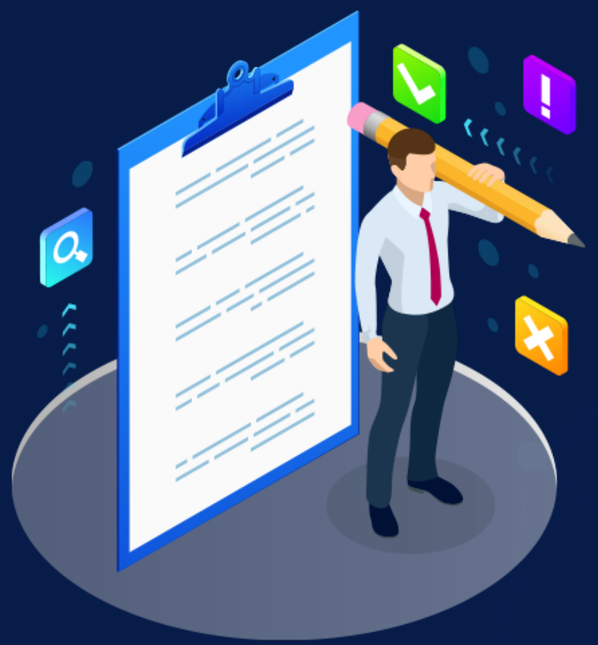
Objectives

- Perform server log analysis to assist businesses in identifying and analyzing critical business errors
- Identify potential customers and their domains

Prerequisites

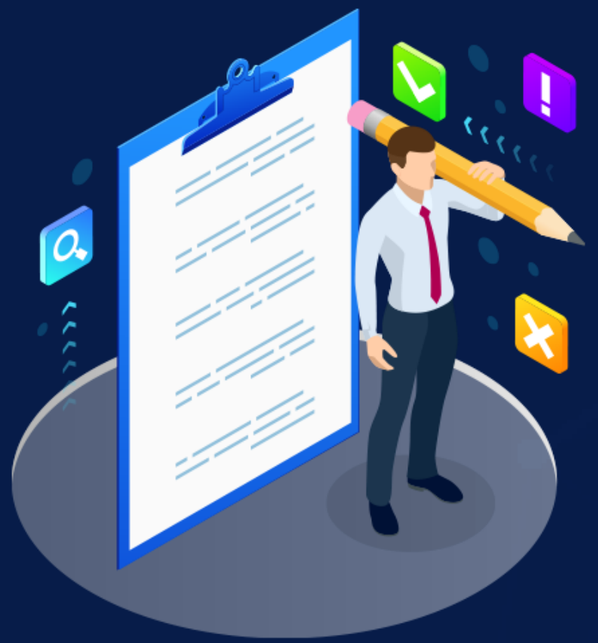
- Python
- PySpark
- Hadoop
- Spark

Industry Relevance



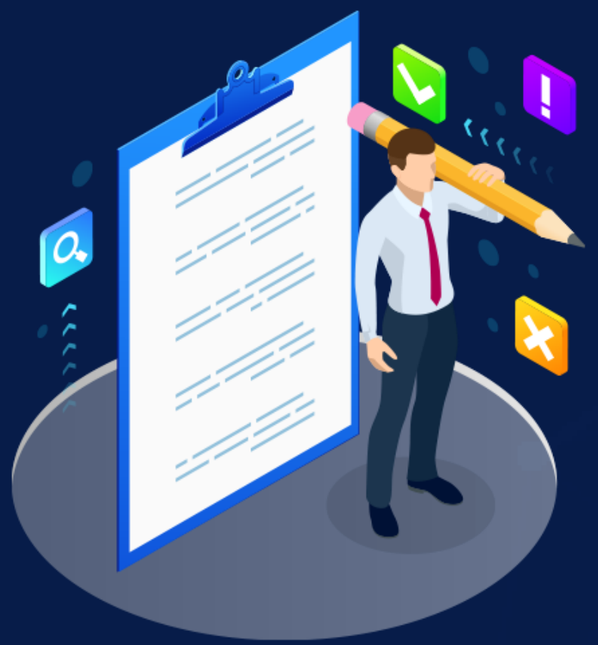
- **Python:** It is used for web development, data science and data analysis, machine learning, startups, and the finance industry
- **PySpark**: It writes Spark apps in Python.
- **Hadoop**: It is used in public sectors, such as intelligence, science, cyber security, and defense.
- **Spark:** It is used for machine learning and streaming data.

Problem Statement



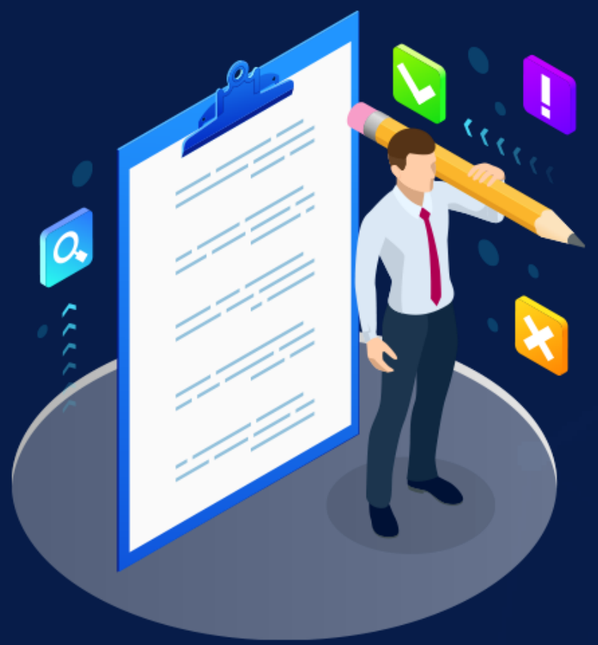
- The Apache services such as Hadoop, Spark, Tomcat, and Hive run on most data engineering servers throughout the world. All the services follow the same pattern because they are all open source. You are a data engineer who works for a start-up named "Hadoop Analytics," which serves major clientele.
- You have been assigned to one of their prestigious clients to resolve a production issue. As you are dealing with Hadoop, you are familiar with the working of logs. The server's information is stored in the logs along with the information listed below:
- 1. Resource details
- 2. Identification of the person who accessed the logs
- 3. Date and time the logs were accessed
- 4. Specifications on any problems that emerge
- 5. Information about the final product

Tasks to Perform



- Perform the following tasks on the dataset provided using PySpark:
- 1. Status code analysis:
    - Read the log file as an RDD in PySpark
    - Consider the sixth element as it is a “request type”
    - Replace the “single quote" with a blank
    - Convert each word into a tuple of (word,1)
    - Apply the “reduceByKey“ transformation to count the values
    - Display the data
  2. Arrange the result in descending order and display
  3. Identify the top 10 frequent visitors of the website

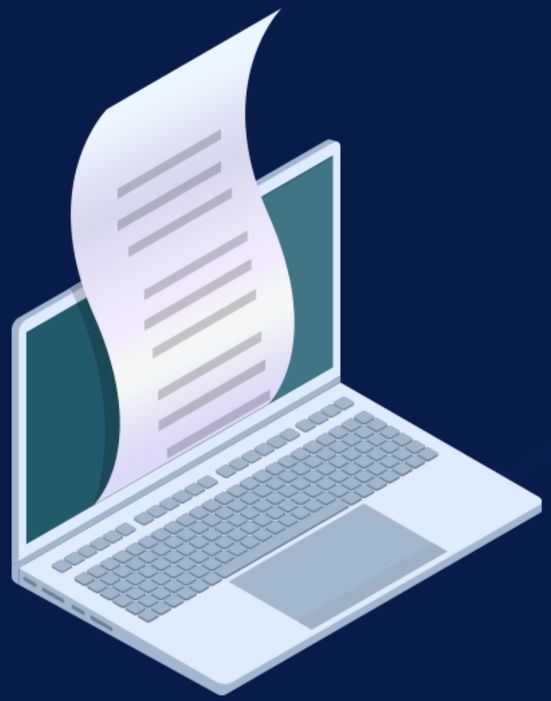
Tasks to Perform



- Perform the following tasks on the dataset provided using PySpark:

4. Identify the top 10 missing (does not exist) URLs using these steps:
    - Read the log file as an RDD in PySpark
    - Identify the URLs for which the server is returning the 404-request code and display the data
5. Identify the traffic (total number of HTTP requests received per day)
    - Read the log file as an RDD in PySpark
    - Fetch the DateTime string and replace "[" with blank
    - Get the date string from the DateTime
    - Identify HTTP requests using the map function
6. Identify the top 10 endpoints that transfer maximum content in megabytes and display the data

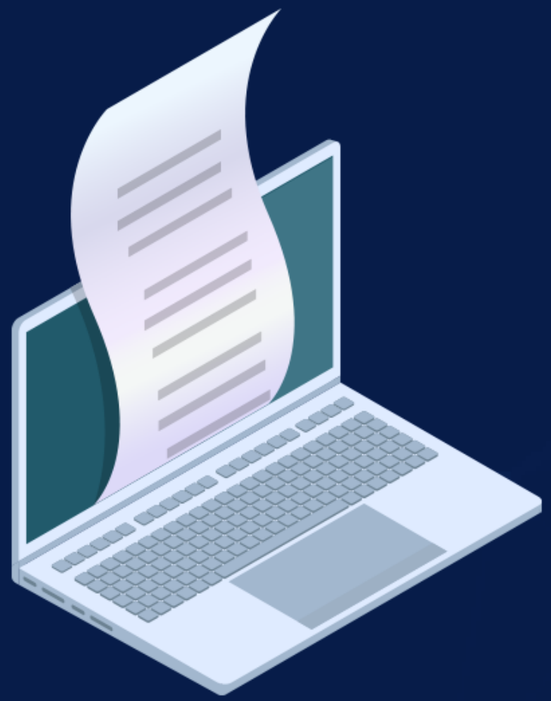
Project Outcome



- This project is designed to help gain an understanding of working with the error log of a website and user interaction on different modules of a website.

- You should be able to analyze the dataset for this project to create a report. You will be able to use PySpark, do analyses, and obtain the desired results.

Submission Process



1. Complete the project in the Simplilearn lab
2. Complete each task listed in the problem statement
3. Take screenshots of the results for each question and the corresponding code
4. Save it as a document and submit it using the assessment tab
5. Tap the "Submit" button (this will present you with three choices)
6. Attach three files and then click "Submit“

- **Note:** Be sure to include screenshots of the output

Thank You