# Big Data Hadoop and Spark Developer

Course-End Project Problem Statement

# Apache Server Log Analysis

**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 all are open source. You are a data engineer, who works for a start-up named **Hadoop Analytics**, which serves major clientele.

You have been assigned 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. Details on the resources that were used
2. The identity of the person who accessed the logs
3. The date and time the logs were accessed
4. Specifics on any problems that emerge
5. Information about the final product

**Objective:**
Perform server log analysis to assist businesses in identifying and analyzing critical business errors as well as potential customers and their domains

**Dataset to be Used:** server-access-log.txt

**Dataset Description:**
**Snippet:**
10.1.2.3 - rehg [10/Nov/2021:19:22:12 -0000] "GET /sematext.png HTTP/1.1" 200 3423

**The following elements are present in the dataset:**

1. **%h**: resolved into **10.1.2.3** – the IP address of the remote host that made the request.

2. **%l**: remote log name provided by **identd**, the hyphen is provided, which is a

value that can be logged when the information provided by the logging directive is not found or can't be accessed.

3. **%u**: resolved into **rehg**, the user identifier determined by the HTTP authentication.

4. **%t**: the date and time of the request with the time zone, in the above case it is **[10/Nov/2021:19:22:12 -0000]**

5. \"%r\": the first line of the request inside double quotes, in the above case it is: **"GET /sematext.png HTTP/1.1"**

6. **%>s**: the status code reported to the client. This information is crucial because it determines whether the request was successful or not.

7. **%b**: the size of the object sent to the client, in our case the object was the **sematext.png** file and its size was **3423** bytes.

**Steps Overview:**
**Step 1:** Upload the **server-access-log** file to the HDFS

**Step 2:** Perform the below tasks on the uploaded dataset

1. Status code analysis
2. Arrange the result in descending order
3. Frequent visitor recognition
4. Missing URL recognition
5. Traffic recognition
6. Endpoint recognition

**Task to be performed:**
**Task 1:** Status code analysis
**Aim**: To analyze which status code values appear how many times
**Steps to perform:**

1. Read the log file as an RDD in PySpark
2. Consider the sixth element as it is **request type**
   a. Split the line by space which will return an array
   b. Consider the sixth element of the array which is the request type.
   c. Replace the **single quote** with blank

3. Convert each word into a tuple of (word,1)
4. Apply **reduceByKey** transformation to count the values for the same key
5. Display the data stored in the RDD

**Task 2:** Arrange the result in descending order
**Aim:** To sort the result of **Task 1** in descending order by request count
**Steps to perform:**

1. Sort RDD using the **sortBy** function
2. Display the sorted data stored in the RDD

**Task 3:** Frequent visitor recognition
**Aim:** To identify the top 10 frequent visitors of the website
**Steps to perform:**

1. Read the log file as an RDD in pySpark
2. Identify the frequent visitors using the map function
3. Display the data stored in the RDD

**Task 4:** Missing URL recognition
**Aim:** To identify the top 10 missing (does not exist) URLs
**Steps to perform:**

1. Read the log file as an RDD in pySpark
2. Identify the URLs for which the server is returning the 404-request code
3. Display the data stored in the RDD

**Task 5:** Traffic recognition
**Aim:** To identify the traffic (total number of HTTP requests received per day)
**Steps to perform:**

1. Read the log file as an RDD in pySpark
2. Fetch the DateTime string and replace **[** with blank
3. Get the Date string from the DateTime
4. Identify HTTP requests using the map function
5. Display the data stored in the RDD

**Task 6:** Endpoint recognition
**Aim:** To identify the top 10 endpoints that transfer maximum content in megabytes
**Steps to be perform:**

1. Read the log file as an RDD in pySpark

2. Identify a maximum number of endpoints using the map function
3. Display the data stored in the RDD