**Objective:**

The objective is to enable participants to get a feel of working with production datasets and analyzing data using tools in Hadoop ecosystem.

This POC requires extracting required columns from Web Server logs, processing data using tools like Map Reduce, Hive, Apache Spark & Kafka and making the data available in HBase for access by front-end applications.

**Input Dataset:**

**Input dataset is a Web Server log which has the following details:**

IP Address,

Timestamp,

Request Type,

Resource Requested,

Request Status,

Total bytes,

Referrer URL, Search key words

User browser

**Problem Statement:**

1] Map Reduce Assignment:

Create Java Map Reduce Programs to:

1. Find total entries from IP Addresses for each request status (Job1)

IP request status count

192.168.1.1 200 100

192.168.1.1 400 80

192.168.1.1 300 50

1. Sort the output of previous assignment by IP Address in ascending order and request status count in descending order (Job2).

(Note: Use Secondary sort algorithm for sorting)

1. Store the intermediate data between Job1 and Job2 in sequence files

SequenceFileOutputFormat (for Job1)

SequenceFileInputFormat (for Job2)

2] Apache Pig and Hive Assignment:

1. Parse the web log and extract the following columns (using Apache Pig)

IP Address, Date, resource requested, request status

1. Create Hive tables to store following data partitioned by **year** and **month**

IP Address, Date, resource requested, request status

1. Write a query to find **Top 5** IP addresses having request status >= 400 (by transaction count) each month-year

3] Apache Spark / Kafka / HBase Assignment:

1. Parse the web log and extract following columns (using spark)

IP Address, Date, resource requested, Request Status

1. Create a kafka topic called **WebLogEntry** and publish records to the topic
2. Create a Spark Kafka consumer which will read the messages from Kafka topic and
3. Populate a HBase table called WebLogEntries if Request status < 400
4. Save records to a output directory called Rejects if Request status >= 400