Secure HDFS setup and access

Use a Linux environment (VM or bare-metal) to explore, configure and install the following:

1. Set up a kerberos server (KDC) which follows Kerberos version 5 protocol
2. Set up Hadoop Distributed File System (HDFS) that runs in cluster mode (1 node). Hadoop version please use v3.2.1. Hadoop using YARN as resource manager is optional, but is a plus
3. Configure HDFS running in security mode, i.e., support only Kerberos and Token as access authentication
4. Configure HDFS with user- and group- access control, i.e. every non-root user can only read/write data based on his own user ID and user group
5. Configure HDFS to be remote accessible
6. When accessing HDFS remotely, a correct keytab from KDC must be used
7. Write a java or python code to connect to HDFS enforcing Kerberos authentication

Please provide a report on the steps taken, problem encountered, your solutions (even they are not fully successful), and the java/python code that you implemented

Log analysis and result visualization

You are given a log file generated by a code repository management system that records multiple user login and code commitments.

Part 1. Data Extraction from log file

You should extract the following information:

1. General information: IP address, timestamp, Web Browser information. E.g.

```
172.27.234.3 - - [13/Oct/2019:02:58:15 +0000] "GET / HTTP/1.1" 302 99 ""
"Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/79.0.3928.4 Safari/537.36"
```

You should extract:

IP: 172.27.234.3

Timestamp: 13/Oct/2019:02:58:15

Web Browser: Mozilla/5.0 (Windows NT 10.0; Win64; x64)

2. Project information: For user committed projects, extract username, and project name. e.g.

```
172.27.235.75 - - [30/Oct/2019:05:12:45 +0000] "GET /abc-ll/ccc-
DB/commits/905ed3275b96ecc9e19e94300f0c1a947d46c12f/sign
atures HTTP/1.1" 200 17 "http://172.28.225.2/abc-ll/ccc-DB/tree/master/cccDbTest"
"Mozilla/5.0 (Macintosh; Intel Mac OS
X 10_14_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/77.0.3865.120
Safari/537.36"
```

User name: abc-ll

Project name: ccc-DB

Note:

a. http://172.28.225.2/abc-ll/ccc-DB the format is http://IP_ADDRESS/USER_NAME/PROJECT_NAME
b. The following patterns are NOT considered as projects. You should exclude these patterns: “users”, “assets”, “uploads”, “admin”, “dashboard”, etc.

```
172.27.234.3 - - [13/Oct/2019:02:58:15 +0000] "GET /users/sign_in HTTP/1.1" 200
4307 "" "Mozilla/5.0 (Windows NT 10.0; W
in64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/79.0.3928.4
Safari/537.36"
```

```
172.27.234.3 - - [13/Oct/2019:02:58:16 +0000] "GET
/assets/webpack/commons~pages.ldap.omniauth_callbacks~pages.omniauth_
callbacks~pages.sessions~pages.sessions.new.e8265136.chunk.js HTTP/1.1" 200 3635
"http://172.28.225.2/users/sign_in" "Mo
```

```
zilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/79.0.3928.4 Safari/537.36"
```

```
172.27.234.3 - - [13/Oct/2019:02:58:33 +0000] "GET /uploads/-
/system/group/avatar/47/aaa-logo.png?width=15 HTTP/1.1" 200
 45189 "http://172.28.225.2/aaa/connection-manager" "Mozilla/5.0 (Windows NT 10.0;
Win64; x64) AppleWebKit/537.36 (KHTML
, like Gecko) Chrome/79.0.3928.4 Safari/537.36"
```

```
172.27.234.13 - - [14/Oct/2019:02:35:25 +0000] "GET /admin/users HTTP/1.1" 200
11868 "http://172.28.225.2/admin" "Mozill
a/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/79.0.3938.0 Safari/537.36"
```

```
172.27.234.14 - - [14/Oct/2019:02:44:57 +0000] "GET /dashboard/projects HTTP/1.1"
200 10778 "http://172.28.225.2/aaa/aaa
_sing_ml" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/77.0.3865.90 Safari/5
37.36"
```

Please write a code in Python or Java, and stores the information in database (save in CSV is also acceptable)

Part2: Data Visualization

Plot the following graphs:

1. For the last 7 days, the unique IP addresses that are recorded in the system. The plotting is in Pie chart, the size denotes the percentage of the total count of the IP over all IPs. Only those total count greater than 5% of the total circle should be plotted (do not plot the tiny regions). It will be a plus if mouse hovering is implemented, where the details (IP address, and total count) will be shown alongside the pie sector
2. For the last 30 days, the users that are active. “Active” means that user should upload the code at least three times per month. Use Pie chart to plot this. Only those total count greater than 5% of the total circle should be plotted (do not plot the tiny regions).
3. Continue with the plotting in item 2: when click the section, it should plot a bar graph, which plots for this user, across the last 30 days, the total count of code upload on each day. In the plot, it should also show the total number of unique project that the user upload on this day.