|  |  |
| --- | --- |
| **Assignment 1** | |
|  | Download the Batting.csv dataset from WebCampus. |
|  | Connect to FDU VPN. |
|  | Login to Ambari by navigating to <http://132.238.7.193:8080> |
|  | Enter your user name and password. |
|  | Using Ambari, start the HDFS view. |
|  | In HDFS view, navigate to your user folder, which is in the form  **/usr/stXX**  where stXX is your user name. |
|  | Using Ambari, upload Batting.csv to your folder in HDFS |
|  | Open the Pig View. |
|  | Substitute st65 in the script below with your user name. |
|  | **Run the script listed below without Tez** |
|  | batting = LOAD '/user/st65/Batting.csv' USING PigStorage(',');  raw\_runs = FILTER batting BY $1>0;  runs = FOREACH raw\_runs GENERATE $0 AS playerID, $1 AS year, $8 AS runs;  grp\_data = GROUP runs BY (year);  max\_runs = FOREACH grp\_data GENERATE group as grp, MAX(runs.runs) AS max\_runs;  join\_max\_runs = JOIN max\_runs BY ($0, max\_runs), runs BY (year, runs);  join\_data = FOREACH join\_max\_runs GENERATE $0 AS year, $2 AS playerID, $1 AS runs;  DUMP join\_data; |
|  | Download Results |
|  | Download Logs |
|  | In the log file find out date and time when the script started and ended execution. Calculate how long it took the script to complete. |
|  | Now run the same script **with** Tez by checking the “Execute on Tex” box and running the script. |
|  | Download Results |
|  | Download Logs |
|  | From the log file, find out how long it took the script to complete the tasks. |
|  | Find out if Tez allows the Pig script to execute faster and by how much. Explain why. |

Course Project. Pig and PigLatin.

Assignment 2 (for extra credit)

|  |  |
| --- | --- |
| 1. | Modify the script above so that the raw\_runs variable will contain rows for a particular year. Find year by your user ID in the mapping table below. |
| 2. | Display the contents of raw\_runs on the screen. |

Assignment 3 (for extra credit)

|  |  |
| --- | --- |
| 1. | Continue from Assignment 1 for Extra Credit. |
| 2. | Use the CONCAT function as described here <http://pig.apache.org/docs/r0.15.0/func.html#eval-functions> , concatenate first 4 columns in raw\_data using the underscore symbol ‘\_’ as a separator. |
| 3. | Display the results. |

User ID to Year Mapping (to be used for Assignment 2 and 3).

|  |  |
| --- | --- |
| **User ID** | **Year** |
| st1 | 1871 |
| st2 | 1872 |
| st3 | 1873 |
| st4 | 1874 |
| st5 | 1875 |
| st6 | 1876 |
| st7 | 1877 |
| st8 | 1878 |
| st9 | 1879 |
| st10 | 1880 |
| st11 | 1881 |
| st12 | 1882 |
| st13 | 1883 |
| st14 | 1884 |
| st15 | 1885 |
| st16 | 1886 |
| st17 | 1887 |
| st18 | 1888 |
| st19 | 1889 |
| st20 | 1890 |
| st21 | 1891 |

|  |  |
| --- | --- |
| **User ID** | **Year** |
| st22 | 1892 |
| st23 | 1893 |
| st24 | 1894 |
| st25 | 1895 |
| st26 | 1896 |
| st27 | 1897 |
| st28 | 1898 |
| st29 | 1899 |
| st30 | 1900 |
| st31 | 1901 |
| st32 | 1902 |
| st33 | 1903 |
| st34 | 1904 |
| st35 | 1905 |
| st36 | 1906 |
| st37 | 1907 |
| st38 | 1908 |
| st39 | 1909 |
| st40 | 1910 |
| st41 | 1911 |
| st42 | 1912 |
| st43 | 1913 |
| st44 | 1914 |
| **User ID** | **Year** |
| st45 | 1915 |
| st46 | 1916 |
| st47 | 1917 |
| st48 | 1918 |
| st49 | 1919 |
| st50 | 1920 |
| st51 | 1921 |
| st52 | 1922 |
| st53 | 1923 |
| st54 | 1924 |
| st55 | 1925 |
| st56 | 1926 |
| st57 | 1927 |
| st58 | 1928 |
| st59 | 1929 |
| st60 | 1930 |
| st61 | 1931 |
| st62 | 1932 |
| st63 | 1933 |
| st64 | 1934 |
| st65 | 1935 |