Big Data Infrastructures – Fall 2018
HDFS + Hbase Homework

The dataset provided in this lab, **weblogs_hbase.txt**, contains web server logs. Each line is composed as following: IP address, year and number of visits during each month.

Task 1: Copy the Weblogs dataset to HDFS folder: /bdi_2018/<YOUR_USERNAME>
Task 2: Write Java code to read and store the data as follows (HINT: use *Create_Hbase_Table.java*):

- Connect to Hbase and create the table "weblogs_<YOUR_USER>"
- Create two column families: 'Months' and 'Statistics'
- Read Weblogs data from HDFS
- Write the data into the Hbase table as follows: The row key is the concatenation of IP address and the year. Insert the data into the table as shown in the below example:

	Months											**Statistics**
RowKey	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Oct**	**Nov**	**Dec**	**Active**
88.87.2.192 |2011	12	1	1	5	1	2	1	15	7	8	3	1

- If the visit value is greater than zero, create month columns in column family "Months" and insert the visit value. Do not insert the number of visits equal to 0 (as shown in the above example: There is no 'Sep' column).
- Create column Active in column family "Statistics", and insert 1 if the total number of visits is greater than 20, otherwise insert 0

Task 3: Execute the following queries using Hbase shell:

- Retrieve the entire content of the table using the "scan" command
- Restrict the scan results to retrieve only keys between "'0.32.85.668|2012" and "'01.660.68.623|2012"
- Count the total number of web logs in the table

Task 4: Write Java code to perform the following queries (HINT: use *Hbase_Queries.java*):

- Retrieve only the contents of the Columns: "Jan" and "Feb" from the row key: 06.305.307.336|2012
- Create a new ip and year, and fill in the table with the same values as the row with key: 01.660.70.74|2012
- Delete the row with key: 88.88.324.601|2012

Deliverables:

1. Java code to read and store the data (Task 2)
2. Snapshots of Hbase queries (Task 3)
3. Java code containing the queries (Task 4)

Send the solution before **midnight 28th of November 2018 to rana@exascale.info**