**General instructions**

I. Your notebook should consist of 2 main parts.

1. **The HiveQL code.** You should write the code of your query in a file using [%%writefile](http://ipython.readthedocs.io/en/stable/interactive/magics.html#cellmagic-writefile) magic command. It isn't necessary to write all query in 1 cell. You can use "%%writefile -a <my\_file.hql>" and append the file.
2. **Executing the query.** When your query is successfully written into the file, you should execute it. You can do this using the following code:

! hive-f<my\_file.hql>

The executing code must be the last cell of your notebook.

II. Please, don't create the Hive database in the notebook will be submitted. You can create it in the terminal or in the other notebook. In the submission notebook you should use only

USE <some\_database>;

III. When you create managed tables, please you default location.

**Dataset's description**

The purpose of this assignment is to calculate some statistics of the stackoverflow.com website.

The input dataset contains two parts: users and posts (questions and answers) from StackOverflow site. Posts are in XML format, but it doesn't matter: you can interpret it as a text of independent lines, one post per each line.

Line example:

<row Id="11" PostTypeId="1" AcceptedAnswerId="1248" CreationDate="2008-07-31T23:55:37.967" Score="1106" ViewCount="115048" Body="&lt;p&gt;Given a specific &lt;code&gt;DateTime&lt;/code&gt; value, how do I display relative time, like:&lt;/p&gt;&#xA;&#xA;&lt;ul&gt;&#xA;&lt;li&gt;2 hours ago&lt;/li&gt;&#xA;&lt;li&gt;3 days ago&lt;/li&gt;&#xA;&lt;li&gt;a month ago&lt;/li&gt;&#xA;&lt;/ul&gt;&#xA;" OwnerUserId="1" LastEditorUserId="1136709" LastEditorDisplayName="user2370523" LastEditDate="2015-12-29T02:08:37.450" LastActivityDate="2016-07-13T23:23:58.537" Title="How can relative time be calculated in C#?" Tags="&lt;c#&gt;&lt;datetime&gt;&lt;datediff&gt;&lt;relative-time-span&gt;" AnswerCount="33" CommentCount="4" FavoriteCount="508" CommunityOwnedDate="2009-09-04T13:15:59.820" />

This line corresponds to the real post here: <https://stackoverflow.com/questions/11/calculate-relative-time-in-c-sharp>

So, the lines not started with the "row" tags should be ignored. The valid row contains the following fields and their order is not defined:

* Id (integer) - id of the post
* PostTypeId (integer: 1 or 2) - 1 for questions, 2 for answers
* CreationDate (date) - post creation date in the format "YYYY-MM-DDTHH:MM:SS.ms"
* Tags (string, optional) - list of post tags, each tag is wrapped with html entities '&lt;' and '&gt;'
* OwnerUserId (integer, optional) - user id of the post's author
* ParentId (integer, optional) - for answers - id of the question
* Score (integer) - score (votes) of a question or an answer, can be negative (!)
* FavoriteCount (integer, optional) - how many times the question was added in the favorites

The second part of the dataset contains StackOverflow users. The format is similar to the posts, example:

<row Id="4" Reputation="26638" CreationDate="2008-07-31T14:22:31.317" DisplayName="Joel Spolsky" LastAccessDate="2016-12-10T22:12:46.367" WebsiteUrl="http://www.joelonsoftware.com/" Location="New York, NY" AboutMe="&lt;p&gt;I am:&lt;/p&gt;&#xA;&#xA;&lt;ul&gt;&#xA;&lt;li&gt;the co-founder and CEO of &lt;a href=&quot;http://stackexchange.com">Stack Exchange&lt;/a&gt;&lt;/li&gt;&#xA;&lt;li&gt;the co-founder of &lt;a href=&quot;http://www.fogcreek.com" rel=&quot;nofollow&quot;&gt;Fog Creek Software&lt;/a&gt;&lt;/li&gt;&#xA;&lt;li&gt;the creator and chairman of the board of &lt;a href=&quot;http://trello.com" rel=&quot;nofollow&quot;&gt;Trello&lt;/a&gt;&lt;/li&gt;&#xA;&lt;li&gt;owner of Taco, the most famous Siberian Husky on the Upper West Side.&lt;/li&gt;&#xA;&lt;/ul&gt;&#xA;&#xA;&lt;p&gt;You can find me on Twitter (as &lt;a href=&quot;http://twitter.com/spolsky" rel=&quot;nofollow&quot;&gt;@spolsky&lt;/a&gt;) or on my rarely-updated blog, &lt;a href=&quot;http://joelonsoftware.com" rel=&quot;nofollow&quot;&gt;Joel on Software&lt;/a&gt;.&lt;/p&gt;&#xA;" Views="66761" UpVotes="779" DownVotes="96" ProfileImageUrl="https://i.stack.imgur.com/C5gBG.jpg?s=128&g=1" AccountId="4" />

This line is an account of Joel Spolsky: <https://stackoverflow.com/users/4/joel-spolsky>

The fields are the following and their order is also not defined:

* Id (integer) - user id
* Reputation (integer) - user's reputation
* CreationDate (string) - creation date in the format "YYYY-MM-DDTHH:MM:SS.ms"
* DisplayName (string) - user's name
* Location (string, options) - user's country
* Age (integer, optional) - user's age

**Hints for the Task1:**

1. To create a regular expression, which describes strings containing two patterns, where order of the patterns is not defined use the following so-called ‘positive lookahead assertion’ with “?=” group modifier. For example, both strings “Washington Irving” and “Irving Washington” match the pattern: “(?=.\*Washington)(?=.\*Irving)”.
2. To capture groups use round brackets. So, the pattern: “(?=.\*(Washington))(?=.\*(Irving))” captures “Washington” and “Irving” from both strings: "William Arthur Irving Washington was an English first-class Cricketer" and: “Washington Irving was an American writer”.
3. If any group is optional, use “?” modifier. So, the pattern: “(?=.\*(Washington))(?=.\*(Irving))?” captures “Washington” from the string “George Washington”.
4. Use ‘\b’ to specify boundaries of words and increase accuracy of your pattern. For example: pattern “(?=.\*\bID=(\d+))(?=.\*\bUserID=(\d+))” captures “1” and “2” from the string "ID=1 UserID=2", whereas pattern without ‘\b’: “(?=.\*ID=(\d+))(?=.\*UserID=(\d+))” returns the wrong groups: “2” and “2”.
5. In Hive pattern for the external table in SERDEPROPERTIES "input.regex" should describe the whole input string, so add “.\*” at the end of the pattern.
6. You can test your pattern with Hadoop Streaming, using a simple mapper which outputs the input strings not matched to your pattern. Analyse the result to improve the pattern. Escape backslashes and double quotes and the pattern from the mapper can be used in SERDEPROPERTIES "input.regex" variable unchanged. Example of the mapper:

#!/usr/bin/env python

import sys

import re

regexp = '.\*?(?=.\*\\bID=\"(\\d+)\")(?=.\*\\bUserID=\"(\\d+)\").\*'

for line in sys.stdin:

if not re.match(regexp, line.strip()):

print line.strip()

**For the tasks 2, 3, etc.** use the tables 'posts' and 'users' from the database *'stackoverflow\_'*. 'Posts' is partitioned by year and by month, columns correspond to the fields of the lines described above:

* id INT
* post\_type\_id TINYINT
* date STRING
* owner\_user\_id INT
* parent\_id INT
* score INT
* favorite\_count INT
* tags ARRAY<STRING>
* year STRING (partition)
* month STRING (partition)

'Users' contains the following columns:

* id INT
* reputation INT
* creation\_date STRING
* display\_name STRING
* location STRING
* age INT

Proceed with Hive assignment tasks.

If you want to deploy the environment on your own machine, please use [bigdatateam/hive-course2](https://hub.docker.com/r/bigdatateam/hive-course2/) Docker container.