Stack Exchange Data Analysis

Introduction

Stack Exchange has released its data-dumps of all its publically available contents including all the stack exchange communities like stack Overflow, Super user, Ask Ubuntu, Server Fault, etc. The data includes posts, users, comments, badges, postfeedbacks, postHistory, postTags, postTypes, reviewTasks, tags, votes, etc. Analyze Stack Exchange data and generate insights. Process posts data and develop program to solve below mentioned problems.

Data Download Link

https://archive.org/details/stackexchange
https://archive.org/download/stackexchange
https://www.dropbox.com/s/i5dfouinpxpyj9l/Posts.xml?dl=0

Sample Post Data

<row Id="41" PostTypeId="1" AcceptedAnswerId="44" CreationDate="2014-05-14T11:15:40.907" Score="28" ViewCount="1897" Body="<p>R has many libraries which are aimed at Data Analysis (e.g. JAGS, BUGS, ARULES etc..), and is mentioned in popular textbooks such as: J.Krusche, Doing Bayesian Data Analysis; B.Lantz, "Machine Learning with R".</p>

<p>I've seen a guideline of 5TB for a dataset to be considered as Big Data.</p>

<p>My question is: Is R suitable for the amount of Data typically seen in Big Data problems?
Are there strategies to be employed when using R with this size of dataset?</p>
" OwnerUserId="136" LastEditorUserId="118" LastEditDate="2014-05-14T13:06:28.407" LastActivityDate="2015-04-12T05:00:23.663" Title="Is the R language suitable for Big Data" Tags="<bigdata><r>" AnswerCount="8" CommentCount="1" FavoriteCount="13" />

Format of Posts Data

- Id
- PostTypeId (listed in the PostTypes table)
  1. Question
  2. Answer
  3. Orphaned tag wiki
  4. Tag wiki excerpt
  5. Tag wiki
  6. Moderator nomination
  7. “Wiki placeholder” (seems to only be the election description)
  8. Privilege wiki
- AcceptedAnswerId (only present if PostTypeId is 1)
- Parent ID (only present if PostTypeId is 2)
- CreationDate
- DeletionDate (only non-null for the SEDE PostsWithDeleted table. Deleted posts are not present on Posts. Column not present on data dump.)
- Score
- ViewCount (nullable)
- Body (as rendered HTML, not Markdown)
- OwnerUserId (only present if user has not been deleted; always -1 for tag wiki entries, i.e. the community user owns them)
- OwnerDisplayName (nullable)
- LastEditorUserId (nullable)
- LastEditorDisplayName (nullable)
- LastEditDate="2009-03-05T22:28:34.823" - the date and time of the most recent edit to the post (nullable)
- LastActivityDate="2009-03-11T12:51:01.480" - the date and time of the most recent activity on the post. For a question, this could be the post being edited, a new answer was posted, a bounty was started, etc.
- Title (nullable)
- Tags (nullable)
- AnswerCount (nullable)
- CommentCount
- FavoriteCount
- ClosedDate (present only if the post is closed)
- CommunityOwnedDate (present only if post is community wikied)

KPIs

1. Count the total number of questions in the available data-set and collect the questions id of all the questions
2. Monthly questions count – provide the distribution of number of questions asked per month
3. Provide the number of posts which are questions and contains specified words in their title (like data, science, nosql, hadoop, spark)
4. The trending questions which are viewed and scored highly by the user – Top 10 highest viewed questions with specific tags
5. The questions that doesn't have any answers – Number of questions with "0" number of answers
6. Number of questions with more than 2 answers
7. Number of questions which are active for last 6 months
8. Questions which are marked closed for each category – provide the distribution of number of closed questions per month
9. The most scored questions with specific tags – Top 10 questions having tag hadoop, spark
10. List of all the tags along with their counts
11. Number of question with specific tags (nosql, big data) which was asked in the specified time range (from 01-01-2015 to 31-12-2015)
12. Average time for a post to get a correct answer