Cloud Computing

MapReduce Solutions to Typical Big Data Analytics Problems

Dell Zhang

Birkbeck, University of London

2018/19

StackExchange

Questions

Stack Overflow is a question and answer site for professional and enthusiast programmers. It's 100% free, no registration required.

Here's how it works:

Anybody can ask a question

Anybody can answer

The best answers are voted up and rise to the top

posts

```
<row Id="6939296" PostTypeId="2" ParentId="6939137"
CreationDate="2011-08-04T09:50:25.043" Score="4" ViewCount=""
Body="&lt;p&gt;You should have imported Poll with &lt;code&gt;
from polls.models import Poll&lt;/code&gt;&lt;/p&gt;&#xA;"
OwnerUserId="634150" LastActivityDate="2011-08-04T09:50:25.043"
CommentCount="1" />

<row Id="6939304" PostTypeId="1" AcceptedAnswerId="6939433"
CreationDate="2011-08-04T09:50:58.910" Score="1" ViewCount="26"
Body="&lt;p&gt;Is it possible to gzip a single asp.net 3.5 page? my
site is hosted on IIS7 and for technical reasons I cannot enable gzip
compression site wide. does IIS7 have an option to gzip individual pages or
will I have to override OnPreRender and write some code to compress the
output?&lt;/p&gt;&#xA;" OwnerUserId="743184"
LastActivityDate="2011-08-04T10:19:04.107" Title="gzip a single asp.net page"
Tags="&lt;asp.net&gt;&lt;iis7&gt;&lt;gzip&gt;"
AnswerCount="2" />
```

comments

```
<row Id="2579740" PostId="2573882" Text="Are you getting any results? What
are you specifying as the command text?" CreationDate="2010-04-04T08:48:51.347"
UserId="95437" />
```

users

```
<row Id="352268" Reputation="3313" CreationDate="2010-05-27T18:34:45.817"
DisplayName="orangeoctopus" EmailHash="93fc5e3d9451bcd3fdb552423ceb52cd"
LastAccessDate="2011-09-01T13:55:02.013" Location="Maryland" Age="26"
Views="48" UpVotes="294" DownVotes="4" />
```

Outline

- Data Summarisation
- Data Filtering
- Data Organisation
- Data Joining

Data Summarisation

- Numerical Summarisation
  - Count the number of users from each state.

Data Summarisation

- Numerical Summarisation
  - Given a list of user's comments, determine the first and last time each user commented and the total number of comments from that user.
  - Given a list of user's comments, determine the mean and standard deviation of comment lengths per hour of day.
  - Given a list of user's comments, determine the median of comment lengths per hour of day.

Data Summarisation

- Inverted Indexing
  - Given a set of user's comments, build an inverted index of Wikipedia URLs to a set of answer post IDs .

Data Filtering

- Distributed Grep
  - We'd like to parallelize the regular expression search across a larger body of text.
- Random Sampling
  - We'd like to grab a subset of our larger data set in which each record has an equal probability of being selected.

Data Filtering

- Top-K
  - Given a list of user information, output the information of the top ten users based on reputation.

Data Filtering

- Distinct
  - Given a list of user's comments, determine the distinct set of user IDs.

Data Organisation

- Relational (Structured) to Hierarchical
  - Given a list of posts and comments, create a structured XML hierarchy to nest comments with their related post.
  - Given the output of the previous example, perform a self-join operation to create a question, answer, and comment hierarchy.

Data Organisation

- Partitioning
  - Given a set of user information, partition the records based on the year of last access date, one partition per year.
- Binning
  - Given a set of StackOverflow posts, bin the posts into four bins based on the tags hadoop, pig, hive, and hbase. Also, create a separate bin for posts mentioning hadoop in the text or title.

Data Organisation

- Total Order Sorting
  - The user data in our StackOverflow data set is in the order of the account's creation. Instead, we'd like to have the data ordered by the last time they have visited the site.

Data Organisation

- Shuffling
  - Given a large data set of StackOverflow comments, anonymize each comment by removing IDs, removing the time from the record, and then randomly shuffling the records within the data set.

Data Joining

- Reduce-Side Join
  - Given a set of user information and a list of user's comments, enrich each comment with the information about the user who created the comment.

Data Joining

- Map-Side Join (Composite Join)
  - Given two large formatted data sets of user information and comments, enrich the comments with user information data.

Data Joining

- In-Memory Join (Replicated Join)
  - Given a small set of user information and a large set of comments, enrich the comments with user information data.

Data Joining

- Cartesian Product
  - Given a groomed data set of StackOverflow comments, find pairs of comments that are similar based on the number of like words between each pair.