



Hbase



Winnie Jao, XuanZhe Li, Saral Bhagat,

Daniel Margason, Andrew Gonser



Queries

“If you want the answer—ask the question.”

— Lorii Myers

- a. Find out the paper with the most co-author
 - b. Find out how many third-level co-authors does David DeWitt have.
 - c. Which proceeding in 2005 had the most distinct number of authors?
 - d. Find out at what level is Moshe Vardi from Joseph M. Hellerstein.
-

Data

“War is ninety percent data.”
-Napoleon Bonaparte

SQLShare:

DBLP_ATTRIBUTES

DBLP_LINKS

DBLP_OBJECTS

Implementation Overview

- Database Model: HBase
 - API: Happybase
 - Environment: Amazon EC2
 - Programming language: Python
-

Changes to our data model

Discard `publish_year`, `publish_in`, `publish_type` and `link_id` from `paper_info` table.
Add new `paper_title` to `paper_info` table.

Paper_info					
<u>paper_id</u>	link_id	publish_year	publish_type	publish_in	author_list



Paper_info		
<u>paper_id</u>	paper_title	author_list

Changes to our data model (cont.)

Discard `proceeding_id` from `journal_info` table.

Journal_info		
<u>journal_name</u>	proceeding_id	paper_list



Journal_info	
<u>journal_name</u>	paper_list

Changes to our data model (cont.)

Add author_name to author_info table.

auhtor_info	
<u>author_id</u>	paper_list



author_info		
<u>author_id</u>	author_name	paper_list

Final Data Model

Paper_info		
<u>paper_id</u>	paper_title	author_list

author_info		
<u>author_id</u>	author_name	paper_list

Journal_info	
<u>journal_name</u>	paper_list

Proceeding_info			
<u>proceeding_id</u>	<u>proceeding_name</u>	year	paper_list

Demo



Find the paper with most co-authors

1. Open connection
2. Loop through each paper, parse the author list
3. Keep track of the count of co-authors of each paper, and the maximum count of co-authors
4. Display the paper title and list of author names of paper that has maximum count of co-authors

Result:

```
connection = happybase.Connection(host="localhost", port=9090, autoconnect=False)
connection.open()
# data table
table = connection.table('paper_info')
author_table = connection.table('author_info')
# dictionary to store paper_id and number of co-author
count_table = {}
# keep track of max number of co-author
max_count = 0
# build the dictionary
for x in range(1000):
    X = str(x+1)
    row = table.row(X, columns=['paper:author_list'])
    if bool(row):
        # parse author list to count how many co-author this paper has
        author_list = row['paper:author_list']
        parse = author_list.split(',')
        count = len(parse)
        count_table[X] = count
        if count > max_count:
            max_count = count

print "The following papers has maximum number of co-author, " + str(max_count) + '\n'

for ID, count in count_table.iteritems():
    if count == max_count:
        row = table.row(ID, columns=['paper:paper_title', 'paper:author_list'])
        print "paper title: " + row['paper:paper_title']
        author_list = row['paper:author_list']
        parse = author_list.split(',')
        author_list = ""
        for author in parse:
            author_row = author_table.row(author, columns=['data:author_name'])
            if bool(author_row):
                author_name = author_row['data:author_name']
                if author_list != "":
                    author_list = author_list + ", " + author_name
                else:
                    author_list = author_name

print "author list: " + author_list + '\n'
```

The following papers has maximum number of co-author, 8

paper title: Taggers for Parsers.

author list: John Adcock, Anthony R. Cassandra, Yoshihiko Gotoh, Eugene Charniak, Jeremy Katz, Glenn Carroll, Michael L. Littman, John McCann

paper title: Multiprocessing.

author list: John D. McGregor, David Hemmendinger, Virgil Wallentine, Arthur M. Riehl, Carolyn McCreary, Roy P. Pargas, Charles J. Fleckenstein, Helen Gill

Which proceeding in 2005 had the most distinct number of authors?

- Connect to database
- Access `proceeding_info`, Finding all matching proceeding in 2005, then parsing the value in `paper_list` of those rows into a list object
 - For each paper object in list, comparing the `paper_id` in `paper_info` table, find distinct authors and count
 - Hold the maximum count value
- The result is shown as following:

```
The following proceedings in 2005 has distinct author count of 1395  
proceeding name: HICSS
```

What would we change next time?

“Be the change that you wish to see in the world.”

— Mahatma Gandhi

- Try to load all the data into Hbase so we can have enough data to complete each query.
 - Try additional technologies to make the queries easier, such as HiveQL
-

Lessons Learned

“When life brings you full circle, pay attention.
There's a lesson there.”

— Mandy Hale, *Life, Love, and a Dash of Sass*

- Start early
- Do more research on new tools for Hbase



THANK YOU!