

Google Query



Web UI Tutorial & Managing
Stackoverflow Database

Athena Li
Lancy Mao
Elena Lopez

Agenda

1. Introduction

2. Basic Tutorial

3. Specific Application

4. Question & Answer



Introduction



Google BigQuery

Introduction

What is BigQuery?

- Externalization of Dremel
 - High performance and scalable query engine on the cloud
- First to identify the drawbacks of MapReduce
 - Not ideal to query large, distributed datasets in real time
 - Extremely slow
- Supports a subset of SQL for querying and retrieving data
- End of the Big Data pipeline

BigQuery Strengths

- Deliver blazing fast query performance on distributed data sets that are stored across thousand of servers
- Interactive data analysis tool for large datasets
- Explore record level data in real time



Google BigQuery

Basic Tutorial



Google BigQuery

Create project

The screenshot shows the Google Cloud IAM & admin interface for the "Big data project". The left sidebar lists various IAM-related options: IAM, Identity, Quotas, Service accounts, Labels, Settings, Encryption keys, Identity-Aware Proxy, and Roles. The "IAM" option is selected and highlighted in blue. The main content area displays the "Permissions for project 'Big data project'". It includes a note about granting permissions to the entire project and its resources, mentioning beta roles and deprecated features. A search bar and a "View by" dropdown set to "Members" are also present. A table lists one member: LANCYMAO9@gmail.com, who is assigned the "Owner" role. A red box highlights the "Manage resources" button at the bottom of the sidebar.

Google APIs Big data project

IAM & admin

IAM

+ ADD - REMOVE

Permissions for project "Big data project"

These permissions affect the entire "Big data project" project and all of its resources. To grant permissions, add a member and then select a role for them. Members can be people, domains, groups, or service accounts.

Some roles are in beta development and might be changed or deprecated in the future. [Learn more ↗](#).

Filter by name or role

View by: Members

Type	Members	Role(s)
<input type="checkbox"/>	LANCYMAO9@gmail.com	Owner

Manage resources



Google BigQuery

Create project

≡ Google APIs

Manage resources [CREATE PROJECT](#) [DELETE](#)

Filter by name, ID, project number, or label [Columns](#)

<input type="checkbox"/> Project name	Project ID	⋮
<input type="checkbox"/> Big data project	instant-keel-185814	⋮

Resources pending deletion



Google BigQuery

Create project

The project ID is the globally unique identifier for your project. You cannot change the project ID after the project is created. It will be used in SQL query to refer to the database and table.

New Project

i You have 9 projects remaining in your quota. [Learn more.](#)

Project name ?

Your project ID will be bigdata-185921 ? [Edit](#)

[Create](#)[Cancel](#)

Google BigQuery

Create project

After creating project, go to Google Cloud Platform, and click BigQuery in Resources section to go to query page.

This screenshot shows the IAM & admin permissions page for the 'Big data project'. It lists three members with their roles: AthenaXLee@gmail.com (Multiple), LANCYMAO9@gmail.com (Owner), and lopezelenamaria37@gmail.com (Multiple). A red box highlights the 'Google Cloud Platform' link at the bottom left of the page.

This screenshot shows the Google Cloud Platform dashboard for the 'Big data project'. The 'DASHBOARD' tab is selected. In the 'Resources' section, 'BigQuery' is listed with '2 datasets'. A red box highlights the 'BigQuery' link. Another red box highlights the 'Big data project' dropdown in the top navigation bar.

Create database

After creating project, you can click “create new dataset” to easily create a blank database in our project.

The screenshot shows the Google BigQuery web interface. At the top left is the "Google BigQuery" logo. Below it is a red button labeled "COMPOSE QUERY". To its right are links for "Query History" and "Job History". A search bar is labeled "Filter by ID or label". A sidebar on the left lists "Big data project" sections: "Stackoverflow" (selected and highlighted with a red box), "wbhealth", and "bigquery-public-data". On the right, under "Dataset Details: Stackoverflow", there is a "Description" field with placeholder text "Describe this dataset..." and a "Details" section with "Edit" buttons for "Edit" and "Switch to project". A "Create new dataset" button is also highlighted with a red box. At the bottom of the sidebar are "Refresh" and "Edit" buttons.



Google BigQuery

Create database

Create Dataset

Dataset ID



Data location



Data expiration

Never In days.

OK

Cancel



Google BigQuery

Create table

After database is created, you can click “+” after the database name to create as many table as you want.

COMPOSE QUERY

Query History

Job History

Filter by ID or label ?

Big data project ▼

- ▶ Stackoverflow + ▾
- ▶ wbhealth

Dataset Details: Stackoverflow

Description

Describe this dataset...

Details

Default Table Expiration	Never	Edit
Labels	None	Edit



Google BigQuery

Create table

This is the platform where I store raw data which is data source for my new table, remember the path to the file and paste it.

The screenshot shows the Google Cloud Platform Storage browser interface. The left sidebar has options for Storage, Browser, Transfer, and Settings. The main area is titled 'Browser' and shows a list of files under the path 'Buckets / bigdata_msba / stackoverflow'. A red box highlights this path. The table below lists five files, all named 'posts_questions' followed by a long string of zeros and ones, indicating they are gzip-compressed CSV files. The columns are Name, Size, Type, Storage class, and Last modified.

Name	Size	Type	Storage class	Last modified
posts_questions000000000000	456.3 MB	application/octet-stream	Multi-Regional	11/12/17, 10:21 PM
posts_questions000000000001	457.22 MB	application/octet-stream	Multi-Regional	11/12/17, 10:21 PM
posts_questions000000000002	457.45 MB	application/octet-stream	Multi-Regional	11/12/17, 10:21 PM
posts_questions000000000003	456.29 MB	application/octet-stream	Multi-Regional	11/12/17, 10:21 PM



Google BigQuery

Create table

Create Table

Source Data Create from source Create empty table

Repeat job

Select Previous Job



Location

Google Cloud Storage



gs://bigdata_msba/stackoverflow/posts_questions*



File format

CSV



[View Files](#)

Destination Table

Table name

Stackoverflow



. posts_questions



Table type

Native table



Schema Automatically detect



Name

id

Type

INTEGER

Mode

REQUIRED



creation_date

TIMESTAMP

REQUIRED



userid

INTEGER

REQUIRED



[Add Field](#)

[Edit as Text](#)

There are several sources you can choose, for example, file from your local machine or from Google Cloud Platform. For example, we want to import data which stored in Google Cloud Platform.

Then we should specify the file location in the Google Cloud Platform.

Google Cloud Storage URIs begin with "gs://" and specify the bucket and object you want to load.

Example: gs://mybucket/path/to/mydata.csv. You can use a wildcard to load multiple files



Google BigQuery

Several Remarks on Loading Data

Query data without loading it

- Public datasets:
 - Public datasets are datasets stored in BigQuery and shared with the public.
- Shared datasets:
 - You can share datasets stored in BigQuery. If someone has shared a dataset with you, you can run queries on that dataset without loading the data.
- External data sources:
 - You can skip the data loading process by creating a table that is based on an external data source.
- Stackdriver log files:
 - Cloud Logging provides an option to export log files into BigQuery.
- Stream the data one record at a time.
 - Typically used when you need the data to be available immediately.

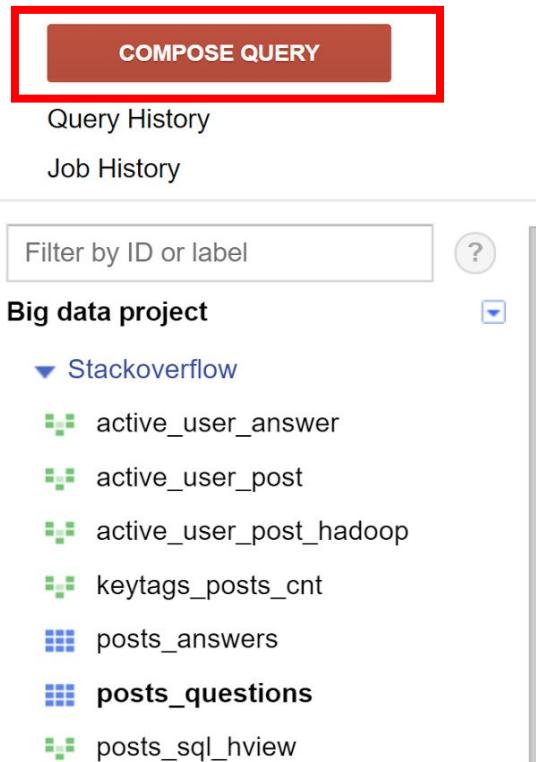
Query data by loading it:

- From Google Cloud Storage
- From a readable data source (such as your local machine)
- By inserting individual records using streaming inserts
- Using DML statements to perform bulk inserts
- Using a Google Cloud Dataflow pipeline to write data to BigQuery



Write, Execute, and Save Queries

Just click “Compose Query”



The screenshot shows the Google BigQuery web interface. On the left, there's a sidebar with a "COMPOSE QUERY" button highlighted by a red box. Below it are "Query History" and "Job History" sections, followed by a search bar labeled "Filter by ID or label". A dropdown menu titled "Big data project" is open, showing "Stackoverflow" with several sub-options: "active_user_answer", "active_user_post", "active_user_post_hadoop", "keytags_posts_cnt", "posts_answers", "posts_questions", and "posts_sql_hview". On the right, the main panel displays "Table Details: posts_questions". It includes tabs for "Schema", "Details", and "Preview". The "Schema" tab is active, showing the following table structure:

Field	Type	Nullable	Description
id	INTEGER	REQUIRED	Describe this field...
title	STRING	NULLABLE	Describe this field...
body	STRING	NULLABLE	Describe this field...
accepted_answer_id	INTEGER	NULLABLE	Describe this field...
answer_count	INTEGER	NULLABLE	Describe this field...
comment_count	INTEGER	NULLABLE	Describe this field...
community_owned_date	TIMESTAMP	NULLABLE	Describe this field...
creation_date	TIMESTAMP	NULLABLE	Describe this field...

Buttons for "Refresh", "Query Table", "Copy Table", "Export Table", and "Delete Table" are located at the top right of the details panel.



Google BigQuery

Write, Execute, and Save Queries

New Query will be on the top of the screen with table schema below, so it will be more convenient to check fields in tables while writing queries. And we can save query and view.

The screenshot shows the Google BigQuery interface. At the top, there's a navigation bar with tabs for "Query Editor" and "UDF Editor". Below the navigation bar is a "New Query" input field containing the following SQL query:

```
1 SELECT FROM [instant-keel-185814:Stackoverflow.posts_questions] LIMIT 1000
```

The "SQL" tab is selected. Below the query input, there are several buttons: "RUN QUERY", "Save Query" (which is highlighted with a red box), "Save View" (also highlighted with a red box), "Format Query", and "Show Options". A small red exclamation mark icon is located on the right side of the toolbar.

Below the toolbar, the text "Ctrl + Enter: run query, Tab or Ctrl + Space: autocomplete." is displayed. The interface then displays "Table Details: posts_questions". Under this heading, there are buttons for "Refresh", "Query Table", "Copy Table", "Export Table", and "Delete Table".

At the bottom, there are three tabs: "Schema", "Details" (which is selected and highlighted with a red box), and "Preview". The "Schema" tab shows the table structure:

id	INTEGER	REQUIRED	Describe this field...
title	STRING	NULLABLE	Describe this field...
body	STRING	NULLABLE	Describe this field...



Google BigQuery

Write, Execute, and Save Queries

You can also **save query** for future use. If you click “Query history” on the upper left side, you can see query history, saved queries, and project queries. Edit the saved queries and re-run them.

The screenshot shows the Google BigQuery interface. On the left, there's a sidebar with a red 'COMPOSE QUERY' button at the top. Below it are two buttons: 'Query History' (which is highlighted with a red box) and 'Job History'. Further down are sections for 'Filter by ID or label' and a dropdown menu 'Big data project' which is set to 'Stackoverflow'. Under 'Big data project', there are three items: 'Stackoverflow', 'active_user_an...', and 'active_user_post'. On the right, the main area has a title 'Queries'. Below it are three tabs: 'Query History', 'Saved Queries' (which is highlighted with a red box), and 'Project Queries'. To the right of the tabs are buttons for 'Sort by: Date' and '1 - 7 of 7'. Below the tabs is a 'Filter queries' dropdown. The main content area lists three queries: 'Active answer user', 'Active post user-hadoop', and 'Active posts user'. Each query has an 'Edit Query' button (highlighted with a red box) and a delete 'X' icon to its right.

Query	Action
Active answer user	Edit Query
Active post user-hadoop	Edit Query
Active posts user	Edit Query



Google BigQuery

Write, Execute, and Save Queries

Save view: if you want to create a table based on queries on other tables, you can just click “save view” to create a view without coding.

Save View

Project: Big data project (instant-keel-185814)

Dataset: Stackoverflow

Table ID: posts_cnt_yr

OK Cancel

Big data project

Stackoverflow

- posts_answers
- posts_cnt_yr
- posts_questions
- users



Google BigQuery

Write, Execute, and Save Queries

Error message will change once you correct the error, so it will avoid running incorrect queries several times.
Also you can see how much data the query will process.

```
1 ▾ SELECT
2   u.display_name,
3   pq.owner_user_id,
4   COUNT(id) AS post_cnt
5 ▾ FROM
6   [instant-keel-185814:Stackoverflow.post]
7 ▾ JOIN
8   [instant-keel-185814:Stackoverflow.user]
9 ▾ ON
10  pq.owner_user_id = u.id
11 ▾ WHERE
```

Error: 3.9 - 3.10: Ambiguous field reference id.

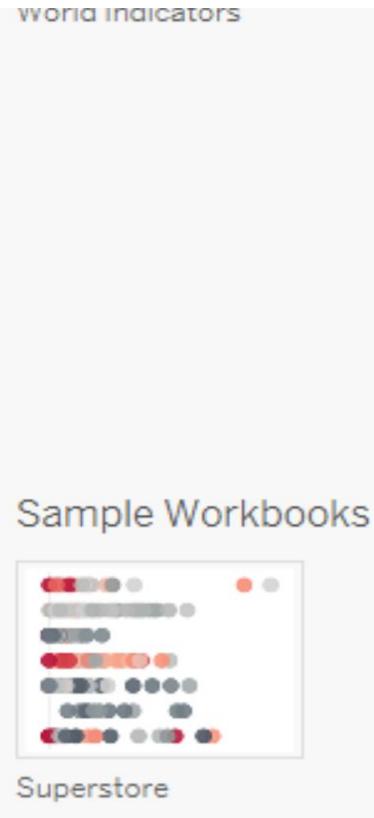
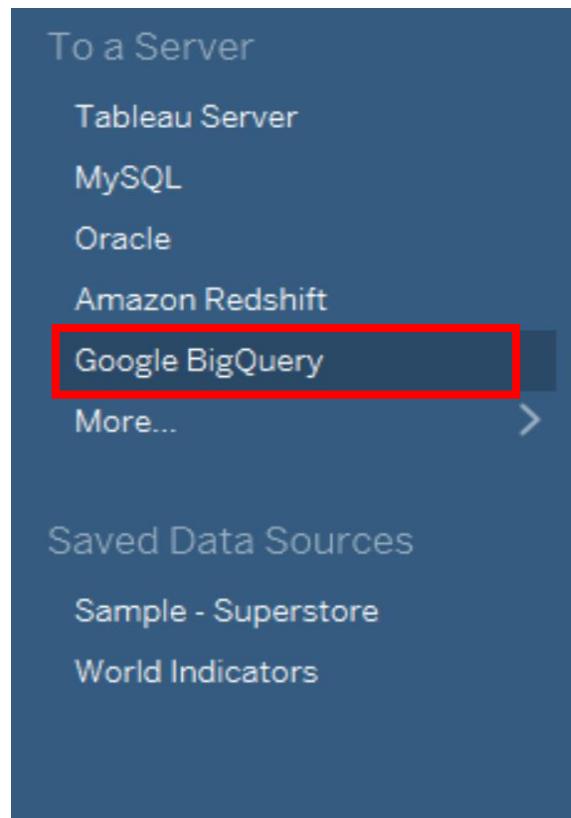
```
1 ▾ SELECT
2   u.display_name,
3   pq.owner_user_id,
4   COUNT(pq.id) AS post_cnt
5 ▾ FROM
6   [instant-keel-185814:Stackoverflow.posts_questions] pq
7 ▾ JOIN
8   [instant-keel-185814:Stackoverflow.users] u
9 ▾ ON
10  pq.owner_user_id = u.id
11 ▾ WHERE
```

Valid: This query will process 476 MB when run.



Google BigQuery

Link with other software, e.g. Tableau



The screenshot shows the 'Connections' dialog box in Tableau. A red box highlights the 'Dataset' dropdown, which is set to 'Stackoverflow'. The 'Table' dropdown lists the following tables:

- active_user_answer
- active_user_post
- active_user_post_hadoop
- keytags_posts_cnt
- posts_answers
- posts_questions
- posts_sql_hview
- users
- New Custom SQL
- New Union
- Use Legacy SQL



Google BigQuery

Stackoverflow Application



Google BigQuery

Stackoverflow Database

This dataset was last updated August, 2017. Originally, there are 16 tables, and we choose the following three to include in our database:

Table ID	posts_questions	posts_answers	users
Table Size	21.8 GB	17.7 GB	1.10 GB
Number of Rows	14,458,875	22,668,556	7,617,191



Google BigQuery

Post questions count by tags by year-month

```
1 SELECT * FROM
2 (SELECT 'SQL' AS tag, YEAR(creation_date) AS Year, MONTH(creation_date) AS Month, COUNT(id) AS post_cnt
3 FROM Stackoverflow.posts_questions
4 WHERE tags LIKE '%sql%' GROUP BY tag,Year,Month),
5 (SELECT 'Hadoop' AS tag, YEAR(creation_date) AS Year, MONTH(creation_date) AS Month,COUNT(id) AS post_cnt
6 FROM Stackoverflow.posts_questions
7 WHERE tags LIKE '%hadoop%'GROUP BY tag,Year,Month),
8 (SELECT 'MySQL' AS tag, YEAR(creation_date) AS Year, MONTH(creation_date) AS Month,COUNT(id) AS post_cnt
9 FROM Stackoverflow.posts_questions
10 WHERE tags LIKE '%mysql%'GROUP BY tag,Year,Month),
11 (SELECT 'Hive' AS tag, YEAR(creation_date) AS Year, MONTH(creation_date) AS Month,COUNT(id) AS post_cnt
```

Valid: This query will process 591 MB when run.

RUN QUERY ▾

Save Query

Save View

Format Query

Show Options

Query complete (1.5s elapsed, cached)

Results Explanation Job Information

Download as CSV

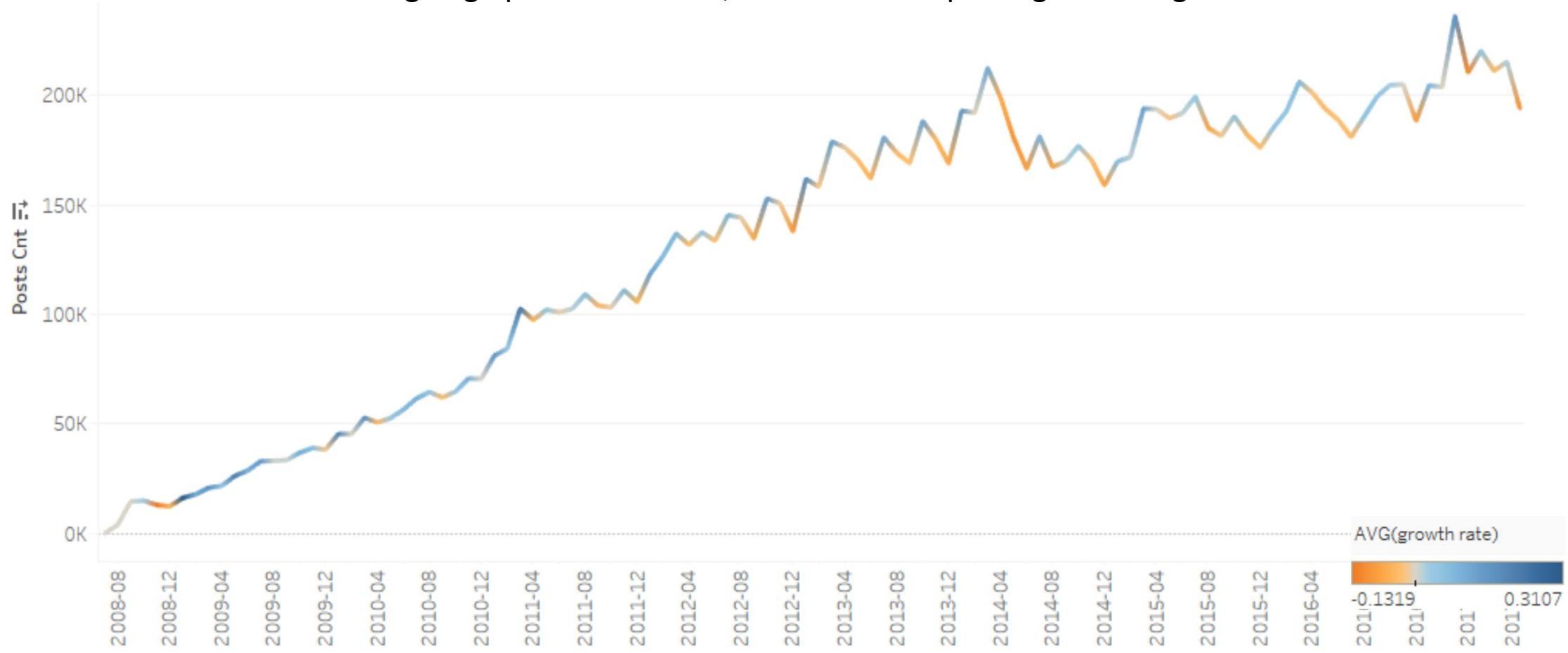
Row	tag	Year	Month	post_cnt	
1	AWS	2008	8	4	
2	AWS	2008	9	5	
3	AWS	2008	10	14	
4	AWS	2008	11	4	
5	AWS	2008	12	2	
6	AWS	2009	1	10	



Google BigQuery

Post questions count by year

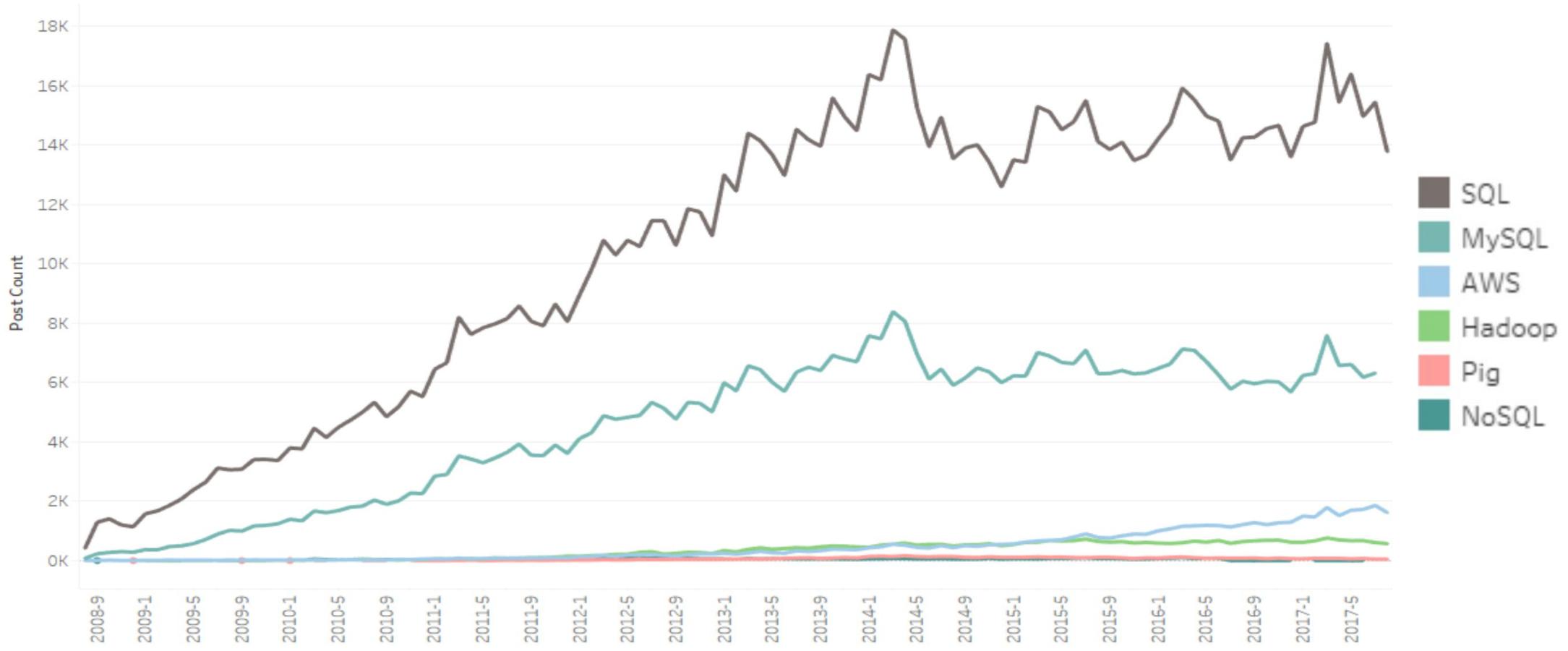
The trend is going upward overtime, but the rate of posting is slowing down



Google BigQuery

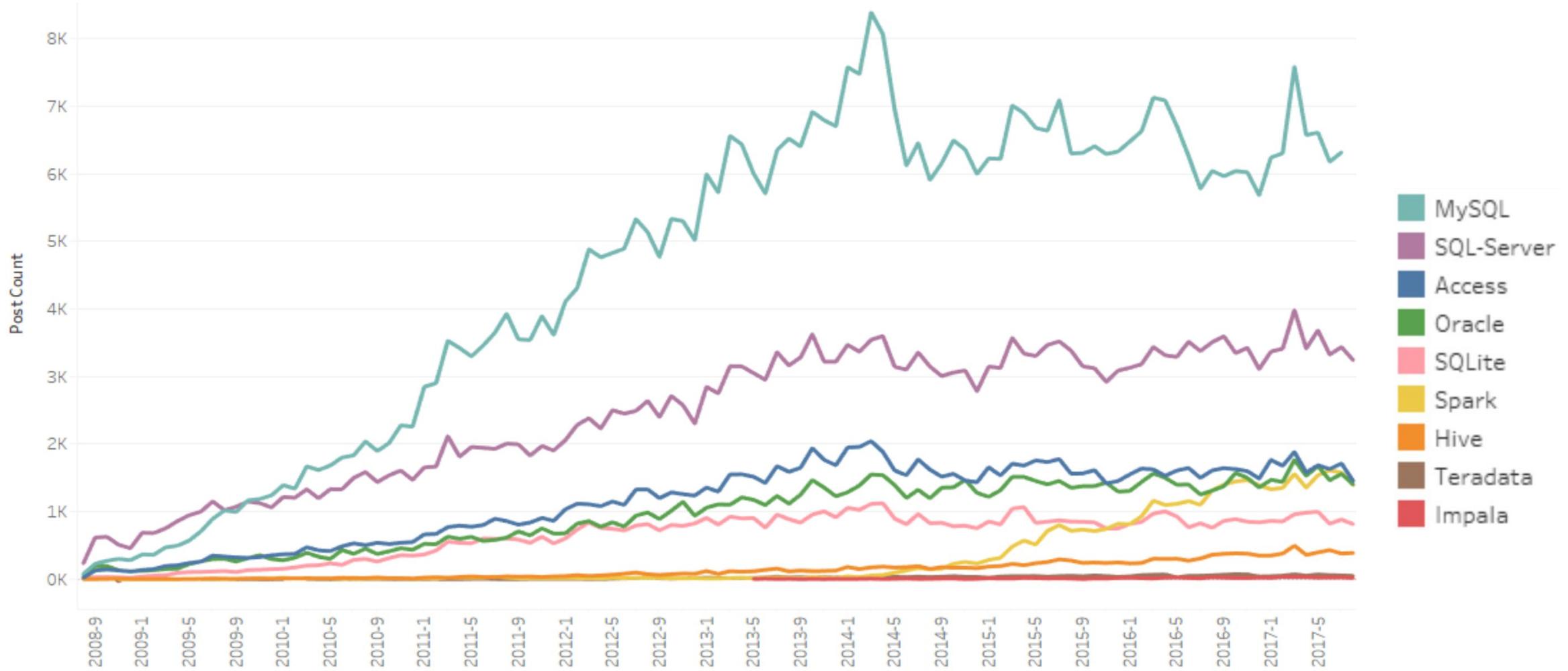
Question count by tags: Tools covered in class

SQL and MySQL are still the most popular tools



Google BigQuery

Question count by tags: Popular relational DBMS



Google BigQuery

SQL-Related Qs with highest view in 2017

```
1 ▾ SELECT
2     id,title,view_count
3 ▾ FROM
4     [instant-keel-185814:Stackoverflow.posts_questions]
5 ▾ WHERE
6     YEAR(creation_date)==2017 AND
7     tags LIKE '%sql%'
8 ▾ ORDER BY
9     view_count DESC
10    LIMIT 1;
```

Valid: This query will process 1.41 GB when run.

RUN QUERY

Save Query

Save View

Format Query

Show Options

Query complete (1.0s elapsed, cached)

Results

Explanation

Job Information

Download as CSV

Row	id	title	view_count	
1	41645309	MySQL Error: : 'Access denied for user 'root'@'localhost'	60176	



Google BigQuery

The Post Question

MySQL Error: : 'Access denied for user 'root'@'localhost'



A yellow banner with a cartoon illustration of a person working at a desk under a sun and umbrella. The text "Love remote work? Find it on a new kind of career site" is displayed next to the illustration. The Stack Overflow Jobs logo and a "Get started" button are on the right.

asked 10 months ago

viewed 95,430 times

active 20 days ago



\$./mysqladmin -u root -p 'redacted'

Enter password:

8



mysqladmin: connect to server at 'localhost' failed error:

'Access denied for user 'root'@'localhost' (using password: YES)'

7

How can I fix this?

mvsal sal database database-connection

BLOG

 The Cliffs of Insanity: MySQL Technologies on the Edge

Looking f



Google BigQuery

Hadoop-related Qs with most favorite in 2017

```
1 ▾ SELECT
2     id,title,favorite_count
3 ▾ FROM
4     [instant-keel-185814:Stackoverflow.posts_questions]
5 ▾ WHERE
6     YEAR(creation_date)==2017 AND
7     tags LIKE '%hadoop%'
8 ▾ ORDER BY
9     favorite_count DESC
10    LIMIT 1;
```

Valid: This query will process 1.33 GB when run.

RUN QUERY

Save Query

Save View

Format Query

Show Options

Query complete (0.5s elapsed, cached)

Results

Explanation

Job Information

Download as CSV

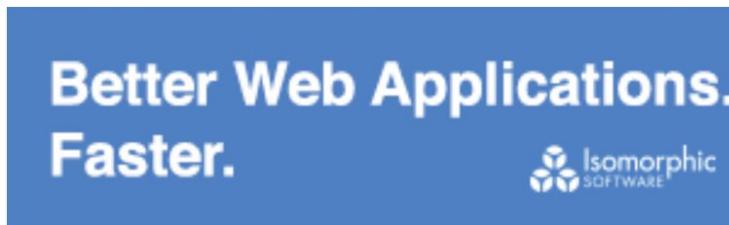
Row	id	title	favorite_count	
1	43270820	How to restart a failed task on Airflow	5	



Google BigQuery

The Post Question

How to restart a failed task on Airflow



[Live Showcase >](#)



asked 7 months ago

viewed 1,565 times

active 1 month ago



7



I am using a **LocalExecutor** and my dag has **3 tasks** where task(C) is dependant on task(A). Task(B) and task(A) can run in parallel something like below

A-->C



5

So **task(A)** has failed and but **task(B)** ran fine. Task(C) is yet to run as task(A) has failed.

My question is **how do i re run Task(A) alone so Task(C) runs** once Task(A) completes and Airflow UI marks them as success.

python hadoop airflow apache-airflow bigdata

BLOG

The Cliffs of Ins... Technologies or

Microsoft

Bring your favo... to our open &



Google BigQuery

Active user with most Qs in 2017

```
1 ▾ SELECT
2     pq.owner_user_id,u.display_name,u.reputation,u.location,COUNT(pq.id) AS post_cnt
3 ▾ FROM
4     [instant-keel-185814:Stackoverflow.posts_questions] pq
5 ▾ JOIN
6     [instant-keel-185814:Stackoverflow.users] u
7 ON pq.owner_user_id = u.id
8 WHERE YEAR(pq.creation_date)==2017
9 GROUP BY u.display_name, pq.owner_user_id, u.reputation,u.location
10 ORDER BY post_cnt DESC
11 LIMIT 10;
```

Valid: This query will process 572 MB when run.

Row	userid	u_display_name	u_reputation	u_location	post_cnt
1	1223975	Alexander Mills	8041	San Francisco, CA, United States	243
2	156458	Tim	23388		237
3	258483	Dims	7573	Moscow, Russia	227
4	1235929	Dave	1181		209
5	5421539	Zhao Yi	2516		207



Google BigQuery

Active user with most Hadoop-related Qs in 2017

```
1 ▾ SELECT
2     pq.owner_user_id, u.display_name, u.reputation, u.location, u.up_votes, u.down_votes, COUNT(pq.id) AS post_cnt
3 ▾ FROM
4     [instant-keel-185814:Stackoverflow.posts_questions] pq
5 ▾ JOIN
6     [instant-keel-185814:Stackoverflow.users] u
7 ON pq.owner_user_id = u.id
8 WHERE YEAR(pq.creation_date)==2017 AND pq.tags LIKE '%hadoop%'
9 GROUP BY u.display_name, pq.owner_user_id, u.reputation, u.up_votes, u.down_votes, u.location
10 ORDER BY post_cnt DESC
11 LIMIT 20;
```

Valid: This query will process 1.03 GB when run.

Row	pq_owner_user_id	u_display_name	u_reputation	u_location	post_cnt
1	2235335	Jain Hemant	62	Anmedabad, Gujarat, India	31
2	4287344	oula alshiekh	114		28
3	7119501	Sidhartha	143	Bangalore, Karnataka, India	27
4	5636416	earl	130		23
5	3544612	Saurab	125		22
6	3438473	CuriousMind	1302		21
7	1460514	SUDARSHAN	137	Bangalore, Karnataka, India	20
8	4751033	Basil Paul	62	Chennai, Tamil Nadu, India	19
9	3454410	Shafiq	1551		15
10	3956731	KayV	1608	Gurgaon, India	15



Google BigQuery

Active user with most answers in 2017

```
SELECT
  pa.owner_user_id AS userid, u.display_name, u.reputation, u.location, COUNT(pa.id) AS answer_cnt
FROM
  [instant-keel-185814:Stackoverflow.posts_answers] pa
JOIN
  [instant-keel-185814:Stackoverflow.users] u
ON pa.owner_user_id = u.id
WHERE YEAR(pa.creation_date)==2017
GROUP BY userid, u.display_name, pa.owner_user_id, u.reputation, u.location
ORDER BY answer_cnt DESC
LIMIT 20;
```

Row	userid	u_display_name	u_reputation	u_location	answer_cnt
1	1144035	Gordon Linoff	579631	New York, United States	5940
2	2901002	jezrael	153048	Bratislava, Slovakia	3084
3	3732271	akrun	285053		2632
4	3832970	Wiktor Stribiżew	211872	Warsaw, Poland	2084
5	2336654	piRSquared	84859	Bellevue, WA, United States	2042



Google BigQuery

Accepted answers & User Reputation

```
SELECT acc.userid, acc.name, acc.u.creation_date AS creation_date, acc.u.location AS location,
acc.u.up_votes AS up_votes,
acc.u.reputation AS reputation, acc.acc_cnt AS accepted_cnt, a.a_cnt AS answer_cnt
FROM
(SELECT pa.owner_user_id AS userid, u.display_name AS name, u.creation_date, u.location,
u.up_votes, u.reputation, COUNT(pq.accepted_answer_id) AS acc_cnt
FROM [instant-keel-185814:Stackoverflow.posts_questions] pq
JOIN [instant-keel-185814:Stackoverflow.posts_answers] pa
ON pq.accepted_answer_id = pa.id
JOIN [instant-keel-185814:Stackoverflow.users] u
ON pa.owner_user_id = u.id
GROUP BY userid, name, u.creation_date, u.location, u.up_votes, u.reputation
ORDER BY acc_cnt DESC) acc
JOIN
(SELECT pa.owner_user_id AS userid, COUNT(pa.id) AS a_cnt
FROM [instant-keel-185814:Stackoverflow.posts_answers] pa
GROUP BY userid
ORDER BY a_cnt DESC) a ON acc.userid=a.userid
ORDER BY acc.acc_cnt DESC;
```

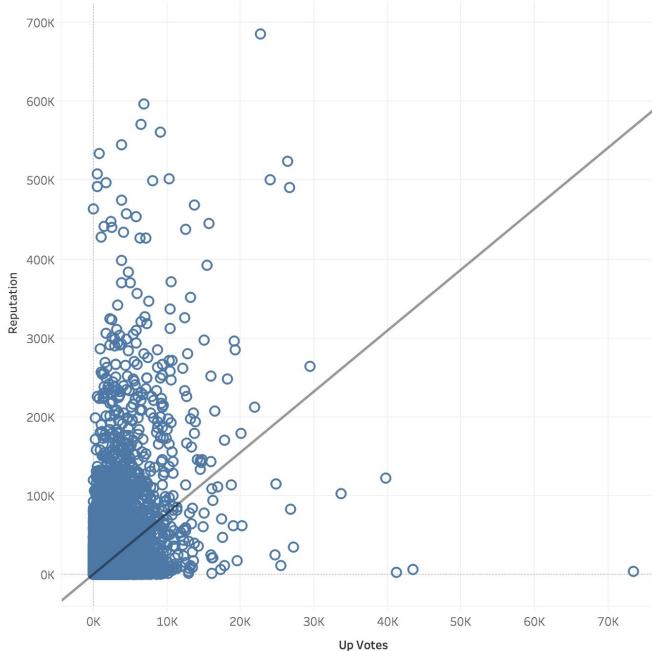


Google BigQuery

Accepted answers & User Reputation

Row	acc_userid	acc_name	creation_date	location	up_votes	reputation	accepted_cnt	answer_cnt
1	22656	Jon Skeet	2008-09-26 12:05:05 UTC	Reading, United Kingdom	15865	969386	20898	33911
2	1144035	Gordon Linoff	2012-01-11 19:53:57 UTC	New York, United States	8129	579631	17235	37988
3	29407	Darin Dimitrov	2008-10-19 16:07:47 UTC	Sofia, Bulgaria	1946	751190	12702	21503
4	100297	Martijn Pieters	2009-05-03 14:53:57 UTC	Cambridge, United Kingdom	5480	579284	12695	17571
5	115145	CommonsWare	2009-05-31 16:20:08 UTC	Pennsylvania, United States	9059	666690	11723	19573

Number of UpVotes v. User Reputation



Google BigQuery

Questions?

