



RIPE NCC

RIPE NETWORK COORDINATION CENTRE

RIPE Atlas in Google BigQuery

Stephen Strowes <sds@ripe.net> | 2020-10-26 | Academic and NREN Session

RIPE Atlas



- Continuously measures the Internet
 - 145 million traceroutes per day,
 - 190 million DNS queries per day,
 - 500 million pings per day
- This becomes unwieldy!





Announcing...

From today,

We're opening public access to our data on BigQuery



The screenshot shows the Google Cloud Platform BigQuery interface. The top navigation bar includes the Google Cloud Platform logo, the project name 'Serious Research Project', and a search bar. The left sidebar contains a navigation menu with options like Query history, Saved queries, Job history, Transfers, Scheduled queries, Reservations, BI Engine, and Resources. The Resources section is expanded, showing a tree view of data sets: 'serious-research-project' > 'ripence-atlas' > 'measurements' > 'samples'. The 'samples' data set is selected, and its contents are listed: 'dns', 'http', 'ntp', 'ping', 'sslicert', and 'traceroute'. The main area is the Query editor, which contains a SQL query:

```
1 select msm_id, count(*) result_count
2 from `ripence-atlas`.samples.ping
3 group by msm_id
4 order by result_count desc
```

 Below the query editor, there is a 'Valid' status indicator and buttons for 'Run', 'Save query', 'Save view', 'Schedule query', and 'More'. A message indicates 'This query will process 40.6 MB when run.' The 'Query results' section shows the query is complete (0.7 sec elapsed, 40.6 MB processed). Below this, there are tabs for 'Job information', 'Results', 'JSON', and 'Execution details'. The 'Results' tab is active, displaying a table with 11 rows and 3 columns: 'Row', 'msm_id', and 'result_count'. The table data is as follows:

Row	msm_id	result_count
1	1012	40299
2	1016	40185
3	1015	40102
4	1004	40058
5	1030	40046
6	1009	40038
7	1010	39993
8	1019	39992
9	1005	39924
10	1013	39906
11	1031	39878

At the bottom of the results table, there are controls for 'Rows per page' (set to 100), '1 - 100 of 9206', and navigation buttons for 'First page', '<', '>', and 'Last page'.

BigQuery: project & datasets



- <https://console.cloud.google.com/bigquery?project=ripencc-atlas>
 - **measurements** dataset:
 - public measurement results for DNS, traceroute, ping, HTTP, SSL, and NTP
 - dataset is current: streaming results
 - starting from 1 January 2020; we'll increase the date range as we backfill
 - **samples** dataset:
 - 1% of results from 1 week in October 2020, for each of the above
 - play with it!
- You need a Google account
 - we're covering storage, queries are on you



Why BigQuery?

- Can manage data like this without flinching
- Few hardware limitations
- Queries are often fast, but not real-time
 - think: $O(\text{seconds} \text{ — minutes})$, not $O(\text{milliseconds})$
- Allows iterative exploration of the data
 - ask for results related to ANY measurement
 - combine with other datasets you have access to

More soon



- Currently we have measurement data
- We are planning to add other common resources soon
 - probe meta data
 - BGP data

This is an active *beta* project



- We've been building experience but in reality:
 - we need to know what you use this for
 - what data works, and what does not?
 - what common datasets could we build and store?

More Information



- Information:
 - <https://labs.ripe.net/tools/ripe-atlas-on-bigquery/ripe-atlas-on-bigquery>
 - https://labs.ripe.net/Members/stephen_strowes/announcing-ripe-atlas-data-on-google-bigquery
- Docs, HOWTOs, schemas, and so forth:
 - <https://github.com/RIPE-NCC/ripe-atlas-bigquery/>



Questions



atlas-bq@ripe.net