RIPE NCC
RIPE NETWORK COORDINATION CENTRE

# BigQuery for Analysis of RIPE Atlas Data

Stephen Strowes <sds@ripe.net> | RACI, RIPE 79 | 2019-10-15

# RIPE Atlas Data

- We collect a lot of measurement data
- We provide no general means to filter or analyse that data

# Google’s BigQuery

- General platform for data analysis
- Experimenting with traceroute data
  - + smaller datasets
- Feedback loop: “is this useful?”

# Google’s BigQuery

- Can manage Atlas data without flinching
- Few hardware limitations
- Queries are fast, but not real-time
  - think: O(seconds — minutes), not O(milliseconds)
  - responsive enough to iterate during data exploration
- Opens up the full dataset in new ways

# Google's BigQuery

- Provides an SQL interface for querying data

```
1  WITH measurements AS
2  (
3    SELECT   rh.from AS ip, prbId AS prb_id, rtt
4    FROM     `data-test-194508.prod.traceroute_atlas_prod`, UNNEST(hops) h, UNNEST(resultHops) rh
5    WHERE    startTime >= "2019-10-10T00:00:00"
6    AND      startTime <  "2019-10-13T00:00:00"
7    AND      rh.from = "2001:67c:2e8:3::c100:a4"
8  )
9
10 SELECT   prb_id, approx_quantiles(rtt, 4) AS q, AVG(rtt) as m
11 FROM     measurements
12 GROUP BY prb_id
13 ORDER BY m ASC
```

# Google’s BigQuery

Query complete (5.7 sec elapsed, 213.3 GB processed)

Job information | Results | JSON | Execution details

| Row | prb_id | q | m |
|---|---|---|---|
| 1 | 6544 | 0.17299999296665192 | 0.24722453696584265 |
| | | 0.20900000631809235 | |
| | | 0.23100000619888306 | |
| | | 0.2639999985694885 | |
| | | 0.5789999961853027 | |
| 2 | 6539 | 0.2029999941587448 | 0.2487708332568959 |
| | | 0.2290000021457672 | |
| | | 0.24300000071525574 | |
| | | 0.2630000114440918 | |
| | | 0.5370000004768372 | |
| 3 | 6598 | 0.2329999953508377 | 0.3392430555627301 |
| | | 0.30000001192092896 | |

# Google’s BigQuery

- Arbitrary code
  - code running per-row becomes embarrassingly parallellisable
- Network data
  - the language doesn’t know what a subnet is
  - it does understand ranges, so prefix matching is easy
  - *longest*-prefix matching is trickier but can still be fast over on large datasets

# Direction of Travel

- Ultimately we'd like to allow access to our data via this interface
  - **Storage cost**: on the NCC
  - **Computation cost**: on anybody who wants to query the data

# Questions

Stephen Strowes <sds@ripe.net>
@sdstrowes