A Better Schema For Paris Traceroute

AIMS 2018

Ya Chang (yachang@google.com), Peter Boothe (pboothe@google.com)
We work at Google but are not speaking on behalf of Google

Overview

- What is Paris Traceroute?
- M-Lab's traceroute data
- Proposed schemas for PT
  - Current
  - Proposal 1
  - Proposal 2

Paris Traceroute History

Originally Proposed by

Brice Augustin, Xavier Cuvellier, Benjamin Orgogozo, Fabien Viger, Timur Friedman, Matthieu Latapy, Clémence Magnien, and Renata Teixeira, "Avoiding traceroute anomalies with Paris traceroute", in *Proc. Internet Measurement Conference*, October 2006

History of PT on the platform

- Set up on M-Lab platform from May, 2013
- Raw data stored on Google cloud storage
  - https://console.developers.google.com/storage/browser/m-lab/
- Parsed into BigQuery
  - https://bigquery.cloud.google.com/project/measurement-lab
- M-Lab data is now processed as 100% open source!
  - An opportunity for change
  - https://github.com/m-lab/etl/

Traceroutes per year for half a decade

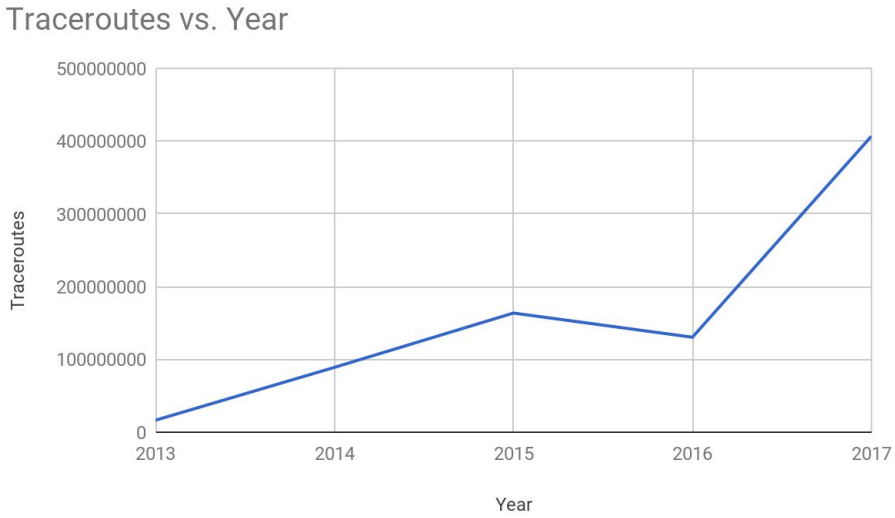
Total number of traces: 1 billion

Number of rows (hops) in the DB: 18 billion

3 billion rows for the first 10 weeks

of 2018

Expect to add 15 billion rows in 2018



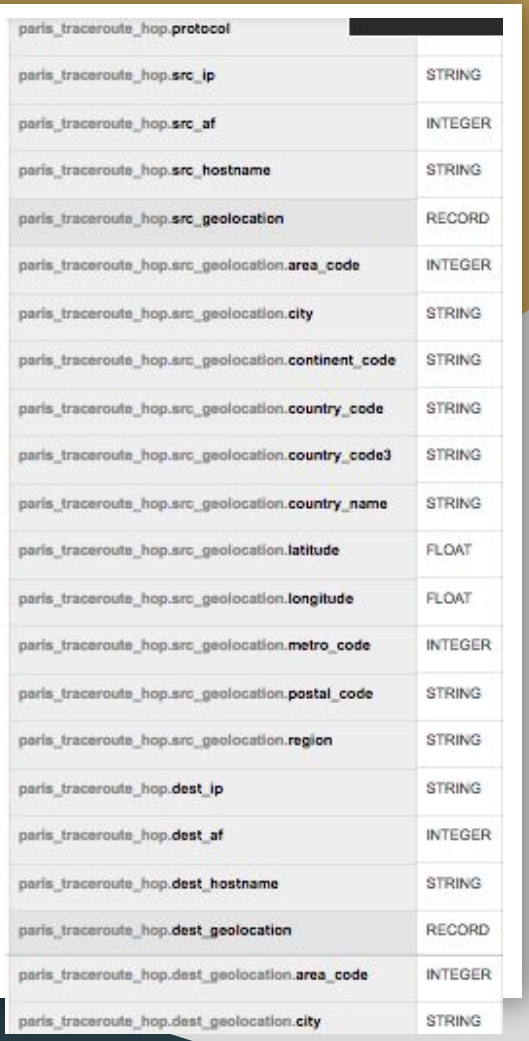
Problems to solve

1. The multiple hops for the same PT test are represented as multiple rows in BigQuery
2. The order of the hops within the same PT test can not be reconstructed directly through BigQuery
3. Frequently, people want to track AS paths
4. Users (you) complain about usability as a result.

Current schema

https://bigquery.cloud.google.com/table/measurement-lab:public.traceroute?pli=1&tab=schema

test_id	STRING
project	INTEGER
log_time	TIMESTAMP
type	INTEGER
connection_spec	RECORD
connection_spec.client_af	INTEGER
connection_spec.client_application	STRING
connection_spec.client_browser	STRING
connection_spec.client_hostname	STRING
connection_spec.client_ip	STRING
connection_spec.client_kernel_version	STRING
connection_spec.client_os	STRING
connection_spec.client_version	STRING
connection_spec.data_direction	INTEGER
connection_spec.server_af	INTEGER
connection_spec.server_hostname	STRING
connection_spec.server_ip	STRING
connection_spec.server_kernel_version	STRING
connection_spec.client_geolocation	RECORD
connection_spec.client_geolocation.area_code	INTEGER
connection_spec.client_geolocation.city	STRING

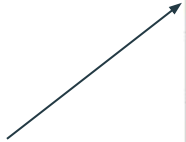


paris_traceroute_hop.protocol	
paris_traceroute_hop.src_ip	STRING
paris_traceroute_hop.src_af	INTEGER
paris_traceroute_hop.src_hostname	STRING
paris_traceroute_hop.src_geolocation	RECORD
paris_traceroute_hop.src_geolocation.area_code	INTEGER
paris_traceroute_hop.src_geolocation.city	STRING
paris_traceroute_hop.src_geolocation.continent_code	STRING
paris_traceroute_hop.src_geolocation.country_code	STRING
paris_traceroute_hop.src_geolocation.country_code3	STRING
paris_traceroute_hop.src_geolocation.country_name	STRING
paris_traceroute_hop.src_geolocation.latitude	FLOAT
paris_traceroute_hop.src_geolocation.longitude	FLOAT
paris_traceroute_hop.src_geolocation.metro_code	INTEGER
paris_traceroute_hop.src_geolocation.postal_code	STRING
paris_traceroute_hop.src_geolocation.region	STRING
paris_traceroute_hop.dest_ip	STRING
paris_traceroute_hop.dest_af	INTEGER
paris_traceroute_hop.dest_hostname	STRING
paris_traceroute_hop.dest_geolocation	RECORD
paris_traceroute_hop.dest_geolocation.area_code	INTEGER
paris_traceroute_hop.dest_geolocation.city	STRING

Proposal 1: repeated fields

Make this field repeated and save the multiple hops of the same PT test in a single row of BigQuery table.

Add index of hops in the repeated field for reconstruction purpose.



connection_spec.server_geolocation.**postal_code**	STRING
connection_spec.server_geolocation.**region**	STRING
paris_traceroute_hop	RECORD
paris_traceroute_hop.**protocol**	STRING
paris_traceroute_hop.**src_ip**	STRING
paris_traceroute_hop.**src_af**	INTEGER
paris_traceroute_hop.**src_hostname**	STRING
paris_traceroute_hop.**src_geolocation**	RECORD

Proposal 2: more metadata

Add more meta information, such as ASN, to the IPs in the path

IP-ASN information is from Maxmind (updated monthly) and Caida Routeviews (updated daily)

Log_time: timestamp
IP: string
ASN
{
maxmind: int64
caida_routeviews: int64
}

Use cases

Do you want to use this traceroute data?

For what? How? How can we make that easier?