Big Data with rubygems.org Download Data

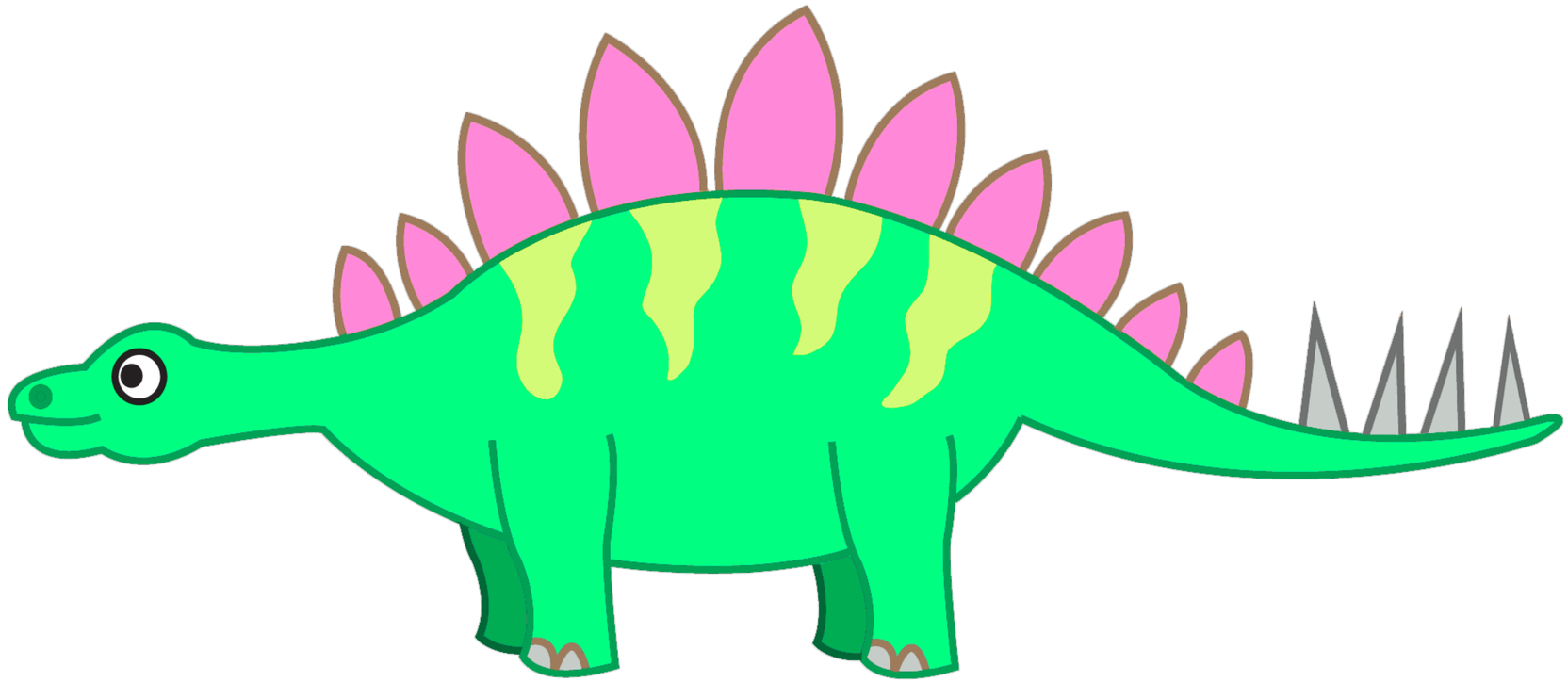
Aja Hammerly

Aja Hammerly

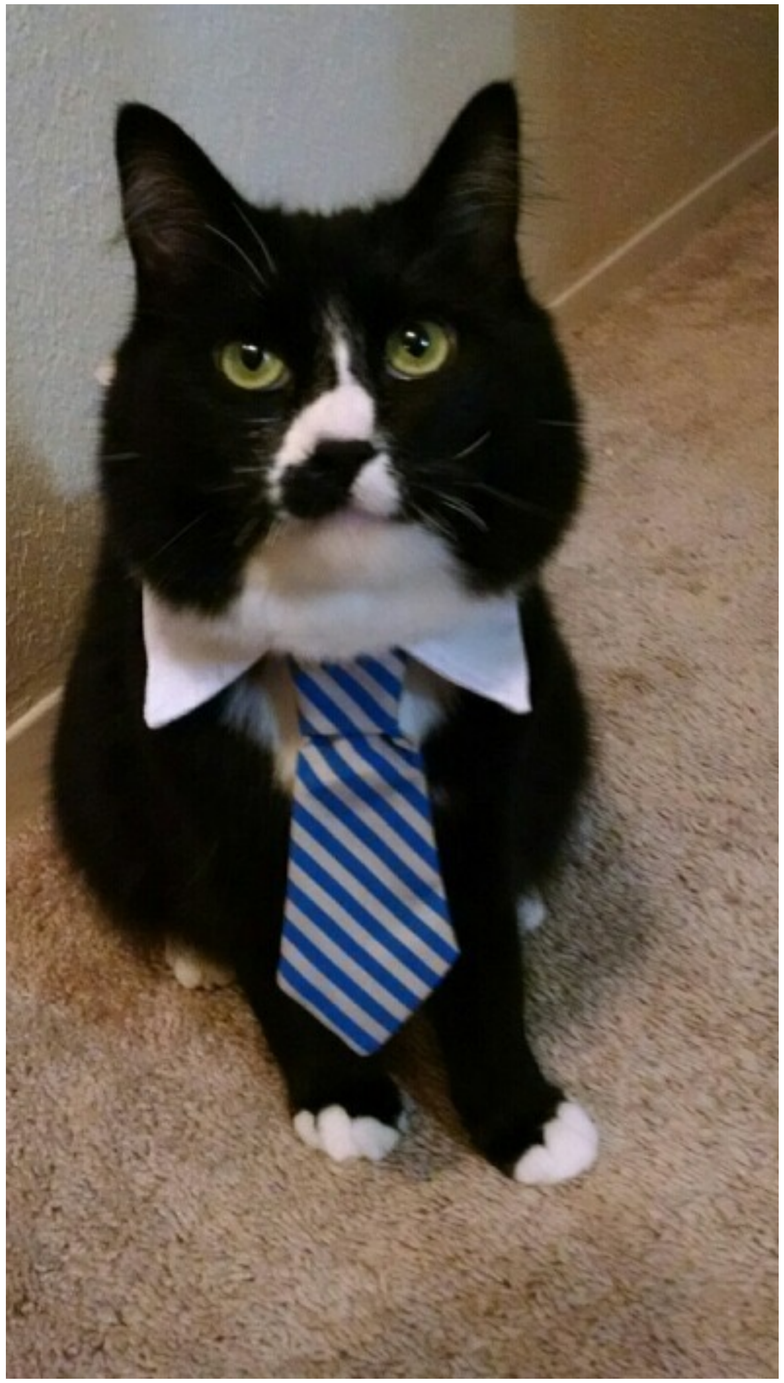
http://github.com/thagomizer

@thagomizer_rb

http://www.thagomizer.com



Google Cloud Platform



Lawyer Cat Says:

Any code is copyright Google and licensed Apache V2

Big Data

DATA

Big Data

Storage is Cheap

Intimidating

OMG Statistics

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

Machine Learning

Exploratory

Rubygems Download Data

Overview

rubygems

Column Name	Type
id	integer
name	varchar
created_at	datetime
updated_at	datetime
slug	varchar

Column Name	Type
id	integer
name	varchar
created_at	datetime
updated_at	datetime
slug	varchar

126,007

gem_downloads

Column Name	Type
id	integer
rubygem_id	integer
version_id	integer
count	bigint

883,848

dependencies

Column Name	Type
id	integer
requirements	varchar
rubygem_id	integer
version_id	integer
scope	varchar
created_at	datetime
updated_at	datetime
unresolved_name	varchar

Column Name	Type
id	integer
requirements	varchar
rubygem_id	integer
version_id	integer
scope	varchar
created_at	datetime
updated_at	datetime
unresolved_name	varchar

3,638,968

linksets

Column Name	Type
id	integer
rubygem_id	integer
home	varchar
wiki	varchar
docs	varchar
mail	varchar
code	varchar
bugs	varchar
created_at	datetime
updated_at	datetime

125,932

versions

Column Name	Type	Column Name	Type
id	integer	authors	text
rubygem_id	integer	description	text
size	integer	summary	text
position	integer	requirements	text
number	varchar	platform	varchar
indexed	boolean	full_name	varchar
prerelease	boolean	licenses	varchar
latest	boolean	required_ruby_version	varchar
yanked_at	datetime	required_rubygems_version	varchar
built_at	datetime	info_checksum	varchar
updated_at	datetime	metadata	hstore
created_at	datetime	sha256	varchar

Column Name	Type	Column Name	Type
id	integer	authors	text
rubygem_id	integer	description	text
size	integer	summary	text
position	integer	requirements	text
number	varchar	platform	varchar
indexed	boolean	full_name	varchar
prerelease	boolean	licenses	varchar
latest	boolean	required_ruby_version	varchar
yanked_at	datetime	required_rubygems_version	varchar
built_at	datetime	info_checksum	varchar
updated_at	datetime	metadata	hstore
created_at	datetime	sha256	varchar

Column Name	Type	Column Name	Type
id	integer	authors	text
rubygem_id	integer	description	text
size	integer	summary	text
position	integer	requirements	text
number	varchar	platform	varchar
indexed	boolean	full_name	varchar
prerelease	boolean	licenses	varchar
latest	boolean	required_ruby_version	varchar
yanked_at	datetime	required_rubygems_version	varchar
built_at	datetime	info_checksum	varchar
updated_at	datetime	metadata	hstore
created_at	datetime	sha256	varchar

Column Name	Type	Column Name	Type
id	integer	authors	text
rubygem_id	integer	description	text
size	integer	summary	text
position	integer	requirements	text
number	varchar	**platform**	varchar
indexed	boolean	full_name	varchar
prerelease	boolean	**licenses**	varchar
latest	boolean	**required_ruby_version**	varchar
yanked_at	datetime	**required_rubygems_version**	varchar
built_at	datetime	info_checksum	varchar
updated_at	datetime	metadata	hstore
created_at	datetime	sha256	varchar

Column Name	Type	Column Name	Type
id	integer	**authors**	text
rubygem_id	integer	description	text
size	integer	summary	text
position	integer	requirements	text
number	varchar	**platform**	varchar
indexed	boolean	full_name	varchar
prerelease	boolean	**licenses**	varchar
latest	boolean	**required_ruby_version**	varchar
yanked_at	datetime	**required_rubygems_version**	varchar
built_at	datetime	info_checksum	varchar
updated_at	datetime	metadata	hstore
created_at	datetime	sha256	varchar

Column Name	Type	Column Name	Type
id	integer	**authors**	text
rubygem_id	integer	description	text
size	integer	summary	text
position	integer	requirements	text
number	varchar	**platform**	varchar
indexed	boolean	**full_name**	varchar
prerelease	boolean	**licenses**	varchar
latest	boolean	**required_ruby_version**	varchar
yanked_at	datetime	**required_rubygems_version**	varchar
built_at	datetime	info_checksum	varchar
updated_at	datetime	**metadata**	hstore
created_at	datetime	sha256	varchar

757,920

Asking Questions

Domain Knowledge

Hypothesis

Examples

The gem with the most downloads is rails.

MiniTest is more popular than Rspec.

Gems released in the last year require ruby > 2.0.

Rails 3 is still more popular than rails 4.

Fewer gems are released during summer.

Largish Data

BigQuery

What

Why

How

I ❤️ BigQuery

SQL

Fast

Scales

Complex Enough

Demo

Vocabulary

Dataset

Table

Import

Streaming

gcloud

pg

```
require 'pg'
require 'gcloud'

ENV["GOOGLE_CLOUD_PROJECT"] = "rubygems-bigquery"
ENV["GOOGLE_CLOUD_KEYFILE"] = "#{key_path}"
```

```
gcloud      = Gcloud.new
bigquery    = gcloud.bigquery
bq_database = bigquery.dataset "rubygems"
```

```
postgres = PG.connect dbname: "rubygems"
```

```ruby
bq_table ||= bq_database.create_table("gems") do |s|
  s.integer   "id"
  s.string    "name"
  s.timestamp "created_at"
  s.timestamp "updated_at"
end
```

```
columns = %w[id name created_at updated_at]
```

```ruby
postgres.exec("SELECT * FROM rubygems") do |pg_table|
  pg_table.each do |row|
    hashed_row = Hash[columns.zip(row.values)]
    bq_table.insert(data)
  end
end
```

```ruby
postgres.exec("SELECT * FROM rubygems") do |pg_table|
  pg_table.each do |row|
    hashed_row = Hash[columns.zip(row.values)]
    bq_table.insert(data)
  end
end
```

```ruby
postgres.exec("SELECT * FROM rubygems") do |pg_table|
  pg_table.each do |row|
    hashed_row = Hash[columns.zip(row.values)]
    bq_table.insert(data)
  end
end
```

```ruby
postgres.exec("SELECT * FROM rubygems") do |pg_table|
  pg_table.each do |row|
    hashed_row = Hash[columns.zip(row.values)]
    bq_table.insert(hashed_row)
  end
end
```

Zip & Hash[]

```
[ key1, key2, key3, key4]
[ val1, val2, val3, val4]
```

zip

```
[ key1, key2, key3, key4]
[ val1, val2, val3, val4]

[[     ,     ], [     ,     ],
 [     ,     ], [     ,     ]]
```

```
[ key1, key2, key3, key4]

[ val1, val2, val3, val4]


[[key1,val1], [key2,val2],
 [key3,val3], [key4,val4]]
```

```
[[key1, val1], [key2, val2],
 [key3, val3], [key4, val4]]
```

Hash::[]

```
Hash[[key1, val1],
     [key2, val2],
     [key3, val3],
     [key4, val4]]
```

```
{ key1 => val1,
  key2 => val2,
  key3 => val3,
  key4 => val4 }
```

```
Hash[keys.zip(values)]
```

```ruby
postgres.exec("SELECT * FROM rubygems") do |pg_table|
  pg_table.each do |row|
    hashed_row = Hash[columns.zip(row.values)]
    bq_table.insert(hashed_row)
  end
end
```

Batch

Formats

CSV

JSON

Avro

CSV

```
require 'pg'
require 'csv'
require 'gcloud'
```

```
postgres = PG.connect dbname: "rubygems"

cols = %w[id requirements created_at updated_at
rubygem_id version_id scope]
```

```
query = "SELECT #{cols.join(',')} FROM dependencies"

CSV.open(csv_path, "wb") do |csv|
  postgres.exec(query) do |pg_table|
    pg_table.each do |row|
      csv << row.values
    end
  end
end
```

```
storage = Gcloud.new.storage
bucket  = storage.bucket "goruco2016-bg-files"

bucket.create_file csv_path, "dependencies.csv"
```

Import

COMPOSE QUERY

Query History

Job History

rubygems-bigquery

- rubygems

Public Datasets

- bigquery-public-d…
- bigquery-public-d…
- bigquery-public-d…
- bigquery-public-d…
- gdelt-bq:hathitrus…
- gdelt-bq:internet…
- lookerdata:cdc
- nyc-tlc:green
- nyc-tlc:yellow

Create Table

Source Data

Location: Google Cloud Storage — gs://goruco2016_bq_files/dependencies.csv

File format: CSV

View Files

Destination Table

Table name: rubygems . dependencies

Table type: Native table

Schema

Name	Type	Mode
id	INTEGER	NULLABLE
requirements	STRING	NULLABLE
created_at	TIMESTAMP	NULLABLE
updated_at	TIMESTAMP	NULLABLE
rubygems_id	INTEGER	NULLABLE
version_id	INTEGER	NULLABLE
scope	STRING	NULLABLE

Add Field

Edit as Text

rubygems-bigquery
rubygems
Public Datasets
bigquery-public-d...
bigquery-public-d...
bigquery-public-d...
bigquery-public-d...
gdelt-bq:hathitrus...
gdelt-bq:internet...
lookerdata:cdc
nyc-tlc:green
nyc-tlc:yellow

Name	Type	Mode
id	INTEGER	NULLABLE
requirements	STRING	NULLABLE
created_at	TIMESTAMP	NULLABLE
updated_at	TIMESTAMP	NULLABLE
rubygems_id	INTEGER	NULLABLE
version_id	INTEGER	NULLABLE
scope	STRING	NULLABLE

Add Field

Edit as Text

Options

Field delimiter: (•) Comma ( ) Tab ( ) Pipe ( ) Other [ ] ?

Header rows to skip: 0 ?

Number of errors allowed: 0 ?

Allow quoted newlines: [ ] ?

Allow jagged rows: [ ] ?

Ignore unknown values: [ ] ?

Write preference: Write if empty ?

Create Table

What Now?

rubygems

Simple

Rails has the most downloads.

Which gem has the most downloads?

```
SELECT name, count
FROM [rubygems.downloads]
JOIN rubygems.gems
ON rubygems.gems.id =
    rubygems.downloads.rubygem_id
ORDER BY count DESC
LIMIT 5
```

name	count
rake	107,076,261
rack	100,955,906
multi_json	100,171,080
json	95,715,131
bundler	93,085,862

```
SELECT name, sum(count) as total
FROM [rubygems.downloads]
JOIN rubygems.gems
ON rubygems.gems.id =
    rubygems.downloads.rubygem_id
GROUP BY name
ORDER BY total DESC
LIMIT 5
```

name	count
rake	214,152,212
rack	201,911,759
multi_json	200,342,260
json	191,430,173
bundler	186,172,479

How many downloads does Rails have?

```
SELECT name, sum(count) as total
FROM [rubygems.downloads]
JOIN rubygems.gems
ON rubygems.gems.id =
     rubygems.downloads.rubygem_id
WHERE name = 'rails'
```

name	total
rails	137,635,731

Minitest is more popular than Rspec.

```
SELECT name, sum(count) as total
FROM [rubygems.downloads]
JOIN rubygems.gems
ON rubygems.gems.id =
     rubygems.downloads.rubygem_id
GROUP BY name
HAVING name IN ('minitest', 'rspec')
```

name	total
minitest	101151246
rspec	77293803

Gems released in the last year require ruby > 2.

```
SELECT
  required_ruby_version,
  COUNT(*) AS total
FROM
  rubygems.versions
WHERE
  created_at > DATE_ADD(CURRENT_TIMESTAMP(), -1, "YEAR")
GROUP BY
  required_ruby_version
ORDER BY
  total DESC
```

name	total
>= 0	95,857
>= 1.9.3	9,069
>= 2.0.0	4,624
>= 2.0	1,648
>= 2.1.0	1,432

Complex

Rails 3 has more downloads than the other Rails major versions.

```
SELECT name,
       REGEXP_EXTRACT(number,r'(\d\.)') AS major,
       sum(rubygems.downloads.count) AS total
FROM [rubygems.versions]
JOIN rubygems.gems ON
  rubygems.gems.id =
   rubygems.versions.rubygem_id
JOIN rubygems.downloads ON
  rubygems.versions.rubygem_id =
   rubygems.downloads.rubygem_id
WHERE rubygems.gems.name = 'rails'
GROUP BY name, major
ORDER BY major
```

```
SELECT name,
       REGEXP_EXTRACT(number,r'(\d\.)') as major,
       sum(rubygems.downloads.count) as total
FROM [rubygems.versions]
JOIN rubygems.gems ON
  rubygems.gems.id =
   rubygems.versions.rubygem_id
JOIN rubygems.downloads ON
  rubygems.versions.rubygem_id =
   rubygems.downloads.rubygem_id
WHERE rubygems.gems.name = 'rails'
GROUP BY name, major
order by major
```

```
REGEXP_EXTRACT(number,r'(\d\.)') as major
```

version	downloads
0	2,890,350,351
1	2,064,535,965
2	3,991,436,199
3	**16,378,651,989**
4	12,662,487,252
5	963,450,117

version	downloads
0	2,890
1	2,064
2	3,991
3	**16,378**
4	12,662
5	963

Gems released in the last year require ruby > 2.

```
SELECT
  required_ruby_version,
  COUNT(*) AS total
FROM
  rubygems.versions
WHERE
  created_at > DATE_ADD(CURRENT_TIMESTAMP(), -1, "YEAR")
GROUP BY
  required_ruby_version
ORDER BY
  total DESC
```

```
SELECT
  REGEXP_EXTRACT(required_ruby_version,
    r'(.*?\d\.?)') AS version,
  COUNT(*) AS total
FROM
  rubygems.versions
WHERE
  created_at > DATE_ADD(CURRENT_TIMESTAMP(), -1, "YEAR")
GROUP BY
  version
ORDER BY
  total DESC
```

name	total
>= 0	95,851
>= 1	13,080
>= 2	12,944
~> 2	2,040
> 2	49

Thank You