

O'REILLY®

# Strata

Making Data Work

CONFERENCE

# Google Cloud for Data Crunchers

Chris Schalk, Developer Advocate, Cloud  
@cschalk

Ryan Boyd, Developer Advocate, Apps  
@ryguyrg, <http://profiles.google.com/ryan.boyd>

strataconf.com

# Agenda

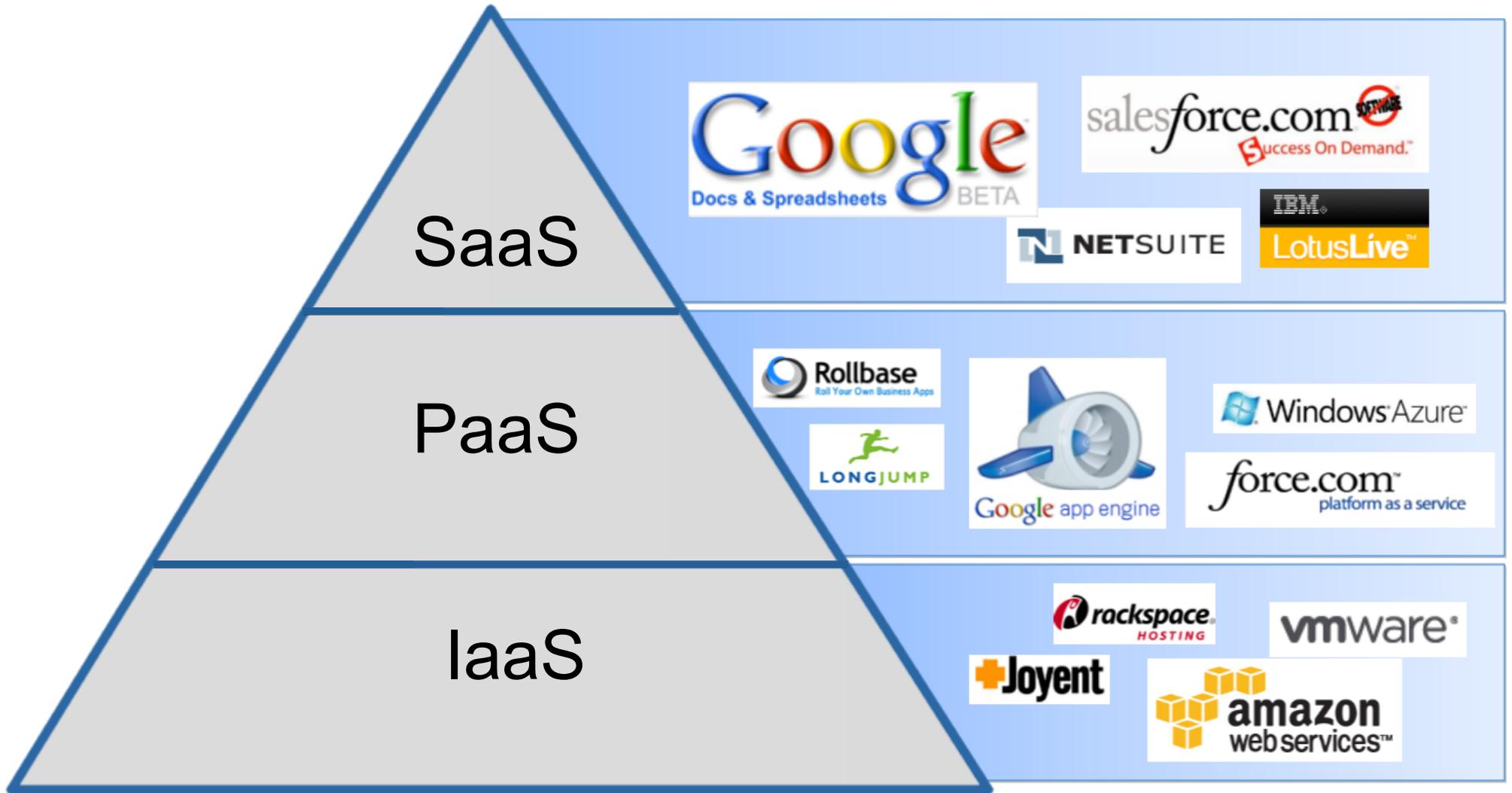
- Google App Engine
- Google Storage for Developers
- Prediction API
- BigQuery
- Google Fusion Tables
- Q&A during break



# Google App Engine



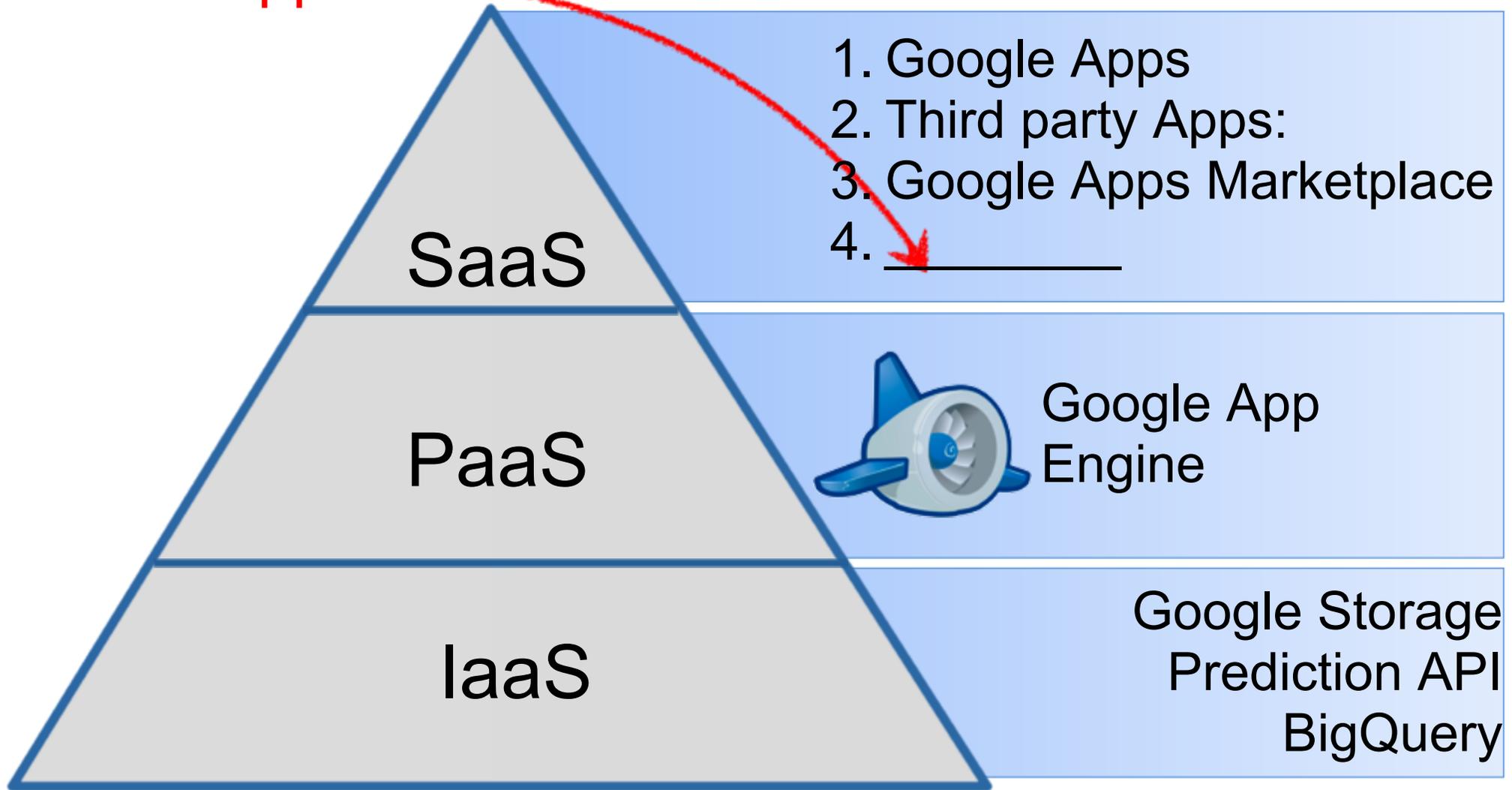
# Cloud Computing Defined



Source: Gartner AADI Summit Dec 2009

# Google's Cloud Offerings

Your Apps



# Google App Engine

Build and run your web apps on Google's infrastructure

- Easy to **build**
- Easy to **maintain**
- Easy to **scale**



*Focus on building your app, let us wear the pagers!*



# Cloud Development in a Box

- SDK & “The Cloud”
- Hardware
- Networking
- Operating system
- Application runtime
  - Java, Python, (go)
- API Services
- Fault tolerance
- Load balancing



# App Engine Services



# App Engine for Data Crunchers

- High Performance Image Serving
- Offline Data Processing
  - Push and Pull Queues
  - Backends
- Prospective Search (Matcher API)
- Mapper API (Reduce coming soon)
- Full Text Search (coming soon)



# Introducing Pull Queues

## Augments the existing "push" queues in App Engine

- Pull queues allow a task consumer to process tasks outside of App Engine's default task processing system
- Can manipulate tasks using simple API calls from an app
- Or externally via REST api

## New feature Introduced for Python & Java

- <http://code.google.com/appengine/docs/java/taskqueue/overview-pull.html>
- <http://code.google.com/appengine/docs/python/taskqueue/overview-pull.html>



# Introducing Backends

## Create offline processes that run indefinitely

- Special App Engine instances that have no request deadlines
- Each backend instance has a unique URL to use for requests
  - Start and run continuously in the background, or be driven by Task Queue tasks or Cron jobs
- Fully configurable via config file (backends.xml | backends.yaml)
  - Number of instances
  - Memory
  - CPU class
  - Public or Private
- Designed for long running **data** crunching

## Introduced for Python & Java

- <http://code.google.com/appengine/docs/java/backends/>
- <http://code.google.com/appengine/docs/python/backends/>



# Prospective Search (Matcher API)

Allows an app to register a set of queries to match against a stream of documents

## Experimental new feature for Python & Java

- <http://code.google.com/appengine/docs/java/prospectivesearch/>
- <http://code.google.com/appengine/docs/python/prospectivesearch/>

## Example:

- <http://code.google.com/p/im-tracker/>



# Mapper API

## First component of App Engine's MapReduce toolkit

- Large scale data manipulation
- Examples include:
  - Report generation
  - Computing statistics and metrics ...

## Blog Introduction

- <http://googleappengine.blogspot.com/2010/07/introducing-mapper-api.html>



# Full Text Search

Full text search coming to App Engine!

## Features

- Core search API
- Datastore search integration
- REST API to use the service outside of App Engine app code

## Status

- In progress

## Google IO 2011 Session

- <http://www.google.com/events/io/2011/sessions/full-text-search.html>



# Google App Engine Summary

- Easy to use, highly scalable Web Application Platform
- Can serve as Web front end to any data intensive application
- Has offline processing capabilities through Task Queues and Backend processes
- New data features
  - Prospective Search
  - Mapper (Reduce coming)
  - Full Text Search (Coming soon)

<http://code.google.com/appengine>

# Google Storage for Developers

Store your data in Google's cloud



# What Is Google Storage?

**Store** your data in Google's cloud

- any format, any amount, any time

**Control access** to your data

- private, shared, or public

**Easily integrate** with your app

- Google APIs + 3rd party tools & libs



# Google Storage Technical Details

## RESTful API

- **Verbs:** GET, PUT, POST, HEAD, DELETE
- **Resources:** identified by URL:  
<http://commondatastorage.googleapis.com/bucket/object>
- **Compatible with S3**

## Buckets

- Flat containers (no bucket hierarchy)



# Performance and Scalability

## Object types and size

- **Objects of any type** and 100GB+ / Object
- **Unlimited numbers of objects** and 1,000s of buckets
- **Range-get** support for data retrieval

## Replication

- **All data replicated** to multiple US data centers
  - Leveraging Google's worldwide network for data delivery

## Consistency

- **“Read-your-writes”** data consistency



# Security and Privacy Features

## Authenticated downloads from a web browser

- Sharing with individuals
- Group sharing via Google Groups
- Sharing with Google Apps domains

## Permissions set on Buckets or Objects

- **READ** (an object, or list a bucket's contents)
- **WRITE** (applicable to buckets, for upload/delete/etc)
- **FULL\_CONTROL** (read/write ACLs on objects or buckets)



# Tools

## Google Storage Manager

Send Feedback | mattg@google.com | Help | Sign out

Google code labs Google Storage for Developers

Places Home mrgtest123 Refresh New Folder Upload Delete

Drag and drop frequently used buckets and folders here for quicker access

Name	Size	Last Updated	Share Publicly
Test 1			
Test 2			
script1.png	368 KB	12:15 pm	✓
script2.png	473 KB	12:16 pm	✓
script3.png	585 KB	12:16 pm	✓

## gsutil

```
dhcp-172-19-3-109:~ wferrell$ gsutil
SYNOPSIS
gsutil [-d] [-h header]... command args

-d option shows HTTP protocol detail.

-h option allows you to specify additional HTTP headers, for example:
gsutil -h "Cache-Control:public,max-age=3600" -h "Content-Type:gzip" cp * g
s://bucket

Commands:
Concatenate object content to stdout:
cat [-h] uri...
-h Prints short header for each object.
Copy objects:
cp [-a canned_acl] [-t] [-z ext1,ext2,...] src_uri dst_uri
- or -
cp [-a canned_acl] [-t] [-z extensions] uri... dst_uri
-a Sets named canned_acl when uploaded objects created (list below).
-t Sets MIME type based on file extension.
-z 'txt,html' Compresses file uploads with the given extensions.
Get ACL XML for a bucket or object (save and edit for "setacl" command):
```

# Some Early Google Storage Adopters

vmware®

 syncplicity

 QTECH

 APPIRIO™

 SnapABug



VivU

 Cloud Sherpas

memeo

widgetbox

theguardian

XYLABS

# Google Storage usage within Google

Google  
BigQuery

Google  
Prediction API

Panoramio  
from Google

picnik

google.org  
Haiti Relief Imagery

Google patents  
USPTO data



double  
click

Partner Reporting



Partner Reporting

# Google Prediction API

Google's prediction engine in the cloud



# Introducing the Google Prediction API

- Google's sophisticated machine learning technology
- Available as an on-demand RESTful HTTP web service



# What is machine learning?

A set of algorithms that learn patterns from data and make intelligent decisions



**Inputs**



Predictive  
Model



**Output**



# How can Prediction be used?

The Prediction API is essentially a bundle of machine learning and statistical analysis algorithms. It can do two types of things in general:

- **Regression** -- Finding patterns in existing data and extrapolating assuming the previous patterns hold (user behaviors, for example)
- **Classification** -- Putting unknown objects in categories based on the features of that object and what categories previous objects were labelled as.



# How can Prediction be used?



Data  
Classification



Customer  
Sentiment



Content  
Moderation



Product  
Recommendation



Automatic  
Tagging



Message  
Routing

# Using the Prediction API



**1. Upload**

Upload your training data to Google Storage

**2. Train**

Build a model from your data

**3. Predict**

Make new predictions

**4. Adapt**

Feed real-time data updates

Marketplaces

Google Apps

Products

- Accounting & Finance
- Admin Tools
- Calendar & Scheduling
- Customer Management
- Document Management
- EDU
- Productivity
- Project Management
- Sales & Marketing
- Security & Compliance
- Workflow

Professional Services

- Archiving & Discovery Implementation
- Custom Application Development
- EDU Specialists
- Google Analytics
- Medium-Large Business Implementation
- Small Business Implementation
- Support & Managed Services
- Training & Change Management

Enterprise Search

Products

- Content Connectors
- OneBox Modules
- Search Extensions

Professional Services

- GSA Deployment
- Google Mini Deployment
- Custom Development
- Training
- GeoSpatial Solutions

The Google Apps Marketplace offers products and services designed for Google users, including installable apps that integrate directly with Google Apps. Installable apps are easy to use because they include single sign-on, Google's universal navigation, and some even include features that integrate with your domain's data.

### Featured Apps

**Concur Breeze – Free Mobile and Web Expense Reporting**

Concur Breeze is designed specifically to help small and mid-sized businesses take the hassle out of expense reporting, allowing your employees to spend more time making your business successful.



**Simple Expense Reporting**

• Try popular & notable apps



**SAP StreamWork**

SAP StreamWork is a collaborative decision-making solution that brings together the people, information, and proven business approaches to drive fast, meaningful results.



**ERPLY**

ERPLY offers web-based software for managing your points of sale, inventory, relationships and billing.



**Gantt Project**

Gantt.com is a powerful, web-based Project Management Tool that requires no software to be installed and it completely integrates with Google Docs.

**"Tops" in Google Apps**

Top Installed this week

1. [Insightly: Free simple CRM and Project Management](#)  
★★★★★ 321 reviews
2. [Manymoon: Free Social Productivity, Project Management & Task Management](#)  
★★★★★ 194 reviews
3. [Zoho CRM \(3 users free\)](#)  
★★★★★ 41 reviews
4. [MailChimp](#)  
★★★★★ 23 reviews
5. [myERP.com \(Free 2 users\) Accounting CRM Sales Inventory](#)  
★★★★★ 72 reviews

Top installed

1. [Manymoon: Free Social Productivity, Project Management & Task Management](#)  
★★★★★ 194 reviews
2. [Insightly: Free simple CRM and Project Management](#)  
★★★★★ 321 reviews
3. [Aviary Design Suite \(Free\)](#)  
★★★★★ 14 reviews
4. [OffiSync - \[FREE\] Integrate Microsoft Office with Google Apps](#)  
★★★★★ 572 reviews
5. [Zoho CRM \(3 users free\)](#)  
★★★★★ 41 reviews

Recently added

- [Zoho Books - Accounting and bookkeeping software](#)

# A Prediction API Example

Automatically determine application recommendations

- **Goal:** Increase relevancy on the Apps Marketplace via recommendations
- **Customers:** Businesses of various sizes and industries using Google Apps around the world
- **Data:** Sampling of previous installs of applications
- **Outcome:** Predict applications which would be appropriate for a new customer visiting the site



# Step 1: Upload

Upload your training data to Google Storage

**Create a CSV file with training data:**

```
"SlideRocket","EDUCATION","us","en","10","5"  
"MailChimp","BUSINESS","us","en","7","0"  
"MailChimp","STANDARD","se","sv","1","0"  
"Smartsheet","BUSINESS","us","en","13","4"
```

**Upload it to Google Storage:**

```
gsutil cp installs gs://myappdata/
```



## Step 2: Train

Create a new model by training on data

### To train a model (asynchronously):

POST /prediction/v1.3/training

```
{"id": "myappdata/installs"}
```

### To check training status:

GET /prediction/v1.3/training/myappdata%2Finstalls

```
{"kind": "prediction#training",  
  "modelInfo": {  
    "modelType": "classification",  
    "classificationAccuracy": 0.xx  
  },  
  "trainingStatus": "DONE"}
```



# Step 3: Predict

Apply the trained model to make predictions on new data

## To make a prediction:

POST /prediction/v1.3/training/myappdata%2Finstalls/predict

```
{ "input": { "csvInstance" : [
  "EDUCATION", "us", "en", "10", "0" ]}}
```

```
{ "kind" : "prediction#output",
  "outputLabel": "Manymoon",
  "outputMulti": [
    {"label": "Manymoon", "score": x.xx},
    {"label": "OffiSync", "score": x.xx},
    {"label": "Zoho CRM", "score": x.xx},
    {"label": "MailChimp", "score": x.xx}]}
```



# Demo!

<http://appsmarketplace-predict.appspot.com>



# Prediction API Capabilities

## Data

- **Input:** numeric and unstructured text
- **Output:** up to hundreds of discrete categories, or continuous values

## Training

- Many machine learning techniques
- Automatically selected
- Performed asynchronously



# Prediction API - Pricing

## Free Quota (for trial/development)

- 100 predictions/day, 5MB trained/day
- Available for 6 months

## Paid Usage

- \$10/month per project includes 10,000 predictions
- Additional predictions are \$0.50 per 1,000
- \$0.002 per MB trained (max size per dataset is 100MB)



# Prediction API - Getting Started

## Hosted Demo Models make it easy:

- Language Identifier
- Tag Categorizer
- Sentiment Predictor

## Building your own:

- Libraries and Samples in 7 Languages: Java, .NET, Objective-C, Go, PHP, Ruby, Python
- Easy REST-based APIs

<http://code.google.com/apis/predict/>



# Google BigQuery

Interactive analysis of large datasets in Google's cloud



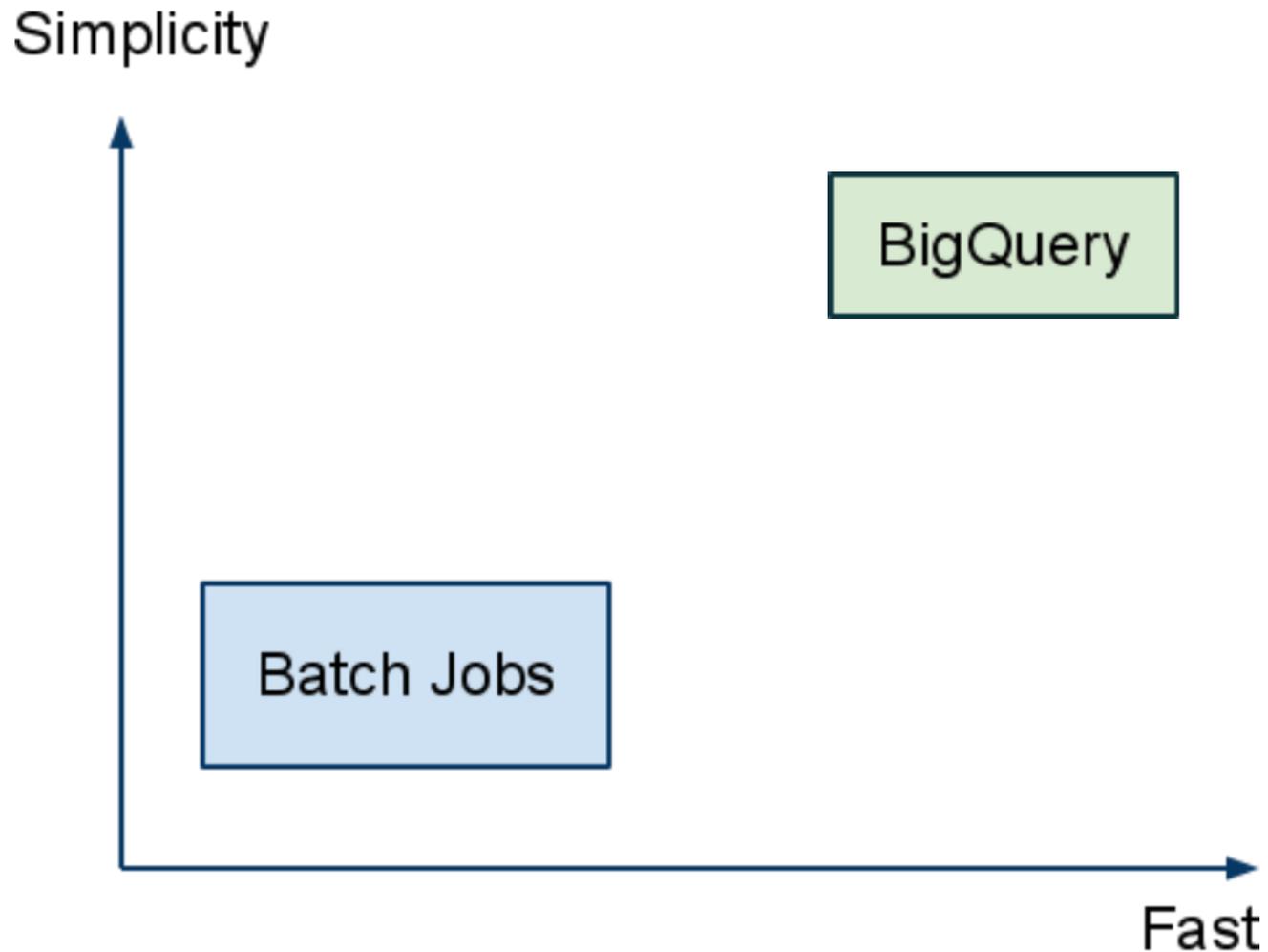
# Introducing Google BigQuery

- Google's large data adhoc analysis technology
  - Analyze massive amounts of data in seconds
- Simple SQL-like query language
- Flexible access
  - REST APIs, JSON-RPC, Google Apps Script



# Why BigQuery?

Working with large data is a challenge



# Many Use Cases ...



Interactive  
Tools



Spam



Trends  
Detection



Web  
Dashboards



Network  
Optimization



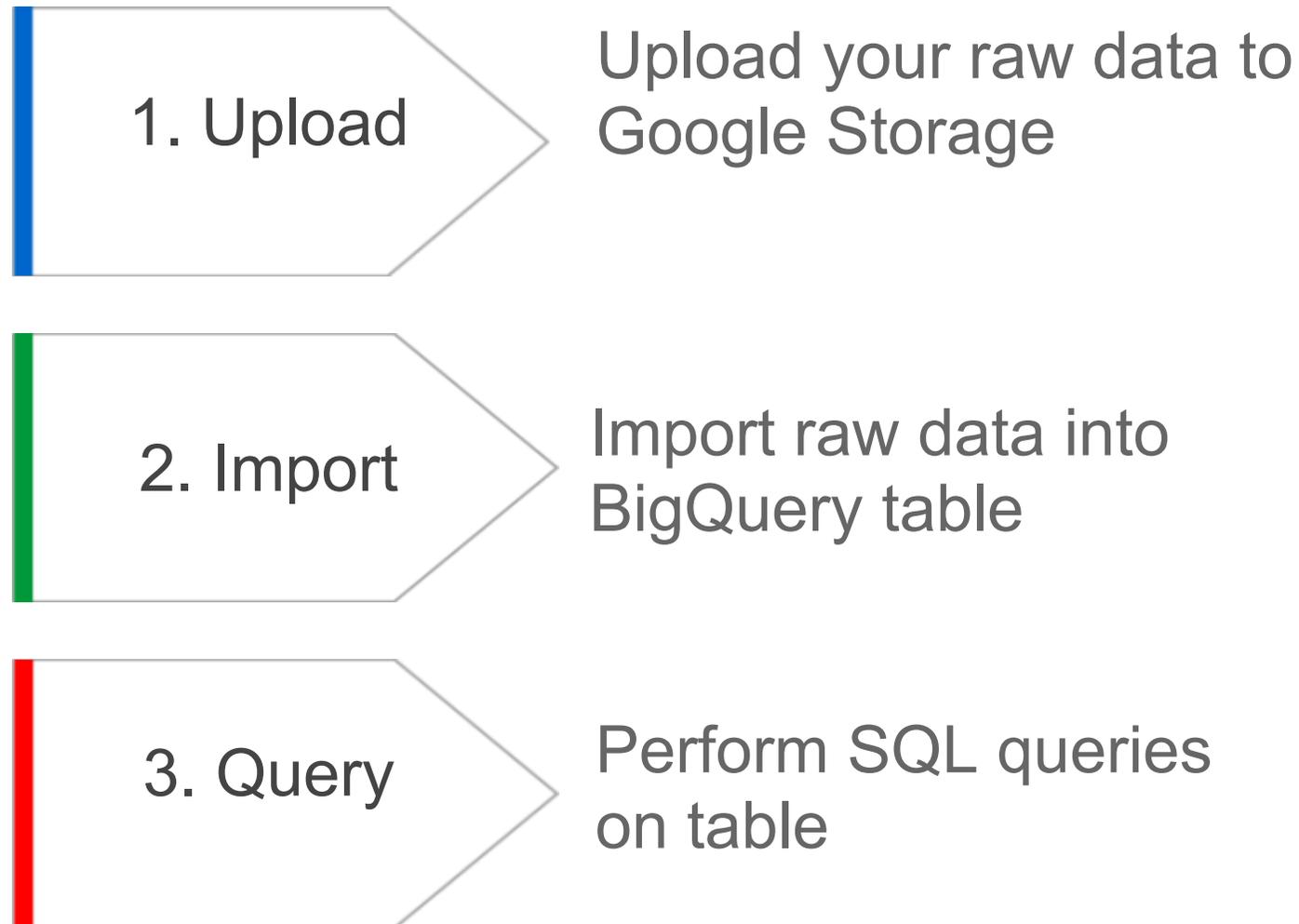
# Key Capabilities of BigQuery

- Scalable: Billions of rows
- Fast: Response in seconds
- Simple: Queries in SQL
- Web Service
  - REST
  - JSON-RPC
  - Google App Scripts



# Using BigQuery

Another simple three step process...



# Writing Queries

Compact subset of SQL

- **SELECT ... FROM ...  
WHERE ...  
GROUP BY ... ORDER BY ...  
LIMIT ...;**

Common functions

- Math, String, Time, ...

Additional statistical approximations

- TOP
- COUNT DISTINCT



# BigQuery via REST

GET /bigquery/v1/tables/{table name}

GET /bigquery/v1/query?q={query}

Sample JSON Reply:

```
{
  "results": {
    "fields": { [
      {"id": "COUNT(*)", "type": "uint64"}, ... ]
    },
    "rows": [
      {"f": [{"v": "2949"}, ...]},
      {"f": [{"v": "5387"}, ...]}, ... ]
    }
  }
}
```

Also supports JSON-RPC



# Security and Privacy

## Standard Google Authentication

- Client Login
- AuthSub
- OAuth

## HTTPS support

- protects your credentials
- protects your data

Relies on Google Storage to manage access



# Using BigQuery Shell Demo

```
Editor
title          STRING NULL
id             INT64 NULL
is_bot        BOOL NULL
comment       STRING NULL
num_characters INT32 NULL
is_minor      BOOL NULL

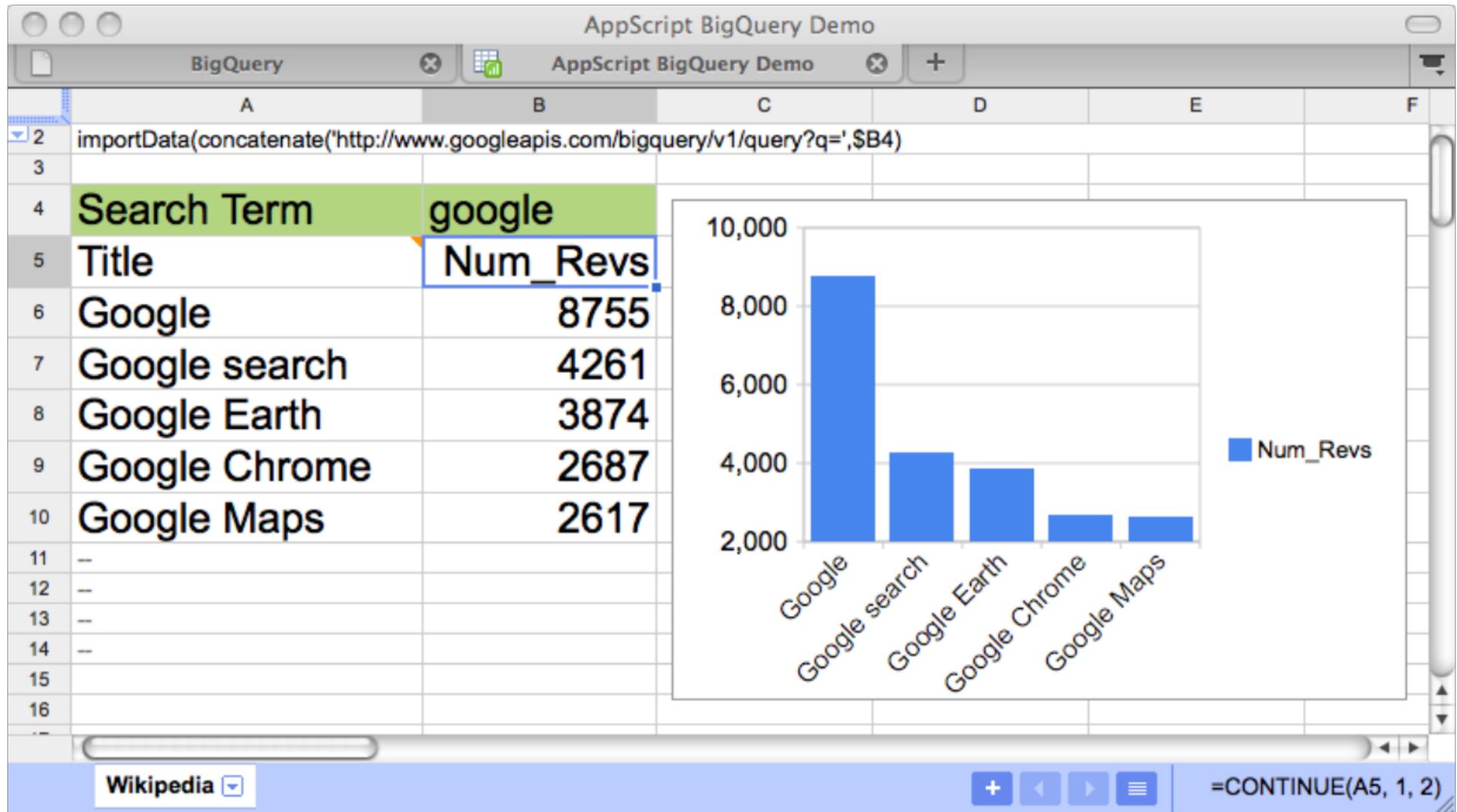
? SELECT TOP(title, 5), COUNT(*) FROM [bigquery.test.001/tables/wikipedia]
> WHERE wp_namespace = 0;
Execution time: 10.953 seconds
5 rows

TOP(title, 5)          COUNT(*)
-----
George W. Bush        43652
List of World Wrestling Entertainment employees  30572
Wikipedia             29726
United States         27433
Michael Jackson       23245

? |
```



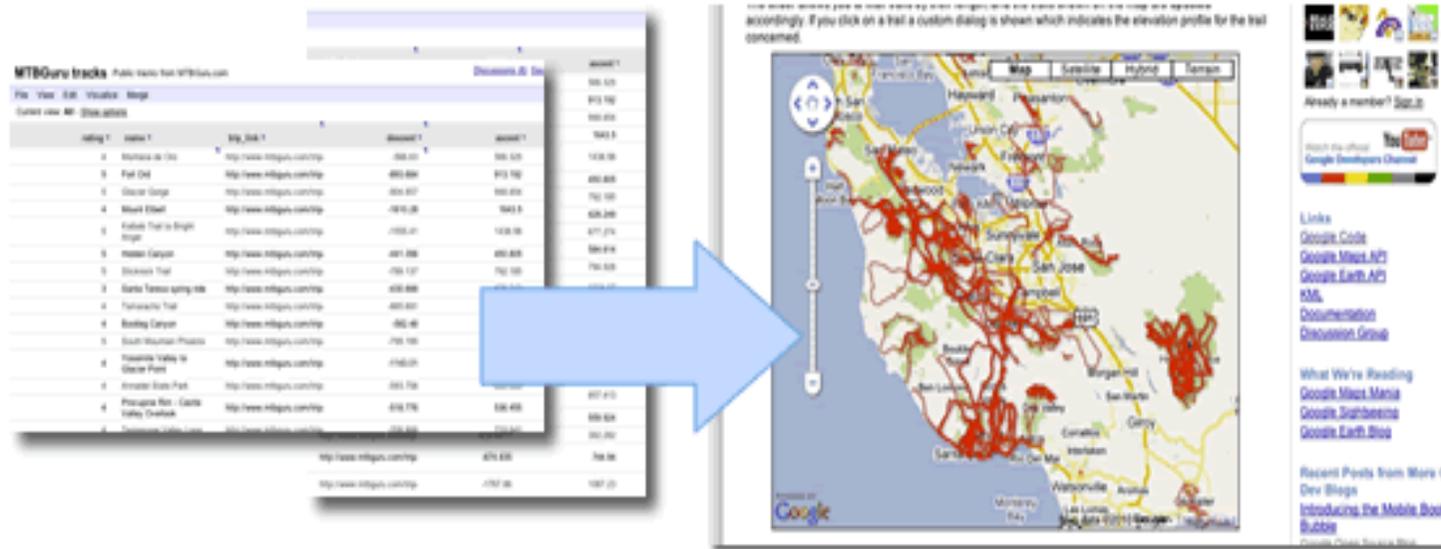
# BigQuery from a Spreadsheet



# Google Fusion Tables



# Fusion Tables, What is it?



Instantly visualize geographic data in a map or chart

- [Map Visualization: The richest and poorest places in England](#)
- [Chart Visualization: GDP per Capita](#)

# Fusion Tables Details

## Manage large collections of tabular data in the cloud:

- Visualizations
- Filters, Aggregation, Merge
- ACL, Collaboration, Discuss Data
- REST API
- Google Maps integration

## Import:

- Delimited Text (CSV, TSV, TXT)
- Spreadsheets (Google Docs, XLS, OpenOffice)
- KML

## Export:

- CSV
- KML



# Ways to access Fusion Tables



# Fusion Tables User Interface

Import | Manage | Visualize | Export | Share

Google fusion tables GDP, literacy rate , Wikipedia Get link Permissions Share (new)

File View Edit **Visualize** Merge

Current view: All - S **Table** 1 - 100 of 183 [Next »](#)

Country ▾	GDP per capita ▾	Literacy rate ▾	
Afghanistan	\$934.00		
Albania	\$7,169.00	99	
Algeria	\$6,885.00	75.4	
Angola	\$6,181.00	67.4	
Antigua and Barbuda	\$17,308.00	99	
Argentina	\$14,525.00	97.6	
Armenia	\$4,983.00	99.7	
Australia	\$38,663.00	99	
Austria	\$38,567.00	99	
Azerbaijan	\$9,540.00	99.5	
Bahamas, The	\$25,807.00		
Bahrain	\$27,214.00	88.8	
Bangladesh	\$1,487.00	53.5	
Barbados	\$22,272.00	99.7	
Belarus	\$12,750.00	99.7	
Belgium	\$35,534.00	99	
Belize	\$7,841.00	75.1	
Benin	\$1,440.00	40.5	

Map  
Intensity map  
Line  
Line (sample)  
Bar  
Pie  
Scatter  
Timeline (date, text, number)

# Google Maps API: Fusion Tables Layer

## Map & Query Your Data with Google Maps API (v3)

```
var myLayer =  
  new google.maps.FusionTablesLayer({  
    query: { select: 'Location', from: TABLEID } });
```



# Google Maps API: Fusion Tables Layer

## Easily add conditions to your queries

```
var myLayer = new google.maps.FusionTablesLayer({
  query: {
    select: 'Location', from: TABLEID,
    where: 'Age > 35' }
});
```

[Demo: Chicago Homicide Google Map \(V3\)](#)

[Demo: Mapping thousands of points with a query with parameter](#)



# Fusion Tables SQL API

```
SELECT * FROM 790805 ORDER BY ST_Distance(  
address, LatLng(37.7832, -122.4035))  
LIMIT 5;
```

```
name, address
```

```
MOSCONE CENTER WEST - BSMT KITCHEN, "800 HOWARD ST BASEMENT, 94103"
```

```
MOSCONE CENTER WEST - 2ND FLR. PANTRY, "800 HOWARD ST 2ND FLOOR, 94103"
```

```
MOSCONE CENTER WEST - 3 FLR. PANTRY, "800 HOWARD ST 3/F, 94103"
```

```
MOSCONE CENTER WEST - 1ST FLR. PANTRY, "800 HOWARD ST 1ST FLOOR, 94103"
```

```
ELAN EVENT VENUE, "839 HOWARD ST , 94103"
```

# Fusion Tables Summary

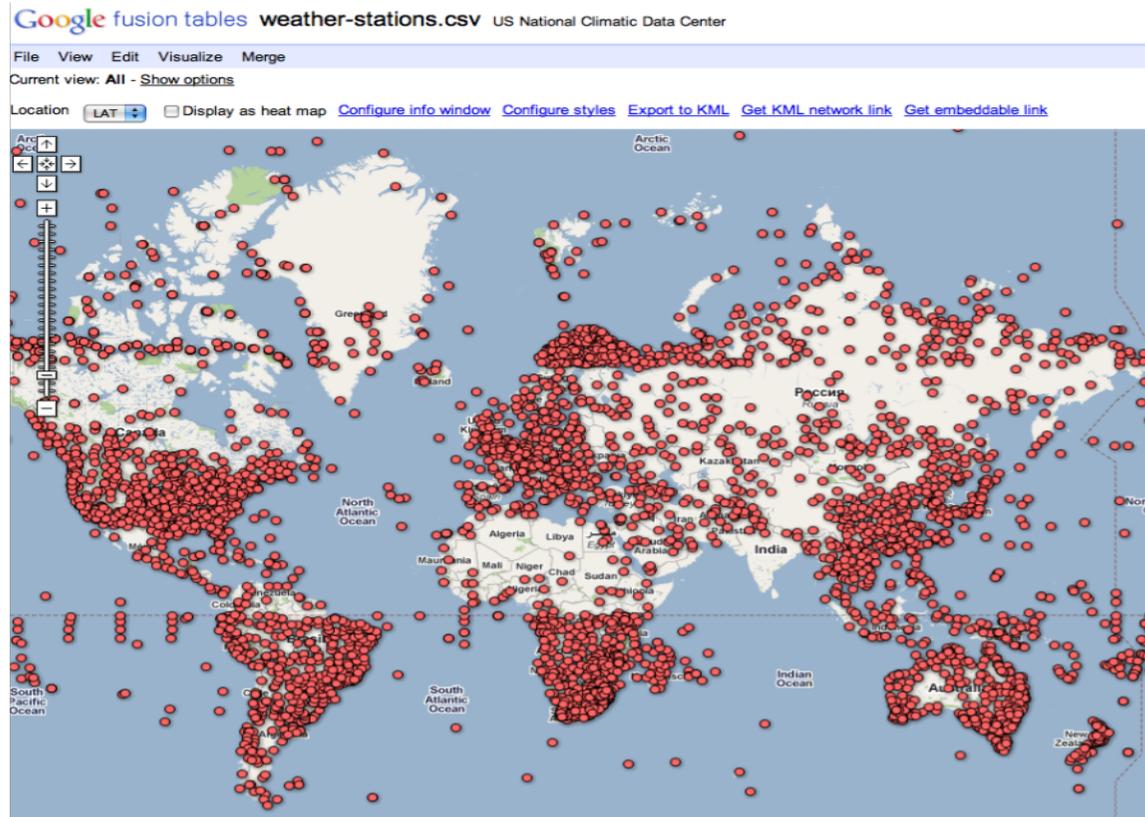
## A versatile and powerful data analysis and visualization platform

- Import | Manage | Visualize | Export | Share
- Works with large geographic datasets
- Access, map, query, and share your data multiple ways
  - Web App
  - Google Maps Fusion Tables Layer
  - Fusion Tables SQL API

Fusion Table API Documentation: <http://goo.gl/SIWR7>



# Integrated Demo!



Fusion Table Weather Stations Data combined with BigQuery and Visualization API

# Google Refine

Clean your data: <http://code.google.com/p/google-refine>

Google refine disasters Permalink Open... Export ▾ Help

Facet / Filter Undo / Redo 0

Refresh Reset All Remove All

17823 rows Extensions: Freebase ▾

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

**Drought** change 18 choices Sort by: name count Cluster

- Epidemic 1179
- Extreme temperature 361
- Flood 3512
- Industrial Accident 1190
- Insect infestation 83
- Mass Movement Dry 4
- Mass Movement Wet 12
- Mass movement dry 48
- Mass movement wet 505
- Miscellaneous accident 1133
- Storm 3206
- Transport Accident 4155

	All	00011969	00001969	Afghanistan	Paktia province	Drought	Drought2	Column	Column2	48000	0.2
1.	☆	29072006	29072006	Afghanistan	Emam Sahib (Kunduz distri ...	Earthquake (seismic activity)	Earthquake (ground shaking)		1	935	2
2.	☆	13122005	13122005	Afghanistan	Hindu Kush	Earthquake (seismic activity)	Earthquake (ground shaking)		5	501	2
3.	☆	8102005	8102005	Afghanistan	Nangarhar, Jalalabad prov ...	Earthquake (seismic activity)	Earthquake (ground shaking)		1	0.05	2
4.	☆	<a href="#">edit</a> 8072004	18072004	Afghanistan	Paktia province	Earthquake (seismic activity)	Earthquake (ground shaking)		2	1040	2
5.	☆	10042003	10042003	Afghanistan	Yakabagh (Takhar province ...	Earthquake (seismic activity)	Earthquake (ground shaking)		1	1001	2
6.	☆	12042002	12042002	Afghanistan	Dawabi, Khojakeder (Nahri ...	Earthquake (seismic activity)	Earthquake (ground shaking)		50	6150	2
7.	☆	25032002	25032002	Afghanistan	Nahrin (Baghlan province) ...	Earthquake (seismic activity)	Earthquake (ground shaking)		1000	91228	2
8.	☆	3032002	3032002	Afghanistan	Dakhli-Ezeu (Hindu Kush m ...	Earthquake (seismic activity)	Earthquake (ground shaking)		150	3513	2
9.	☆	1062001	1062001	Afghanistan	Jabul Saraj, Gumbahar, Pa ...	Earthquake (seismic activity)	Earthquake (ground shaking)		4	270	2
10.	☆	25022001	25022001	Afghanistan	Faizabad region (Badakhsh ...	Earthquake (seismic activity)	Earthquake (ground shaking)				2

# Summary

## Google App Engine

- Easily host your data intensive Web Apps

## Google Storage for Developers

- Highly performant cloud storage

## Prediction API

- Machine learning for the masses

## BigQuery

- High speed data analysis for extremely large data

## Google Fusion Tables

- Easily visualize your Geo data



O'REILLY®

# Strata

Making Data Work

CONFERENCE

# Thank You!

Chris Schalk, Developer Advocate, Cloud  
@cschalk

Ryan Boyd, Developer Advocate, Apps  
@ryguyrg, <http://profiles.google.com/ryan.boyd>

strataconf.com