Google Cloud for Data Crunchers

Chris Schalk, Developer Advocate, Cloud
@cschalk

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

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

- **IaaS**
- **PaaS**
- **SaaS**

Source: Gartner AADI Summit Dec 2009
Your Apps

1. Google Apps
2. Third party Apps:
3. Google Apps Marketplace
4. __________

Google's Cloud Offerings

- IaaS (Infrastructure as a Service)
- PaaS (Platform as a Service)
- SaaS (Software as a Service)

Google App Engine

Google Storage Prediction API
BigQuery
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
  - Java, Python, (go)
- Application runtime
- API Services
- Fault tolerance
- Load balancing
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

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

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


Example:

Mapper API

First component of App Engine’s MapReduce toolkit
- Large scale data manipulation
- Examples include:
  - Report generation
  - Computing statistics and metrics …

Blog Introduction
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

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
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

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

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

APPIRIO™
SnapABug

UNITED STATES NAVY

ViV:U
Cloud Sherpas

meneo
widgetbox

the guardian
OXYLABS
Google Storage usage within Google

Google BigQuery

Google Prediction API

Panoramio
picnik
Haiti Relief Imagery
Google patents
USPTO data

doubleclick
Partner Reporting

Partner Reporting

YouTube

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

“Tous pour un, un pour tous, c'est notre devise.”

“french”
What is machine learning?

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

Google Prediction API

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
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.

**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.

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

```json
{ "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/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 ...
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...

1. Upload
   Upload your raw data to Google Storage

2. Import
   Import raw data into BigQuery table

3. Query
   Perform SQL queries on table
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

```
? 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

<table>
<thead>
<tr>
<th>Search Term</th>
<th>Num_Revs</th>
</tr>
</thead>
<tbody>
<tr>
<td>google</td>
<td></td>
</tr>
<tr>
<td>Google</td>
<td>8755</td>
</tr>
<tr>
<td>Google search</td>
<td>4261</td>
</tr>
<tr>
<td>Google Earth</td>
<td>3874</td>
</tr>
<tr>
<td>Google Chrome</td>
<td>2687</td>
</tr>
<tr>
<td>Google Maps</td>
<td>2617</td>
</tr>
</tbody>
</table>

Bar chart showing the number of reviews for different Google-related terms.
Google Fusion Tables
Fusion Tables, What is it?

- Map Visualization: The richest and poorest places in England
- Chart Visualization: GDP per Capita

Instantly visualize geographic data in a map or chart.
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

- Your Data
- Web App
- APIs
## Fusion Tables User Interface

- **Import**
- **Manage**
- **Visualize**
- **Export**
- **Share**

### Table: GDP, literacy rate

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

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

<table>
<thead>
<tr>
<th>name</th>
<th>address</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOSCONE CENTER WEST - BSMT KITCHEN</td>
<td>&quot;800 HOWARD ST BASEMENT, 94103&quot;</td>
</tr>
<tr>
<td>MOSCONE CENTER WEST - 2ND FLR. PANTRY</td>
<td>&quot;800 HOWARD ST 2ND FLOOR, 94103&quot;</td>
</tr>
<tr>
<td>MOSCONE CENTER WEST - 3 FLR. PANTRY</td>
<td>&quot;800 HOWARD ST 3/F, 94103&quot;</td>
</tr>
<tr>
<td>MOSCONE CENTER WEST - 1ST FLR. PANTRY</td>
<td>&quot;800 HOWARD ST 1ST FLOOR, 94103&quot;</td>
</tr>
<tr>
<td>ELAN EVENT VENUE</td>
<td>&quot;839 HOWARD ST, 94103&quot;</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Drought</th>
<th>17823 rows</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td></td>
<td>1. 20072096</td>
</tr>
<tr>
<td></td>
<td>2. 13122005</td>
</tr>
<tr>
<td></td>
<td>3. 81202005</td>
</tr>
<tr>
<td></td>
<td>4. 10042003</td>
</tr>
<tr>
<td></td>
<td>5. 12042002</td>
</tr>
<tr>
<td></td>
<td>7. 25032200</td>
</tr>
<tr>
<td></td>
<td>8. 3032002</td>
</tr>
<tr>
<td></td>
<td>9. 1082001</td>
</tr>
<tr>
<td></td>
<td>10. 25022001</td>
</tr>
</tbody>
</table>
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
Thank You!

Chris Schalk, Developer Advocate, Cloud
@cschalk

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