Google Cloud for Data Crunchers

Patrick Chanezon, Developer Advocate, Cloud
@chanezon, chanezon@google.com

Ryan Boyd, Developer Advocate, Apps
@ryguyrg, rboyd@google.com

Kirrily Robert, Data Engineer, Freebase.com
@skud, skud@google.com
Agenda

• Google App Engine
• Google Storage for Developers
• Prediction API
• BigQuery
• Google Fusion Tables
• Google Refine
Google App Engine
What is cloud computing?
Cloud Computing Defined

Source: Gartner AADI Summit Dec 2009
Google's Cloud Offerings

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

- Your Apps

- IaaS
- PaaS
- SaaS

- Google App Engine
- Google Storage Prediction API
- BigQuery
Google App Engine

- Easy to build
- Easy to maintain
- Easy to scale
Cloud development in a box

- SDK & “The Cloud”
- Hardware
- Networking
- Operating system
  - Java, Python
- Static file serving
- Services
- Fault tolerance
- Load balancing
App Engine Services

- Memcache
- Datastore
- URL Fetch
- Mail
- XMPP
- Task Queue
- Images
- Blobstore
- User Service
Always free to get started

~5M pageviews/month

- 6.5 CPU hrs/day
- 1 GB storage
- 650K URL Fetch calls/day
- 2,000 recipients emailed
- 1 GB/day bandwidth
- 100,000 tasks enqueued
- 650K XMPP messages/day
Purchase additional resources *

* free monthly quota of ~5 million page views still in full effect
Google App Engine for Business
Same scalable cloud hosting platform. Designed for the enterprise.

• **Enterprise application management**
  – Centralized domain console

• **Enterprise reliability and support**
  – 99.9% Service Level Agreement
  – Premium Developer Support

• **Hosted SQL**
  – Managed relational SQL database in the cloud

• **SSL on your domain**
  – Including "naked" domain support

• **Secure by default**
  – Integrated Single Sign On (SSO)

• **Pricing that makes sense**
  – Pay only for what you use

* Hosted SQL and SSL on your domain available later this year
App Engine for Data Crunchers

• High Performance Image Serving
• OpenId/Oauth integration
• Increased quotas
  • > 1k entities per query
  • 10” task queues
• Async UrlFetch
• Mapper API (Reduce coming soon)
• Channel API
• Matcher API
Mapper API

• First component of App Engine’s MapReduce toolkit
• Large scale data manipulation
• Examples include:
  • Report generation
  • Computing statistics and metrics …
• Python Example:
  • http://blog.notdot.net/2010/05/Exploring-the-new-mapper-API
• Java Example:
Channel API

• Allows for Server Push (Comet) to browser
  • Blog post announcement:
    • http://googleappengine.blogspot.com/2010/05/app-engine-at-google-io-2010.html
  • External coverage:
    • Sneak Peak from an early trusted tester
      • http://bitshaq.com/2010/09/01/sneak-peak-gae-channel-api/
  • Demo code for Dance Dance Robot available here:
    • http://code.google.com/p/dance-dance-robot/
    • Also see: https://groups.google.com/group/google-appengine-java/browse_thread/thread/6fa09953ffae2cd3/c1db7de5fdb82b65?pli=1#
Matcher API

- Allows an app to register a set of queries to match against a stream of documents
  - Trusts Testers, Python only
  - Group post announcement:
    - http://groups.google.com/group/google-appengine/msg/40021537e2e58962
  - Docs:
- Demo code:
  - http://code.google.com/p/google-app-engine-samples/source/browse/#svn/trunk/matcher-sample
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

• You control access to your data
  ◦ private, shared, or public

• Access via Google APIs or 3rd party tools/libraries
Google Storage Technical Details

RESTful API
• Verbs: GET, PUT, POST, HEAD, DELETE
• Resources: identified by URI, like:
  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, 1000s 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, allows upload/delete/etc)
• FULL_CONTROL (read/write ACLs on objects or buckets)
Tools

Google Storage Manager

**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 * gs://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):
Google Storage Benefits

High Performance and Scalability
Backed by Google infrastructure

Strong Security and Privacy
Control access to your data

Easy to Use
Get started fast with Google & 3rd party tools
Some Early Google Storage Adopters
Google Storage usage within Google

Google BigQuery

Google Prediction API

Panoramio by Google

picnik

google.org Haiti Relief Imagery

Google patents USPTO data

doubleclick

Partner Reporting

Partner Reporting

YouTube™
Google Storage - Availability

Limited preview in US* currently
- 100GB free storage and network per account
- Sign up for wait list at
  - [http://code.google.com/apis/storage/](http://code.google.com/apis/storage/)

* Non-US preview available on case-by-case basis
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.”
A virtually endless number of applications...

<table>
<thead>
<tr>
<th>Customer Sentiment</th>
<th>Transaction Risk</th>
<th>Species Identification</th>
<th>Message Routing</th>
<th>Diagnostics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Churn Prediction</td>
<td>Legal Docket Classification</td>
<td>Suspicious Activity</td>
<td>Work Roster Assignment</td>
<td>Inappropriate Content</td>
</tr>
<tr>
<td>Recommend Products</td>
<td>Political Bias</td>
<td>Uplift Marketing</td>
<td>Email Filtering</td>
<td>Career Counseling</td>
</tr>
</tbody>
</table>

... and many more ...
How does it work?

1. TRAIN
The Prediction API finds relevant **features** in the sample data during training.

<table>
<thead>
<tr>
<th>&quot;english&quot;</th>
<th>The quick brown fox jumped over the lazy dog.</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;english&quot;</td>
<td>To err is human, but to really foul things up you need a computer.</td>
</tr>
<tr>
<td>&quot;spanish&quot;</td>
<td>No hay mal que por bien no venga.</td>
</tr>
<tr>
<td>&quot;spanish&quot;</td>
<td>La tercera es la vencida.</td>
</tr>
</tbody>
</table>

2. PREDICT
The Prediction API later searches for those **features** during prediction.

| ? | To be or not to be, that is the question. |
| ? | La fe mueve montañas. |
Introducing the Google Prediction API

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.

Gantter Project
Gantter.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
Using the Prediction API

A simple three step process...

1. Upload
   Upload your training data to Google Storage

2. Train
   Build a model from your data

3. Predict
   Make new predictions
Step 1: Upload

Upload your training data to Google Storage

- Training data: outputs and input features
- Data format: comma separated value format (CSV), result in first column

"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 to Google Storage

gsutil cp installs gs://appdata/
Step 2: Train
Create a new model by training on data

To train a model:

POST prediction/v1.1/training?data=appdata%2Finstalls

Training runs asynchronously. To see if it has finished:

GET prediction/v1.1/training/appdata%2Finstalls

{"data":{
  "data":"appdata/installs",
  "modelinfo":"estimated accuracy: 0.xx"}}
Step 3: Predict

Apply the trained model to make predictions on new data

POST prediction/v1.1/query/appdata%2Finstalls/predict

{ "data":{
   "input": { "mixture" : [
   "EDUCATION","us","en","10","0" ] } }
}

{ data : {
    "kind" : "prediction#output",
    "outputLabel":"Manymoon",
    "outputMulti" : [
    {"label":"OffiSync", "score": x.xx}
    {"label":"Zoho CRM", "score": x.xx}
    {"label":"MailChimp", "score": x.xx}]}}
Demo!
Predicting apps for a 501-1,000 seat educational institution
Predicting apps for a 501-1,000 seat educational institution
### Demo Screenshots

#### Predicting apps for a small business

<table>
<thead>
<tr>
<th>App Name</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SherpaTools for Google Apps (FREE!)</td>
<td>-24.8841</td>
</tr>
<tr>
<td>User Renamer for Google Apps</td>
<td>-25.4215</td>
</tr>
<tr>
<td>Power Panel for Google Apps</td>
<td>-26.0944</td>
</tr>
<tr>
<td>Dito Directory - Shared Contacts Manager</td>
<td>-26.7904</td>
</tr>
<tr>
<td>PromovigoPanel for Google Apps Administration</td>
<td>-26.9243</td>
</tr>
<tr>
<td>Aviary Design Suite (Free)</td>
<td>-27.1832</td>
</tr>
<tr>
<td>EasyBib Bibliography</td>
<td>-27.5909</td>
</tr>
<tr>
<td>OffiSync - [FREE] Integrate Microsoft Office with Google Apps</td>
<td>-27.7593</td>
</tr>
<tr>
<td>SurveyMonkey</td>
<td>-27.9503</td>
</tr>
<tr>
<td>Mojo Helpdesk</td>
<td>-28.4835</td>
</tr>
<tr>
<td>eStreamDesk Helpdesk</td>
<td>-28.5436</td>
</tr>
<tr>
<td>PlanbookEdu Lesson Planner</td>
<td>-28.7191</td>
</tr>
<tr>
<td>Jobite Recruiting Software</td>
<td>-28.7334</td>
</tr>
<tr>
<td>Creately - Online Diagramming and Design</td>
<td>-28.8297</td>
</tr>
<tr>
<td>Brokkit Learning Platform</td>
<td>-28.8383</td>
</tr>
<tr>
<td>Backups for Google Apps</td>
<td>-28.8383</td>
</tr>
</tbody>
</table>

**Perform a Prediction**

From here, you can predict the applications which would be appealing for a Google Apps domain:

- **Google Apps Edition**
  - **BUSINESS**
  - **EDUCATION**
- **Country**
  - **United States**
- **Language**
  - **English**
- **Company Type**
  - **Professional Services**
- **Company Size**
  - **50 or less**
Demo Screenshots

Predicting apps for a small business
Prediction API Capabilities

Data
• Input Features: numeric or unstructured text
• Output: up to hundreds of discrete categories, or continuous values

Training
• Many machine learning techniques
• Automatically selected
• Performed asynchronously

Access from many platforms:
• Web app from Google App Engine
• Apps Script (e.g. from Google Spreadsheet)
• Desktop app
Prediction API - Pricing

Free Quota in 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
• Absolute limit of 60,000 predictions per day
• $0.002 per MB trained (max size per dataset is 100MB)
Google Storage - Availability

Limited preview in US* currently
• Sign up for wait list at
  • http://code.google.com/apis/predict/

* Non-US preview available on case-by-case basis
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

Simplicity

Fast

Batch Jobs

BigQuery
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:

```json
{
    "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
- OAuth
- AuthSub

HTTPS support
- protects your credentials
- protects your data

Relies on Google Storage to manage access
Large Data Analysis Example

Wikimedia Revision History

BigQuery

```
SELECT TOP(title, 5), COUNT(*)
FROM [wikipedia]
WHERE wp_namespace = 0;
```

Wikimedia Revision history data from:
http://download.wikimedia.org/enwiki/latest/enwiki-latest-pages-meta-history.xml.7z
Using BigQuery Shell

Python DB API 2.0 + B. Clapper's sqlcmd
http://www.clapper.org/software/python/sqlcmd/

```sql
SELECT TOP(title, 5), COUNT(*) FROM [bigquery.test.001/tables/wikipedia]
WHERE wp_namespace = 0;
```

Execution time: 10.953 seconds

5 rows

<table>
<thead>
<tr>
<th>TOP(title, 5)</th>
<th>COUNT(*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>George W. Bush</td>
<td>43652</td>
</tr>
<tr>
<td>List of World Wrestling Entertainment employees</td>
<td>30572</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>29726</td>
</tr>
<tr>
<td>United States</td>
<td>27433</td>
</tr>
<tr>
<td>Michael Jackson</td>
<td>23245</td>
</tr>
</tbody>
</table>
BigQuery from a Spreadsheet

Here is a screenshot of a spreadsheet showing a query to BigQuery. The query is set to import data from a URL and then perform a search with the term 'google'. The results show the number of reviews for various Google products:

<table>
<thead>
<tr>
<th>Search Term</th>
<th>Num_Revs</th>
</tr>
</thead>
<tbody>
<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>
Google Fusion Tables
Google Fusion Tables

- Manage large collections of tabular data in the cloud
  - 100 Mb tables
  - Filters, Aggregation, Merge
  - ACL, Collaboration, Discuss Data
  - Visualizations
- REST API
  - Geo queries
- Maps Integration
  - FusionTablesLayer
Google Fusion Tables

Google fusion tables weather-stations.csv US National Climatic Data Center

File  View  Edit  Visualize  Merge

Current view: All  Show options

Location  LAT  LAT  Display as heat map  Configure info window  Configure styles  Export to KML  Get KML network link  Get embeddable link

![Map of world with numerous red dots indicating weather stations](image-url)
Google Visualization API
Google Visualization API

- Collection of JavaScript Visualization components
  - Some from Google (Chart Tools)
  - Some from other developers
  - Share the same wire protocol for Data Sources
Example: Weather data

- US National Climatic Data Center
  - weather data at stations around the globe since 1929
  - Stored in Google Storage
  - Created a Table for Bigquery
  - Upload Weather Station coordinates in Fusion Tables
- App Engine App
  - Maps API to display weather station Maps
  - Bigquery to query average temperature in January
  - A bit of Python to create a JSON Data Source
  - Visualization API
- Just an example: rinse, repeat, enhance!
Example: Weather data
Google Refine
Google Refine

- Power tool for working with messy data
  - Cleanup
  - Transform
  - Augment
  - (Link with FreeBase)
- Desktop software for now
- http://code.google.com/p/google-refine/
Google Refine

---

Google Refine interface showing a table with columns for Contractor Name, Type of Contract, Date of Award, Start Date, End Date, and Total value of Contract. The table contains data for various contractors and contracts with dates ranging from 2009 to 2011.

- ASAP SOFTWARE EXPRESS INC DELL MARKETING L.P.
- BMC SOFTWARE DISTRIBUTION INCORPORATED
- GOVCORPORATION INCORPORATED
- ITS CORPORATION
- SENET INTERNATIONAL CORPORATION
- IT FEDERAL SALES LIMITED LIABILITY COMPANY
- REDHAWK IT SOLUTIONS LLC

---

Options for filtering and sorting data are also visible in the interface, including Facet, Text facet, Numeric facet, Timeline facet, Scatterplot facet, Custom text facet, Custom numeric facet, and Customized facets.
Recap

- **Google App Engine**
  - Easy to build, deploy and manage web apps
- **Google Storage**
  - High speed data storage on Google Cloud
- **Prediction API**
  - Google's machine learning technology
- **BigQuery**
  - Interactive analysis of very large data sets
- **Google Fusion Tables**
  - Manage collections of tabular data in the cloud
- **Google Refine**
  - Power tool for working with messy data
- **Google Visualization**
  - Collection of JavaScript Visualization
More information

http://code.google.com/apis/
http://code.google.com/more/table/