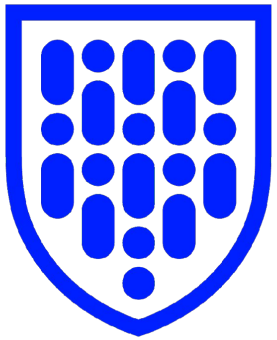


Institute of Data

2024

© 2024 Institute of Data

Data Science and AI

Module 3
Part 2:

APIs

© 2024 Institute of Data

Institute of Data

Agenda: Module 3 Part 2

- What is an API?
- APIs for data services
- APIs for analytic services
- APIs for visualisation services
- APIs for cognitive services
- Creating an API

© 2024 Institute of Data

Institute of Data

What is an API?

- Definition, examples
- Interfaces
- Authentication protocols
- Documentation

© 2024 Institute of Data

What is an API?

- What does “API” stand for?
  - Application Programming Interface

- Examples?
  - automation in Microsoft Office
    - e.g. generating a Word document or an Outlook reminder from another application
  - high-level database drivers
    - e.g. PyMongo
  - programming libraries for mobile & wearable devices
  - programmable web services
  - other?

© 2024 Institute of Data

Use Cases for APIs

- integrate remote data access
  - repetitive analyses of an **evolving dataset**
  - **up-to-the-moment** forecasting

- **integrate** familiar functionality
  - location sharing using Google Maps
  - simplified app login via Facebook
  - in-app purchases
  - in-app YouTube viewing

© 2024 Institute of Data

Some Popular Web Service APIs

Name	Nature	URL
Facebook	Networking, marketing	https://developers.facebook.com/tools/
Amazon S3	Cloud storage, Big Data analytics	https://aws.amazon.com/s3/
LinkedIn	Networking	https://developer.linkedin.com/
eBay	E-commerce	https://developer.ebay.com/
Google API Console	Data access & analytics, e-commerce, etc.	https://developers.google.com/apis-explorer/#p/
New York Times	News	http://developer.nytimes.com/

© 2024 Institute of Data

Interfaces for Web Service APIs

- SOAP
  - *Simple Object Access Protocol*
  - early, widespread web service protocol
  - exposes components of application logic as services
  - XML
- REST
  - *Representational State Transfer*
  - now > 70% of public APIs
  - accesses data
  - variety of data formats, coupled with JSON
  - generally faster and uses less bandwidth
  - easier to integrate with existing websites

Overview of RESTful API Description Languages:
Overview of RESTful API Description Languages - Wikipedia

roll your own:
REST API Tutorial
API Management - Amazon API Gateway - AWS

© 2024 Institute of Data

HTTP

- HyperText Transfer Protocol
- underlies RESTful APIs

- 4 major methods
  - GET fetches data from web server
  - PUT edits data on web server
  - POST adds new data
  - DELETE removes data

- HTTP Status Codes
  - 1xx informational
  - 2xx success
  - 3xx redirection
  - 4xx client error
  - 5xx server error

HTTP Status Codes

© 2024 Institute of Data

Elements of an API call

- ***endpoint***
  - URL of a server page that provides data or functionality via ***requests*** and ***responses***
- ***protocol***
  - the communication standard for passing requests to an endpoint
- ***authentication***
  - secure **identification** of user making request
  - if a developer creates an app for other users, the app needs to obtain **authorisation** from the owner of the API for both the developer's access *and* the user's access

© 2024 Institute of Data

Authentication Protocols

- HTTP Basic Access Authentication
  - username + password
  - transmitted in header of HTTP request
  - weakly encoded, no encryption
- OAuth 1.0
  - uses encrypted tokens
- OAuth 2.0
  - simpler, more robust than OAuth 1.0

© 2024 Institute of Data

OAuth 2.0

- token-based
  - e.g. ***client_id*** & ***client_secret***
  - allows a 3rd-party app to access a user's/developer's account **without knowing the account password**
  - allows an end-user to access an API via *your* app, using *their* token

- redirect URL
  - **registered** when app created
  - OAuth 2.0 service **returns user to this URL** after authorising (and issuing a user token)
  - protects access token from **interception**

Background - OAuth 2.0 Simplified

© 2024 Institute of Data

Developer Access

- some API's have **a developer mode** that may allow access without requesting a user token
- options for connect/request include:
  - use developer's *user_id* and *password*
  - use *app_id*, developer's *client_id*, developer's *secret*
- access granted **may** include
  - read developer's posts, comments, profile, etc.
  - post to developer's account
  - read other users' posts, comments, profiles, etc.

© 2024 Institute of Data

Python Libraries: Utilities

requests

- HTTP library (“elegant and simple”)
- http://www.python-requests.org/en/latest/
- returns JSON-formatted byte strings

json

- JSON ↔ lists, dictionaries
- https://docs.python.org/2/library/json.html

untangle, xmltodict

- parses XML to Pythonic data structures

BeautifulSoup (bs4)

- parses HTML, XML to Pythonic data structures

© 2024 Institute of Data

Python Libraries: API Wrappers

- simplify usage of APIs by introducing a Python API into the loop
- use data types & structures familiar to Python developers

pyfacebook

linkedin

praw (Reddit)

bucketstore (Amazon S3)

python-forecastio (weather)

foursquare (location-based networking)

GooPyCharts (Google Charts)

indeed (indeed.com)

kiteconnect (stock trading)

pymaps (Google Maps)

pymed (PubMed)

pyspotify (Spotify)

newsapi

rottentomatoes (crowd-based movie reviews)

sportradar (sport APIs)

tesserocr (OCR)

bowshock (NASA)

geopy (geocoding)

List of Python API Wrappers and Libraries

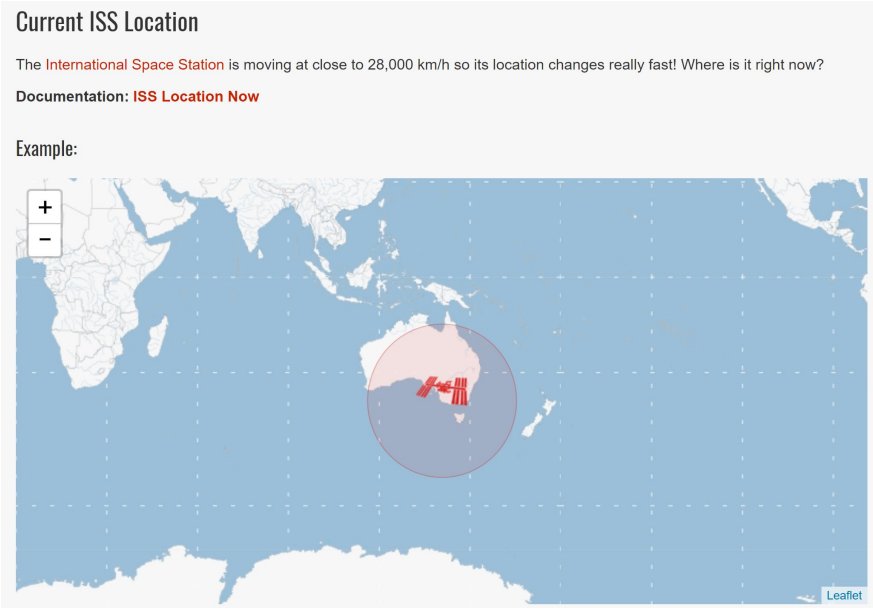
© 2024 Institute of Data

Lab 3.2.1: Querying the ISS

- Purpose:
  - To become familiar with basic API requests and responses

- Resources:
  - API for the International Space Station:
    OpenNotify
    Open Notify -- API Documentation
  - HTTP response codes
    HTTP Status Codes

- Materials:
  - ‘Lab 3.2.1.ipynb’



© 2024 Institute of Data

Institute of Data

Extracting Data from APIs

- Reddit API
- Google Public Data and BigQuery API

© 2024 Institute of Data

Institute of Data

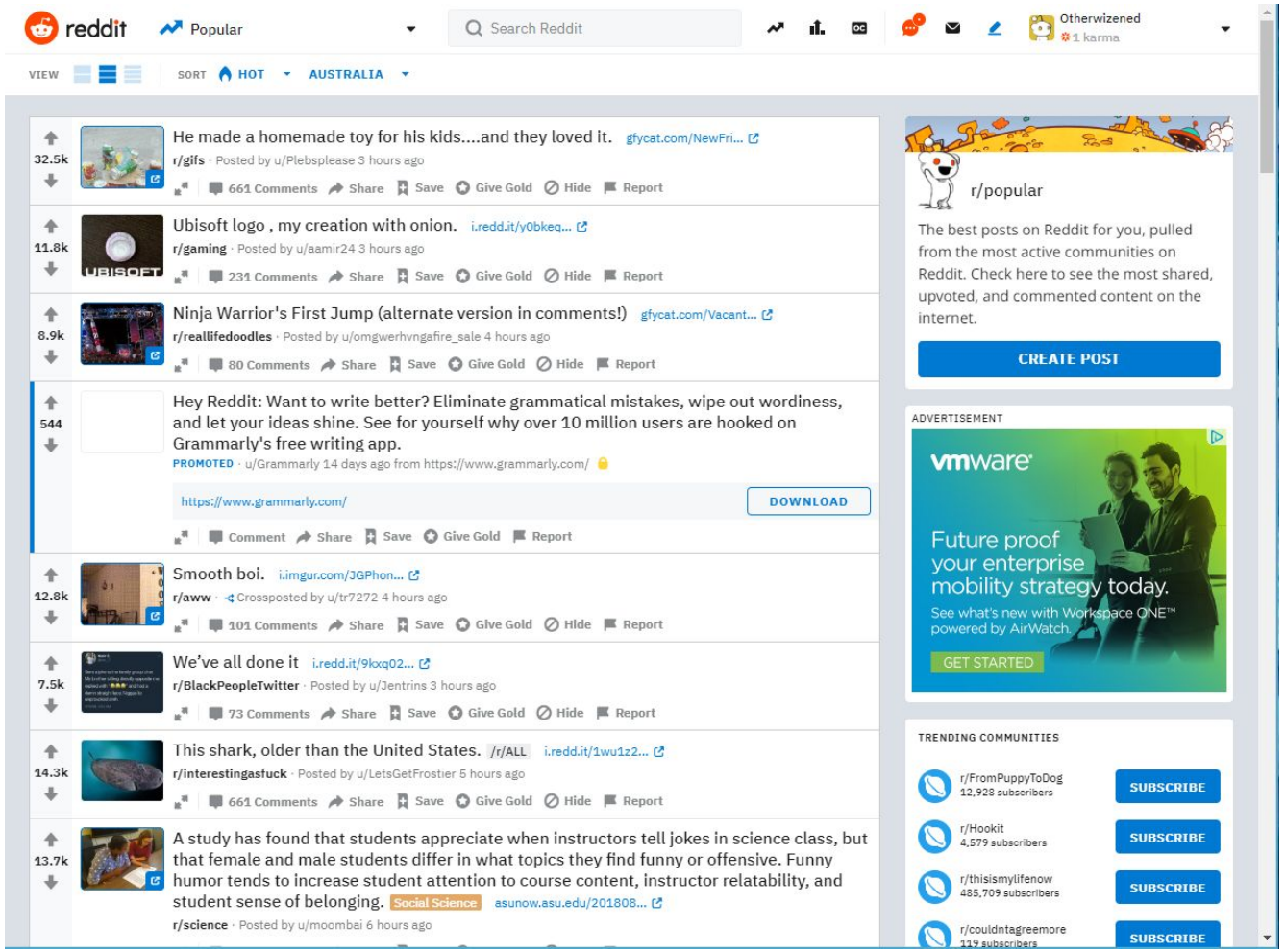
Reddit API

- Introduction to Reddit
- API structure
- Developer access
- Reddit API: Using Python

© 2024 Institute of Data

Reddit

- why Reddit?
  - good example of a social media product
  - rich content
  - large user base
  - highly structured API
  - immediately accessible

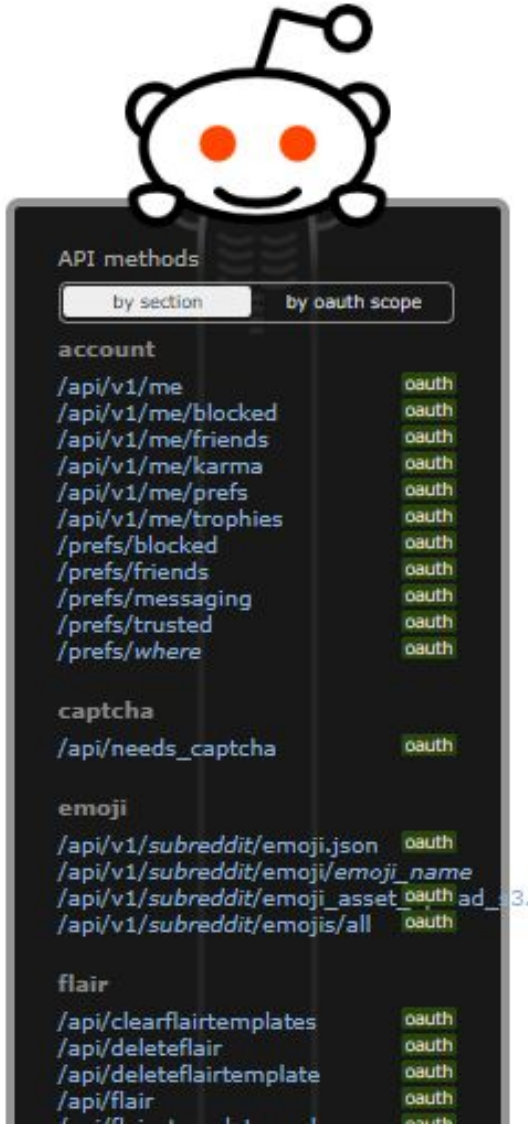


[faq - reddit.com]

© 2024 Institute of Data

Reddit API

- *Account* endpoints:
  - *me*, *me/friends*, *me/prefs*, ...
- Links & comments endpoint:
  - *comment*, *vote*, *report*, ...
- *Listing* endpoints:
  - categories
    - *hot*, *new*, *random*, ...
  - navigation (pagination) and filtering
    - *before*, *after*, *count*, *show*
- and many more ...



[reddit.com: api documentation]

© 2024 Institute of Data

Reddit API: Developer Access

1. Open a Reddit user account
2. Create a Reddit app
3. Register the app for API access
4. Store your credentials
   - for accessing your account:
     - user name
     - password
   - for authenticating your app:
     - user agent (information describing your app)
     - client ID (a unique identifier for your app)
     - client secret (secure token for authorising your app to access the API)

© 2024 Institute of Data

Reddit API: Using Python

- install PRAW package
- import praw
- create a connection object (to Reddit API)
- invoke API methods on the connection object
  - send requests that GET or PUT data to/from Reddit objects
- do something with data!

r/popular

faq - reddit.com

Quick Start - PRAW 7.7.1 documentation

© 2024 Institute of Data

Lab 3.2.2: Mining Social Media with Reddit

- Purpose:
  - To develop skills in using a media-rich API
- Resources:
  - Python library for Reddit API: ***PRAW***
    Quick Start - PRAW 7.7.1 documentation
- Materials:
  - 'Lab 3.2.2.ipynb'

© 2024 Institute of Data

Institute of Data

Google Cloud Platform

- public data sets / BigQuery
- APIs based on data science products

© 2024 Institute of Data

Google Cloud Platform

Google Cloud SDK	• gcloud CLI overview | Google Cloud CLI Documentation • Initializing the gcloud CLI | Google Cloud CLI Documentation
Google Cloud Platform	• GitHub - GoogleCloudPlatform/python-docs-samples: Code samples used on cloud.google.com • https://googlecloudplatform.github.io/google-cloud-python/ • https://googlecloudplatform.github.io/google-cloud-python/latest/
Google API Client Libraries	• https://developers.google.com/api-client-library/
Google BigQuery	• BigQuery public datasets Google Cloud • Query a public dataset with the Google Cloud console • BigQuery API Client Libraries | Google Cloud • Query a public dataset with the BigQuery client libraries | Google Cloud • https://github.com/googleapis/google-cloud-python/tree/main/bigquery

© 2024 Institute of Data

Google Public Data sets

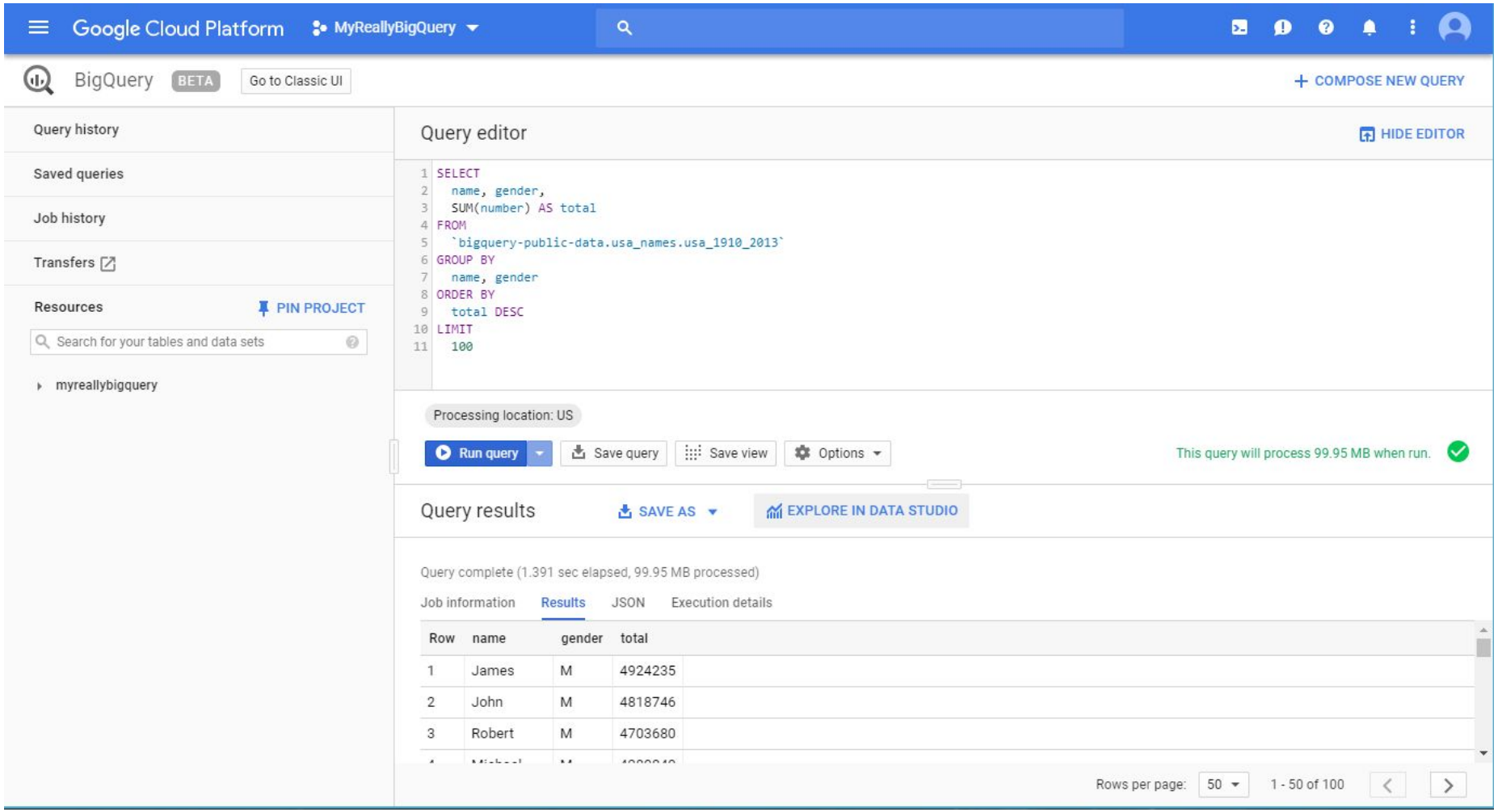
- accessible via Google BigQuery
- free for 1st TB / month
- subject areas:
  - genomics
  - medicine & epidemiology
  - geo imagery (Earth science, weather, etc.)
  - transport & service utilisation
  - annotated images
  - etc.
- Datasets and pre-built solutions | Google Cloud

© 2024 Institute of Data

Google BigQuery

Quickstart to BigQuery Web UI:

[Query a public dataset with the Google Cloud console | BigQuery]



© 2024 Institute of Data

BigQuery API: Authentication

Service accounts

- for client apps that you will run
  - e.g. dev/test, batch processing pipelines
- authentication via your service credentials

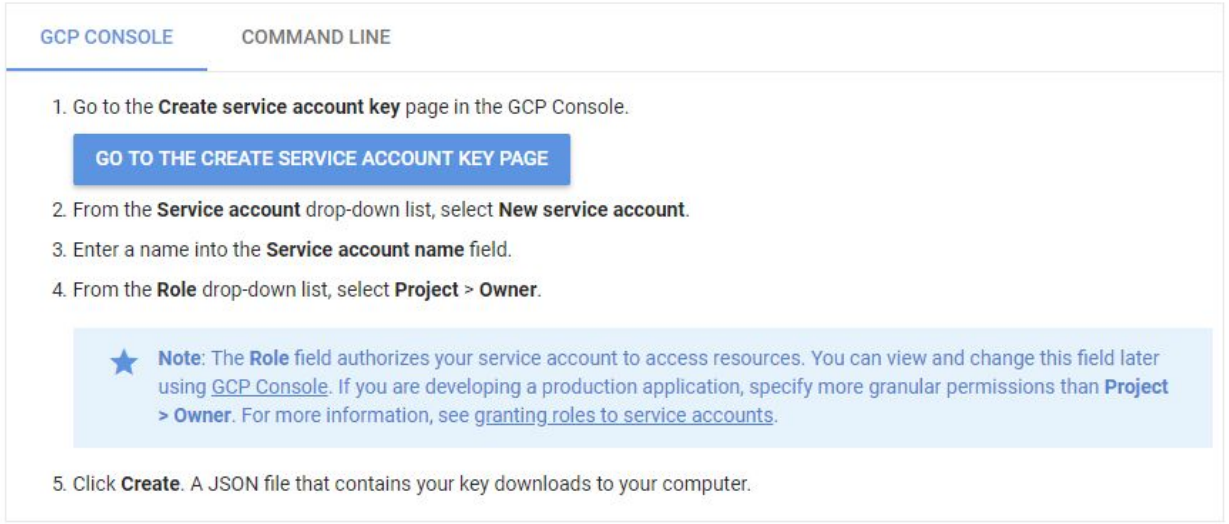
User accounts

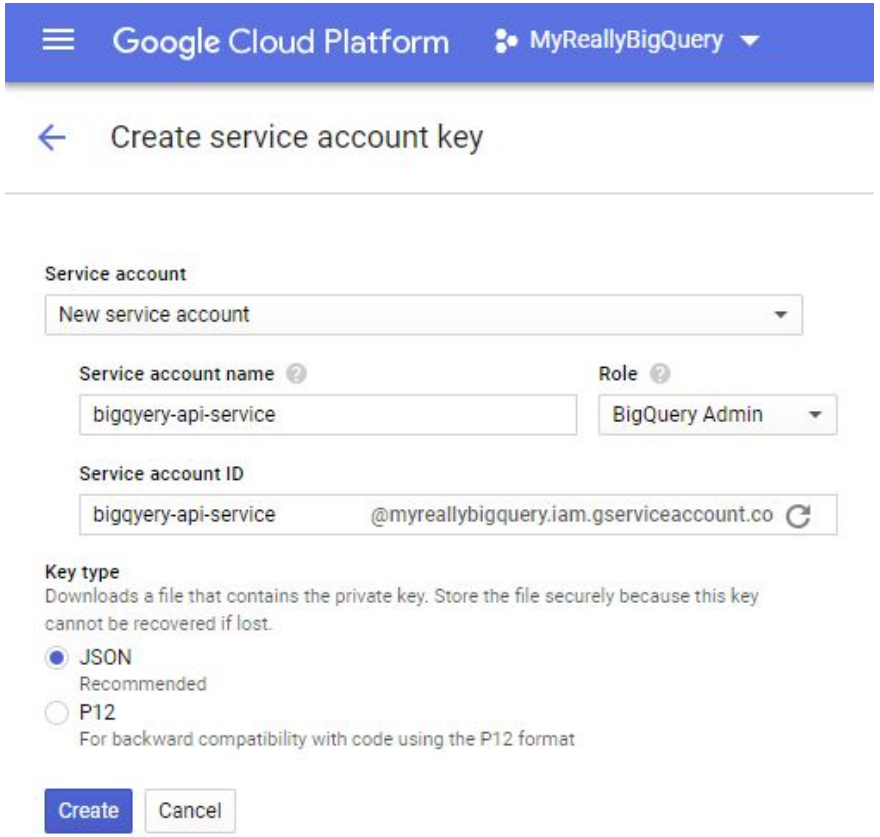
- for apps you create for other end-users
  - e.g. data products
- authentication via end-users credentials
  - app can only access BigQuery tables that the end-user is authorised to access
  - end-user gets billed for queries

Authenticate to BigQuery | Google Cloud

© 2024 Institute of Data

BigQuery API: Authentication – cont'd

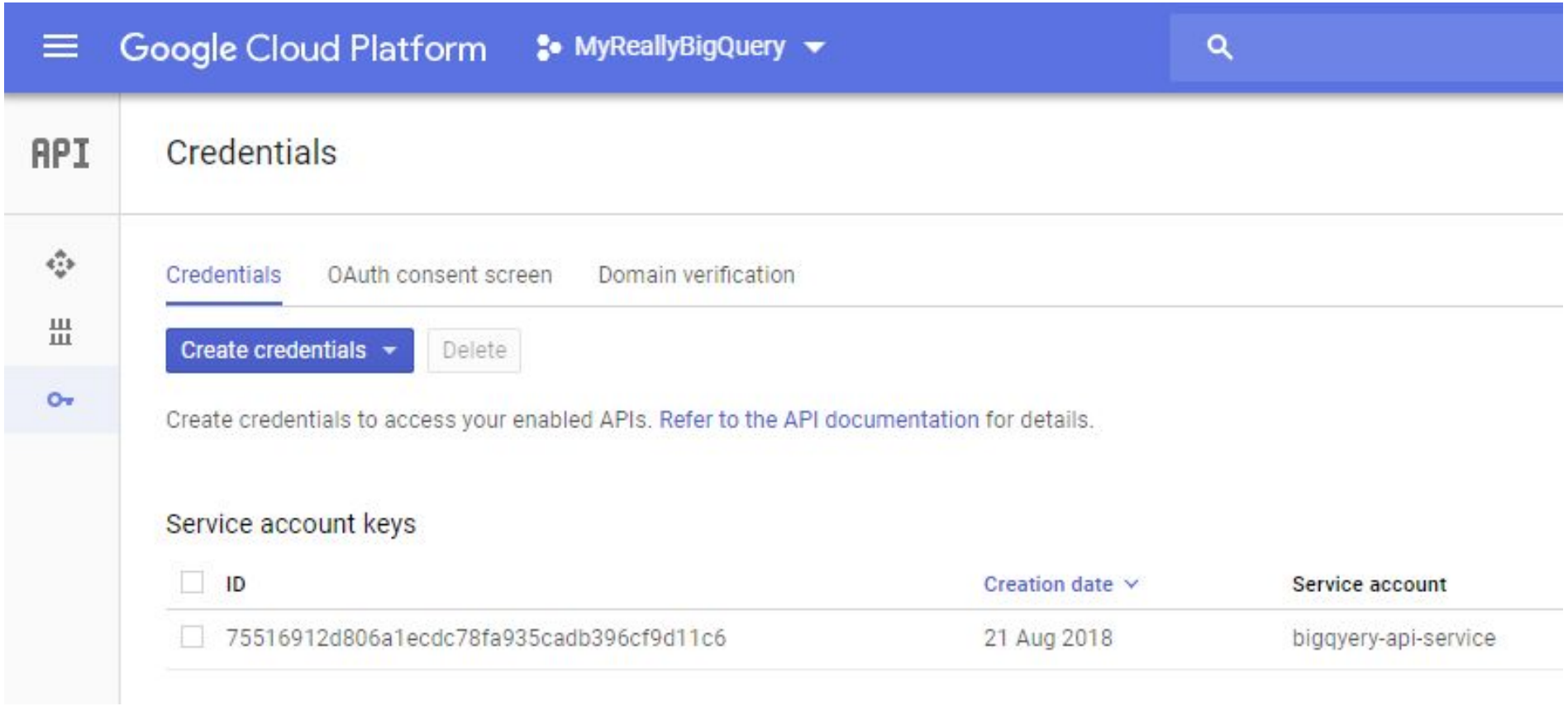




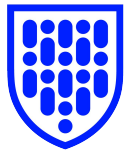
Authentication methods at Google

© 2024 Institute of Data

BigQuery API: Authentication – cont'd



© 2024 Institute of Data



Google BigQuery API: Top-Level Object

client object:

- connection
  - authenticated connection to the BigQuery service
  - determines credentials
    - implicitly from the environment,
    - or directly via *from_service_account_json* and *from_service_account_p12*
- project
  - top-level container
  - tied to billing
  - can provide default access control across all its datasets
  - access control list (ACL)
    - grants reader / writer / owner permission to one or more entities
    - must be managed using the Google Developer Console (not API)

© 2024 Institute of Data

BigQuery API Object Hierarchy

bigquery

- .projects
- .datasets
  - .get, .delete, .insert, .list, .update, ...
- .tabledata
- .tables
- .jobs
  - .get, .cancel, .insert, .list, .query, ...
- . . .

Google APIs Explorer

© 2024 Institute of Data

Lab 3.2.3: Big Data Analytics with BigQuery

- Purpose:
  - (1) To learn how to the Google BigQuery Web UI for discovering public data sets and performing basic analytics.
  - (2) To become proficient with the Google BigQuery API for wrangling Google's public datasets.

- Materials:
  - 'Lab 3.2.3.ipynb'

© 2024 Institute of Data

Lab 3.2.3 – cont'd

- Python packages :
  - pyarrow (pip)
  - google-cloud-bigquery (conda-forge)
  - google-cloud-storage (conda-forge)

- Resources:
  - Google BigQuery Public Datasets BigQuery public datasets | Google Cloud
  - BigQuery UI Query a public dataset with the Google Cloud console | BigQuery
  - Python client for BigQuery API https://github.com/GoogleCloudPlatform/google-cloud-python/tree/master/bigquery

© 2024 Institute of Data

Discussion

- Extracting data using APIs
  - applications?

© 2024 Institute of Data

Lab/ HOMEWORK

1. Create a mini-project based on any skills from the course so far:
   - select an interesting public data set or form a question you are interested answer and identify data needed to answer the question
   - use Jupyter Notebook to access, analyse and visualise the data
2. Prepare a 5-minute presentation
   - use Jupyter Notebook
   - organise as:
     - question
     - dataset & analysis
     - conclusion
3. plan to present to the class

© 2024 Institute of Data

Presentations

- each team
  - 5 minute presentation

© 2024 Institute of Data

Analytics-Based APIs

- **Google**
  - Google Analytics | Google for Developers
  - Computer vision: AI applications | Google Cloud
  - Product Directory: AI & Machine Learning | Google Cloud
- **IBM Watson**
  - IBM Watson
  - GitHub - watson-developer-cloud/python-sdk: :snake: Client library to use the IBM Watson services in Python and available in pip as watson-developer-cloud
  - https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=SP&infotype=PM&htmlfid=LBS03048USEN&attachment=LBS03048USEN.PDF

© 2024 Institute of Data

Analytics-Based APIs – cont'd

- AWS
  - AWS SDK for Python (Boto3)
    - low-level ("client") and high-level ("resource") APIs for all AWS products
  - Google APIs Explorer
- Azure
  - Code samples, Cognitive Services API, etc.
    - https://docs.microsoft.com/en-us/python/azure/?view=azure-python
  - Python API Browser
    - https://docs.microsoft.com/en-au/python/api/?view=azure-python

© 2024 Institute of Data

Machine Vision APIs

- use cases:
  - autonomous vehicles
  - industrial control & QA
  - face recognition
  - number plate recognition
  - biometric identity verification
  - print & handwriting transcription
  - image annotation
    - detecting and labelling objects or themes in an image

© 2024 Institute of Data

Creating APIs

- Why would a data scientist/engineer want to create their own API?
  - for building an interface to your data product
  - for enforcing control over how your application's data and services can be used
  - for isolating the IP that your data product is based on

- References:
  - Application Programming Interfaces - Full Stack Python

© 2024 Institute of Data

Discussion

More APIs

- List of Free APIs (Rapid API)
  23 Free Public APIs for Developers & Free Alternatives List - April, 2024 | RapidAPI

- Public APIs List
  ApisList

- toddmotto Public APIs
  https://github.com/toddmotto/public-apis

© 2024 Institute of Data

HOMEWORK

1. Investigate a data or analytic API for one of the following:
    - AWS
    - Microsoft Azure
    - IBM Cloud
2. Create a Jupyter notebook that demonstrates some basic operations (e.g. transporting, querying, or visualising data).

NOTES:

- The offerings of these platforms are myriad and complex. It may not be obvious which API you need to use at first, so try to start with published code examples.
- APIs (and the libraries that wrap them) change. Online examples may not work as documented.

© 2024 Institute of Data

Questions?

© 2024 Institute of Data

End of Presentation!

© 2024 Institute of Data