GENERAL ASSEMBLY

Automated Serverless Web Scraping on Google Cloud

By: Jason Yang
January 6, 2021

About Me

Jason Yang

Data Product Manager @ Hearst

- Cosmopolitan, ELLE, SF Chronicle

Data Scientist @ InMarket

- Consumer intelligence, digital advertising

github.com/jasonwithcoffee/jobmatch/tree/master/etl

linkedin.com/in/jasonwithcoffee

Summary

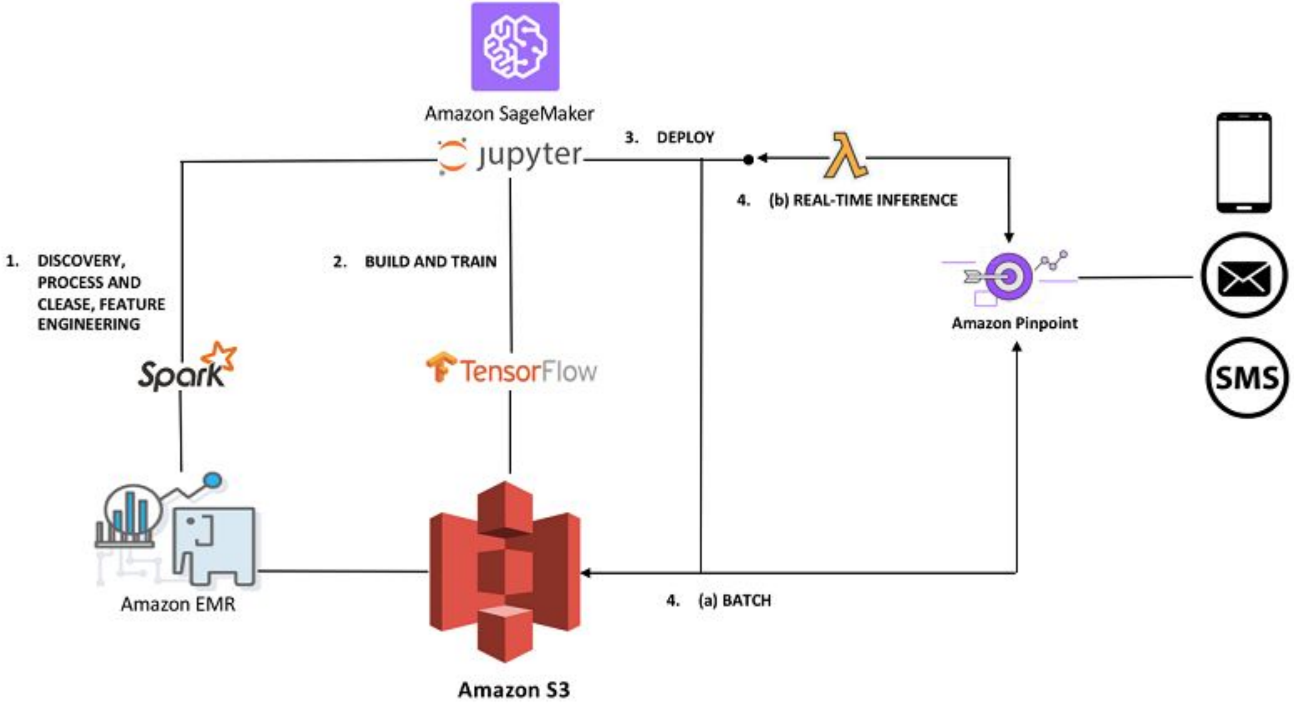
1. Machine Learning on AWS and Google Cloud (15 min)
2. Automated Serverless Web Scraping on Google Cloud (20-30 min)
3. Q&A (15-25 min)

Machine Learning on AWS and Google Cloud

Machine Learning on AWS and Google Cloud

- Select appropriate ML approach
- Identify appropriate services
- Design and implement, scalable, cost-optimized, and reliable ML solutions

Example: Amazon Pinpoint campaigns driven by machine learning on Amazon SageMaker



https://aws.amazon.com/blogs/machine-learning/amazon-pinpoint-campaigns-driven-by-machine-learning-on-amazon-sagemaker/

AWS ML Specialty

A data scientist uses logistic regression to build a fraud detection model. While the model accuracy is 99%, 90% of the fraud cases are not detected by the model.

What action will definitively help the model detect more than 10% of fraud cases?

A. Using undersampling to balance the dataset
B. Decreasing the class probability threshold
C. Using regularization to reduce overfitting
D. Using oversampling to balance the dataset

AWS ML Specialty

A data scientist uses logistic regression to build a fraud detection model. While the model accuracy is 99%, 90% of the fraud cases are not detected by the model.

What action will definitively help the model detect more than 10% of fraud cases?

A. Using undersampling to balance the dataset
B. Decreasing the class probability threshold
C. Using regularization to reduce overfitting
D. Using oversampling to balance the dataset

Google Cloud Professional ML Engineer

You work for a maintenance company and have built and trained a deep learning model that identifies defects based on thermal images of underground electric cables. Your dataset contains 10,000 images, 100 of which contain visible defects.

How should you evaluate the performance of the model on a test dataset?

A. Calculate the Area Under the Curve (AUC) value
B. Calculate the number of true positive results predicted by the model.
C. Calculate the fraction of images predicted by the model to have a visible defect.
D. Calculate the Cosine Similarity to compare the model's performance on the test dataset to the model's performance on the training dataset.

Google Cloud Professional ML Engineer

You work for a maintenance company and have built and trained a deep learning model that identifies defects based on thermal images of underground electric cables. Your dataset contains 10,000 images, 100 of which contain visible defects.

How should you evaluate the performance of the model on a test dataset?

A. **Calculate the Area Under the Curve (AUC) value**

B. Calculate the number of true positive results predicted by the model.

C. Calculate the fraction of images predicted by the model to have a visible defect.

D. Calculate the Cosine Similarity to compare the model’s performance on the test dataset to the model’s performance on the training dataset.

Jason’s Review of Cloud ML Certifications

- Learn about AWS, Google Cloud, and Azure
- Learn about Cloud ML
  - Select appropriate ML approach
  - Identify appropriate services
  - Design and implement, scalable, cost-optimized, and reliable ML solutions
- Learn about data science and machine learning

- Certification for job hunting

Automated Serverless Web Scraping on Google Cloud

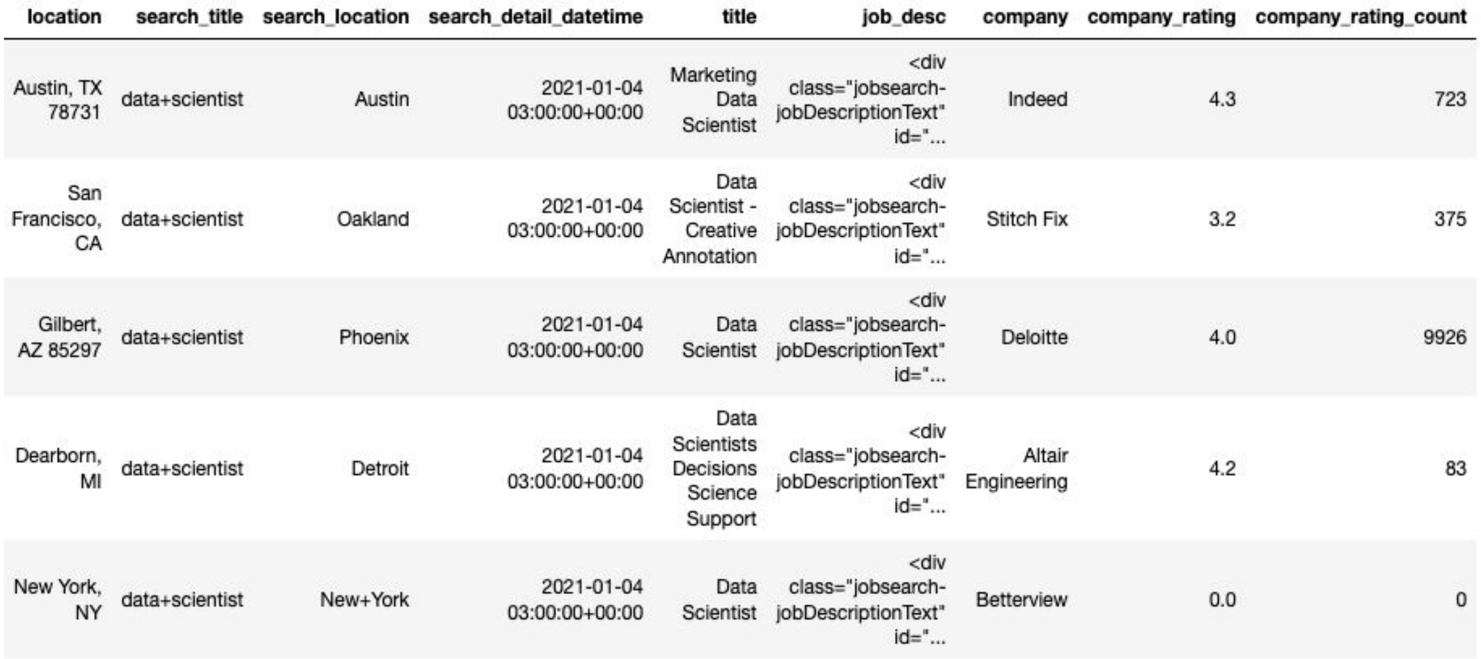
Automated Serverless Web Scraping on Google Cloud

Why?

What is the differences for

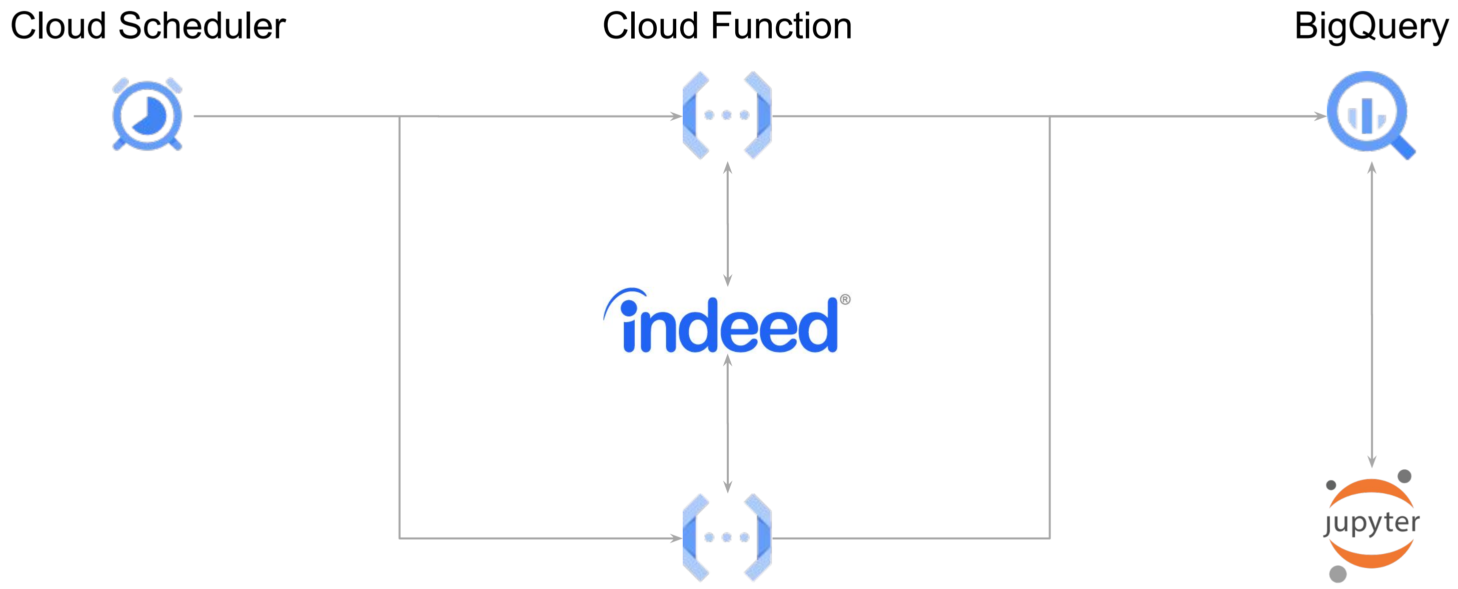
BI Analyst, BI Developer, Data Analyst, Data Scientist, Data Engineer, Machine Learning Engineer

Web Scraped Data

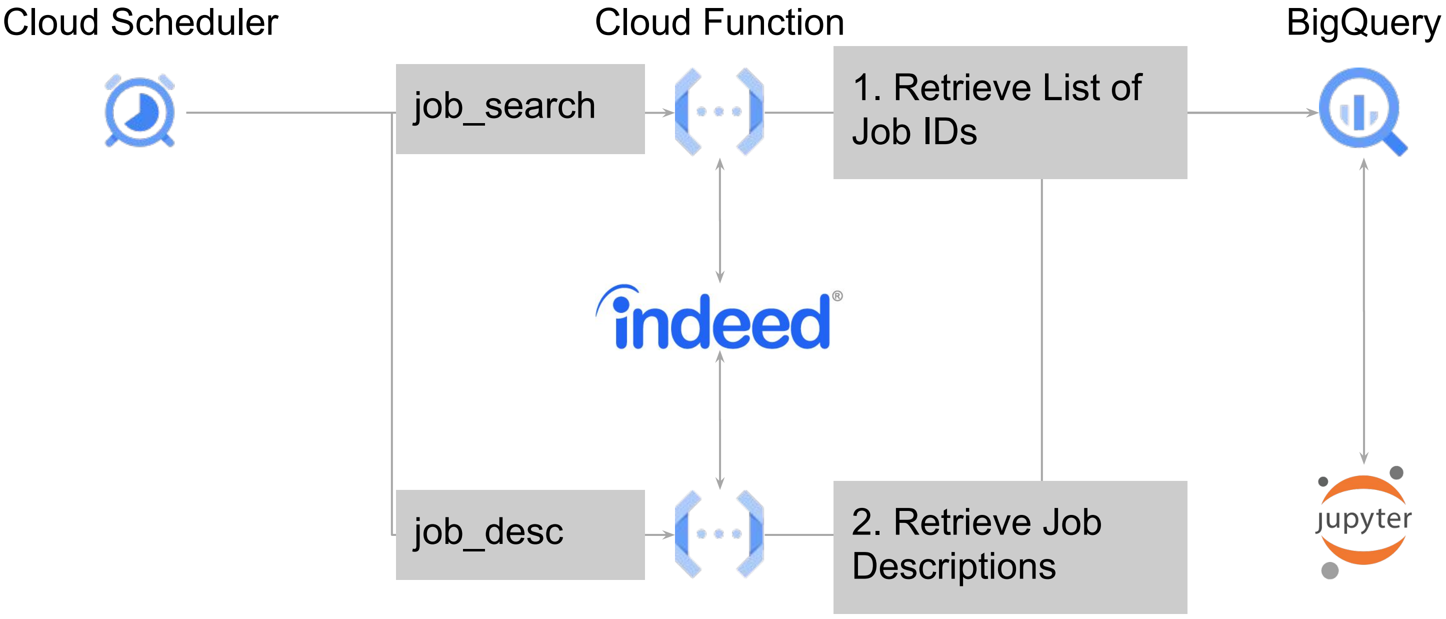


location	search_title	search_location	search_detail_datetime	title	job_desc	company	company_rating	company_rating_count
Austin, TX 78731	data+scientist	Austin	2021-01-04 03:00:00+00:00	Marketing Data Scientist	<div class="jobsearch-jobDescriptionText" id="...	Indeed	4.3	723
San Francisco, CA	data+scientist	Oakland	2021-01-04 03:00:00+00:00	Data Scientist - Creative Annotation	<div class="jobsearch-jobDescriptionText" id="...	Stitch Fix	3.2	375
Gilbert, AZ 85297	data+scientist	Phoenix	2021-01-04 03:00:00+00:00	Data Scientist	<div class="jobsearch-jobDescriptionText" id="...	Deloitte	4.0	9926
Dearborn, MI	data+scientist	Detroit	2021-01-04 03:00:00+00:00	Data Scientists Decisions Science Support	<div class="jobsearch-jobDescriptionText" id="...	Altair Engineering	4.2	83
New York, NY	data+scientist	New+York	2021-01-04 03:00:00+00:00	Data Scientist	<div class="jobsearch-jobDescriptionText" id="...	Betterview	0.0	0

Architecture Diagram

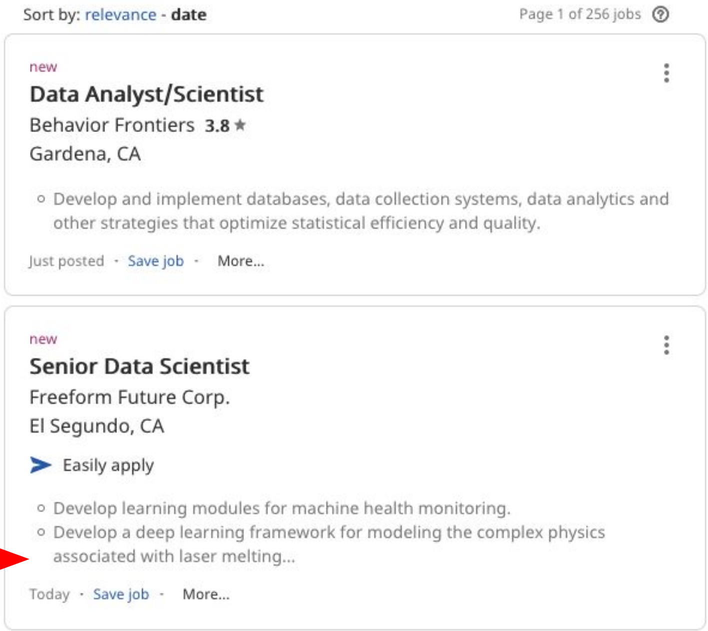


Architecture Diagram



GCF: job_search - Retrieve List of Job IDs

https://www.indeed.com/jobs?q=data+scientist&l=los+angeles&sort=date



Sort by: relevance - date

Page 1 of 256 jobs

new

Data Analyst/Scientist

Behavior Frontiers 3.8 ★

Gardena, CA

- Develop and implement databases, data collection systems, data analytics and other strategies that optimize statistical efficiency and quality.

Just posted · Save job · More...

new

Senior Data Scientist

Freeform Future Corp.

El Segundo, CA

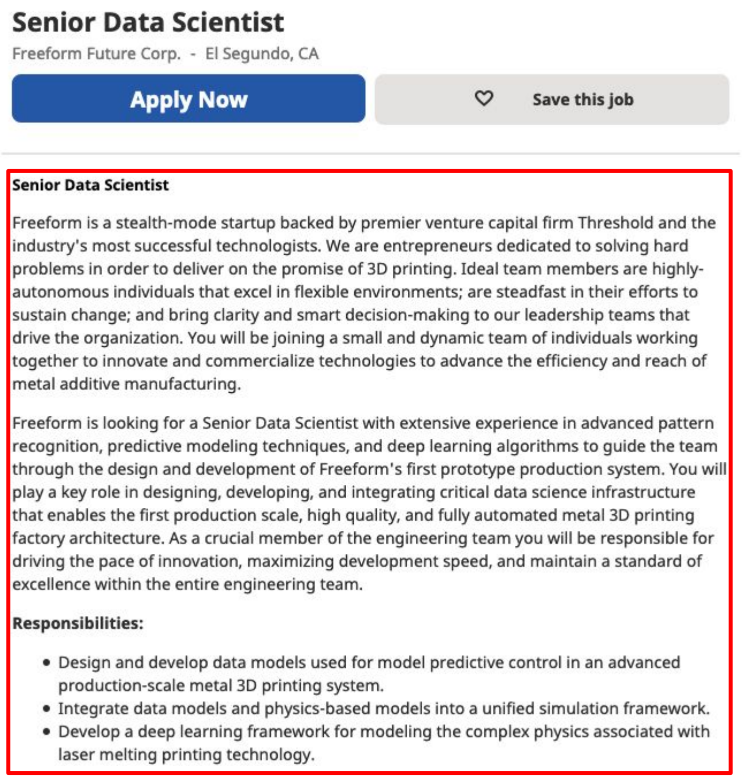
Easily apply

- Develop learning modules for machine health monitoring.
- Develop a deep learning framework for modeling the complex physics associated with laser melting...

Today · Save job · More...

GCF: job_desc - Retrieve List of Job Descriptions

https://www.indeed.com/viewjob?jk=158beffdc5845669



Senior Data Scientist

Freeform Future Corp. - El Segundo, CA

Senior Data Scientist

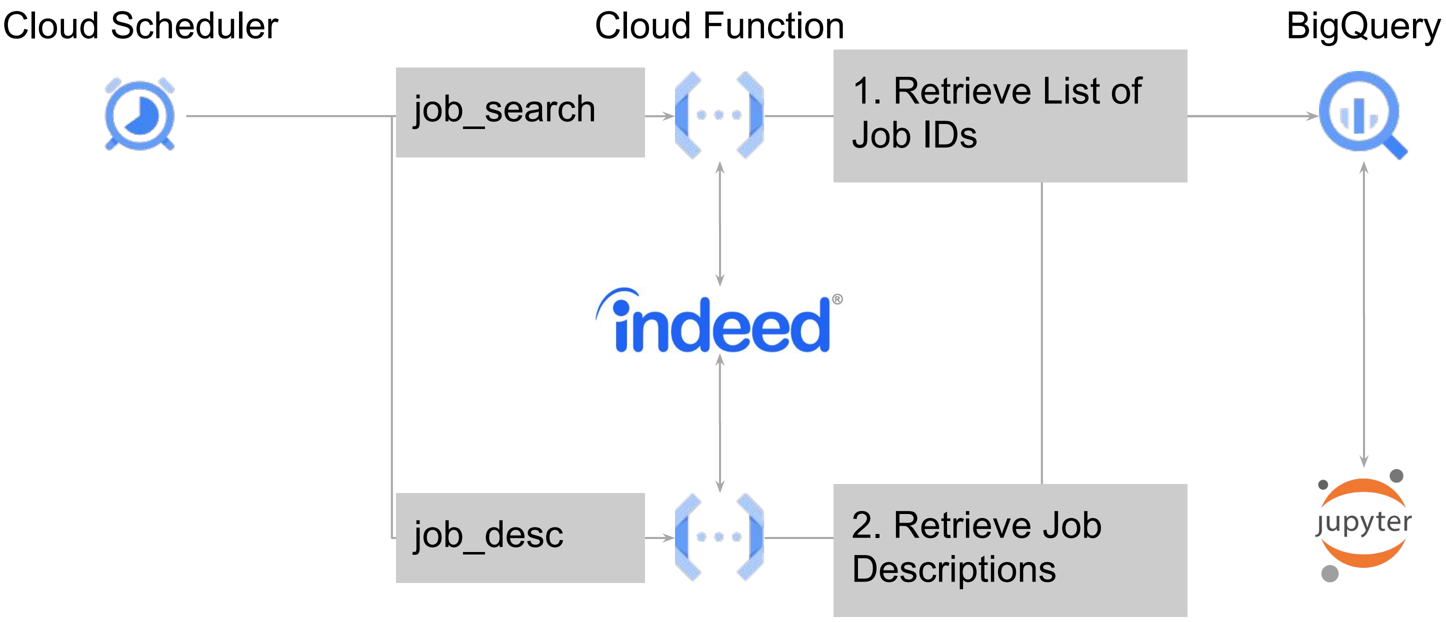
Freeform is a stealth-mode startup backed by premier venture capital firm Threshold and the industry's most successful technologists. We are entrepreneurs dedicated to solving hard problems in order to deliver on the promise of 3D printing. Ideal team members are highly-autonomous individuals that excel in flexible environments; are steadfast in their efforts to sustain change; and bring clarity and smart decision-making to our leadership teams that drive the organization. You will be joining a small and dynamic team of individuals working together to innovate and commercialize technologies to advance the efficiency and reach of metal additive manufacturing.

Freeform is looking for a Senior Data Scientist with extensive experience in advanced pattern recognition, predictive modeling techniques, and deep learning algorithms to guide the team through the design and development of Freeform's first prototype production system. You will play a key role in designing, developing, and integrating critical data science infrastructure that enables the first production scale, high quality, and fully automated metal 3D printing factory architecture. As a crucial member of the engineering team you will be responsible for driving the pace of innovation, maximizing development speed, and maintain a standard of excellence within the entire engineering team.

Responsibilities:

- Design and develop data models used for model predictive control in an advanced production-scale metal 3D printing system.
- Integrate data models and physics-based models into a unified simulation framework.
- Develop a deep learning framework for modeling the complex physics associated with laser melting printing technology.

Architecture Diagram



Cloud Scheduler

Cloud Function

BigQuery

job_search

job_desc

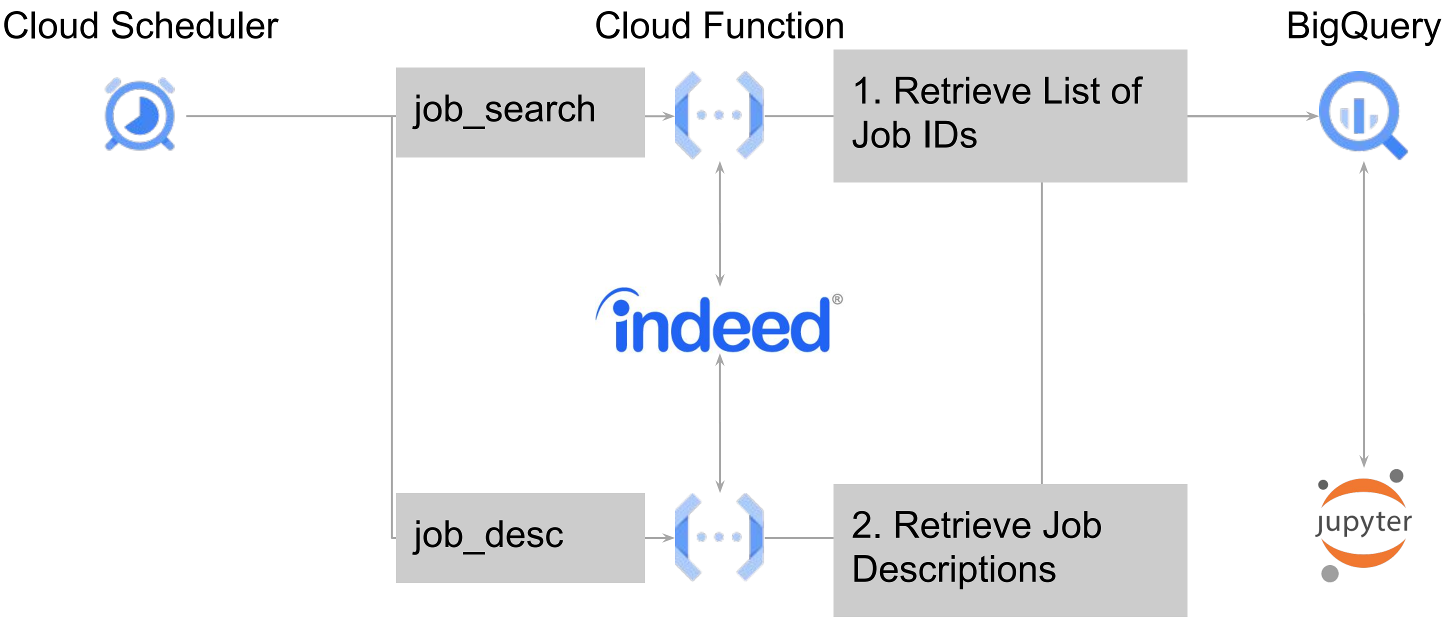
1. Retrieve List of Job IDs

2. Retrieve Job Descriptions

Automated Serverless Web Scraping on Google Cloud

DEMO

Architecture Diagram



Cloud Scheduler

Cloud Function

BigQuery

job_search

1. Retrieve List of Job IDs

job_desc

2. Retrieve Job Descriptions

Q&A

COST SAVING
TRACKING
NONE
IN HEAD
LEVELER/ CALCULATOR
PHOTOS TO STORE
PHOTOS PRINT STORE
INSPIRATION
SKILL LEVELS