



UNIVERSITY OF
TORONTO

INSPIRING
INCLUSIVE
EXCELLENCE

Article Recommendation System

University of Toronto 3760 Term Project | April 2022
Adon/Steven

BUSINESS CASE

- Build an article recommendation system that **EVERY** publisher can use

Canada's hospital capacity crisis will remain long after the pandemic is over

To solve the country's capacity problem, experts say, leaders need to finally confront the deeper flaws in how Canadian health care is structured

Related articles

Ontario hospitals keeping mandatory COVID-19 vaccination for staff, some for visitors
FEBRUARY 17, 2022 

Alberta health minister says COVID-19 pressure remains high but easing on hospitals
FEBRUARY 10, 2022 

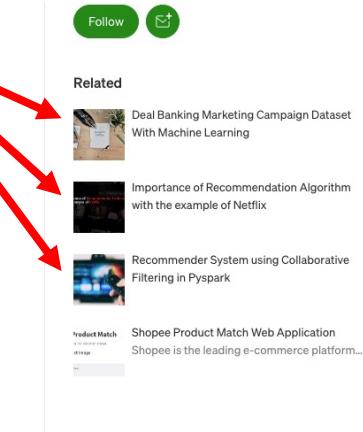
Shortage of blood-test materials leaves hospitals scrambling to treat patients
FEBRUARY 11, 2022 

More from Canada CANADA >
OPINION
For Alberta's Jason Kenney, the bad news never ends 

Article Recommendation System Using Python

The main motivation for me to pick up this topic for my post is probably because I learned the value of reading late in life, and after I did, I regret not doing so since my school. Now by reading, I mean books and or articles outside my education curriculum. Most of them know the value of reading, as for me, it really changed my viewpoint to various things in life that I was taught by family and society around me. It introduced a new perspective, and I started to question the norms, which otherwise I never did. Most importantly, when I started making reading a daily habit, it also trained my mind to be analytical and make decisions based on critical thinking.

In earlier years, between 2009–2013, I was only interested in reading books; however, over the past few years, I realized some good reading stuff that exists over the internet, like Scholarly writeups, Ribbon Farm, etc., motivates me.



The screenshot shows a blog post titled "Article Recommendation System Using Python". At the top right, there are "Follow" and "Email" buttons. Below the title, under the heading "Related", are three links with small thumbnail images:

- Deal Banking Marketing Campaign Dataset With Machine Learning
- Importance of Recommendation Algorithm with the example of Netflix
- Recommender System using Collaborative Filtering in PySpark

At the bottom of the sidebar, there is a link to "Shopee Product Match Web Application" with the note "Shopee is the leading e-commerce platform...".

WHY PUBLISHER NEED IT?

- Help viewers to find desired articles quickly
 - **Better** customer satisfaction
- Increase viewer engagement
 - **Higher** impression/ads revenue
 - **More** subscribers

REVENUE AND COST ANALYSIS

- Assumptions:

- Publisher gets 1M page views and 10M impressions on monthly basis
- Subscriber conversion 0.002% per page views
- RPM (Revenue Per thousand Impressions) is \$10
- LTV (Lifetime Value) per subs \$400

Recommendation System Revenue:

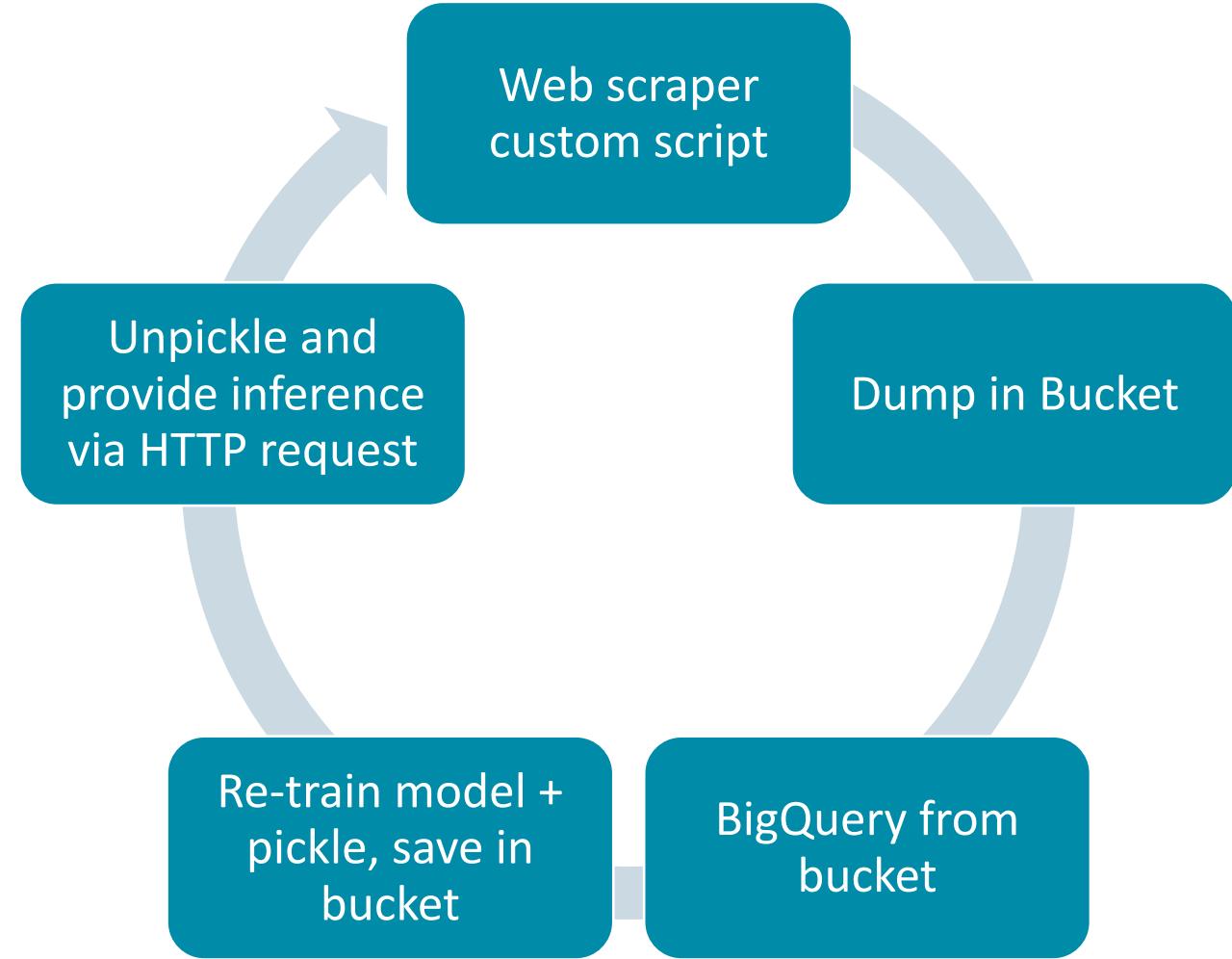
	Average Case	Best Case
% of User Clicks Recommended article	5%	10%
Extra Impressions	500,000	1,000,000
Extra RPM(\$)	<u>\$5,000</u>	<u>\$10,000</u>
Extra Subscribers	1	2
Extra Subs Revenue	<u>\$400</u>	<u>\$800</u>
Total Extra Revenue	<u>\$5,400</u>	<u>\$10,800</u>

WHY CLOUD SERVICE?

- Most of publishers have already moved to online and they are already using website analytic tool (eg. Google analytic) to collect data
- # of pageviews are unpredictable and it keeps changing over time, # of inference from ML model need **scale up/down** depending on pageviews
- Highly sensitive data can be **encrypted** and **protected** by VPN
- The same recommender system be **replicated** to a new publisher quickly

SOLUTION ARCHITECTURE OVERVIEW

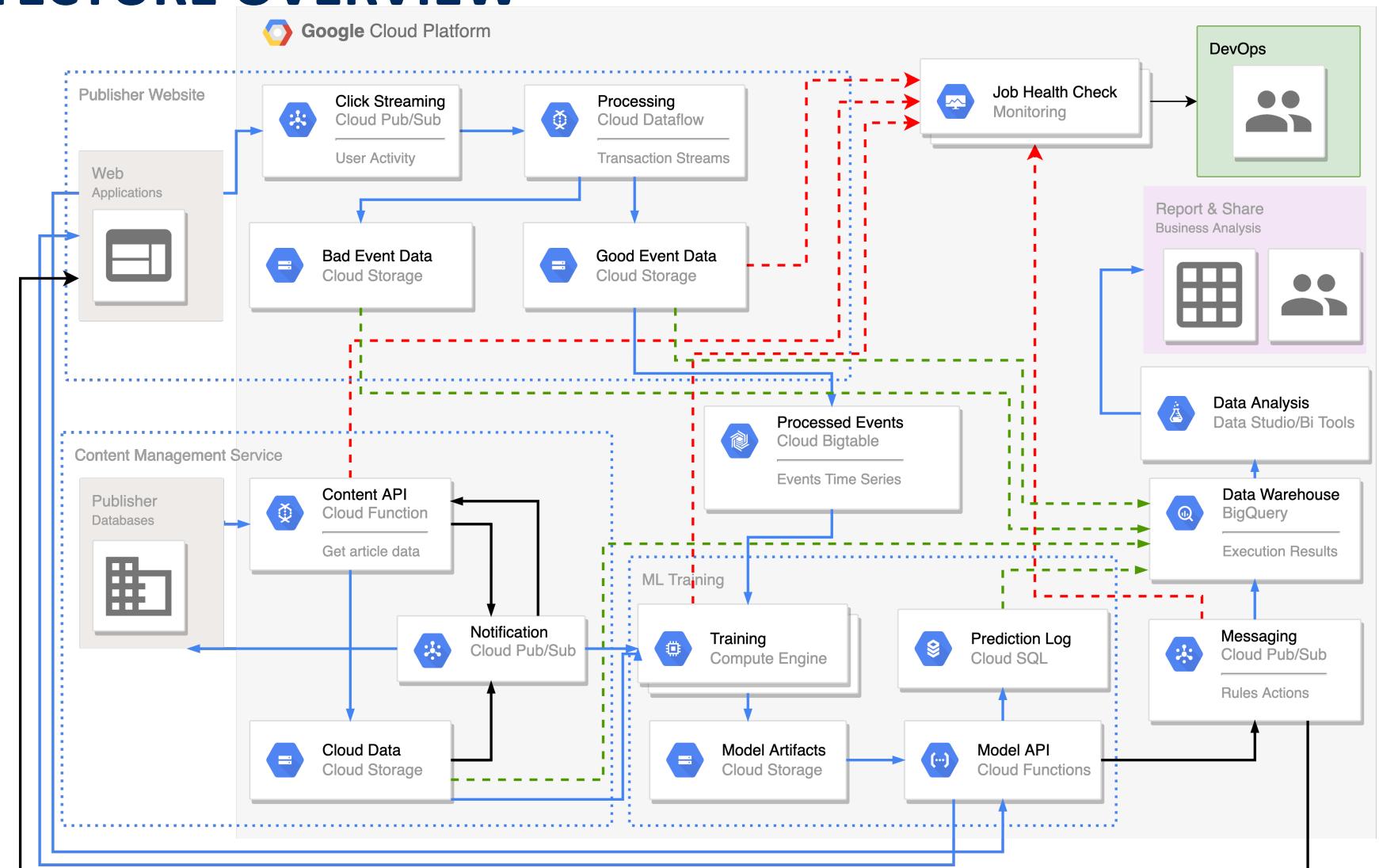
- (low resolution summary)



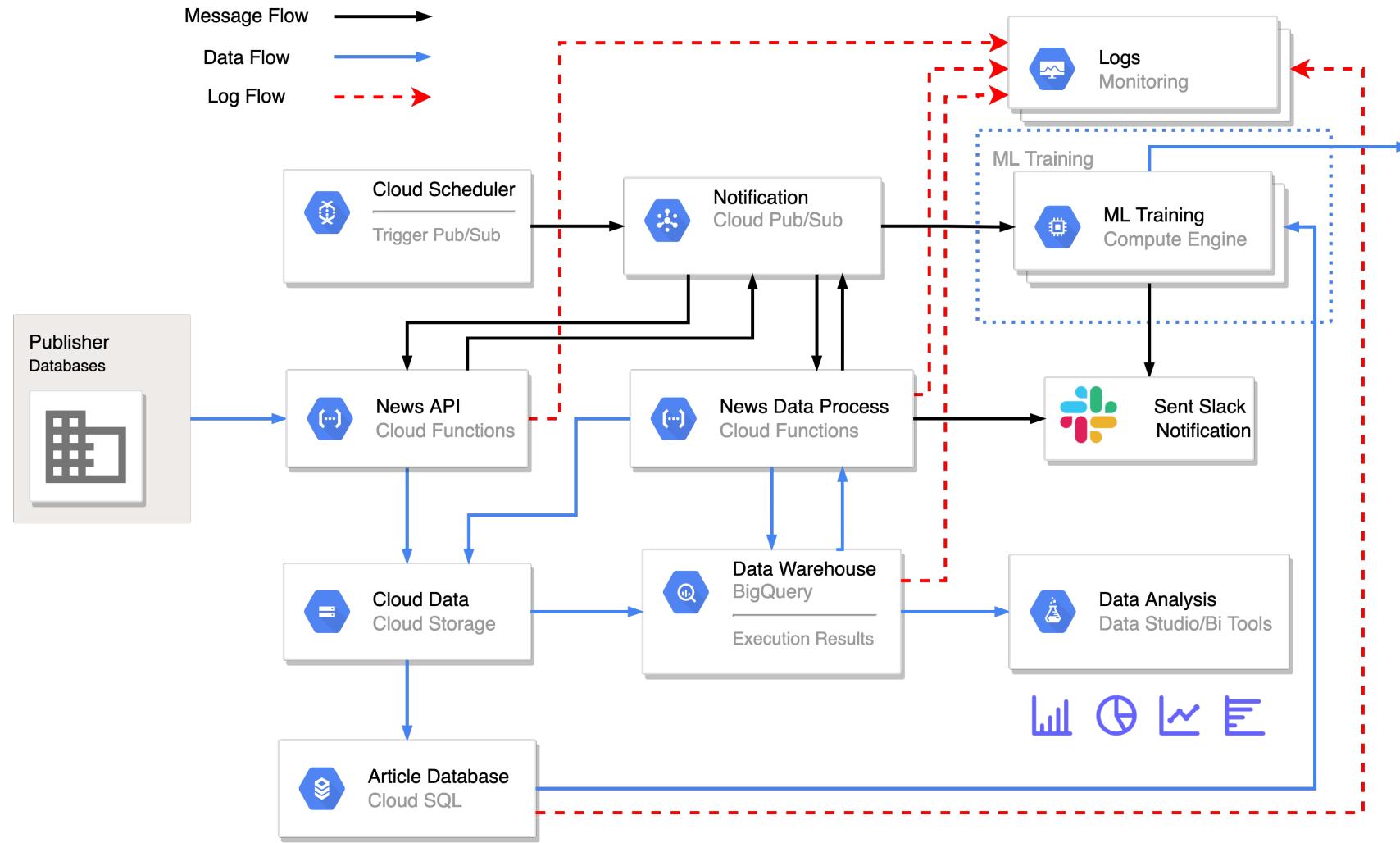
SOLUTION ARCHITECTURE OVERVIEW

The solution is implemented in GCP using:

- *Cloud Functions*
- *Cloud Storage*
- *BigQuery*
- *Cloud SQL*
- *Pub/sub scheduler*



CONTENT MANAGEMENT



IMPLEMENTATION

- Cloud Scheduler
- Pub/sub #1
- Cloud function #1 (News API)
 - Get News from Publisher API
 - Save csv files to Cloud Storage
- Pub/sub #2
- Cloud Function #2
 - Run BigQuery get data from Cloud Storage
 - Process Data
 - Save Processed Data to Cloud Storage
- Update data to Cloud SQL
- Pub/Sub #3
- Train model

SCHEDULER JOBS

Name	Region	State	Description	Frequency	Target
newsapi	northamerica-northeast1	Enabled	call news api	30 */6 * * *	Topic : projects/uoft3760project/topics/API-trigger

API-function Version 18, deployed at Apr 10, 2022, 12:49:28...

content-bigquery Version 4, deployed at Apr 10, 2022, 12:46:20 A...

content-bigquery Version 4, deployed at Apr 10, 2022, 12:46:20 A...

```

1 import base64
2 import json
3 import os
4 from google.cloud import storage
5 from newsapi import NewsApiClient
6 import pandas as pd
7 from datetime import datetime, timedelta
8 from google.cloud import pubsub_v1
9
10
11 # Instantiates a Pub/Sub client
12 publisher = pubsub_v1.PublisherClient()
13 PROJECT_ID = os.getenv('uoft3760project')
14
  
```

```

1 import requests
2 from google.cloud import storage
3 from google.cloud import bigquery
4 import db_dtypes
5 import pandas as pd
6 from datetime import datetime
7
8 bqclient = bigquery.Client()
9
10 def upload_blob(bucket_name, source_file):
11     """Uploads a file to the bucket."""
12
  
```

IMPLEMENTATION CONT. (BIGQUERY)

Create table

Source

Create table from Google Cloud Storage

Select file from GCS bucket or use a URI pattern * articlebuckets/*.csv

File format CSV

Source Data Partitioning

Destination

Project * uoft3760project

Dataset * newsapi

Table * Newtables

Unicode letters, marks, numbers, connectors, dashes or spaces

Table type External table

Explorer + ADD DATA

Type to search

Viewing pinned projects.

- uoft3760project
 - Saved queries (2)
 - content_summary_table
 - newsapi
 - Newsfeed
 - processed_articles
 - bigrquery-public-data
 - housecanary-com
 - listenbrainz
 - uoftproject-steven

EDITOR X PROCESS... X NEWSFEED X

Newsfeed QUERY SHARE

SCHEMA DETAILS

Table schema

Field name	Type	Mode	P
int64_field_0	INTEGER	NULLABLE	
source	STRING	NULLABLE	
author	STRING	NULLABLE	
title	STRING	NULLABLE	
description	STRING	NULLABLE	
url	STRING	NULLABLE	
urlToImage	STRING	NULLABLE	
publishedAt	TIMESTAMP	NULLABLE	
content	STRING	NULLABLE	

EDITOR X PROCESS... X NEWSFEED X

Newsfeed PROCESS... X

Type to search

Viewing pinned projects.

uoft3760project

Saved queries (2)

Project queries

2022-04-08 14:30:58

processed_article_q...

content_summary_table

newsapi

Newsfeed

processed_articles

bigrquery-public-data

housecanary-com

listenbrainz

uoftproject-steven

X PROCESS... X NEWSFEED X

Newsfeed PROCESS... X

1 SELECT * FROM `uoft3760project.newsapi.processed_articles`

Processing location: northamerica-northeast2

Query results SAVE RESULTS EXPLORE DATA

JOB INFORMATION RESULTS JSON EXECUTION DETAILS

id cagecategory headline

news-com-au general Motorcyclist films shocking crash after li

google-news general Show HN: An easy way to drag and drop

new-york-magazine general Are the Sanctions Against Russia Workin

the-globe-and-mail general Ottawa's tax on banks and life insurers to and Mail

the-globe-and-mail general Your daily horoscope: April 8 - The Globe



IMPLEMENTATION CONT. (CLOUD STORAGE)

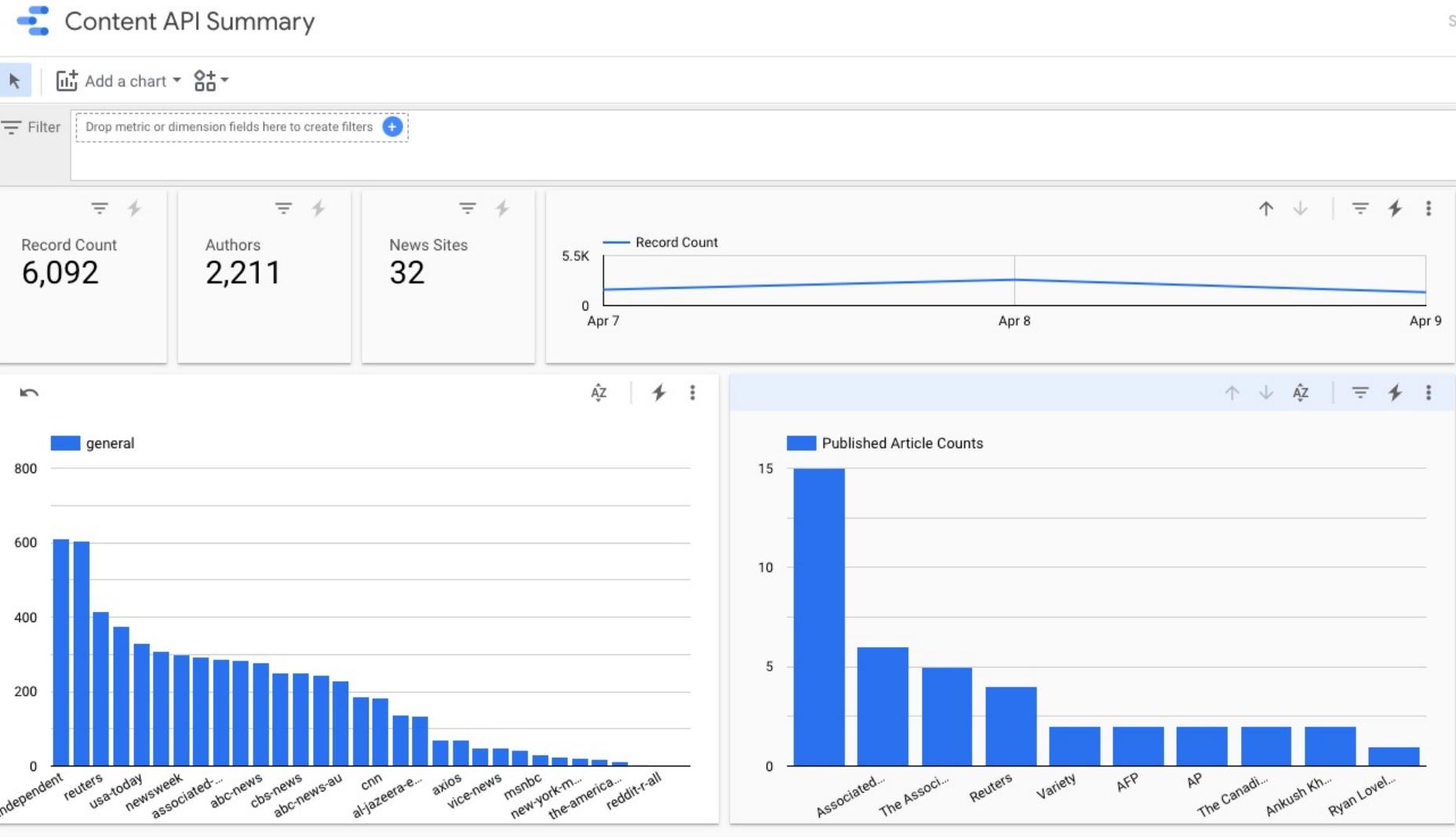
The screenshot shows the Google Cloud Platform Bucket details page for a bucket named 'articlebuckets'. The bucket is located in 'northamerica-northeast2 (Toronto)' with a 'Standard' storage class, 'Not public' public access, and no protection. The 'OBJECTS' tab is selected, showing a list of objects. The objects are all named 'abc-news-au2022_04_08-30.c...' and are text/csv files, each 77.4 KB in size, created on April 8, 2022. The objects are listed in descending order of creation date.

Name	Size	Type	Created	Storage class
abc-news-au2022_04_08-30.c...	78.3 KB	text/csv	Apr 8, 202...	Standard
abc-news-au2022_04_08-30.c...	77.6 KB	text/csv	Apr 8, 202...	Standard
abc-news-au2022_04_08-30.c...	77.6 KB	text/csv	Apr 8, 202...	Standard
abc-news-au2022_04_08-30.c...	77.5 KB	text/csv	Apr 8, 202...	Standard
abc-news-au2022_04_08-30.c...	77.4 KB	text/csv	Apr 8, 202...	Standard
abc-news-au2022_04_08-30.c...	77.4 KB	text/csv	Apr 8, 202...	Standard
abc-news-au2022_04_08-30.c...	77.4 KB	text/csv	Apr 8, 202...	Standard

WORD TO VECTOR MODEL

- We are using `sklearn.feature_extraction.text.TfidfVectorizer` and `sklearn.metrics.pairwise_distances`
 - `TfidfVectorizer` takes all the “words” from each article to create a matrix of term frequency
 - `Pairwise distances` is used to calculate the Euclidian distance between two matrices
- The final model is weighted:
 - 0.6x headline and short description
 - 0.2x category
 - 0.1x author
 - 0.1x publish day

DATA ANALYSIS



Thank you!
Any Questions?