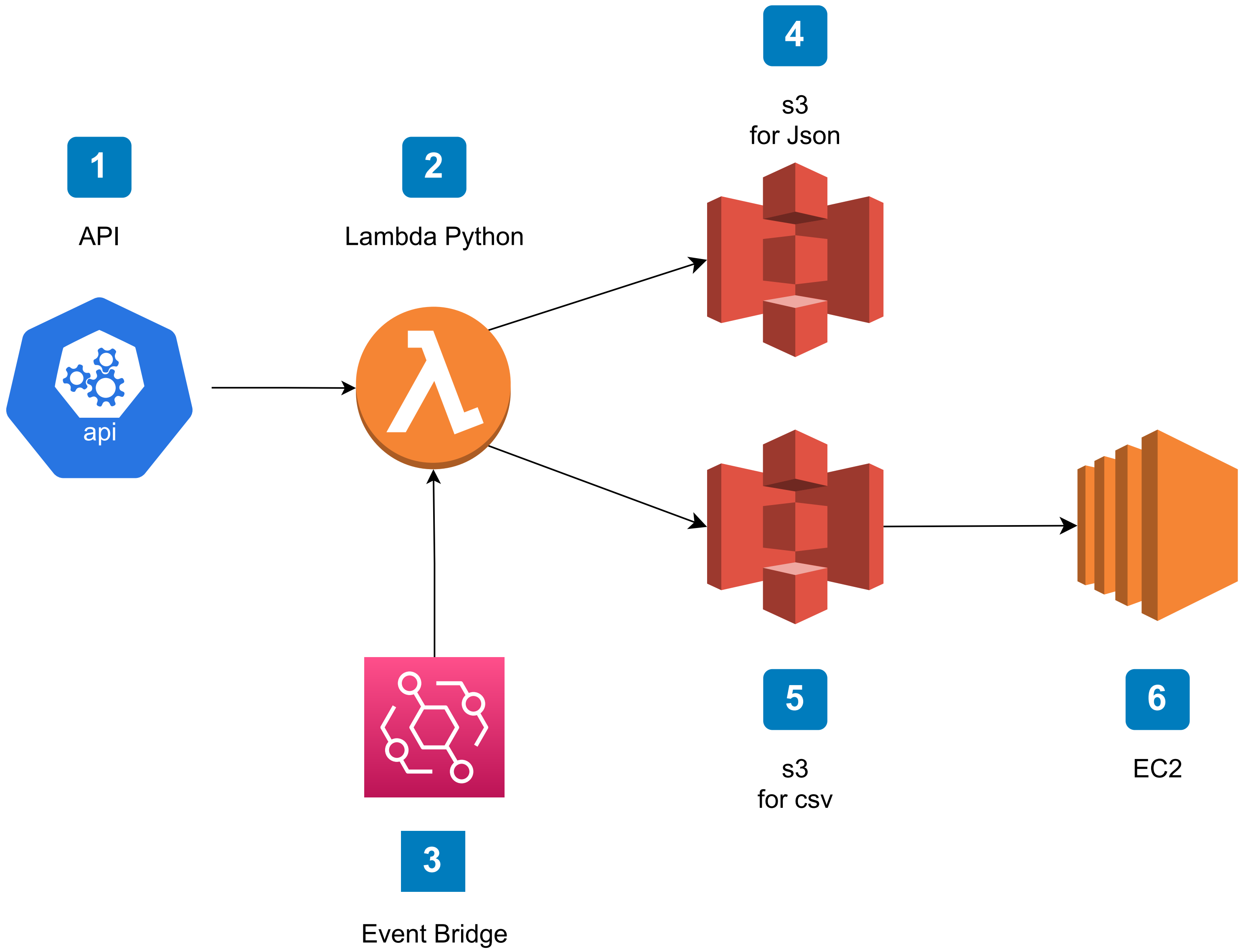
Data Processing with AWS

Yash Revannavar
s0586034



0 Introduction:

This project on AWS is for data processing. The main goal is to gather data from the News Api so that it can be stored and queried as needed.

1 API:

The source of the data is the News API, whereas I will fetch the data from the API to process it further.

2 Lambda python:

This lambda is made up of Python code that gathers data from the News API and stores it as a JSON file with data and a time stamp in the first S3 and with the option to store data in a CSV format in the second S3.

3 Event Bridge:

The data from the Python code will be sent to the S3 bucket after every 4 hours with the help of Event Bridge.

4 S3 for Json:

The data will be stored in the Json format as a backup option.

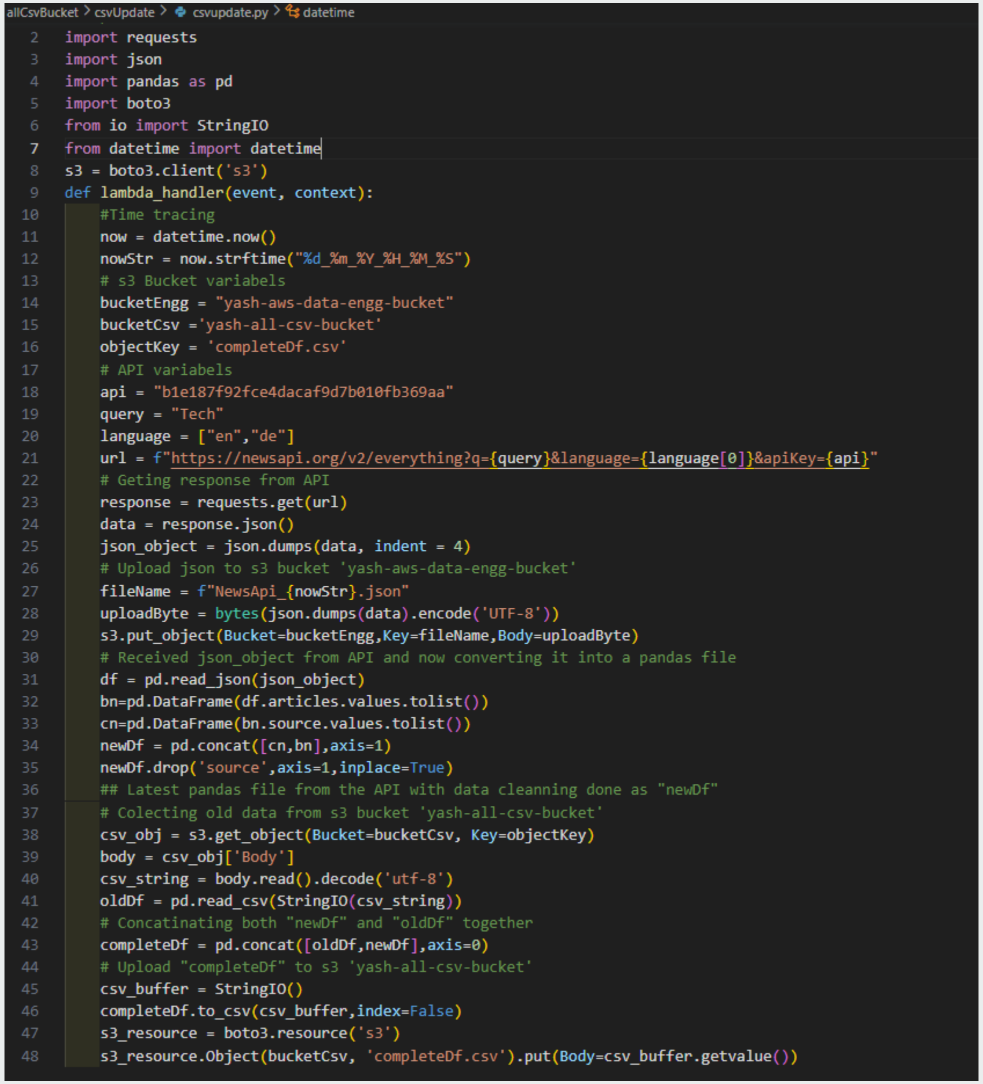
5 S3 for csv:

Here, the lambda function returns the data in a csv format so that it can be queried.

6 EC2:

On the home page for the Public IPv4 address, the data is requested according to the user.

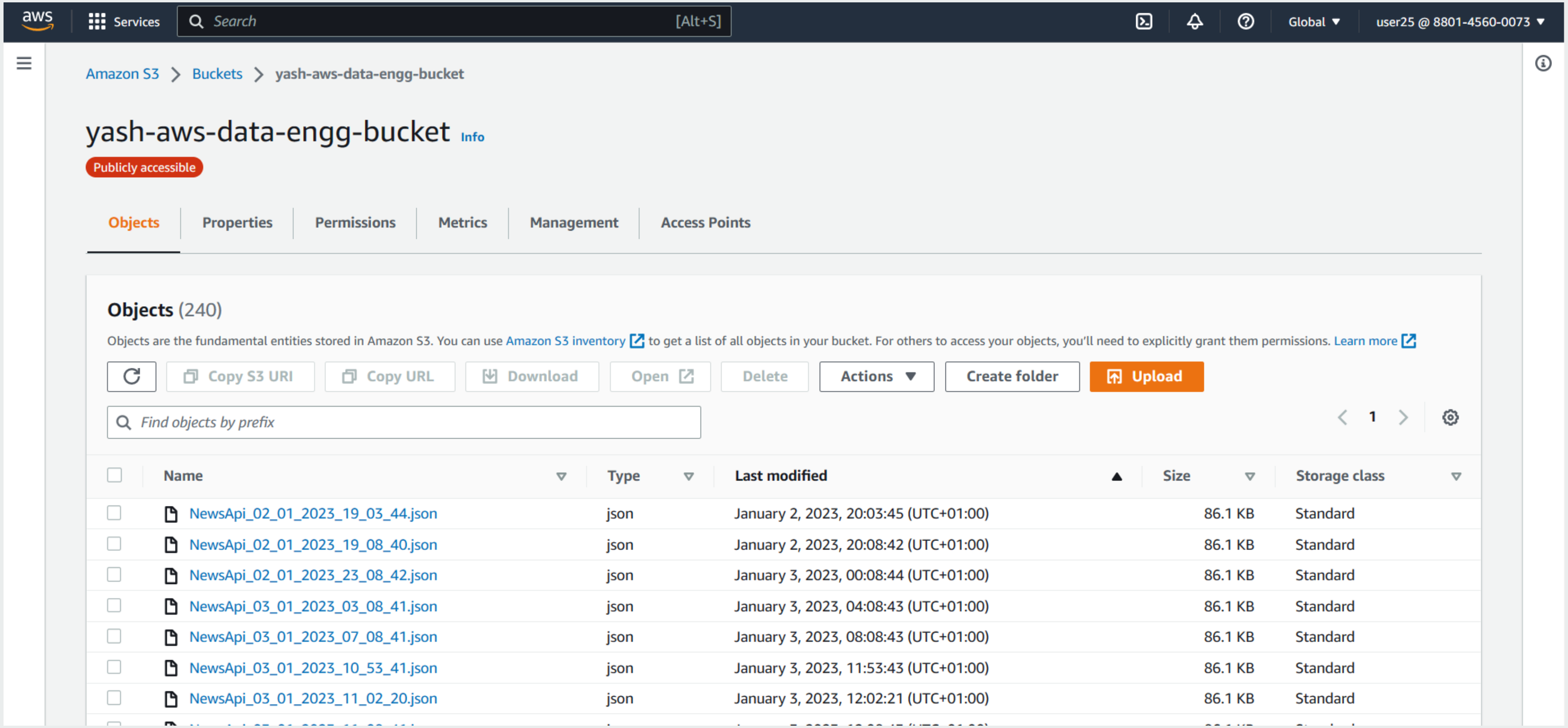
2 Lambda python:
This lambda is made up of Python code that gathers data from the News API and stores it as a JSON file with data and a time stamp in the first S3 and with the option to store data in a CSV format in the second S3.



```python
import requests
import json
import pandas as pd
import boto3
from io import StringIO
from datetime import datetime
s3 = boto3.client('s3')
def lambda_handler(event, context):
    #Time tracing
    now = datetime.now()
    nowStr = now.strftime("%d_%m_%Y_%H_%M_%S")
    # s3 Bucket variabels
    bucketEngg = "yash-aws-data-engg-bucket"
    bucketCsv ='yash-all-csv-bucket'
    objectKey = 'completeDf.csv'
    # API variabels
    api = "b1e187f92fce4dacaf9d7b010fb369aa"
    query = "Tech"
    language = ["en","de"]
    url = f"https://newsapi.org/v2/everything?q={query}&language={language[0]}&apiKey={api}"
    # Geting response from API
    response = requests.get(url)
    data = response.json()
    json_object = json.dumps(data, indent = 4)
    # Upload json to s3 bucket 'yash-aws-data-engg-bucket'
    fileName = f"NewsApi_{nowStr}.json"
    uploadByte = bytes(json.dumps(data).encode('UTF-8'))
    s3.put_object(Bucket=bucketEngg,Key=fileName,Body=uploadByte)
    # Received json_object from API and now converting it into a pandas file
    df = pd.read_json(json_object)
    bn=pd.DataFrame(df.articles.values.tolist())
    cn=pd.DataFrame(bn.source.values.tolist())
    newDf = pd.concat([cn,bn],axis=1)
    newDf.drop('source',axis=1,inplace=True)
    ## Latest pandas file from the API with data cleanning done as "newDf"
    # Colecting old data from s3 bucket 'yash-all-csv-bucket'
    csv_obj = s3.get_object(Bucket=bucketCsv, Key=objectKey)
    body = csv_obj['Body']
    csv_string = body.read().decode('utf-8')
    oldDf = pd.read_csv(StringIO(csv_string))
    # Concatinating both "newDf" and "oldDf" together
    completeDf = pd.concat([oldDf,newDf],axis=0)
    # Upload "completeDf" to s3 'yash-all-csv-bucket'
    csv_buffer = StringIO()
    completeDf.to_csv(csv_buffer,index=False)
    s3_resource = boto3.resource('s3')
    s3_resource.Object(bucketCsv, 'completeDf.csv').put(Body=csv_buffer.getvalue())
```

4

S3 for Json:
From the date 02-01-2023, data is recorded in this "yash-aws-data-engg-bucket" in json format for every 4 hours with the use of Event Bridge.



Amazon S3 > Buckets > yash-aws-data-engg-bucket

yash-aws-data-engg-bucket Info

Publicly accessible

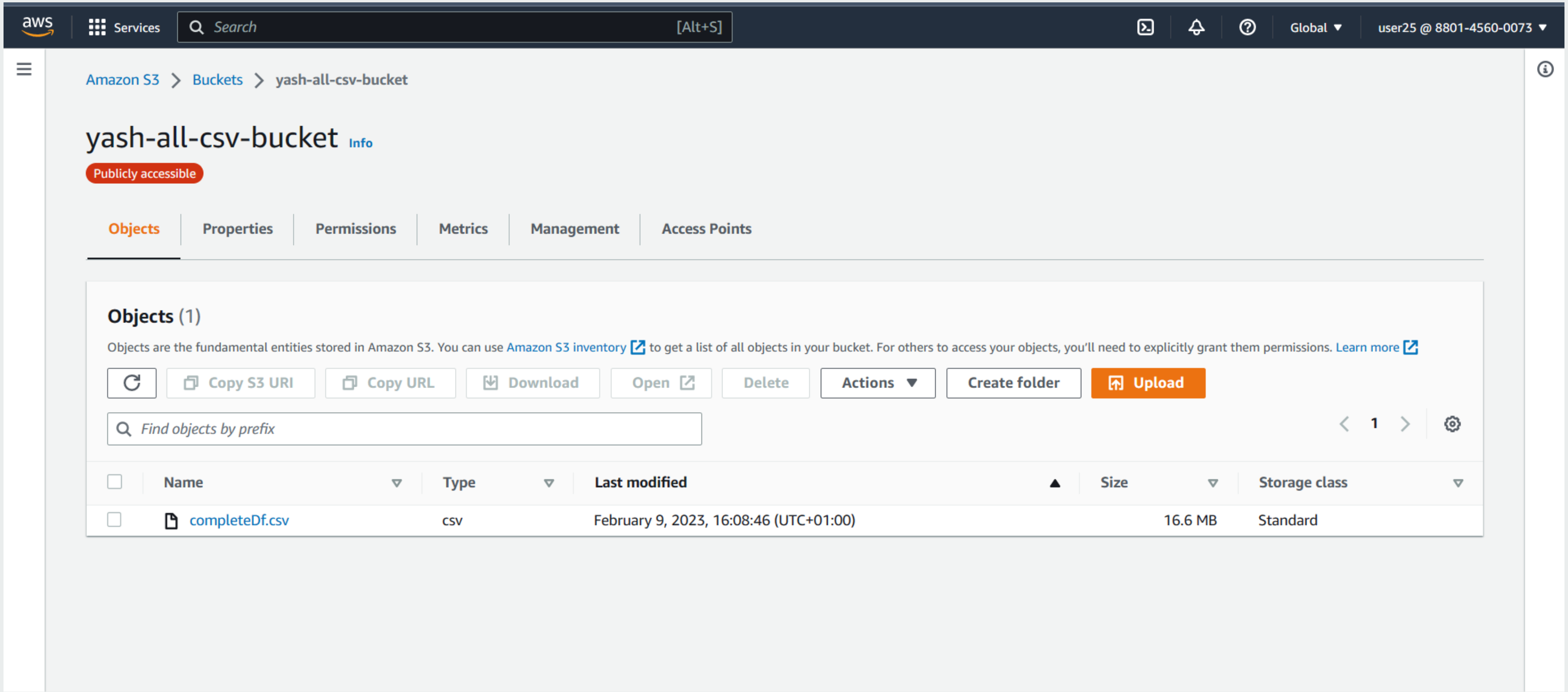
Objects (240)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	NewsApi_02_01_2023_19_03_44.json	json	January 2, 2023, 20:03:45 (UTC+01:00)	86.1 KB	Standard
□	NewsApi_02_01_2023_19_08_40.json	json	January 2, 2023, 20:08:42 (UTC+01:00)	86.1 KB	Standard
□	NewsApi_02_01_2023_23_08_42.json	json	January 3, 2023, 00:08:44 (UTC+01:00)	86.1 KB	Standard
□	NewsApi_03_01_2023_03_08_41.json	json	January 3, 2023, 04:08:43 (UTC+01:00)	86.1 KB	Standard
□	NewsApi_03_01_2023_07_08_41.json	json	January 3, 2023, 08:08:43 (UTC+01:00)	86.1 KB	Standard
□	NewsApi_03_01_2023_10_53_41.json	json	January 3, 2023, 11:53:43 (UTC+01:00)	86.1 KB	Standard
□	NewsApi_03_01_2023_11_02_20.json	json	January 3, 2023, 12:02:21 (UTC+01:00)	86.1 KB	Standard

5

S3 for csv:
In this "yash-all-csv-bucket" the data is stored in a single csv and updated for every 4 hours with the help of Event Bridge from the date 02-01-2023.



yash-all-csv-bucket

Objects (1)

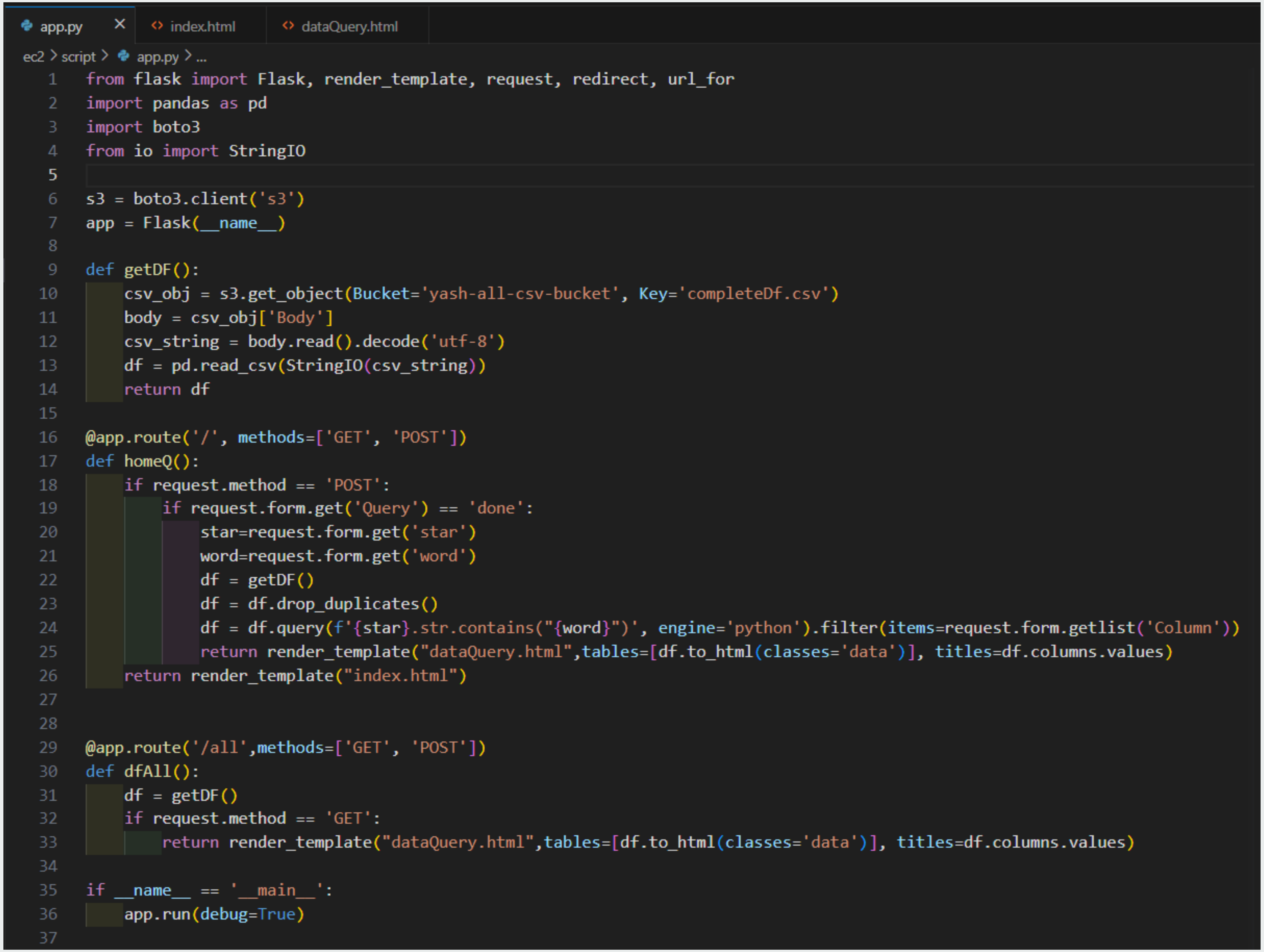
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified ▲	Size	Storage class
□	completeDf.csv	csv	February 9, 2023, 16:08:46 (UTC+01:00)	16.6 MB	Standard

6

EC2:

EC2 runs on an Ubuntu system using Python code that makes use of flask, pandas, and boto3. Pandas is used to manipulate the data, Flask is used to host the website, and boto3 is used to interface with AWS.

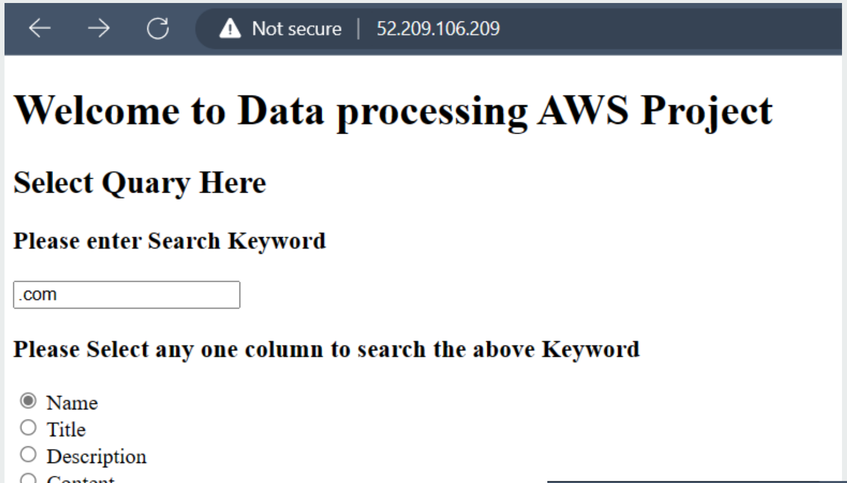


```python
from flask import Flask, render_template, request, redirect, url_for
import pandas as pd
import boto3
from io import StringIO

s3 = boto3.client('s3')
app = Flask(__name__)

def getDF():
    csv_obj = s3.get_object(Bucket='yash-all-csv-bucket', Key='completeDf.csv')
    body = csv_obj['Body']
    csv_string = body.read().decode('utf-8')
    df = pd.read_csv(StringIO(csv_string))
    return df

@app.route('/', methods=['GET', 'POST'])
def homeQ():
    if request.method == 'POST':
        if request.form.get('Query') == 'done':
            star=request.form.get('star')
            word=request.form.get('word')
            df = getDF()
            df = df.drop_duplicates()
            df = df.query(f'{star}.str.contains("{word}")', engine='python').filter(items=request.form.getlist('Column'))
            return render_template("dataQuery.html",tables=[df.to_html(classes='data')], titles=df.columns.values)
    return render_template("index.html")


@app.route('/all',methods=['GET', 'POST'])
def dfAll():
    df = getDF()
    if request.method == 'GET':
        return render_template("dataQuery.html",tables=[df.to_html(classes='data')], titles=df.columns.values)

if __name__ == '__main__':
    app.run(debug=True)

```

6

EC2:

The user chooses a keyword, a column name to search the keyword in, and the column that must be displayed in the result in the form in screenshot A. The outcome of the form filled out in screenshot B.



Welcome to Data processing AWS Project

Select Quary Here

Please enter Search Keyword

.com

Please Select any one column to search the above Keyword

● Name
○ Title
○ Description
○ Content

Choose the columns which you are intrested in:

☑ Name
□ Author
☑ Title
□ Description
☑ url
□ url To Image
☑ Published At
□ Content

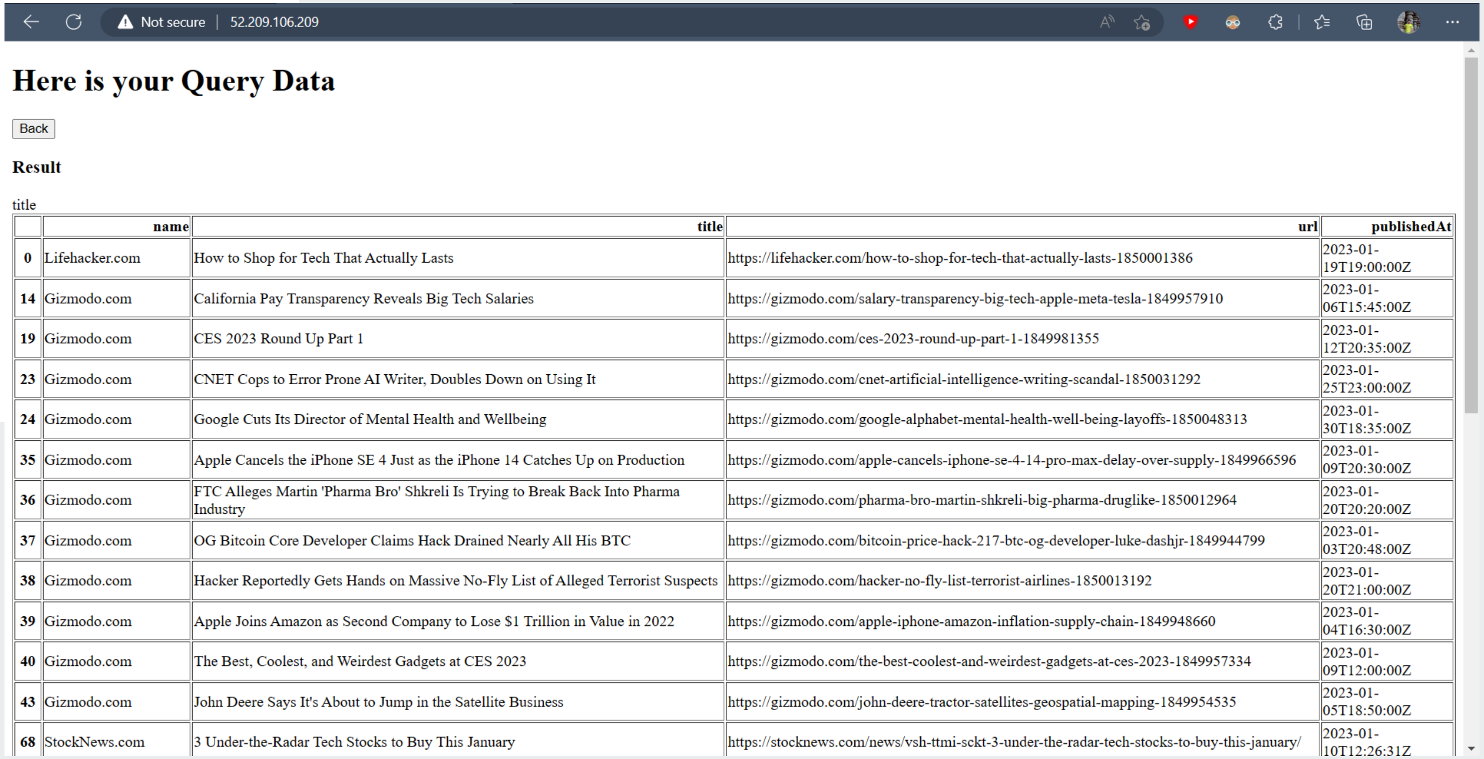
done

A

screenshot

B

screenshot



Here is your Query Data

Back

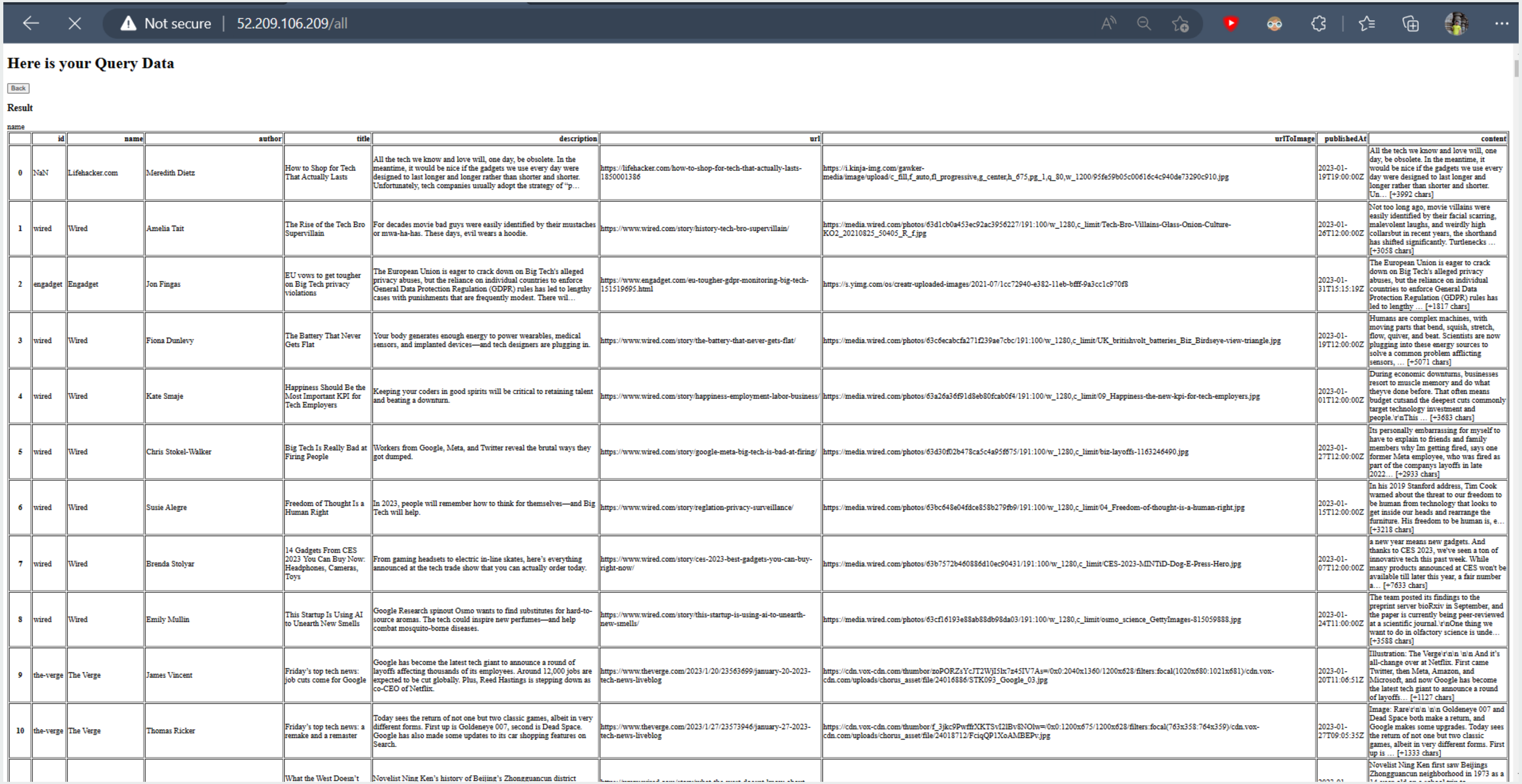
Result

title

	name	title	url	publishedAt
0	Lifehacker.com	How to Shop for Tech That Actually Lasts	https://lifehacker.com/how-to-shop-for-tech-that-actually-lasts-1850001386	2023-01-19T19:00:00Z
14	Gizmodo.com	California Pay Transparency Reveals Big Tech Salaries	https://gizmodo.com/salary-transparency-big-tech-apple-meta-tesla-1849957910	2023-01-06T15:45:00Z
19	Gizmodo.com	CES 2023 Round Up Part 1	https://gizmodo.com/ces-2023-round-up-part-1-1849981355	2023-01-12T20:35:00Z
23	Gizmodo.com	CNET Cops to Error Prone AI Writer, Doubles Down on Using It	https://gizmodo.com/cnet-artificial-intelligence-writing-scandal-1850031292	2023-01-25T23:00:00Z
24	Gizmodo.com	Google Cuts Its Director of Mental Health and Wellbeing	https://gizmodo.com/google-alphabet-mental-health-well-being-layoffs-1850048313	2023-01-30T18:35:00Z
35	Gizmodo.com	Apple Cancels the iPhone SE 4 Just as the iPhone 14 Catches Up on Production	https://gizmodo.com/apple-cancels-iphone-se-4-14-pro-max-delay-over-supply-1849966596	2023-01-09T20:30:00Z
36	Gizmodo.com	FTC Alleges Martin 'Pharma Bro' Shkreli Is Trying to Break Back Into Pharma Industry	https://gizmodo.com/pharma-bro-martin-shkreli-big-pharma-druglike-1850012964	2023-01-20T20:20:00Z
37	Gizmodo.com	OG Bitcoin Core Developer Claims Hack Drained Nearly All His BTC	https://gizmodo.com/bitcoin-price-hack-217-btc-og-developer-luke-dashjr-1849944799	2023-01-03T20:48:00Z
38	Gizmodo.com	Hacker Reportedly Gets Hands on Massive No-Fly List of Alleged Terrorist Suspects	https://gizmodo.com/hacker-no-fly-list-terrorist-airlines-1850013192	2023-01-20T21:00:00Z
39	Gizmodo.com	Apple Joins Amazon as Second Company to Lose $1 Trillion in Value in 2022	https://gizmodo.com/apple-iphone-amazon-inflation-supply-chain-1849948660	2023-01-04T16:30:00Z
40	Gizmodo.com	The Best, Coolest, and Weirdest Gadgets at CES 2023	https://gizmodo.com/the-best-coolest-and-weirdest-gadgets-at-ces-2023-1849957334	2023-01-09T12:00:00Z
43	Gizmodo.com	John Deere Says It's About to Jump in the Satellite Business	https://gizmodo.com/john-deere-tractor-satellites-geospatial-mapping-1849954535	2023-01-05T18:50:00Z
68	StockNews.com	3 Under-the-Radar Tech Stocks to Buy This January	https://stocknews.com/news/vsh-ttmi-sckt-3-under-the-radar-tech-stocks-to-buy-this-january/	2023-01-10T12:26:31Z

6

EC2:
All of the data gathered from the S3 bucket "yash-all-csv-bucket" is displayed when the Public IPv4 address + "/all" is entered.



Here is your Query Data

Result

name

	id	name	author	title	description	url	urlToImage	publishedAt	content
0	NaN	Lifehacker.com	Meredith Dietz	How to Shop for Tech That Actually Lasts	All the tech we know and love will, one day, be obsolete. In the meantime, it would be nice if the gadgets we use every day were designed to last longer and longer rather than shorter and shorter. Unfortunately, tech companies usually adopt the strategy of “p…	https://lifehacker.com/how-to-shop-for-tech-that-actually-lasts-1850001386	https://i.kinja-img.com/gawker-media/image/upload/c_fill,f_auto,fl_progressive,g_center,h_675,pg_1,q_80,w_1200/95fe59b05c00616c4c940de73290c910.jpg	2023-01-19T19:00:00Z	All the tech we know and love will, one day, be obsolete. In the meantime, it would be nice if the gadgets we use every day were designed to last longer and longer rather than shorter and shorter. Un… [+3992 chars]
1	wired	Wired	Amelia Tait	The Rise of the Tech Bro Supervillain	For decades movie bad guys were easily identified by their mustaches or mwa-ha-has. These days, evil wears a hoodie.	https://www.wired.com/story/history-tech-bro-supervillain/	https://media.wired.com/photos/63d1cb0a453ec92ac3956227/191:100/w_1280,c_limit/Tech-Bro-Villains-Glass-Onion-Culture-KO2_20210825_50405_R_f.jpg	2023-01-26T12:00:00Z	Not too long ago, movie villains were easily identified by their facial scarring, malevolent laughs, and weirdly high collarsbut in recent years, the shorthand has shifted significantly. Turtlenecks … [+3058 chars]
2	engadget	Engadget	Jon Fingas	EU vows to get tougher on Big Tech privacy violations	The European Union is eager to crack down on Big Tech's alleged privacy abuses, but the reliance on individual countries to enforce General Data Protection Regulation (GDPR) rules has led to lengthy cases with punishments that are frequently modest. There wil…	https://www.engadget.com/eu-tougher-gdpr-monitoring-big-tech-151519695.html	https://s.yimg.com/os/creatr-uploaded-images/2021-07/1cc72940-e382-11eb-bfff-9a3cc1c970f8	2023-01-31T15:15:19Z	The European Union is eager to crack down on Big Tech's alleged privacy abuses, but the reliance on individual countries to enforce General Data Protection Regulation (GDPR) rules has led to lengthy … [+1817 chars]
3	wired	Wired	Fiona Dunlevy	The Battery That Never Gets Flat	Your body generates enough energy to power wearables, medical sensors, and implanted devices—and tech designers are plugging in.	https://www.wired.com/story/the-battery-that-never-gets-flat/	https://media.wired.com/photos/63c6ecabcfa271f239ae7cbc/191:100/w_1280,c_limit/UK_britishvolt_batteries_Biz_Birdseye-view-triangle.jpg	2023-01-19T12:00:00Z	Humans are complex machines, with moving parts that bend, squish, stretch, flow, quiver, and beat. Scientists are now plugging into these energy sources to solve a common problem afflicting sensors, … [+5071 chars]
4	wired	Wired	Kate Smaje	Happiness Should Be the Most Important KPI for Tech Employers	Keeping your coders in good spirits will be critical to retaining talent and beating a downturn.	https://www.wired.com/story/happiness-employment-labor-business/	https://media.wired.com/photos/63a26a36f91d8eb80fcab0f4/191:100/w_1280,c_limit/09_Happiness-the-new-kpi-for-tech-employers.jpg	2023-01-01T12:00:00Z	During economic downturns, businesses resort to muscle memory and do what theyve done before. That often means budget cutsand the deepest cuts commonly target technology investment and people.\r\nThis … [+3683 chars]
5	wired	Wired	Chris Stokel-Walker	Big Tech Is Really Bad at Firing People	Workers from Google, Meta, and Twitter reveal the brutal ways they got dumped.	https://www.wired.com/story/google-meta-big-tech-is-bad-at-firing/	https://media.wired.com/photos/63d30f02b478ca5c4a95f675/191:100/w_1280,c_limit/biz-layoffs-1163246490.jpg	2023-01-27T12:00:00Z	Its personally embarrassing for myself to have to explain to friends and family members why Im getting fired, says one former Meta employee, who was fired as part of the companys layoffs in late 2022… [+2933 chars]
6	wired	Wired	Susie Alegre	Freedom of Thought Is a Human Right	In 2023, people will remember how to think for themselves—and Big Tech will help.	https://www.wired.com/story/reglation-privacy-surveillance/	https://media.wired.com/photos/63bc648e04fdce858b279fb9/191:100/w_1280,c_limit/04_Freedom-of-thought-is-a-human-right.jpg	2023-01-15T12:00:00Z	In his 2019 Stanford address, Tim Cook warned about the threat to our freedom to be human from technology that looks to get inside our heads and rearrange the furniture. His freedom to be human is, e… [+3218 chars]
7	wired	Wired	Brenda Stolyar	14 Gadgets From CES 2023 You Can Buy Now: Headphones, Cameras, Toys	From gaming headsets to electric in-line skates, here’s everything announced at the tech trade show that you can actually order today.	https://www.wired.com/story/ces-2023-best-gadgets-you-can-buy-right-now/	https://media.wired.com/photos/63b7572b460886d10ec90431/191:100/w_1280,c_limit/CES-2023-MINTiD-Dog-E-Press-Hero.jpg	2023-01-07T12:00:00Z	a new year means new gadgets. And thanks to CES 2023, we've seen a ton of innovative tech this past week. While many products announced at CES won't be available till later this year, a fair number a… [+7633 chars]
8	wired	Wired	Emily Mullin	This Startup Is Using AI to Unearth New Smells	Google Research spinout Osmo wants to find substitutes for hard-to-source aromas. The tech could inspire new perfumes—and help combat mosquito-borne diseases.	https://www.wired.com/story/this-startup-is-using-ai-to-unearth-new-smells/	https://media.wired.com/photos/63cf16193e88ab88db98da03/191:100/w_1280,c_limit/osmo_science_GettyImages-815059888.jpg	2023-01-24T11:00:00Z	The team posted its findings to the preprint server bioRxiv in September, and the paper is currently being peer-reviewed at a scientific journal.\r\nOne thing we want to do in olfactory science is unde… [+3588 chars]
9	the-verge	The Verge	James Vincent	Friday’s top tech news: job cuts come for Google	Google has become the latest tech giant to announce a round of layoffs affecting thousands of its employees. Around 12,000 jobs are expected to be cut globally. Plus, Reed Hastings is stepping down as co-CEO of Netflix.	https://www.theverge.com/2023/1/20/23563699/january-20-2023-tech-news-liveblog	https://cdn.vox-cdn.com/thumbor/zoPORZsYcJT2WjI5lx7z45IV7As=/0x0:2040x1360/1200x628/filters:focal(1020x680:1021x681)/cdn.vox-cdn.com/uploads/chorus_asset/file/24016886/STK093_Google_03.jpg	2023-01-20T11:06:51Z	Illustration: The Verge\r\n\n \n\n And it's all-change over at Netflix. First came Twitter, then Meta, Amazon, and Microsoft, and now Google has become the latest tech giant to announce a round of layoffs… [+1127 chars]
10	the-verge	The Verge	Thomas Ricker	Friday’s top tech news: a remake and a remaster	Today sees the return of not one but two classic games, albeit in very different forms. First up is Goldeneye 007, second is Dead Space. Google has also made some updates to its car shopping features on Search.	https://www.theverge.com/2023/1/27/23573946/january-27-2023-tech-news-liveblog	https://cdn.vox-cdn.com/thumbor/f_3jkc9PwffrXKTSvI2lBv8NOlw=/0x0:1200x675/1200x628/filters:focal(763x358:764x359)/cdn.vox-cdn.com/uploads/chorus_asset/file/24018712/FciqQP1XoAMBEPv.jpg	2023-01-27T09:05:35Z	Image: Rare\r\n\n \n\n Goldeneye 007 and Dead Space both make a return, and Google makes some upgrades. Today sees the return of not one but two classic games, albeit in very different forms. First up is … [+1333 chars]
				What the West Doesn’t	Novelist Ning Ken’s history of Beijing’s Zhongguancun district				Novelist Ning Ken first saw Beijings Zhongguancun neighborhood in 1973 as a