Paul Gillet & Kevin Auberson
Groupe : L03GrA
Date : 21.03.2024

TASK 1: USE THE WEB CONSOLE TO CREATE AN S3 BUCKET AND UPLOAD AND DOWNLOAD OBJECTS (FILES)

Buckets name:
Paul Gillet : ist-gra-gillet-bucket
Kevin Auberson: ist-gra-auberson-bucket

TASK 2: USE THE AWS COMMAND-LINE INTERFACE TO MANAGE BUCKETS AND OBJECTS

> 3.1.List all available regions:

```
>aws ec2 describe-regions
{
    "Regions": [
        {
            "Endpoint": "ec2.ap-south-1.amazonaws.com",
            "RegionName": "ap-south-1",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.eu-north-1.amazonaws.com",
            "RegionName": "eu-north-1",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.eu-west-3.amazonaws.com",
            "RegionName": "eu-west-3",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.eu-west-2.amazonaws.com",
            "RegionName": "eu-west-2",
```

```
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.eu-west-1.amazonaws.com",
            "RegionName": "eu-west-1",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.ap-northeast-3.amazonaws.com",
            "RegionName": "ap-northeast-3",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.ap-northeast-2.amazonaws.com",
            "RegionName": "ap-northeast-2",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.ap-northeast-1.amazonaws.com",
            "RegionName": "ap-northeast-1",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.ca-central-1.amazonaws.com",
            "RegionName": "ca-central-1",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.sa-east-1.amazonaws.com",
            "RegionName": "sa-east-1",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.ap-southeast-1.amazonaws.com",
            "RegionName": "ap-southeast-1",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.ap-southeast-2.amazonaws.com",
            "RegionName": "ap-southeast-2",
```

```
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.eu-central-1.amazonaws.com",
            "RegionName": "eu-central-1",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.us-east-1.amazonaws.com",
            "RegionName": "us-east-1",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.us-east-2.amazonaws.com",
            "RegionName": "us-east-2",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.us-west-1.amazonaws.com",
            "RegionName": "us-west-1",
            "OptInStatus": "opt-in-not-required"
        },
        {
            "Endpoint": "ec2.us-west-2.amazonaws.com",
            "RegionName": "us-west-2",
            "OptInStatus": "opt-in-not-required"
        }
    ]
}
```

3.2.Display account attributes:

```
> aws ec2 describe-account-attributes
{
    "AccountAttributes": [
        {
            "AttributeName": "supported-platforms",
            "AttributeValues": [
                {
                    "AttributeValue": "VPC"
```

```
                }
            ]
        },
        {
            "AttributeName": "vpc-max-security-groups-per-interface",
            "AttributeValues": [
                {
                    "AttributeValue": "5"
                }
            ]
        },
        {
            "AttributeName": "max-elastic-ips",
            "AttributeValues": [
                {
                    "AttributeValue": "5"
                }
            ]
        },
        {
            "AttributeName": "max-instances",
            "AttributeValues": [
                {
                    "AttributeValue": "20"
                }
            ]
        },
        {
            "AttributeName": "vpc-max-elastic-ips",
            "AttributeValues": [
                {
                    "AttributeValue": "20"
                }
            ]
        },
        {
            "AttributeName": "default-vpc",
            "AttributeValues": [
                {
                    "AttributeValue": "vpc-049e2f8e56e0bafef"
```

```
                }
            ]
        }
    ]
}
```

3.3.Display available EC2 Instance types:

```
> aws ec2 describe-instance-type-offerings
...
        {
            "InstanceType": "r7iz.16xlarge",
            "LocationType": "region",
            "Location": "us-east-1"
        },
        {
            "InstanceType": "m6in.xlarge",
            "LocationType": "region",
            "Location": "us-east-1"
        },
        {
            "InstanceType": "c6a.4xlarge",
            "LocationType": "region",
            "Location": "us-east-1"
        },
        {
            "InstanceType": "r6a.xlarge",
            "LocationType": "region",
            "Location": "us-east-1"
        },
        {
            "InstanceType": "dl1.24xlarge",
            "LocationType": "region",
            "Location": "us-east-1"
        }
    ]
}
```

3.4.List all S3 buckets in the account:

```
> aws s3 ls
2024-02-16 15:53:19 acmedata-rms
2024-03-01 12:53:35 athena-queries-ist-rms
2024-03-17 15:44:45 bucket1-721449c17af4fc6b2ee37abc80049ae0
2024-03-17 15:44:43 bucket2-a14876721e3c566424f83072bcd05d94
2024-03-17 15:44:43 bucket3-772ade4372b7a3338bf2c99ca077cf00
2024-03-21 16:50:51 grk-charbonnier
2024-03-21 17:04:19 grk-charbonnier-2
2024-03-21 17:16:46 grk-charbonnier-website
2024-03-01 15:19:08 heigvd-ist-prepared
2024-03-21 16:56:53 ist-gra-auberson-bucket
2024-03-21 16:52:36 ist-gra-gillet-bucket
2024-03-21 16:59:21 ist-grb-butty-bucket
2024-03-21 16:55:22 ist-grb-muhlemann
2024-03-21 16:57:04 ist-grc-bonzon-bucket
2024-03-21 17:27:14 ist-grc-bonzon-bucketcli
2024-03-21 16:56:34 ist-grd-laurenti
2024-03-21 16:57:37 ist-grd-panchaud-bucket
2024-03-21 16:52:22 ist-gre-giuliano
2024-03-21 17:05:47 ist-gre-giuliano-1
2024-03-21 16:59:18 ist-gre-schaller
2024-03-21 17:07:07 ist-gre-schaller-blabla
2024-03-21 16:52:11 ist-grg-ansermoz-1
2024-03-21 16:58:12 ist-grg-ansermoz-2
2024-03-21 17:15:26 ist-grh-bloch-bucket
2024-03-21 17:04:47 ist-grh-strefeler-bucket
2024-03-21 17:02:10 ist-gri-piccin-bucket1
2024-03-21 17:34:16 ist-gri-piccin-bucket2
2024-03-21 17:06:26 ist-grj-billeter-bucket
2024-03-21 17:05:34 ist-grj-ronquillo-bucket
2024-03-21 17:27:41 ist-l03grl-ferchichi
2024-02-15 16:35:52 ist-rms-bucket
2024-02-16 16:31:03 ist-rms-meteo
```

4.1.Create a new bucket.

```
> aws s3 mb s3://ist-gra-gillet-bucket-2
make_bucket: ist-gra-gillet-bucket-2
```

> 4.2.Upload an object

```
> aws s3 cp .\results.csv s3://ist-gra-gillet-bucket-2
upload: .\results.csv to s3://ist-gra-gillet-bucket-2/results.csv
```

> 4.3.List the objects in the bucket

```
> aws s3 ls s3://ist-gra-auberson-bucket2
2024-03-21 17:40:30        342 lab.csv
```

> 4.4.Copy the object

```
aws s3 cp s3://ist-gra-auberson-bucket/lab.csv s3://ist-gra-auberson-bucket2
copy: s3://ist-gra-auberson-bucket/lab.csv to s3://ist-gra-auberson-
bucket2/lab.csv
```

> 4.5.Delete the copied object

```
aws s3 rm s3://ist-gra-auberson-bucket2/lab.csv
delete: s3://ist-gra-auberson-bucket2/lab.csv
```

> 5.2.Make a copy of the object and move the copy into the folder.

```
aws s3 cp s3://ist-gra-auberson-bucket/lab.csv s3://ist-gra-auberson-
bucket/ist-gra-auberson-folder/
copy: s3://ist-gra-auberson-bucket/lab.csv to s3://ist-gra-auberson-
bucket/ist-gra-auberson-folder/lab.csv
```

> 5.3.What happens if you move an object to a folder that does not exist?

It creates the folder then copies the object in it

TASK 3

> 2. On which URL is your new website reachable?

Gillet: http://ist-gra-gillet-bucket.s3-website-us-east-1.amazonaws.com/
Auberson: http://ist-gra-auberson-bucket.s3-website-us-east-1.amazonaws.com/

TASK 4

> When was the latest crawl?

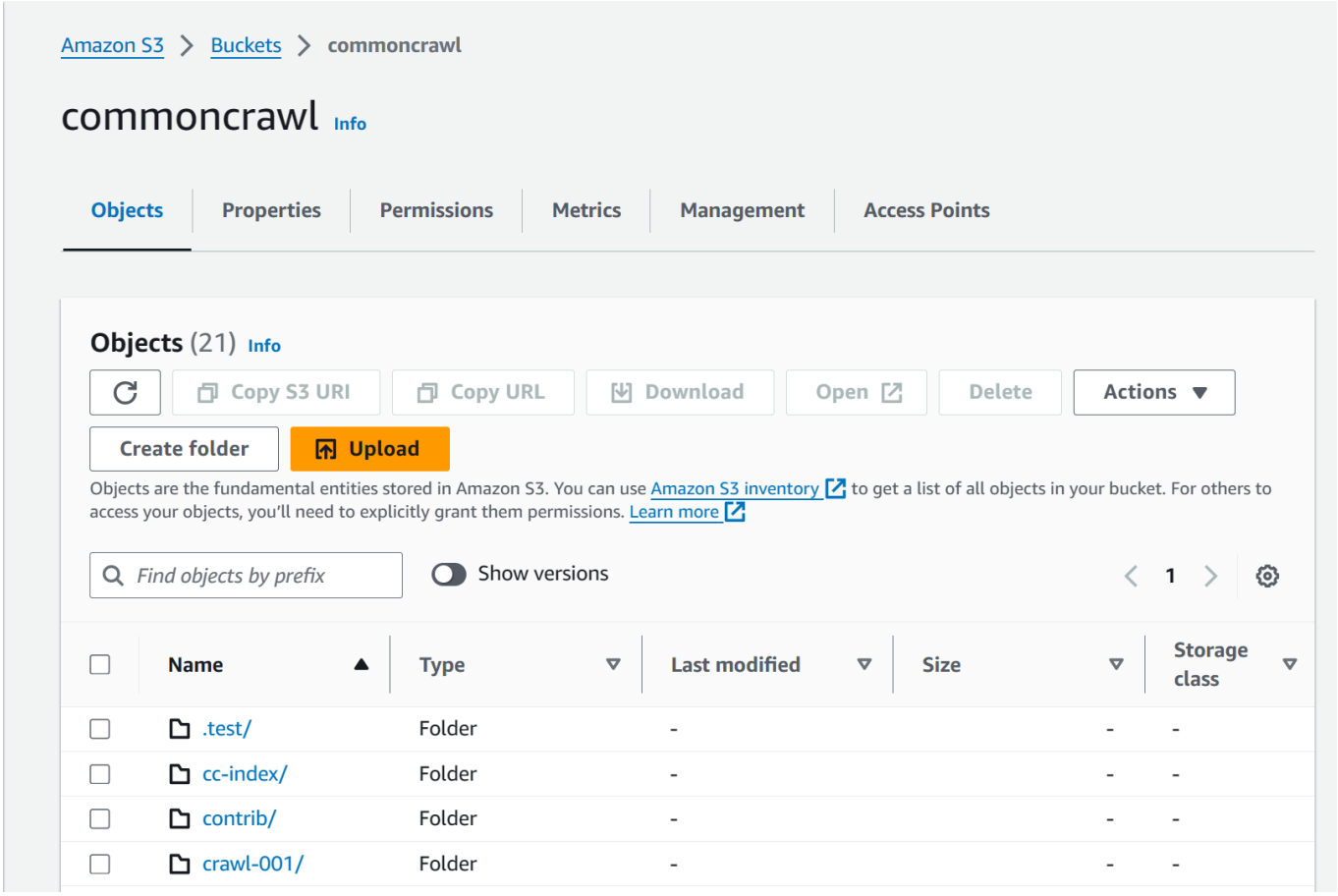
Feburary/March 2024

> What is the bucket name?
>
> commoncrawl

> Under which prefix is the latest crawl stored?

CC-MAIN-2024-10

> Log into the AWS S3 Management Console. Replace the browser URL with commoncrawl



> Navigate to the root folder of the latest crawl. Click on the object index.html. Click the Open button to load it into your browser. What is the URL of this object?

https://commoncrawl.s3.us-east-1.amazonaws.com/crawl-data/CC-MAIN-2024-10/index.html?response-content-disposition=inline&X-Amz-Security-Token=IQoJb3JpZ2luX2VjEIr%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCWV1LXdlc3QtMyJGMEQCIGcVV0AtmTKjaPwAXPhxuZOu2JN3eAUTV6R3VNH8aYECAiAIU5NJKG%2BMrEIUKt7sd%2B3pvfGLRs68QxLcjORQgVunIiqGAwij%2F%2F%2F%2F%2F%2F%2F%2F%2F%2F8BEAAaDDg1MTcyNTU4MTg1MSIMC9vdJc%2BtNf0thS9kKtoCIUfKlt3S%2BkNZssJCTGhi%2FnESYZqsww%2Bsc1wpkJT4TiLOg7UBXARquJgsBI1YnIgQaaGlX3jL%2F4gU6PnqmGqSu

tYIQNeXLHesarM%2F3RoGzMRUI%2BcAqOGJY95YI9IMCO1jlWAp52MIxQgsj8x97DJ19X4W
AkJbcdb1ZxjCPfx0CI%2FkHfRFLI71vhCkDc5fzRK%2F2Qgyg3Y0M9Z4I7inVQ39UsZyinNz0Qy
6i5eIjtIr%2FadoUIlXqYhoBJttuPMhGS20XER6skuLJdzTOFOGG45I%2BPHFnKtFS0qwAtZhQe
t8BfV%2FM%2BR7FJBSQiDDGeOZxwcLIAFdV19GyVK0br3%2BQq9NLHOEFQe8gJm%2F1Ji
8OQjcWqaWAfZwsCQgAeCgSPGz7bJE7eKjpCcHDLcS0tY2VvVNabrMN3TyjDiaL4mNN7FT67
%2BQOWODRSfYiuOOi6e5FZ4qNTxKzAUEBMHWTzDCp4qwBjq0AqE%2BYGVWLMviWcyZ
PQDDn7pgXnZjQpIWZhb6Eq%2B%2Boh71D7kEoBWnM8XDPPe2jhL2Kin%2Fr9ZjT0qfgwlsG
FX%2FjGHrOiKBPyegavF0vCOhwoiEewuLDusOHUAI64ZYP%2FY0juAtO9F740N5%2BCujn9
5McB%2FejcRJcRQaKEBadFHEg8Dhsugh8RCVnbNBy4A2A5VBuQjkS3ODVr%2FrZSAuLv0S
i7LbnojtibSNKo5Tnz6eaVtMsW0V4iE%2FL565TnjtL389I6z69ENDc%2BMM9dloeJ4D8s4HJrFo
C2HIvHpFfmKONfvzzuJDjDJTgzRXX2xKZAsI5jJTRe5wlVG1oIV8awKRV8NYg%2Fo0r5XEq3k
MDXErSgIUlXJd%2BpJ4%2F7otHj6wvFebSL1J5IhmcYa%2Ff2K%2FwYygdlpCsCjF&X-Amz-
Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20240326T095613Z&X-Amz-
SignedHeaders=host&X-Amz-Expires=300&X-Amz-
Credential=ASIA4MTWM7YNSPF7CPNF%2F20240326%2Fus-east-
1%2Fs3%2Faws4_request&X-Amz-
Signature=3de72deabe801ea2105562f79579b834eaf7acf15ac7838225349aae5909a39d

> What are WARC, WAT and WET files (look at the Get Started guide)?

- WARC : files which contains the raw crawl data
- WAT : files which contains the metadata about the records stored in the WARC format.
- WET : files which contains the plaintext about the records stored in the WARC format.

> What is the typical size of a WARC file (ballpark)?

The WARC standard recommends 1 GB as target size

> Why is it not sufficient to just store the WARC, WAT and WET files in the bucket? What other type of file is needed?

Index files are also stored to make search and accessibility easier and imporve the overall usability of the dataset

> What storage classes have the Common Crawl developers chosen to store the data?

The Common Crawl developers have chosen the Amazon S3 Standard (S3 Standard) storage classe.

TASK 5

> You are a data engineer at Coop responsible for a data product. The product is the global Coop sales data, which is shared with several Coop departments. Your task is to add each

> week the new sales data of the week to an S3 bucket.
> How would you do this task? Describe your thought process.

At first, we need to extract the new raw sales data from the databases. We can then verify the data format and if needed transform to unify it. Once the data is prepared we can load it on a S3 bucket on Amazon Web Service.

As this task is kind of redunant, we should have scripts do that and simply have reports of the execution and status of the save state.