Objektové úložiště S3 úvod

Jiří Sitera

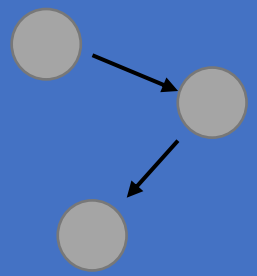
EurOpen.CZ, 15.5.2023

Proč se tím zabývat?

Remote storage protocol

backup

Datové workflow



Backend

S3

Kam S3 patří

- `s5cmd cp s3://bucket/inputfile .`
- `S5cmd cp outputfile s3://bucket/`
- Něco jako FTP, přes HTTP, REST API
- Standard z cloudového světa (interoperabilita, podpora v SW)
- Primární zaměření: dynamický, rychlý, efektivní přístup k datům
- Podpora ACL

Amazon	Protokol/ ekvivalent	Vlastnosti
S3 (simple storage)		Nejlevnější, největší ochrana proti výpadku, globálně dostupný, výkon pro lineární čtení/zápis
EBS (elastic block storage)	RBD, i-scsi	
EFS (elastic filesystem)	NFS v4.1	POSIX FS, asi 3x dražší, IOPS, určité omezení na počet souborů a velikost
FSx Lustre	LustreFS	Pro HPC
EMR (elastic Map-Reduce)	Hadoop	

Kde ho můžu získat?

AWS
$$
SaaS

https://aws.amazon.com/s3/

DU CESNET
akademická služba
budoucí repozitáře

https://du.cesnet.cz/cs/navody/object_storage/

Vlastní server

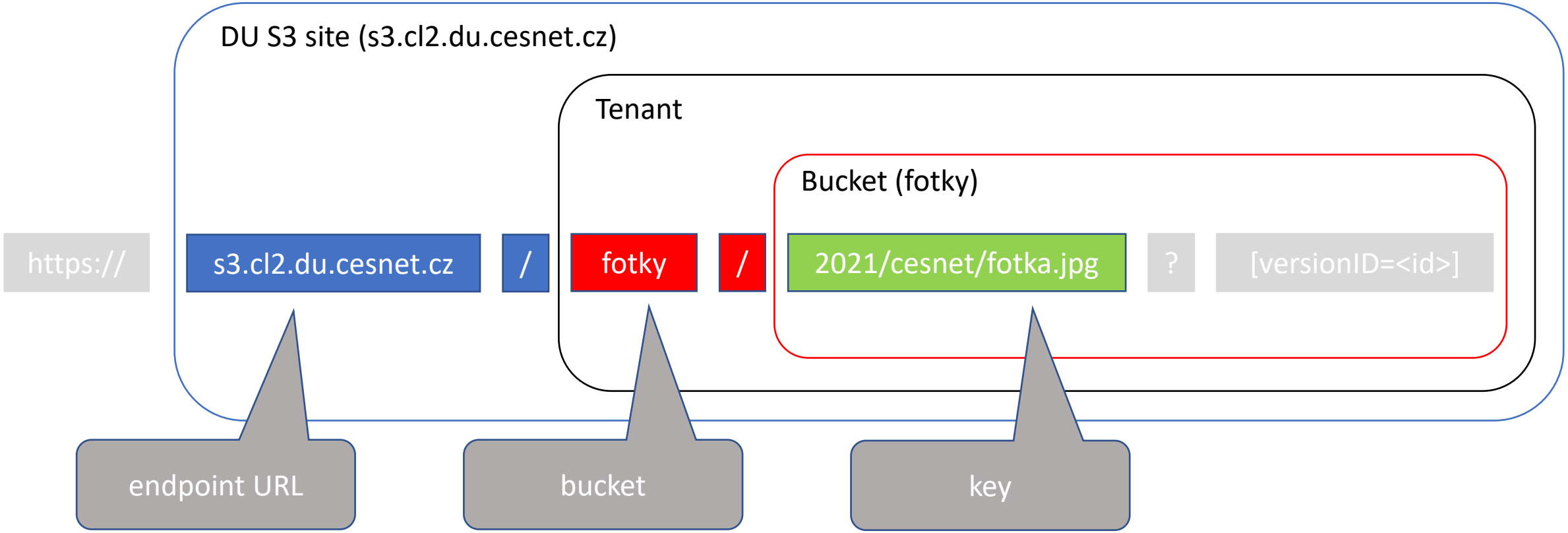
MinIO

https://min.io/

Objekty v S3

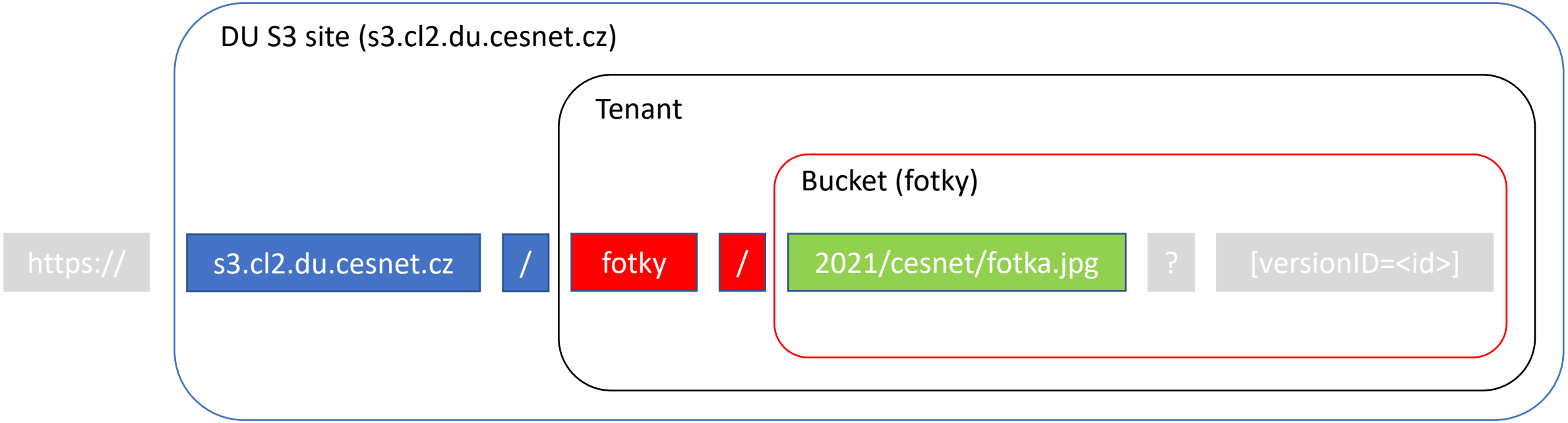
- Objekty jsou organizovány v entitách zvaných bucket
  - Bucket má unikátní (viz dále) jméno
  - Některé vlastnosti (například zapnutí/vypnutí versování) se přidělují celému bucketu
  - Bucket si vyrábí a ruší uživatel
  - Bucket zvládne velké množství objektů (1M), bucketů si může uživatel udělat 1000
- Objekt má jméno označované jako klíč
  - Jméno obecně nemá interpretaci, obsah bucketu nemá hierarchii
  - Používá se konvence s lomítkem, tj. typu `2021/cesnet/fotka.jpg` se interpretuje jako hierarchická struktura (adresáře)
  - Může a nemusí se používat finta s prázdným objektem reprezentujícím adresář

Identifikace objektu



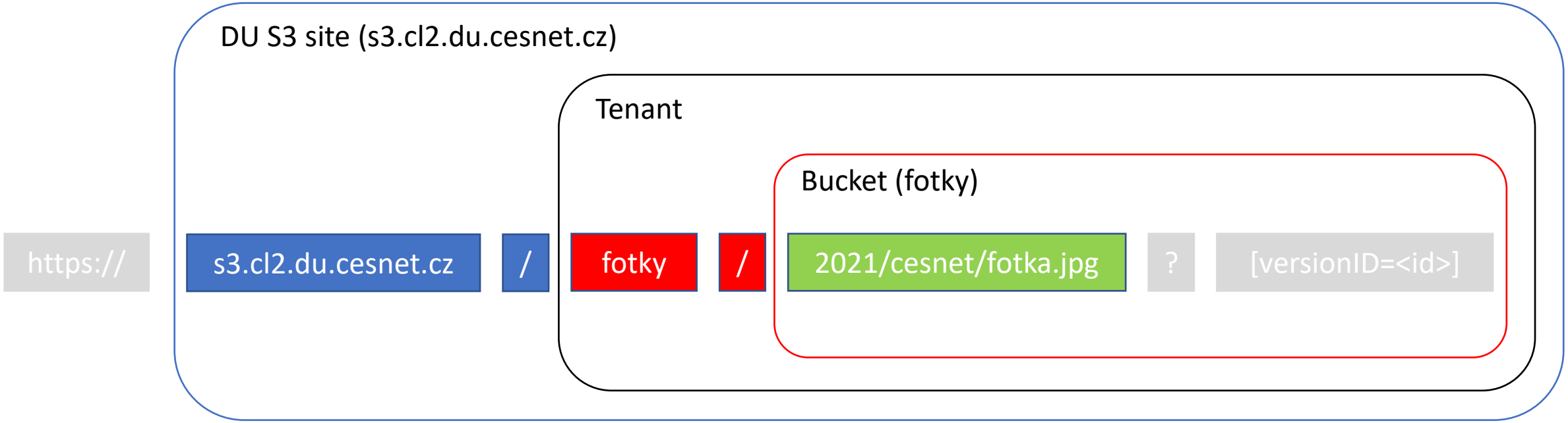
- Název bucketu nesmí mít velká písmenka a dovoleno je jen málo speciálních znaků (dovolena je pomlčka)
- Klíč může obsahovat skoro všechno (UTF-8), limit 1024 znaků
- Verzování je defaultně vypnuto, nastavit spolu s politikou retence

Multi tenancy



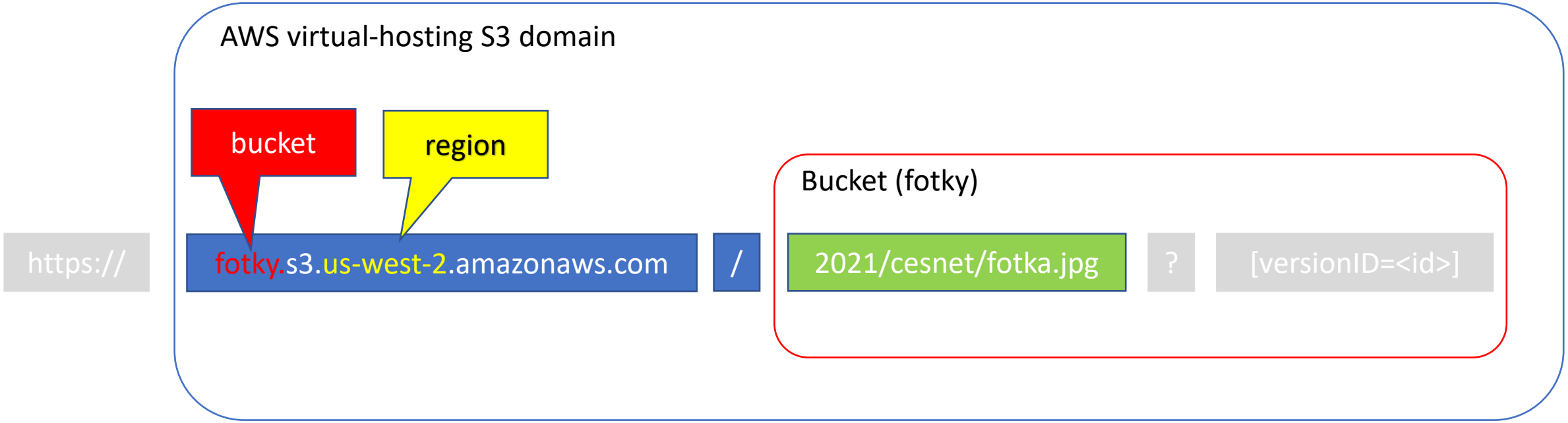
- Tenant poskytuje oddělený prostor pro pojmenování bucketů (jinak by byly globálně unikátní napříč uživateli)
  - U autentizovaného přístupu se tenant odvodí od uživatelského access key, v public URL musí být tenant:bucket, např. `sitera:fotky (reálně 0ba6811479730b2e4c22aa672d4706bacee9ded6:fotky)`
  - Řízení přístupu (ACL) funguje pouze uvnitř tenantu (návrh na změnu)

Multi site



- AWS má globální distribuční model (+ něco jako CDN) - regiony
- Standalone implementace (CEPH, MINIO) také nějak řeší
- Poznámka: CESNET DU (implementace CEPH) lze požádat o dvě site se synchronizací dat na pozadí
- Toto je path-style URL

Virtual-hosting style URL



- Bucket = DNS name
- AWS podporuje virtual-hosting style URL, path-style URL postupně vytlačuje

Model konzistence dat

- Eventual consistency model
  - Zajištěno není nic, propagace změn běží paralelně k dalším operacím
  - Např. smazání bucketu se v následujícím výpisu seznamu bucketů nemusí projevit, změna konfigurace bucketu/objektu se nemusí projevit hned
- Strong read-after-write consitency
  - Po úspěšném ukončení operace následující operace změny respektují.
  - Například pro PUT je garantováno, že po skončení přepisu objektu všechny následující GET dostanou nová data.
  - Platí pro vytvoření a smazání objektu a následnou operaci LIST.
- Žádné zámky
  - V případě dvou souběžných PUT vyhraje jeden podle timestampu.
  - Update objektu je atomický – GET vrací starou nebo novou verzi.

Metadata a přístup k objektům

- Atributy = pevně přirazené objektu
  - Některé mají systémový význam, třeba klíč
  - Klíč nejde změnit, neexistuje atomické přejmenování objektu
- Tagy = lze měnit, dokonce atomicky
  - Nelze podle nich vybírat/vyhledávat objekty
- Vyhledávat objekty lze pouze podle klíče/jména
  - Správně je mít někde vedle SW/databázi
  - Je podporováno využít struktury jména, např. hledat *leden*

Náhled na CLI

AAI S3 obecně

```
"user": "test_metacentrum$user1",
"access_key": "6087WG4NVHHGUNOSC43F",
"secret_key": "BQ1Q9UQusLAcSC0LhMPDeWzdiRje2LU5bvOoK9DT"
```

- Proměnná či konfigurační soubor obsahují access a secret key, jedno je vlastně jméno a druhé heslo
- User se v konfiguraci cli nepoužívá, ale je z něj vidět jméno tenantu

Základní postupy – s5cmd

.aws/credentials

```
[default]
aws_access_key_id = xxxxxxxxxxxxxxxxxxxx
aws_secret_access_key = xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
```

Endpoint

```
export S3ARGS="--endpoint-url=https://s3.cl2.du.cesnet.cz"
```

Vytvoření bucketu

```
s5cmd $S3ARGS mb s3://data-input
```

Zjištění existujících bucketů

```
s5cmd $S3ARGS ls
```

(https://du.cesnet.cz/cs/navody/object_storage/cesnet_s3/start)

Alternativně `aws s3 […]`
https://du.cesnet.cz/cs/navody/object_storage/awscli/

Základní postupy s5cmd

Kopie dat do bucketu

```
s5cmd $S3ARGS cp file.jpg s3://data-input/
s5cmd $S3ARGS cp directory s3://data-input/
```

List obsahu bucketu

```
s5cmd $S3ARGS ls s3://data-input/*
```

Vyhledání a download podle jména

```
s5cmd $S3ARGS cp s3://data-input/*20201114* .
```

Vypsání zabraného místa v bucketu

```
s5cmd $S3ARGS du s3://data-input/*
```

https://github.com/peak/s5cmd

Kam se rozhlížet dále

Zajímavá témata

- CLI? Podívejte se na MinIO mc cli [MinIO client]
- Zálohování? Podívejte se na restic [restic.net]
- Web stránka? Podívejte se na AWS S3 javascript exlorer [aws-s3-js-explorer]
- Presigned URL
- S3 jako backend pod POSIX FS
  - Hodně kompromisů
  - Typické objektové úložiště (CEPH) poskytuje blokové zařízení (RBD)
- WORM (AWS object lock)
  - Retention period - nelze odstranit dříve než za stanovenou dobu
  - Legal hold – lze odstranit, vhodné pro splnění legislativních podmínek
- Lifecycle policy
- S3 based (static) website delivery

Presigned URL

- Vlastník objektu může vytvořit dočasné „samonosné“ URL
- Alternativa k nastavení ACL
- Download i upload

```
aws s3 presign s3://bucket/file [--expires-in <seconds>]
```

```
https://s3.cl2.du.cesnet.cz/datainput/fotky/20210223_152623.mp4?AWSAccessKeyId=UHGNF
Q8G60H8VB8450D9&Signature=2Iy8Gr9g%2B%2FiXraO8sYs%2Fmlli5J8%3D&Expires=1684064768
```

S3 přes souborový systém

- **s3fs** – objekty v S3 prezentovány jako soubory, metadata uložena v tagu, lokální cache, write-on-close, user-space mount přes FUSE
- **s3ql** – používá S3 jako backend, má v něm vlastní strukturu (vytvoření přes s3ql.mkfs), objekt = blok, lokální cache, write-on-close
- **rclone** – utilita s mnoha funkcemi včetně FUSE (rclone mount)
- **s3backer** – z S3 udělá blokový device

Vlastnosti

- POSIX like přístup (v případě s3ql i náhodný přístup), šifrování, atd., ale:
  - Neumožní vícenásobné připojení RW
  - Pravděpodobně to není rozumné řešení z hlediska celkové koncepce (vs RBD)
  - Virtuální FS versus cloudový koncept nad HTTP (retry)

Take out – Co a k čemu je S3

- Neformální standard (API), FTP přes HTTP, platform agnostic
- Cloud native design (scalable, oddělení rolí, portabilita, multicloud)
- Objekty na jedné hromadě, ale lze používat jako soubory
- Levné, vysoce dostupné, robustní, lineární výkon
- Podle metadat nelze vyhledávat, silná ACL, presigned URL, versování, WORM

Děkuji za pozornost