

Data a datové služby

David Antoř

1. 10. 2026

Datová úložiště MetaCentra z pohledu výpočtáře

- scratch
 - pro zpracovávaná data v rámci aktuální úlohy
- disková pole MetaCentra aka „home“
 - pro zpracovávaná data mezi úlohami
 - souborový systém
 - limitované kapacity
 - staging pro úlohu: typicky pouze kopírování
 - můžu počítat nad daty přímo z diskového pole bez stagingu?
 - pokud úloha jednorázově čte data „od začátku do konce“
 - jinak spíš ne, staging do scratche bude efektivnější

S3 (Simple Storage Service)

- Amazon: Původní služba pro uložení dat ve formě objektů
- DU: S3-kompatibilní služba implementovaná v Ceph

objekt

- analogie souboru v souborovém systému

bucket

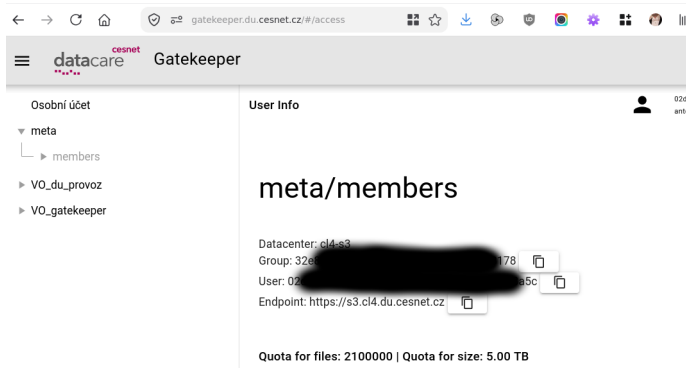
- jmenný prostor pro objekty
- lze si představit jako kořenový adresář (top-level adresář)

Objektová úložiště pro výpočtáře

- objektová úložiště protokolem S3
- určená pro vědecká data ve fázi zpracování
 - důvod je použít: pokud se nevejdete na diskové pole
 - „od TB po PB“
 - cena za to: není to souborový systém, ale S3
 - „pro staging místo cp použiju jiný příkaz“
 - například s5cmd nebo rclone
 - maximální dostupný výkon všech komponent
 - přistupujete přímo k úložišti
- stejně můžete použít pro výpočty i data z jiné/vlastní virtuální organizace

Přístup do S3 MetaCentra

- <https://gatekeeper.cesnet.cz/>



The screenshot shows the Gatekeeper web interface. The browser address bar displays `gatekeeper.du.cesnet.cz/#/access`. The page header includes the **datacare cesnet** logo and the title **Gatekeeper**. On the left, a sidebar menu shows the navigation structure: **Osobní účet**, **meta** (expanded), **members**, **VO_du_provoz**, and **VO_gatekeeper**. The main content area is titled **User Info** and displays details for the **meta/members** group. The information includes: **Datacenter:** cl4-s3, **Group:** 32e...178, **User:** 02...a5c, and **Endpoint:** `https://s3.cl4.du.cesnet.cz`. At the bottom, it states **Quota for files: 2100000 | Quota for size: 5.00 TB**. A user profile icon in the top right corner shows the name **02a anti**.

Speciality na S3: bucket versioning

- jak se chránit před změnou dat nebo nechtěným smazáním
- bucket versioning: Ceph ukládá každou verzi zvlášť (i smazání je verze)
- variace:
 - oddělené identity (i pro jednoho uživatele) – dual-identity WORM pattern
 - aplikační oprávnění v účtu A, admin oprávnění v účtu B
 - účet A není schopen smazat verze nebo jimi manipulovat
 - varianta full Write-Once-Read-Many (WORM)
 - způsob ukládání dat, kde zapsaná data nelze přepsat ani smazat po stanovenou dobu (nebo navždy)
 - legal lock
 - musí být zapnut při vytvoření bucketu (nelze dodatečně)

Speciality na S3: šifrování na klientské straně

- silná ochrana před únikem dat
- S3 bucket lze opatřit šifrovanou vrstvou na straně S3 klienta
 - data jsou na úložiště přenášena a ukládána zašifrovaná
- ideální S3 klient pro šifrování na klientské straně je rclone
 - pro potřeby skupin je možné využít sdílený bucket se šifrovanými daty
 - máme nástroj pro výrobu šablony šifrovaného bucketu a její další sdílení koncovým uživatelům
 - manažer nastaví master key a zašifruje config a nasdílí uživateli
 - uživatel do šablony doplní svoje kredence a může použít sdílenou šifrovanou vrstvu/prostor

Další služby úložišť

- předávárna souborů FileSender
 - lze volat přes API z MetaCentra
 - „chci někomu předat výsledky ke stažení“
- sync-and-share (ownCloud/CesBox)
 - pro data pro výpočty?
 - srsly...?

Objektová úložiště pro výpočtáře – novinky

- cloud tým zprovoznil službu Onedata
 - „velké úložiště s hezkým rozhraním“
- více v části Tomáše Svobody
- data z Onedata se ukládají na objektové úložiště
 - to ale uživatel nevidí

- Národní repozitářová platforma
 - distribuovaný, multi-tenant systém pro vytváření repozitářů
 - dovoluje vytváření repozitáře z připravených komponent „jako službu“
- repozitář
 - v první řadě úložiště dat s (obsahovými) metadaty
 - obsahuje „citovatelné datové sady“

Hledání dat v repozitářích

- jestli máte svůj oborový repozitář, hledejte rovnou v něm
- pokud nemáte, hledat datové sady lze v Národním metadatovém adresáři
- NMA (<https://nma.eosc.cz/>)
 - agregátor metadat
 - dovede k repozitáři s daty
- repozitář s daty umí jemnější vyhledávání

Jak si zvolit vhodný repozitář pro ukládání

- dle zvyklostí vědecké disciplíny
- obecná vodítka
 - první volba: známý komunitou uznávaný repozitář
 - pokud není, tak institucionální nebo catch-all
 - pracovní skupiny EOSC budou ustavovat své oborové repozitáře

Catch-all datový repozitář

- <https://datarepo.eosc.cz/>
- repozitář poslední volby
 - pro individuální vkladatele, kteří nejsou zapojení ve výzkumných komunitách
 - nebo pro malé komunity
- s obecným metadatovým modelem
- záznamy mohou být zveřejněny/zveřejněna pouze metadata/neveřejné
 - lze nastavit embargo
- lze přiřadit DOI
- pokud reprezentujete větší organizovanou komunitu uživatelů, chcete specifický repozitář

Ukládání do repozitáře

- kdy ukládat data do repozitáře
- je třeba vyvážit
 - „co nejdříve“
 - data už se nemají měnit
 - data mají hodnotu do budoucna
 - chceme umět prokázat jejich původ
 - chceme s nimi pevně svázat metadata
 - „ale ne dřív“
 - např. pokud většinu z nich zahodíme
 - nebo se budou výrazně měnit

Jak použít repozitář ve výpočtech

- drobná data lze stáhnout ručně
- integrace do Galaxy
- pro příkazovou řádku `nrp-cmd`
 - funguje pro CESNET-Invenio verze 14 a výše
 - tedy např. pro catch-all repozitář
 - `module add nrp-cmd`
 - `nrp-cmd add https://datarepo.eosc.cz datarepo`
 - `nrp-cmd create-record | get-record`

- pokud nevystačíte s kapacitou diskových polí MetaCentra, lze použít objektová úložiště
- jestli jsou data trvale užitečná, měla by spíše končit v repozitáři

Shrnutí a kontakty

- při výběru služby úložišť se podívejte na základní vlastnosti
 - zda se hodí pro zamýšlený účel
- pokud si nejste jisti, napište
- kontakty a dokumentace
 - meta@cesnet.cz
 - obecný kontakt support@cesnet.cz
 - <https://docs.du.cesnet.cz>
 - <https://docs.nrp.eosc.cz>

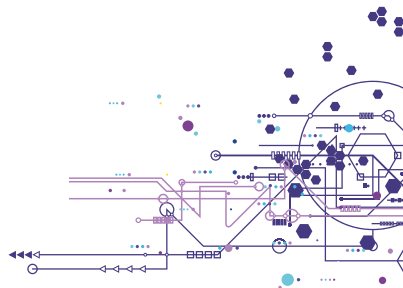


Lemon curry?

The logo for e-infra.cz, consisting of the text 'e-infra.cz' inside a circle, which is flanked by two curved lines on either side.

e-infra.cz

Backup slides



Co se stalo na CL4 I

- „pro získání pocitu, v čem spočívá problém“
- a bez ambice na technickou přesnost
- jak funguje Erasure Coding pool, řekněme 12+4
 - každý kus dat se rozdělí na 12 dílů a k nim se dopočítají 4 redundantní („parity“) pro případ opravy
 - ty se zapíší na 16 různých disků
 - čtení: když je těch 12 dílů v pořádku, tak se přečtou a hotovo
 - když by se kterýkoli poškodil, dá se dopočítat z ostatních (math&magic)
 - u každého kusu se drží kontrolní součet, aby se poznalo, že je OK
 - čtení: normálně se vezme těch 12 dílů a složí se
 - pokud by některý byl poškozen, tak se opraví a zapíší

Co se stalo na CL4 II

- chyba <https://tracker.ceph.com/issues/80848>
 - nahlásil kolega blízký vývoji jádra cephu, který chybu zvládl reprodukovat
 - je možné, že jsme první instance na světě, kde se problém projevil
- plugin CLAY dělá zápisy na disky
 - každý zápis musí sedět na 32bajtovou adresní hranici
 - před zápisem se zkontroluje, jestli políčko sedí na hranici, pokud ne, tak v políčku zůstane původní obsah
 - z pohledu zapisovaného souboru tedy náhodný
- Ceph Squid vs. Tentacle
 - Squid zarovnával na 32 bajtů
 - Tentacle na 64
 - takže cca polovina zápisů se netrefí a zahodí se
 - žádná chyba, žádné varování

Co se stalo na CL4 III

- když storage server naběhl na nové verzi
 - každý zápis uložil správná data, ale špatné parity
 - dokud se nic nepokazí, čtení dá správná data
 - ale chybí pojistka s paritami pro případ poškození některého dílu
 - při čtení se může narazit na chybu, ta se opravuje z parity
 - pokud je potřeba opravit parity, tak se spočtou špatně a tentokrát zapíšou na disk
 - samotné úložiště dostává k zapsání „platná data“, jen jiná, než mělo
- proč se to dělá: zarovnání na 64 bajtů zrychlí operace
 - používá se funkce `rebuild_aligned`, která zajišťuje, aby se buffer zarovnal
 - CLAY předpokládá, že se zarovnané pole už nikam nepřesune, `rebuild_aligned` ho přesune

Co se stalo na CL4 IV

- čeho se to týká
 - pouze erasure coding poolu ve specifických kombinacích disků
 - u nás prakticky na CL4 (S3 a některé pooly pro RBD)
- co děláme teď
 - potřebujeme detailně určit, co vše mohlo být poškozeno
 - a jak/jestli to jde opravit
 - to závisí na řadě okolností, bude potřeba připravit řadu scénářů
 - data zapsaná po upgradu
 - data před upgradem, se kterými Ceph manipuloval (zejména jejich parity)
 - jak rozpoznat korektní data (přestalo platit „Ceph uživateli poskytne pouze korektní data“)