cesnet

DATOVÁ ÚLOŽIŠTĚ CESNET, FAIR DATA

David Antoš

CESNET

11. 5. 2018

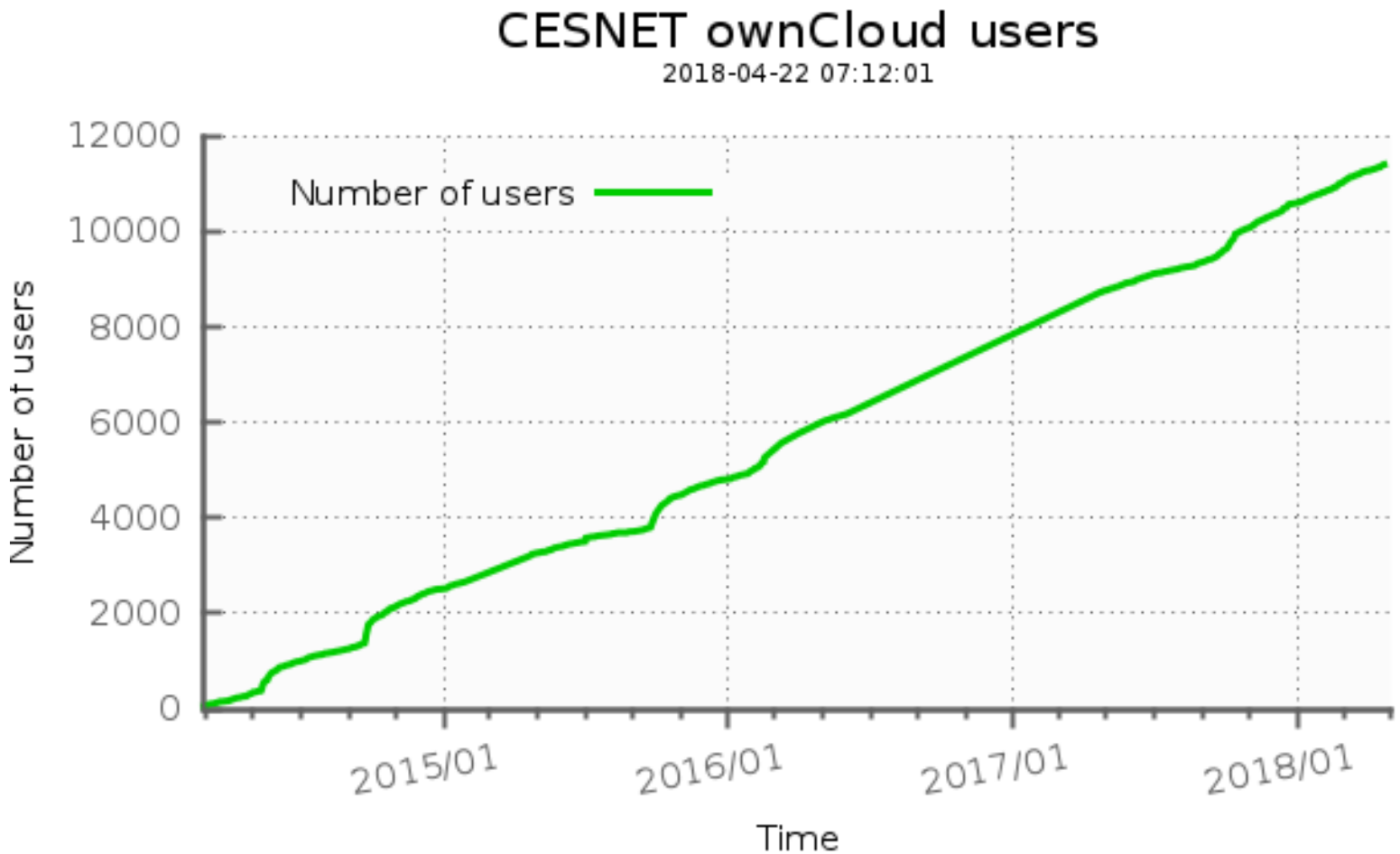
Přehled

- typické možnosti použití datových úložišť
- současná a budovaná infrastruktura
- přesun na nové úložiště
- připravované služby
  - komunitní budování úložiště
  - dlouhodobé archivy
- FAIR data – o čem mluví mezinárodní komunita

„Instantní“ služby úložišť

- jednorázový přenos souborů
  - „moc velkých do mailu“
  - FileSender
  - `http://filesender.cesnet.cz`
- synchronizace souborů mezi svými počítači a mobilními zařízeními
  - s webovým rozhraním
  - s možností sdílení dat
  - “sync’n’share”
  - ownCloud
  - `https://owncloud.cesnet.cz`

ownCloud – počet uživatelů



Možnosti využití úložišť

- zálohy
  - uživatelé mají primární data u sebe
  - na úložiště odkládají zálohu pro případ havárie
  - buď pro zálohování jednotlivých strojů
  - nebo i agregovaně – IT oddělení zálohuje celou katedru
- archivace
  - uživatelé na úložiště odkládají cenná primární data
  - data nejsou často využívána
  - uživatelé nemají prostředky pro jejich uchovávání
  - individuální přístup koncových uživatelů vs. „laboratorní archivář“

Možnosti využití úložišť

- sdílení dat
  - distribuovaný tým potřebuje společně pracovat nad většími objemy dat, případně je zveřejňovat
  - typicky koncoví uživatelé
- „něco jiného“
  - distribuce obsahu, jiné speciální aplikace
- realizováno přístupem k souborovému systému nebo objektovému úložišti

Přístup k souborovým úložištím

- uživatelé MetaCentra mají na úložiště přístup v jeho rámci
- pro ostatní: základní služba na registraci přes `http://du.cesnet.cz`
  - pro individuální data, nejvýše jednotky TB
- jiné použití po konzultaci s uživatelskou podporou
  - „založíme vám virtuální organizaci (VO)“
    - tj. skupinu uživatelů, která má správce a členy
  - správce VO dohodne se správcem úložiště technické parametry
    - objem dat, počet replik, . . .
  - správce VO rozhoduje o členství jednotlivých uživatelů ve VO

Stávající infrastruktura

- hierarchická úložiště pořízená v letech 2011–13
  - Plzeň, Jihlava, Brno
- celkem 22 PB hrubé kapacity médií
- připojena do MetaCentra
  - resp. „poskytují zdroje VO MetaCentrum“
  - v Meta jako „. . . -archive“
- infrastruktura dosluhuje a je obměňována

Nová úložiště

- úložiště v Ostravě
  - hierarchický systém
  - 5 PB disků, 17 PB pásek
  - připojený do MetaCentra
- plány (projekt OP VVV)
  - 2018 diskové pole a cluster pro objektové úložiště
  - 2019 cluster pro objektové úložiště

Stěhování

- úložiště v Plzni (du1) bude odstaveno
- data uživatelů MetaCentra přesuneme do Ostravy
  - typicky jsou to data trvalé hodnoty
  - po dobu přesunu budou znepřístupněna
    - desítky hodin
- v ostatních VO postupně oslovujeme uživatele, aby
  - začali zálohovat na nové úložiště
  - nebo si přesunuli data
- nepřesouváme vše sami, u záloh to nemá smysl

Zálohy vs. archivy

- úložiště není nekonečné
- rozdělili jsme proto dva hlavní typy použití
  - data trvalé hodnoty – archiv
    - je omezen kvótou, tj. objemem dat
  - zálohy
    - jsou omezeny časem uložení
    - dnes jeden rok
    - pak je správce úložiště smí smazat(!)
    - samozřejmě předtím vlastníka dat upozorní
    - při rozumné zálohovací strategii se data rotují
- platí na novém úložišti (= Ostrava)
  - na starých je vše považováno za archiv

Komunitní úložiště

- spolupracovat se skupinami, které mají velká data
  - ukládání, zpracování
- komunitní modely budování úložišť
  - s HSM není snadné podpořit model „a co kdybych vám koupil disky/pásky/. . . “
  - jsme monolitický poskytovatel *kapacity*
- chceme umět podpořit scénář typu „chtěl bych na tři místa uložit 500 TB dat“
  - „tak si kupte 1500 TB a zapojte se k nám“
  - „500 budete mít přímo u sebe, další dvě repliky zařídí infrastruktura, zbylou kapacitu úložiště použijeme pro repliky ostatních“
  - a provozovali bychom dostatečnou infrastrukturu pro pokrytí nárazových a dočasných potřeb

Služba archivu pro LTP

- implementace systému temného archivu
  - pro dlouhodobé ukládání dat
  - s garancemi binární korektnosti
  - a pravidelnými kontrolami konzistence
- použitelného jako spolehlivý backend pro LTP systémy dle standardu OAIS
- služba je koncipována jako nadstandardní
- API + webové rozhraní + dokumentace
- probíhá implementace

FAIR data

- Findable, Accessible, Interoperable, Reusable
- prosté ukládání souborů nezaručuje, že data budou v budoucnu užitečná
- FAIR principy byly formulovány cca 2014–15, publikovány 2016
  - a jsou dnes propagovány organizacemi jako FORCE11, Evropskou komisí, National Institutes of Health (US), Australian National Data Service, . . .
  - jsou formulovány nezávisle na disciplíně
- jsou vnímány jako nástroj (seznam témat k zamyšlení?), který pomůže
  - objevování znalostí a inovaci
  - sdílení a znovupoužití dat
  - strojovému zpracování dat

FAIR a strojové zpracování

- „Sovětští vědci jsou přesvědčeni, že do konce příští pětiletky vytvoří automatický stroj, který bude schopen vypracovat národohospodářský plán SSSR a provést jeho analýzu.“
  - – zpráva ze začátku 50. let
- Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals.
  - – The FAIR Guiding Principles for scientific data management and stewardship, https://www.nature.com/articles/sdata201618

Findable

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

Accessible

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
  - A1.1 The protocol is open, free, and universally implementable
  - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

- všimněte si: metadata jsou věčná, data mohou zmizet

Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

- stručně: v metadatech používejte řízené slovníky a specifické odkazy
  - „X je řízeno Y“ je lepší než „X souvisí s Y“
  - samozřejmě formálním jazykem

Reusable

- R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage license
  - R1.2. (Meta)data are associated with detailed provenance
  - R1.3. (Meta)data meet domain-relevant community standards

Co FAIR je a není

- FAIR $\neq$ veřejně přístupný
- FAIR $\neq$ dostupný zdarma
  - jen má být jasné, jak se věci mají
- FAIR je užitečný checklist i pro tvorbu DMP
  - „nad čím bychom se měli zamyslet“
  - ne „vše z toho je třeba splnit dokonale“
- FAIR není standard
- FAIR je svatý grál
- FAIR je proces
  - level of FAIRness
  - FAIRification (ehm)

Jak s FAIR můžeme pomoci

- umíme ukládat data (i velká)
- učíme se repozitáře a dlouhodobé uchovávání
- potkáváme se s komunitami, které se v tom (taky) plácají
  - a spolupracujeme s nimi
- umíme poradit s „technickou“ částí metadat

Shrnutí

- typické scénáře použití datových úložišť
- služby nad rámec „dalšího úložiště v MetaCentru“
- stávající a plánovaná infrastruktura
- stěhování dat a změny v jejich organizaci
  - archiv vs. zálohy
- nové služby
- FAIR