NÁSKOK
DÍKY
ZNALOSTEM

PROFINIT

NDBI047
Aplikace bigdatových technologií v data science

Úvodní cvičení

Jan Hučín

22. 2. 2019

Osnova cvičení

1. Důležité odkazy
2. Co je potřeba umět
3. Úvod do Hadoopu pro lidi
4. Cvičný cluster, přihlášení
5. První kroky na clusteru

PROFINIT

Důležité odkazy

Materiály k výuce

- https://github.com/profinit/MFF-BDT

Dokumentace ke cvičnému clusteru

- https://wiki.metacentrum.cz/wiki/Hadoop
- https://www.metacentrum.cz/cs/Sluzby/Hadoop/

Přihlášení na cluster

- žádost podat na www.metacentrum.cz/cs/hadoop, skupina UK:MFF:NDBI047
- ssh na hador.ics.muni.cz

Co je potřeba umět (a mít)

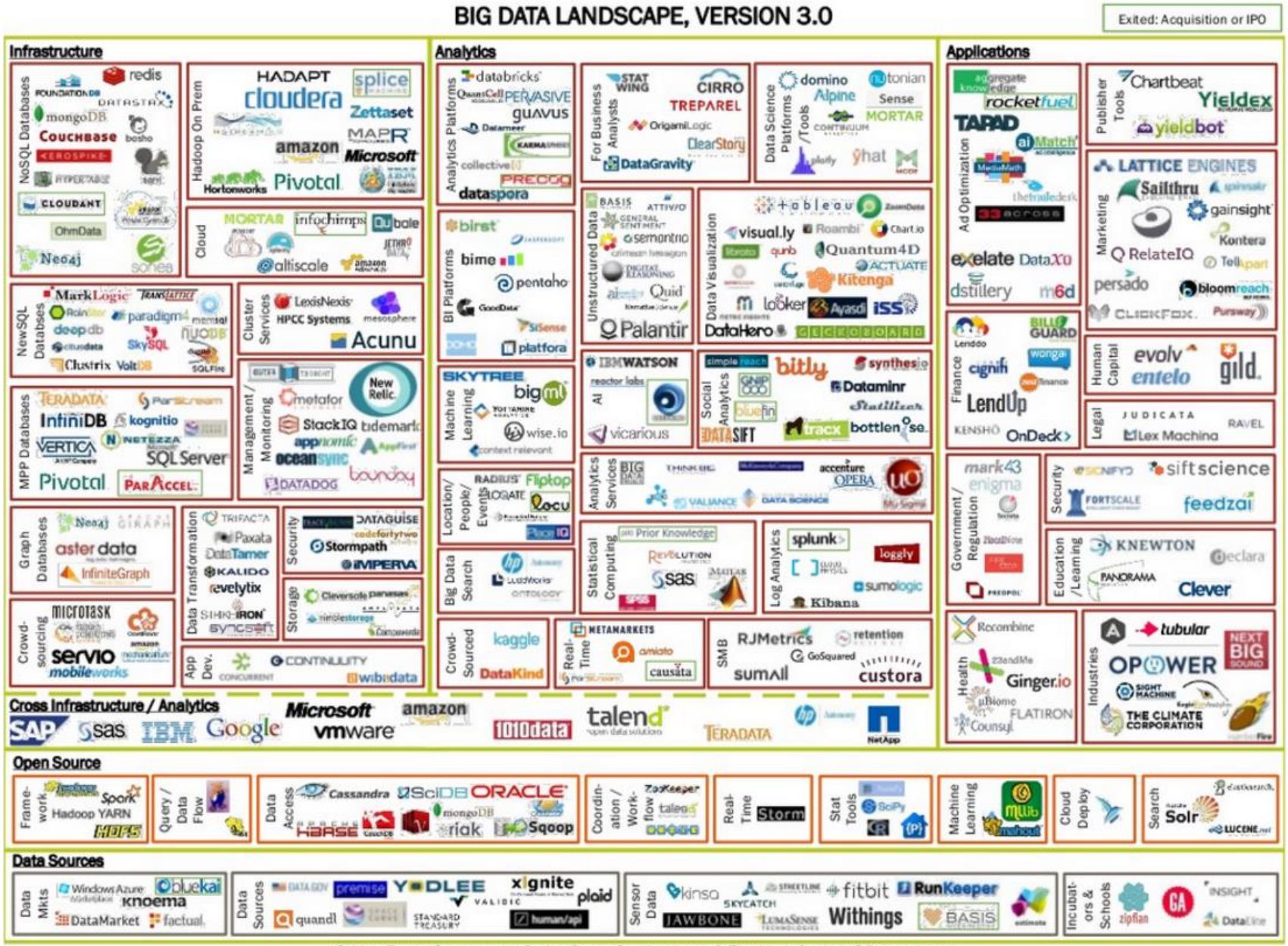
- Linux – práce v příkazové řádce, základní příkazy, práva
- SQL – SELECT s agregací, JOIN, INSERT, CREATE
- Python – definice funkce, typ list, řetězce, for, if-else
- regulární výrazy – základy, jednoduché substituce

- čím víc umíte, tím víc si předmět užijete

- přístup na internet
- klient pro SSH (např. Putty)
- webový prohlížeč

Úvod do Hadoopu pro lidi

Big data neznamená Hadoop



PROFINIT

Apache Hadoop

- Systém (Framework) pro distribuované ukládání a distribuované zpracování velkých datových souborů
- Východiska:
  - některé datové soubory jsou opravdu velké
  - mnoho slabších strojů udělá dohromady víc než jeden výkonný
  - horizontálně škálovatelný systém má mnoho výhod
  - stroje se občas rozbijí

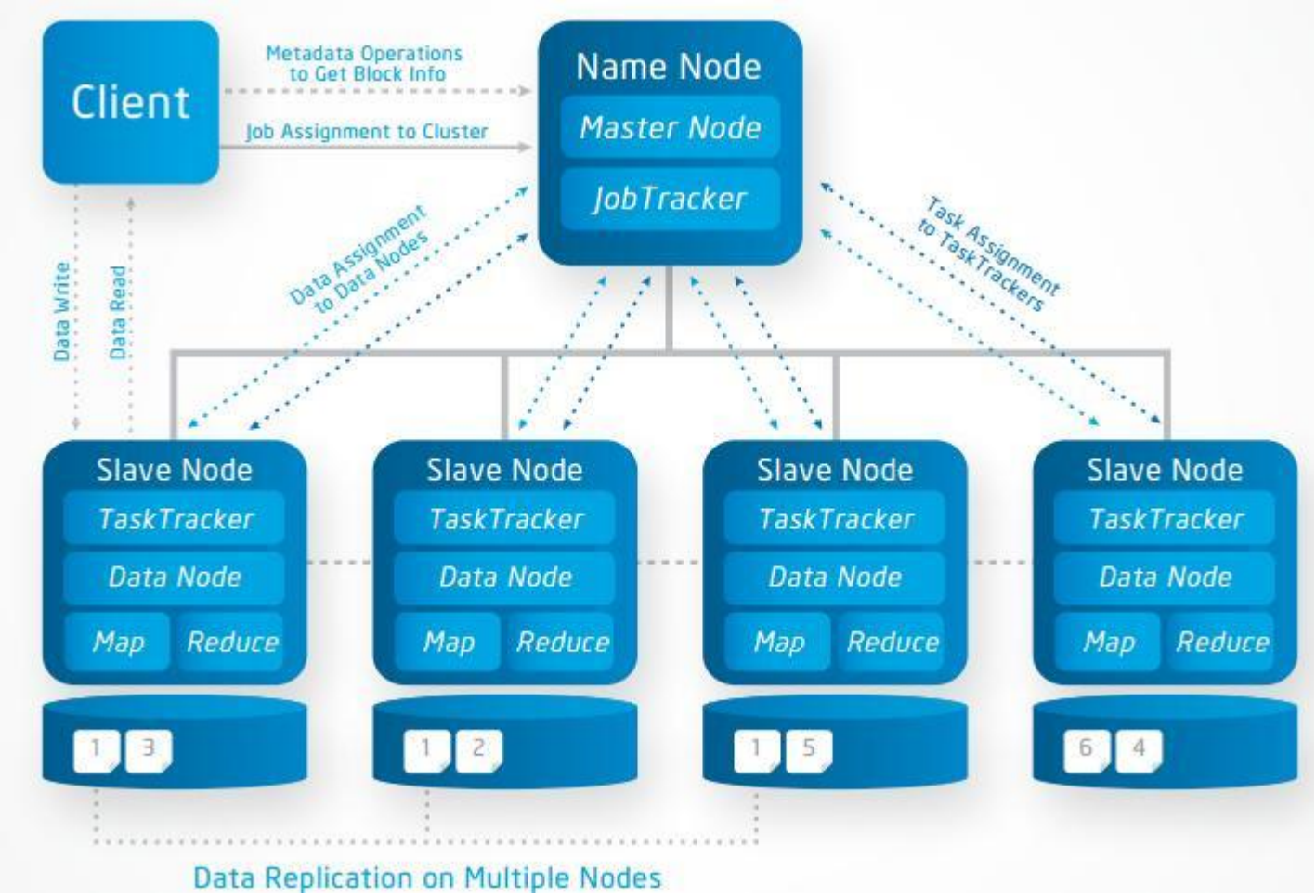
Apache Hadoop

- Systém (Framework) pro distribuované ukládání a distribuované zpracování velkých datových souborů
- Jak se to řeší:
  - některé datové soubory jsou opravdu velké
    → rozdělení dat mezi několik strojů
    → preference sekvenčního čtení
  - mnoho slabších strojů udělá dohromady víc než jeden výkonný
    → rozdělení práce mezi několik strojů, paralelizace
    → data zpracovat tam, kde jsou uložena
  - horizontálně škálovatelný systém má mnoho výhod
    → rozumné přidělování zdrojů
    → přidání nebo odebrání stroje
  - stroje se občas rozbijí
    → replikace souborů (defaultně 3 kopie)
- **Úzké hrdlo – přesuny dat po síti**

PROFINIT

Hadoop a cluster

- **cluster** – skupina nodů s přidělenými rolemi
- **node** – „stroj“, pracovní jednotka s vlastním OS



Hadoop – komponenty

- správa úložiště – **HDFS** (Hadoop Distributed File System)
- resource manager – YARN
- SQL databáze – **Hive**, Impala
- noSQL databáze – HBase
- výpočty – **MapReduce**, **Spark**
- scheduling – Oozie
- streaming – Storm
- export/import – Sqoop
- atd.

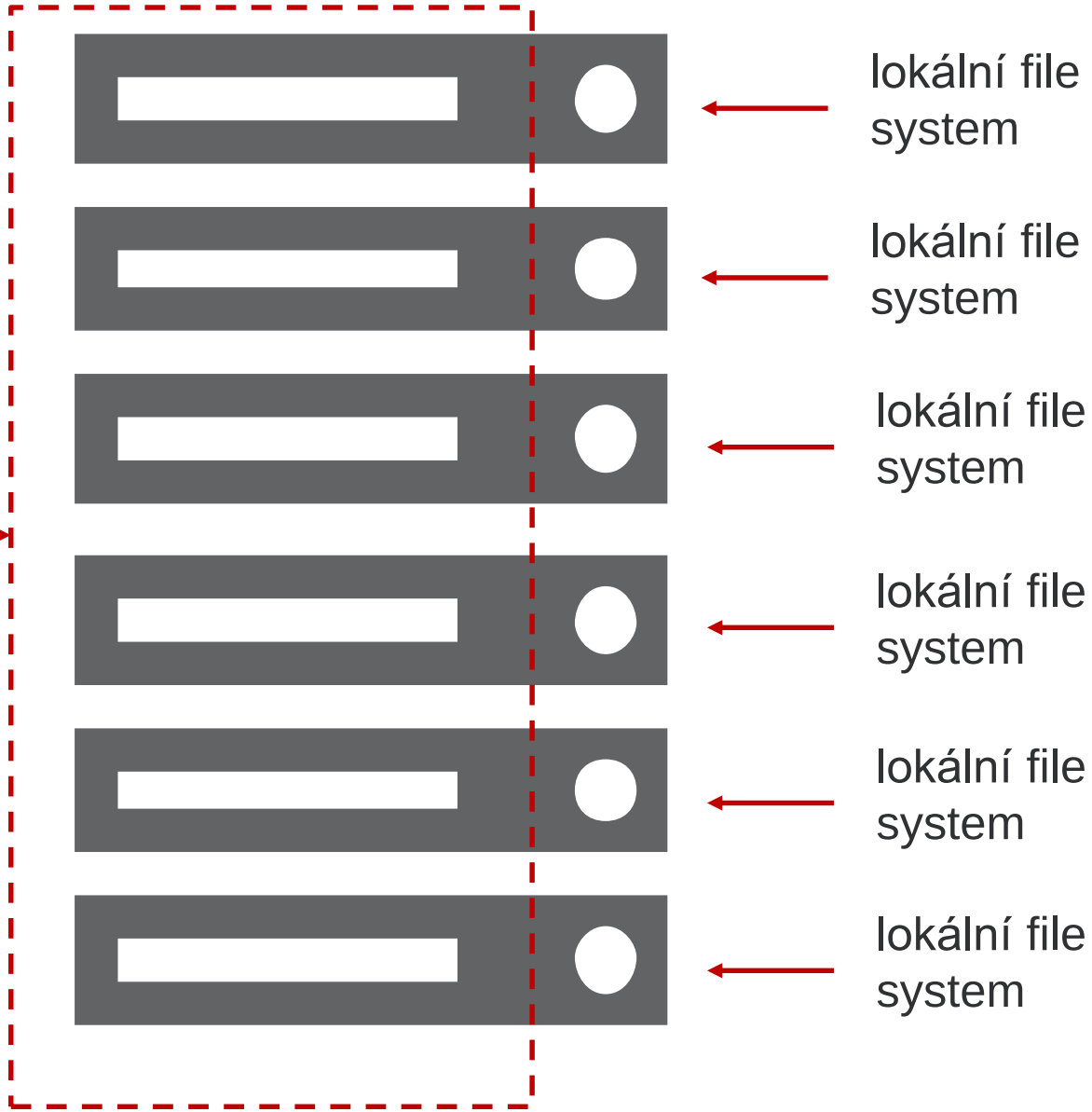
Distribuce – řešení závislosti

cloudera

HORTONWORKS®

PROFINIT

HDFS



PROFINIT

HDFS

- Optimalizovaný pro
  - velké soubory
  - sekvenční čtení
  - paralelní zpracování
- Špatný pro
  - spoustu malých souborů
  - náhodný přístup
  - nízkolatenční přístup

YARN

- **Y**et **A**nother **R**esource **N**egotiator
- Plánovač a alokátor zdrojů
  - paměť
  - CPU
  - počet vláken
  - síť...
- Většinou transparentní – nevyžaduje zásah uživatele
- Požadavky na zdroje ale může uživatel upřesnit (Spark)
- Ne všechny aplikace YARN využívají

PROFINIT

Hive a Impala

- Emulace SQL světa na Hadoopu nad HDFS:
  - adresáře ~ databáze
  - podadresáře ~ tabulky
  - HiveQL jako dialekt SQL
- Hive má pomalý start, jednoduché dotazy trvají dlouho

PROFINIT

MapReduce

- Paradigma pro paralelní zpracování
- Cyklus MR:
  - načti data
  - transformuj data do párů (klíč, hodnota) – fáze **map**
  - shromáždi páry se stejným klíčem – fáze **shuffle**
  - proveď výpočet (agregaci) odděleně pro každý klíč – fáze **reduce**
  - výsledek zapiš
- Pomalé, ale účinné
- Využíval např. Hive

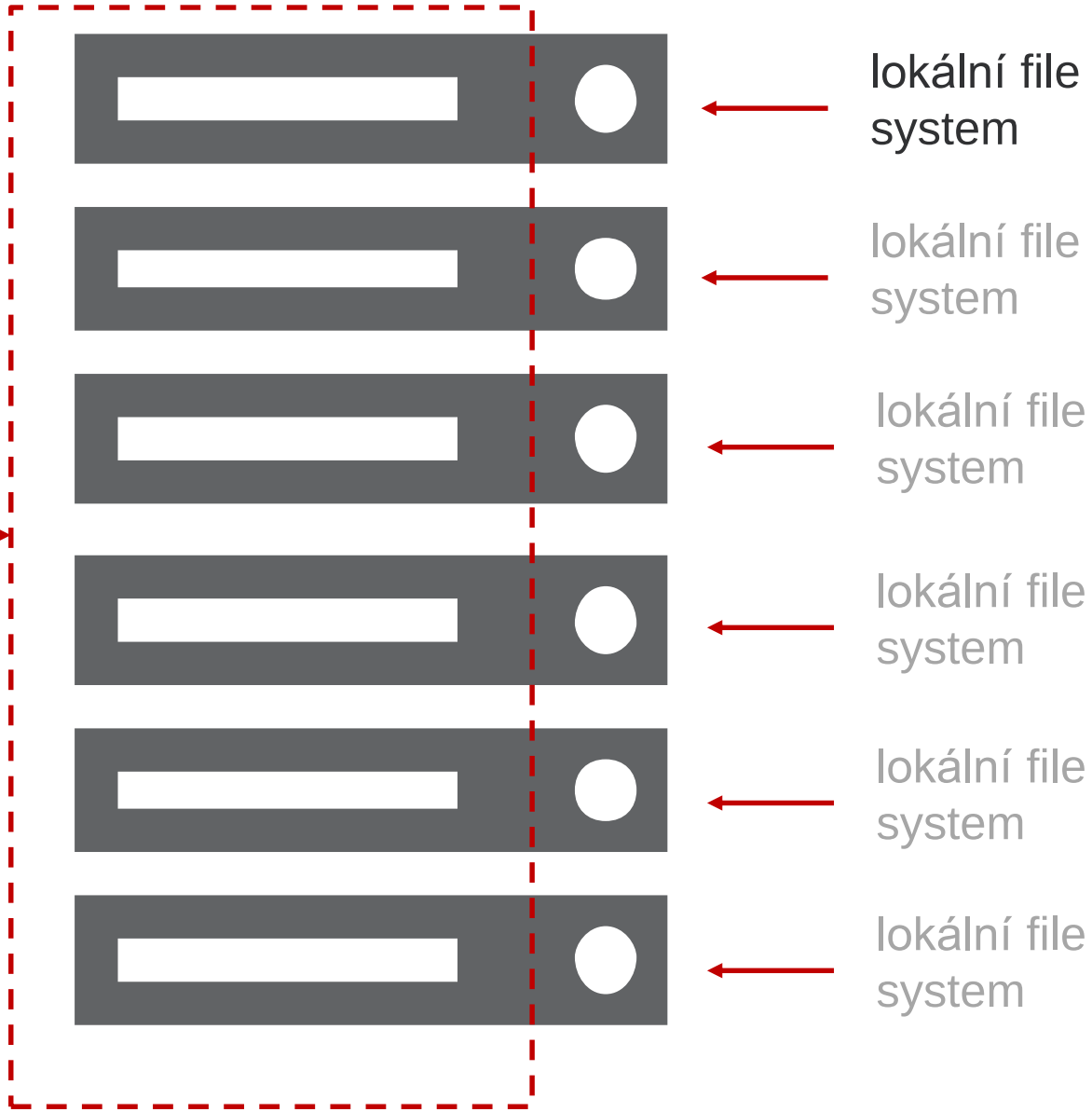
Spark

- Framework pro paralelní zpracování in-memory
- Podobný princip jako MapReduce, ale mezivýsledky drží v paměti
- Rychlý, ale závislý na dostatku zdrojů
- Oblíbený → mnoho nadstaveb:
  - Spark SQL
  - Spark Streaming
  - Spark ML
  - GraphX

PROFINIT

Cluster a první kroky

HDFS



PROFINIT

HDFS příkazy

`hdfs dfs -akce parametry`

Akce

- `ls` → výpis adresáře
- `mkdir` → vytvoření adresáře
- `cp` → kopírování v rámci HDFS
- `mv` → přesun v rámci HDFS
- `rm` → mazání souboru nebo adresáře

Domovský adresář na HDFS

`/user/login`

Adresář nelze změnit – neexistuje příkaz `-cd` !

PROFINIT

HDFS příkazy

`hdfs dfs -akce parametry`

Akce

- `put` → kopírování z lokálního FS na HDFS
- `get` → kopírování z HDFS na lokální FS
- `cat` → výpis obsahu souboru
- `chmod` → změna přístupových práv

Příklady

```
hdfs dfs -mkdir work
hdfs dfs -put data/*.csv work
```

PROFINIT

Samostatná práce

viz zadání cvičení 01_CLUSTER_BASICS

github.com/profinit/MFF-BDT

PROFINIT

Díky
za pozornost

PROFINIT
NÁSKOK DÍKY ZNALOSTEM

Profinit EU, s.r.o.
Tychonova 2, 160 00 Praha 6

Telefon
+ 420 224 316 016

Web
www.profinit.eu

LinkedIn
linkedin.com/company/profinit

Twitter
twitter.com/Profinit_EU