NÁSKOK
DÍKY
ZNALOSTEM

PROFINIT

A4M33BDT
Technologie pro velká data

První kroky na clusteru

Jan Hučín ft. Tomáš Kadlec & Petr Paščenko

15.3.2017

Osnova cvičení

- Přihlášení na cluster
- Lokální filesystem versus HDFS
- Cvičné úlohy
- Administrace clusteru přes Ambari
- Začínáme používat Hive

PROFINIT

Uvědomíme si, že...

- ... pracujeme na jednorázově propůjčeném clusteru.
- ... abychom se víc naučili, máme trochu větší práva.
- ... je nás hodně, sdílíme stejný prostor a stejné prostředky.
- ... každý má své tempo a svou úroveň dovedností.
- ... dnes zkoušíme základy, složitější úkoly budou později.

PROFINIT

Připojení na cluster

1. **Příkazová řádka**
2. **Webové rozhraní**
   - Ambari (hlavně administrace)
   - Hue (hlavně dotazování a skriptování)

› **Ad 1.** Příkazová řádka

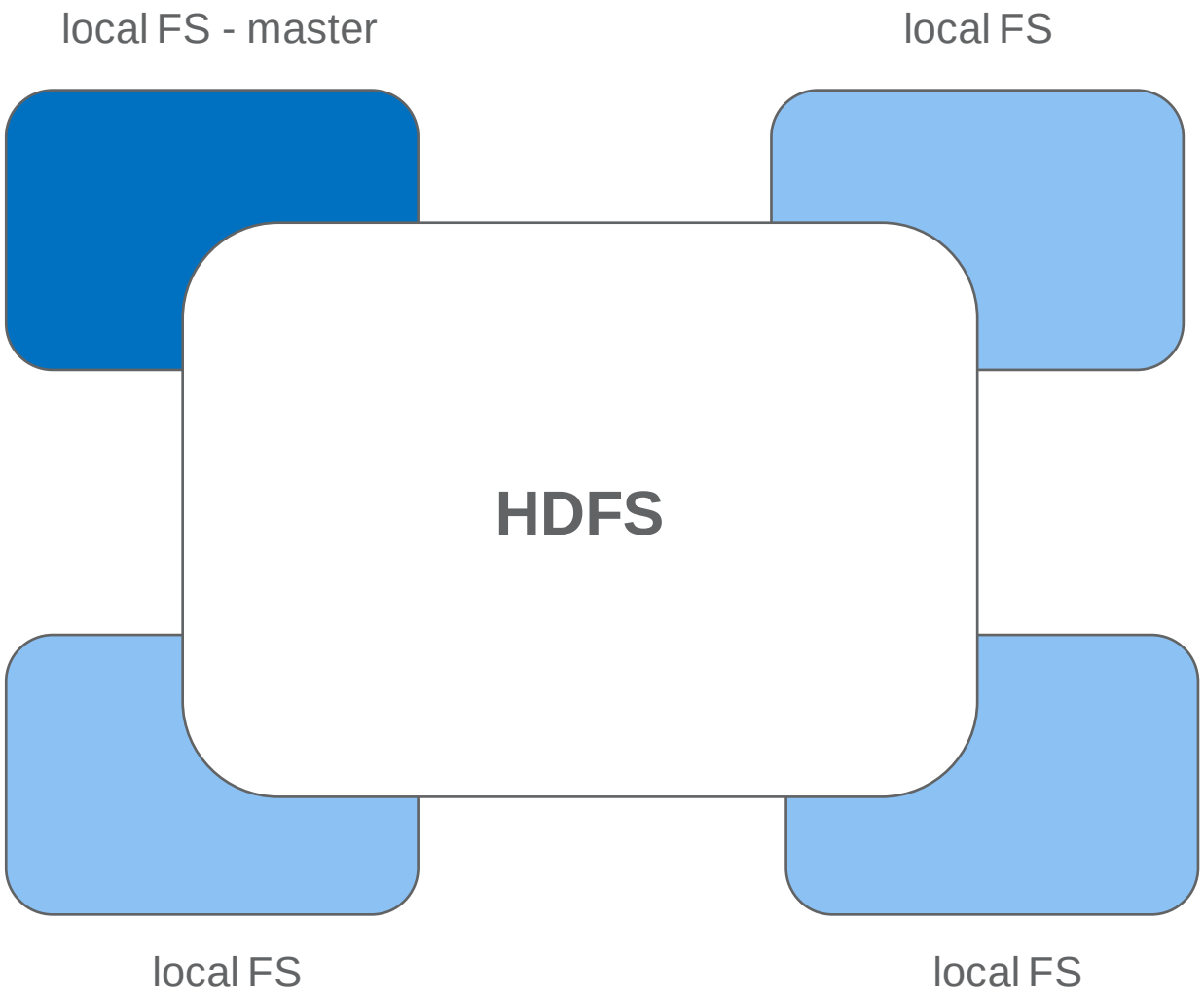
› PuTTY: 212.24.144.105

› login + heslo

› **Ad 2.** bude později

PROFINIT

Lokální filesystem versus HDFS



- (Jen) část úložiště každého nodu je vyhrazena pro HDFS.

Lokální filesystem versus HDFS

local FS

/home/*user_dir*

```
mkdir /home/novakm/pracovni
cp data1.csv ..
ls -l
```

HDFS

/user/*user_dir*

```
hadoop fs –mkdir user/novakm/pracovni
hadoop fs –get data1.csv /temp
hadoop fs –ls –h
```

› Příkazy týkající se HDFS začínají „hadoop fs“ nebo „hdfs dfs“.

PROFINIT

HDFS příkazy (hadoop fs)

- `-ls`
  obsah adresáře
- `-put`
  kopírování local FS ☞ HDFS
- `-get`
  kopírování HDFS ☞ local FS
- `-cp`
  kopírování v rámci HDFS
- `-cat`
  kopírování na standardní výstup
- `-rm`
  mazání souboru
- `-chmod`
  změna práv

☞ /home/_data/manuals

PROFINIT

Cvičné úlohy

Využijeme Linux a regulární výrazy (grep, sed, wc, ...)

› Zkopírovat na local FS do svého adresáře z HDFS (/user/_data) soubory s teplotami a ratingy šachistů.

1. Kolik mužů a žen z ČR (CZE) je v souboru s ratingy? Porovnejte počty i pro jiné země.
2. Zvolte si některé město USA s měřicí stanicí. Který den v červenci je pro poledne nejvyšší teplotní normál a kolik to je?
3. Zvolte si měsíc, den a hodinu.
   - Ze souboru s teplotami vyfiltrujte jen řádky pro tento okamžik.
   - Zjistěte, kolik jich je.
   - Pro zdatné: omezte řádky jen na název stanice a teplotu, seřaďte podle názvu stanice.
   - Uložte řádky do nějakého souboru na lokální FS (váš prac. adresář).
   - Zkopírujte tento soubor na HDFS do svého prac. adresáře.
4. V kolika verších bible se vyskytuje slovo „dog“ a v kolika „cat“?

PROFINIT

Zpracování souborů přímo z HDFS

- `hadoop fs -cat` *`soubor`* `|` *`příkaz Linuxu`*
- vhodné pro **textové** soubory: malý soubor nebo jednorázový průzkum souboru (počet řádků, výpis několika prvních řádků...)

PROFINIT

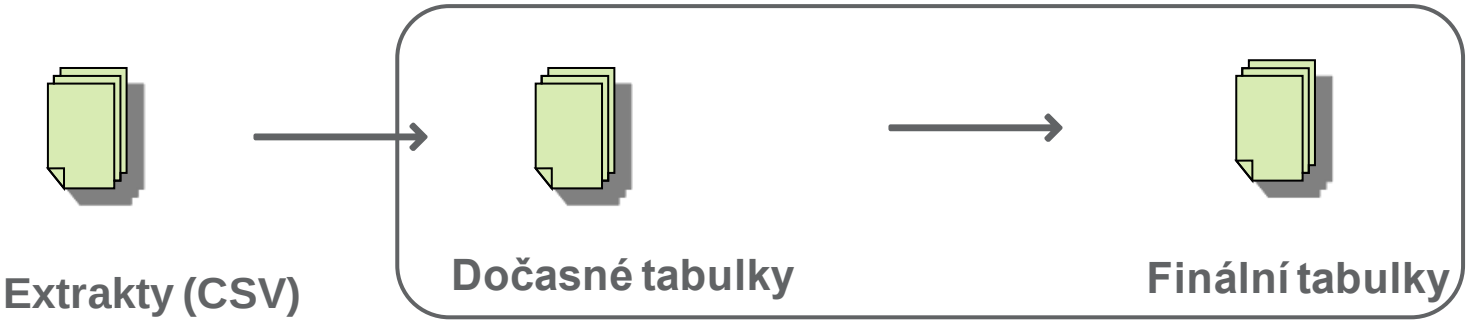
Správa clusteru

Webové rozhraní Ambari (nebo Hue)

- http://212.24.144.105:8080
- login + heslo

PROFINIT

Hive – vytvoření tabulky



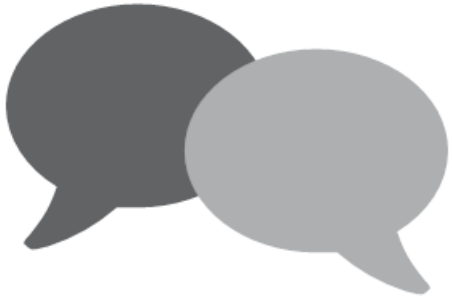
PROFINIT

Hive – úvodní cvičení

- Vytvoříme externí tabulku pro teploty.
- Nejprve si **každý zkopírujeme** adresář /user/_data/teploty-usa **k sobě** (v rámci HDFS).

```
CREATE EXTERNAL TABLE jprijmeni_temp (
    pole formát
    )
formát řádků (oddělovače, ukončení)
formát dat
umístění adresáře s daty
vlastnosti (např. vynechat první řádek)
;
```

PROFINIT



Diskuze

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