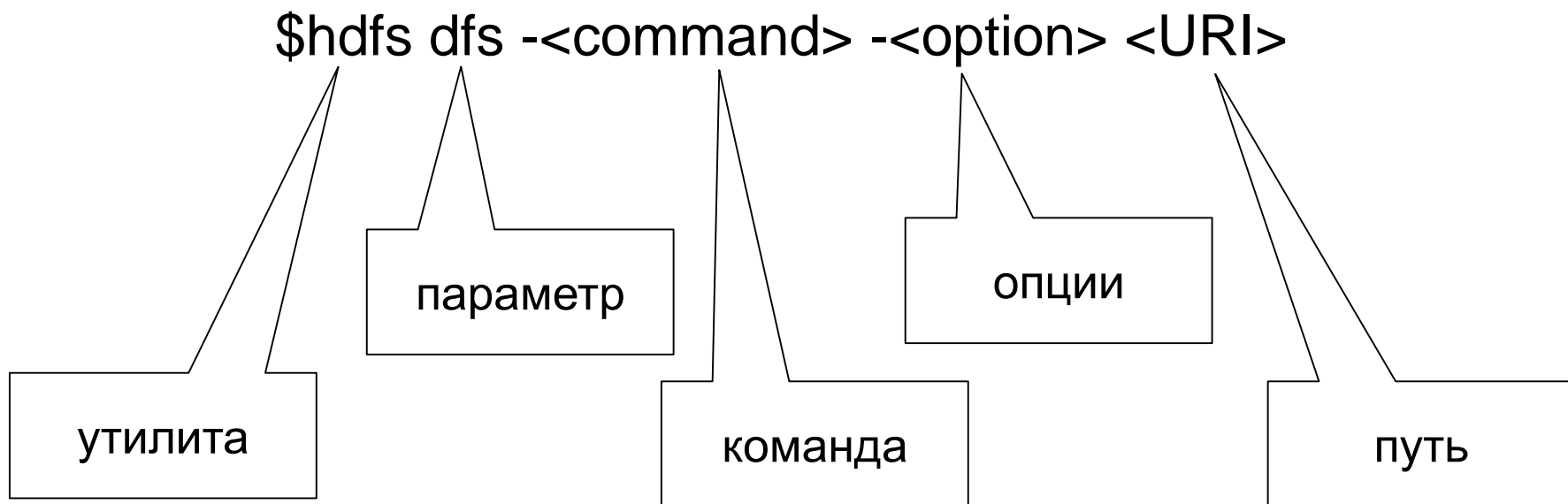


# Hadoop-2

# Структура команды

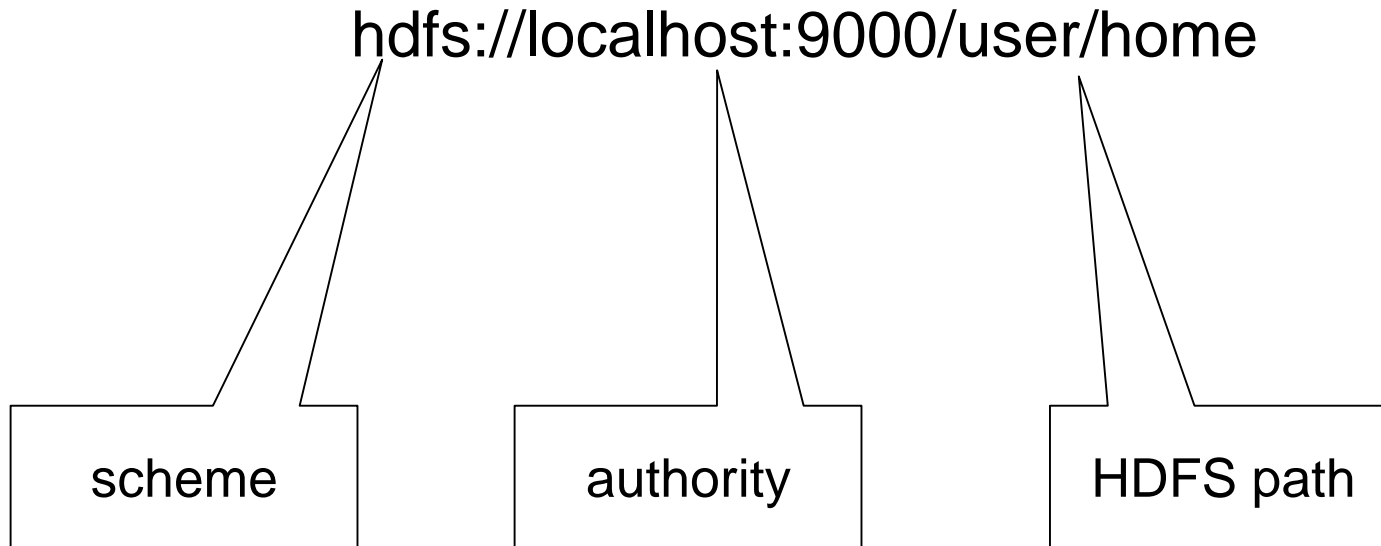


## Пример:

```
hdfs dfs -ls /  
hdfs dfs -ls file://
```

<URI> -- путь в виде URI-схемы. URI (/ˌjuː ɑːr 'aɪ/ англ. Uniform Resource Identifier) — унифицированный (единообразный) идентификатор ресурса. символьная строка, позволяющая идентифицировать какой-либо ресурс: документ, изображение, файл, службу, ящик электронной почты и т. д. Прежде всего, речь идёт о ресурсах сети Интернет и Всемирной паутины. URI предоставляет простой и расширяемый способ идентификации ресурсов. Расширяемость URI означает, что уже существуют несколько схем идентификации внутри URI, и ещё больше будет создано в будущем.

# URI



## **local:**

`hdfs dfs -ls file:///`

`hdfs dfs -ls file:///usr/local`

## **HDFS:**

`hdfs dfs -ls hdfs://localhost:9000/user/home`

# Основные команды shell: -ls

hdfs dfs -ls file:///usr/local

```
hduser@vasya-Virtual-Machine:~$ hdfs dfs -ls file:///usr/local
Found 10 items
drwxr-xr-x - root root 4096 2020-07-31 19:27 file:///usr/local/bin
drwxr-xr-x - root root 4096 2020-07-31 19:27 file:///usr/local/etc
drwxr-xr-x - root root 4096 2020-07-31 19:27 file:///usr/local/games
drwxr-xr-x - hduser hadoop 4096 2020-11-26 16:30 file:///usr/local/hadoop
drwxr-xr-x - root root 4096 2020-07-31 19:27 file:///usr/local/include
drwxr-xr-x - root root 4096 2020-07-31 19:27 file:///usr/local/lib
drwxr-xr-x - root root 4096 2020-07-31 19:27 file:///usr/local/man
drwxr-xr-x - root root 4096 2020-07-31 19:27 file:///usr/local/sbin
drwxr-xr-x - root root 4096 2020-07-31 19:31 file:///usr/local/share
drwxr-xr-x - root root 4096 2020-07-31 19:27 file:///usr/local/src
hduser@vasya-Virtual-Machine:~$
```

ls -- самая популярная команда, выводит листинг директории и статистику файлов. Если запустим команду для определенной директории, то увидим ее содержимое:

- сколько элементов в данной директории items
- файлы, которые содержатся в директории
- статистика -- указание прав доступа и типа файла (поддиректории), фактор репликации (у директории -- нет), принадлежность файла (пользователь и группа), время создания.
- Параметры команды:
  - r -- вывести статистику по поддиректориям

## Основные команды shell: -mkdir

hdfs dfs -mkdir file:///usr/local/hadoop/hadoop\_temp

hdfs dfs -mkdir hdfs://localhost:9000/data

hdfs dfs -ls hdfs://localhost:9000/data/

```
hduser@vasya-Virtual-Machine:~$ hdfs dfs -ls hdfs://localhost:9000/  
Found 1 items  
drwxr-xr-x  - hduser supergroup          0 2020-11-26 21:07 hdfs://localhost:9000/data
```

# Основные команды shell: -cat -text

## Команда чтения файлов cat

Выводит источник в стандартный поток вывода -- stdout.

- вывести весь файл hdfs dfs -cat /dir/file.txt -- для больших файлов очень долго и не понятно для чего
- перенаправление потока вывода через каналы pipe в утилиты less, head, tail
- получить первые 100 строк из файла  
hdfs dfs -cat /dir/file.txt | head -n 100

## Команда чтения файлов text

Аналог команды cat, но разархивирует архивы:

- hdfs dfs -cat /dir/file.gz -- непонятный текст
- hdfs dfs -text /dir/file.gz -- понятный текст
- Если файл не заархивирован, то работа команд text и cat будет одинакова.

## Основные команды shell: -tail

Выводит последние строки файла:

`hdfs dfs -cat /dir/file.txt | tail --` плохо

`hdfs dfs -tail /dir/file.txt` — хорошо

В первом случае читается весь файл и только потом от него отрезаются последние строки для вывода. Для большого файла может потребоваться на чтение много времени -- прочитать все блоки файла со всех серверов. Только с последнего блока вывести последние строки. Во втором случае нет лишних действий, будут выведены последние строки из последнего блока.

Полезна для просмотра, например, логов.

## Основные команды shell: Копирование данных

`cp` -- копирование файла из одного места в другое. Годится только для небольших файлов!

```
hdfs dfs -cp /dir/file1 /otherDir/file2
```

Копирует файл из одной директории в другую. Т.к. файл состоит из блоков, то каждый блок копируется на новое место. Если файл большой и состоит из многих блоков, то копирование такого файла будет выполняться долго. Разумно копировать небольшие файлы, иначе для больших файлов потребуется много времени.

`distcp` -- копирование больших файлов или множества файлов.

```
hdfs distcp /dir/file1 /otherDir/file2
```

это команда не часть команды `dfs`. При использовании команды `distcp` блоки файла копируются параллельно и независимо друг от друга. Файл копируется быстро. Для больших файлов.



## Основные команды shell: Копирование данных

`mv` -- перемещение файла из одного места в другое

```
hdfs dfs -mv/dir/file1 /otherDir
```

Физически перемещение данных не происходит. Указание Namenode, что файл теперь будет храниться в другой директории.

`put`, `copyFromLocal` -- копирование локального файла в директорию HDFS

```
hdfs dfs -put localfile /dir/file
```

`get`, `copyToLocal` -- копирование файла из HDFS в локальную файловую систему

```
hdfs dfs -get /dir/file localfile
```

## Основные команды shell: -rm

`rm` -- удаляет файл в корзину. При случайном удалении большого файла его очень трудно восстановить. Поэтому файл сначала помещается в корзину. Корзина -- специальная для каждого пользователя директория. Файлы из нее можно удалить или скопировать на прежнее место. Автоматическая очистка корзины задана в параметрах HDFS. Если нужно освободить место в HDFS, то можно удалить файл из корзины.

`hdfs dfs -rm /dir/file` -- удалить файл в корзину

`hdfs dfs -rm -skipTrash /dir/file` -- удалить файл без помещения в корзину

`hdfs dfs -rm -r /dir` -- удалить рекурсивно всю директорию (поддиректории)

## Команда вывода статистики в shell

`du` -- размер файла или директории в байтах `du -h` -- размер файла или директории в удобно-читаемом формате

```
hdfs dfs -du -h /dir
```

В HDFS файлы большие, поэтому удобнее использовать ключ `-h`.

## Права доступа в HDFS

- ограничения на уровне файла/директории -- сходство с моделью прав в POSIX -- Read (r), Write (w), Execute (x) -- Разделяется на пользователя, группу и всех остальных
- Права пользователя определяются исходя из прав той ОС, где он запускает клиентское приложение.
- У файлов есть владелец и группа, которой он принадлежит.

**Спасибо за  
внимание!**

