HDFS

План семинара

1. Обзор HDFS
2. Устройство Hadoop кластера АТП
3. Интерфейс командной строки
4. API HDFS
5. GUI HDFS

Общие особенности HDFS

HDFS = Hadoop **Distributed** File System

- **Ненадёжные** машины, ненадёжная сеть

Hadoop — **экосистема** приложений (технологический стек)

HDFS — «традиционная» **файловая система** стека

- **Единая точка отказа** (namenode)
  - Hot standby нет; есть «прохладный» backup
- Репликация:
  - Single master («единственный ведущий»)
  - Синхронная или асинхронная

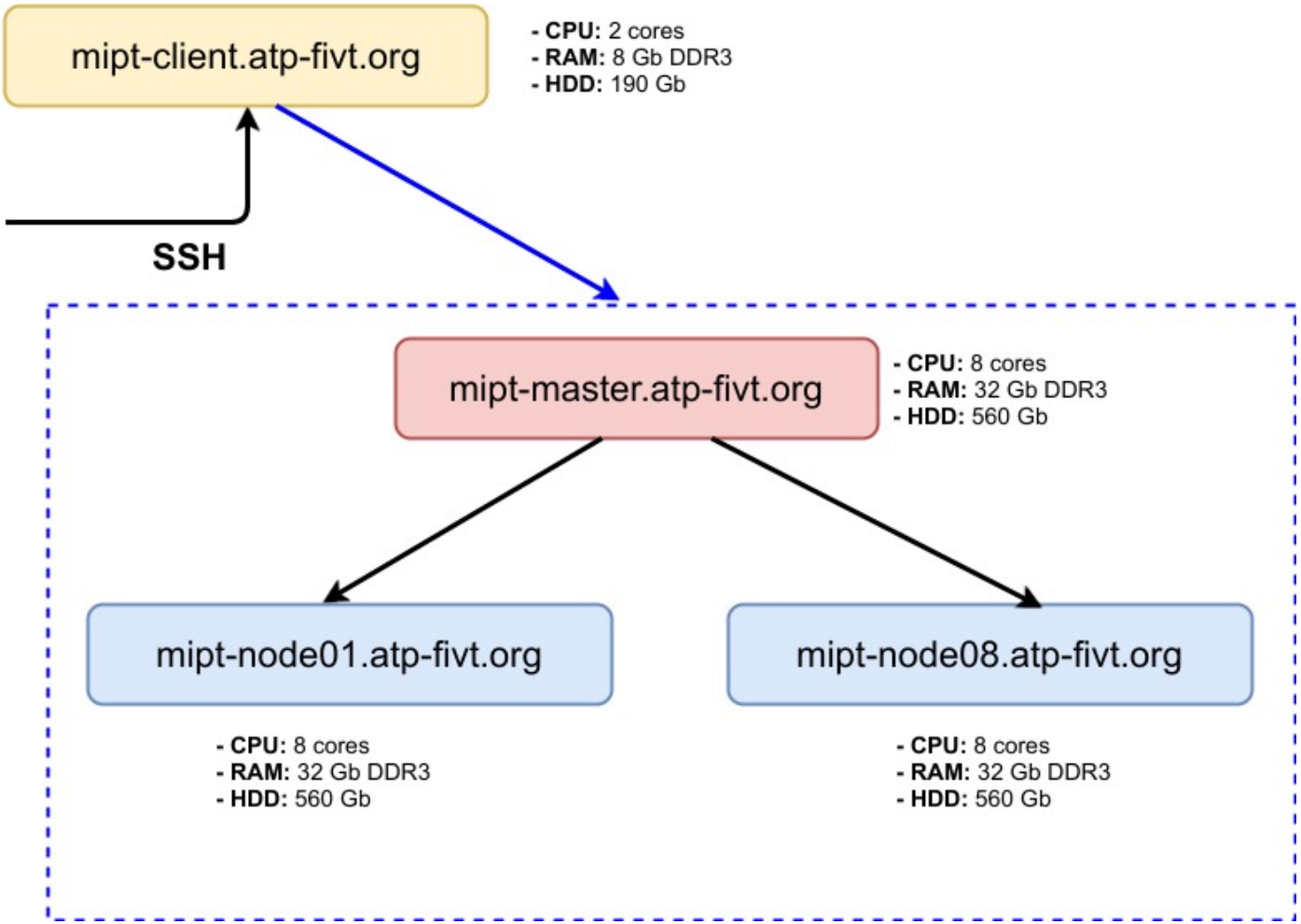
Транзакционность отсутствует

У нескольких vendors есть свои решения всех проблем

- Но рынок уменьшается

CLOUDERA

Кластер АТП




```
ssh USER@mipt-client.atp-fivt.org
```

Есть авторизация по публичному ключу

Кластер АТП

Зайти на datanode (XX = 01..08)

```
sudo -u hdfsuser ssh hdfsuser@mipt-nodeXX.atp-fivt.org
```

SSH proxy (доступ к IP-порту удалённой машины)

Делает PORT кластера доступным как localhost:LPORT

```
ssh -fN -L LPORT:mipt-master.atp-fivt.org:PORT USER@mipt-client.atp-fivt.org
```

HTTP-запросы удобно выполнять с помощью

```
curl
```

Команды HDFS DFS

hdfs dfs	Список команд
hdfs dfs -help [команда]	Справка

Команды передаются как первый флаг. Например, **hdfs dfs** -ls /

Некоторые доступные команды:

- ls
- mv
- cp
- rm
- cat
- tail
- mkdir
- du
- put — Поместить локальный файл в HDFS
- setrep — Изменить фактор репликации
- …

HDFS API

Java API. Наиболее полный и производительный
https://gitlab.com/pd2020-supplementary/8xx-GLOBAL/-/tree/master/practice/code/hdfs_example

HTTP REST API

```
curl -i "http://mipt-master.atp-fivt.org:50070/webhdfs/v1/data/wiki/en_articles_part/articles-part?op=OPEN"
```

Документация: http://hadoop.apache.org/docs/r1.2.1/webhdfs.html

Python API: hdfscli
https://gitlab.com/pd2020-supplementary/8xx-GLOBAL/-/blob/master/practice/07-hdfs.md#python-api

Документация:
https://hdfscli.readthedocs.io/en/latest/quickstart.html#reading-and-writing-files

GUI для доступа к HDFS

Hadoop web GUI: http://mipt-master.atp-fivt.org:50070

Apache HUE: `http://mipt-node03:8888`
`hue_user` / `hue_userpd`
(см. ранее, как сделать proxy)

Полезные материалы

Обзор архитектуры HDFS
https://hadoop.apache.org/docs/r2.5.2/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html

Hadoop Java API
https://hadoop.apache.org/docs/r2.5.2/api/index.html

HDFS command line
https://hadoop.apache.org/docs/r2.5.2/hadoop-project-dist/hadoop-common/FileSystemShell.html

HdfsCLI (Python API)
https://hdfscli.readthedocs.io/en/latest/