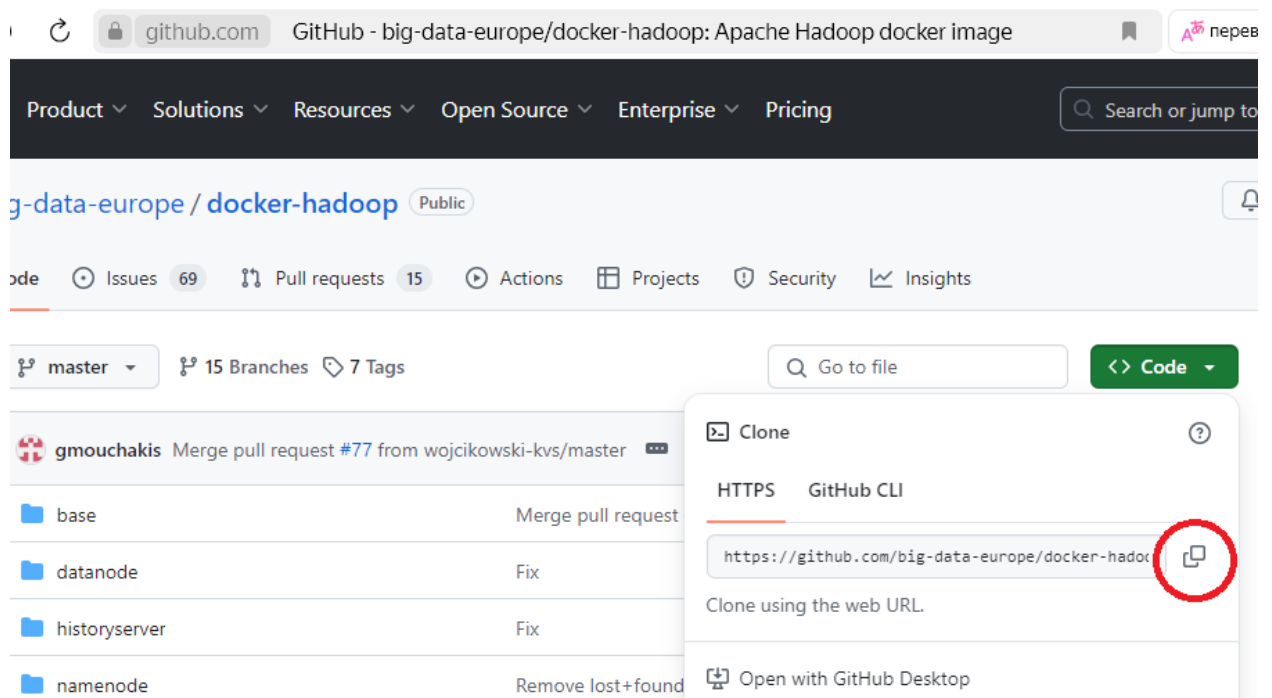
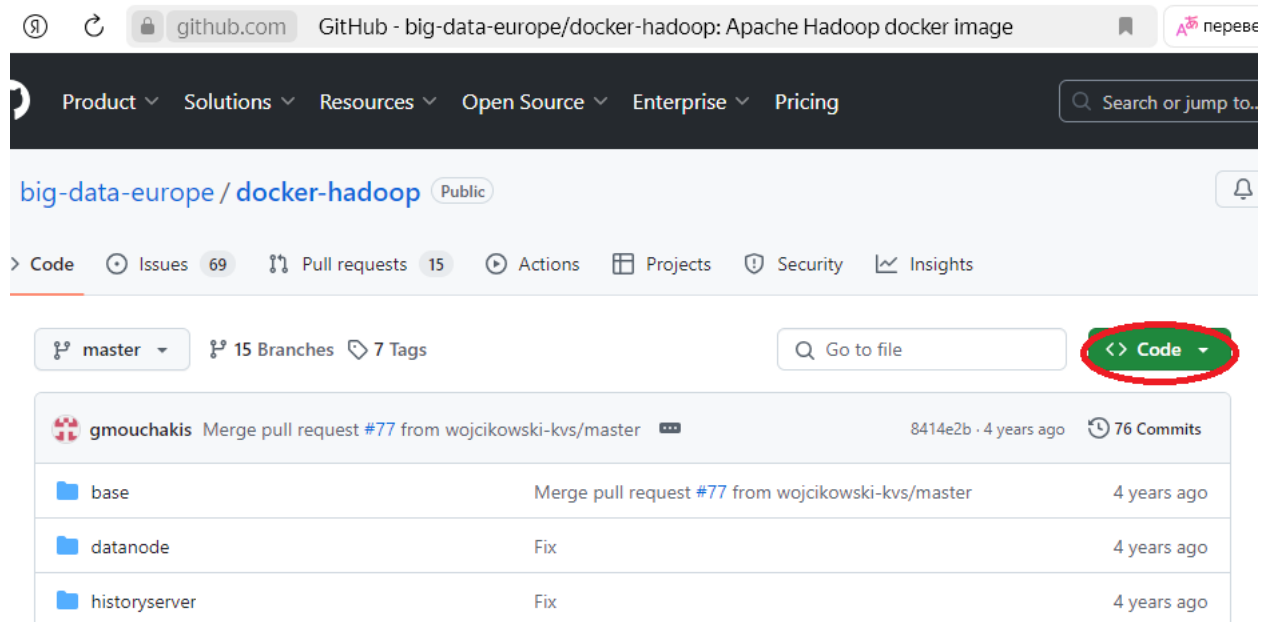


## Установка Hadoop в Ubuntu

Разворачивать Hadoop и др. будем в контейнерах используя docker-compose.

<https://github.com/WtCrow/docker-hadoop-spark-hive2-jupyter>



Клонируем репозиторий на Ubuntu server:

```
user@hadoop-server:~$ git clone https://github.com/WtCrow/docker-hadoop-spark-hive2-jupyter.git
Cloning into 'docker-hadoop'...
remote: Enumerating objects: 539, done.
remote: Counting objects: 100% (189/189), done.
remote: Compressing objects: 100% (23/23), done.
remote: Total 539 (delta 169), reused 166 (delta 166), pack-reused 350 (from 1)
Receiving objects: 100% (539/539), 108.00 KiB | 349.00 KiB/s, done.
Resolving deltas: 100% (251/251), done.
user@hadoop-server:~$
```

Переходим в скачанную папку:

```
user@hadoop-server:~$ cd docker-hadoop-spark-hive2-jupyter/
user@hadoop-server:~/docker-hadoop-spark-hive2-jupyter$ ls
docker-compose.yml example.ipynb hadoop-hive.env README.md spark_conf
user@hadoop-server:~/docker-hadoop-spark-hive2-jupyter$
```

В папке находятся директории для каждого компонента Hadoop и **docker-compose.yml** – файл настройки многоконтейнерного приложения, настроим его открыв в редакторе.

Настройки нейм-ноды не трогаем:

```
GNU nano 6.2 docker-compose.yml
version: "3"

services:
  namenode:
    image: bde2020/hadoop-namenode:2.0.0-hadoop3.2.1-java8
    container_name: namenode
    restart: always
    ports:
      - 9870:9870
      - 9000:9000
    volumes:
      - hadoop_namenode:/hadoop/dfs/name
    environment:
      - CLUSTER_NAME=test
    env_file:
      - ./hadoop.env
```

Дата-нода по умолчанию одна,

```
GNU nano 6.2                                docker-compose.yml
  env_file:
    - ./hadoop.env

  datanode:
    image: bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8
    container_name: datanode
    restart: always
    volumes:
      - hadoop_datanode:/hadoop/dfs/data
    environment:
      SERVICE_PRECONDITION: "namenode:9870"
    env_file:
      - ./hadoop.env

  resourcemanager:
    image: bde2020/hadoop-resourcemanager:2.0.0-hadoop3.2.1-java8
    container_name: resourcemanager
    restart: always
```

добавим еще две и дадим им уникальные имена и назначим свои volume:

```
  datanode1:
    image: bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8
    container_name: datanode1
    restart: always
    volumes:
      - hadoop_datanode1:/hadoop/dfs/data
    environment:
      SERVICE_PRECONDITION: "namenode:9870"
    env_file:
      - ./hadoop.env
```

```
  datanode2:
    image: bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8
    container_name: datanode2
    restart: always
    volumes:
      - hadoop_datanode2:/hadoop/dfs/data
    environment:
      SERVICE_PRECONDITION: "namenode:9870"
    env_file:
      - ./hadoop.env

  datanode3:
    image: bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8
    container_name: datanode3
    restart: always
    volumes:
      - hadoop_datanode3:/hadoop/dfs/data
    environment:
      SERVICE_PRECONDITION: "namenode:9870"
    env_file:
      - ./hadoop.env
```

И добавим в конце файла наши volume:

```

- ./hadoop.env
volumes:
  hadoop namenode:
  hadoop_datanode1:
  hadoop_datanode2:
  hadoop_datanode3:
  hadoop_historyserver:

```

Сохраняем изменения в файле и запускаем, предварительно установив docker и docker-compose:

```

user@hadoop-server:~/docker-hadoop$ sudo snap install docker
[sudo] password for user:
docker 24.0.5 from Canonical✓ installed
user@hadoop-server:~/docker-hadoop$

```

```

user@hadoop-server:~/docker-hadoop$ sudo apt install docker-compose
Чтение списков пакетов... Готово
Построение дерева зависимостей... Готово
Чтение информации о состоянии... Готово
Будут установлены следующие дополнительные пакеты:
  bridge-utils containerd dns-root-data dnsmasq-base docker.io pigz python3-docker
  python3-dockerpty python3-docopt python3-dotenv python3-texttable python3-websocket r
  ubuntu-fan
Предлагаемые пакеты:
  ifupdown aufs-tools cgroupfs-mount | cgroup-lite debootstrap docker-doc rinse zfs-fus
  | zfsutils
Следующие НОВЫЕ пакеты будут установлены:
  bridge-utils containerd dns-root-data dnsmasq-base docker-compose docker.io pigz
  python3-docker python3-dockerpty python3-docopt python3-dotenv python3-texttable
  python3-websocket runc ubuntu-fan
Обновлено 0 пакетов, установлено 15 новых пакетов, для удаления отмечено 0 пакетов, и 4
е обновлено.
Необходимо скачать 75,8 МБ архивов.
После данной операции объём занятого дискового пространства возрастёт на 286 МБ.
Хотите продолжить? [Д/н]

```

```

user@hadoop-server:~/docker-hadoop$ sudo docker-compose up -d
[+] Running 15/15
✔ Network docker-hadoop_default Created
✔ Volume "docker-hadoop_hadoop_datanode3" Created
✔ Volume "docker-hadoop_hadoop_datanode1" Created
✔ Volume "docker-hadoop_hadoop_datanode2" Created
✔ Container datanode2 Started
✔ Container datanode1 Started
✔ Container resourcemanager Started
✔ Container historyserver Started
✔ Container datanode3 Started
✔ Container nodemanager Started
✔ Container spark-master Started
✔ Container namenode Started
✔ Container spark-worker Started
✔ Container jupyter-notebook Started
✔ Container hue Started

```

Дожидаемся запуска всех контейнеров и открываем в браузере страницу админки Hadoop:

```
user@hadoop-server:~$ sudo docker ps
CONTAINER ID   IMAGE                                     COMMAND                  CREATED        STATUS              PORTS
0ce7e3b41768   bde2020/hadoop-resourcemanager:2.0.0-hadoop3.2.1-java8   "/entrypoint.sh /run..." 19 minutes ago   Up About a minute (unhealthy) 8088/tcp
88888ffcdf39   bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8          "/entrypoint.sh /run..." 19 minutes ago   Up 19 minutes (healthy)      9864/tcp
0d13ba27c834   bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8          "/entrypoint.sh /run..." 19 minutes ago   Up 19 minutes (healthy)      9864/tcp
6896432c9d0a   bde2020/hadoop-nodemanager:2.0.0-hadoop3.2.1-java8       "/entrypoint.sh /run..." 19 minutes ago   Up About a minute (unhealthy) 8042/tcp
a26495e130b3   bde2020/hadoop-historyserver:2.0.0-hadoop3.2.1-java8     "/entrypoint.sh /run..." 19 minutes ago   Up About a minute (unhealthy) 8188/tcp
50dd81a149a8   bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8          "/entrypoint.sh /run..." 19 minutes ago   Up 19 minutes (healthy)      9864/tcp
741bd87e49c4   bde2020/hadoop-namenode:2.0.0-hadoop3.2.1-java8          "/entrypoint.sh /run..." 19 minutes ago   Up 19 minutes (healthy)      0.0.0.0:9000->9000/tcp, ::9000->9000/tcp, 0.0.0.0:9870->9870/tcp, ::9870->9870/tcp
user@hadoop-server:~$
```

← ⓘ ↺ 🔴 10.10.10.121:50070

Namenode information

🔖

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities ▾

# Overview 'namenode:8020' (active)

Started:	Tue Oct 29 04:59:09 UTC 2024
Version:	2.7.4, rcd915e1e8d9d0131462a0b7301586c175728a282
Compiled:	2017-08-01T00:29Z by kshvachk from branch-2.7.4
Cluster ID:	CID-74d2a272-2c9f-46fb-a6eb-0e73653f72d5
Block Pool ID:	BP-460939484-172.18.0.3-1730104026474

# Summary

Security is off.

Safemode is off.

28 files and directories, 3 blocks = 31 total filesystem object(s).

# Summary

Security is off.

Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 27.81 MB of 44.87 MB Heap Memory. Max Heap Memory is 475.63 MB.

Non Heap Memory used 44.48 MB of 45.59 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

объем кластера

Configured Capacity:	55.59 GB
Configured Remote Capacity:	0 B
DFS Used:	72 KB (0%)
Non DFS Used:	20.96 GB
DFS Remaining:	31.73 GB (57.08%)
Block Pool Used:	72 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	3 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)

"живые" ноды

Во вкладке Datanodes, можно посмотреть на ноды:

192.168.1.178:9870

Namenode information

90%

перевести

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

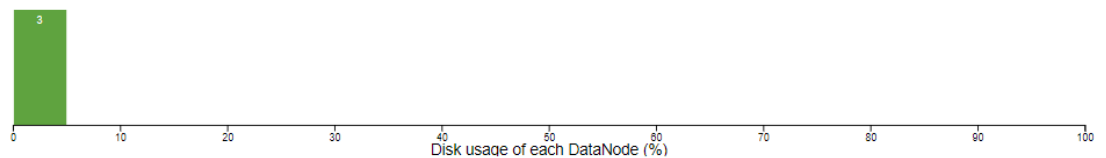
Startup Progress

Utilities

## Datanode Information

✓ In service    ⬇ Down    ⌚ Decommissioning    ⚡ Decommissioned    ⚡ Decommissioned & d  
➡ Entering Maintenance    ⚡ In Maintenance    ⚡ In Maintenance & d

### Datanode usage histogram



192.168.1.178:9870 Namenode information

In operation

Show 25 entries Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✓ 0d13ba27c834:9866 (172.18.0.3:9866)	http://0d13ba27c834:9864	1s	10m	18.53 GB	0	28 KB (0%)	3.2.1
✓ 38888ffcd39:9866 (172.18.0.8:9866)	http://38888ffcd39:9864	1s	10m	18.53 GB	0	28 KB (0%)	3.2.1
✓ 50dd81a149a8:9866 (172.18.0.4:9866)	http://50dd81a149a8:9864	1s	10m	18.53 GB	0	24 KB (0%)	3.2.1

Showing 1 to 3 of 3 entries

объем каждой ноды

пока ничего не хранится

Previous 1 Next

Entering Maintenance

No nodes are entering maintenance.

Теперь загрузим данные, например, откроем в браузере <https://openflights.org/data>

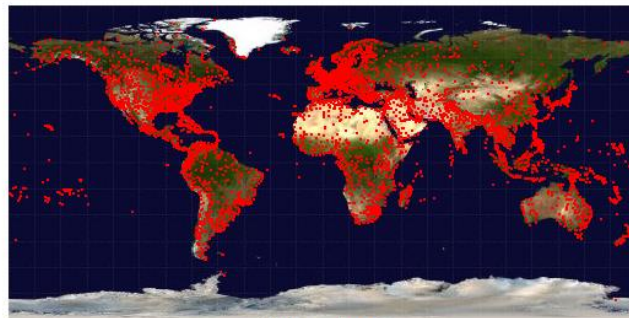
openflights.org OpenFlights: Airport and airline data

пересказать перевести

## Airport, airline and route data

Navigation: [Airport](#) | [Airline](#) | [Route](#) | [Plane](#) | [Country](#) | [Schedule](#) | [Other](#) | [License](#)

### Airport database



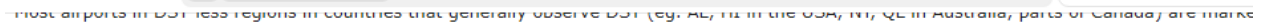
(click to enlarge)

As of January 2017, the OpenFlights Airports Database contains **over 10,000** airports, train stations and ferry terminals spanning the globe, as shown in the above. Each entry contains the following information:

**Airport ID** Unique OpenFlights identifier for this airport.  
**Name** Name of airport. May or may not contain the **City** name.  
**City** Main city served by airport. May be spelled differently from **Name**.

Копируем ссылки:





### Sample entries

507, "London Heathrow Airport", "London", "United kingdom", "LHR", "EGLL", 51.4706, -0.461941, 83, 0, "E", "Europe/London", "airport"  
26, "Kugaaruk Airport", "Pelly Bay", "Canada", "YBB", "CYBB", 68.534401, -89.808098, 56, -7, "A", "America/Edmonton", "airport", "OurA  
3127, "Pokhara Airport", "Pokhara", "Nepal", "PKR", "VNPK", 28.200899124145508, 83.98210144042969, 2712, 5.75, "N", "Asia/Katmandu",  
8810, "Hamburg Hbf", "Hamburg", "Germany", "ZMB", "\N, 53.552776, 10.006683, 30, 1, "E", "Europe/Berlin", "station", "User"

Try it out: [Airport Search](#) (new window)

*Note:* The Airport Search window above is a part of [OpenFlights](#). You will not be able to add or edit airports unless you are logged in.

**Download**

To download the current data dump from [GitHub](#) as a very straightforward CSV (comma-separated value) file, suitable for use in spreadsheets e

Download: [airports.dat](#) (Airports only, high quality)

Download: [airports-extended.dat](#) (Airports, train stations and ferry terminals, including user co

Creating and maintaining this database has required and continues to require an *immense amount* of work. We need your support to keep this database up-to-date: just click on the PayPal link to the right (Visa, MasterCard, American Express and Discover also accepted). We suggest **US\$50**, but any amount at all is welcome, and you may use the data for free if you feel that you are unable to pay. If you do donate, please specify in the comments if you would like a itemized receipt for business expense or tax purposes.

Support O  
Donation  
Buy N  
MasterCard VISA

The GitHub copy is only a sporadically updated static snapshot of the live OpenFlights database (see [revision log](#)). If you would like an up-to-the

И скачиваем на сервер:

```
user@hadoop-server:~$ wget https://raw.githubusercontent.com/jpatokal/openflights/master/data/airports.dat
--2024-10-25 04:56:00-- https://raw.githubusercontent.com/jpatokal/openflights/master/data/airports.dat
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.111.133, 185.199.108.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.111.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1127225 (1,1M) [text/plain]
Saving to: 'airports.dat'

airports.dat      100%[=====>]      1,07M  2,51MB/s      in 0,4s

2024-10-25 04:56:01 (2,51 MB/s) - 'airports.dat' saved [1127225/1127225]

user@hadoop-server:~$
```

```
user@hadoop-server:~$ wget https://raw.githubusercontent.com/jpatokal/openflights/master/data/airports-extended.dat
--2024-10-25 04:56:49-- https://raw.githubusercontent.com/jpatokal/openflights/master/data/airports-extended.dat
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.111.133, 185.199.108.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.111.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1670162 (1,6M) [text/plain]
Saving to: 'airports-extended.dat'

airports-extended.dat 100%[=====>] 1,59M 2,93MB/s in 0,5s

2024-10-25 04:56:50 (2,93 MB/s) - 'airports-extended.dat' saved [1670162/1670162]

user@hadoop-server:~$
```

Затем копируем файлы в контейнер:



```

user@hadoop-server:~$ sudo docker cp airports.dat namenode:/
Successfully copied 1.13MB to namenode:/
user@hadoop-server:~$ sudo docker cp airports-extended.dat namenode:/
Successfully copied 1.67MB to namenode:/
user@hadoop-server:~$

```

Проверяем, файлы внутри контейнера:

```

user@hadoop-server:~$ sudo docker exec -it namenode /bin/bash
root@741bd87e49c4:/# ls
KEYS                boot                hadoop              lib64               proc               sbin               usr
airports-extended.dat dev                  hadoop-data         media               root               srv               var
airports.dat         entrypoint.sh       home                mnt                 run                 sys
bin                  etc                  lib                  opt                  run.sh             tmp
root@741bd87e49c4:/#

```

Затем кладем файлы в HDFS:

```

root@741bd87e49c4:/# hdfs dfs -put airports.dat /
2024-10-25 05:07:41,064 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@741bd87e49c4:/# hdfs dfs -put airports-extended.dat /
2024-10-25 05:09:10,403 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
root@741bd87e49c4:/#

```

Проверяем в админке:

The screenshot shows the Hadoop Admin interface for browsing HDFS. The 'Utilities' menu is open, showing options like 'Browse the file system', 'Logs', 'Log Level', 'Metrics', 'Configuration', and 'Process Thread Dump'. The 'Browse Directory' page displays a table of files in the root directory. The 'Replication' column is highlighted, showing a value of 3 for both files. The files are 'airports-extended.dat' (1.59 MB) and 'airports.dat' (1.08 MB), both owned by 'root' and 'supergroup'.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	root	supergroup	1.59 MB	Oct 25 10:09	3	128 MB	airports-extended.dat
-rw-r--r--	root	supergroup	1.08 MB	Oct 25 10:07	3	128 MB	airports.dat

Фактор репликации 3, значит файл должен быть на всех трех нодах:

## In operation

Show  entries

Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✓0d13ba27c834:9866 (172.18.0.3:9866)	<a href="http://0d13ba27c834:9864">http://0d13ba27c834:9864</a>	1s	46m	18.53 GB <div><div></div></div>	2	2.72 MB (0.01%)	3.2.1
✓38888ffcdf39:9866 (172.18.0.8:9866)	<a href="http://38888ffcdf39:9864">http://38888ffcdf39:9864</a>	1s	10m	18.53 GB <div><div></div></div>	2	2.72 MB (0.01%)	3.2.1
✓50dd81a149a8:9866 (172.18.0.4:9866)	<a href="http://50dd81a149a8:9864">http://50dd81a149a8:9864</a>	1s	46m	18.53 GB <div><div></div></div>	2	2.73 MB (0.01%)	3.2.1

Можно прочитать содержимое файла:

```
root@741bd87e49c4:/#hdfs dfs -cat /airports.dat
```

```
2591,\N,\N,\N,"airport","OurAirports"
14100,"Ramon Airport","Eilat","Israel","ETM","LLER",29.723694,35.011416,288,\N,\N,
N,"airport","OurAirports"
14101,"Rustaq Airport","Al Masna'ah","Oman","MNH","OORQ",23.640556,57.4875,349,\N,
N,\N,"airport","OurAirports"
14102,"Laguindingan Airport","Cagayan de Oro City","Philippines","CGY","RPMY",8.61
203,124.456496,190,\N,\N,\N,"airport","OurAirports"
14103,"Kostomuksha Airport","Kostomuksha","Russia",\N,"ULPM",64.61799621579999,30.
87000274699997,681,\N,\N,\N,"airport","OurAirports"
14104,"Privolzhskiy Air Base","Astrakhan","Russia",\N,"XRAP",46.396,47.893,-66,\N,
N,\N,"airport","OurAirports"
14105,"Kubinka Air Base","Kubinka","Russia",\N,"UUMB",55.611695,36.650002,614,\N,\
,\N,"airport","OurAirports"
14106,"Rogachyovo Air Base","Belaya","Russia",\N,"ULDA",71.61669921880001,52.47829
1873,272,\N,\N,\N,"airport","OurAirports"
14107,"Ulan-Ude East Airport","Ulan Ude","Russia",\N,"XIUW",51.849998474121094,107
73799896240234,1670,\N,\N,\N,"airport","OurAirports"
14108,"Krechevitsy Air Base","Novgorod","Russia",\N,"ULLK",58.625,31.3850002288818
6,85,\N,\N,\N,"airport","OurAirports"
14109,"Desierto de Atacama Airport","Copiapo","Chile","CPO","SCAT",-27.2611999512,
70.7791976929,670,\N,\N,\N,"airport","OurAirports"
14110,"Melitopol Air Base","Melitopol","Ukraine",\N,"UKDM",46.880001,35.305,0,\N,\
,\N,"airport","OurAirports"
root@741bd87e49c4:/#
```

Установка и настройка распределенной файловой системы завершена.

