

## Задание 1

В этом задании мы начали с настройки Hadoop, hive, derby и hive. Затем мы загрузили данные `bi` из kaggle, доступ к которым можно получить по следующей ссылке

<https://www.kaggle.com/datasets/therohk/india-headlines-news-dataset>. Поскольку название было немного длиннее, мы сократили его, чтобы упростить использование. Мы переименовали набор данных в `india-news-headlines.csv`. Теперь, когда у нас был набор данных, пришло время загрузить его в Hadoop. Мы создали каталог в Hadoop с помощью командной строки, используя `hdfs dfs -mkdir /data`. Это должна была быть наша рабочая директория для доступа к данным и работы с ними. После того как это было успешно сделано (об этом свидетельствуют кластеры на рисунке 1.1), мы загрузили загруженный набор данных в Hadoop (об этом свидетельствует рисунок 1.2).

Переведено с помощью DeepL.com (бесплатная версия)

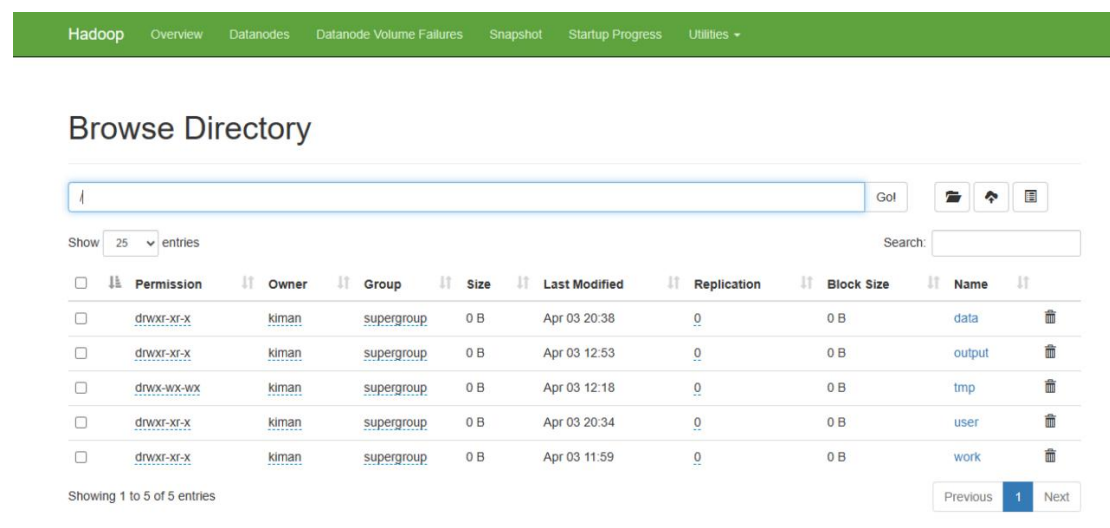


Figure 1.1

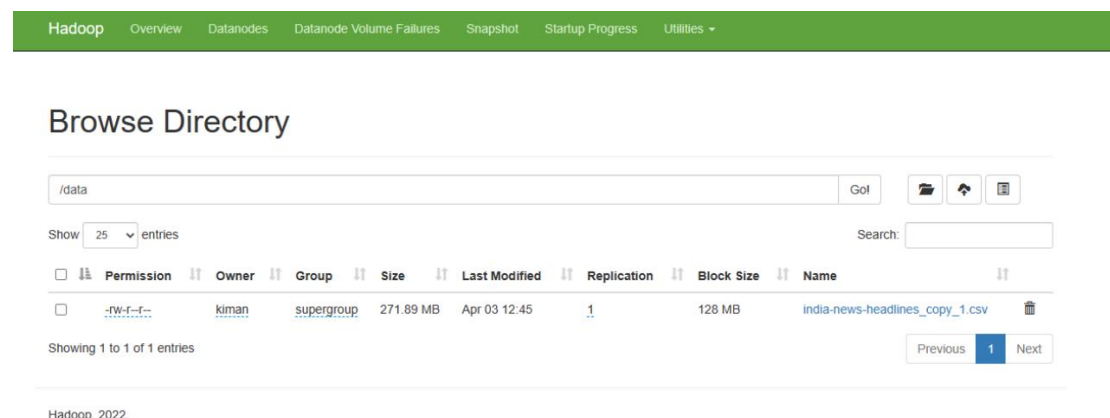


Figure 1.2

В hive мы создали базу данных с именем news с помощью команды CREATE DATABASE news; далее мы создали внешнюю таблицу, указывающую на Hadoop, с помощью следующего SQLscript

```
CREATE EXTERNAL TABLE india_news_headlines (  
    publish_date STRING,  
    headline_category STRING,  
    headline_text STRING  
)
```

```
ROW FORMAT DELIMITED
```

```
FIELDS TERMINATED BY ','
```

```
LOCATION '/data/';
```

Затем мы загрузили данные в таблицу из Hadoop, используя следующее

```
LOAD DATA INPATH '/data/india-news-headlines.csv' INTO TABLE india_news_headlines;
```

Чтобы убедиться, что данные присутствуют;

```
hive> SELECT * FROM india_news_headlines LIMIT 10;  
2024-04-03T20:38:45,766 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value  
2024-04-03T20:38:45,766 INFO [main] org.apache.hadoop.hive ql.session.SessionState - Updating threa  
2024-04-03T20:38:51,960 INFO [f88c01d3-cc30-4e71-9060-2ae5827f6b58 main] org.apache.hadoop.hive.com  
ost:9000/tmp/hive/kiman/f88c01d3-cc30-4e71-9060-2ae5827f6b58/hive_2024-04-03_20-38-45_792_780723776  
23776775533699-1  
2024-04-03T20:38:52,364 INFO [f88c01d3-cc30-4e71-9060-2ae5827f6b58 main] org.apache.hadoop.conf.Con  
se mapreduce.task.ismap  
OK  
publish_date    headline_category    headline_text  
20010102        unknown              "Status quo will not be disturbed at Ayodhya; says Vajpayee"  
20010102        unknown              "Fissures in Hurriyat over Pak visit"  
20010102        unknown              "America's unwanted heading for India?"  
20010102        unknown              "For bigwigs; it is destination Goa"  
20010102        unknown              "Extra buses to clear tourist traffic"  
20010102        unknown              "Dilute the power of transfers; says Riberio"  
20010102        unknown              "Focus shifts to teaching of Hindi"  
20010102        unknown              "IT will become compulsory in schools"  
20010102        unknown              "Move to stop freedom fighters' pension played"  
Time taken: 6.598 seconds, Fetched: 10 row(s)
```

-- Используйте предложение WITH для временных вычислений

```
WITH temp_table AS (  
    SELECT publish_date, headline_category, headline_text  
    FROM india_news_headlines  
    WHERE publish_date BETWEEN '20010102' AND '20010110'
```

)

-- Use GROUP BY to aggregate data

SELECT headline\_category, COUNT(\*) AS num\_headlines

FROM temp\_table

GROUP BY headline\_category

ORDER BY num\_headlines DESC

LIMIT 10; -- Use LIMIT to restrict the number of rows returned

```
2024-04-03 21:06:27,376 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 24.075 sec
MapReduce Total cumulative CPU time: 30 seconds 604 msec
Ended Job = job_1712163883084_0001
Launching Job 2 out of 2
2024-04-03 21:06:30,790 INFO [f88c01d3-cc30-4e71-9060-2ae5827f6b58 main] org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces<number>
2024-04-03 21:06:33,211 INFO [f88c01d3-cc30-4e71-9060-2ae5827f6b58 main] org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
Starting Job = job_1712163883084_0002, Tracking URL = http://DESKTOP-9KESAV4:8088/proxy/application_1712163883084_0002/
Kill Command = C:\hadoop\bin\mapred job -kill job_1712163883084_0002
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-04-03 21:06:50,054 Stage-2 map = 0%, reduce = 0%
2024-04-03 21:06:59,637 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 3.093 sec
2024-04-03 21:07:11,167 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 8.17 sec
MapReduce Total cumulative CPU time: 8 seconds 170 msec
Ended Job = job_1712163883084_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 38.604 sec HDFS Read: 285134463 HDFS Write: 503 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 8.17 sec HDFS Read: 8398 HDFS Write: 362 SUCCESS
Total MapReduce CPU Time Spent: 38 seconds 774 msec
OK
unknown 630
city.patna 5
india 2
entertainment.english.hollywood 2
entertainment.hindi.bollywood 2
city.bengaluru 2
city.delhi 1
edit-page 1
business.india.business 1
Time taken: 114.991 seconds, Fetched: 0 row(s)
2024-04-03 21:07:13,659 INFO [f88c01d3-cc30-4e71-9060-2ae5827f6b58 main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: f88c01d3-cc30-4e71-9060-2ae5827f6b58
2024-04-03 21:07:13,660 INFO [f88c01d3-cc30-4e71-9060-2ae5827f6b58 main] org.apache.hadoop.hive.ql.session.SessionState - Resetting thread name to main
hive>
```

Для измерения производительности мы выполнили следующий запрос

EXPLAIN

WITH temp\_table AS (

SELECT publish\_date, headline\_category, headline\_text

FROM india\_news\_headlines

WHERE publish\_date BETWEEN '20010102' AND '20010110'

)

SELECT headline\_category, COUNT(\*) AS num\_headlines

FROM temp\_table

GROUP BY headline\_category

ORDER BY num\_headlines DESC

LIMIT 10;

```
2024-04-03T21:10:08,382 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Using the default value passed in for log id: f88c01d3-cc30-4e71-9060-2ae5827f6
2024-04-03T21:10:08,382 INFO [main] org.apache.hadoop.hive.q1.session.SessionState - Updating thread name to f88c01d3-cc30-4e71-9060-2ae5827f6b58 main
OK
STAGE DEPENDENCIES:
  Stage-1 is a root stage
  Stage-2 depends on stages: Stage-1
  Stage-0 depends on stages: Stage-2
STAGE PLANS:
  Stage: Stage-1
    Map Reduce
      Map Operator Tree:
        TableScan
          alias: india_news_headlines
          Statistics: Num rows: 5890458 Data size: 2850981888 Basic stats: COMPLETE Column stats: NONE
        Filter Operator
          predicate: publish_date BETWEEN '20010102' AND '20010110' (type: boolean)
          Statistics: Num rows: 654495 Data size: 316775603 Basic stats: COMPLETE Column stats: NONE
        Select Operator
          expressions: headline_category (type: string)
          outputColumnNames: headline_category
          Statistics: Num rows: 654495 Data size: 316775603 Basic stats: COMPLETE Column stats: NONE
        Group By Operator
          aggregations: count()
          keys: headline_category (type: string)
          mode: hash
          outputColumnNames: _col0, _col1
          Statistics: Num rows: 654495 Data size: 316775603 Basic stats: COMPLETE Column stats: NONE
        Reduce Output Operator
          key expressions: _col0 (type: string)
          sort order: +
          Map-reduce partition columns: _col0 (type: string)
          Statistics: Num rows: 654495 Data size: 316775603 Basic stats: COMPLETE Column stats: NONE
          value expressions: _col1 (type: bigint)
```

```
Stage: Stage-0
Fetch Operator
  limit: 10
Processor Tree:
  ListSink

Time taken: 0.434 seconds, Fetched: 79 row(s)
```

Время выполнения запроса составляет 0,434 секунды. Это время включает в себя время, затраченное на планирование, выполнение задания MapReduce и получение результатов.