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)

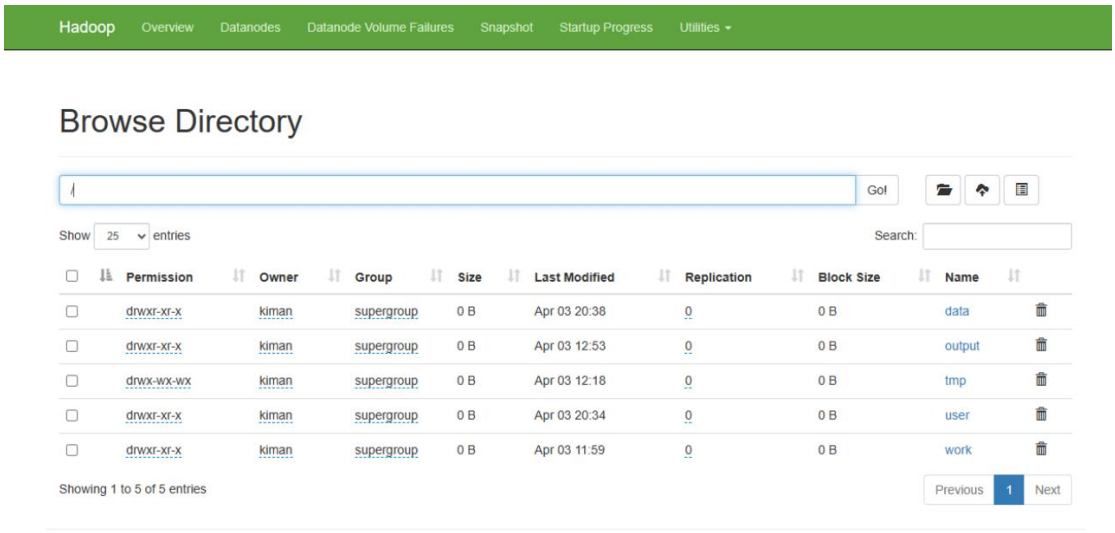


Figure 1.1

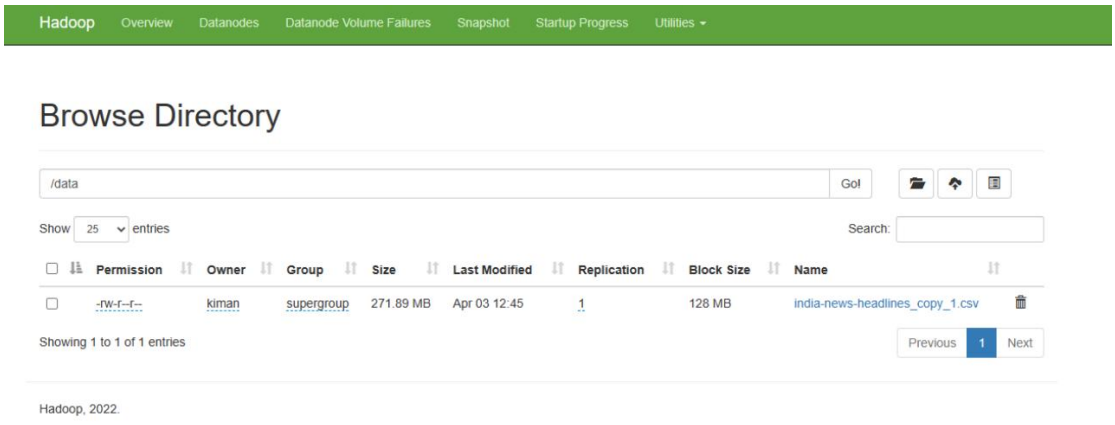


Figure 1.2

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 flayed"
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

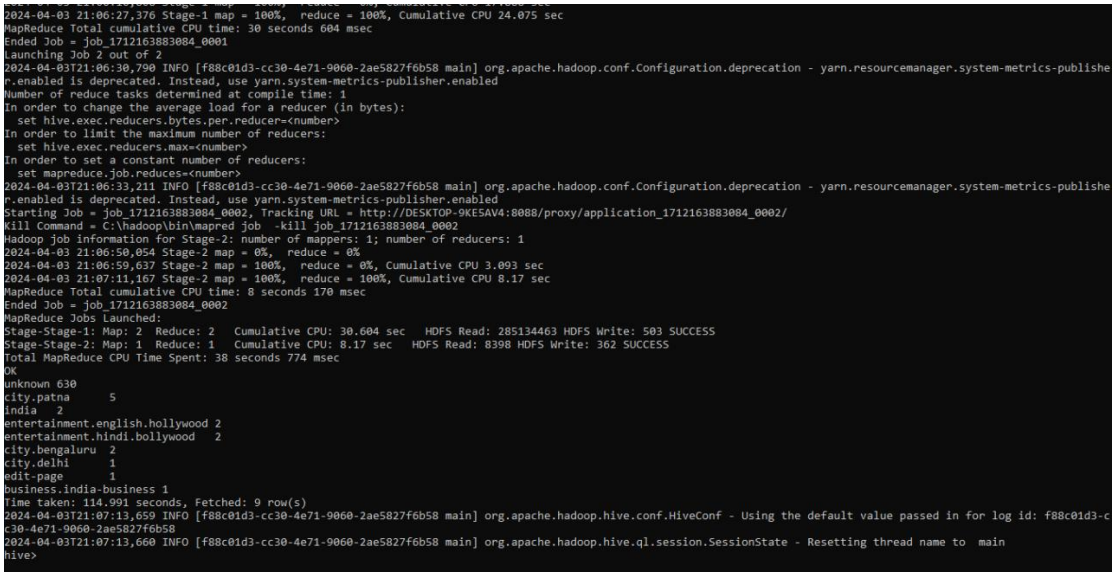
SELECT headline_category, COUNT(*) AS num_headlines

FROM temp_table

GROUP BY headline_category

ORDER BY num_headlines DESC

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



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.ql.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.