|  |
| --- |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение высшего образования **«МИРЭА − Российский технологический университет»**  **РТУ МИРЭА** |

**Институт информационных технологий (ИИТ)**

**Кафедра прикладной математики (ПМ)**

**ОТЧЕТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ**

по дисциплине «Методы анализа данных»

**Практическое занятие № 2**

|  |  |  |  |
| --- | --- | --- | --- |
| Студент группы ИНБО-01-17 | *ИМБО-02-22, Ким Кирилл Сергеевич* | (подпись) | |
| Преподаватель | *Морошкин Никита Андреевич, преподаватель* | (подпись) | |
| Отчет представлен | «\_\_\_»\_\_\_\_\_\_\_\_202\_\_г. | |  | |

Москва 2024 г.

Практическая работа №2

Цель работы

В этой практической работе вам предстоит импортировать и экспортировать таблицы из РСУБД в HDFS (Hadoop Distributed Filesystem), используя Sqoop.

Импортирование таблицы в HDFS с использованием Sqoop.

Начнем с получения информации о базах данных и таблицах в них перед импортированием таблиц в HDFS.

Выполнения практической работы

В терминале авторизуйтесь в MariaDB и выберите базу данных labs. БД: labs. Авторизация прошла успешно, то приглашение для ввода сменится на "MariaDB [labs]>". Ввели команду для проверки того, какие базы данных доступны. (Рисунок 1)

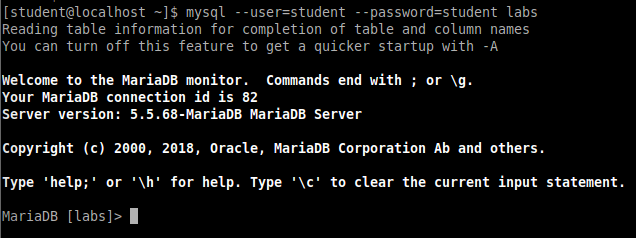


Рисунок — Список баз данных в MariaDB

Затем введите команду для просмотра таблиц в базе labs: (Рисунок 2)

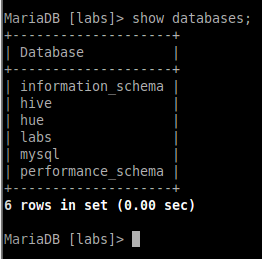


Рисунок — Список названий таблиц в БД labs

На экране отображены таблицы authors и posts. Эти таблицы мы будем импортировать и экспортировать через команды Sqoop показано на Рисунке 3.

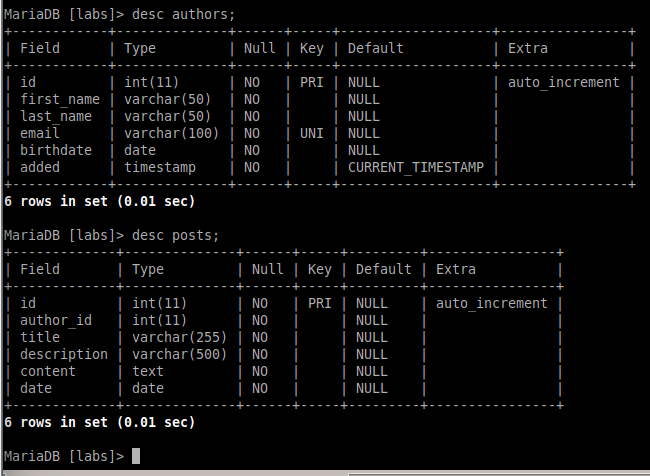


Рисунок — Структура таблиц authors и posts

Посмотрим на структуру таблиц authors и post, а также на несколько записей оттуда (Рисунок 4).

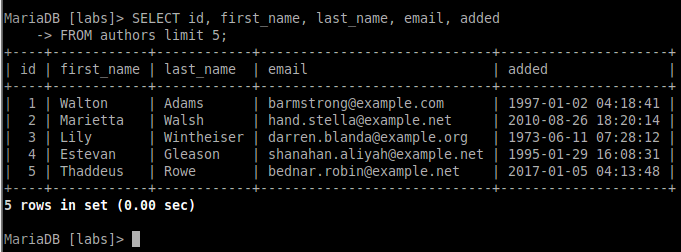


Рисунок — Результат запроса

Для выхода из MariaDB напишите quit и нажмите Enter (Рисунок 5).

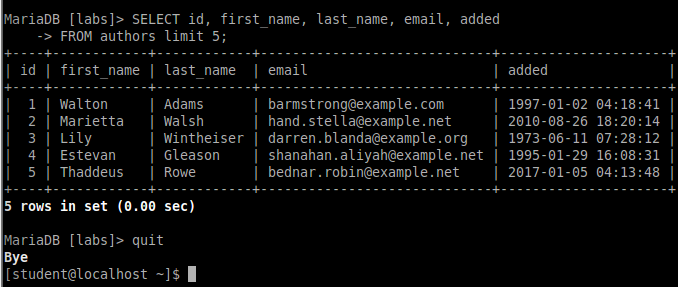


Рисунок — Выход из MariaDB

Для получения помощи по базовым командам sqoop запустите следующую команду (Рисунок 6).

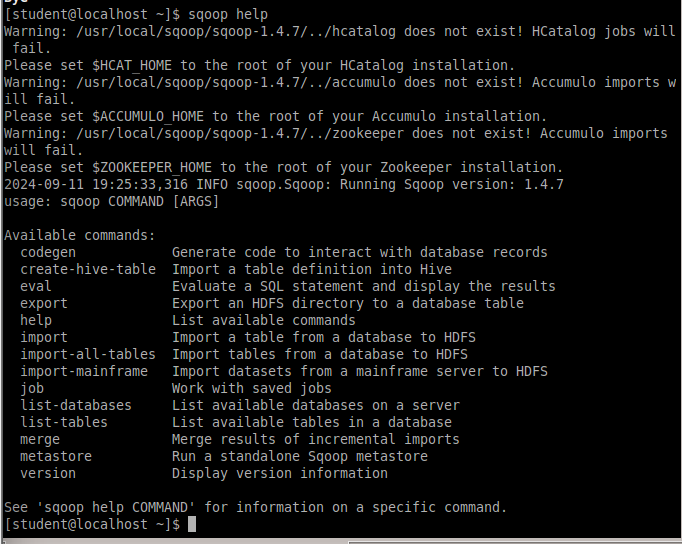


Рисунок 6 — Помощь по базовым командам

Для получения детализированной информации по каждой подкоманде, введите ее название после help. Например, для получения справки по команде import, выполните следующую команду (Рисунок 7):

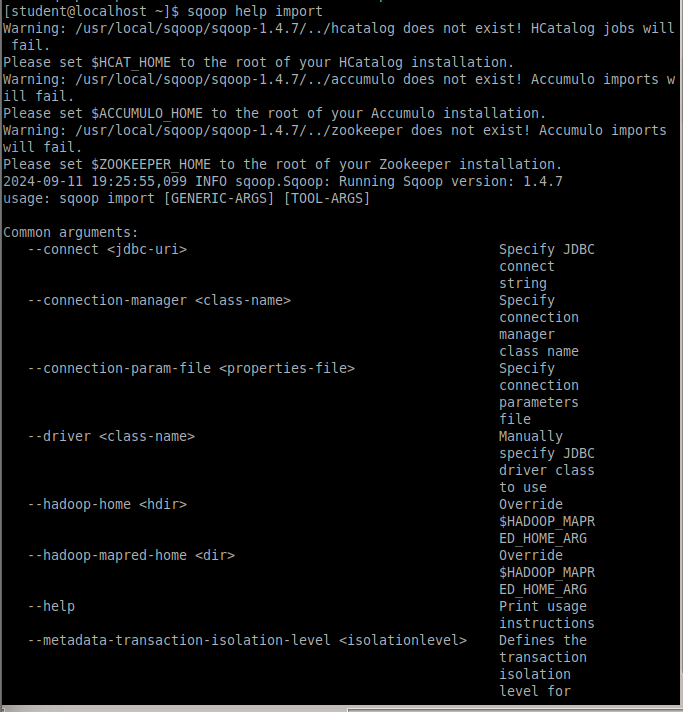


Рисунок 7 — Помощь по команде import

Просмотр списка баз данных в MariaDB и таблиц в базе данных labs осуществляется следующей командой (Рисунок 8):

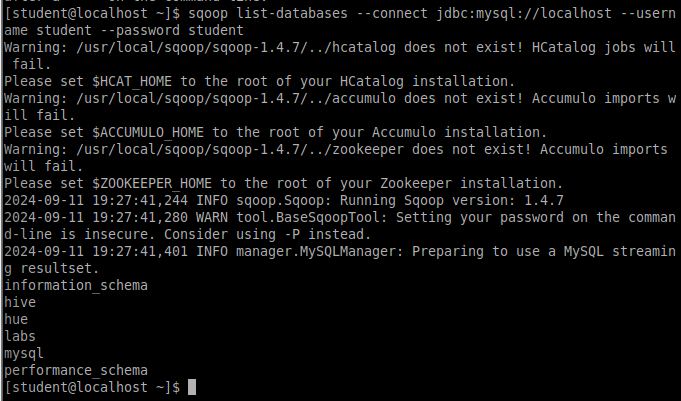


Рисунок 8 — Список баз данных

Импортируем все таблицы в базе labs используя команду import-all-tables (Рисунок 9).

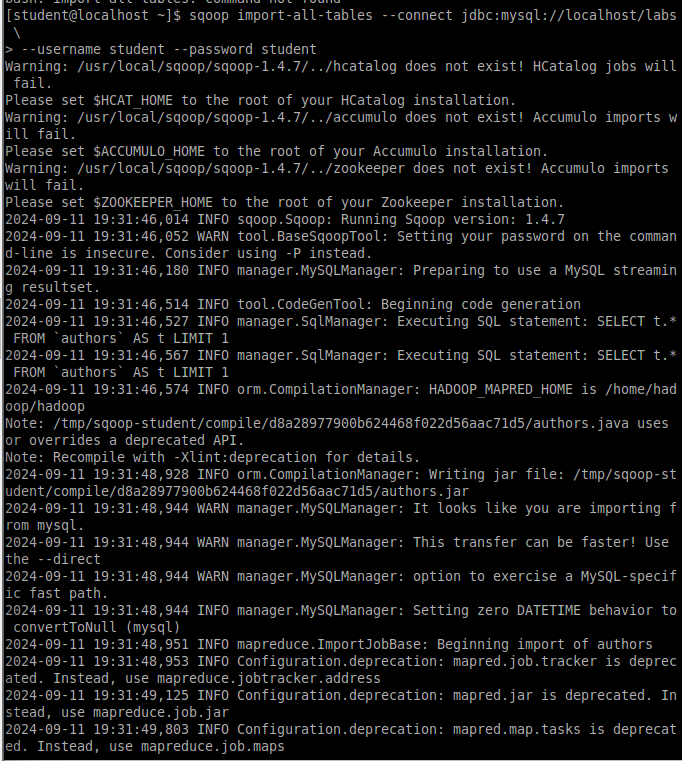


Рисунок 9 — Импорт таблиц

Выполните эту команду для получения таблицы posts из базы labs и сохранения последней в HDFS (Рисунок 10):

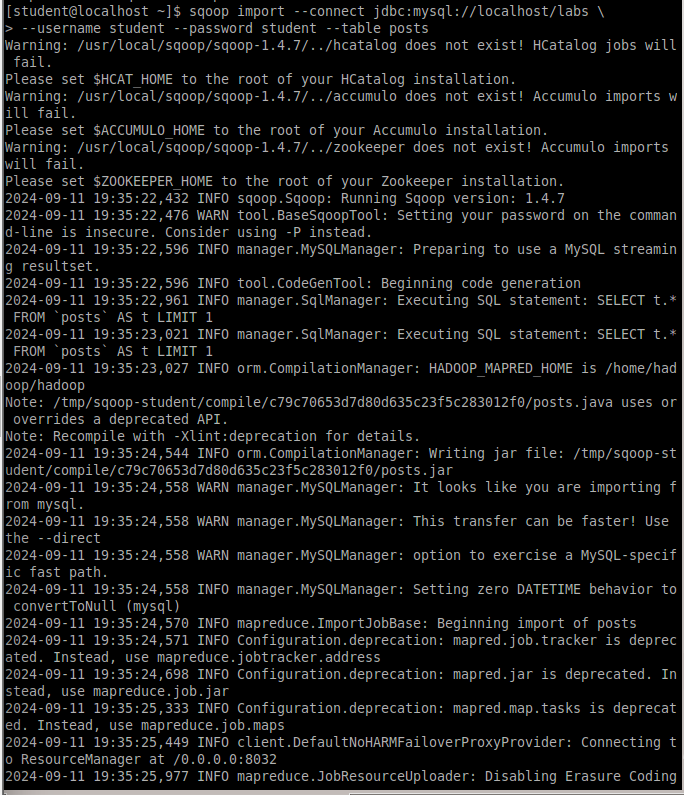
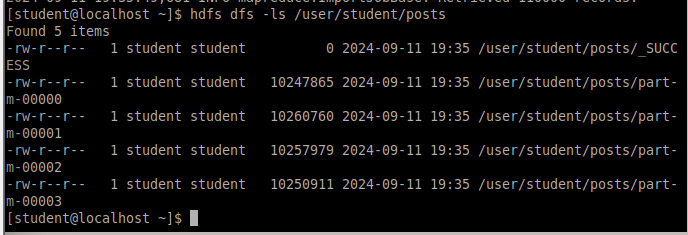


Рисунок 10 — Импорт таблицы posts

После выполнения этой команды создастся директория posts в домашней директории HDFS /user/student, а данные сохранятся как показано ниже (Рисунок 11):



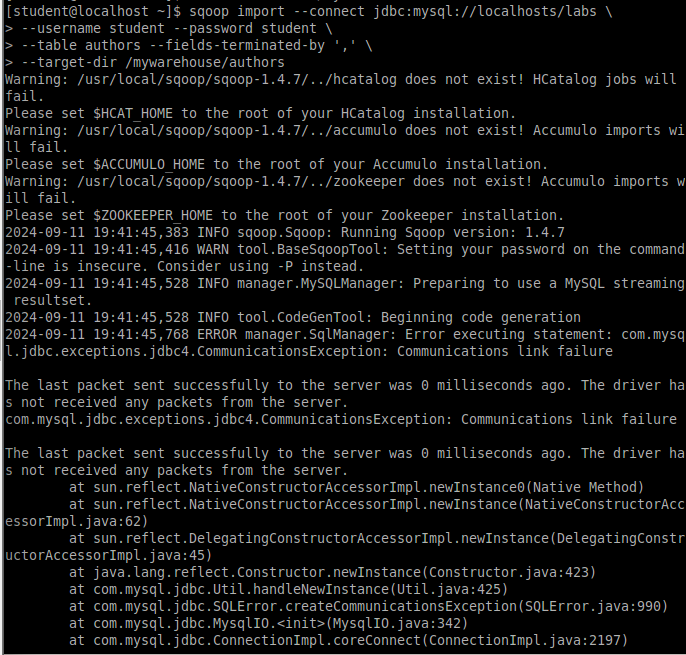
**Рисунок 11** — **Список постов в HDFS**

Создаем целевую директорию в HDFS для импортирования данных в нее (Рисунок 12).



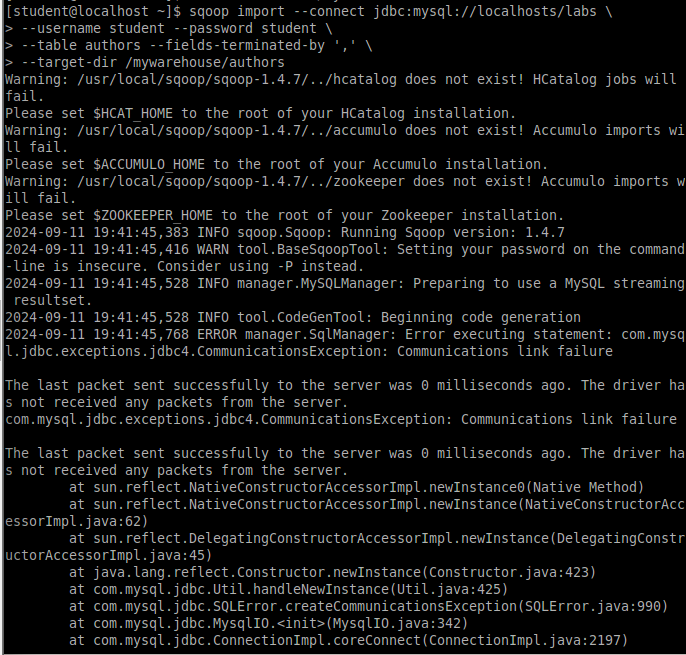
**Рисунок 12** — **Создание целевой директории**

Импортируем таблицу authors и сохраняем ее в директорию HDFS, которую мы создали выше, используя ‘,’ для разделения полей (Рисунок 13).



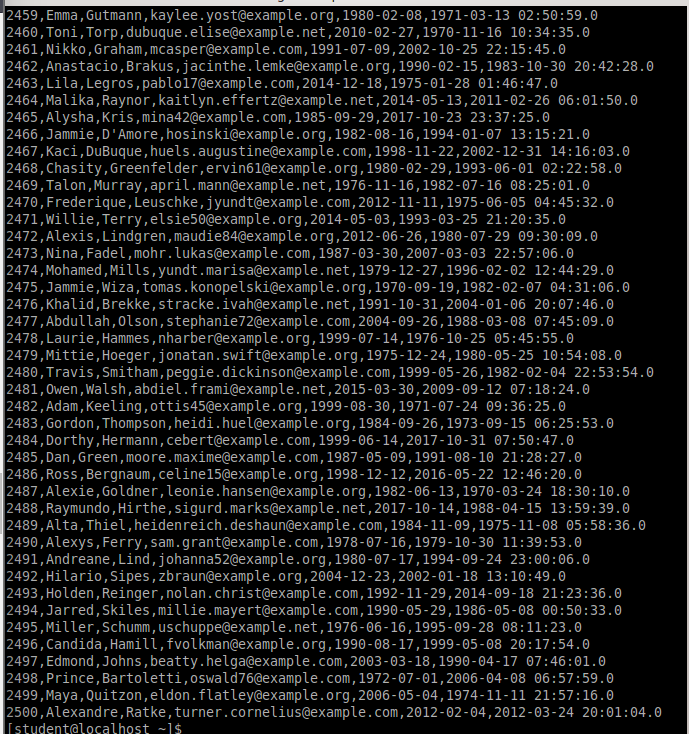
**Рисунок 13** — **Импорт таблицы authors**

Для проверки воспользуемся командой hdfs по целевой директории (Рисунок 14).



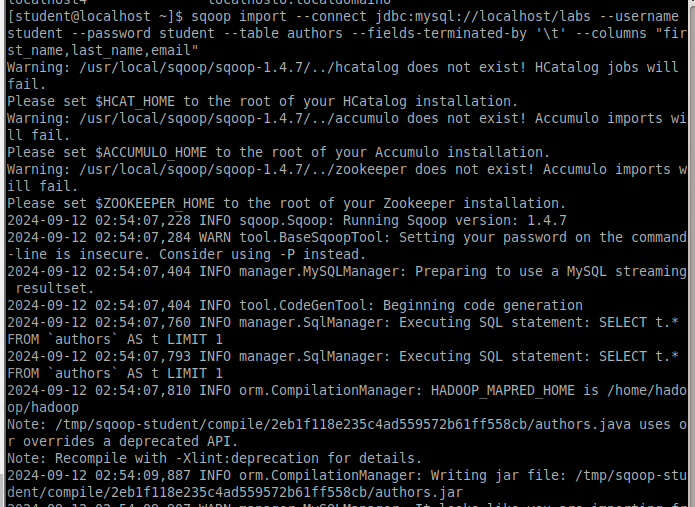
**Рисунок 14. Список файлов в /mywarehouse/authors**

Если запустить команду cat, вы сможете увидеть, что каждая строка данных отделена от соседней знаком ‘,’, что отличается от того, что было в файле posts (там были табуляции) в HDFS.



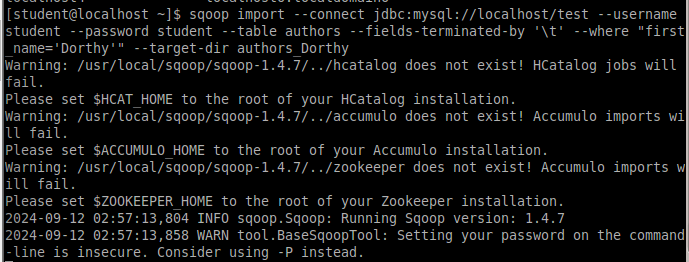
**Рисунок 15 – Вывод командой cat**

Импортирование определенных столбцов осуществляется указанием аргумента --columns для папки authors в домашней директории HDFS. Импортируемые столбцы: first\_name, last\_name, email



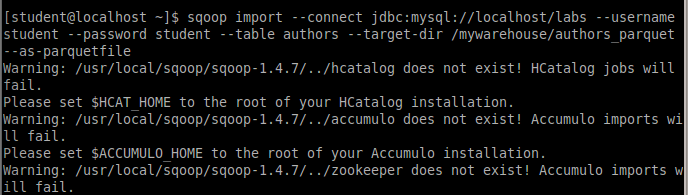
**Рисунок 16 – Импорт некоторых столбцов**

Импортируем только подходящие строки с помощью --where. Для примера импортируем строки из таблицы authors, где first\_name указано, как ‘Dorthy’.

****

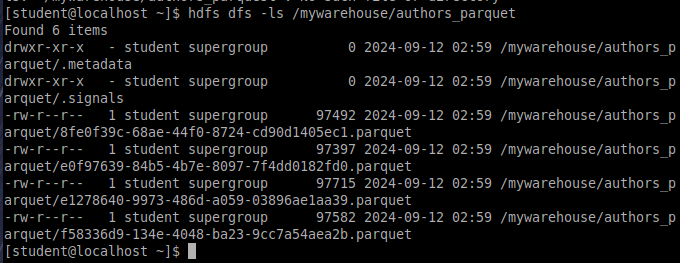
**Рисунок 17 – Импорт некоторых столбцов**

Импортируем таблицу, используя другой формат файла, вместо простого текстового. Импортируем таблицу authors в файл формата Parquet:



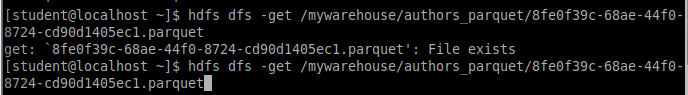
**Рисунок 18 – Импорт в формат Parquet**

Смотрим результат при помощи отображения содержимого целевой директории HDFS (Рисунок 19):

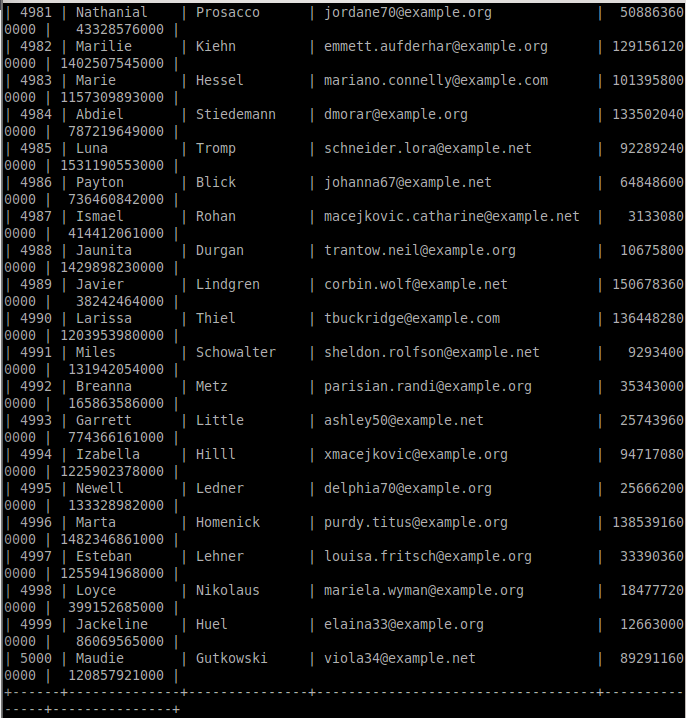


**Рисунок 19 – Список файлов Parquet**

Каждому файлу Parquet дано уникальное имя, например, 8fe0f39c-68ae-44f08724-cd90d1405ec1.parquet. Данные в этом формате нельзя просмотреть напрямую, так как формат двоичный. Для просмотра записей в таких файлах используйте команду parquet-tools show. Сначала нужно получить файл parquet на локальный узел, а затем запустить команду parqet-tools:

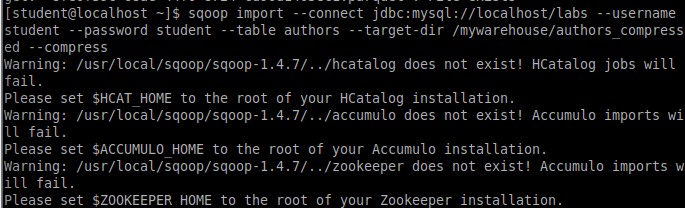


**Рисунок 20 – Получили файл**

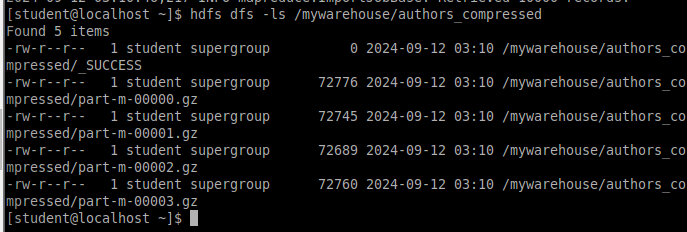


**Рисунок 21 – Список записей**

Импортируем таблицу authors, используя сжатие (аргумент --compress или -z):

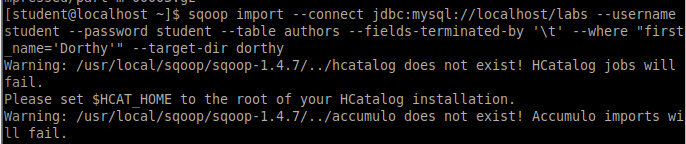
****

**Рисунок 22 – Импорт с сжатием**

****

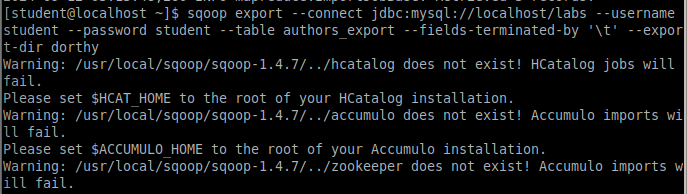
**Рисунок 23 – Список сжатых файлов**

Импортируем строки, где first\_name указано, как «Dorthy» (мы уже делали это в шаге 15) и сохраняем результат в папку dorthy в домашней директории HDFS:

****

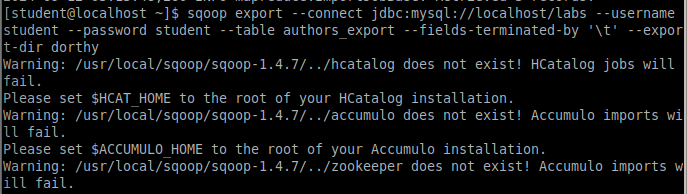
**Рисунок 24 – Импорт строк Dorthy**

Экспортируем сохраненную папку dorthy в качестве таблицы в базу данных labs:

****

**Рисунок 25 – Экспорт папки в labs**

Посмотрим содержимое экспортированных записей в MariaDB:

****

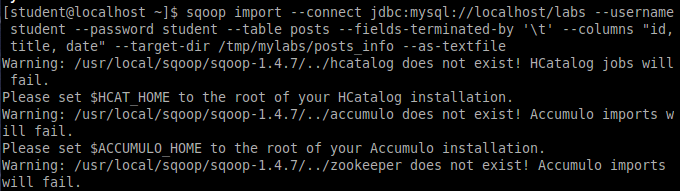
**Рисунок 26 – Экспортированные записи в MariaDB**

**Самостоятельная работа**

Создайте целевую директорию в HDFS для импортирования данных таблицы из базы labs (/tmp/mylabs).

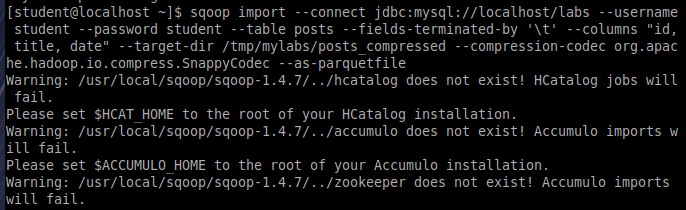
Из таблицы posts импортируйте первичный ключ, заголовки постов, дату публикации в директорию HDFS /tmp/mylabs/posts\_info. Сохраните файл в текстовом формате с табуляциями в качестве разделителей.

Подсказка: вам необходимо выяснить названия нужных столбцов, используя команды sqoop, а не напрямую в MariaDB.



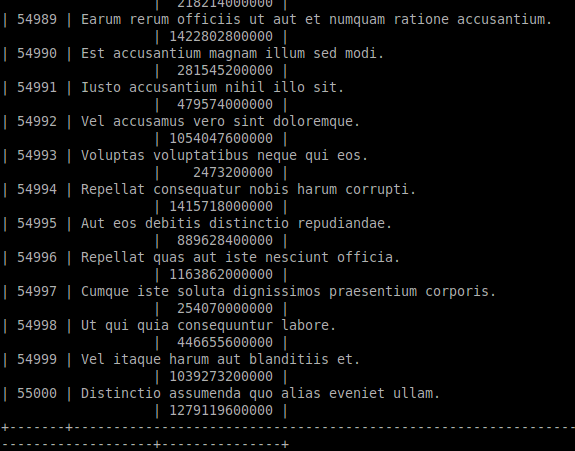
**Рисунок 27 – Импорт данных в текстовом формате**

Теперь сохраните то же самое, только в формате parquet со сжатием snappy в папке /tmp/mylabs/posts\_compressed.



**Рисунок 28 – Импорт данных в текстовом формате**

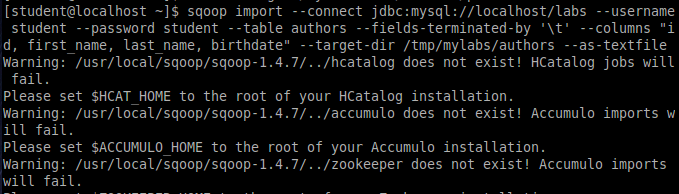
Из окна терминала отобразите несколько записей из тех, что вы только что проимпортировали.



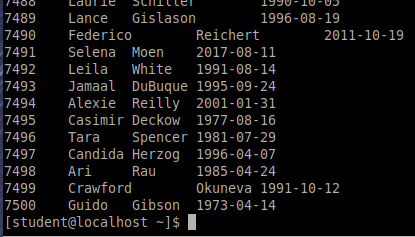
**Рисунок 29 – Просмотр записей**

Импортируйте столбцы id, first\_name, last\_name, birthdate из таблицы authors в домашнюю директорию HDFS. Файл сохраните в текстовом формате с табуляциями в качестве разделителей.

Подсказка: если директория authors уже существует в домашней директории HDFS, удалите ее и только потом выполните команду import.



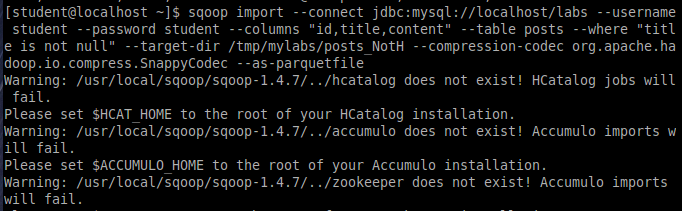
**Рисунок 30 – Импорт данных**



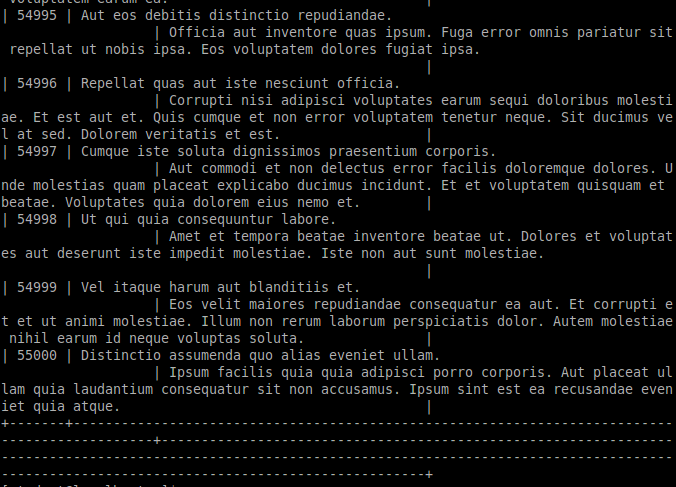
**Рисунок 31 – Просмотр записей**

Импортируйте и сохраните в директории /tmp/mylabs/posts\_NotN те посты, название которых не NULL. Импортируйте только следующие столбцы: id, title, content. Сохраните файл в формате parquet, а для сжатия используйте кодек snappy.

Подсказка: выражение SQL для проверки того, что значение в столбце не NULL: "название\_столбца is not NULL"



**Рисунок 32 – Импорт данных в формате Parquet**



**Рисунок 33 – Просмотр данных в формате Parquet**