Лабораторная работа 5.1. Развертывание и настройка Hadoop Анализ данных с использованием экосистемы Hadoop

Цель работы: получить практические навыки развертывания одноузлового кластера Hadoop, освоить базовые операции с распределенной файловой системой HDFS, выполнить загрузку и простейшую обработку данных, а также научиться выгружать результаты для последующего анализа и визуализации во внешней среде (Jupyter Notebook / Google Colab).

Оборудование и программное обеспечение:

- Предустановленная виртуальная машина с OS Linux (Ubuntu 20.04+) или система контейнеризации Docker.
- Установленный **Java Development Kit (JDK)** версии 8 или 11.
- Дистрибутив **Hadoop 3**.
- Внешняя среда для анализа: **Google Colab** или **Jupyter Notebook**.
- Python-библиотеки для анализа (в Colab/Jupyter): pandas, seaborn, matplotlib.
- Система контроля версий **Git** и аккаунт на **GitHub** (или GitLab/Bitbucket).

Теоретические основы

Экосистема Hadoop — это набор программных инструментов с открытым исходным кодом для распределенного хранения и обработки больших наборов данных (Big Data). Ключевыми компонентами являются:

- **HDFS (Hadoop Distributed File System):** распределенная, отказоустойчивая файловая система, предназначенная для хранения очень больших файлов на кластерах из стандартного оборудования.
- **YARN (Yet Another Resource Negotiator):** система для управления ресурсами кластера и планирования выполнения заданий.
- **MapReduce:** программная модель для распределенных вычислений.
- **Apache Spark:** современный, высокопроизводительный фреймворк для обработки больших данных, который часто используется вместо или вместе с MapReduce для ускорения вычислений за счет обработки данных в оперативной памяти.

ETL (Extract, Transform, Load) — это процесс, в ходе которого данные извлекаются из одного или нескольких источников, преобразуются в нужный формат и загружаются в целевое хранилище. в контексте Hadoop этот процесс позволяет интегрировать данные из традиционных баз данных (например, PostgreSQL) в озеро данных на HDFS для дальнейшего масштабного анализа.

Порядок выполнения работы

Подготовка окружения:

- запустите предоставленную виртуальную машину или Docker-контейнеры, убедитесь в работоспособности Hadoop (команда jps), Spark и PostgreSQL.
- проверьте доступ к веб-интерфейсам HDFS (http://localhost:9870) и YARN (http://localhost:8088).

Извлечение и загрузка данных в HDFS (Extract & Load):

- выберите свой вариант задания из таблицы ниже и скачайте соответствующий CSV-файл по прямой ссылке.
- напишите Python-скрипт для создания таблицы в PostgreSQL и загрузки в нее данных из вашего CSV-файла. это имитирует наличие данных в операционной реляционной СУБД.
- напишите PySpark-скрипт, который подключается к вашей таблице в PostgreSQL, считывает данные в DataFrame, а затем сохраняет их в HDFS в формате Parquet. Parquet — это колоночный формат хранения, оптимизированный для аналитических запросов в Spark.

Обработка данных (Transform):

- Используйте утилиту **Hadoop Streaming** для выполнения простой задачи обработки данных (например, подсчет строк, фильтрация). Это позволяет использовать любой исполняемый файл или скрипт в качестве mapper и reducer.
- *Пример: подсчет строк в файле.*

```
hadoop-streaming.jar \
-input /user/your_user/lab_data/input.csv \
-output /user/your_user/lab_data/output \
-mapper "/bin/cat" \
-reducer "/usr/bin/wc -l"
```

- Выполните аналитическую задачу из вашего варианта, написав простые скрипты на Bash (например, с использованием grep, awk, sort, uniq) или Python, которые будут выполнять роль mapper и reducer.

Выгрузка результатов:

- Просмотрите результат выполнения задачи в HDFS командой hdfs dfs -cat.
- Выгрузите результирующий файл (или файлы) из HDFS на локальную файловую систему командой hdfs dfs -get.

Часть 3. Анализ и визуализация

Подготовка среды. Запустите Jupyter Notebook или откройте Google Colab.

Загрузка данных. Загрузите выгруженный из HDFS файл в вашу аналитическую среду.

Анализ и визуализация:

- Используя pandas, преобразуйте данные в удобный для анализа формат (DataFrame).
- С помощью matplotlib и seaborn постройте графики и диаграммы, которые наглядно представляют результаты вашего анализа (согласно заданию).
- Напишите текстовые выводы, объясняющие, что изображено на графиках и какие бизнес-инсайты можно из этого извлечь.

Задания для самостоятельной работы: кейсы анализа данных

№	Бизнес-кейс	Источник данных (CSV)	Аналитическая задача с использованием Hadoop (визуализировать результаты)
1	Анализ фондового рынка	S&P 500 Stock Data https://www.kaggle.com/datasets/camnugent/sandp500	Рассчитать среднюю цену закрытия и объем торгов по компаниям за год (Hive SQL)
2	Анализ транзакций	Credit Card Fraud Detection https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud	Сгруппировать транзакции по типу и посчитать мошеннические и легитимные операции (MapReduce)
3	Анализ онлайн-ритейла	Online Retail Dataset https://archive.ics.uci.edu/ml/datasets/Online+Retail	Найти топ-10 товаров по количеству продаж (Hive)
4	Задержки авиарейсов	US Flight Delays https://www.kaggle.com/datasets/usdot/flight-delays	Средняя задержка вылета по аэропортам (Spark)
5	Рынок недвижимости	Melbourne Housing Market https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot	Средняя цена и подсчет объектов по району (HiveQL)
6	Сердечно-сосудистые заболевания	Cardiovascular Disease Dataset https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset	Средний возраст, вес, рост пациентов с заболеванием и без (MapReduce)
7	YouTube видео	YouTube Trending Videos https://www.kaggle.com/datasets/datasnaek/youtube-new	Топ-5 каналов по количеству лайков (Hive)
8	Продажи видеоигр	Video Game Sales https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings	Общие продажи по жанрам и платформам (MapReduce)
9	ДТП в США	US Accidents https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents	Количество аварий по серьезности и погоде (Hive)
10	Потребление энергии	Electricity Load Diagrams https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014	Среднее почасовое потребление по месяцам (Spark)
11	Температуры по городам	Global Land Temperatures https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data	Средняя температура для стран (Hive)

№	Бизнес-кейс	Источник данных (CSV)	Аналитическая задача с использованием Hadoop (визуализировать результаты)
12	Зарплаты Data Scientist	Data Science Salaries https://www.kaggle.com/datasets/thedevastator/data-science-job-salaries	Средняя зарплата по уровню опыта (MapReduce)
13	Фильмы IMDb	IMDb Top 1000 https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset	Топ-10 режиссеров по среднему рейтингу (Hive)
14	Отток клиентов в телеком	Telco Customer Churn https://www.kaggle.com/datasets/blastchar/telco-customer-churn	Процент оттока по типам контрактов (MapReduce)
15	Доклад о счастье 2019	World Happiness Report https://www.kaggle.com/datasets/unsdsn/world-happiness	Средний балл счастья и ВВП по регионам (Hive)
16	Популярность Spotify	Spotify Tracks https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db	Средние показатели "энергичность" и "танцевальность" по жанру (MapReduce)
17	Аренда жилья Airbnb Нью-Йорк	Airbnb New York https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data	Средняя цена и количество предложений по району (Hive)
18	Пассажиры Титаника	Titanic Dataset https://www.kaggle.com/c/titanic/data	Процент выживших по классу билета и полу (MapReduce)
19	Качество красного вина	Wine Quality Dataset https://archive.ics.uci.edu/ml/datasets/Wine+Quality	Средний уровень алкоголя по уровню качества (Hive)
20	Велопрокат в Сеуле	Seoul Bike Sharing Demand https://www.kaggle.com/datasets/ymq2010/seoul-bike-sharing-demand	Среднее количество арендованных велосипедов по времени года (Spark)
21	Успеваемость студентов	Student Performance Dataset https://archive.ics.uci.edu/ml/datasets/Student+Performance	Средние оценки по времени в пути до школы (Hive)
22	Погода в Дели	Delhi Weather Daily https://www.kaggle.com/datasets/muthuj7/weather-dataset	Найти месяц с самой высокой средней температурой (MapReduce)

№	Бизнес-кейс	Источник данных (CSV)	Аналитическая задача с использованием Hadoop (визуализировать результаты)
23	Зарплаты IT в Европе	IT Salaries Europe https://www.kaggle.com/datasets/radmirzosimov/it-salaries-in-europe	Средняя зарплата по специализациям (Hive)
24	Медицинские страховые выплаты	Medical Insurance Dataset https://www.kaggle.com/datasets/mirichoi0218/insurance	Средняя сумма выплат для курящих и некурящих (MapReduce)
25	Диабет	Diabetes Dataset https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database	Средний ИМТ для людей с и без диабета (Hive)
26	Цены на авокадо	Avocado Prices Dataset https://www.kaggle.com/datasets/neuromusic/avocado-prices	Регион с наивысшей средней ценой (MapReduce)
27	Отзывы на музыкальные инструменты	Amazon Musical Instruments Reviews https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews	Средний рейтинг и количество отзывов для топ-10 рецензентов (Hive)
28	Рынок криптовалют	Cryptocurrency Market Data https://www.kaggle.com/datasets/mczielinski/bitcoin-historical-data	Средняя рыночная капитализация топ-5 криптовалют (MapReduce)
29	Алмазы	Diamonds Dataset https://www.kaggle.com/datasets/shivam2503/diamonds	Средняя цена по типам чистоты (Hive)
30	Землетрясения	Earthquake Data https://www.kaggle.com/datasets/usgs/earthquake-database	Тип землетрясения с максимальной средней магнитудой (Spark)

Правила оформления отчета

Отчет о выполненной работе должен быть представлен в виде единого документа (в формате .pdf или .docx) и содержать следующие разделы:

1. **Титульный лист:** с указанием названия вуза, дисциплины, темы работы, номера варианта, ФИО студента и преподавателя.
2. **Цель работы:** переформулированная цель из данного методического указания.
3. **Краткое описание бизнес-кейса:** описание задачи, которую вы решали в рамках своего варианта.
4. **Описание конвейера данных (ETL):**
    - **Extract:** описание процесса загрузки исходного CSV-файла в таблицу PostgreSQL. Привести код Python-скрипта и скриншот с данными в pgAdmin.
    - **Load to HDFS:** описание процесса извлечения данных из PostgreSQL и загрузки их в HDFS с помощью PySpark. Привести код PySpark-скрипта и скриншот веб-интерфейса HDFS, показывающий загруженные файлы в формате Parquet.
5. **Аналитическая обработка в Spark (Transform):**
    - **Код:** листинг кода PySpark-скрипта, выполняющего аналитическую задачу вашего варианта.
    - **Результат:** скриншот вывода Spark-скрипта и скриншот веб-интерфейса HDFS, показывающий результирующий CSV-файл.
6. **Визуализация и выводы:**
    - **Код:** листинг кода Python-скрипта для построения визуализаций.
    - **Визуализации:** вставленные в отчет графики и диаграммы.
    - **Выводы:** развернутый текстовый анализ полученных результатов. Объясните, какие бизнес-инсайты можно извлечь из проведенного анализа.
7. **Общий вывод по работе:** краткие итоги, описывающие полученные навыки и понимание работы с экосистемой Hadoop.

Критерии оценки работы (10-балльная система)

- **10 баллов (отлично):**
    - работа выполнена в полном объеме, продемонстрирован полный конвейер ETL (PostgreSQL -> HDFS -> Spark -> HDFS).
    - код на Python/PySpark чистый, эффективный и хорошо прокомментирован.
    - анализ результатов глубокий, выводы логичны и подкреплены визуализациями.
    - отчет безупречно оформлен.

- **8-9 баллов (хорошо):**
  - работа выполнена полностью, но в коде есть незначительные недочеты или недостаточно комментариев.
  - анализ результатов верный, но мог бы быть более детальным.
  - все необходимые элементы отчета присутствуют.
- **6-7 баллов (удовлетворительно):**
  - работа выполнена, но не в полном объеме (например, данные загружены напрямую в HDFS, минуя PostgreSQL, или отсутствует визуализация).
  - код работает, но его структура неоптимальна.
  - выводы поверхностны.
- **4-5 баллов (посредственно):**
  - выполнены только базовые части задания (например, только загрузка данных без обработки).
  - код содержит ошибки, мешающие его полному выполнению.
- **1-3 балла (неудовлетворительно):**
  - работа не соответствует минимальным требованиям, студент не смог справиться с основными задачами.
- **0 баллов:**
  - работа не сдана или является плагиатом.