**СОДЕРЖАНИЕ**

[Введение 3](#_Toc194259623)

[1. Теоретическая часть 4](#_Toc194259624)

[1.1 Описание исследуемой области 4](#_Toc194259625)

[1.2 Информация об источнике и о наборе данных 5](#_Toc194259626)

[1.3 Описание процесса сбора данных 5](#_Toc194259627)

# Введение

Загрязнение воздуха является глобальной проблемой, которая представляет серьезную угрозу здоровью и благополучию людей. Воздействие загрязнения воздуха связано с целым рядом проблем со здоровьем, включая сердечно-сосудистые заболевания, инсульт и рак легких. Воздействие загрязнения воздуха на здоровье населения отражается в тревожной статистике смертей, вызванных им. Согласно последним исследованиям, загрязнение воздуха ежегодно приводит к миллионам смертей во всем мире.

В данной работе рассматривается жизненный цикл анализа данных включающий сбор, хранение, обработку и визуализацию информации о загрязнении воздуха и связанных с ним заболеваниях. Используются технологии Apache Sqoop, Hive, Spark и MariaDB, что позволяет эффективно обрабатывать большие объёмы данных и получать аналитические выводы.

Цель работы — разработка конвейера данных на основе технологий Big Data для анализа взаимосвязи между загрязнением воздуха и уровнем заболеваемости. В качестве среды для развертывания инструментов используется VirtualBox.

В работе решаются следующие задачи:

* Сбор данных из открытых источников (Kaggle).
* Построение конвейера для предобработки и анализа данных.
* Визуализация результатов для выявления взаимосвязей между загрязнением воздуха и уровнем заболеваемости.

# Теоретическая часть

## Описание исследуемой области

Анализ влияния загрязнения воздуха на уровень заболеваний представляет собой важное направление в экологической эпидемиологии и здравоохранении. Исследуемая область охватывает взаимосвязь между качеством атмосферного воздуха и показателями здоровья населения, включая:

Область исследования:

* Основные загрязнители: PM2.5, PM10, диоксид азота (NO2), диоксид серы (SO2), озон (O3);
* Заболевания: астма, сердечно-сосудистые, онкологические;
* Географический охват: глобальный анализ с акцентом на урбанизированные территории.

Актуальность исследования обусловлена:

1. Ростом урбанизации и промышленного производства;
2. Увеличением нагрузки на экосистемы;
3. Необходимостью доказательной базы для принятия управленческих решений.

В контексте курсовой работы анализ данных позволяет:

* Строить прогнозные модели;
* Оптимизировать систему экологического мониторинга.

## Информация об источнике и о наборе данных

В данной курсовой работе в конвейер поступит файл формата csv, который содержит статистику смертности, связанной с загрязнением воздуха, по различным странам и регионам за период с 1990 по 2019 год. Данные включают следующие ключевые показатели.

Данные, представленные в этом файле, имеют 6840 записей и 7 полей:

* id — номер записи;
* Entity — страна/регион;
* Year — год;
* Total Deaths for Air Pollution — общее количество смертей от загрязнения воздуха;
* Total Deaths for Outdoor Air Pollution — смертность от атмосферного воздуха;
* Total Deaths for Household Air Pollution from Solid Fuels — смертность от бытового загрязнения (топливо);
* Death Rate from Air Pollution Per 100000 — коэффициент смертности на 100 тыс. человек.

## Описание процесса сбора данных

Процесс сбора данных для анализа включал следующие этапы:

1. Получение данных: данные загружаются из репозитория Kaggle (<https://www.kaggle.com/datasets/abmsayem/air-pollution/data>) в формате CSV. Для этого используются библиотеки Python, такие как pandas, которые позволяют удобно работать с табличными данными.
2. Архитектура конвейераи предобработка данных: построение конвейера на основе изученных технологий для предобработки и маршрутизации выбранных данных.

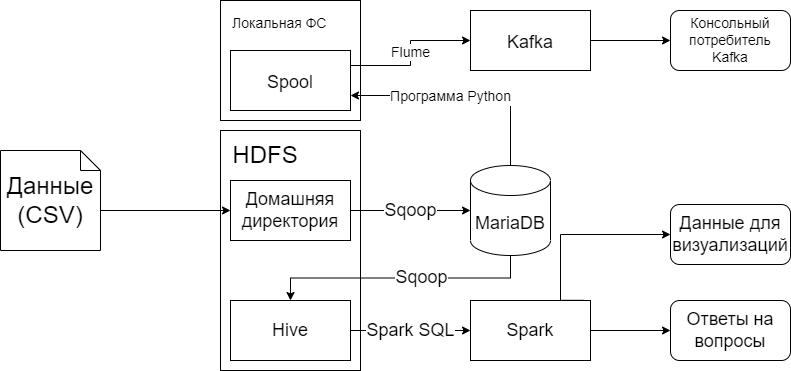


Рисунок 1.1 — Схема конвейера

1. Анализ данных и визуализация: после предобработки данные подвергаются анализу с использованием статистических методов и методов машинного обучения.

Инструменты и методы:

* Hadoop (HDFS) — Хранение данных;
* Hive — SQL-интерфейс для структурирования;
* Spark — Распределённая обработка и анализ;
* Python — (Pandas/Matplotlib) Визуализация.