Архитектура решения

Система анализа данных COVID-19 на основе PySpark.

1

Загрузка данных

metadata.csv

2

Предобработка

Очистка, унификация

3

Анализ качества

Пропуски, аномалии

4

SQL-аналитика

5 запросов

5

PySpark обработка

UDF, фильтрация

6

Визуализация

4 графика

Технологии

- Apache Spark (PySpark)
- Spark SQL
- Matplotlib / Seaborn
- Google Colab

Ключевые операции

- UDF функции
- Оконные функции SQL
- Фильтрация и группировка
- Сохранение в Parquet

Ключевые статистики

Результаты анализа датасета COVID-19

670

Всего записей

Полный объем данных для анализа

2019-2023

Период данных

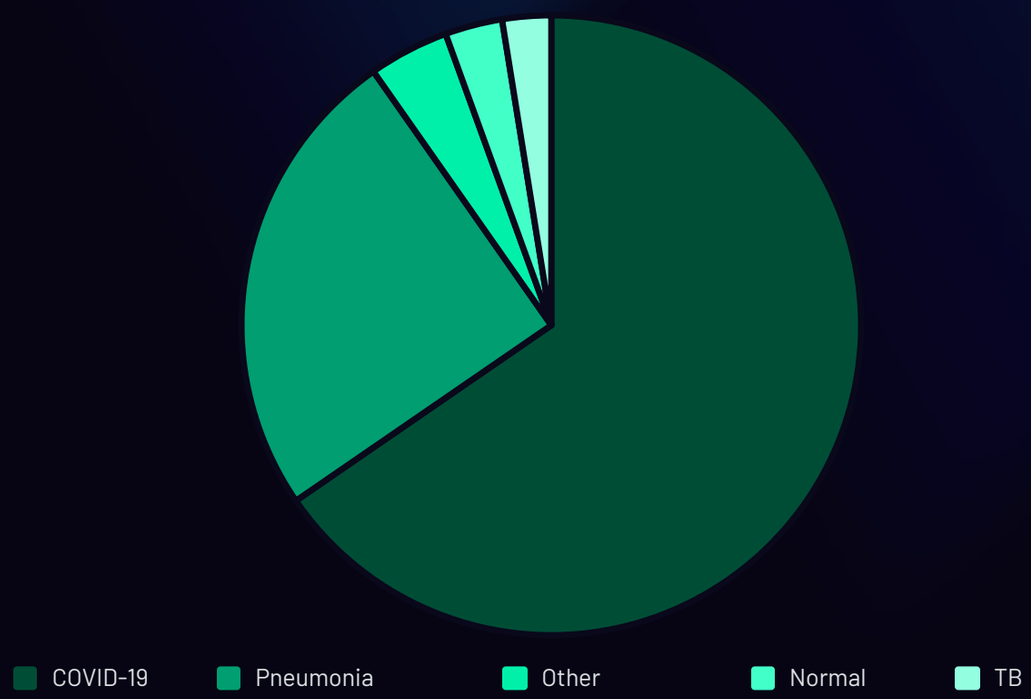
Охват пандемии и послепандемийного периода

5

Категорий

Различные типы диагнозов, классифицированных в датасете

Распределение диагнозов



SQL-аналитика

- 5 запросов
- Оконные функции
- Временные тренды

PySpark

- 2 UDF функции
- 5 фильтраций
- Формат Parquet

Визуализация

- Круговая
- Столбчатая
- Heatmap

Визуализации и выводы

Результаты эпидемиологического анализа

Созданные визуализации

- Круговая диаграмма
- Столбчатая диаграмма
- График трендов
- Heatmap

Ключевые выводы

- COVID-19 доминирует — 65% случаев
- Высокая доля пожилых пациентов
- Пик заболеваемости в 2020-2021 годах

Использованные техники

- Matplotlib + Seaborn
- Конвертация Spark → Pandas
- Настройка цветов

Технические достижения

- Полный цикл Big Data
- Работа с PySpark
- SQL-аналитика

Применение

- Эпидемиологический мониторинг
- Анализ групп риска
- Планирование здравоохранения