

Разработка системы анализа медицинских изображений для эпидемиологического мониторинга COVID-19

Итоговый проект
По дисциплине «Инфраструктура Big Data»

Выполнил:
студент группы М24-525
Лапшин Степан Сергеевич

Архитектура решения

1. Источник данных:

- metadata.csv с рентген-снимками пациентов

2. Анализ качества

- Просмотр распределения пропущенных значений
- Выявление аномальных значений

2. Предобработка данных:

- Очистка пропусков
- Унификация диагнозов
- Удаление дубликатов
- Создание возрастных категорий

3. Хранилище и обработка:

- PySpark DataFrame → SQL-аналитика + UDF для категорий
- Сохранение в формат Parquet для оптимизации

4. Аналитика и визуализация:

- SQL-запросы
- Визуализации: диаграммы, графики, тренды, heatmap

Источник

Анализ качества

Предобработка

Обработка и
хранилище

Аналитика и
визуализация



Ключевые статистики

Данные показывают распределение пациентов по диагнозам, полу, возрасту, а также временные пики исследований и наиболее часто используемые проекции снимков

Диагноз	Количество записей	Количество пациентов	Пол (M/F)	Топ-возраста	Основной пик (год/месяц)	Основные проекции
COVID-19	482	319	330 / 152	94, 93	2020/1 (347)	PA, AP, AP Supine
Pneumonia	170	114	97 / 73	90, 80	2018/11 (92)	PA, AP, L
Other	88	43	59 / 29	78, 75, 70	2020/1 (76)	AP Supine, PA
Tuberculosis	11	11	8 / 3	78, 70, 58	2018/11 (11)	PA, AP

Визуализации

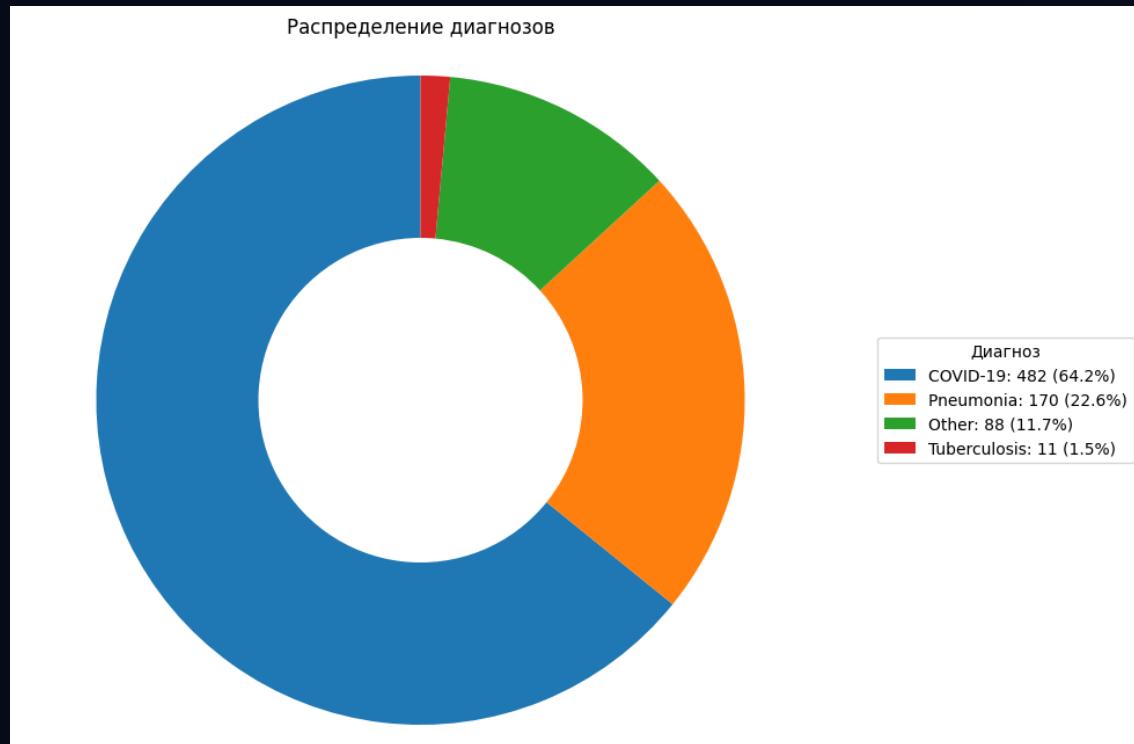
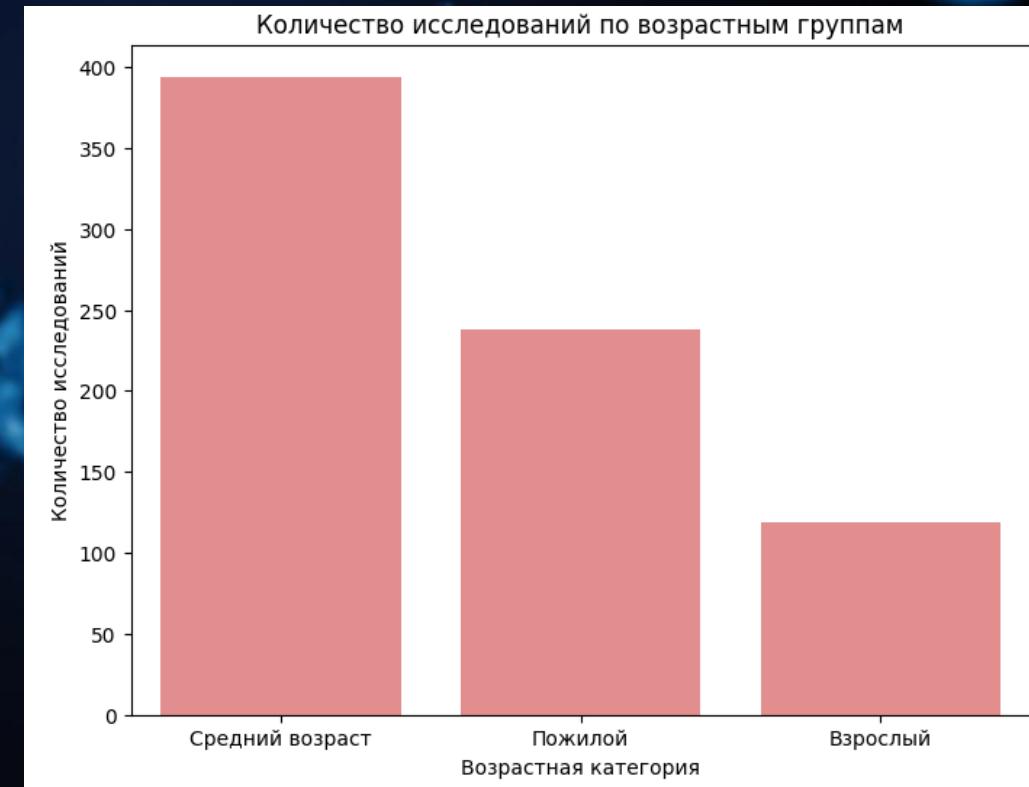
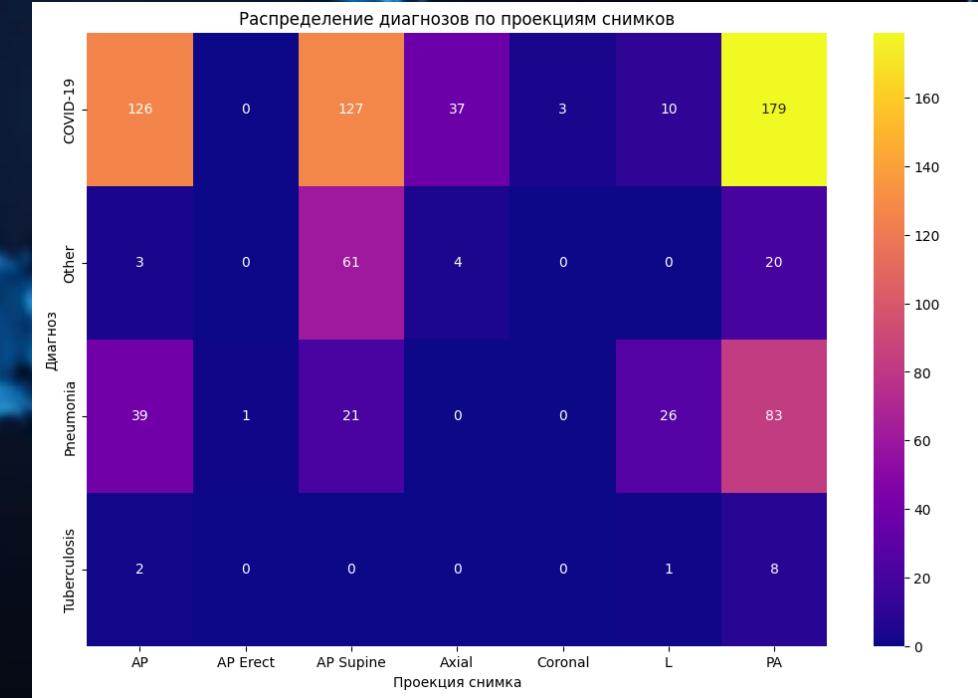
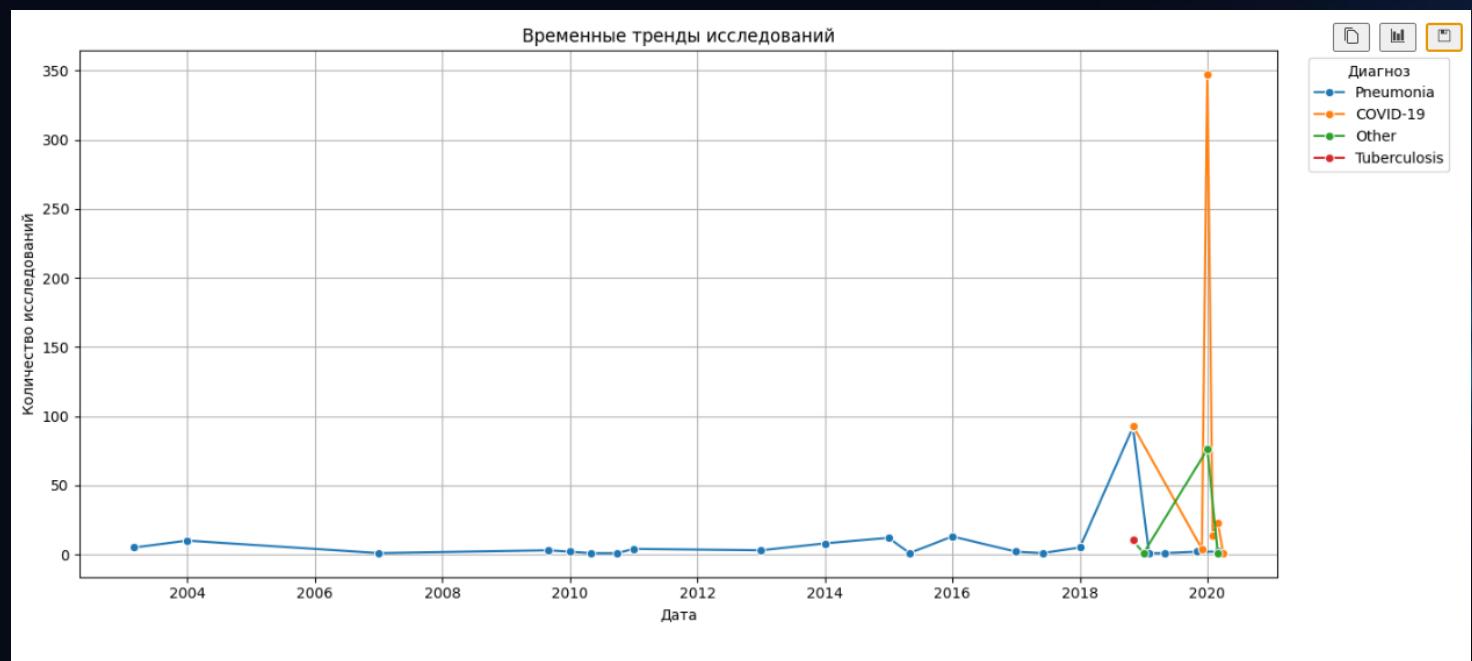


График показывает, что подавляющее большинство случаев приходится на COVID-19 (более 60%), тогда как пневмония занимает второе место с существенно меньшей долей. Остальные диагнозы встречаются значительно реже, что подчёркивает доминирующее влияние COVID-19 в представленном распределении.



Наибольшее количество исследований приходится на группу среднего возраста, за ней следует пожилой возраст. Исследования для взрослых проводятся реже.

Визуализации



В 2020 году наблюдается резкий рост числа исследований, что связано с пандемией COVID-19, в то время как до этого года количество исследований оставалось на стабильном низком уровне.

COVID-19 и пневмония являются наиболее часто встречающимися диагнозами, при этом большинство снимков сделано в проекциях АР и РА.

Вывод по работе

Проект является итоговой работой по дисциплине «Инфраструктура Big Data» и позволяет применить на практике полученные знания о распределённых вычислениях и работе с PySpark

Работа проводилась на реальном медицинском наборе данных, что позволило решать задачи, близкие к профессиональной практике анализа данных

Рекомендации по улучшению системы:

1. Внедрение строгой валидации данных.
2. Автоматизация унификации диагнозов.
3. Регулярное обновление базы данных.

Проект позволяет применить знания в области Big Data и аналитики для решения задач реального медицинского анализа.

