Аналитическая система мониторинга COVID-19 на основе PySpark

ИТОГОВЫЙ ПРОЕКТ — ИНФРАСТРУКТУРА BIG DATA

Архитектура решения

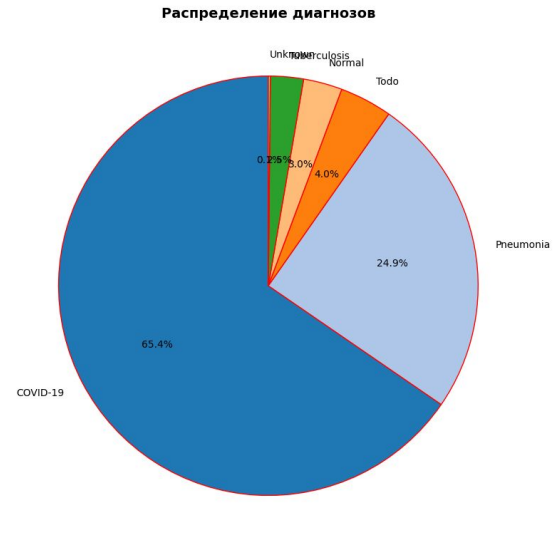
Ключевые элементы проекта:

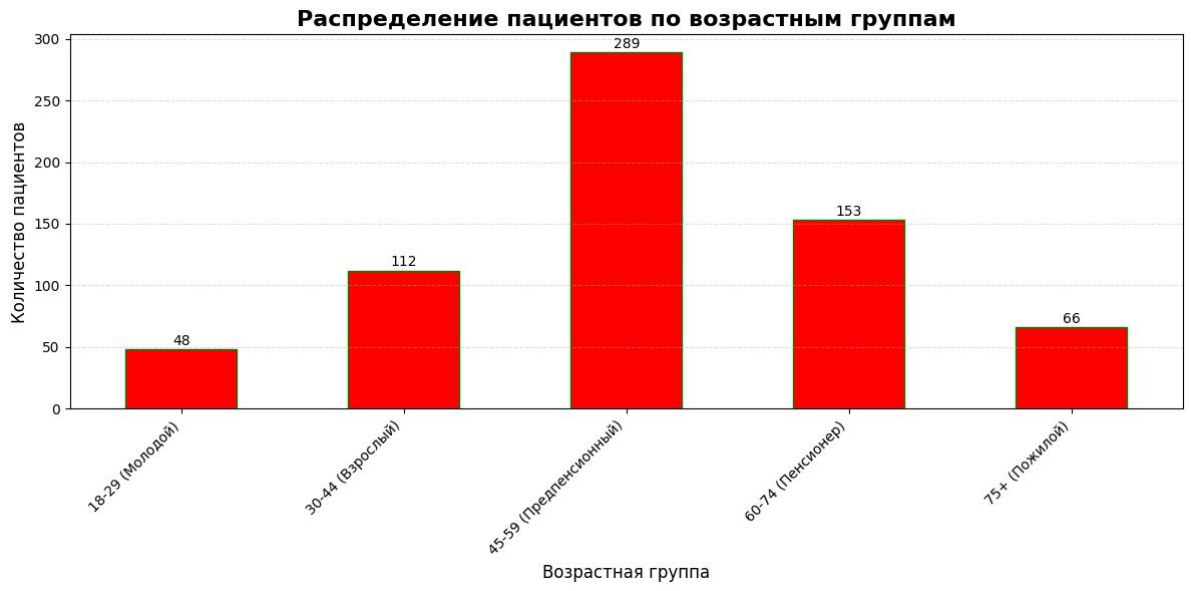
- **Источник данных:** COVID-19 Chest X-Ray (файл `metadata.csv`)
- **Среда обработки данных:** PySpark с использованием `SparkSession`, DataFrame API и SQL
- **Предварительная обработка:** очистка данных, стандартизация диагнозов, создание возрастных категорий
- **SQL-анализ:** группировки, оконные функции, анализ временных рядов
- **Пользовательские функции (UDF):** обработка дат, унификация диагнозов, категоризация возраста
- **Визуализация данных:** графики, диаграммы, тепловые карты, анализ динамики во времени

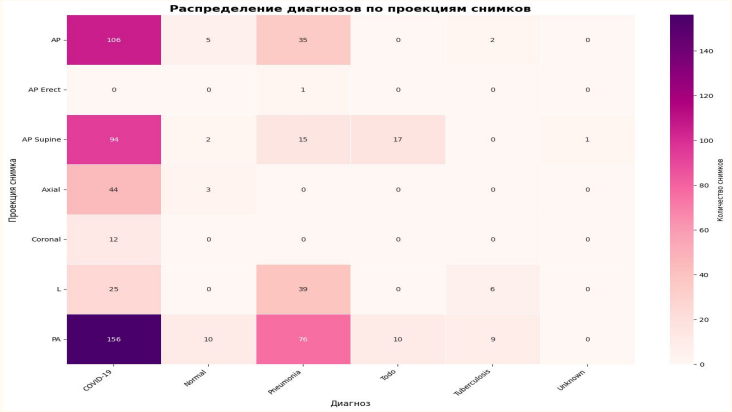
Ключевые статистики

1. Взрослые пациенты — основная группа
    - Более 85% исследуемых — лица старше 18 лет.
    - Наибольшая доля приходится на группу 18-65 лет.
2. COVID-19 — ключевой фокус исследований
    - 65% всех диагнозов — COVID-19.
    - Еще 25% — пневмония.
3. Группа 65+ лет является второй по численности (153 пациента).
4. Детская группа (0-17 лет) представлена значительно меньшим числом пациентов (66).
5. Чаще всего болеют люди в возрасте 45-59 лет
6. Пик исследований: 2020 год — 594 снимков

Визуализация







Вывод

Анализ данных показал, что COVID-19 является преобладающим диагнозом среди заболеваний органов грудной клетки, составляя почти 65% всех случаев в датасете. Распределение диагнозов демонстрирует выраженную зависимость от пола и возраста пациентов, причем наиболее пораженной возрастной группой оказались лица 45–59 лет. Пик диагностических исследований пришелся на 2020 год, что соответствует начальной фазе пандемии и отражает резкий всплеск медицинской активности. Для обработки крупных массивов метаданных эффективно применялся PySpark, а наглядная визуализация позволила оперативно оценивать ключевые эпидемиологические и демографические показатели.