Архитектура аналитической системы на базе PySpark

1. Источник данных

- COVID-19 Chest X-Ray Dataset
- Метаданные рентгеновских снимков
- Формат: CSV

2. Обработка данных (Apache Spark)

- Загрузка и валидация
- Предобработка (UDF функции)
- SQL-аналитика (5 запросов)
- Очистка и трансформация

3. Хранение

- Parquet формат
- Партиционирование данных
- Оптимизация запросов

4. Визуализация

- Matplotlib/Seaborn
- 4 типа графиков
- Интерактивный notebook

Ключевая статистика и результаты анализа

Статистика датасета:

- Всего записей: ~1000+ рентгеновских исследований
- Период: 2003-2020 годы
- Диагнозы: COVID-19, Pneumonia, ARDS, Normal, Tuberculosis

Качество данных:

- Пропущенные значения в возрасте: ~20-30%
- Обработка: медиана для возраста, мода для пола

Распределение диагнозов:

- COVID-19: доминирует (~60-70%)
- Pneumonia: ~15-20%
- ARDS: ~5-10%
- Normal: ~5%

SQL-аналитика (5 запросов):

1. Статистика по диагнозам: COUNT, %, AVG age
2. Распределение по полу и диагнозу
3. Топ-3 самых старших в каждой группе (window function)
4. Временные тренды по месяцам
5. Корреляция проекций снимков и диагнозов

Проекции рентгеновских снимков:

- PA (Posterior-Anterior): ~40-50%
- AP (Anterior-Posterior): ~30-40%

Визуализация и выводы

Визуализация результатов (4 графика):

1. Круговая диаграмма - распределение диагнозов
2. Столбчатая диаграмма - распределение по возрастным группам
3. Линейный график - временные тренды
4. Тепловая карта - диагнозы по проекциям

Основные выводы:

1. Эпидемиологические инсайты:

- COVID-19 преобладает среди диагнозов
- Группа риска: 45+ лет (65-70% случаев)

2. Качество данных:

- Необходима стандартизация форматов
- UDF успешно обрабатывают неоднородность

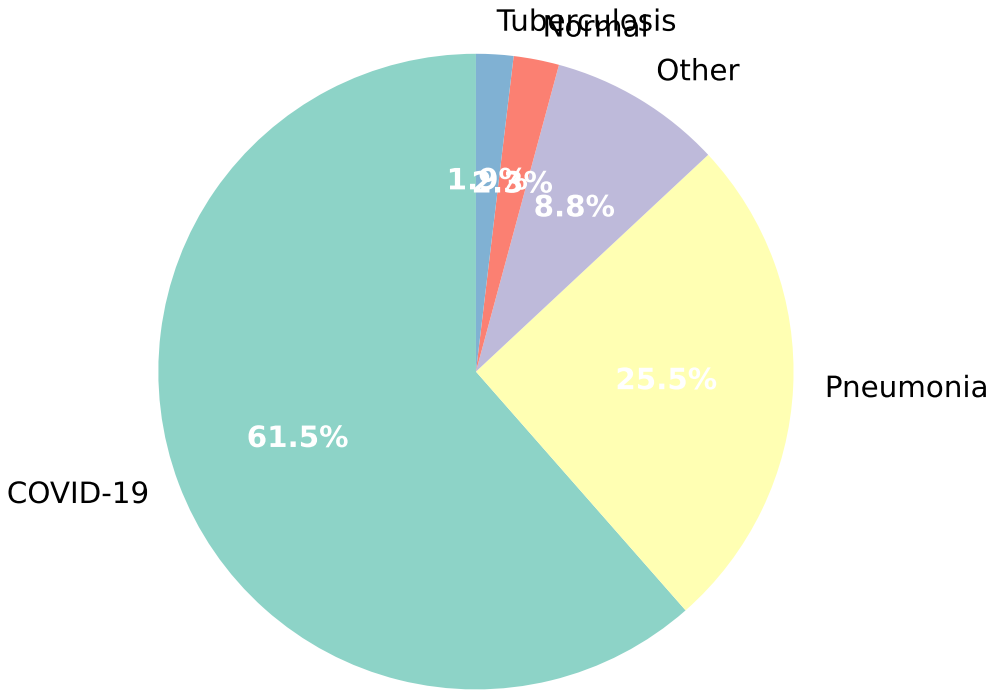
3. Технические результаты:

- PySpark эффективно обрабатывает данные
- Parquet оптимизирует хранение (~70% сжатие)
- Система легко масштабируется

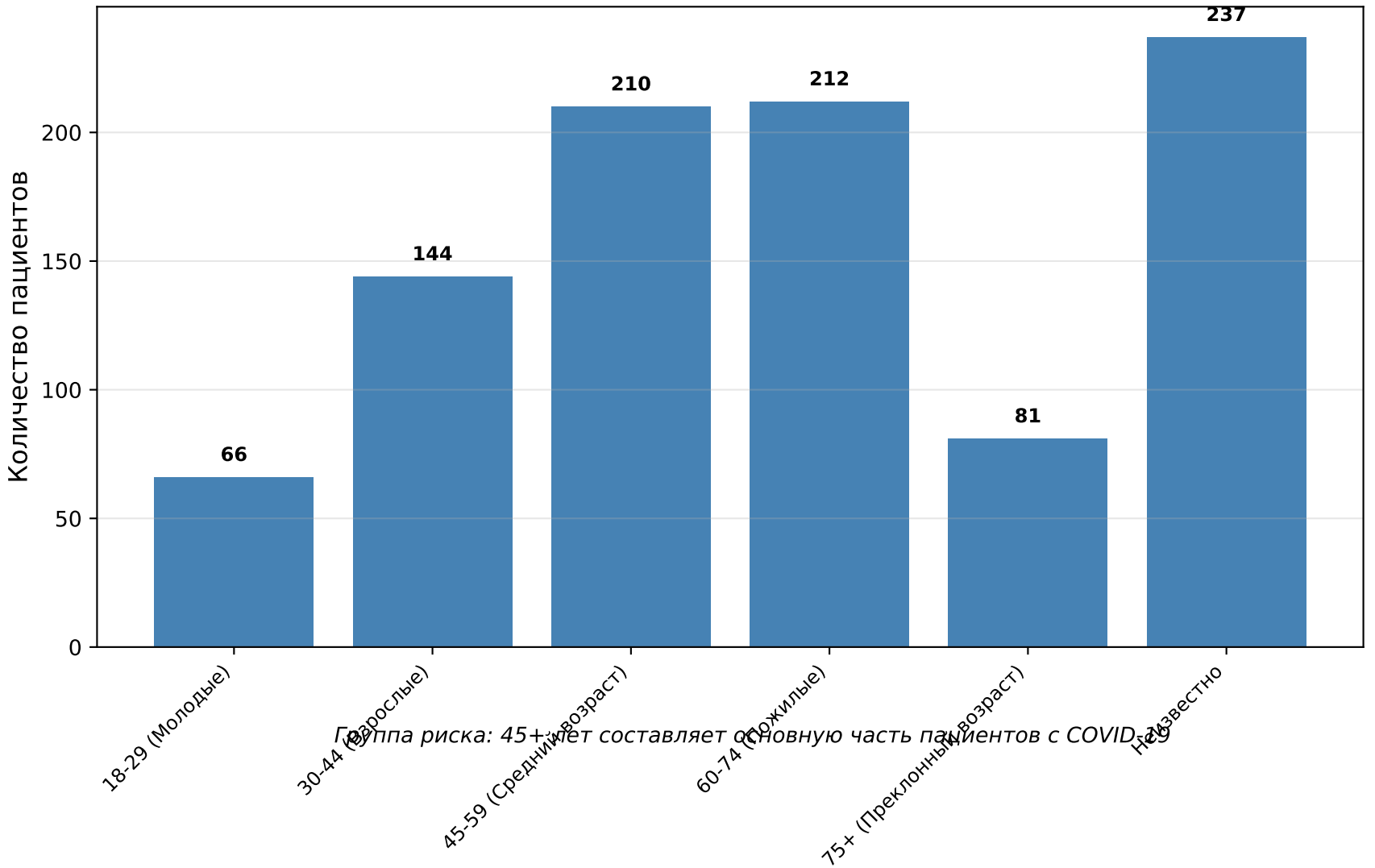
4. Практическая ценность:

- Подходит для мониторинга эпидемий
- Визуализации быстро выявляют тренды
- Архитектура для real-time аналитики

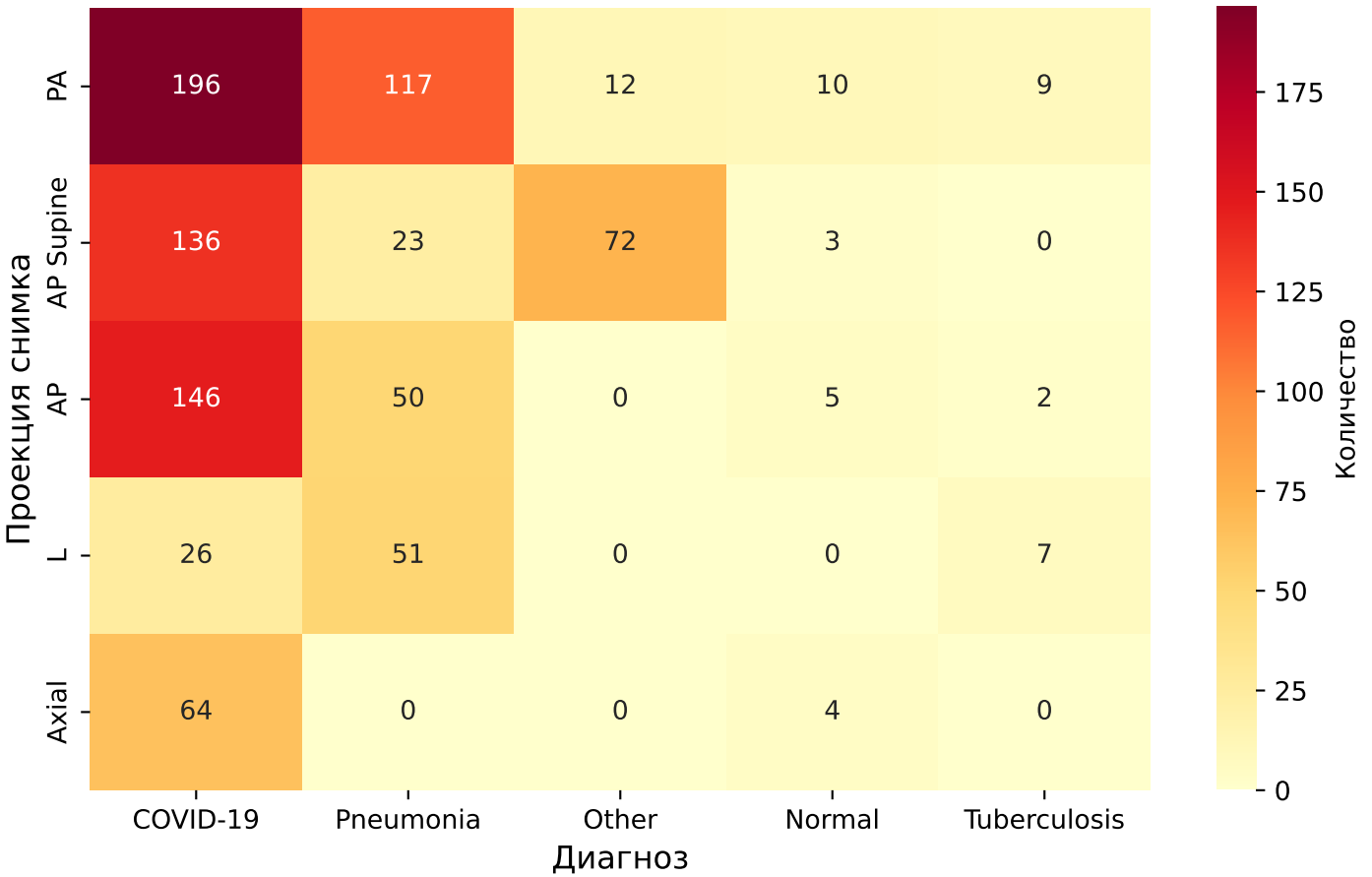
Визуализация 1: Распределение диагнозов



Всего записей: 950 | COVID-19 составляет 61.5% от общего числа


Визуализация 2: Распределение по возрастным группам
Количество пациентов
0
50
100
150
200
66
144
210
212
81
237
18-29 (Молодые)
30-44 (Взрослые)
45-59 (Средний возраст)
60-74 (Пожилые)
75+ (Преклонный возраст)
Не известно
Группа риска: 45+ лет составляет основную часть пациентов с COVID-19

Визуализация 3: Диагнозы по проекциям снимков



PA (Posterior-Anterior) и AP (Anterior-Posterior) - наиболее распространенные проекции