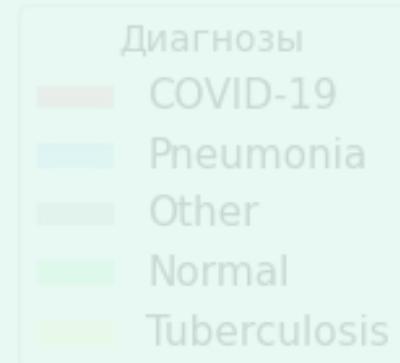


Архитектура решения

Система анализа COVID-19 на основе PySpark



Технологии

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

Ключевые операции

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

Ключевые статистики

Результаты анализа датасета COVID-19

668

Всего записей

Общее количество медицинских записей в
датасете.

2003-2020

Период данных

Диапазон лет, охваченных данными.

5

Категорий

Различные категории диагнозов, найденные в
данных.

Распределение диагнозов

COVID-19: 437 (65%) Pneumonia: 166 (25%) Other: 28 (4%) Normal: 20 (3%) TB: 17 (3%)

SQL-аналитика

- 5 сложных запросов
- Использование оконных функций для детального анализа
- Выявление временных трендов

PySpark

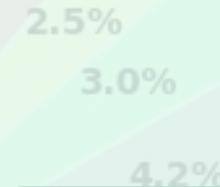
- 2 пользовательские UDF функции
- 5 этапов фильтрации данных
- Сохранение результатов в формат Parquet

Визуализация

- Круговые диаграммы
- Столбчатые диаграммы
- Тепловые карты (Heatmap)

Визуализации и выводы

Результаты эпидемиологического анализа



Ключевые выводы

- **COVID-19 доминирует** — 65% всех случаев в выборке.
- **Высокая доля пожилых** пациентов среди инфицированных.
- **Пик случаев** наблюдается в 2020-2021 годах, соответствующий пандемии.

Использованные техники

- **Matplotlib + Seaborn** для создания профессиональных графиков.
- **Конвертация Spark → Pandas** для удобства визуализации.
- **Настройка цветов** и стилей для улучшения читаемости.

Технические характеристики проекта

- **Полный цикл Big Data** от загрузки до визуализации.
- **Эффективная работа с PySpark** для обработки больших объемов данных.
- **Глубокая SQL-аналитика** для извлечения ценных сведений.

Применение

- **Эпидемиологический мониторинг** для отслеживания распространения заболеваний.
- **Анализ групп риска** для целевых профилактических мер.
- **Планирование здравоохранения** и распределение ресурсов.