Разработка системы анализа медицинских изображений

для эпидемиологического мониторинга COVID-19

Автор проекта:
Краснов Данила

Стек:

- Hadoop + Spark + PySpark
- Визуализация результатов в Jupyter

Архитектура системы

Raw CSV Data

Сырые исходные данные в формате CSV. Это таблицы с данными, собранными из различных источников (например, медицинские данные, логи, транзакции), которые будут загружены в систему для дальнейшей обработки.

HDFS (Hadoop FS)

Хранилище больших данных. Hadoop Distributed File System (HDFS) позволяет хранить большие объёмы файлов, разбивая их на блоки и распределяя между серверами в кластере для отказоустойчивости и масштабируемости.

Hive Metastore DB

База метаданных Hive хранит информацию о таблицах, схемах и форматах файлов в HDFS. Позволяет обращаться к данным в HDFS с помощью SQL-подобных запросов через Hive или Spark.

Spark (PySpark)

Мощный движок для обработки больших данных. PySpark позволяет быстро анализировать данные из HDFS или Hive, выполнять трансформации, агрегации, строить витрины данных, машинное обучение и многое другое.

Visualization (Jupyter)

Инструмент для визуализации и анализа данных. Jupyter Notebook позволяет писать Python-код, строить графики, таблицы и интерактивные дашборды для анализа результатов обработки данных в Spark.

Оптимизации производительности

- **Партиционирование:**
  - *По признакам:*
    - finding_grouped
    - age_group
- **Бакетирование (Bucketing):**
  - *По столбцам:*
    - sex
    - view
  - *Ускоряет JOIN и фильтры*
- **Сортировка внутри бакетов:**
  - По age

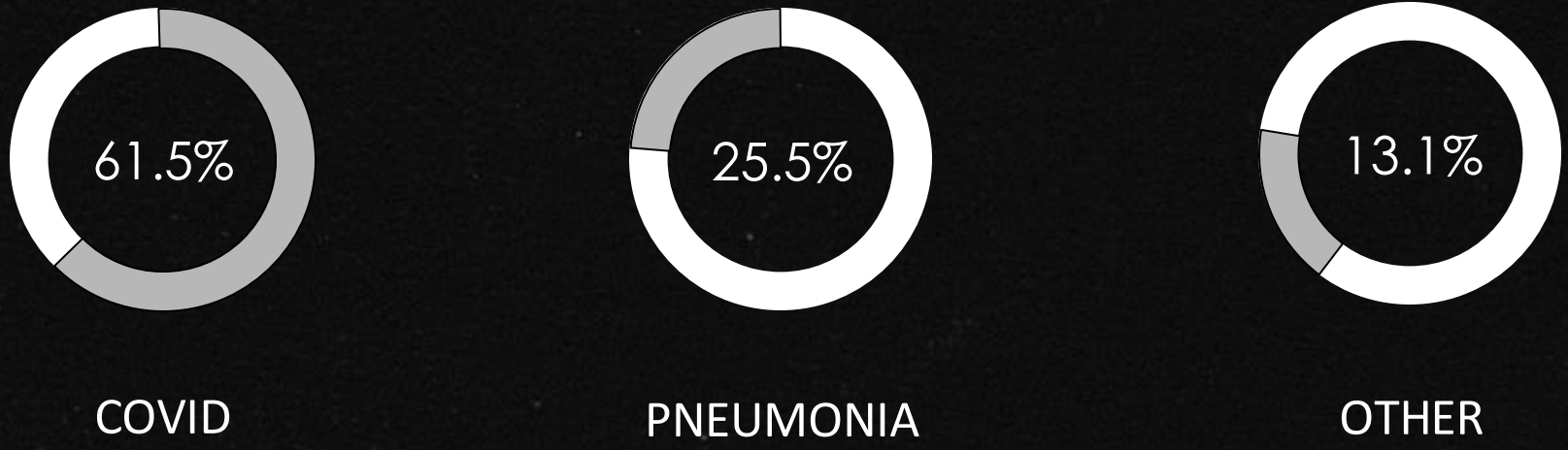
***Результат**: более быстрые запросы и экономия ресурсов кластера.*

Ключевые выводы

- **COVID-19:**
  - Явный показатель понижение сатурации
- **Основные виды снимков позволяющие диагностировать заболевание:**
  - PA (Posteroanterior)
  - AP (Anteroposterior)
- **Доля COVID-19 среди всех диагнозов:**
  - Около 61.26%

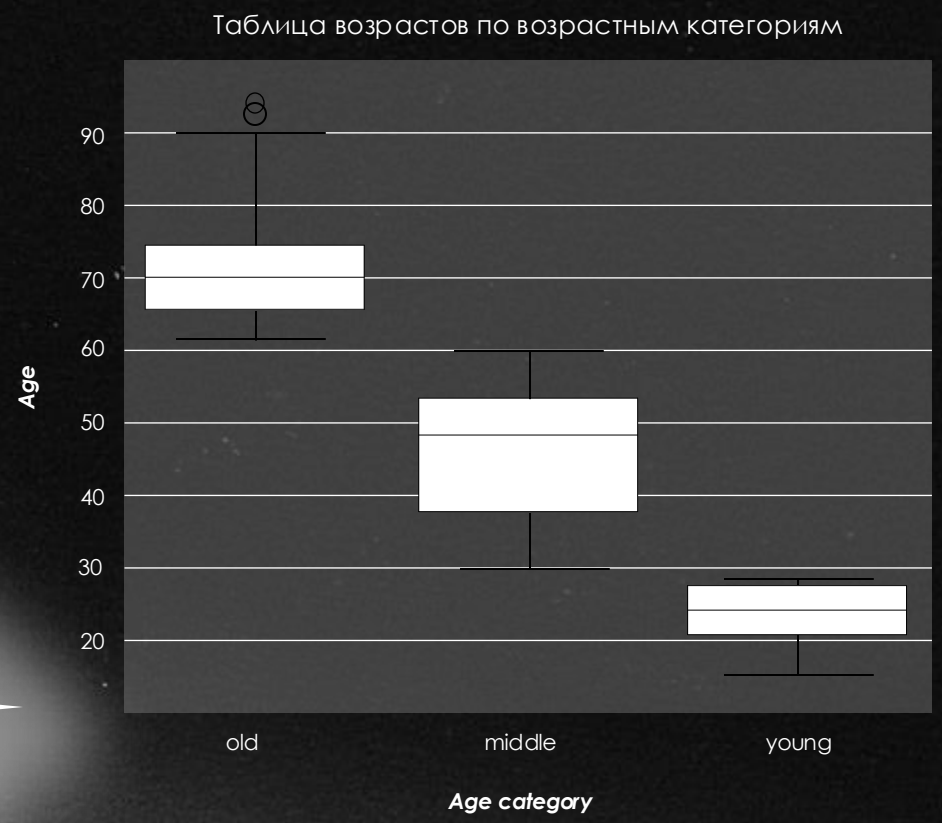
Примеры визуализаций:

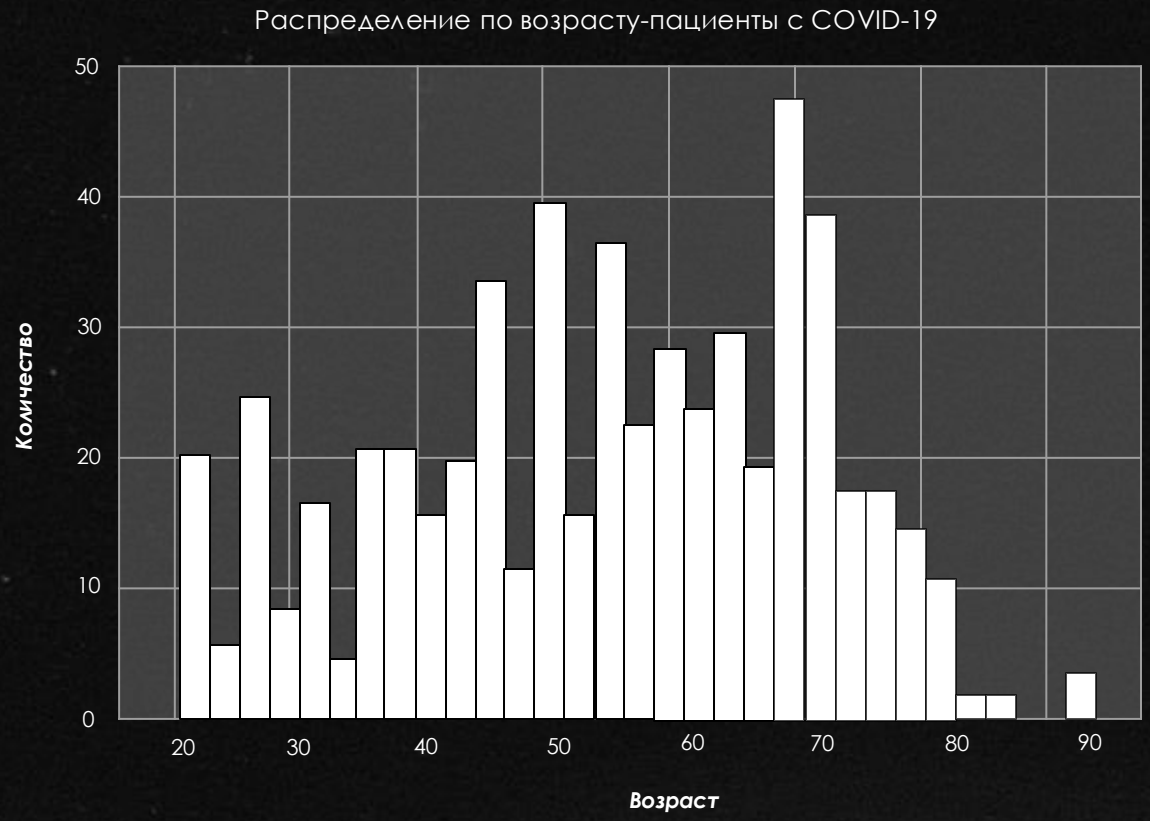
Распределение диагнозов



Примеры визуализаций:

Гистограмма возраста





Аналитический отчет по проекту

Принятые решения

- Использование Spark SQL для масштабируемой аналитики
- Выбор Hive Metastore для управления таблицами
- Партиционирование и бакетирование для ускорения аналитики
- Использование Docker Docker-compose для простоты развертывания

Интерпретация результатов

- Высокий процент COVID-19 у мужчин старшего возраста
- Выявлены явные показатели COVID-19

Рекомендации

➢ Улучшить качество данных:

- Заполнить пропуски
- Унифицировать диагнозы ещё глубже

➢ Рассмотреть более сложные ML-модели:

- Классификация по признакам
- Прогнозирование исхода болезни