Разработка системы анализа медицинских изображений для эпидемиологическ ого мониторинга COVID-19

Архитектура системы

Была развернута система из нескольких контейнеров на базе Docker образов:

1. bde2020/hadoop-datanode:2.0.0-hadoop2.7.4-java8
   для namenode и datanode
2. apache/spark:latest
   для spark-master и spark-worker
3. jupyter/pyspark-notebook
   для jupyter контейнера

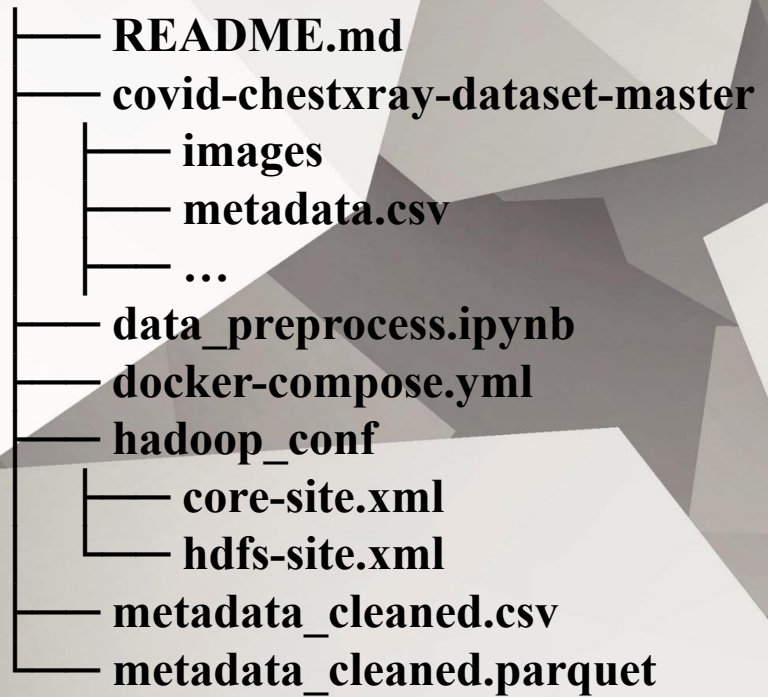
Каждый элемент подключен к единой сети

Хранение и доступ к данным обеспечены HDFS, поверх которого настраивается Hive слой SQL-аналитики для работы с таблицами и метаданными. Spark в свою очередь является вычислительным движком в для обработки данных, который подключается к Hive метаданным и HDFS, загружает данные из Hive таблиц, обрабатывает и отдаёт результаты.

Вводные данные

За основу был взят реальный датасет медицинских изображений, включающий в себя задачи, с которыми сталкиваются специалисты в области CV и аналитики больших данных.

Структура проекта



Структура данных

1. Изображения: рентгеновские снимки (PNG/JPEG) пациентов с COVID-19, пневмонией и другими патологиями.
2. Метаданные: файл metadata.csv с полями:
   - patientid (идентификатор пациента);
   - age (возраст, есть не для всех записей);
   - sex (пол, не для всех записей);
   - finding (диагноз: COVID-19, SARS, Pneumonia и другие);
   - view (проекция снимка: PA, AP и так далее);
   - date (дата исследования).

Ключевые выводы

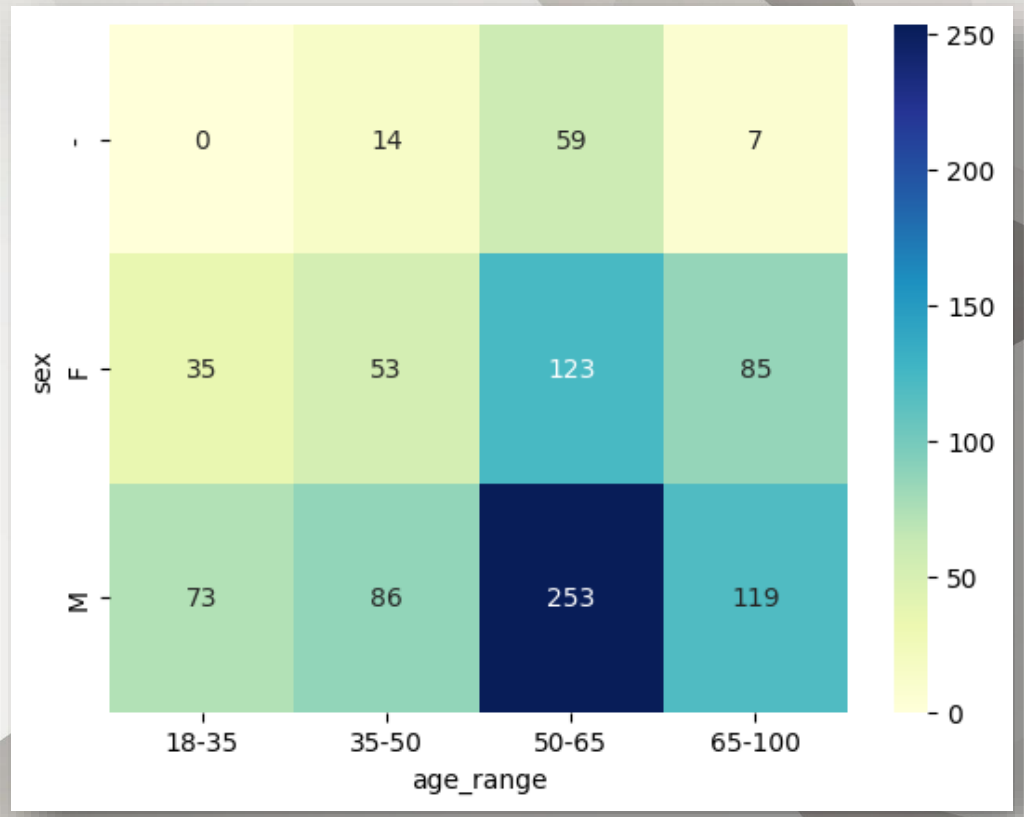
Проблемы и препроцессинг данных

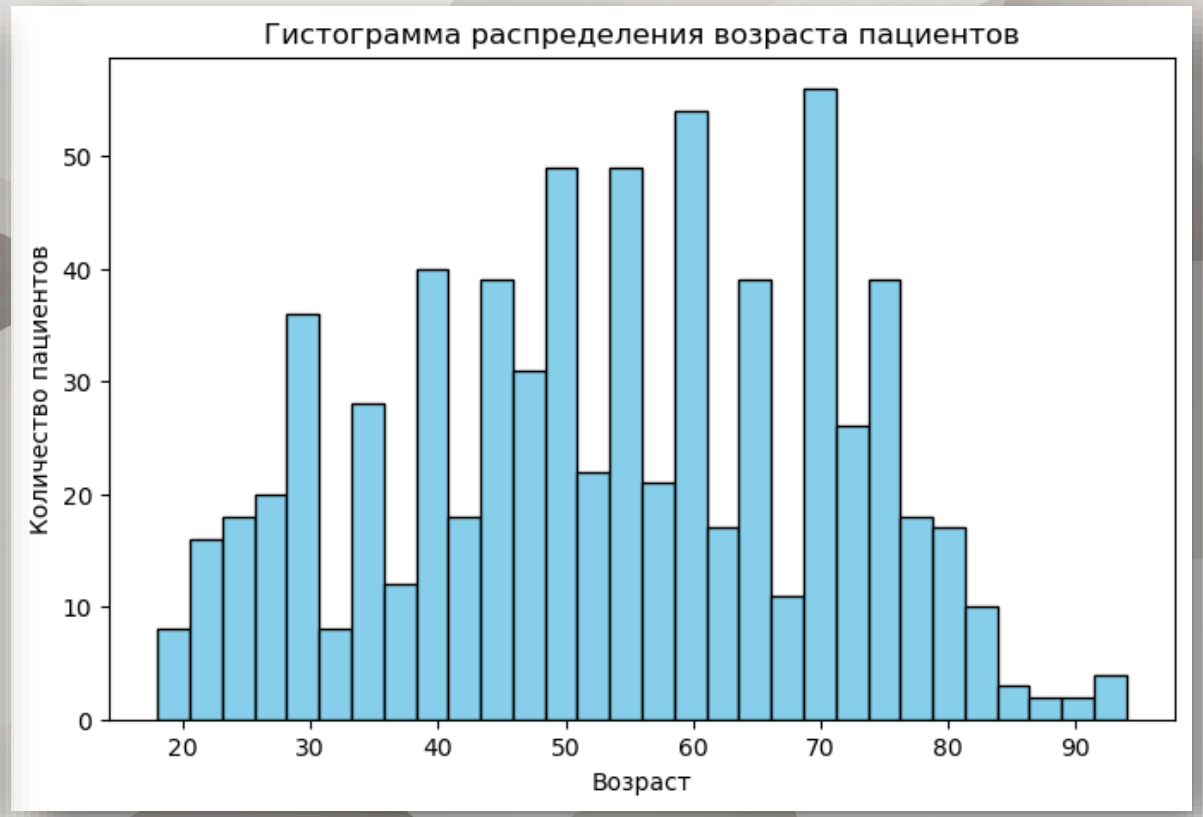
- Пропущенные значения обнаружились во всех фичах
- Грубые выбросы в offset (<0), temperature (>40), pO2_saturation (>100%)
- Колонки view, finding, date и location были унифицированы
- Созданы новые фичи age_range и temperature_range для более глубокой аналитики. На них в последующем также можно строить предиктивные модели

Статистика на текущих данных

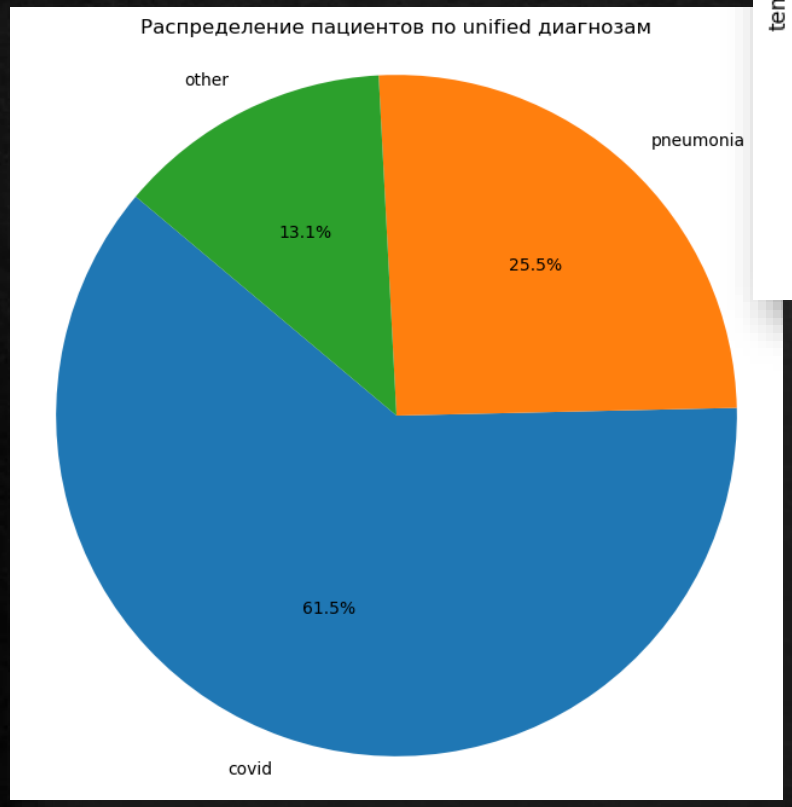
- Более 70% пациентов старше 50 лет
- Чаще заболевали люди в возрасте 50-65 лет, из которых больше половины – мужчины
- У подавляющего большинства пациентов температура была умеренно высокой [37.6; 38.5) во время болезни
- Из трех категорий заболеваний (pneumonia, covid, other) пациенты имели covid в ~60% случаев

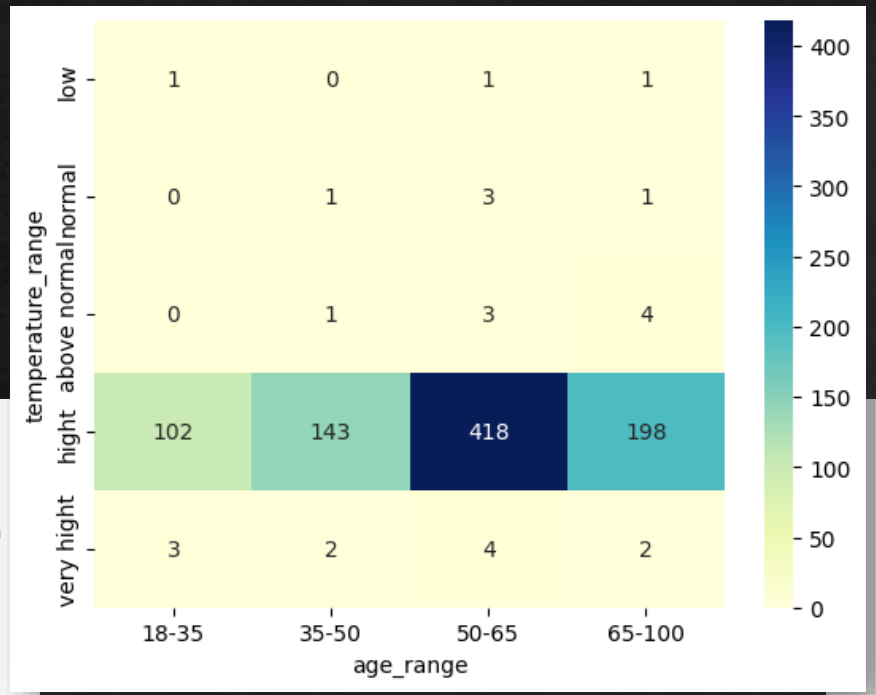
Визуализация

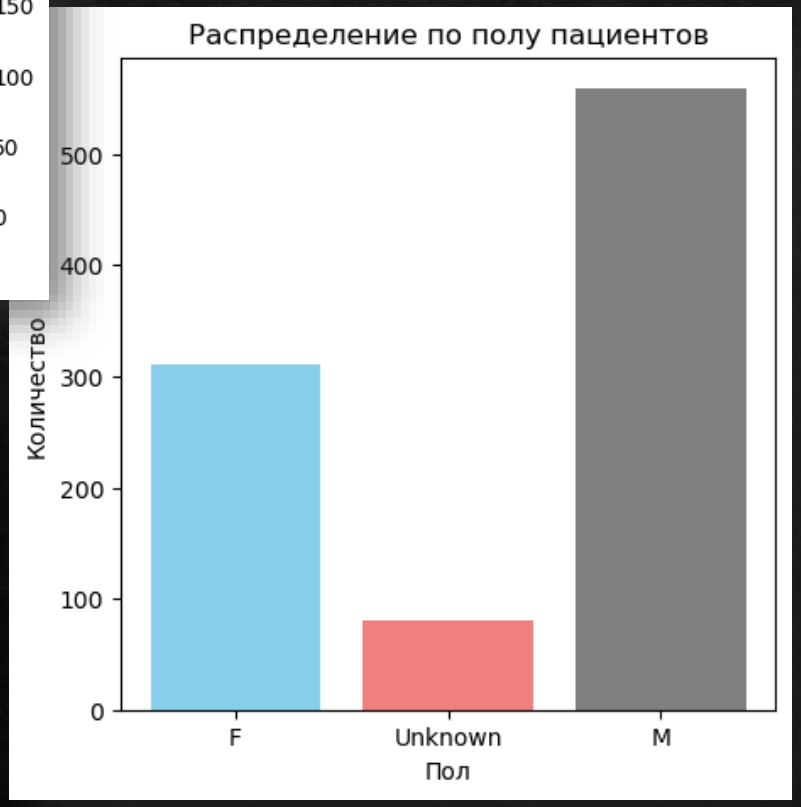




Визуализация







Аналитический отчет

Резюмируя, можно заключить, что болезни больше подвержено мужское население старше 50 лет. Однако, судя по данным, к такому перевесу нужно относиться с осторожностью, поскольку в представленной выборке мужчин в целом больше (около 60%). Неполнота, как метаданных, так и выборки в целом – основная трудность.

В рамках доработки препроцессинга имеет смысл использовать ML-подходы для улучшенного заполнения пропусков и корректировки выбросов. В зависимости от дальнейшей задачи генерация дополнительных фичей также не будет лишней.

Для улучшения развернутой системы можно внедрить bash-скрипты с начальными настройками с запуском во время подъема контейнеров в автоматическом режиме (не создавать папки и не добавлять права мануально), а также добавить индексы в Hive.