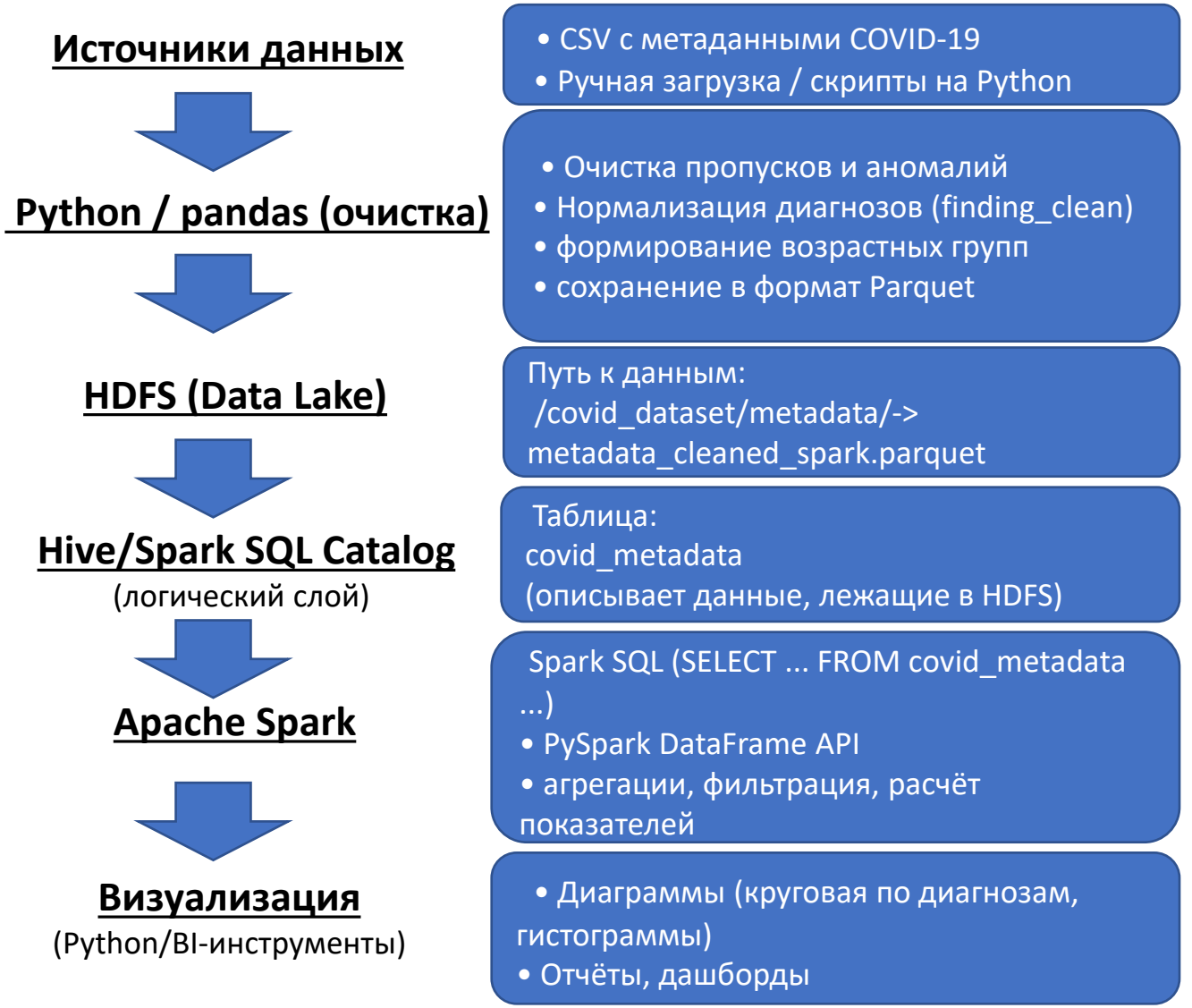
Система мониторинга данных пациентов с COVID-19

Проект по дисциплине «Базы данных для компьютерного зрения»
Автор: Пластуненко Л.Ю.

1. Архитектура системы мониторинга

- В проекте реализована прототипная архитектура системы обработки медицинских метаданных:
- **HDFS** — хранение очищенных данных в формате Parquet;
- **(Hive)** — слой метаданных и логических таблиц над HDFS;
- **Apache Spark** — выполнение SQL-запросов и аналитики;
- **Визуализация** — построение графиков и отчётов (Python/BI-инструменты).



2. Организация и оптимизация хранения (HDFS/Hive)

Для эффективной работы аналитических запросов в проекте использованы:

- Хранение данных в **колоночном формате Parquet** в HDFS
- Логическая разбивка по каталогам (*/covid_dataset/metadata,/covid_dataset/analytics*);
- Концептуально возможно применение **партиционирование** данных по ключевым признакам (дате и/или диагнозу).

Пример структуры:
/covid_dataset/metadata:
/covid_dataset/metadata/date=2020-01-22/
/covid_dataset/metadata/date=2020-01-23/

В промышленной системе таблица covid_metadata в Hive могла бы быть партиционирована по:

- дате исследования (PARTITIONED BY (date)
- по диагнозу (PARTITIONED BY (finding_clean).

Примечание: Это позволило бы:
•ускорить выборки по датам/диагнозам за счёт чтения только нужных партиций;
•снизить нагрузку на кластер при выполнении тяжёлых запросов.

3. Ключевая статистика и проблемы данных

Запрос

```
SELECT finding_clean, COUNT(*) AS cnt
FROM covid_metadata
GROUP BY finding_clean
ORDER BY cnt DESC
```

Результат

- COVID-19 — 584 записей
- Pneumonia — 226
- Unknown — 84
- No finding — 22
- Tuberculosis — 18
- SARS — 16

Вывод

•Основной диагноз в выборке — **COVID-19** (584 записей), что составляет примерно **две трети** всех случаев.
•На втором месте — **пневмония** (226 записей), остальные диагнозы (туберкулёз, SARS, отсутствие находок) представлены значительно реже.

```
SELECT age_category, finding_clean, COUNT(*) AS cnt
FROM covid_metadata
GROUP BY age_category, finding_clean
ORDER BY age_category, cnt DESC
```

finding_clean	young	middle	old
COVID-19	33	341	210
Pneumonia	17	143	66
Unknown	0	84	0
No finding	3	11	8
Tuberculosis	7	9	2
SARS	6	3	7

•Наибольшее число случаев COVID-19 приходится на категорию **«middle»** — 341 запись.
•В категории **«old»** также наблюдается значительное количество COVID-19 — 210 записей.
•В группе **«young»** случаев COVID-19 заметно меньше (33 записи), что подчёркивает смещение выборки в сторону пациентов среднего и старшего возраста.

```
SELECT finding_clean,
    ROUND(AVG(age), 1) AS avg_age,
    COUNT(*) AS cnt
FROM covid_metadata
GROUP BY finding_clean
HAVING cnt >= 10
ORDER BY avg_age
```

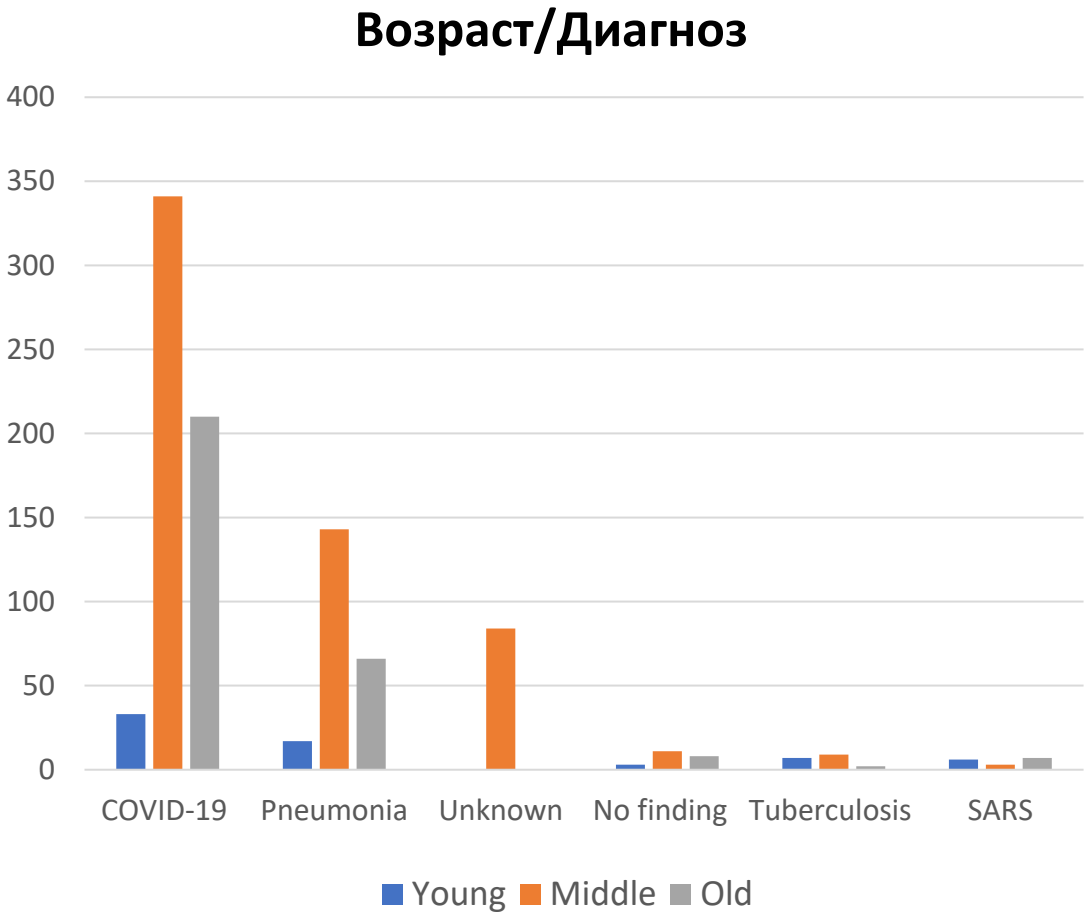
•Tuberculosis — 43.1 (18 записей)
•Pneumonia — 49.2 (226)
•SARS — 50.1 (16)
•No finding — 52.5 (22)
•Unknown — 54.0 (84)
•COVID-19 — 55.8 (584)

•Средний возраст пациентов с **COVID-19** — около **56 лет**, что является максимальным значением среди рассмотренных диагнозов.
•Пациенты с **обычной пневмонией** в среднем моложе (около 49 лет), а с **туберкулёзом** — ещё моложе (около 43 лет).
•Таким образом, в имеющемся датасете пациенты с COVID-19 представлены преимущественно в **старших возрастных группах**.

4. Проблемы качества данных

Проблема	Признак	Вывод
1.Массовые пропуски в датах	По агрегации по датам: SELECT date, COUNT(*) FROM covid_df GROUP BY date ORDER BY date я получил: **date = NULL - 528 случаев для COVID-19**.	Для значительной части записей с COVID-19 отсутствует дата исследования: **528 случаев имеют NULL вместо даты**. Это существенно ограничивает возможности временного анализа и построения корректной динамики.
2.Пропуски в клинических признаках	По **printSchema()** видно, что многие показатели (temperature, pO2_saturation, leukocyte_count и др.) в первых строках **NULL**, и по опыту этого датасета — заполнены очень редко.	Большинство клинических полей (температура, насыщение кислородом, лабораторные показатели крови) заполнены только для части пациентов, что препятствует использованию этих признаков в широкомасштабном анализе.
3. Неоднородность и «Unknown» в диагнозах	Из первой таблицы видно **finding_clean = 'Unknown' - 84 записи**.	В ряде случаев диагноз указан неявно или противоречиво, что после нормализации приводит **к категории Unknown (84 записи).** Это отражает исходную неоднородность и неполноту текстовых описаний находок.
4. Повторяющиеся пациенты/серии снимков	Из covid_df.show(5) видно, что **один patientid встречается несколько раз с разными датами**: patientid=2, age=65, несколько дат (2020-01-22, 25, 27, 28)	Один и тот же пациент может встречаться в данных несколько раз (серии исследований в разные даты). Это важно учитывать при интерпретации частот (число записей не равно числу уникальных пациентов).

5. Визуализация результатов



6. Аналитический отчёт(выводы) по результатам проекта

Использован стек **HDFS + Parquet + Spark** в Docker для демонстрации реальной Big Data-архитектуры и возможности масштабирования.

Основной диагноз - **COVID-19** (584 записей), далее **пневмония** (226), остальные диагнозы существенно реже.

Случаи COVID-19 сосредоточены в **среднем и старшем возрасте** (категории *middle* и *old*), средний возраст пациентов с COVID-19 ≈ **56 лет**, что выше, чем при пневмонии и туберкулёзе.

Выборка ориентирована на более возрастных пациентов, что нужно учитывать при интерпретации результатов

Рекомендации по улучшению:

Повысить полноту заполнения ключевых полей (возраст, дата, клинические параметры) и доработать нормализацию диагнозов.

Внедрить полноценный слой **Hive/метаданных** с партиционированием по дате/диагнозу для ускорения запросов.

На основе агрегатов построить дашборды и в перспективе использовать данные для моделей машинного обучения.