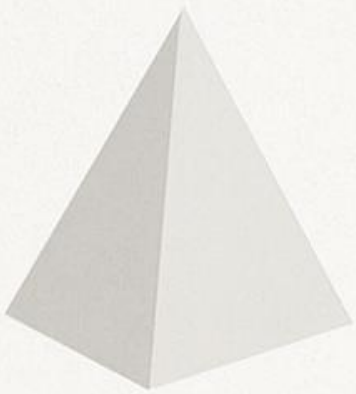
Разработка системы анализа медицинских изображений

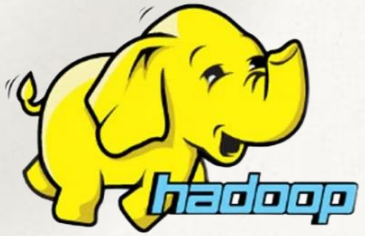
Для эпидемиологического мониторинга COVID-19

Автор проекта: Холин Никита



Архитектура системы

- **HDFS** — распределённое хранилище Hadoop, разбивает файлы на блоки и хранит их на разных серверах для масштабируемости и отказоустойчивости.
- **Hive Metastore DB** — база метаданных Hive, хранит информацию о таблицах и схемах, даёт доступ к данным в HDFS через SQL-подобные запросы (Hive, Spark).
- **Spark (PySpark)** — движок для быстрой обработки больших данных: трансформации, агрегации, витрины данных, ML.
- **Visualization (Jupyter)** — ноутбук для анализа и визуализации: Python-код, графики, таблицы, интерактивные дашборды по данным из Spark.



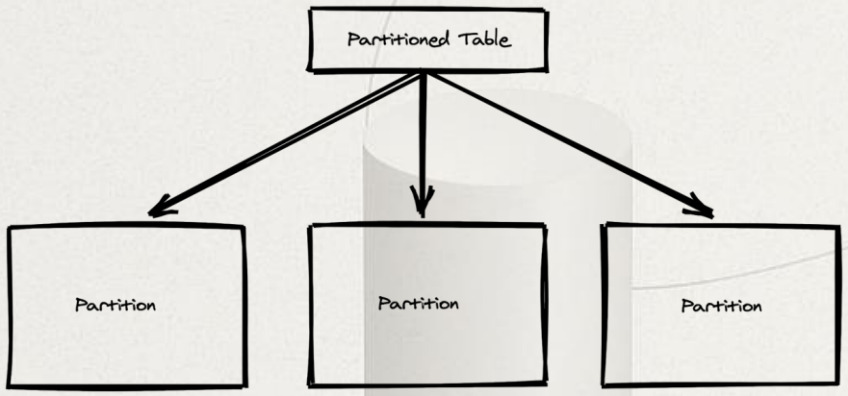
Оптимизации производительности

Партиционирование:

1. Данные делятся на папки по признакам:
finding - например, тип находки или диагноза
age_group - возрастная группа

2. Это помогает быстро отсеивать ненужные данные при запросах.

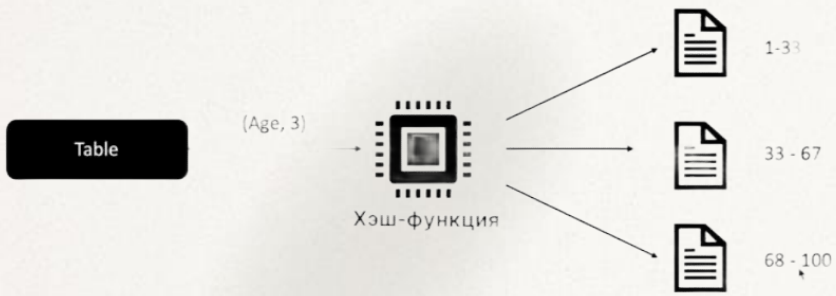
```
df.write \
    .mode("overwrite") \
    .partitionBy("finding", "age_group") \
    .bucketBy(8, "sex", "view") \
    .sortBy("age") \
    .format("parquet") \
    .saveAsTable("covid_metadata_partitioned")
```



Бакетирование (Bucketing)

1. Данные внутри партиций разбиваются на фиксированное число бакетов по столбцам:
sex — пол
view — например, тип снимка или угол обзора

2. Ускоряет операции JOIN и фильтрацию — данные с одним значением будут в одних и тех же бакетах.

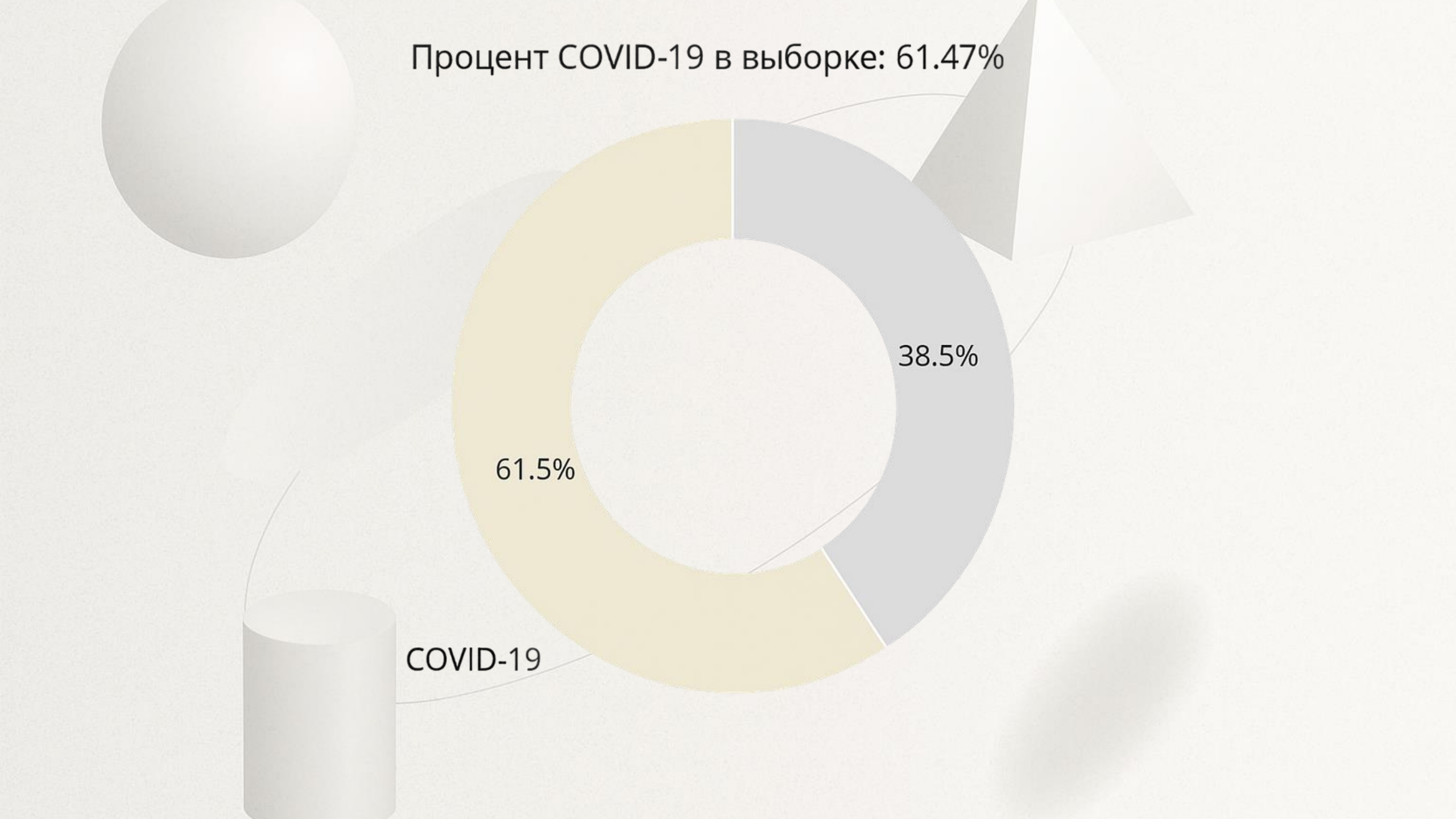


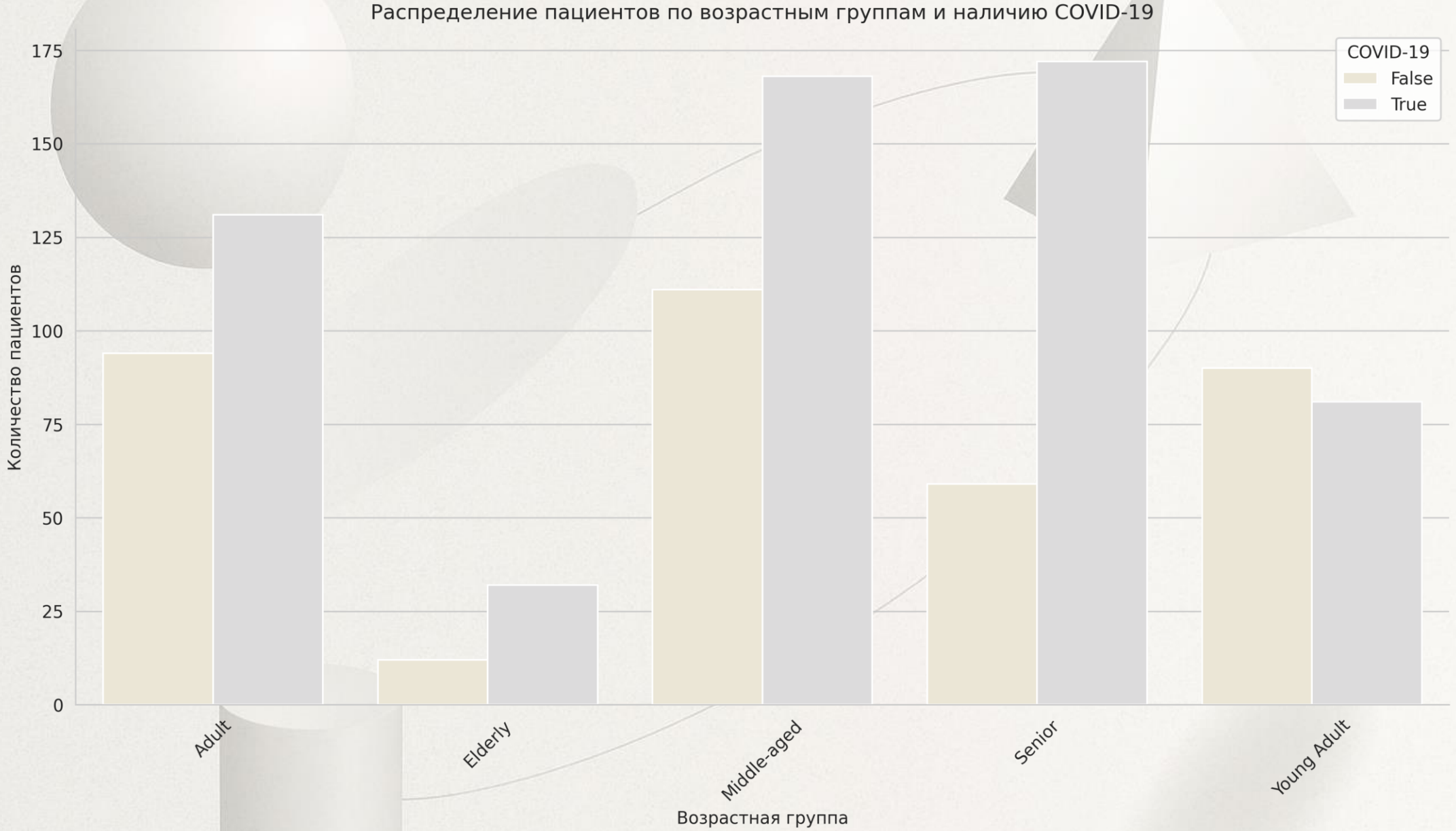
Ключевые выводы

- Явный показатель понижение сатурации у пациенотов с COVID-19

- Пациенты с COVID-19 в основном сосредоточены в возрастных группах "Young Adult" и "Adult", что может указывать на более высокую восприимчивость или диагностику в этих возрастах

- Доля COVID-19 среди всех диагнозов: около 61.5%


Процент COVID-19 в выборке: 61.47%
38.5%
61.5%
COVID-19


Распределение пациентов по возрастным группам и наличию COVID-19
COVID-19
False
True
Количество пациентов
175
150
125
100
75
50
25
0
Adult
Elderly
Middle-aged
Senior
Young Adult
Возрастная группа