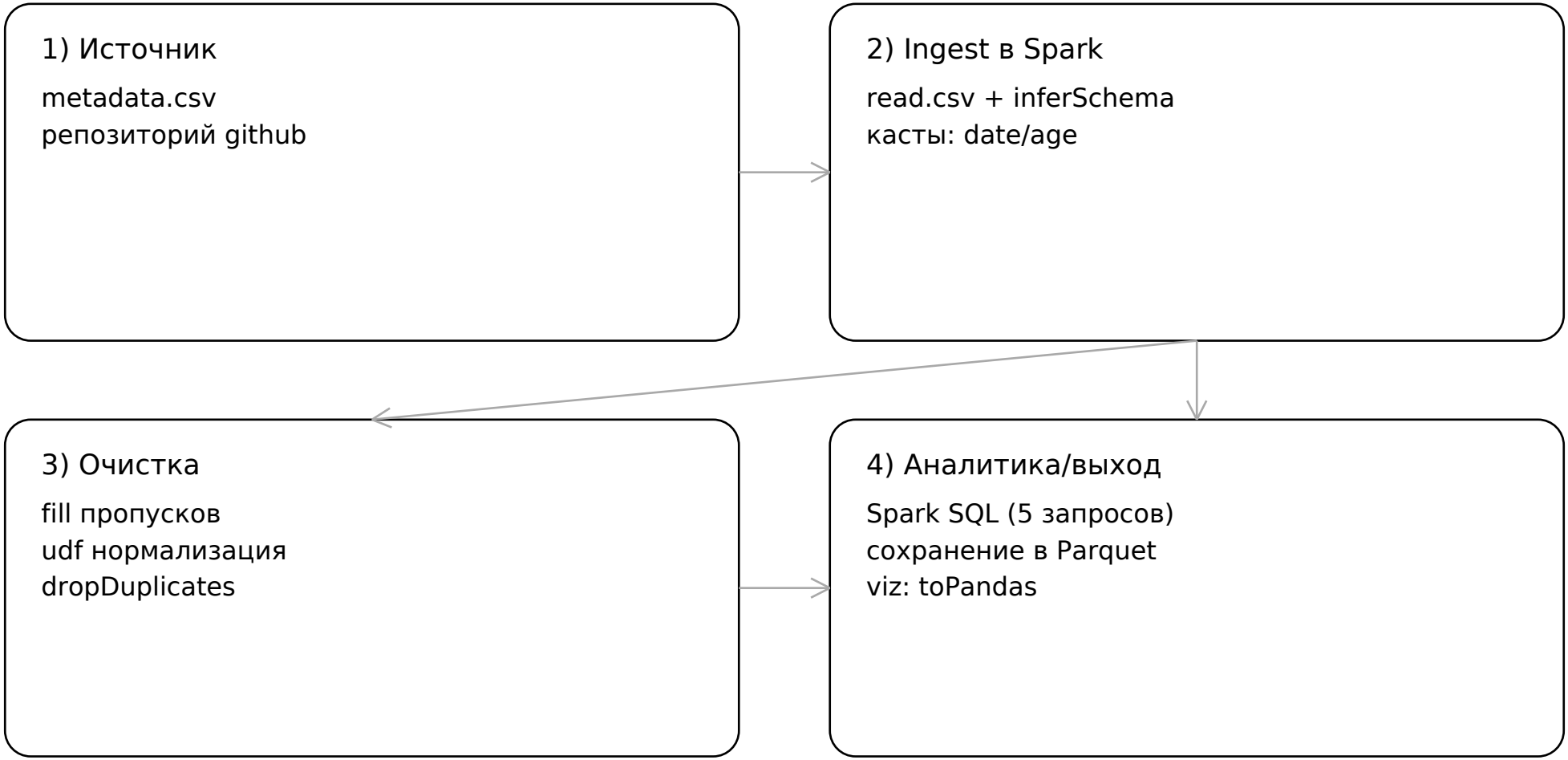
Архитектура решения

Цель: мониторинг COVID-19 по метаданным рентген-снимков (Spark + SQL + визуализации).



- витрина: parquet, partitionBy(finding_std)
- графики: агрегаты в spark -> pandas -> matplotlib/seaborn

Ключевые статистики и качество данных

- finding привёл к 4 классам: COVID-19 / Pneumonia / Normal / Other
- age: медиана (approxQuantile), sex: мода (самый частый)
- аномалии: возраст <0 и >110 убрал; даты парсил to_date (часть стала null)
- дубликаты: patientid + study_date + view + finding_std

Точные цифры/таблицы считаются в ноутбуке (зависят от версии metadata.csv).

Визуализации и выводы

- круговая диаграмма: распределение диагнозов
- столбчатая: распределение по возрастным группам
- линейный график: тренд исследований по месяцам
- heatmap: проекция снимка (view) vs диагноз (finding)

- данные неидеальные: пропуски и перекос по времени (датасет собирался вручную)
- для прод-решения нужен слой data quality + правила валидации на входе