Национальный исследовательский
Томский государственный университет

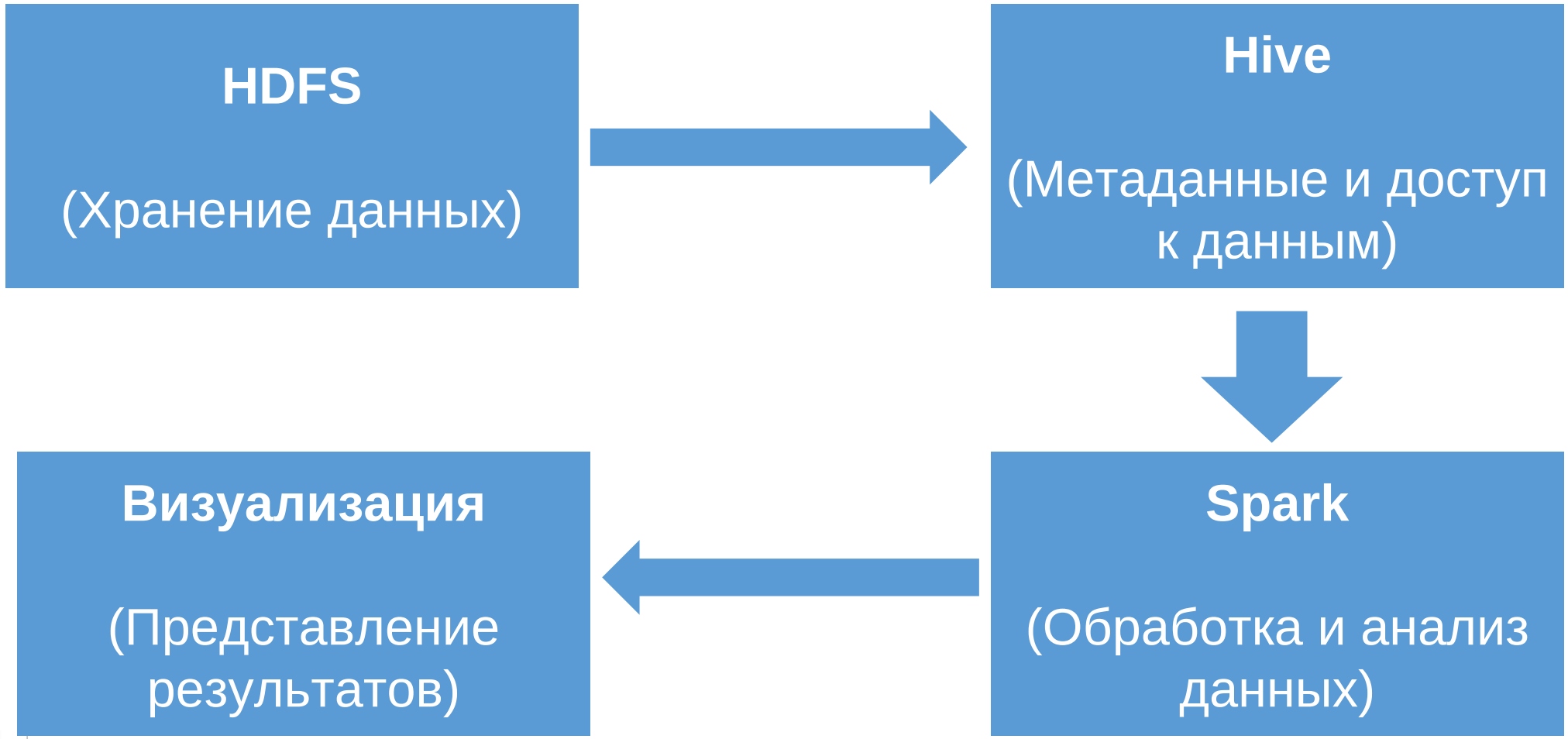
Курс: Базы данных для компьютерного зрения

Проект: Разработка системы анализа медицинских изображений для эпидемиологичесого мониторинга COVID-19

Студент: Иванов Кирилл

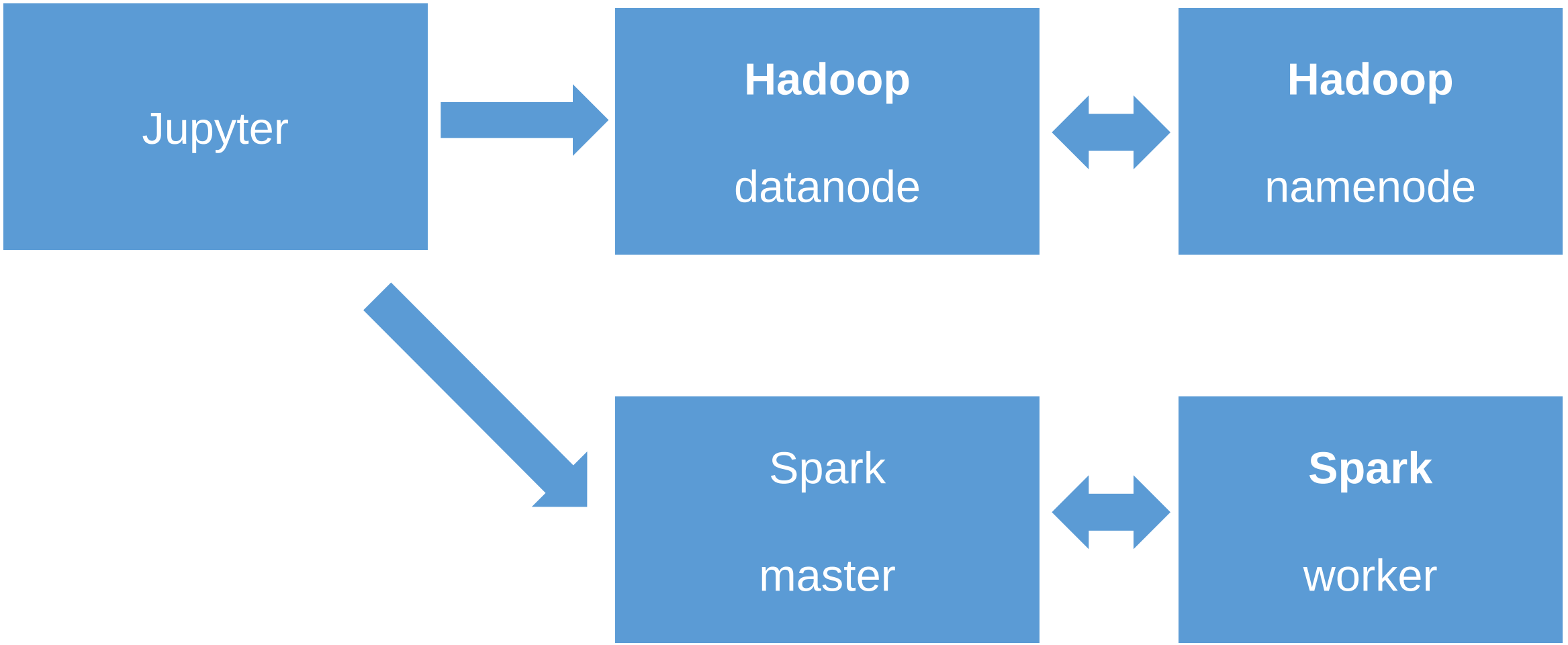
Архитектура системы

Потоки данных



Архитектура системы

Модули



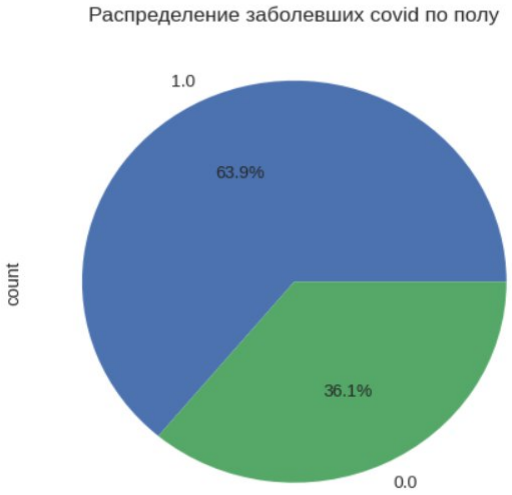
Ключевые выводы

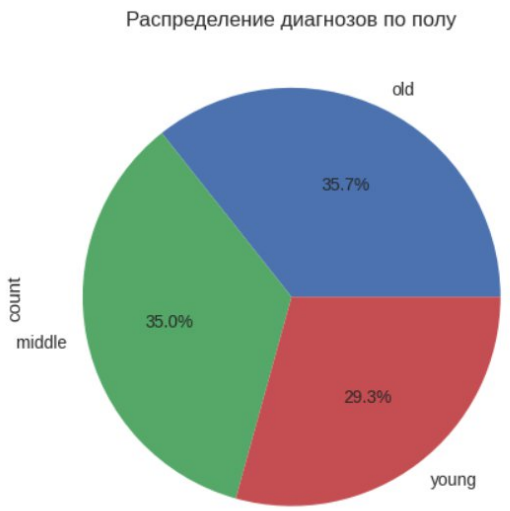
- 70% всех диагнозов - COVID
- 64% заболевших - мужчины
- распределение по категориям заболевших приблизительно одинаковое: 30-35% в группах «до 30», «30-60» и «60+»

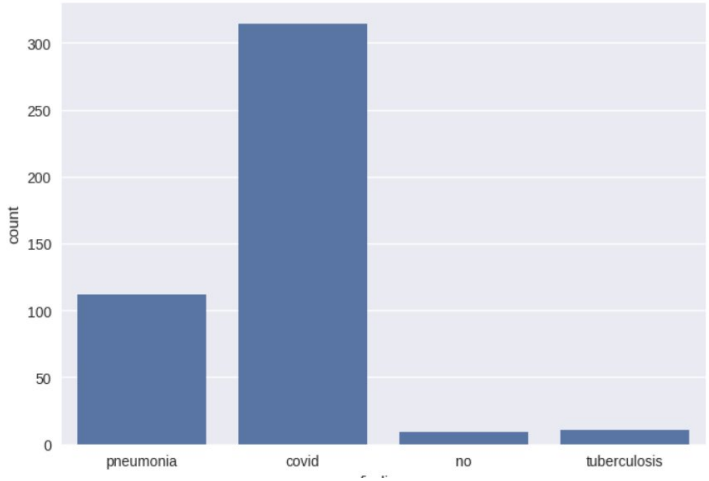
Проблемы данных

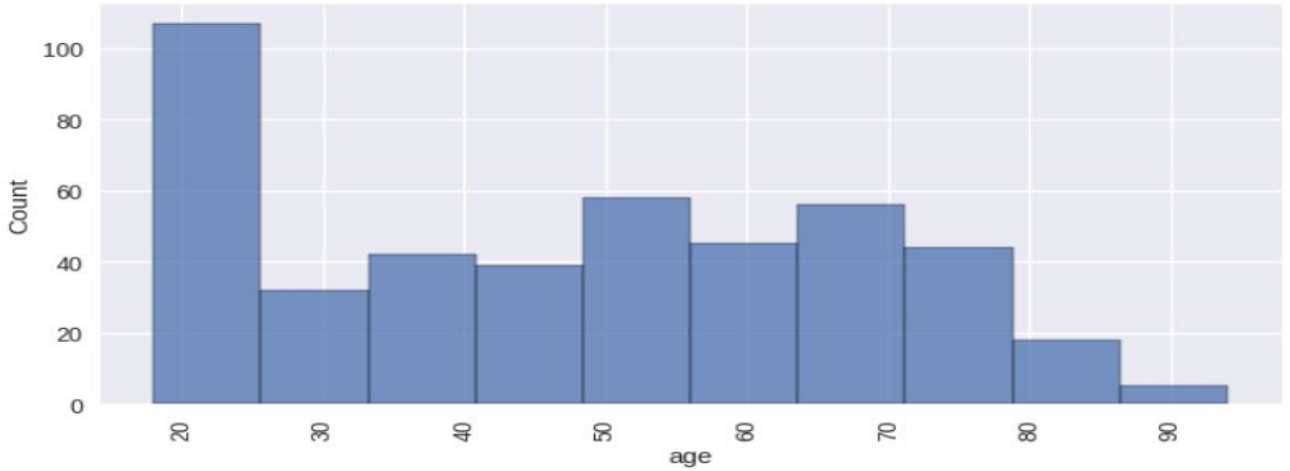
- большое количество пропусков по анализам данных, по некоторым параметрам данные отсутствуют в 90%
- в ключевых объектах (возраст, пол) также большое количество пропусков - около 25%
- некоторые файлы со снимками отсутствуют в реестре картинок

Статистика в данных









Аналитический отчет

При сохранении данных в Hive было применено:

- **партиционирование** по полю *finding* (диагноз) - механизм, позволяющий физически отделить группы строк в таблице по значению, что позволит при анализе одного заболевания:
  - повысить скорость выполнения запросов (обращение только к конкретной папке)
  - эффективное управление данными
  - экономия ресурсов
  - организация хранения
- **бакетирование** по полям *age* (возраст) и *sex* (пол) - улучшит работу с join -операциями и сортировкой данных, позволит эффективно распределить задачи между рабочими узлами