APACHE SPARK: ETL, SQL аналитика и машинное обучение

С помощью PySpark DataFrame API реализован комплексный пайплайн для очистки, стандартизации и обогащения исходных методанных. Это включает обработку пропусков, удаление дубликатов, унификацию тестовых полей и создание новых аналитических признаков.

Используя Spark SQL, выполнены сложные аналитические запросы к оптимизированным данным в HDFS. Проводился глубокий анализ статистических распределений, применялись агрегации, фильтрации и оконные фунуции для выявления закономерностей в данных о COVID-19.

Spark MLlib использовалась для подготовки данных и моделированию и ообучению модели: создание прогностической модели для определения наличия COVID-19 на основе обработанных метаданных.

Модель Random Forest Classifier показала наилучшие результаты, достигнув AUC = 0.9228, что свидетельствует о ее высокой предсказательной способности.