Схема работы

В основе проекта лежит Apache Hadoop HDFS - распределенная файловая система, обеспечивающая надежное и масштабируемое хранение больших объемов данных. Для эффективной работы с метаданными и ускорения аналитических запросов реализовано партиционирование данных непосредственно в HDFS.

Исходные методанные были загружены в HDFS. После ETL-обработки в Spark, очищенные обогащенные данные были сохранены обратно в HDFS в формате Parquet.

Мы применили партиционирование по полям year (год) и month (месяц), что позволяет Spark - запросам с фильтрацией по дате считывать только необходимые подмножества данных, значительно повышая производительность.