УДК 004.67

РАЗРАБОТКА ХРАНИЛИЩА ДАННЫХ ДЛЯ РЕШЕНИЯ ЗАДАЧ БИОИНФОРМАТИКИ

Сазыкин Г.А. (ИТМО),
Научный руководитель – кандидат физико-математических наук, доцент
Токман М.А. (ИТМО)

Введение

Современная биоинформатика представляет собой науку, важной частью которой является обработка большого объема различных структурированных и неструктурированных данных: последовательности ДНК, определенные в результате секвенирования, данные об экспрессии генов, информация о вариациях, полученных в результате выравнивания. Такие данные необходимо грамотно хранить, иметь возможность быстро находить нужные данные и настраивать ETL-процессы для их преобразования и дальнейших исследований. В рамках данной работы требуется разработать хранилище данных, включающее в себя компоненты для объектного хранения файлов различных форматов, аналитическую СУБД для написания запросов на языке SQL, СУБД для хранения и накопления метаданных, а также сравнить классический подход к обработке данных биоинформатики с помощью специальных утилит и подход, который предоставляет реализованное хранилище данных.

Основная часть

В работе рассматривается подход к хранению и обработке данных о вариациях последовательности генов - файлов формата VCF, содержащие также информацию о генотипах образцов и представленных в сжатом виде; на основе данных о вариантах доступно создание файла, представляющий собой индекс для осуществления быстрого поиска. Источником информации является проект “1000 genomes”, содержащий открытые данные о генетических вариациях человека [1]. Для применения классических подходов к обработке таких данных может быть осуществлен перевод представленных файлов в форматы, которые обычно используются для обработки Big Data – например, колоночный формат Apache Parquet или Apache Iceberg; перевод файлов в подходящий формат является отдельным ETL-процессом, настроенным с помощью инструмента Airflow. Такой подход позволяет читать данные из объектного хранилища в аналитическую СУБД для написания пользовательских SQL-запросов, в случае необходимости использовать распределенные системы с возможностью масштабирования. В качестве объектного хранилища был выбран инструмент S3 Minio, в качестве базы данных для аналитических запросов – колоночная СУБД ClickHouse; информация о наличии, изменении, удалении файлов в объектном хранилище содержится в СУБД Postgres. В рамках работы также сравниваются результаты, удобство и время выполнения запросов к данным о вариантах, полученные с помощью утилиты bcftools и в результате выполнения запросов на языке SQL.

Выводы

Разработка хранилища предоставляет широкие возможности для исследования данных о вариантах, дальнейшего масштабирования [2], настройки пользовательских процессов по преобразованию данных, а также визуализации. Предложенный подход позволит применять распространённые инструменты для обработки информации и исследований в биоинформатике со стороны аналитиков и инженеров данных.

Список использованных источников

1. The International Genome Sample Resource [Электронный ресурс] // EMBL-EBI. Режим доступа: https://www.internationalgenome.org (дата обращения: 26.02.2026).
2. Клеппман М. Высоконагруженные приложения. Программирование, масштабирование, поддержка // Астана: Спринт Бук, 2025. – 640 с.: ил. – (Серия “Бестселлеры O’Reilly”)