УДК 004.654

Обзор методов и решений эффективного хранения и обработки гетерогенных данных структурированных и полуструктурированных данных

Першинов А.В. (Университет ИТМО), **Ковальчук М.А.** (Университет ИТМО),
Научный руководитель – доцент, кандидат технических наук Насонов Д.А.
(Университет ИТМО)

Введение. Объемы создаваемых, фиксируемых и воспроизводимых данных продолжают быстро расти: мгновенные данные, малые данные, большие данные, данные в реальном времени. С каждым годом объём данных увеличивается в геометрической прогрессии.

С увеличением общего количества данных так же увеличивается и их сложность. Данные, полученные из разных источников, отражающие разные процессы или свойства называются разнородными, или же, гетерогенными.

Эффективные методы хранения и обработки гетерогенных данных имеют решающее значение в создании искусственного интеллекта. Гетерогенность данных обуславливается различной структурой. Соответственно, для хранения каждого вида таких данных характерна определённая технология хранения данных [1].

Для дальнейшего развития и эксплуатации сильного искусственного интеллекта необходимо разработать эффективные методы хранения и обработки сложных, гетерогенных данных.

Основная часть. Существует множество различных методов для хранения различных данных. При работе с неструктурированными данными используются NoSQL системы. При работе с гетерогенными данными используются различные технологии, в зависимости от вида данных [1].

1) Геостационарные данные хранятся в PostgreSQL/PostGIS [2].
2) Для хранения графовых структур данных используются Neo4j, ArangoDB, Dgraph [3].
3) Документы формата JSON хранятся с помощью MongoDB [3].
4) Хранение временных рядов осуществляется с помощью InfluxDB, Postgres+TimescaleDB, RedisTimeSeries [3].
5) Для финансовых данных подходят системы Bloomberg, Statista, Eikon [3].

В прикладных задачах встречаются такие наборы гетерогенных данных, которые могут содержать в себе несколько различных источников данных. Для корректной работы таких систем необходимо создать эффективную систему обработки данных, основанную на оптимальном наборе технологий, кроме того, эти технологии должны быть связаны между собой.

В данный момент нет универсального инструмента для развёртывания системы хранения и обработки гетерогенных данных. Создание такой системы поможет разработчикам более эффективно создавать

Систематизация различных подходов к хранению гетерогенных данных необходима для создания представления о наиболее удачном выборе технологий для различных комбинаций видов данных. Имея полную картину того, какими способами нужно хранить разнородные данные можно, на её основе, создать единую систему хранения данных, основанную на наиболее подходящих технологий их хранения.

Выводы. Проблема обработки гетерогенных данных остро встает в задачах больших данных. Создание единого интерфейса управления системами хранения гетерогенных данных ускорит процесс разработки сложных информационных систем.

Список использованных источников:

1. D. Ganesh Chandra, “BASE analysis of NoSQL database,” *Future Generation Computer Systems*, vol. 52, 2015, doi: 10.1016/j.future.2015.05.003.
2. A. F. Sveen, “Efficient storage of heterogeneous geospatial data in spatial databases,” *J Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0262-8.
3. L. Wang, “Heterogeneous Data and Big Data Analytics,” *Automatic Control and Information Sciences*, vol. 3, no. 1, 2017, doi: 10.12691/acis-3-1-3.
4. N. I. Pikuleva, A. S. Khafizova, and R. A. Khakov, “Query Processing for Heterogeneous Data Sources in SQL Generator,” in *2020 International Multi-Conference on Industrial Engineering and Modern Technologies, FarEastCon 2020*, 2020. doi: 10.1109/FarEastCon50210.2020.9271101.
5. A. Chakrabarti and M. Jayapal, “Data transformation methodologies between heterogeneous data stores: A comparative study,” in *DATA 2017 - Proceedings of the 6th International Conference on Data Science, Technology and Applications*, 2017. doi: 10.5220/0006438802410248.

Першинов А.В. (автор) Подпись

Ковальчук М.А. (автор) Подпись

Насонов Д.А. (научный руководитель) Подпись