УДК 659.118

АРХИТЕКТУРА СОВРЕМЕННЫХ СИСТЕМ УПРАВЛЕНИЯ НАУЧНЫМ ЭКСПЕРИМЕНТОМ В ОБЛАСТИ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

Кайда Анастасия Юрьевна[1],
ayk13@tpu.ru

Савельев Алексей Олегович[1],
sava@tpu.ru

[1] Национальный исследовательский Томский политехнический университет,
Россия, 634050, г. Томск, пр. Ленина, 30.

История развития систем управления экспериментом насчитывает порядка 25 лет – от небольших прототипов десктопных приложений до масштабируемых распределенных систем. EMS-системы становятся все более популярными, а их устройство – все более сложным. Новым этапом развития EMS-систем послужило начало «эры больших данных», чьи обработка, анализ и хранение происходят в условиях специальной экосистемы. Экосистема больших данных подразумевает наличие новых элементов, которые должны быть учтены при проектировании EMS-системы для поддержки полного жизненного цикла данных. Одним из ключевых вызовов является обобщение тех EMS-систем, которые возникали и изменялись по мере развития сопутствующих технологий, при этом на текущий момент не существует единого паттерна проектирования таких систем ввиду специфики каждого конкретного проекта, под который создается подобная система. В данной работе представлена концептуальная высокоуровневая архитектура EMS-системы как пример унифицированного паттерна, подходящего для экосистемы больших данных. Важно отметить, что предлагаемый паттерн не подразумевает работу с использованием грид-систем.

Ключевые слова: *Big Data, система управления экспериментом, жизненный цикл данных, обработка данных, базы знаний.*

Введение

Системы управления экспериментом

Система управления экспериментом (Experiment Management System, EMS) [1] – система, обеспечивающая сопровождение научного эксперимента на протяжении всех этапов его проведения. Под сопровождением подразумевается обеспечение поддержки полного жизненного цикла данных. EMS-система подразумевает под собой сложный программный комплекс, являющийся связующим звеном между исследователями и данными об эксперименте. Также EMS-системы позволяют планировать и организовывать эксперимент, управлять данными и анализировать полученные результаты. Первые EMS-системы появились задолго до «эры больших данных», однако, несмотря на длительный период своего существования, перед такими системами стоит ряд вызовов, связанных с большими данными:

- экспоненциальное увеличение объема данных;
- расширение перечня используемых форматов (гетерогенность данных);
- использование распределенных гетерогенных вычислительных систем;
- расширение сотрудничества исследователей посредством создания новых научных коллабораций и междисциплинарных исследовательских групп [2].

Жизненный цикл данных

При упоминании решения прикладных задач с использованием больших данных подразумевается одновременный сбор и обработка большого количества различных форматов и типов – гетерогенных данных. Источник данных становится полезным только после того, как выстроена цепочка операций для извлечения информации из данных этого источника начиная от сбора данных и завершая их загрузкой в конечное хранилище или уничтожением. Выделяют следующие основные этапы:

Извлечение данных. Начальная стадия жизненного цикла данных. Данные могут быть предварительно сохранены и извлечены по запросу из источника в цифровой инфраструктуре эксперимента или получены из внешнего автономного источника, такого как веб-ресурс, с использованием программ-пауков [3], API или веб-скрейпинга [4].

Предобработка данных. Принято считать, что качество сырых данных может оказаться достаточно низким. Для оценки и, в перспективе, повышения качества данных, получаемых при последующих стадиях обработки, необходимо провести разведочный анализ (сырых) данных, выявив такие недостатки, как пропуски, ошибки, повторяющиеся записи и выбросы. Выявление и разрешение вышеперечисленного также позволит подготовить данные для будущего использования [5].

Анализ данных. Несмотря на включение разведочного анализа данных в предыдущий этап обработки, после его проведения и оценки качества данных требуется проведение ряда аналитических процедур, связанных с решением прикладной задачи (например, статистический или семантический анализ) в соответствии с типом данных и степенью их структурированности. Данный этап позволяет получить информацию, необходимую для решения прикладных задач [6].

Операционный этап. На данном этапе извлеченная и подготовленная информация применяется для решения прикладных задач. Например, обработанные данные (извлеченная из них информация) могут послужить входными данными для решения прикладных задач в области искусственного интеллекта.

Хранение/уничтожение данных. Последний этап жизненного цикла. Жизненный цикл данных может быть

представлен как конечный отрезок или быть зацикленным в зависимости от конечной цели. Ответ на вопрос, какие данные должны быть сохранены в конце жизненного цикла данных, зависит от цифровой инфраструктуры эксперимента и целей научной группы [7].

Обработка больших массивов данных научных экспериментов

В отличие от построения типовых инженерных решений, создание цифровой инфраструктуры вокруг новой исследовательской группы может носить стохастический характер. На начальном этапе или при переходе на новую концепцию есть вероятность понимания, что ожидаемые объемы оперируемых данных будут быстро увеличиваться и потребовать новых инструментов. Подобные факторы могут существенно влиять как на проведение самого эксперимента в целом, так и на то, какие данные будут сохранены после каждого из этапов жизненного цикла [8].

Общие компоненты EMS-систем

Постановка проблемы

В данной работе предложена концептуальная высокоуровневая архитектура EMS-системы для работы с большими данными, в частности, для проектов, не требующих внедрения грид-систем. Проекты, опирающиеся на грид-системы, зачастую относятся к проектам класса «мегасайенс» [9] и включают в себя уникальный аппаратный комплекс, а также специфичную информационную инфраструктуру. Предполагается, что необходимость поиска общей закономерности в таком классе научных экспериментов все еще остается под вопросом.

Описание компонентов

Предлагаемая архитектура была сформулирована на основе предыдущего опыта авторов и основных проблем, которые наблюдались или могут наблюдаться в ходе проведения экспериментов с большими массивами данных (рисунок). Весь набор обязательных компонентов разделен на две категории. Некоторые из них основаны на существующих решениях с открытым исходным кодом, требующих определенной конфигурации. Остальные элементы могут быть разработаны исследовательскими группами самостоятельно.

База знаний. Ключевой элемент, содержащий ссылки на источники данных и дополнительные метаданные, которые обеспечивают управление огромным разнообразием конфигураций экосистемы больших данных [10]. Допускается, что база знаний может быть выполнена на основе нереляционного документно-ориентированного хранилища или онтологического хранилища.

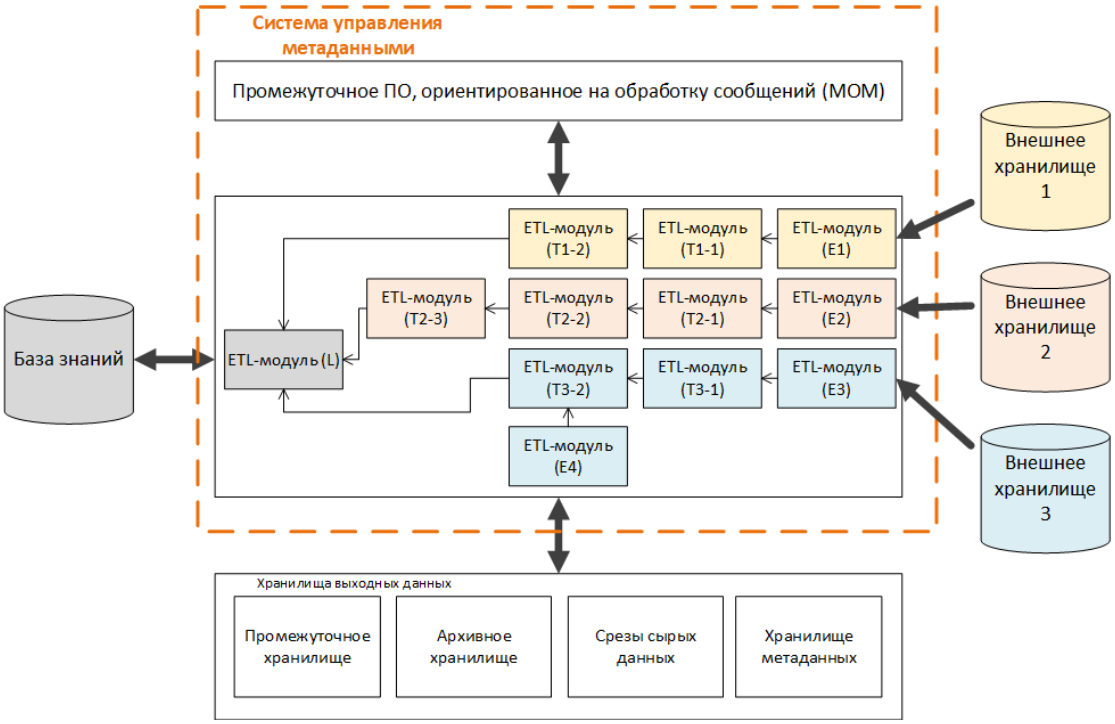

Рисунок. *Концепция высокоуровневой архитектуры EMS-системы*
Figure. *The concept of high-level architecture of the EMS system*

ETL-конвейеры позволяют работать с разнородными данными в гибкой форме, достигнутой благодаря логическому разделению ETL-модулей и их структуры, ориентированной на данные. В случае применения ETL-конвейера для обработки данных изменение цели исследования окажет меньшее влияние на процесс перестройки потока данных (dataflow), чем использование целостного неделимого решения по обработке данных [11].

Система управления метаданными. Экосистема больших данных, организованная для проведения научного эксперимента, состоит из разных отдельных элементов. Потоковая обработка данных должна быть организована таким образом, чтобы захватывать все сопутствующие элементы – модули обработки, источники данных и конечные хранилища. Поддержание потоковой обработки данных вручную при работе с большими мас-

сивами разнородных данных может привести к ошибкам данных, дисбалансу вычислительных нагрузок и перегрузкам памяти при обработке на этапах с низкой скоростью обработки относительно остальных этапов. Система управления метаданными предназначена для связывания компонентов EMS-системы и управления потоком данных в автоматическом режиме [11, 12].

Управляющая утилита. Высокоуровневый пользовательский инструмент для управления потоковой обработкой данных. Управляющая утилита является инструментом, позволяющим запускать, отслеживать и останавливать процесс обработки данных.

Промежуточное программное обеспечение, ориентированное на обработку сообщений (Message-oriented middleware, MOM). Наименьший элемент потока данных, содержащий в себе данные, может быть представлен в виде условного сообщения, но доставка данных посредством обмена такими сообщениями не может выполняться модулем обработки из ETL-конвейера сама по себе. MOM является промежуточным слоем между системой управления метаданными и ETL-модулями, который помогает доставлять данные, контролировать состояние процесса обработки и вести журнал событий [13, 14].

Хранилища данных и метаданных. В связи с тем, что развитие научного эксперимента может привести к возникновению различных по своей структуре, типу и назначению хранилищ на протяжении всего эксперимента, можно утверждать о разнородности интегрируемых в процессе обработки данных. Кроме того, необходимо осуществлять сбор данных, полученных на этапах обработки: срезы необработанных данных, промежуточные результаты обработки данных, предобработанные сырые данные и архивные данные, которые могут быть запрошены позже при необходимости воспроизведения эксперимента [7].

Внешние источники данных. Некоторые эксперименты (например, исследования, сопряженные с работой с данными социальных сетей) могут включать автономные внешние источники данных. Такие источники обязательны в ряде проводимых экспериментов, однако имеется риск внезапной недоступности ресурса по тем или иным причинам. Для проведения экспериментов с учетом вероятности потери важных источников данных необходимо снятие промежуточных «слепков» данных и их хранение, по крайней мере, до конца активной стадии эксперимента при отсутствии необходимости загрузки нового «слепка» [15].

Заключение

Реализация предложенной архитектуры позволит организовать автоматизацию проведения эксперимента на больших массивах гетерогенных данных, в том числе данных разной степени структурированности. Это означает, что потоковую обработку результатов экспериментов с большими данными можно организовывать, управлять ей и даже перестраивать цепочку этапов обработки с меньшими затратами времени и меньшим количеством действий в ручном режиме. В свою очередь, это уменьшит вероятность критической нагрузки на систему и количество ошибок, допущенных по причине человеческого фактора. Следующим шагом является создание испытательного стенда на основе предложенной нами архитектуры и проведение на нем экспериментов, что будет являться доказательством осуществимости концепции (PoC).

СПИСОК ЛИТЕРАТУРЫ

1. Desktop experiment management / Y. Ioannidis, M. Livny, E. Haber, R. Miller, O. Tsatalos, J. Wiener // IEEE Data Engineering Bulletin. – 1993. – V. 16. – № 1. – P. 19–23.
2. ZOO: A desktop experiment management environment / Y. Ioannidis, M. Livny, S. Gupta, N. Ponnekanti // Proceedings of the 22nd VLDB Conference. – Mumbai (Bombay), India, 1996. –P. 274–285. URL: http://www.cs.cmu.edu/~natassa/aapubs/conference/ZOO-desktop-experiment.pdf (дата обращения: 10.01.2023).
3. A fast and powerful scraping and web crawling framework // Scrapy. URL: https://scrapy.org/ (дата обращения: 21.09.2021).
4. Mitchell R. Web scraping with Python. – CA: O'Reilly Media Inc., 2018. – 300 p.
5. Kwak S.K., Kim J.H. Statistical data preparation: management of missing values and outliers // Korean Journal of Anesthesiology. – 2017. – № 70 (4). – P. 407–411. DOI: 10.4097/kjae.2017.70.4.407
6. Bruce P., Bruce A. Practical statistics for data scientists. – CA: O'Reilly Media, Inc., 2017. – 320 p.
7. Bengfort B., Bilbro R., Ojeda T. Applied text analysis with Python. – CA: O'Reilly Media, Inc., 2018. – 332 p.
8. Ferreira D.R. Enterprise systems integration: a process-oriented approach. – Berlin: Springer, 2013. – 401 p.
9. Megascience class installations. URL: https://ckp-rf.ru/megaunu/ (дата обращения: 10.01.2023).
10. Data knowledge base: metadata integration system for HENP Experiments / M. Golosova, M. Grigorieva, V. Aulov, A. Kaida, M. Borodin // 27th International Symposium on Nuclear Electronics & Computing. – Becici, Budva, Montenegro, 30 September – 4 October 2019. URL: https://cds.cern.ch/record/2702958 (дата обращения: 10.01.2023).
11. What is ETL (Extract, Transform, Load)? // IBM. URL: https://www.ibm.com/cloud/learn/etl (дата обращения: 10.01.2023).
12. A survey of general-purpose experiment management tools for distributed systems / T. Buchert, C. Ruiz, L. Nussbaum, O. Richard // Future Generation Computer Systems. – 2015. – V. 45. – P. 1–12. URL: https://doi.org/10.1016/j.future.2014.10.007 (дата обращения: 10.01.2023).
13. Narkhede N., Shapira G., Palino T. Kafka: the definitive guide. – CA: O'Reilly Media, Inc., 2017. – 322 p.
14. Active MQ. URL: https://activemq.apache.org/ (дата обращения: 10.01.2023).
15. Изучение процесса онлайн-радикализации молодежи в социальных медиа (междисциплинарный подход) / А.Ю. Карпова, А.О. Савельев, А.Д. Вильнин, Д.В. Чайковский // Мониторинг общественного мнения: экономические и социальные перемены. – 2020. – № 3 (157). – С. 159–181.

Поступила 20.04.2023 г.
Принята: 03.05.2023 г.

Информация об авторах

Кайда А.Ю., старший преподаватель отделения информационных технологий Инженерной школы информационных технологий и робототехники Национального исследовательского Томского политехнического университета.

Савельев А.О., кандидат технических наук, доцент отделения информационных технологий Инженерной школы информационных технологий и робототехники Национального исследовательского Томского политехнического университета.

UDC 659.118

ARCHITECTURE OF MODERN SYSTEMS OF MANAGING A SCIENTIFIC EXPERIMENT IN THE FIELD OF BIG DATA PROCESSING

Anastasiia Y. Kaida[1],
kaicc@tpu.ru

Aleksei O. Savelev[1],
sava@tpu.ru

[1] National Research Tomsk Polytechnic University,
30, Lenin avenue, Tomsk, 634050, Russia.

The history of the development of experiment managemant systems has about 25 years – from small prototype desktop applications to scalable distributed systems. EMS systems are becoming more popular and more complex. A new stage in the development of EMS systems was the beginning of the «era of big data», which processing, analysis and storage take place in a special ecosystem. The big data ecosystem implies new elements that must be taken into account when designing an EMS system to support the full data lifecycle. One of the key challenges is to generalize those EMS systems that have arisen and changed as related technologies have developed, while, at the moment, there is no single design pattern for such systems due to the features of each specific project for which such a system is created. This paper presents a conceptual high-level architecture of an EMS system as an example of a unified pattern suitable for a big data ecosystem. It is important to note that the proposed pattern does not imply work using grid systems.

Key words: *Big Data, experiment management system, data life cycle, data processing, knowledge bases.*

REFERENCES

1. Ioannidis Y., Livny M., Haber E., Miller R., Tsatalos O., Wiener J., Desktop Experiment Management. *IEEE Data Engineering Bulletin*, 1993, vol. 16, no. 1, pp. 19–23.
2. Ioannidis Y., Livny M., Gupta S., Ponnekanti N. ZOO: a desktop experiment management environment. *Proceedings of the 22nd VLDB Conference*. Mumbai (Bombay), India, 1996. pp. 274–285. Available at: http://www.cs.cmu.edu/~natassa/aapubs/conference/ZOO-desktop-experiment.pdf (accessed: 10 January 2023).
3. A fast and powerful scraping and web crawling framework. *Scrapy*. Available at: https://scrapy.org/ (accessed: 21 September 2021).
4. Mitchell R. *Web Scraping with Python*. CA, O'Reilly Media Inc., 2018. 300 p.
5. Kwak S.K., Kim J.H. Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology*, 2017, no. 70 (4), pp. 407–411. DOI: 10.4097/kjae.2017.70.4.407
6. Bruce P., Bruce A. *Practical statistics for data scientists*. CA, O'Reilly Media, Inc., 2017. 320 p.
7. Bengfort B., Bilbro R., Ojeda T. *Applied text analysis with Python*. CA, O'Reilly Media, Inc., 2018. 332 p.
8. Ferreira D.R. *Enterprise systems integration: a process-oriented approach*. Berlin, Springer, 2013. 401 p.
9. Megascience class installations. Available at: https://ckp-rf.ru/megaunu/ (accessed: 10 January 2023).
10. Golosova M., Grigorieva M., Aulov V., Kaida A., Borodin M. Data knowledge base: metadata integration system for HENP experiments. *27th International Symposium on Nuclear Electronics & Computing*. Becici, Budva, Montenegro, 30 September – 4 October 2019. Available at: https://cds.cern.ch/record/2702958 (accessed: 10 January 2023).
11. What is ETL (Extract, Transform, Load)? *IBM*. Available at: https://www.ibm.com/cloud/learn/etl (accessed: 10 January 2023).
12. Buchert T., Ruiz C., Nussbaum L., Richard O. A survey of general-purpose experiment management tools for distributed systems. *Future Generation Computer Systems*, 2015, vol. 45, pp. 1–12. Available at: https://doi.org/10.1016/j.future.2014.10.007 (accessed: 10 January 2023).
13. Narkhede N., Shapira G., Palino T. *Kafka: the definitive guide*. CA, O'Reilly Media, Inc., 2017. 322 p.
14. *Active MQ*. Available at: https://activemq.apache.org/ (accessed: 10 January 2023).
15. Karpova A.Yu., Savelev A.O., Vilnin A.D., Chaykovskiy D.V. Studying online radicalization of youth through social media (Interdisciplinary Approach). *Monitoring of Public Opinion: Economic and Social Changes Journal*, 2020, no. 3 (157), pp. 159–181.

Received: 20 April 2023.
Reviewed: 3 May 2023.

Information about the authors

Anastasiia Y. Kaida, senior lecturer, National Research Tomsk Polytechnic University.
Aleksei O. Savelev, Cand. Sc., associate professor, National Research Tomsk Polytechnic University.