ЕЗЕРА ОТ ДАННИ(DATA LAKES) КАТО ХРАНИЛИЩЕ ЗА ГОЛЕМИ ДАННИ. АНАЛИЗ НА AWS ПОДХОДА

Гено Стефанов

главен асистент, доктор, катедра Информационни технологии и комуникации, УНСС,
e-mail: genostefanov@unwe.bg

Резюме

Езерата от данни(Data Lakes) станаха неизменна част от Големите данни(Big Data) в ролята си на хранилище за хетерогенни и големи по обем данни. Езерата от данни предлагат всички необходими инструменти за съхранение на Големи данни. Тези инструменти могат да бъдат разнообразни в зависимост от софтуерния доставчик. Amazon Web Services(AWS) като един от големите доставчици на услуги за Големи данни е подходящ избор за анализиране на ползите и предизвикателствата пред съхранението на Големи данни в Езера от данни. Освен това ще се разгледа в детайли подхода на AWS.

Ключови думи: Големи Данни, Езера от данни

DATA LAKES AS DATA STORE FOR BIG DATA. ANALYZING AWS APPROACH.

Geno Stefanov

Abstract

Data Lakes have become an integral part of Big Data in their role as storage for heterogeneous and large-volume data. Data lakes offer all the necessary tools to store Big Data. These tools can vary depending on the software vendor. Amazon Web Services (AWS) as one of the major Big Data service providers is a suitable choice to analyze the benefits and challenges of storing Big Data in Data Lakes. In addition, the AWS approach will be discussed in detail.

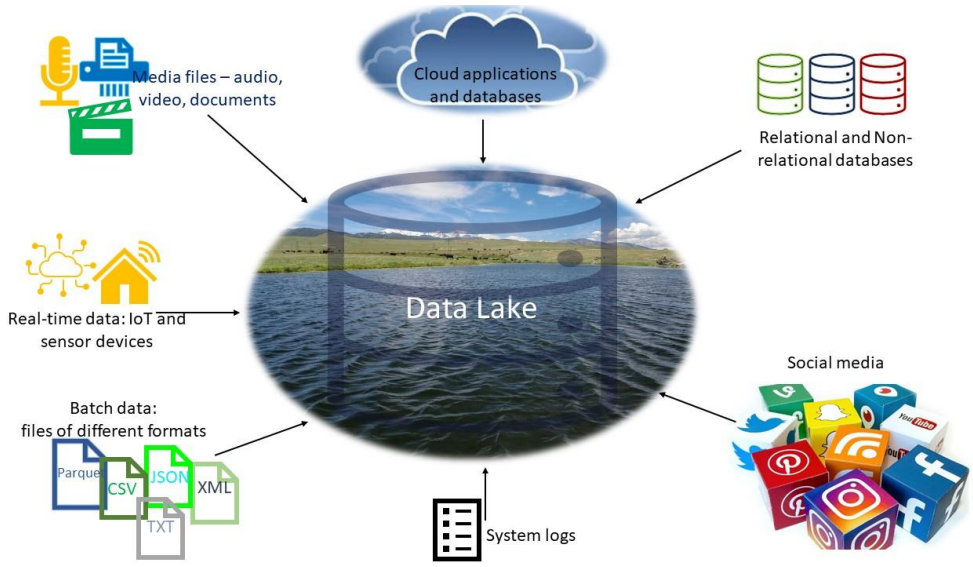
Key words: Big Data, Data Lakes

1.Въведение

В съвременния свят обемът на данни нараства експоненциално, предизвиквайки огромни предизвикателства при управлението, съхранението и анализа на тази информация. Този феномен е известен като "Големи данни" (Big Data). За да се справят с тези предизвикателства, Езерата от данни (Data Lakes) се появяват като нова и иновативна архитектура за съхранение на големи обеми, разнородни и неструктурирани данни. В този доклад ще се разгледа концепцията на езерата от данни и тяхната роля като хранилище за Големи данни.

2.Езера от данни като хранилища за Големи данни

Езеро от данни може да се дефинира като централизирано хранилище, което съхранява големи обеми от сурови и необработени данни в техния естествен формат, без необходимостта от предварително дефинирана схема. Това архитектурно решение позволява съхранението на данни от различни източници, включително структурирани, полуструктурирани и неструктурирани данни, по ефективен и мащабируем начин.



Фигура 1: Езеро от данни

Идеята за Езера от данни за първи път е инициирана от изпълнителния директор на Pentaho Джейм Диксън [3]. Ако складът от данни или базата данни ги представим като бутилка вода, чиста и готова за консумация, тогава Езерата от данни е цяло езеро от данни, което е почистено и готово за употреба. [4] добавя по-задълбочено определение за Езера от данни като „Езерата от данни съхраняват различна информация, като същевременно игнорира почти всичко“. Някои автори смятат, че е необходима нова архитектура на данните [5] в ерата на Големите данни, тъй като тази интензивна изчислителна ера изисква нови идеи и техники за съхраняване и обработка на обемни, разнообразни, променящи се и развиващи се данни.

Езерата от данни са проектирани да обработва Големи данни, което се отнася до големи и сложни масиви от данни, които надхвърлят възможностите на традиционните системи за съхранение и обработка на данни. Езерата от данни предлагат гъвкав и рентабилен подход за съхранение и анализ на данни, позволявайки съхраняването на данни от различни източници.

За разлика от традиционните хранилища за данни, при които структура на данните трябва да е известна[13], преди да се съхранят, езерата от данни поддържат данните в тяхната сурова форма, запазвайки оригиналната детайлност. Тази характеристика позволява на Езерата от данни да съхраняват данни от всякакъв тип и обем, което ги прави подходящи за обработка на непрекъснато нарастващите потоци от данни, генерирани от модерни технологии, като социални медии, IoT устройства, сензори и бизнес приложения[10].

Езерата от данни предлагат всички необходими инструменти за съхранение на Големи данни, за това може да се счита, че те могат да играят съществена роля като хранилище за Големи данни. Тук ще разгледаме предимствата и недостатъците на това Езерата от данни да се използват като хранилище за Големи данни.

Езерата от данни предлагат няколко предимства като хранилище за големи данни:

- Мащабируемост: Езерата от данни могат лесно да нарастват и се справят с нарастващия обем на данни, осигурявайки дългосрочно съхранение без компрометиране на производителността.

- Гъвкавост: Запазвайки суровия формат на данните, езерата от данни предоставят гъвкавост за изследване и анализ на данните, позволявайки на специалистите по данни да извличат информация с различни инструменти.
- Икономичност: Чрез използването на стандартно оборудване и облачни решения, езерата от данни могат да бъдат по-икономични от традиционните решения за съхранение на данни.
- Възможност за обработка в реално време: Езерата от данни позволяват обработка на данни в реално време или близко до реално време, което дава възможност на организациите да взимат бързи решения основани на данни.

Предизвикателства при използването на Езерата от данни. Въпреки своите предимства, езерата от данни представят и някои предизвикателства:

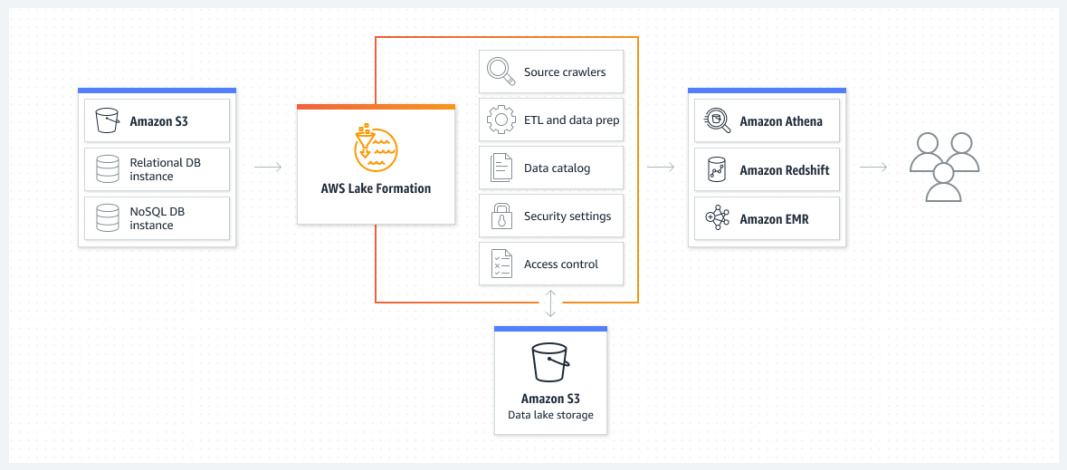
- Управление на данните: Управлението на качеството, сигурността и поверителността на данните в Езерата от данни може да бъде сложно, тъй като липсата на предварително дефинирана схема може да доведе до неизползваемост на данните и грешно тълкуване.
- Управление на метаданните: Управлението на метаданните е от съществено значение в езерата от данни, тъй като помага на потребителите да разберат структурата и контекста на съхранените данни, позволявайки ефективно откриване и анализ на данните.
- Разделение на данните: Ако не се управляват правилно, Езерата от данни могат да се превърнат в “блато” от данни, което ще затрудни или да направи невъзможно управлението на данните.

Най-добри практики при използването на Езерата от данни: За максимизиране на предимствата на Езерата от данни като хранилище за Големи данни, трябва да се следват следните най-добри практики:

- Управление на метаданните: Въвеждането на ефективна стратегия за управление на метаданните позволява каталогизиране на данните, проследяване на техния произход и откриване
- Интеграция с инструменти за обработка на данни: Езера от данни могат да се интегрират с инструменти за обработка на данни като Apache Spark или Hadoop, което ще подобри анализа на данните и тяхната трансформация.
- Управление на сигурността – Трябва да има въведен контрол на достъпа, криптиране и техники за маскиране на данните, за да са защитени чувствителните данни от неоторизиран достъп.

3.AWS решението за Езера от данни

Подходът на AWS за разработване на решения за Езера от данни се нарича AWS Lake Formation. През 2019 г., AWS обявява нова услуга - AWS Lake Formation. AWS Lake Formation е услуга на AWS, която улеснява процеса на създаване, осигуряване и управление на Езера от данни. Тази услуга предоставя комплексен набор от инструменти и функционалности, които позволяват на организациите да интегрират, съхраняват, каталогизират и анализират големи обеми данни от различни източници[15].



Фигура 2: AWS Lake Formation архитектура и функционалности

Основни характеристики на AWS Lake Formation включват:

- Хетерогенни източници на данни: AWS Lake Formation позволява обработката на данни от разнообразни източници, включително бази данни, хранилища за данни, стрийминг данни, лог файлове и други, като се поддържа както пакетно, така и в реално време на приемане.
- Съхранение на данни: С AWS Lake Formation организациите могат да съхраняват големи обеми от сурови, структурирани, полуструктурирани и неструктурирани данни в Amazon S3 (Simple Storage Service) - услугата за съхранение на AWS, която предлага мащабируемост и икономичност.
- Каталог на данни: AWS Lake Formation включва каталог на данни, който автоматично обхожда и каталогизира метаданни от различните източници на данни, съхранени в Езерото от данни. Този каталог позволява на потребителите да откриват, търсят и използват данни по-лесно. Автоматичното каталогизиране на данните става възможно с услугата Crawlers.
- Сигурност и управление на данните: AWS Lake Formation осигурява надеждни механизми за сигурност, включително подробен контрол на достъпа, криптиране и интеграция със системата за управление на идентичност и достъп на AWS (AWS IAM), за контрол на достъпа до ресурсите в Езерото от данни. Също така се поддържа и класификация на данните, което гарантира, че чувствителните данни са подходящо идентифицирани и защитени.
- Трансформация на данните: С помощта на AWS Lake Formation потребителите могат да дефинират задачи за трансформация на данните, използвайки AWS Glue - друга услуга на AWS, която позволява подготовка и трансформация на данните за анализ.
- Интеграция с аналитични услуги: AWS Lake Formation се интегрира безпроблемно с различни аналитични услуги и услуги за машинно обучение в екосистемата на AWS, като например Amazon Athena, Amazon Redshift, Amazon QuickSight и Amazon SageMaker, което позволява на потребителите да извличат ценни аналитични данни от данните, съхранени в Езерото от данни.

AWS Lake Formation предлага редица предимства, включително лесно създаване и управление на Езера от данни, мащабируемост, икономичност и интеграция с широка гама от аналитични и услуги за машинно обучение на AWS. Улеснява процеса на обработка на Големи обеми данни и извличане на ценни знания за организациите, които се нуждаят от вземане на бързи и информирани решения.

4.Заключение

В заключение може да кажем, че Езерата от данни е мощно решение за управление на данни, което позволява на организациите ефективно да обработват Големи данни, което им позволява да получат ценна информация и да вземат ефективно решения, базирани на данни

References

1. Brian Stein, Alan Morrison," The enterprise data lake: Better integration and deeperanalytics, Technology Forecast: Rethinking integration", Issue 1, 2014,
2. Pwint Phyu Khine, Zhao Shun Wang, Data lake: a new ideology in big data era, ITM Web of Conferences, 2018, DOI:10.1051/itmconf/20181703025.
3. James Dixon, Pentaho, Hadoop and Data Lakes. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/
4. Timothy King "The Emergence of Data Lake: Pros and Cons", March 3, 2016, https://solutionsreview.com/data-integration/the-emergence-of-data-lake-pros-and-cons/
5. Delchev, D., Lazarova, V., Big Data Analysis Architecture, Economic Alternatives, 2021, Issue 2, pp. 315-328
6. Boyanov L., The Digital World - The Change, The global digital transformation - enriching or impoverishing humanity, ISBN 978-619-239-637-4, Avangard Prima Publ., Sofia 2021, 188 p.
7. M. Tsaneva, "A Practical Approach For Integrating Heterogeneous Systems," Business management, no. 2, p. 11, 2019.
8. V. Mihova, Common Architecture Design of a Business Information System for Performance Management of the Business Applications, in 3rd International conference on application of information and communication technology and statistics in economy and education ICAICTSEE– 2013, Sofia, Bulgaria, 2013.
9. P. Milev, Technological Issues of Storing Dynamic Data in a Relational Database on Research Projects, Trakia Journal of Sciences, vol. 13, pp. 22-25, 2015
10. E. Karkalikova, A. Murdjeva, Organization of Data in Data Lake – Real-Life Practice, 11th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education ICAICTSEE– 2021, Sofia, Bulgaria.
11. P. Milev, Approach for Analysis and Comparison of Search Query Results in Web Publications, 11th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education ICAICTSEE– 2021, Sofia, Bulgaria
12. Marzovanova M., Building Multi-Touch User Interface, 4TH International Conference on Application of Information and Communication Technology and Statistics in Economy And Education (ICAICTSEE-2014), 2014, (icaictsee.unwe.bg/past-conferences/ICAICTSEE-2014.pdf).
13. Mihova V., Murdjeva A. Metadata for generating a specific data warehouse. International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE-2012), Sofia, Bulgaria, 2012. (icaictsee.unwe.bg/past-conferences/ICAICTSEE-2012.pdf)
14. Dan Wood, Big data requires a big new architecture, Forbes, https://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/#66609cb61157
15. AWS Lake Formation. https://aws.amazon.com/lake-formation/