КАТАЛОГИЗИРАНЕ НА ДАННИ В СРЕДА НА ГОЛЕМИ ДАННИ

Гено Стефанов

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

Резюме

С непрестанното нарастване на обемите от данни и в среда на Големи данни, правилното и навременно каталогизиране на данните играе важна роля в тяхната обработка и откриване. Съвременните среди за Големи данни неминуемо имат някакъв вид каталогизиране на данните под формата на различни по вид технолиги и инструменти. Целта на настоящия доклад е да анализира възможностите и предизвикателствата пред каталогизирането на данни в контекста на Големи Данни. Освен това ще се разгледат конкретни техники и технологии въведени от Amazon Web Services(AWS) - AWS Catalog и AWS Glue Crawler.

Ключови думи: Големи Данни, Каталогизиране на данни. Big Data, Data Catalogs.

DATA CATALOGING IN BIG DATA ENVIRONMENT

Geno Stefanov

Abstract

With the ever-increasing volumes of data and in a Big Data environment, proper and timely cataloging of data plays an important role in its processing and discovery. Modern Big Data environments inevitably have some form of data cataloging in the form of various technologies and tools. The purpose of this report is to analyze the opportunities and challenges of data cataloging in the context of Big Data. In addition, specific techniques and technologies provided by Amazon Web Services (AWS) - AWS Catalog and AWS Glue Crawler - will be analyzed.

Key words: Big Data, Data Catalogs.

1. Въведение

Каталогизирането на данни в областта на Големите данни се отнася до централизираната система за съхранение и управление на метаданни, която съдържа информация за различните набори от данни в екосистемата на организацията. Каталогизирането на данни може да се представи като създаване на инвентар на данните, с помощта на който може да се търсят, оценяват и използват данните от потребителите. Каталогът за данни има решаващо значение за улесняването на управлението на данни, търсенето на данни и използването им в големи и сложни среди с данни, каквито са Големите данни.

2. Каталогизиране на данни

В основата на каталозите за данни са метаданните. Според стандарта ISO 11179 метаданните се дефинират като данни, които описват и дефинират други данни. Всеки потребител на компютър използва метаданни: запазването на документ в папка и даването му на определено име означава активно присвояване на атрибути на метаданни като име на файл и път до файла. Тези метаданни

могат по-късно да се използват за проследяване на файла, когато е необходи. Има различни видове групи метаданни [2], [3]:

- описателни метаданни: насочени към подобряване на откриваемостта и разбирането на съдържанието

- административни метаданни: могат да бъдат разделени на множество подкатегории като технически, съхранение и разрешения. Тези метаданни дават контекст по отношение на техническия аспект като файлов тип и декодиране, чувствителност и атрибути за достъпност.

Някои автори [1] описват каталогът за данни като „Каталогът за данни поддържа инвентаризация на активи с данни чрез откриване, описание и организиране на набори от данни. Каталогът предоставя контекст, за да позволи на анализаторите на данни, специалисти по данни, администратори на данни и други потребители на данни да намерят и разберат подходящ набор от данни с цел извличане на бизнес стойност„.

Тъй като каталогизирането на данните разчита на метаданни, качеството на метаданните е от изключително значение. Множество проучвания са фокусирани върху това как може да се оцени качеството на метаданните [4], [5] и подчертават, че липсата на качествени метаданните може да доведе до намалена възможност за търсене и достъпност [4]. Тъй като метаданните се оказват толкова ключов елемент от използваемостта на каталога за данни, всички инструменти за каталогизиране на данни се очаква да имат някакъв вид модул за управление на метаданни в своя софтуер.

Основни характеристики и ползи от каталогизирането на данни в областта на Големите данни включват:

1. Управление на метаданни: Каталогът за данни съхранява детайлни метаданни за наличните данни в Големите данни, включително информация за източниците на данни, произхода на данните, качеството на данните, схемата и други. Тези метаданни имат за цел да подпомогнат разбирането на съдържанието и контекста на данните, което улеснява оценката им за актуалност и надеждност.

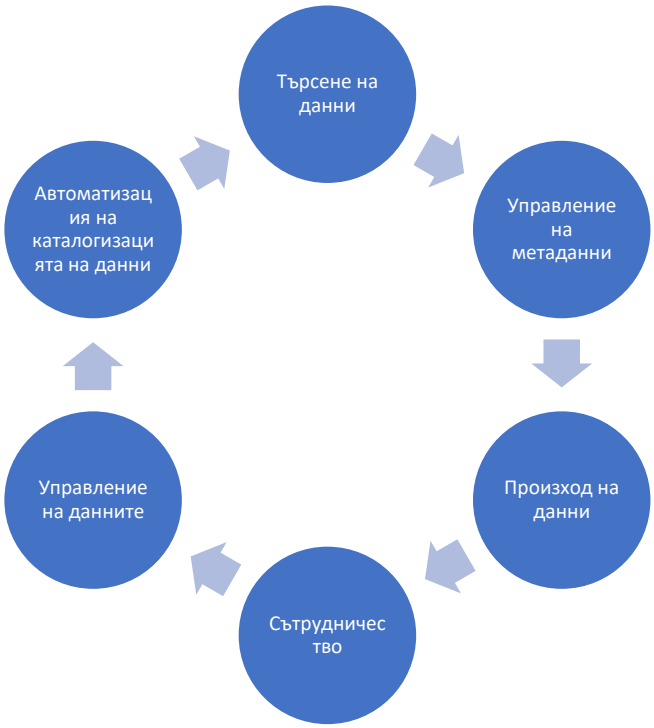
2. Търсене на данни: Каталогизирането на данни позволява лесно и бързо търсене и изследване в каталога на данни, за да се открият подходящи набори от данни за анализи или репорти. Това може значително да намали времето, отделено за търсене на данни и достъп до тях.

3. Произход на данни: Информацията за произхода на данните в каталога показва началото и трансформациите, през които данните преминават по време на обработката. Това помага да се разбере как данните се трансформират и преобразуват, което допринася за по-доброто разбиране на данните.

4. Сътрудничество: Каталозите за данни често предоставят функции за сътрудничество, които позволяват добавянето на анотации, коментари, оценки и рецензии към наборите от данни. Това насърчава сътрудничеството между потребителите на данните, анализаторите на данни и лицата, отговорни за управлението на данните.

5. Управление на данните: Каталозите за данни подпомагат усилията по управление на данните, като налагат политики за достъп до данните, гарантират съответствие с регулациите за данните и осигуряват видимост върху това, кой достъпва и използва данните.

6. Автоматизация на каталогизацията на данни: В средите с Големи данни от изключително значение е да има подходящи инструменти за автоматичното каталогизиране на данните в Каталог на данни. Това ще подобри възможностите за бързо откриване на връзки между данните.



Фигура 1: Основни характеристики и ползи от каталогизирането на данни

3. AWS Data Catalog и AWS Glue Crawler

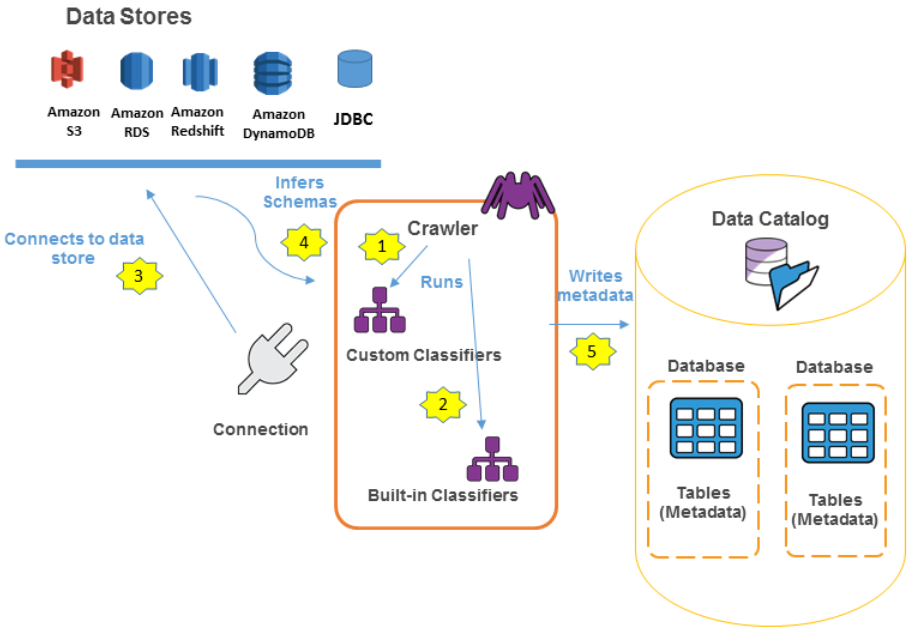
Като един от водещите доставчици на технологични услуги в сферата на Големите данни, AWS(Amazon Web Services) има свое решение за каталог на данни и автоматично каталогизиране на данни. AWS Data Catalog е част ETL инструмента Glue, но може да се използва и самостоятелно като отделна услуга.

Според [6] каталогът с данни на AWS Glue е напълно управлявано хранилище за постоянни метаданни, което позволява съхранението, анотирането и споделянето на метаданни. Той предоставя унифицирано хранилище на метаданни за различни източници на данни и формати на данни, като може да се интегрира с Amazon EMR, както и с Amazon RDS, Amazon Redshift и Redshift Spectrum, Amazon Athena, AWS Lake Formation и всяко приложение, съвместимо с metastore Apache Hive.

Типичното приложение на каталогът за данни на AWS е като част от процесите по извличане, трансформиране и зареждане на данни(ETL). Каталогът за данни на AWS Glue представлява индекс на местоположението, схемата и показателите за време на изпълнение относно данните.

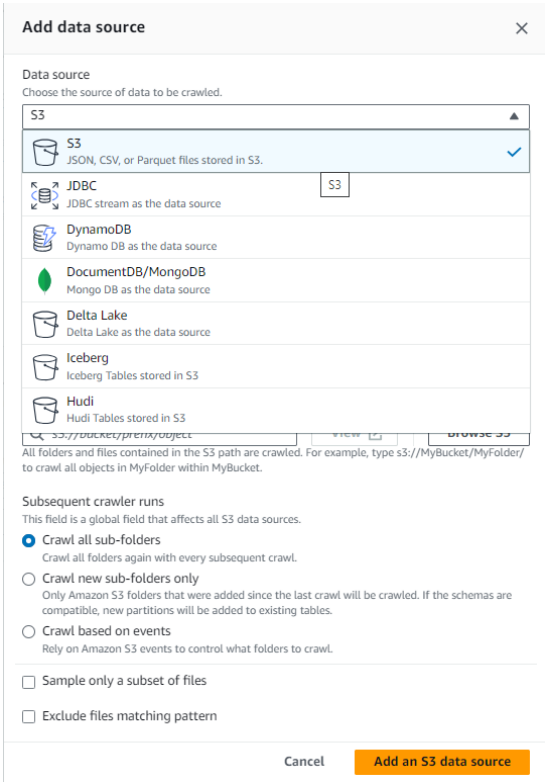
Информацията в каталога за данни се съхранява като таблици с метаданни, където всяка таблица посочва едно хранилище на данни. Обикновено автоматичното управление на тези таблици с метаданни става чрез използването на AWS Crawler.

Следната диаграма представя типичен работен процес, който показва как AWS Glue Crawler си взаимодейства с хетерогенните източници на данни за да ги каталогизира в каталога за данни.



Фигура 2: AWS Glue Crawler и Data Catalog

Една от основните функции на Glue Crawler е да намира данни, без значение дали те са във формати като JSON, CSV, Parquet, или дори бази данни като Amazon RDS или Amazon Redshift. Този автоматичен процес позволява да се сведе до минимум ръчното каталогизиране на данните и ускорява времето за подготовката на данните за последваща обработка. Освен това Glue Crawler предоставя възможности за достъп до разнообразни източници на данни(Фиг. 3).



Фигура 3: AWS Glue Crawler източници на данни

Основната услуга за формиране на Езеро от данни в AWS е S3. AWS Crawler-а е специално разработен да обхожда автоматично данни в S3 и да записва в Data Catalog-а наличните промени или нови метаданни за данните.

AWS Glue Crawler генерира богати метаданни за откритите данни, включително информация за структурата на таблиците и полетата, типовете данни и ключовите връзки. AWS Glue Crawler предоставя автоматизация на процеса по обновление на метаданни, като сканира периодично източниците на данни и актуализира информацията за тях. Също така, услугата е скалируема и може да се адаптира към растящия обем на данни и източници.

Общият работен процес за това как AWS Crawler-а попълва каталога за данни е следния:

1. Изпълняват се така наречените класификатори, които проверяват дали данните могат да бъдат прочетени в определен формат. Важно е да се отбележи, че има възможност за създаване на персонализирани класификатори.

2. Crawler-а се свързва с източника на данни, като е важно да се отбележи, че в зависимост от типа на източника на данни може да се изискват определени свойства за достъп до източника.

3. Изведената схема се създава за откритите данни.

4. Crawler-а записва метаданните в каталога за данни под формата на таблица. Дефиницията на таблица съдържа метаданни за данните. Таблицата се записва в база данни, която е контейнер от таблици в каталога за данни. Атрибутите на таблица включват класификация, която е етикет, създаден от класификатора, който е извел схемата на таблицата.

4. Заключение

В заключение може да кажем, че каталогът за данни е ключов инструмент за ефективното управление, изследване и използване на големи и сложни набори от данни. Той играе съществена роля в улесняването на откриването на данни, насърчаването на сътрудничество, осигуряването на управление на данните и подобряването на разбирането на данните в организацията. В епохата на Големите данни ефективната стратегия за каталогизиране на данни се явява крайъгълен камък за постигане на успех в осъществяването на решения, базирани на Големи данни. AWS Glue Catalog и AWS Glue Crawler успешно се справят с каталогизирането на различни източници и то автоматично

References

1. E. Zaidi, G. De Simoni, R. Edjlali, and A. D. Duncan, “Data Catalogs Are the New Black in Data Management and Analytics,” Gartner, no. December, pp. 1–16, 2017
2. L. Ehrlinger, J. Schrott, M. Melichar, N. Kirchmayr, and W. Wöß, “Data Catalogs: A Systematic Literature Review and Guidelines to Implementation,” in Database and Expert Systems Applications - DEXA 2021 Workshops, 2021, vol. 2, pp. 148–158, doi: 10.1007/978-3-030-87101-7_15.
3. J. Riley, Understanding Metadata. Baltimore: National Information Standards Organization (NISO), 2017.
4. J. Nogueras-Iso, J. Lacasta, M. A. Urena-Camara, and F. J. Ariza-Lopez, “Quality of Metadata in Open Data Portals,” IEEE Access, vol. 9, pp. 60364–60382, 2021, doi: 10.1109/ACCESS.2021.3073455.
5. J. Klímek, “Reflections on: DCAT-AP representation of Czech national open data catalog and its impact,” CEUR Workshop Proc., vol. 2576, no. 19, pp. 1–9, 2019.
6. AWS Best Practices for Building a Data Lake on AWS for Games https://docs.aws.amazon.com/whitepapers/latest/best-practices-building-data-lake-for-games/data-cataloging.html

7. Delchev, D., Lazarova, V., Big Data Analysis Architecture, Economic Alternatives, 2021, Issue 2, pp. 315-328
8. 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.
9. M. Tsaneva, "A Practical Approach For Integrating Heterogeneous Systems," Business management, no. 2, p. 11, 2019.
10. 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.
11. 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
12. 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.
13. 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
14. 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).
15. 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)