Data Discovery

Chidiebere Ogbuchi ♠ Tianheng Zhou ♢
♠chidiebere.ogbuchi@estudiantat.upc.edu ♢tianheng.zhou@estudiantat.upc.edu
Universitat Politecnica de Catalunya, BarcelonaTech, Barcelona, Spain

Introduction

We are currently experiencing the era of big data, where an immense volume of information is captured and stored in various formats. Figure 1 provides an overview of the key components of data discovery in relation to data science and big data itself. As the volume, velocity, and variety of data continue to grow at an unprecedented rate, the demand for data science expertise has surged across organizations and domains. Consequently, the ability to effectively gather and analyze diverse datasets and variables has become essential.

Data discovery refers to the process of extracting valuable insights by employing automated techniques to identify relevant and complementary datasets from vast repositories. This involves exploring and understanding available data sources to uncover new and valuable datasets that enhance the analysis and decision-making processes. In essence, data discovery acts as a bridge between raw data and actionable insights.

The objective of this poster is to emphasize the paramount importance of data discovery in the field of data science. We will explore the role of data lakes, which serve as comprehensive repositories for diverse data sources, and delve into the potential of data discovery tools to democratize the field of data science.

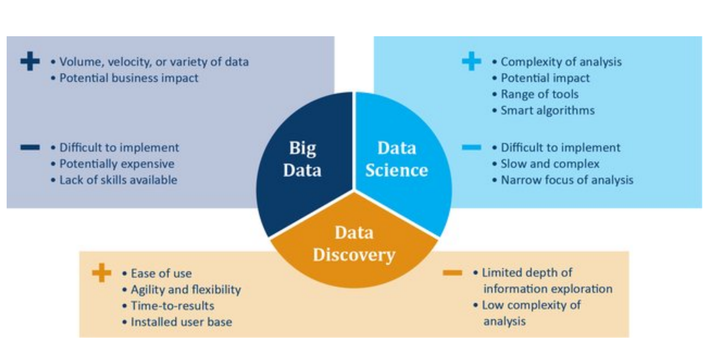


Figure 1. Big Data Discovery is the combination of Big Data, Data Science, and Data Discovery [2]

Literature review

Traditional Data Discovery

These methods formed the foundation of data discovery before the advent of data lakes, providing insights into data sources and relationships. Figure 2b classifies them into catalog and non-catalog methods

Search and Query-Based Methods:

- Search engines: Keyword-based data retrieval.
- Query-based discovery: Using structured queries for finding relevant data.

Data Profiling:

- Analyzing data sets for structure, content, and relationships.
- Identifying anomalies and inconsistencies.

Data Integration and Consolidation:

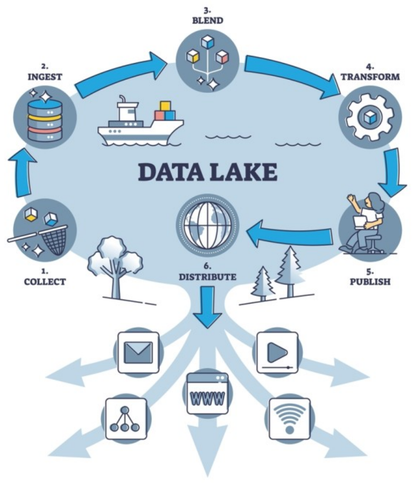
- Combining data from multiple sources.
- Techniques like data mapping and schema matching.

Collaboration and Knowledge Sharing:

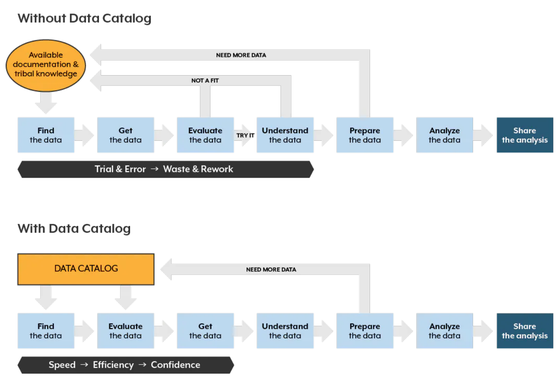
- Sharing expertise and insights.
- Collaboration platforms and communities for knowledge exchange.

A new Architecture: Data lake

Data lakes have emerged as a flexible and scalable solution for managing and accessing diverse data sources. The figure 2a highlights how data lakes store raw and unprocessed data, allowing organizations to handle the volume and complexity of modern data. The Load-First, Model-Later approach shows the potential of data lakes in facilitating the data discovery processes.



(a) Data lake Architecture [5]



(b) Data discovery with & without data catalog [7]

Figure 2. Data Lake & Discovery Processes

Methods

Techniques and Approaches for Data Discovery in Datalake

- Inferring join plans on column compatibility and suitability as join attributes.

- Utilizing semantic information: e.g PEXESO framework (using high-dimensional vectors to embed textual values.), Seeping semantics (leveraging word embeddings to establish meaningful connections between datasets.)

- Learning-based approaches using profiles to capture dataset characteristics for joinability.

- Interactive data discovery for dynamic exploration of data lakes.

Data Discovery Systems, Platforms, and Tools

- Data Exploration and Retrieval Tools: Aurum, Auctus, Amundsen, Waterline Data
- Large-Scale Data Discovery Tools: Zaloni Data Catalog, IBM Watson Discovery
- Data-Driven Domain Discovery Tools: Apache Atlas
- Metadata Management Tools: Collibra Data Catalog, Alation Data Catalog, etc.

These examples represent a variety of tools and platforms available for data discovery, each offering unique features and capabilities to assist users in exploring, retrieving, and managing data within data lakes.

Case Study

In order to illustrate data discovery methods, we analyze the methods used in literatures [3] and [1]. The present objectives as shown in figure 3 is to:

- **Objective 1:** Identify joinable attributes in structured datasets (Some other data formats can be transformed into the structured format).
- **Objective 2:** Identify joinable attributes accurately when semantic characteristics are unknown.
- **Objective 3:** Evaluate the suitability of attributes from multiple tables for joining.

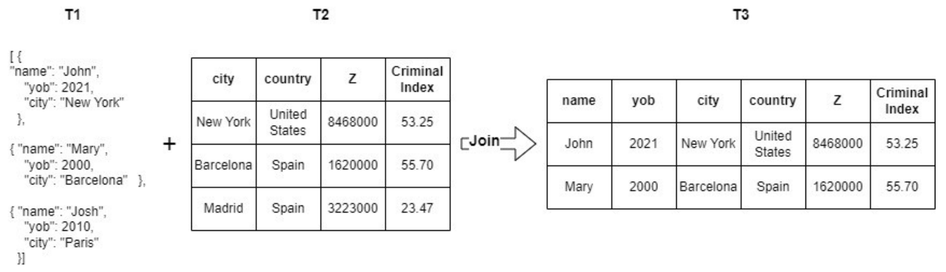


Figure 3. Join Operation with Heterogeneous Data in the Data Lake

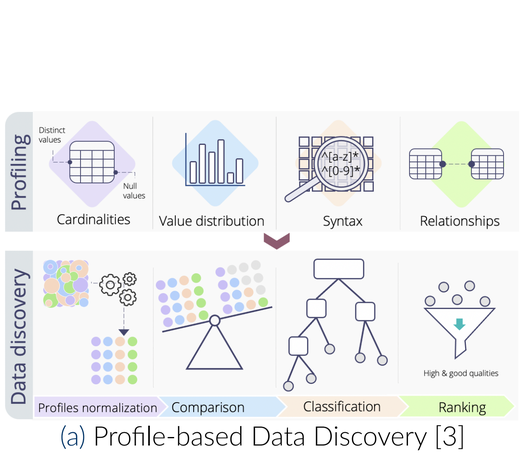
What does this study add?

Data discovery using Profiles [3]

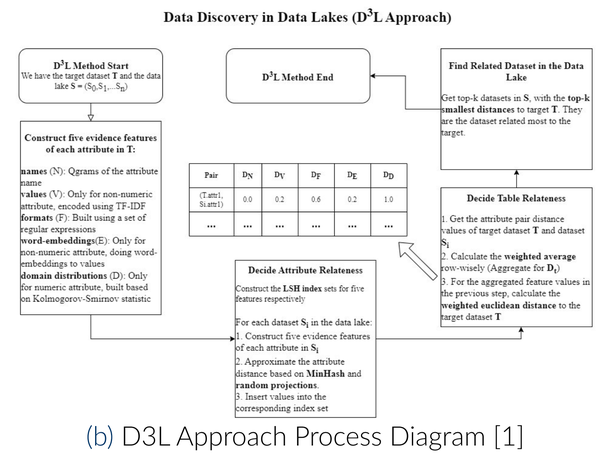
- As depicted in figure 4a, it proposes Profiling methods involving evaluation of cardinalities, value distribution, syntax, and relationships.
- Discovery method such as profile normalization, comparison, classification, and ranking.
- Similarity-based and more scalable, with the use of Apache Spark to leverage efficiency.

Dataset Discovery using Local Sensitive Hashing (D3L Approach) [1]

- Attribute distances are mapped in a 5-dimensional Euclidean space, used for evaluating joinability, which are names, values, formats, word-embeddings, and domain distributions.
- Utilizes LSH-based indexes to find similar attributes using Jaccard and Cosine similarity.
- Distance-based D3L uses the LSH-index to avoid inefficient pair-wise comparison. MinHash and Random Projections approximates the distance calculation to improve computational efficiency.



(a) Profile-based Data Discovery [3]

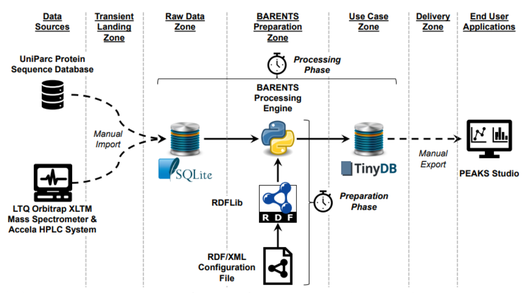


(b) D3L Approach Process Diagram [1]

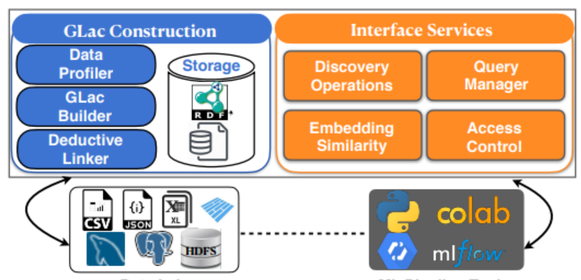
Figure 4. Data Discovery Methods

Gaps and Future Opportunities

- Tailored approaches for data discovery in large-scale and complex data environments.
- Promoting data reuse and exploration techniques for efficient data discovery.
- Knowledge graph technologies for intelligent and context-aware data discovery. e.g BARENT prototype and KGLac Architecture as in figure 5.
- Demand-driven data provisioning for optimized data access and availability in data lakes.
- Advanced search and join algorithms for seamless data lakes exploration and integration.



(a) A Prototypical Implementation of BARENTS used for the Evaluation [6]



(b) The KGLac architecture [4]

Figure 5. Two Context-aware Data Discovery Platforms

Conclusions

Data discovery plays a vital role in enabling informed decision-making and predictive modeling by exploring and integrating relevant data. Techniques such as capturing semantic relationships, enhancing organizational capabilities, and developing advanced algorithms improve the efficiency and effectiveness of data discovery in data lakes. By leveraging these approaches, data exploration and analysis are enhanced, leading to better utilization of data resources and increased value generation. Overall, data discovery is essential for optimizing data lake usage and maximizing the benefits derived from it.

References

[1] A. Bogatu, A. A. A. Fernandes, N. W. Paton, and N. Konstantinou.
Dataset discovery in data lakes, 2020.

[2] Gartner.
Big Data Discovery is the combination of Big Data, Data Science, and Data Discovery., 2017.

[3] J. Gil Flores, S. Nadal, and O. E Romero.
Scalable data discovery using profiles, 2020.

[4] E. Mansour.
A data discovery platform empowered by knowledge graph technologies: Challenges and opportunities, 2021.

[5] Pulse.
Data lake topnotch tech.

[6] Bräcker J. Eichler R. Giebler C. Mitschang B. Stach, C.
Ddemand-driven data provisioning in data lakes, 2021.

[7] D. Wells.
What is a data catalog? data catalog features benefits.