MOHAMMED VI
POLYTECHNIC
UNIVERSITY

UM6P-CS
SCHOOL OF
COMPUTER SCIENCE

Scalable Human-Guided Data Integration

Student :

Jaouhara Chanchaf

Advisor :

Karima Echihabi, UM6P

Pre-doctoral program
2022-2023

24 November 2022

Summary

1 Introduction

2 Problem Statement

3 Literature Review

4 Progress

5 Future work

Introduction

Scalable Human-Guided Data Integration

- **Data Integration**
  Combine data from disparate sources.
- **Related Dataset Discovery**
  Find relevant related dataset.

Dataset Discovery Challenges

1. **Data Modality :**
   Data is available in different formats, structured (e.g. datasets found in relational databases) semi-structured (e.g. XML, CSV and JSON files) and unstructured (e.g. social media data such as text documents, images, audio files etc.)

2. **Data Volume :**
   Data is available in massive collection of dataset that could reach Hundreds of Terra-bytes.

3. **Data Locality :**
   Data is available across different separate repositories (e.g. open data portals, data marketplaces and data lakes)

Problem Statement : Related Dataset Search

Given a data repository $\mathcal{D}$ and a query dataset D_Q characterized with a domain Q. Find all datasets in $\mathcal{D}$ that contains at least one domain S similar to Q.

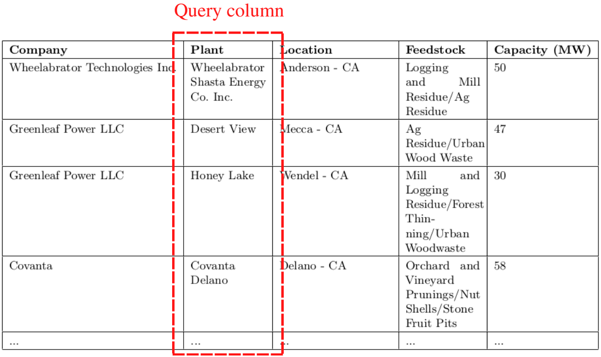
Example :

- Dataset : Relational Table
- Domain : Column
- Relatedness measure : Joinability

Problem Statement : Joinable Tables Search

Given a data repository $\mathcal{D}$ and a query column Q in a query table T_Q. find all tables in $\mathcal{D}$ that can join with T_Q on Q.

Example of Joinable Tables



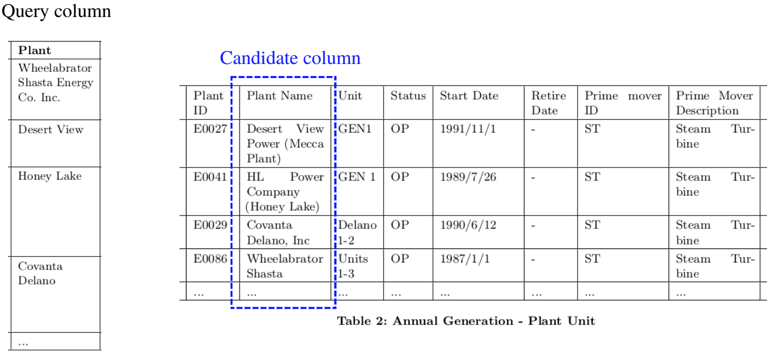
Query column

Company	Plant	Location	Feedstock	Capacity (MW)
Wheelabrator Technologies Inc	Wheelabrator Shasta Energy Co. Inc.	Anderson - CA	Logging and Mill Residue/Ag Residue	50
Greenleaf Power LLC	Desert View	Mecca - CA	Ag Residue/Urban Wood Waste	47
Greenleaf Power LLC	Honey Lake	Wendel - CA	Mill and Logging Residue/Forest Thinning/Urban Woodwaste	30
Covanta	Covanta Delano	Delano - CA	Orchard and Vineyard Prunings/Nut Shells/Stone Fruit Pits	58
...	...	...	...	...

Table 1: U.S. Biomass Power Plants

Figure – Joinable Tables Example from WDC 2015 English Corpus

Example of Joinable Tables



Query column

Plant
Wheelabrator Shasta Energy Co. Inc.
Desert View
Honey Lake
Covanta Delano
...

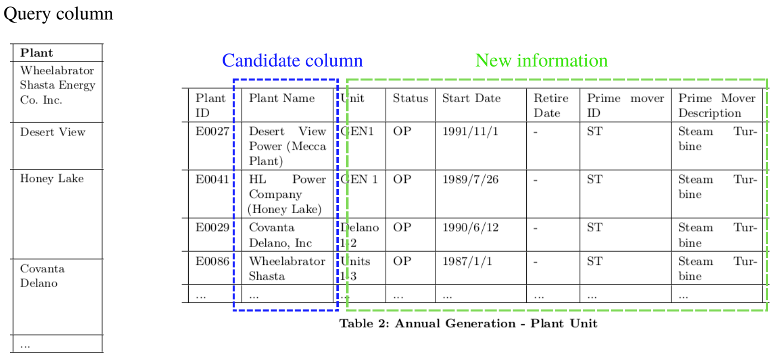
Candidate column

Plant ID	Plant Name	Unit	Status	Start Date	Retire Date	Prime mover ID	Prime Mover Description
E0027	Desert View Power (Mecca Plant)	GEN1	OP	1991/11/1	-	ST	Steam Turbine
E0041	HL Power Company (Honey Lake)	GEN 1	OP	1989/7/26	-	ST	Steam Turbine
E0029	Covanta Delano, Inc	Delano 1-2	OP	1990/6/12	-	ST	Steam Turbine
E0086	Wheelabrator Shasta	Units 1-3	OP	1987/1/1	-	ST	Steam Turbine
...	...	...	...	...	...	...	...

Table 2: Annual Generation - Plant Unit

Figure – Joinable Tables Example from WDC 2015 English Corpus

Example of Joinable Tables



Query column

Plant
Wheelabrator Shasta Energy Co. Inc.
Desert View
Honey Lake
Covanta Delano
...

Candidate column

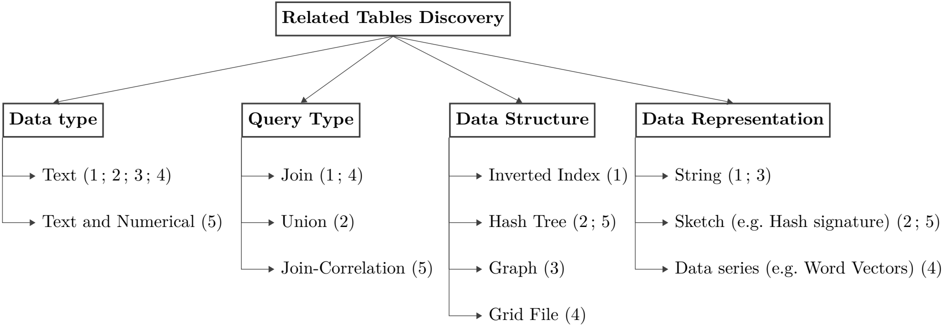
New information

Plant ID	Plant Name	Unit	Status	Start Date	Retire Date	Prime mover ID	Prime Mover Description
E0027	Desert View Power (Mecca Plant)	GEN1	OP	1991/11/1	-	ST	Steam Turbine
E0041	HL Power Company (Honey Lake)	GEN 1	OP	1989/7/26	-	ST	Steam Turbine
E0029	Covanta Delano, Inc	Delano 1-2	OP	1990/6/12	-	ST	Steam Turbine
E0086	Wheelabrator Shasta	Units 1-3	OP	1987/1/1	-	ST	Steam Turbine
...	...	...	...	...	...	...	...

Table 2: Annual Generation - Plant Unit

Figure – Joinable Tables Example from WDC 2015 English Corpus

Literature Review



Related Tables Discovery

- **Data type**
  - Text (1 ; 2 ; 3 ; 4)
  - Text and Numerical (5)
- **Query Type**
  - Join (1 ; 4)
  - Union (2)
  - Join-Correlation (5)
- **Data Structure**
  - Inverted Index (1)
  - Hash Tree (2 ; 5)
  - Graph (3)
  - Grid File (4)
- **Data Representation**
  - String (1 ; 3)
  - Sketch (e.g. Hash signature) (2 ; 5)
  - Data series (e.g. Word Vectors) (4)

Progress

- ↪ Review the literature and identify point of similarity and dissimilarity between proposed frameworks.
- ↪ Identify the key factors that influence the effectiveness and efficiency of a dataset discovery framework.
- ↪ Implement a dataset discovery framework using DSTree (Y. Wang et al., 2013) an existing data structure designed for efficient storage and retrieval of data series.

Kashif : Incremental Joinable Table Search

Incremental Joinable Table Search using Parallel kNN Search

- **Data Type :** Text data in tabular datasets.
- **Query Type :** Join
- **Data Representation :** FastText Word Embeddings
- **Data Representation :** DSTree (Y. Wang et al., 2013)

Kashif : Incremental Joinable Table Search

Choosing k

- A larger k values yield a higher recall.
- The larger is k the longer it takes to answer the query.
- The optimal k value is data dependent.

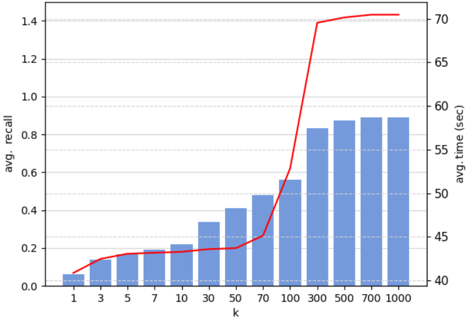


Figure – Kashif Mean query time and avg. recall [25M vectors, 10 queries, query size = 50 - 100]

Kashif : Incremental Joinable Table Search

Choosing k

- A larger k values yield a higher recall.
- The larger is k the longer it takes to answer the query.
- The optimal k value is data dependent.

→ Solution : Set a very high value for k and return results incrementally.

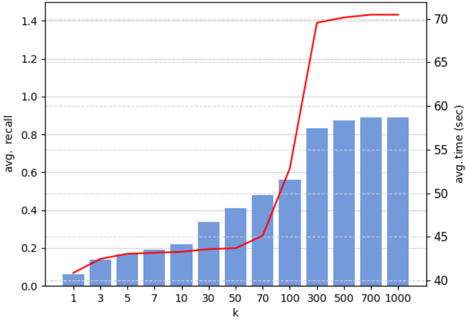
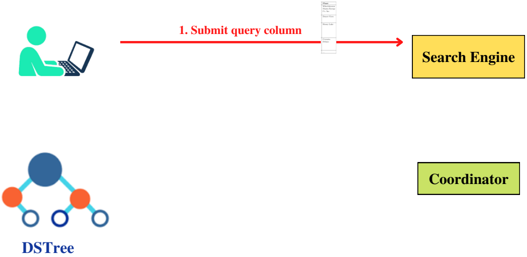
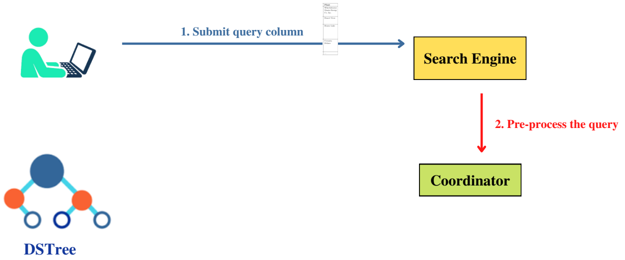


Figure – Kashif Mean query time and avg. recall [25M vectors, 10 queries, query size = 50 - 100]

Kashif : Incremental Joinable Table Search (Parallel kNN Search)

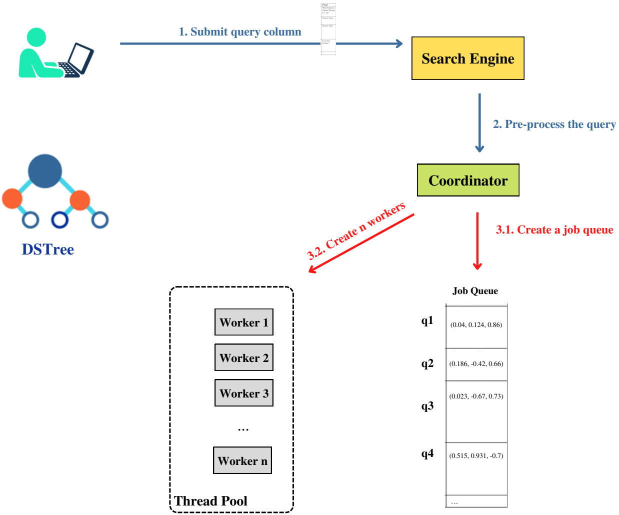


Kashif : Incremental Joinable Table Search (Parallel kNN Search)



Plant
Wheelabrator Shasta Energy Co. Inc.
Desert View
Honey Lake
Covanta Delano
...

Kashif : Incremental Joinable Table Search (Parallel kNN Search)



1. Submit query column

Search Engine

2. Pre-process the query

Coordinator

3.1. Create a job queue

3.2. Create n workers

DSTree

Thread Pool

Worker 1

Worker 2

Worker 3

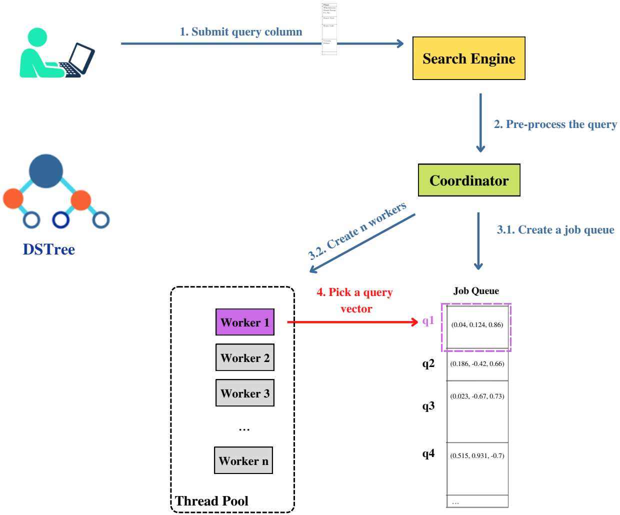
...

Worker n

Job Queue

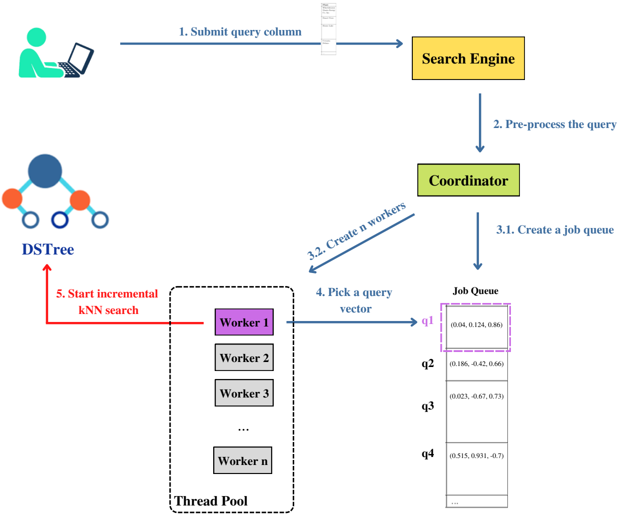
q1	(0.04, 0.124, 0.86)
q2	(0.186, -0.42, 0.66)
q3	(0.023, -0.67, 0.73)
q4	(0.515, 0.931, -0.7)
	...

Kashif : Incremental Joinable Table Search (Parallel kNN Search)



	Job Queue
q1	(0.04, 0.124, 0.86)
q2	(0.186, -0.42, 0.66)
q3	(0.023, -0.67, 0.73)
q4	(0.515, 0.931, -0.7)
	...

Kashif : Incremental Joinable Table Search (Parallel kNN Search)



1. Submit query column

Search Engine

2. Pre-process the query

Coordinator

3.1. Create a job queue

3.2. Create n workers

DSTree

5. Start incremental kNN search

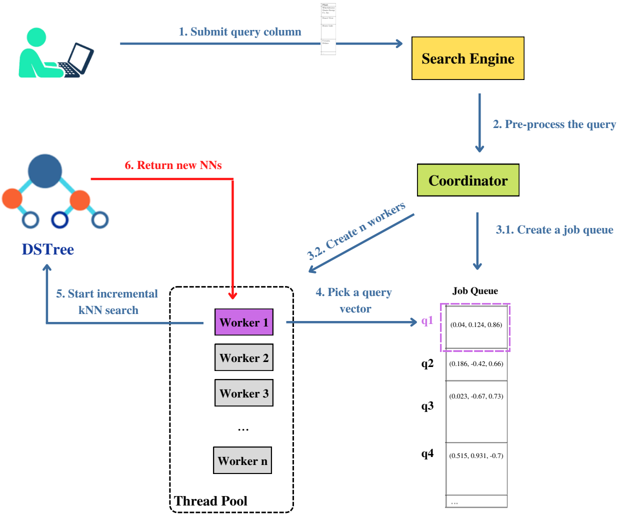
Worker 1
Worker 2
Worker 3
...
Worker n
Thread Pool

4. Pick a query vector

Job Queue

q1 (0.04, 0.124, 0.86)
q2 (0.186, -0.42, 0.66)
q3 (0.023, -0.67, 0.73)
q4 (0.515, 0.931, -0.7)
...

Kashif : Incremental Joinable Table Search (Parallel kNN Search)



1. Submit query column

Search Engine

2. Pre-process the query

Coordinator

3.1. Create a job queue

3.2. Create n workers

4. Pick a query vector

5. Start incremental kNN search

6. Return new NNs

DSTree

Worker 1

Worker 2

Worker 3

...

Worker n

Thread Pool

Job Queue

q1 (0.04, 0.124, 0.86)

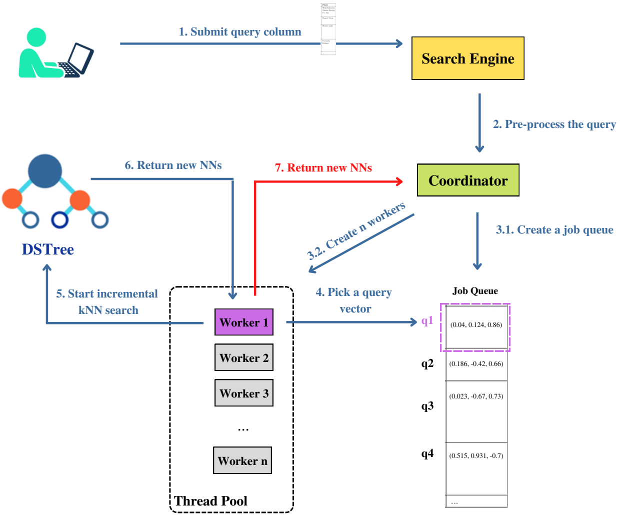
q2 (0.186, -0.42, 0.66)

q3 (0.023, -0.67, 0.73)

q4 (0.515, 0.931, -0.7)

...

Kashif : Incremental Joinable Table Search (Parallel kNN Search)



1. Submit query column

Search Engine

2. Pre-process the query

Coordinator

3.1. Create a job queue

3.2. Create n workers

4. Pick a query vector

5. Start incremental kNN search

6. Return new NNs

7. Return new NNs

DSTree

Worker 1

Worker 2

Worker 3

...

Worker n

Thread Pool

Job Queue

q1 (0.04, 0.124, 0.86)

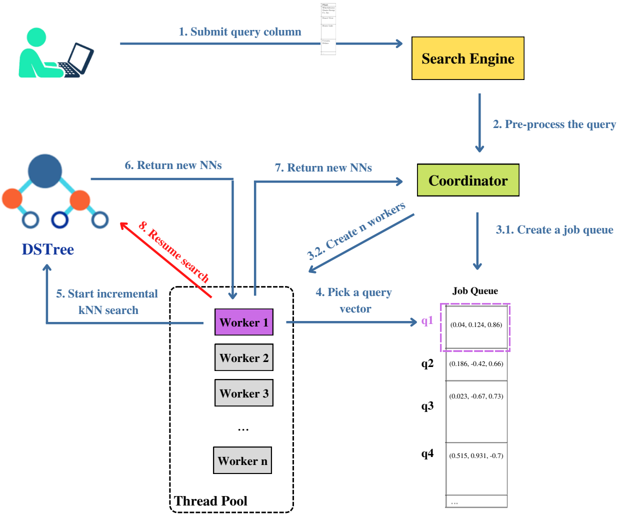
q2 (0.186, -0.42, 0.66)

q3 (0.023, -0.67, 0.73)

q4 (0.515, 0.931, -0.7)

...

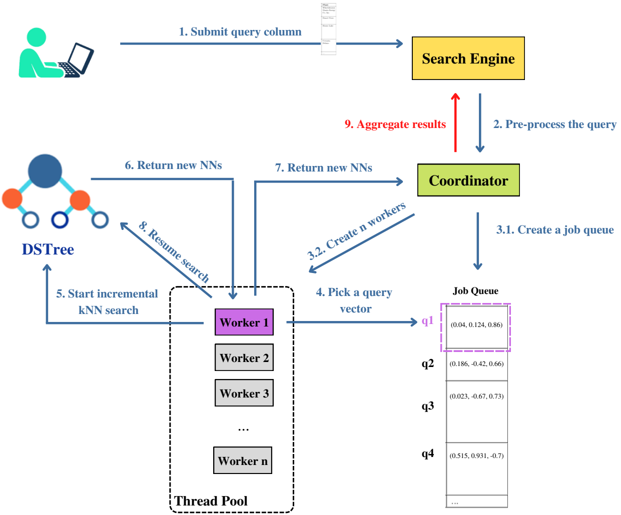
Kashif : Incremental Joinable Table Search (Parallel kNN Search)



Job Queue

q1	(0.04, 0.124, 0.86)
q2	(0.186, -0.42, 0.66)
q3	(0.023, -0.67, 0.73)
q4	(0.515, 0.931, -0.7)
	...

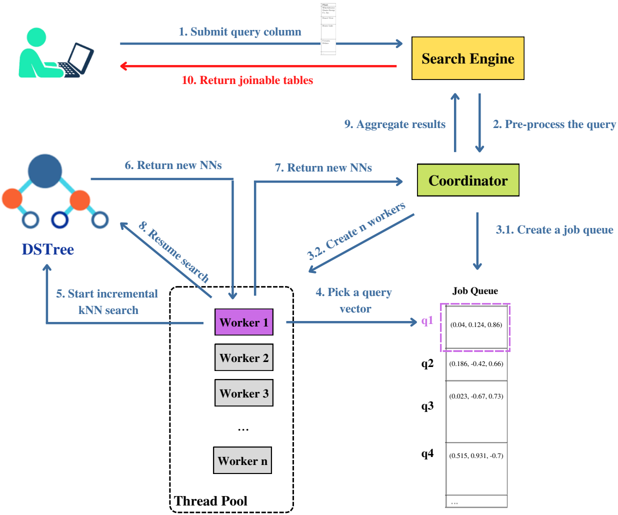
Kashif : Incremental Joinable Table Search (Parallel kNN Search)



Job Queue

q1	(0.04, 0.124, 0.86)
q2	(0.186, -0.42, 0.66)
q3	(0.023, -0.67, 0.73)
q4	(0.515, 0.931, -0.7)
	...

Kashif : Incremental Joinable Table Search (Parallel kNN Search)



Job Queue	
q1	(0.04, 0.124, 0.86)
q2	(0.186, -0.42, 0.66)
q3	(0.023, -0.67, 0.73)
q4	(0.515, 0.931, -0.7)
	...

Kashif : Performance

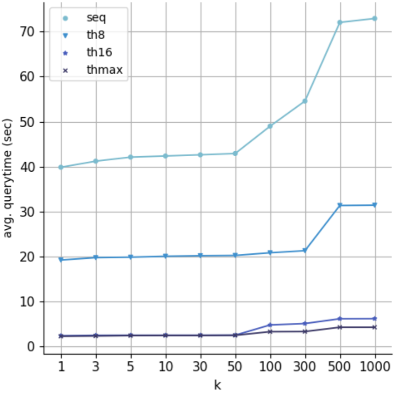


Figure – Kashif Mean Query time [25M vectors, 10 queries, query size = 50 - 100]

Thesis Goals and Future Work

- Provide a systematic literature review of all proposed data discovery frameworks.
- Compare Kashif performance against other frameworks in the literature.
- Extend our work to support search over other data modalities (e.g. Images, Audio files etc.).

Thank you for your attention

References

[1] E. Zhu, D. Deng, F. Nargesian, and R. J. Miller, “Josie : Overlap set similarity search for finding joinable tables in data lakes,” *Proceedings of the 2019 International Conference on Management of Data*, p. 847–864, 2019.

[2] E. Zhu, F. Nargesian, K. Q. Pu, and R. J. Miller, “Lsh ensemble : Internet-scale domain search,” *VLDB Endowment*, vol. 9, no. 12, p. 1185–1196, 2016.

[3] Y. Zhang and Z. G. Ives, “Finding related tables in data lakes for interactive data science,” *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 1951–1966, 2020.

[4] Y. Dong, K. Takeoka, C. Xiao, and M. Oyamada, “Efficient joinable table discovery in data lakes : A high-dimensional similarity-based approach,” *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 456–467, 2021.

[5] A. Santos, A. Bessa, C. Musco, and J. Freire, “A sketch-based index for correlated dataset search,” *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 2928–2941, 2022.