



WarpGate: A Semantic Join Discovery System for Cloud Data Warehouses

Tianji Cong¹, James Gale², Jason Frantz², H. V. Jagadish¹, Çağatay Demiralp²

¹University of Michigan, ²Sigma Computing

Background: Data Discovery

Data discovery over

- Web table corpora
- Corporate data lakes
- Open Data repositories

Goal: Find connections between datasets and go beyond keyword search

Observations

Modern organizations also face data discovery challenges when dealing with structured data (i.e., data in cloud data warehouses).

- In large CDWs, few people have a good understanding of data
- Data is added in a much less modeled state
- Business users may not have the expertise to navigate data
- There are valuable relationships not defined by the relational model

Background: Sigma

The screenshot displays the Sigma database interface. On the left is a navigation sidebar with a search bar and a list of categories including EXAMPLES, ACCOUNT, CAMPAIGN, and USER. The main area shows a table with columns: Name, Billing Street, Billing City, Billing State, Billing Postal Code, and Industry. The table contains 29 rows of data for various companies. At the bottom of the table, it indicates '24,752 Rows - 29 Columns'. The interface also includes a top navigation bar with 'Edit' and 'Explore' buttons, and a user profile icon labeled 'JG'.

Name	Billing Street	Billing City	Billing State	Billing Postal Code	Industry
Affiliated Managers Group, Inc.	183 Carey Park	Pueblo	Colorado	81015	Investment Managers
J P Morgan Chase & Co	968 Forster Center	Chicago	Illinois	60686	n/a
Blackrock MuniYield Quality Fund, Inc.	8 Northridge Avenue	Evansville	Indiana	47705	n/a
Atlantica Yield plc	7 Arapahoe Pass	Rochester	New York	14646	Electric Utilities: Central
Stantec Inc	9857 Tennessee Pass	Washington	District of Columbia	20029	Military/Government/Technical
Quaker Chemical Corporation	745 Leroy Point	Chicago	Illinois	60691	Major Chemicals
VictoryShares US 500 Enhanced Volatility Wtd ETF	17 Sunfield Pass	Hamilton	Ohio	45020	n/a
NAPCO Security Technologies, Inc.	0841 Annamark Pass	Fresno	California	93726	Telecommunications Equipment
Customers Bancorp, Inc	9 Basil Park	Washington	District of Columbia	20420	n/a
The Rubicon Project, Inc.	5 Miller Avenue	Fresno	California	93773	Computer Software: Programming, D
Dicerna Pharmaceuticals, Inc.	30 Mendota Pass	Louisville	Kentucky	40233	Major Pharmaceuticals
Olin Corporation	9 Prairieview Way	Kansas City	Missouri	64193	Major Chemicals
Cubic Corporation	3 Loeprich Court	Fort Myers	Florida	33906	Industrial Machinery/Components
PartnerRe Ltd.	187 Dovetail Pass	Sacramento	California	94263	n/a
Autobyte Inc.	0913 Kim Crossing	Johnstown	Pennsylvania	15906	Computer Software: Programming, D
K2M Group Holdings, Inc.	5 Meadow Vale Court	West Hartford	Connecticut	06127	Medical/Dental Instruments
Movado Group Inc.	91 Vernon Junction	Minneapolis	Minnesota	55458	Consumer Specialties
Ascendis Pharma A/S	713 Eagan Court	Washington	District of Columbia	20022	Major Pharmaceuticals
BB&T Corporation	17 Sommers Terrace	Boise	Idaho	83757	n/a
National CineMedia, Inc.	9 Hoard Hill	El Paso	Texas	88569	Advertising
Knight Transportation, Inc.	97459 Bellfuss Crossing	Pensacola	Florida	32505	Trucking Freight/Courier Services
Gladstone Investment Corporation	14556 Dunning Plaza	Washington	District of Columbia	20557	n/a
J P Morgan Chase & Co	7 Colorado Trail	San Francisco	California	94159	n/a
Provident Financial Services, Inc	0 Summerview Hill	Memphis	Tennessee	38197	Savings Institutions
Shinhan Financial Group Co Ltd	50477 Jackson Place	Albuquerque	New Mexico	87201	Major Banks

Discovery Need in Sigma - LOOKUP

The screenshot shows the Sigma software interface. On the left is a sidebar with a tree view of columns for an 'ACCOUNT' table, including 'Name', 'Billing City', 'Billing State', and 'Billing Postal Code'. The main area displays a data table with columns: 'Name', 'Billing City', 'Billing State', and 'Billing Postal Code'. The 'Name' column contains various company names like 'Affiliated Managers Group, Inc.' and 'J P Morgan Chase & Co'. The 'Billing City' column lists cities like 'Pueblo', 'Chicago', and 'Evansville'. The 'Billing State' column lists states like 'Colorado', 'Illinois', and 'Indiana'. The 'Billing Postal Code' column lists codes like '81015', '60686', and '47705'. An 'ADD LOOKUP' dialog box is overlaid on the table, with a red box highlighting it. The dialog has the following content:

ADD LOOKUP

1. Which column would you like to add?
Select element
Select source

Column to add: Select column
Aggregate: None

2. Map two elements
Select key columns that have matching values in both sources

ACCOUNT Source not selected
+ Add another mapping

KEYS WITH MATCHES
KEYS WITH MULTIPLE MATCHES

Buttons: Cancel, Done

Discovery Need in Sigma - JOIN

SOURCES +

- ACCOUNT
- COMPANIES

FINAL OUTPUT

2 sources, 50 columns

Join with Selected source

ACCOUNT COMPANIES

Join type

Left outer join

Assume referential integrity [Learn more](#)

Join keys

abc Name = abc Name

Create Join

Cancel Preview Output

JOIN OUTPUT

Is Deleted	Master Record Id	Name	Type	Parent Id	Billing Street	Billing City	Billing State
null	null	Affiliated Managers Group, Inc.	null	null	183 Carey Park	Pueblo	Colorado
null	null	J P Morgan Chase & Co	null	null	968 Forster Center	Chicago	Illinois
null	null	Blackrock MuniYield Quality Fund, Inc.	null	null	8 Northridge Avenue	Evansville	Indiana
null	null	Atlantica Yield plc	null	null	7 Arapahoe Pass	Rochester	New York
null	null	Stantec Inc	null	null	9857 Tennessee Pass	Washington	District of Columbia
null	null	Quaker Chemical Corporation	null	null	745 Leroy Point	Chicago	Illinois
null	null	VictoryShares US 500 Enhanced Volatility Wtd ETF	null	null	17 Sunfield Pass	Hamilton	Ohio
null	null	NAPCO Security Technologies, Inc.	null	null	0841 Annamark Pass	Fresno	California
null	null	Customers Bancorp, Inc	null	null	9 Basil Park	Washington	District of Columbia
null	null	The Rubicon Project, Inc.	null	null	5 Miller Avenue	Fresno	California
null	null	Dicerna Pharmaceuticals, Inc.	null	null	30 Mendota Pass	Louisville	Kentucky
null	null	Olin Corporation	null	null	9 Prairieview Way	Kansas City	Missouri
null	null	Cubic Corporation	null	null	3 Loeprich Court	Fort Myers	Florida
null	null	PartnerRe Ltd.	null	null	187 Dovetail Pass	Sacramento	California
null	null	Autobytel Inc.	null	null	0913 Kim Crossing	Johnstown	Pennsylvania
null	null	K2M Group Holdings, Inc.	null	null	5 Meadow Vale Court	West Hartford	Connecticut
null	null	Meadow Group Inc.	null	null	01 Meadow Justice	Mississippi	Mississippi

6

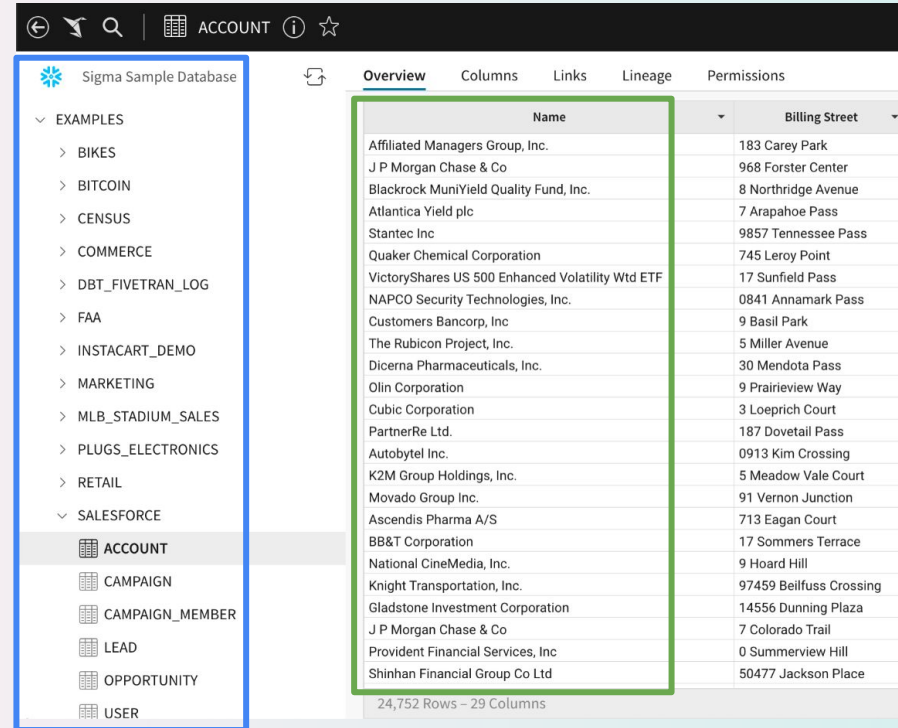
Motivations

- There is a need for surfacing relationships to help business users connect related tables within and across schemas
- Many aspects of data discovery, particularly those pertaining to enterprise settings, are less explored
 - Existing solutions assume (at least) a full pass of underlying data
 - How to estimate joinability without access to full data
 - Value to work flows of business users

Problem Definition

Given a corpus of tables S , a query column c_q from a table, and a constant k , find up to k candidate columns from S in descending order of semantic column joinability.

Hypothesis: Embeddings can encapsulate column semantics and serve as a proxy for their join-ability.

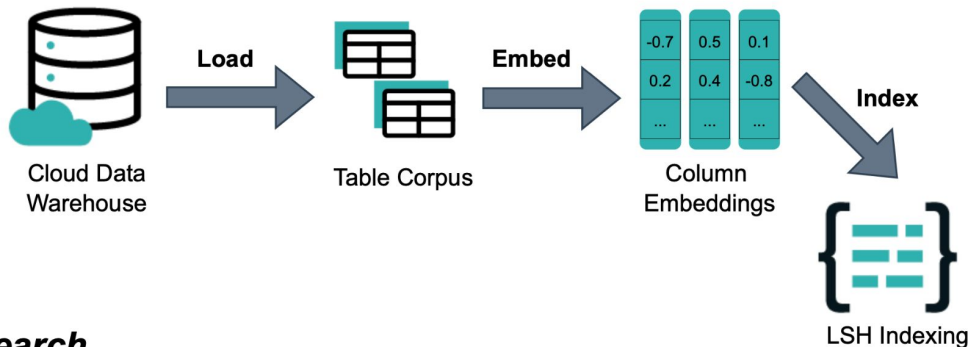


The screenshot shows a database interface with a table of company names and billing streets. The table is titled 'ACCOUNT' and has 24,752 rows and 29 columns. The table is highlighted with a green border. The left sidebar shows a navigation menu with 'ACCOUNT' selected.

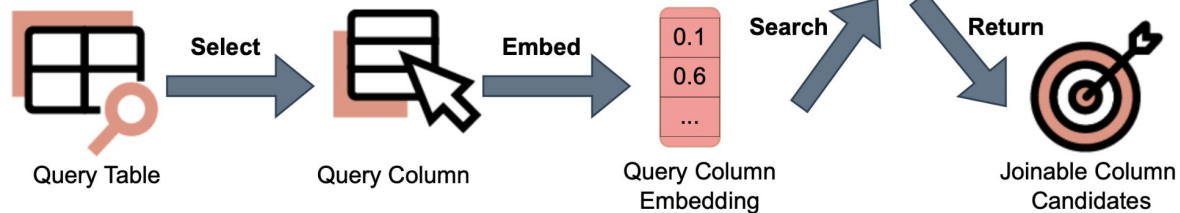
Name	Billing Street
Affiliated Managers Group, Inc.	183 Carey Park
J P Morgan Chase & Co	968 Forster Center
Blackrock MuniYield Quality Fund, Inc.	8 Northridge Avenue
Atlantica Yield plc	7 Arapahoe Pass
Stantec Inc	9857 Tennessee Pass
Quaker Chemical Corporation	745 Leroy Point
VictoryShares US 500 Enhanced Volatility Wtd ETF	17 Sunfield Pass
NAPCO Security Technologies, Inc.	0841 Annamark Pass
Customers Bancorp, Inc	9 Basil Park
The Rubicon Project, Inc.	5 Miller Avenue
Dicerna Pharmaceuticals, Inc.	30 Mendota Pass
Olin Corporation	9 Prairieview Way
Cubic Corporation	3 Loeprich Court
PartnerRe Ltd.	187 Dovetail Pass
Autobyte Inc.	0913 Kim Crossing
K2M Group Holdings, Inc.	5 Meadow Vale Court
Movado Group Inc.	91 Vernon Junction
Ascendis Pharma A/S	713 Eagan Court
BB&T Corporation	17 Sommers Terrace
National CineMedia, Inc.	9 Hoard Hill
Knight Transportation, Inc.	97459 Bellfuss Crossing
Gladstone Investment Corporation	14556 Dunning Plaza
J P Morgan Chase & Co	7 Colorado Trail
Provident Financial Services, Inc	0 Summerview Hill
Shinhan Financial Group Co Ltd	50477 Jackson Place

Solution Overview: WarpGate

Indexing



Search



Solution Overview: Column Embeddings

The embedding representations depend on the embedding model

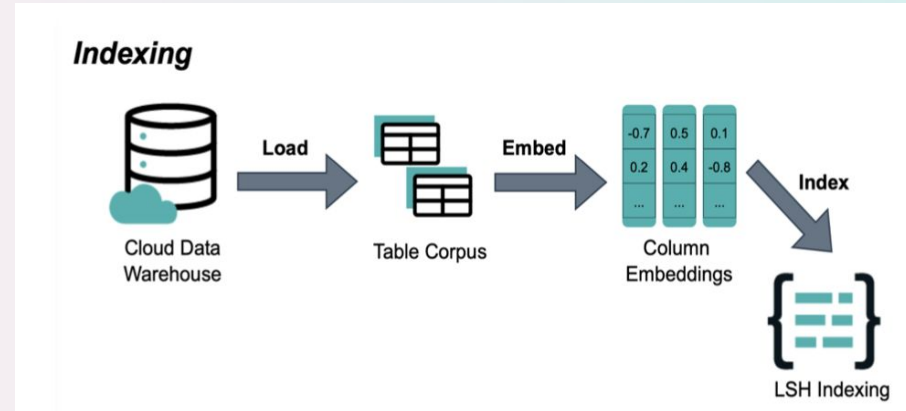
- Whether the model is (pre-)trained over tabular data
- The size of the training corpus
- The efficiency of model inference

Challenge: How to adapt embeddings for data and join discovery?

- Self-supervised contrastive learning

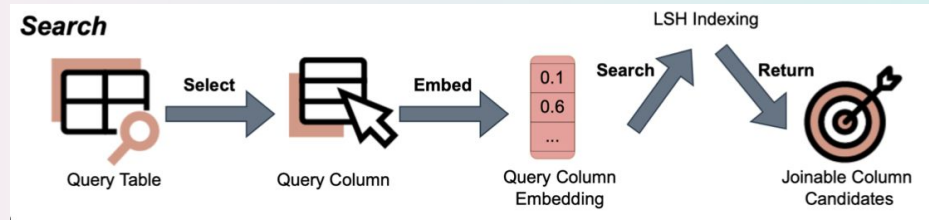
Solution Overview: Indexing

- Index: Locality-sensitive hashing (LSH)^[1]
- Basic idea: Maximize hash collisions for similar inputs
- LSH to approximate cosine similarity: random projection^[2]



Solution Overview: Search

- A user only needs to select a column of interest
- The query column is embedded and hashed
- Search within the sub-universe of embedding vectors with the same hash



Solution Overview: WarpGate Interface

Name	Type	Parent Id
Affiliated Managers Group		
J P Morgan Chase & Co		
Blackrock MuniYield Quality		
Atlantica Yield plc		
Stantec Inc		
Quaker Chemical Corporation		
VictoryShares US 500 Enhanced		
NAPCO Security Technologies		
Customers Bancorp, Inc		
The Rubicon Project, Inc.		
Dicerna Pharmaceuticals, Inc.		
Olin Corporation		
Cubic Corporation		
PartnerRe Ltd.		
Autobytel Inc.		
K2M Group Holdings, Inc.		
Movado Group Inc.		
Ascendis Pharma A/S		
BB&T Corporation		
National CineMedia, Inc.		
Knight Transportation, Inc.		
Gladstone Investment Corporation		
J P Morgan Chase & Co		
Provident Financial Services		

Context menu options:

- Sort
- Filter
- Add new column
- Add column via lookup...
- Duplicate column
- Rename column
- Set description
- Hide column
- Delete column
- Freeze up to column
- Group column
- Column details...
- Transform
- Conditional formatting

Step 1: Right click a column and choose the discovery option

Add lookup

1. What column would you like to add?

Select source

Value source

RECOMMENDED SOURCES
(Matching with Column: ACCOUNT / NAME)

- LEAD / COMPANY 99.6%
- INDUSTRIES / COMPANY NAME 90.9%
- COMPANIES / NAME 88.8%

MANUALLY SELECT A SOURCE

Page 1

Cancel Done

Name	Type	Parent Id
Affiliated Managers Group, ...		
J P Morgan Chase & Co		
Blackrock MuniYield Quali...		
Atlantica Yield plc		
Stantec Inc		
Quaker Chemical Corporati...		
VictoryShares US 500 Enha...		
NAPCO Security Technolog...		
Customers Bancorp, Inc		
The Rubicon Project, Inc.		
Dicerna Pharmaceuticals, I...		
Olin Corporation		
Cubic Corporation		
PartnerRe Ltd.		
Autobytel Inc.		
K2M Group Holdings, Inc.		
Movado Group Inc.		
Ascendis Pharma A/S		
BB&T Corporation		
National CineMedia, Inc.		
Knight Transportation, Inc.		
Gladstone Investment Corp...		
J P Morgan Chase & Co		
Provident Financial Service...		
Shinhan Financial Group C...		
Galena Biopharma, Inc.		
Eastern Company (The)		
Otter Tail Corporation		

Step 2: Select a recommendation

Name	Company (LEAD)
Managers Group, ...	Affiliated Managers Group...
an Chase & Co	J P Morgan Chase & Co
k MuniYield Quali...	Blackrock MuniYield Quali...
Yield plc	Atlantica Yield plc
nc	Stantec Inc
hemical Corporati...	Quaker Chemical Corporat...
ares US 500 Enha...	VictoryShares US 500 Enh...
ecurity Technolog...	NAPCO Security Technolo...
rs Bancorp, Inc	Customers Bancorp, Inc
on Project, Inc.	The Rubicon Project, Inc.
harmaceuticals, I...	Dicerna Pharmaceuticals, ...
oration	Olin Corporation
rporation	Cubic Corporation
e Ltd.	PartnerRe Ltd.
l Inc.	Autobytel Inc.
up Holdings, Inc.	K2M Group Holdings, Inc.
Group Inc.	Movado Group Inc.
Pharma A/S	Ascendis Pharma A/S
rporation	BB&T Corporation
CineMedia, Inc.	National CineMedia, Inc.
ansportation, Inc.	Knight Transportation, Inc.
e Investment Corp...	Gladstone Investment Cor...
an Chase & Co	J P Morgan Chase & Co
t Financial Service...	Provident Financial Servic...

Step 3: Add a recommended column

Experiments: Setup

- Datasets

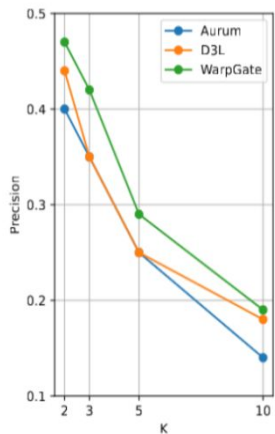
- NextiaJD (XS, S, M, L)
 - Across-schema joins
 - Effects of dataset size
- Spider dataset
 - PK/FK detection
- Sigma Sample Database
 - Ad-hoc discovery

- Baselines

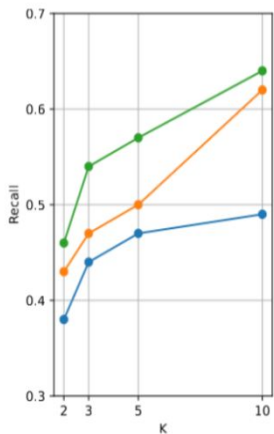
- Aurum
- D³L

	# Tables	# Columns	Avg. # Rows	# Queries	Avg. # Answers
<i>XS</i>	28	257	1,938	35	2.8
<i>S</i>	46	2,553	209,646	177	3.6
<i>M</i>	46	1,067	3,175,904	188	4.4
<i>L</i>	19	541	12,288,165	92	3.6
<i>Spider</i>	70	429	7632	60	1.1
<i>Sigma</i>	98	1,343	2,243,932	TBD	N/A

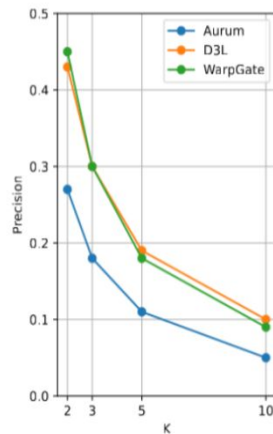
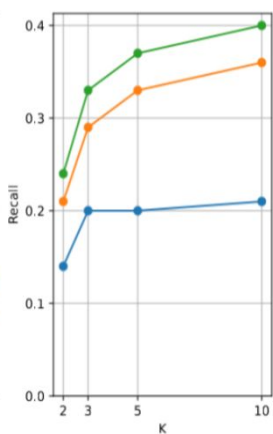
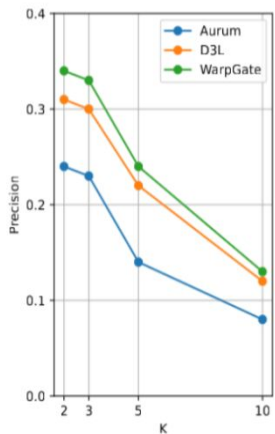
Experiments: Effectiveness



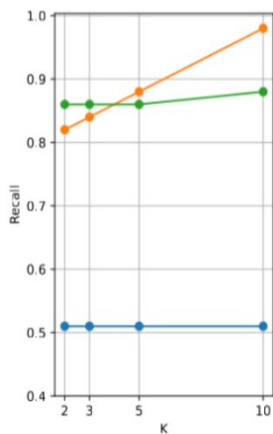
(a) testbedS



(b) testbedM



(c) Spider



- Across-schema joins: WarpGate consistently obtains higher precision and recall compared with two baselines as k increases
- PK/FK detection: WarpGate compares favorably with the ensemble approach (D³L)

Experiments: Efficiency

- Scanning entire data is expensive and won't give us interactive speed
- Index lookup time (in parentheses) is only a portion of end-to-end time

Dataset	D ³ L	WarpGate
testbedS	4.77	3.12 (1.04)
testbedM	57.69	38.73 (8.39)

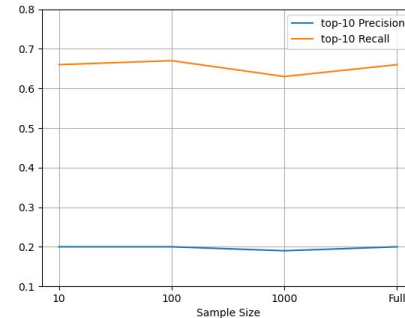
s / query

Experiments: Sample Efficiency

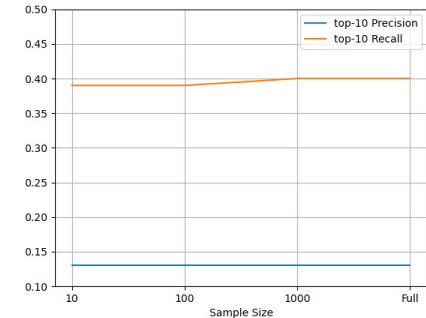
- Random sampling significantly reduces index lookup time & e2e time
- Need as few as 10 samples to get effective embeddings

Sample Size	testbedS	testbedM
10	32 (20)	50 (20)
100	43 (20)	59 (20)
1000	65 (20)	82 (30)

ms / query



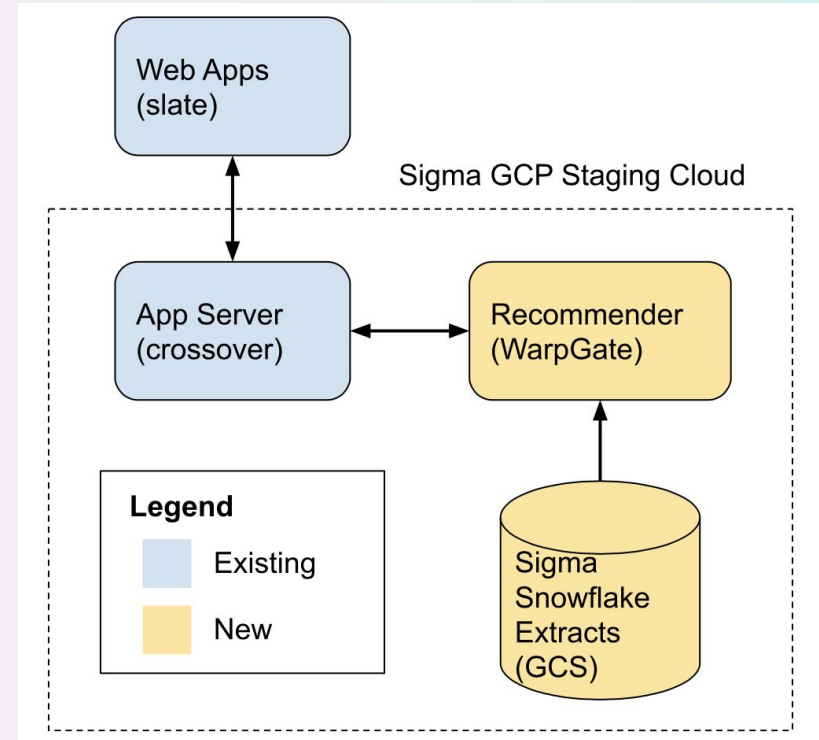
testbedS



testbedM

Roadmap

- **Current Stage**
 - Deploy WarpGate in Sigma staging environment
 - Conduct user study of Sigma internal users
- **Challenges**
 - Envision cpu and memory usage when deploying in containers over K8s
 - How to quickly rebuild the index when service is down



Conclusion

- We present WarpGate, a prototype for semantic join discovery in Sigma Workbooks
- We show that the embedding approach is both effective and sample efficient
- We expect to share more of our experience and lessons of deploying WarpGate in the production environment and evaluate its values to end users