SELECT STAR

Data Discovery: getting more with metadata

Shinji Kim

Data Council Austin, March 2022

For the last 13 yrs, I've been in the shoes of the data producers and consumers

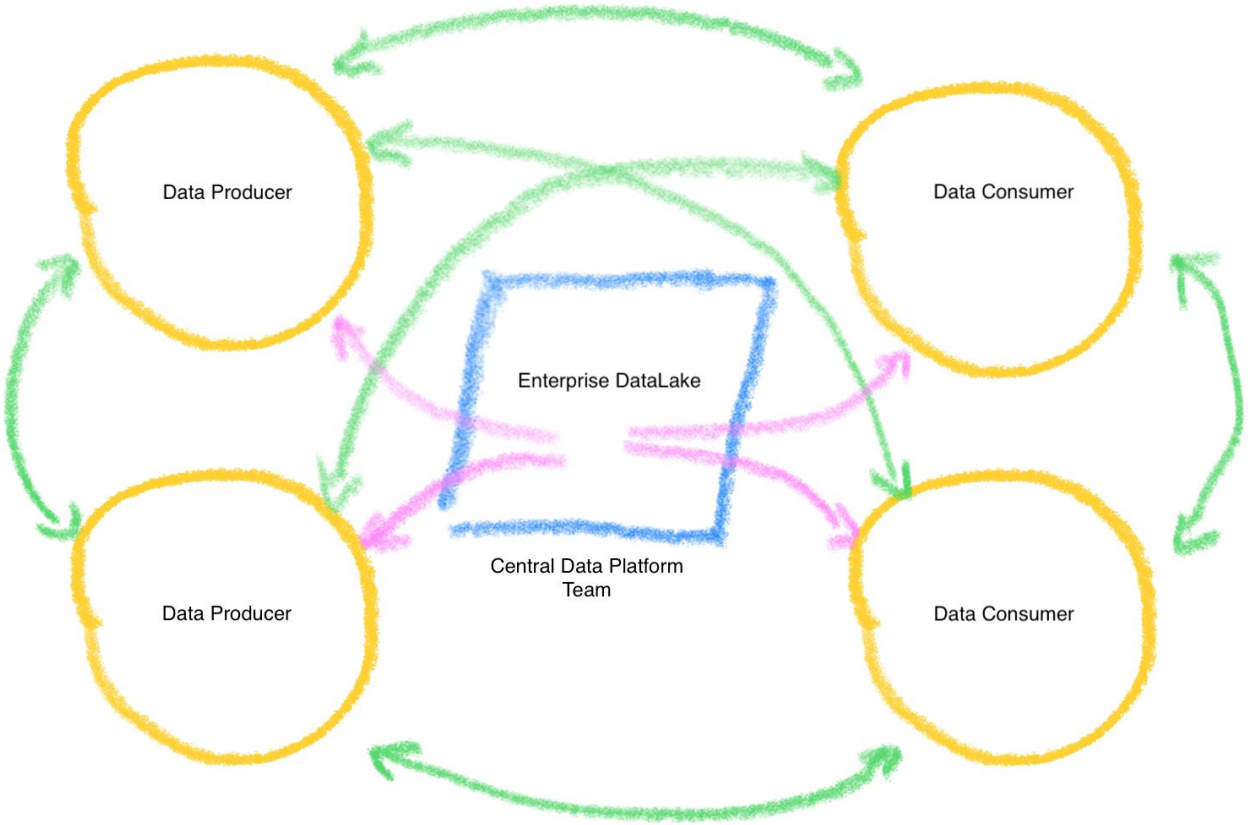
Shinji Kim
Founder & CEO

- Software Engineering, University of Waterloo
- Statistical Analyst, Sun Microsystems Research Labs - Sales Forecasting
- Development DBA, Barclays Capital - Lead Architect, Global IT Database Consolidation
- Growth - Internet Marketing, Facebook - SEM Campaign Optimization
- S&O / Technology Strategy, Deloitte Consulting
- Founder, ShufflePix
- Product Manager, Yieldmo
- Founder & CEO, Concord Systems (acquired by Akamai)
- Sr. Manager, Product Management, Akamai Technologies - IoT Edge Connect

What is Data Discovery?

Find & Understand Data

How can we understand what's going on in the data warehouse?



Source: https://aws.amazon.com/blogs/industries/how-to-create-a-modern-cpg-data-architecture-with-data-mesh/

Metadata + Query Logs provides Context

Metadata + Query Logs provides Context

Metadata = What exists in the data warehouse?

OLIST ☆ 5 schemas, 62 tables

This is a Brazilian ecommerce public dataset of orders made at Olist Store.

Edit

Table (62)	Description
DATASETS.ORDERS	Order information: This is the table contains informatio
DATASETS.ORDER_ITEMS	This dataset includes data about the items purchased v
DATASETS.ORDER_PAYMENTS	
DATASETS.CLOSED_DEALS	After a qualified lead fills in a form at a landing page he Show more
DATASETS.CUSTOMERS	This dataset has information about the customer and it Show more
DATASETS.PRODUCTS	This dataset includes data about the products sold by C
DATASETS.ORDER_REVIEWS	This dataset includes data about the reviews made by t
ANALYTICS.CITY_FIELDS	Which cities contribute the most in terms of customer t
DATASETS.PRODUCT_CATEGORY_NAME_TRANSLATION	This dataset has information Brazilian zip codes and its Show more
DATASETS.SELLERS	This dataset includes data about the sellers that fulfille Show more

Metadata + Query Logs provides Context

Metadata = What exists in the data warehouse?

OLIST 5 schemas, 62 tables

This is a Brazilian ecommerce public dataset of orders made at Olist Store.

Edit

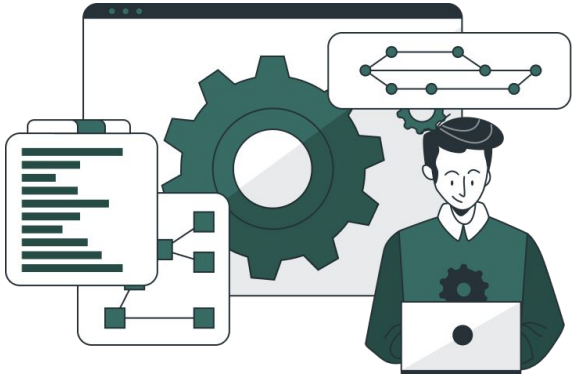
Table (62)	Description
DATASETS.ORDERS	Order information: This is the table contains informatio
DATASETS.ORDER_ITEMS	This dataset includes data about the items purchased v
DATASETS.ORDER_PAYMENTS	
DATASETS.CLOSED_DEALS	After a qualified lead fills in a form at a landing page he Show more
DATASETS.CUSTOMERS	This dataset has information about the customer and it Show more
DATASETS.PRODUCTS	This dataset includes data about the products sold by C
DATASETS.ORDER_REVIEWS	This dataset includes data about the reviews made by t
ANALYTICS.CITY_FIELDS	Which cities contribute the most in terms of customer t
DATASETS.PRODUCT_CATEGORY_NAME_TRANSLATION	This dataset has information Brazilian zip codes and its Show more
DATASETS.SELLERS	This dataset includes data about the sellers that fulfille Show more

Query logs = What has happened to the data objects?

Query (17325)	User
SELECT "source"."substring7149" AS "substring7149", "source"."NEWCOL1" AS "NEWCOL1", "source"."NEWCOL4" AS "NEWCOL4", "source"."NEWCOL3"...	SELECTSTAR_TEST
SELECT "source"."substring7144" AS "substring7144", "source"."NEWCOL1" AS "NEWCOL1", "source"."NEWCOL4" AS "NEWCOL4", "source"."NEWCOL3"...	SELECTSTAR_TEST
SELECT "DBT_TEST"."CIRCULAR_TABLE1"."COL1" AS "COL1" FROM "OLIST"."DBT_TEST"."CIRCULAR_TABLE1" LIMIT 10000	SELECTSTAR_TEST
SELECT "DATASETS"."TABLE_IF_NO_EXISTS"."COL1" AS "COL1" FROM "OLIST"."DATASETS"."TABLE_IF_NO_EXISTS" LIMIT 10000	SELECTSTAR_TEST
SELECT "source"."substring7139" AS "substring7139" FROM (SELECT "DATASETS"."TEST3"."OTHER_COL" AS "OTHER_COL", substring("DATASETS"."TE...	SELECTSTAR_TEST
SELECT "source"."substring7137" AS "substring7137" FROM (SELECT "DATASETS"."B"."COL2" AS "COL2", substring("DATASETS"."B"."COL2", 1, 1234) A...	SELECTSTAR_TEST
SELECT "DBT_TEST"."CIRCULAR_TABLE2"."COL1" AS "COL1" FROM "OLIST"."DBT_TEST"."CIRCULAR_TABLE2" LIMIT 10000	SELECTSTAR_TEST
SELECT "PUBLIC"."PUBLIC_TEST"."COL1" AS "COL1" FROM "OLIST"."PUBLIC"."PUBLIC_TEST" LIMIT 10000	SELECTSTAR_TEST

Understanding of your data can come from analyzing your metadata & query history

1. What data exists today?
2. Where can I find the data?
3. What does this data represent?
4. Where did it come from?
5. Is it up to date?
6. Do others also use this data today?
7. Where is it being used?
8. How is it being used today?
9. What are other related dataset?
10. Who should I ask questions about this?



What can you answer with Metadata + Query log about your database?

1. Data Usage
2. Data Freshness
3. User Behavior
4. User Behavior + Cost
5. Data Dependencies

What can you answer with Metadata + Query log about your database?

1. Data Usage

SNOWFLAKE.ACCOUNT_USAGE.TABLES

Column Name
TABLE_ID
TABLE_NAME
TABLE_SCHEMA_ID
TABLE_SCHEMA
TABLE_CATALOG_ID
TABLE_CATALOG

DATABASE	SCHEMA	TABLE COUNT
DWH	IOS	20
DWH	ANDROID	24
DWH	WEB	25
FIVETRAN	SALESFORCE	50
DATAMART	CUSTOMER	102
DATAMART	PRICING	245
DATAMART	PRICING_V2	0
DATAMART	ADS	122

Empty schema

What can you answer with Metadata + Query log about your database?

1. Data Usage

SNOWFLAKE.ACCOUNT_USAGE.TABLES

Column Name
TABLE_ID
TABLE_NAME
TABLE_SCHEMA_ID
TABLE_SCHEMA
TABLE_CATALOG_ID
TABLE_CATALOG

SNOWFLAKE.ACCOUNT_USAGE.QUERY_HISTORY

Column Name
QUERY_ID
QUERY_TEXT
DATABASE_NAME
SCHEMA_NAME
QUERY_TYPE

DATABASE	SCHEMA	TABLE COUNT	QUERY COUNT (SELECT)
DWH	IOS	20	298
DWH	ANDROID	24	0
DWH	WEB	25	234
FIVETRAN	SALESFORCE	50	31
DATAMART	CUSTOMER	102	12
DATAMART	PRICING	245	25
DATAMART	PRICING_V2	0	0
DATAMART	ADS	122	24

What can you answer with Metadata + Query log about your database?

2. Data Freshness

Did the data arrive successfully?

SNOWFLAKE.ACCOUNT_USAGE.TABLES

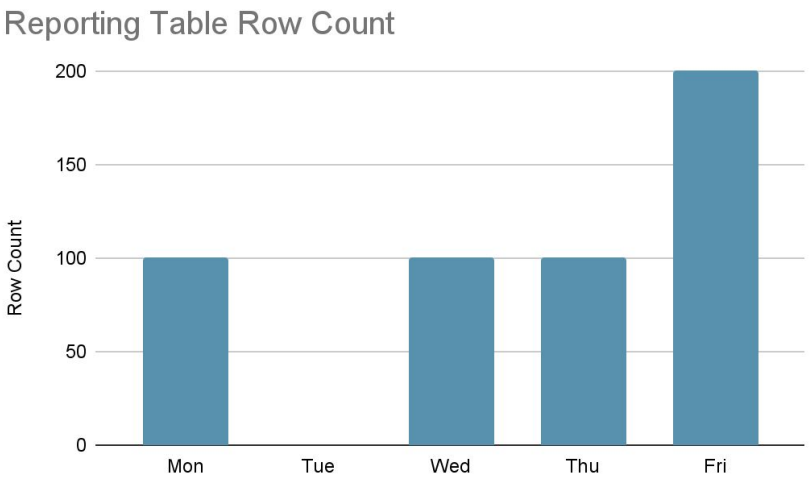
Column Name	Data Type	Description
TABLE_ID	NUMBER	Internal, Snowflake-generated identifier for the table.
TABLE_NAME	TEXT	Name of the table.
TABLE_SCHEMA_ID	NUMBER	Internal, Snowflake-generated identifier of the schema for the table.
TABLE_SCHEMA	TEXT	Schema that the table belongs to.
TABLE_CATALOG_ID	NUMBER	Internal, Snowflake-generated identifier of the database for the table.
TABLE_CATALOG	TEXT	Database that the table belongs to.
ROW_COUNT	NUMBER	Number of rows in the table.
BYTES	NUMBER	Number of bytes accessed by a scan of the table.
RETENTION_TIME	NUMBER	Number of days that historical data is retained for Time Travel.

What can you answer with Metadata + Query log about your database?

2. Data Freshness

Did the data arrive successfully?

→ Did the data get replaced / computed successfully?



What can you answer with Metadata + Query log about your database?

2. Data Freshness

Did the data arrive successfully?

SNOWFLAKE.ACCOUNT_USAGE.TABLES

Column Name	Data Type	Description
TABLE_ID	NUMBER	Internal, Snowflake-generated identifier for the table.
TABLE_NAME	TEXT	Name of the table.
TABLE_SCHEMA_ID	NUMBER	Internal, Snowflake-generated identifier of the schema for the table.
TABLE_SCHEMA	TEXT	Schema that the table belongs to.
TABLE_CATALOG_ID	NUMBER	Internal, Snowflake-generated identifier of the database for the table.
TABLE_CATALOG	TEXT	Database that the table belongs to.
CREATED	TIMESTAMP_LTZ	Date and time when the table was created.
LAST_ALTERED	TIMESTAMP_LTZ	Date and time when the table was last altered by a DDL or DML operation.
DELETED	TIMESTAMP_LTZ	Date and time when the table was dropped.
COMMENT	TEXT	Comment for the table.

What can you answer with Metadata + Query log about your database?

2. Data Freshness

Did the data arrive successfully?

SNOWFLAKE.ACCOUNT_USAGE.QUERY_HISTORY

Column Name	Data Type	Description
QUERY_ID	TEXT	The statement's unique id.
QUERY_TEXT	TEXT	Text of the SQL statement.
DATABASE_NAME	TEXT	Database that was in use at the time of the query.
SCHEMA_NAME	TEXT	Schema that was in use at the time of the query.
QUERY_TYPE	TEXT	DML, query, etc. If the query is currently running, or the query failed, then the query type may be UNKNOWN.
START_TIME	TIMESTAMP_LTZ	Statement start time
END_TIME	TIMESTAMP_LTZ	Statement end time. If the query is still running, the END_TIME is the UNIX epoch timestamp ("1970-01-01 00:00:00"), adjusted for the local time zone. E.g. for Pacific Standard Time, this would be "1969-12-31 16:00:00.000 -0800".
TOTAL_ELAPSED_TIME	NUMBER	Elapsed time (in milliseconds)
BYTES_SCANNED	NUMBER	Number of bytes scanned by this statement.
ROWS_PRODUCED	NUMBER	Number of rows produced by this statement.
COMPILATION_TIME	NUMBER	Compilation time (in milliseconds)
EXECUTION_TIME	NUMBER	Execution time (in milliseconds)

What can you answer with Metadata + Query log about your database?

3. User Behavior

SNOWFLAKE.ACCOUNT_USAGE.QUERY_HISTORY

Column Name
QUERY_ID
QUERY_TEXT
DATABASE_NAME
SCHEMA_NAME
QUERY_TYPE
START_TIME
END_TIME
TOTAL_ELAPSED_TIME
BYTES_SCANNED
ROWS_PRODUCED
COMPILATION_TIME
EXECUTION_TIME

USER_NAME	QUERY_TYPE	SUM (COUNT)	SUM (COUNT) %
dbt	SELECT	209	23%
dbt	CREATE	146	16%
dbt	USE	28	3%
dbt Total			42%
DAVE	SELECT	208	23%
LOOKER	SELECT	201	22%
MODE	SELECT	98	11%
JOHN	SELECT	50	5%
ERIKA	SELECT	40	4%

What can you answer with Metadata + Query log about your database?

4. User Behavior + Cost

SNOWFLAKE.ACCOUNT_USAGE.QUERY_HISTORY

Column Name
QUERY_ID
QUERY_TEXT
DATABASE_NAME
SCHEMA_NAME
QUERY_TYPE
START_TIME
END_TIME
TOTAL_ELAPSED_TIME
BYTES_SCANNED
ROWS_PRODUCED
COMPILATION_TIME
EXECUTION_TIME

USER_NAME	QUERY_TYPE	SUM (COUNT)	SUM (COUNT) %	SUM (EXEC TIME)	SUM (EXEC TIME) %
dbt	SELECT	209	23%		
dbt	CREATE	146	16%		
dbt	USE	28	3%		
dbt Total			42%		
DAVE	SELECT	208	23%		
LOOKER	SELECT	201	22%		
MODE	SELECT	98	11%		
JOHN	SELECT	50	5%		
ERIKA	SELECT	40	4%		

What can you answer with Metadata + Query log about your database?

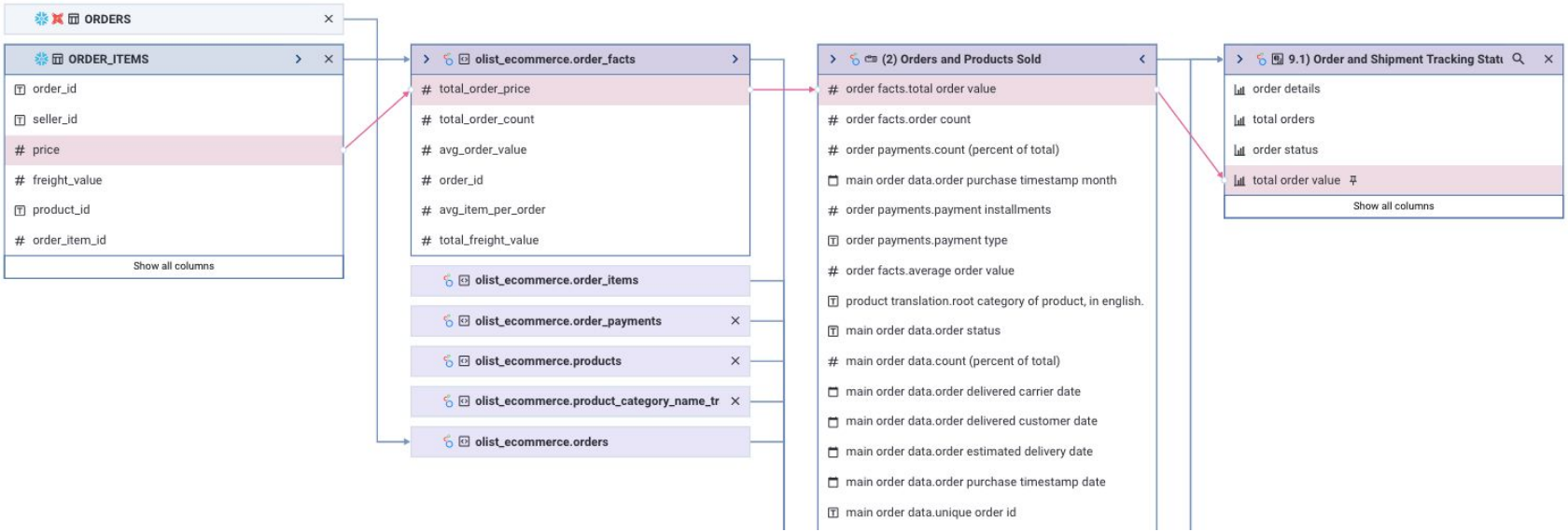
5. Data Dependencies (Data Lineage)

```
GET_DDL('TABLE', 'EMP_COPY')
```

```
CREATE TABLE EMP_COPY as
  SELECT * FROM
  EMPLOYEE.PUBLIC.EMP
where
  DEPARTMENT=10
```

What can you answer with Metadata + Query log about your database?

5. Data Dependencies (Data Lineage)



Friendly Reminder: every database is different

- Usually need Admin-level permissions to access the metadata.
- Although all databases have metadata tables and query logs, the way they provide them may be different.

Snowflake:

- Look under SNOWFLAKE.ACCOUNT_USAGE views
  - ACCOUNT_USAGE.QUERY_HISTORY
  - ACCOUNT_USAGE.TABLES

Redshift:

- Look under systems tables that starts with SVV_ or STL_
  - SVV_TABLE_INFO
  - SVV_TABLES
  - STL_QUERYTEXT
  - STL_DDLTEXT
- Full query log will require enabling audit log & activity logs (turned off by default)

We're just scratching the surface here...

There are more insights you can uncover as you add more metadata

- Who should I notify about this data change?
- What are the tables that require remodeling?
- Which tables need documentation the most?
- and more!

Thank you!

https://selectstar.com

@selectstarhq