

DataHub: Powering LinkedIn's Metadata

Budapest Data Forum
Sep 23, 2020
Mars Lan



“The world’s most valuable
resource is no longer oil, but data.”

—THE ECONOMIST, 2017

The background is a dark blue gradient. It features an abstract pattern of small, scattered squares in various colors (light blue, pink, orange, teal) and thin white vertical lines. The squares are mostly empty outlines, with a few solid-colored ones. The lines are thin and vertical, extending across the width of the image.

“Just like crude oil, data is only
useful after it’s refined.”

—EVERYONE, 2020

MACRO TRENDS

- Data Democratization ⇒ More Organic
 - Data mesh, decentralized data governance, self-service, remote work
- Role Specialization ⇒ More Personas
 - Data scientists/analysts/engineers, AI/ML engineers, business analysts/users, ...
- Explosion of Data Systems & Tools ⇒ More Complexity
 - Hadoop, Spark, Flink, Kafka, Presto, TensorFlow, Elasticsearch, MongoDB
- Adoption of Cloud Computing ⇒ More Data
 - Easier & cheaper than ever to create more data
- Increasing Regulatory Pressure ⇒ More Controls
 - GDPR, CCPA, LGPD, BCBS 239, MiFID II, FRTB, CCAR

PROBLEMS

- Finding data is hard
 - Data lake ⇒ Data swamp
 - Siloed teams ⇒ Siloed data
 - Specialized systems ⇒ Specialized data
- Managing data is hard
 - Governance
 - Compliance
 - Policy-driven data management
- Trusting data is hard
 - Lineage
 - Data availability
 - Data quality & health
 - Data profiling & distribution
 - Ownership & documentation
 - Certification



Metadata,
Metadata,
Metadata

LINKEDIN METADATA JOURNEY

Pull-based (crawlers)
Monolithic app
SQL-based models

OSS WhereHows (V1)

2016

Push-based (Kafka)
App + distributed services
Generalized models

DataHub + GMA (V3)

2019

WhereHows + TMS (V2)

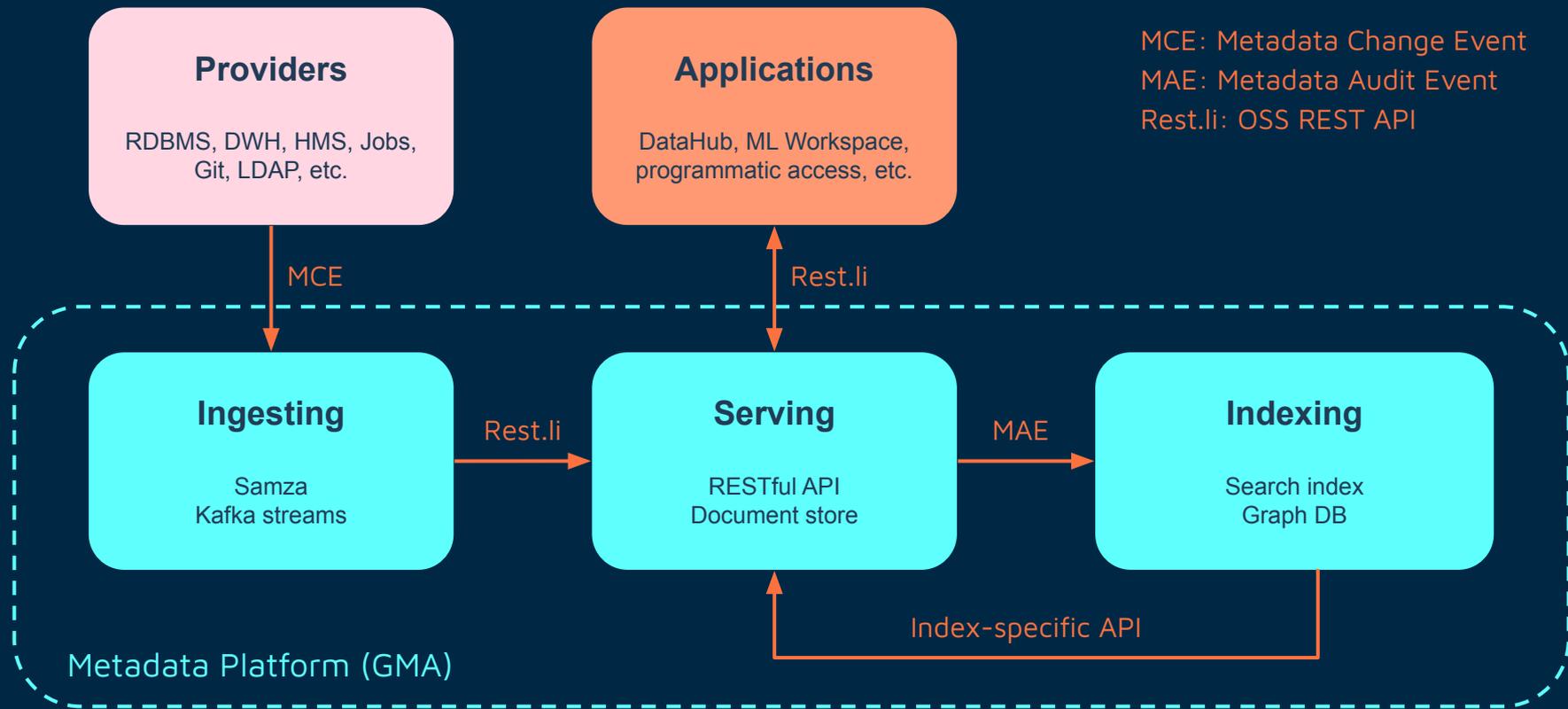
Push-based (Kafka)
App + monolithic service
Opinionated models

2020

OSS DataHub

Docker & K8s
Cloud integration
GraphQL

ARCHITECTURE



METADATA PLATFORM (GMA)

- Scalable

- Web-scale stack
- Distributed storage, indexing & serving
- four/five 9's uptime
- Decentralized metadata modeling

- Enrichable

- Beyond read-only aggregator
- Collaborative edits
- Human curation

- Queryable

- Key-value
- Distributed joins
- Full-text search
- Graph traversal

- Real-time

- Stream-based ingestion
- Event-driven architecture
- Trigger-based applications

WHY STREAMS (KAFKA)?

- Near real-time
 - O(seconds) delay
- Loose coupling
 - Non-blocking, fire-and-forget
- Queuing
 - Smooth out bursty traffic
 - Async consumption
- Schema Compatibility
 - Easy to enforce via schema registry
- Scalable
 - Multi-readers/writers
 - Partition-level parallelism
- Persistent storage
 - Sequential key-value store
 - Bootstrap & backfill

INGESTION MODES

- Existing metadata services

- Crawler: uninstrumentable
- Direct event emission: instrumentable
- Event conversion: existing events

- New metadata services

- DAO: “man-in-the-middle” integration

- Metadata in Git

- Build-time: tooling emits event
- Publish-time: events artifact
- Deploy-time: services/jobs emit events

METADATA MODELING

Nodes

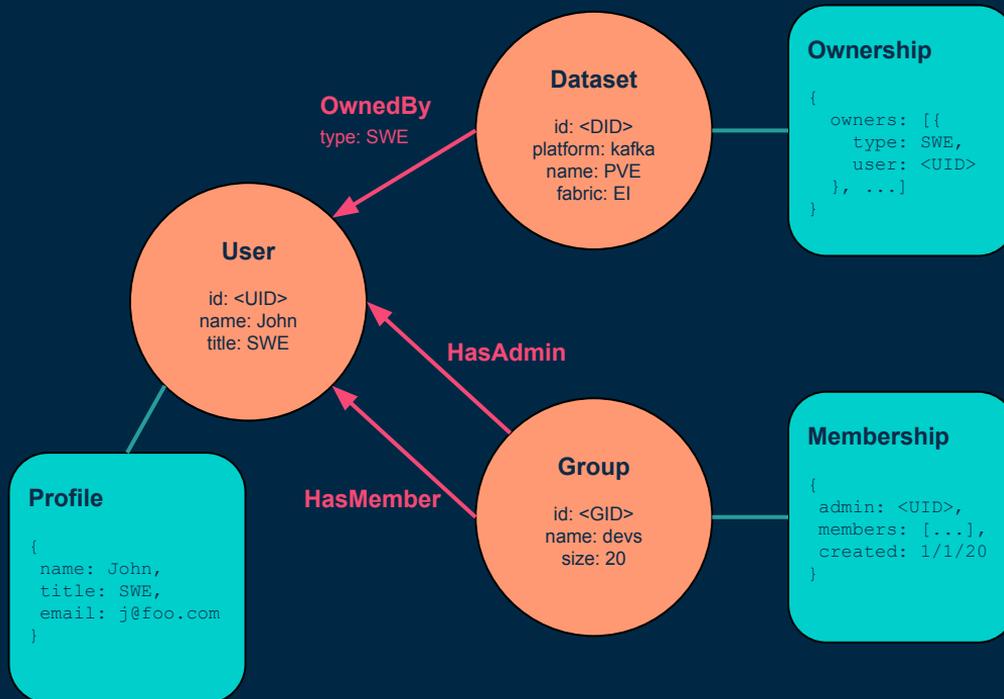
Entities, e.g. datasets, partitions, features, users, groups, experiments, ...

Edges

Relationship, e.g. OwnedBy, DerivedFrom, Contains, HasMember, ...

Documents

Metadata, e.g. ownership, membership, upstreams, configs, compliance metadata, ...



ADOPTION

- LinkedIn (within 18 months)

- Integrated with 40 teams/projects
- 30+ entities, 200+ types of metadata
- Use cases
 - Search & discovery
 - Data privacy compliance
 - Access control
 - Life-cycle management
 - Data Ops
 - AI DevOps

- OSS (within 6 months)

- 7 companies running in production
- 18 companies building POC or seriously evaluating
- Success stories
 - Expedia: Data-driven tech company
 - Saxo Bank: Investment bank going through digital transformation
 - SpotHero: Cloud-native small startup (3 data engs & 50 data users)

OSS ROADMAP

- New Entities & Relationships

- Jobs & flows*
- Dashboards*
- AI features/models*
- Schemas
- Metrics
- Services
- Business glossary

- Integrations

- BI: Looker, Mode, Redash, Superset
- Scheduler: AirFlow, Dagster, Azkaban
- Data Quality: Great Expectations

- Features

- Entity insights
- Data privacy
- Governance
- OIDC & SAML

- Platform

- Strongly consistent secondary index
- Gremlin-based query DAO
- Aspect-specific events
- GraphQL API
- NoSQL backend (e.g. MongoDB)
- OLAP index (e.g. Pinot, Druid)



Data concepts related to "foodie"

[Browse All Data Concepts](#)

Curated information about key data terms presented with relevant context of known data entities.

DATA CONCEPT

Engaged Quality Foodies

A Quality Foodie (QF) is a member who satisfies the criteria of: Reachable - Can be contacted by restaurants about the promotions or new menu item...

Filters

Data Origin

- Prod (15)
- Corp (12)

Platform

- Mysql (24)
- Hdfs (3)

Datasets

Showing 1 - 10 of 27 results

[/demo/teamY/Foodie](#)

Data Origin PROD
 Platform hdfs
 Health 100%

0 - 0

[/demo/teamX/Foodie](#)

Data Origin PROD
 Platform hdfs
 Health 100%

0 - 0

[/nacho/test/Foodie](#)

Data Origin PROD
 Platform hdfs
 Health 100%

0 - 0



Datasets

/demo/rolling_aggregate_orders_2 Hdfs UndefinedFabric: **HOLDEM/WAR**

Health

Last calculated a month ago

100%

[See Details](#)

0 - 0

[Schema](#)[Status](#)[ACL Access](#)[Ownership](#)[Compliance](#)[Dataset Groups](#)[Relationships](#)[Health](#)[Docs](#)

Last Saved: 3 months ago by nkanamar

Owners

Please maintain at least 2 owners.

LDAP Username	Full Name	ID Type	Ownership Type	
nkanamar	Nagarjuna Kanamarlapudi	USER	DataOwner	
pgunnam	Pardhu Gunnam	USER	DataOwner	
+ Add an owner				

Please make changes to save.

Save

[Metrics](#) > [all](#) > [foodie](#) > [app_funnel](#) > [activated](#)

METRICS

activated

number of users activated the app

Owners

[malan, pgunnam](#)Health [®]

Last calculated 5 months ago

100%

[See Details](#)[Overview](#)[Related Entities](#)[Health](#)[Docs](#)

Related entities

This metric is derived from [1 dataset](#) and related to [17 metrics](#)

View the sections below for more details

Derived from datasets (1)

Name	Description	Owners
app_funnel	All lite funnel metrics	pawkumar, jmodi, adchoudh, fmt, kisir

Other metrics from the same dataset (10)

Name	Alias	Description
download	-	number of users downloaded the app
signup_pass_overall	-	number of users signup overall
first_time_signup_pass	-	number of users signup in the first time
first_time_signin	-	number of users first time logged in
first_signin_failed	-	number of users first login failed



Mars Lan

Staff Software Engineer, Systems Infrastructure

[Edit profile](#)

ASK ME ABOUT

[metadata](#)

TL for the metadata team

TEAM

[datahub](#) [gma](#)

MANAGER



Tai Tran

malan@linkedin.com [LinkedIn Profile](#) [Cinco Profile](#)

Access Management

JIT Datasets

Lists

Features

Liked Entities

Followed Entities

Data you might be interested in

Datasets

Ownership

Datasets

Metrics

Charts

Dashboards

ML features

UMP Flows

Metrics Mars Lan Owns

Showing 1 - 10 of 17 results

count_of_events

number of events produced

Bucket	datahub
Formula	COUNT
Dataset	datahub_page_view_event
XLNT Tier	Daily: 2
Tags	datahub, metadata, pageviewevent, route
Frequency	DAILY
Health	100%

count_of_events

number of events produced

Bucket	datahub
Formula	COUNT
Dataset	datahub_search_impression_event
XLNT Tier	Daily: 2
Tags	datahub, metadata, search, impressions
Frequency	DAILY
Health	100%

mie_total_count

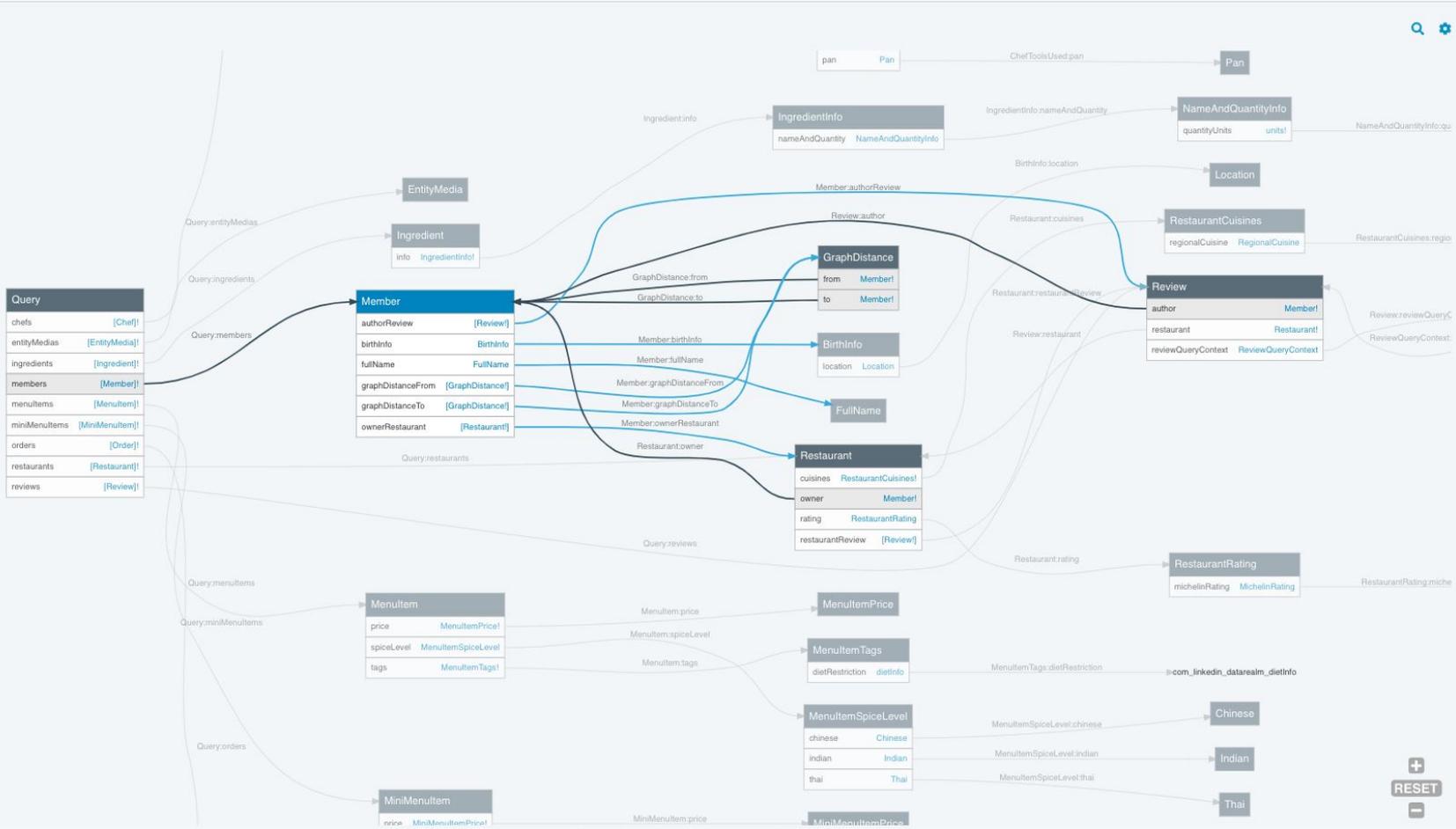
number of mie events

Bucket	tms
--------	-----



Schema Graph View

Schemas [datarealm-restaurants.openapi.graphql](#)



/demo/rolling_aggregate_orders_2
/city
/date
/restaurantName
/no_of_orders
View dataset detail

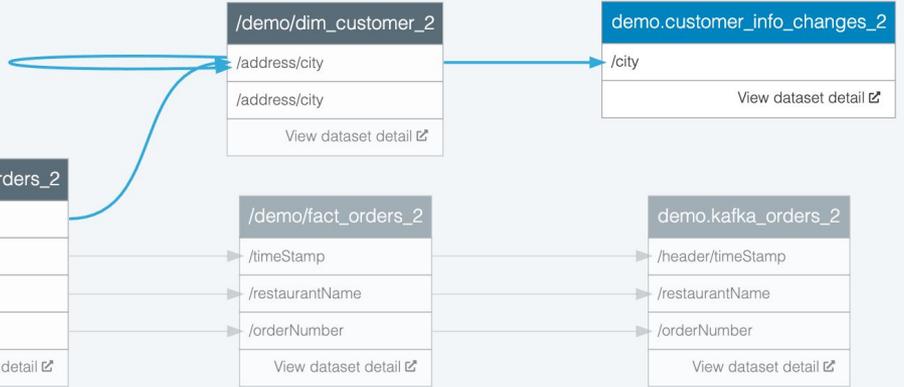
/demo/aggregate_restaurant_orders_2
/city
/date
/restaurantName
/no_of_orders
View dataset detail

/demo/dim_customer_2
/address/city
/address/city
View dataset detail

demo.customer_info_changes_2
/city
View dataset detail

/demo/fact_orders_2
/timeStamp
/restaurantName
/orderNumber
View dataset detail

demo.kafka_orders_2
/header/timeStamp
/restaurantName
/orderNumber
View dataset detail



GitHub: linkedin/datahub

Monthly town hall
Slack workspace
PRs welcome!



Do you have any questions?

malan@linkedin.com
linkedin.com/in/marслан

THANKS

CREDITS: This presentation template was created by [Slidesgo](#),
including icons by [Flaticon](#), and infographics & images by [Freepik](#)
Please keep this slide for attribution