Metadata Use Cases at LinkedIn

Shirshanka Das, Principal Staff Software Engineer, LinkedIn
Projects: LinkedIn DataHub, Apache Gobblin, Dali

@shirshanka

linkedin.com/in/shirshankadas

Community Meeting
Nov 6, 2020

DataHub is really two pieces

DataHub : App

An application for enabling productivity and governance use-cases on top of the metadata mesh

DataHub : GMA

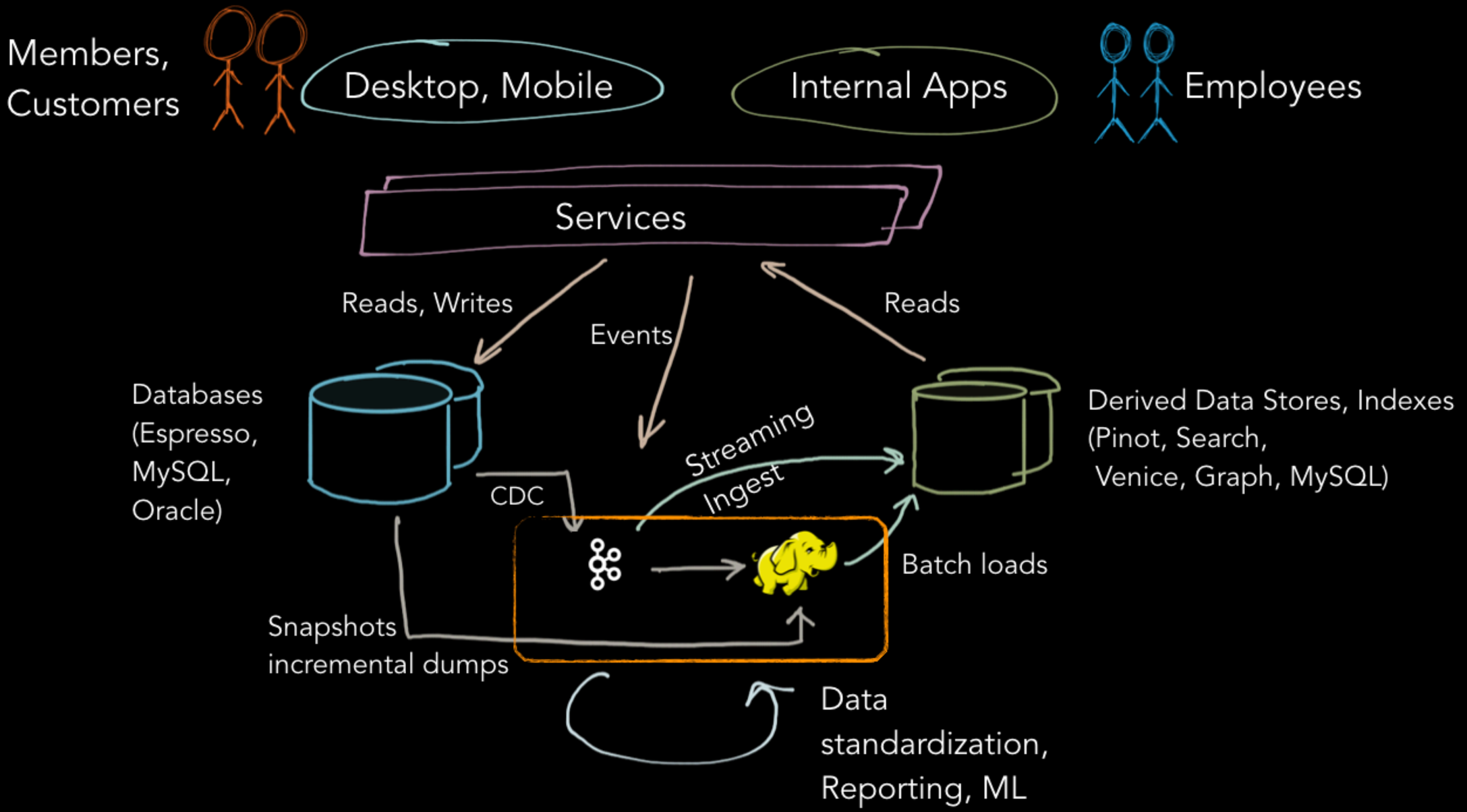
A framework for building a mesh of metadata services

Powered by Metadata

???

DataHub : GMA

LinkedIn's Data Ecosystem

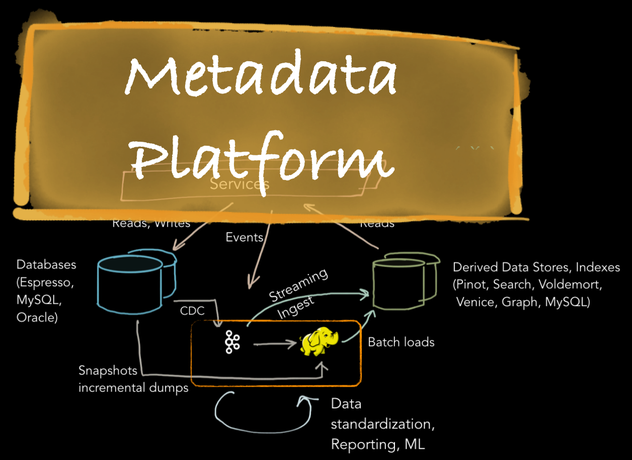


Search and Discovery

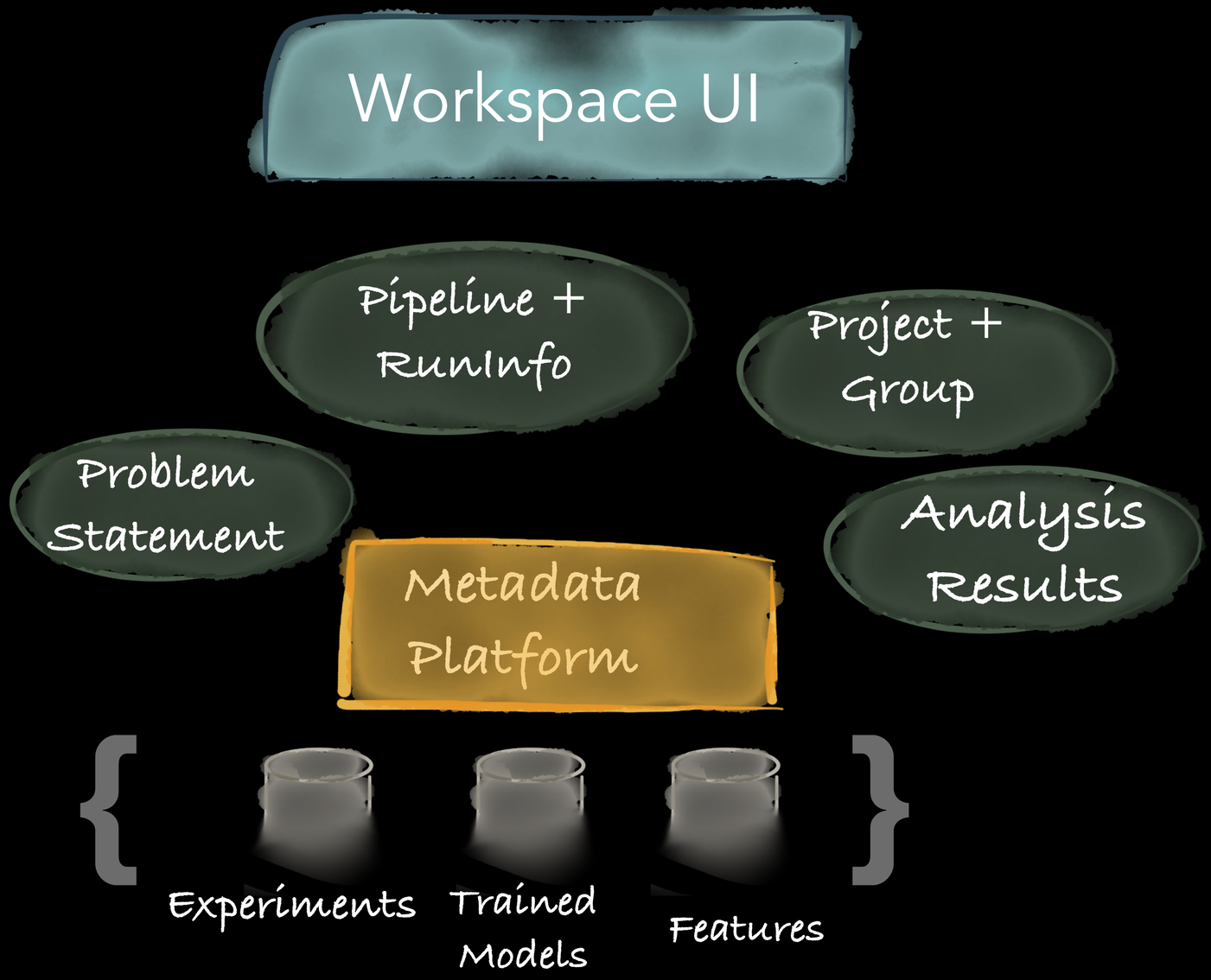
Free form, faceted search / browse for all entities
(datasets, features, metrics, pipelines, people)

Explore relationships between entities (lineage, ownership etc.)

DataHub : App



AI Metadata

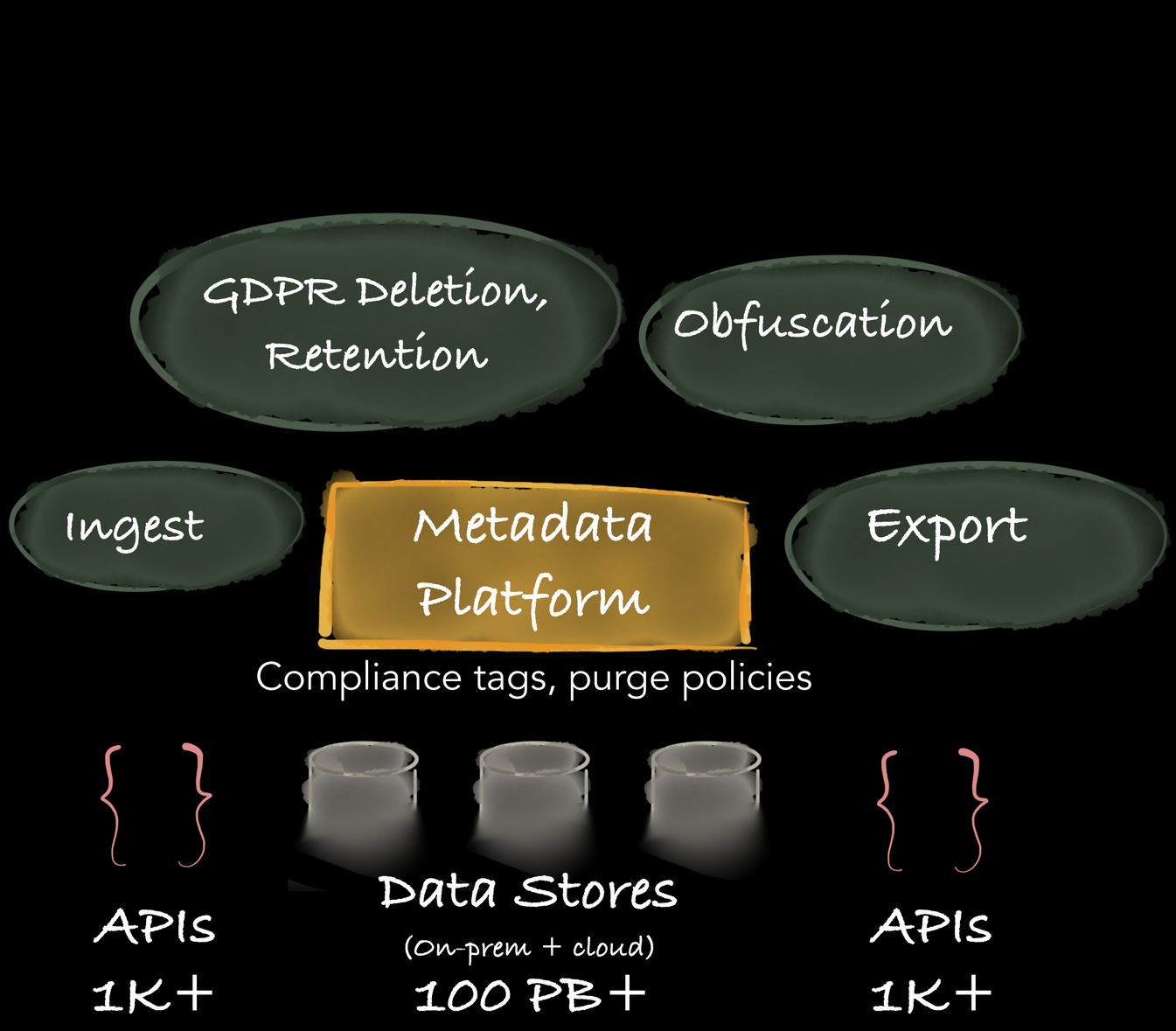


Goals

Reproducibility
Audit-ability
Visibility
Consistency

Integrated with dev workflow
  Metadata should live with code

Compliant Data Management



Apache Gobblin

gobblin.apache.org

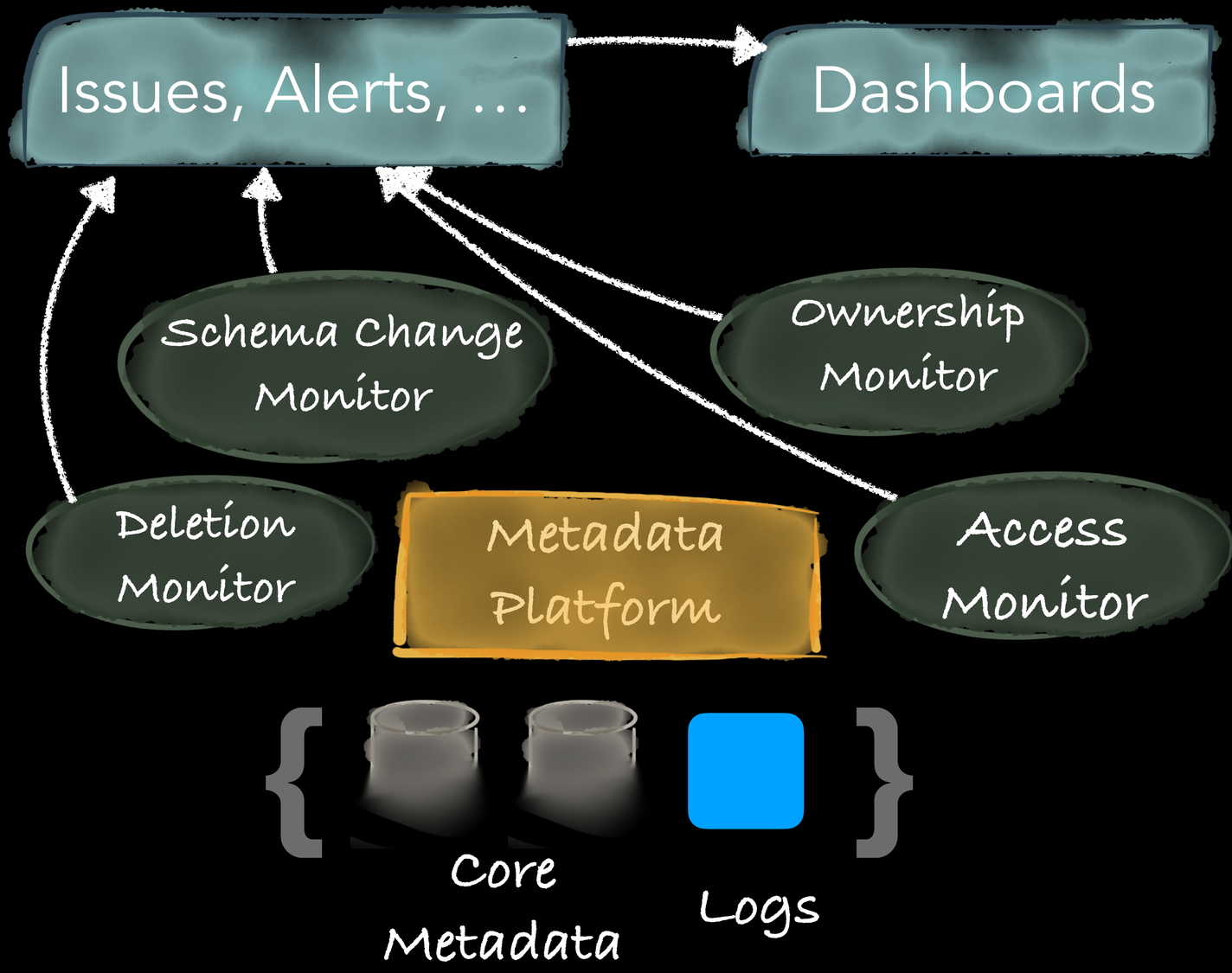
Ingest compliant data from external sources (API-s, data)

Manage data assets lifecycle inside enterprise
- Limited retention
- Fine-grained data deletion

Automatically Obfuscate data based on compliance tags
to create pii-free zone

Export, manage compliant data in external destinations (API-s, data)

Governance Workflows



A Few Scenarios

Ownership of an asset must stay above a threshold

Schema changes must be in sync with metadata changes

GDPR Deletion must happen in time across all systems

Access must be granted in accordance with policies

Powered by Metadata

Search and Discovery beyond just Datasets	AI : Model, Feature reproducibility, explainability	Compliant Data Management	Governance Workflows	Data Quality	Operational Monitoring	… and we're just getting started

DataHub : GMA

Thank You!

Shirshanka Das, Principal Staff Software Engineer, LinkedIn

Projects: LinkedIn DataHub, Apache Gobblin, Dali

@shirshanka

linkedin.com/in/shirshankadas