Data Discovery and Lineage: Integrating streaming data in the public cloud with on-prem, classic datastores and heterogeneous schema types

Barbara Eckman, Ph.D.

Principal Architect

Comcast

Comcast collects, stores, and uses all data in accordance with our privacy disclosures to users and applicable laws.

Our Group's Mission

Gather and organize metadata and lineage from diverse sources to make data universally discoverable, integrate-able and accessible to empower insight-driven decision making

- Dozens of tenants and stakeholders
- Millions of messages/second captured
- Tens of PB of long term data storage
- Thousands of cores of distributed compute

COMCAST

Quickie Quiz

- Does your job involve integrating data across corporate silos/verticals?
- Do you spend more time finding and reformatting data than you do analyzing it?
- When you attempt to integrate your data with another team's data, are you uncertain about what the other team's data means?
- Are you worried that in joining the two datasets, you may be creating "Frankendata"?
- Does your Big Data ecosystem go beyond a single hadoop provider, or even include public cloud and on-prem?

We Answer These Questions!

- Where can I find data about X?
- How is this data structured?
- Who produced it?
- What does it mean?
- How ”mature” is it?
- What attributes in your data match attributes in mine? (e.g., potential join fields)

- How has the data changed in its journey from ingest to where I’m viewing it?
- Where are the derivatives of my original data to be found?

COMCAST

Outline

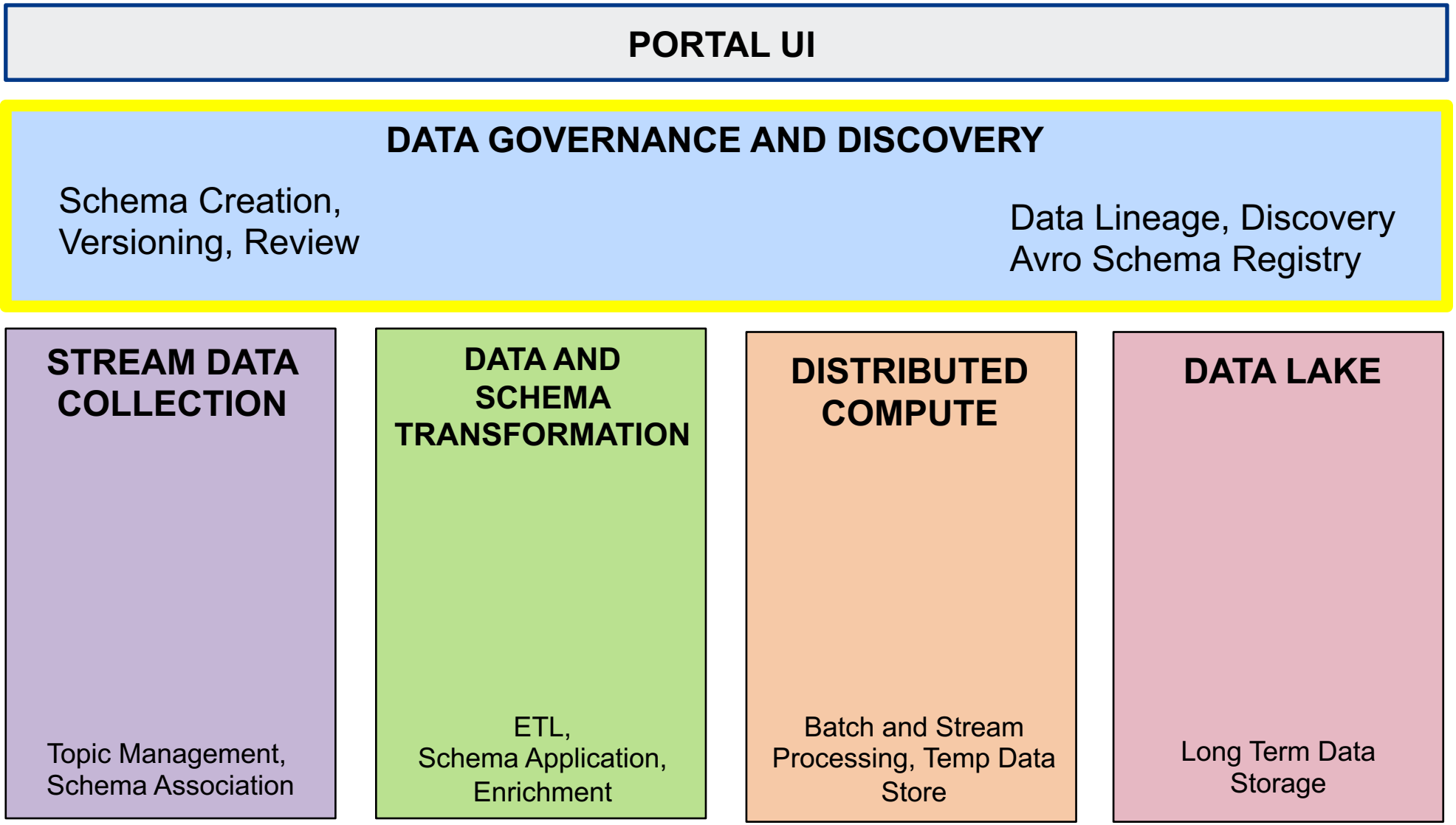
- #TBT** to Strata Data NYC Sept 2017
- Reorganization Yields New Requirements (Dec 2017)
- The Challenge of Legacy Big Data
- New Integrative Data Discovery and Lineage Architecture
- Next steps

** “Throw Back Thursday”

COMCAST

#TBT to Strata Data NYC Sept 2017

Data Platform Architecture, Sept 2017



COMCAST

Building a New Platform for Big Data

- Our Motto (and luxury): Nip chaos in the bud!
- Require well-documented schemas on data ingest
- Build lineage and metadata capture into the data flow
- Separate “team” data lakes from “community” data lake
- Build any additional metadata types as needed
- Heterogeneity is the biggest challenge…

COMCAST

Challenges of Heterogeneity for Building a Metadata Platform

- There are many excellent data discovery tools
  - OS and commercial
- BUT limited in scope of data set types supported
  - Only a certain Big Data ecosystem provider
  - Only RDBMS's, text documents, emails
- We need to add new data set types from multiple providers nimbly!
- We need to integrate metadata from diverse data sets, both traditional Hadoop and AWS
- We need to integrate lineage from diverse loading jobs, both batch and streaming

COMCAST

Strata Data NYC 2017: Key Metadata Technologies

Avro.apache.org

Atlas.apache.org

Apache Avro

Apache Atlas

COMCAST

What are Avro and Atlas?

- A data serialization system
  - **A JSON-based schema language**
  - A compact serialized format
- APIs in a bunch of languages
- Benefits:
  - Cross-language support for dynamic data access
  - **Simple but expressive schema definition and evolution**
  - Built-in documentation, defaults

Apache Atlas

- **Data Discovery, Lineage**
  - Browser UI
  - Rest/Java and kafka APIs
  - Synchronous and **Asynchonous messaging**
  - **Free-text, typed, & graph search**
- Integrated Security (Apache Ranger)
- Schema Registry as well as Metadata Repo

Open Source
Extensible

COMCAST

Strata Data 2017: Atlas Metadata Types

Built-in Atlas Types

- DataSet
- Process
- Hive tables
- Kafka topics

Custom Atlas Entities

- Avro Schemas
  - Reciprocally linked to all other dataset types
- Extensions to Kafka topic
  - sizing parameters
- AWS S3 Object Store

Custom Atlas Processes

- Lineage Processes
  - Avro schema evolution with compatibility
  - Storing data to S3 objects
- Enrichment Processes on streaming data
  - Re-publishing to kafka topics

COMCAST

Reorganization Yields New Requirements (Dec 2017)

New Requirements

- Integrate on-prem data sources' metadata and lineage
  - Traditional warehousing (Teradata/Informatica)
  - RDBMS's
  - Legacy Hadoop Datalake (hive, hdfs)
- End-user annotations
  - Stakeholders, documentation

COMCAST

RDBMS's

Created RDBMS Atlas typedefs

- Instance
- Database (schema)
- Table
- Column
- Index
- Foreign Key

Used for:

- Informatica Metadata Manager, on top of Teradata EDW
- Oracle
- Others to come

Comments:

- Back pointers to parent class at every level of hierarchy
- Load only whitelisted databases to increase signal, reduce noise

COMCAST

End-user annotations: new tag typedefs

Stakeholders
- Individuals
  - Data Business Owner
  - Data Technical Owner
  - Data Steward
  - Delivery Manager
  - Data Architect
- Teams
  - Delivery Team
  - Support Team
  - Data Producer
  - Data Consumer

Documentation
- Name
- Description
- URL

Acknowledgements:
Portal team

COMCAST

Legacy Hadoop Data Lake

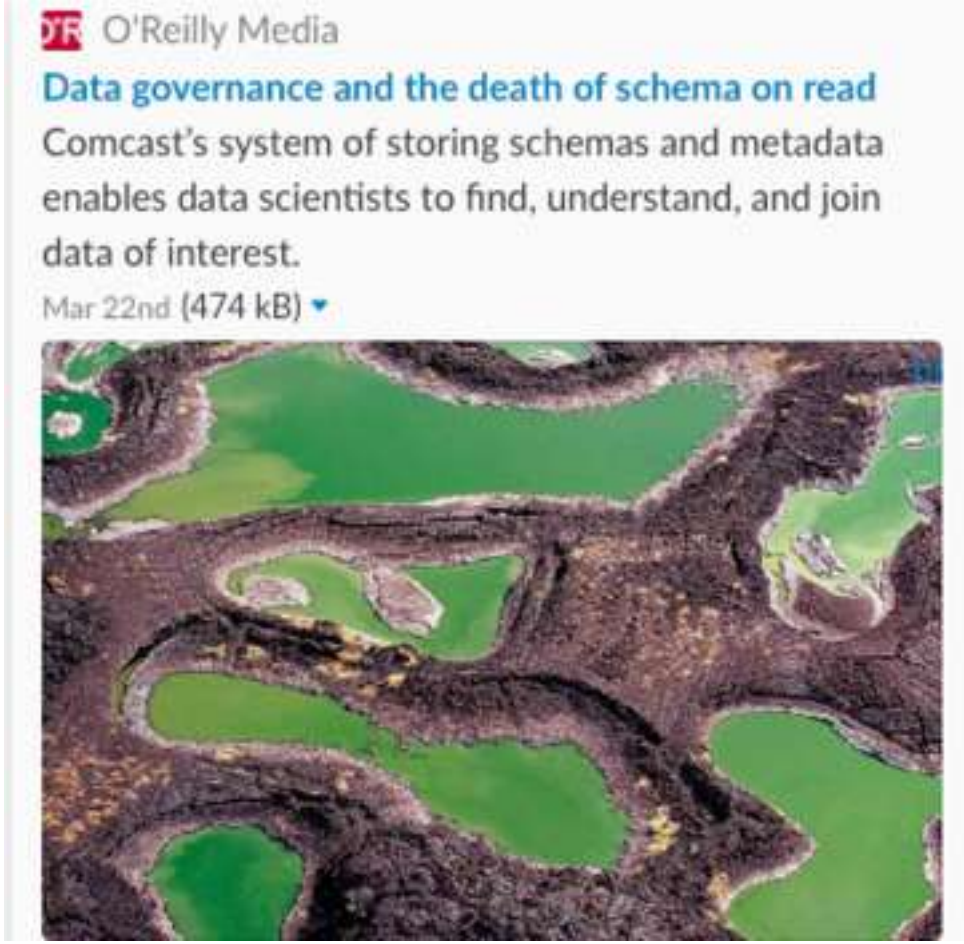
- Apache Atlas comes with built-in hooks for hdfs path, hive table metadata and lineage
  - Event-driven
- Installed Atlas in on-prem data center
- Atlas-to-Atlas Connector consumes from on-prem kafka topic, publishes into central repository
  - Load only whitelisted hive dbs to increase signal, reduce noise
  - Handles multiple Atlas versions

The Challenge of Legacy Big Data: Deconstructing the Big Data Revolution

COMCAST

Suggested Reading

https://www.oreilly.com/ideas/data-governance-and-the-death-of-schema-on-read



COMCAST

Pre-revolutionary Data Management

- Enterprise Data Warehouse is the exemplar
- Single schema to which all incoming data must be transformed when it is written (schema-on-write)
- Often tightly controlled by DBA's/IT department, who owned the schema and often the ETL jobs ingesting data
- Usually modeled in flat relations (RDBMS's)
  - Naturally nested data was "normalized", then "denormalized" to support specific queries (eg sales by region and by month)
- Rigorous data and schema governance

Big Data Bastille Day

Overthrow the self-serving nobility of EDWs and their tightly controlled data representation (and data governance)!

"Data Democratization"

- Anyone can write data to the datalake in any structure (or no consistent structure)
- Data from multiple previously siloed teams could be stored in the same repository.
- Nested structures no longer artificially flattened
- Schemas discovered at time of reading data (schema-on-read)

COMCAST

Post-revolutionary Status

- Data representation and self-service access have blossomed
- Data discovery and semantics-driven data integration have suffered
  - Unable to find data of interest
  - Hard to integrate due to lack of documented semantics
  - Data duplicated many times
- Data gives up none of its secrets until actually read
  - Even when read data has no documentation beyond attribute names, which may be inscrutable, vacuous, or even misleading.
- We need a Post-revolutionary Schema and Data Governance!

COMCAST

A New Integrative Data Discovery and Lineage Architecture

Conquering Legacy Big Data Platform

Many new challenges!

- The lugubriousness of EDW process without the control of schema-on-write
  - Retained journaling, Type 2 of EDW in building legacy data lake
- Identify and reduce redundancy among, say, hive tables
- Identify semantic relationships among existing (de-duped) hive tables
  - Not just attribute names, but data-based ML as well
- Identify what is for community consumption, and what is for individual team use, and maintain distinction
- Begin documentation of existing community tables
- Begin governance of schemas going forward

Dataset Lineage capture

- Generic lineage process typedef
  - Used for both batch and streaming lineage capture
  - Attributes include transforms performed, general-purpose config parameters
  - May be subclassed to add attributes for individual cases

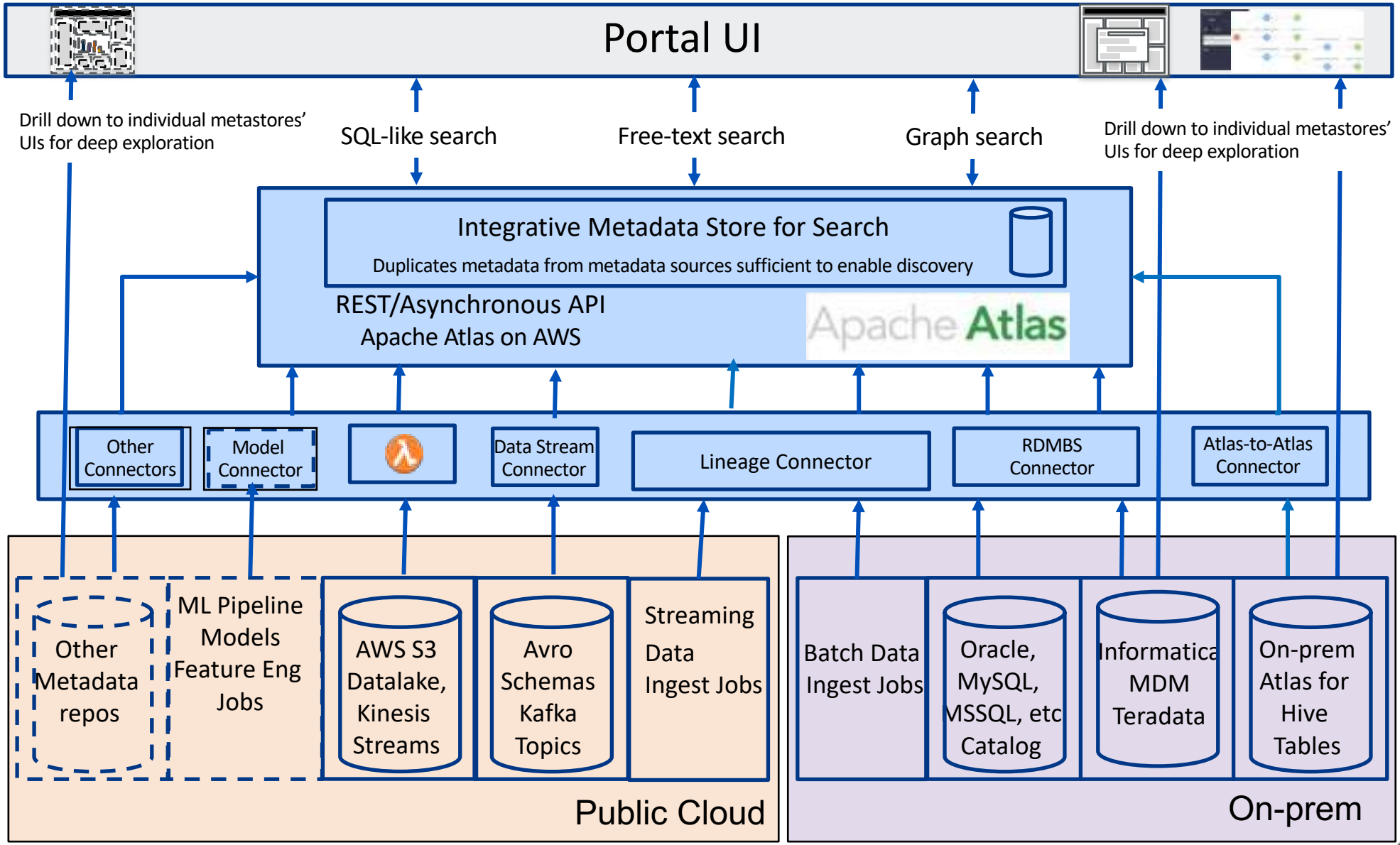
- Lineage capture is event-driven whenever possible
  - In AWS, Cloudwatch event on Glue crawler triggers lambda function
  - In on-prem hadoop, Inotify event on hdfs triggers microservice

- Triggered components assemble requisite info and publish to Atlas lineage connector

Acknowledgements:
Datalake Team

COMCAST

New Integrative Data Discovery and Lineage Architecture



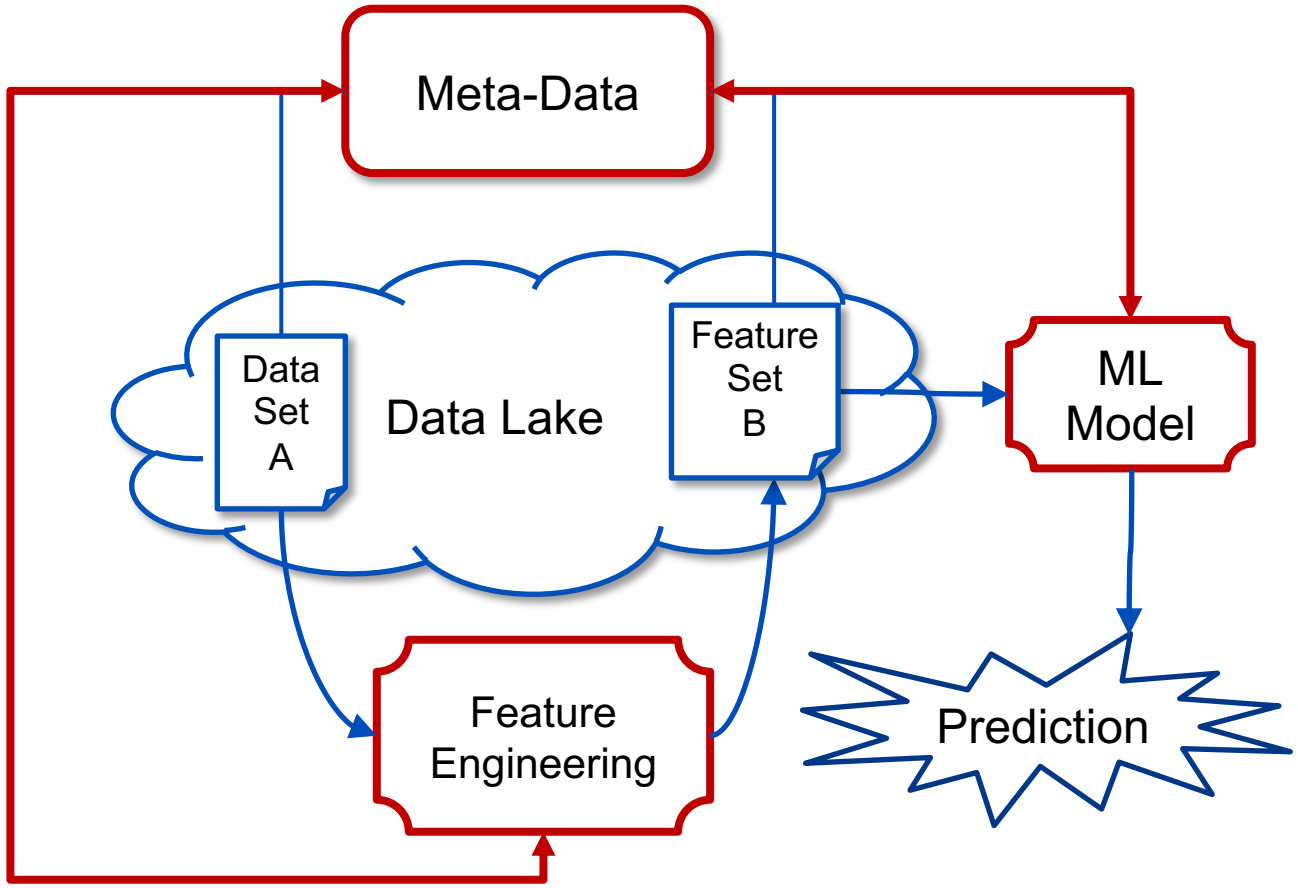
COMCAST

Connectors for all metadata sources

- One java codebase for all sources
- Differ in means of acquiring metadata/lineage, but use the same methods to package data for publishing to Apache Atlas via kafka api
  - RDBMS's (including EDW)
  - Atlas to Atlas (supports different versions)
  - Kafka topics
  - Avro schemas
  - AWS datalake objects
  - Kafka-to-datalake lineage

Next Steps

Metadata repo for discovery and documentation of models



- End-to-end metadata repository
  - Models are first-class objects, captured with rich metadata (eg input file schema, feature set schema, model parameters, etc)
  - Feature engineering jobs are first-class objects, captured with rich metadata (eg model, data quality threshold, input file schema, owner)
  - Build metadata capture on models and feature engineering jobs into the ML pipeline

COMCAST

Extreme scaling for Metadata and Lineage Capture

Currently we build connectors to pull from other sources of metadata and lineage, then push to our metadata repo

Coming: API for community push of metadata, lineage

- Making it easy for anyone to contribute to our repository

COMCAST

Extending avro schema governance to other schema types

- Interactive user app facilitates creation of schemas and enforces compliance with Comcast conventions
  - Each schema is reviewed and approved by at least one human being
- Comcast conventions:
  - Non-vacuous doc comments required to document every attribute
  - All attributes must have default values
  - Unnecessary complexity is discouraged (YAGNI principle)
- Library of commonly used subschemas
  - Available via app, use is encouraged by reviewers

COMCAST

Data Discovery and Lineage:
Integrating streaming data in the public cloud with on-prem, classic datastores and heterogeneous schema types

- #TBT to Strata Data NYC Sept 2017
- Reorganization Yields New Requirements (Dec 2017)
- The Challenge of Legacy Big Data
- New Integrative Data Discovery and Lineage Architecture
- Next steps
- Parting “Gifts”

COMCAST

Comcast Contributions to Apache Atlas OS Community

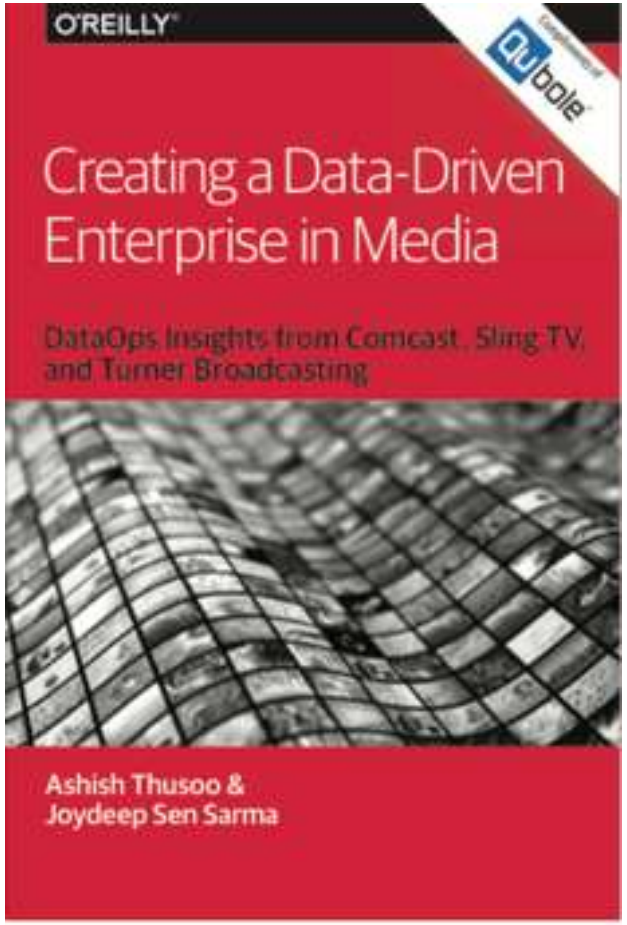
https://issues.apache.org/jira/browse/ATLAS-XXXX

Jira Ticket	Description
ATLAS-2694	Avro schema typedef and support for Avro schema evolution in Atlas
ATLAS-2696	Typedef extensions for Kafka in Atlas
ATLAS-2708	AWS S3 data lake typedefs for Atlas
ATLAS-2709	RDBMS typedefs for Atlas
ATLAS-2724	UI enhancement for Avro schemas and other JSON-valued attributes

More Suggested Reading

Creating A Data-Driven Enterprise in Media

Comcast Chapter:
How a Focus on Customer Experience Led to a Focus on Data Science



Can be reached from: https://www.oreilly.com/ideas/data-governance-and-the-death-of-schema-on-read

COMCAST

My collaborators









Hortonworks

Vadim Vaks
Principal Solutions Architect

COMCAST

Attributions

- Eiffel tower with fireworks photo
  - Yann Caradec, under https://creativecommons.org/licenses/by-sa/2.0/legalcode

COMCAST

Barbara_Eckman@Cable.Comcast.com