2019 W3C Workshop on Data Models for Transportation

September 12th, 2019

Sli.do: #w3ctransport

Welcome/intro: Joshua Shinavier, PhD

Uber

Current problems

- Thousands of datasets and schemas
  - Static, streaming, RPC
- Data sources are not composable
- Strong identifiers, weak semantics
  - Duplicate types, homonyms, synonyms
- Per-language data islands
- Diversity of data modeling conventions

Future problems

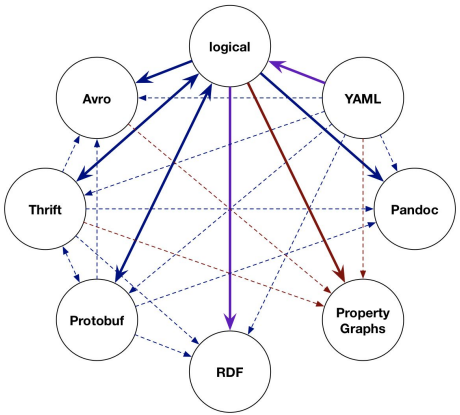
- Ingesting data from external sources
  - Schema alignment tends to be a bottleneck
- Exposing data through public APIs
  - Same problem in reverse
- Extending current schemas
  - Schemas are not composable *unless* designed that way

Schema standardization @Uber

- Controlled vocabularies for the entire company
- Basic type aliases
- Structured types
  - geospatial, time, sensors, money, addresses and user contact info, etc.
- Metadata vocabularies
  - Datasets, services, lineage, privacy, SLAs, deletion and retention, etc.
- Entities and relationships
  - User, Vehicle, Trip, etc.
  - Hundreds of commonly-redefined entities
- Elevate domain-specific RPC and storage schemas to *ontologies*

Data model standardization @Uber

- Data exists in diverse formats
  - Relational data (storage)
    - SQL, CQL, HQL, DQL, …
  - Data interchange formats (RPC, streaming, and storage)
    - Protocol Buffers, Apache Thrift, Apache Avro
- Need additional formats for analysis and visualization
  - Resource Description Framework (RDF)
  - Property Graphs
    - Algebraic Property Graphs
- Common core data model
  - Bridge between data formats
- N-to-n mappings
  - Thrift to Protobuf, Protobuf to Avro, etc.



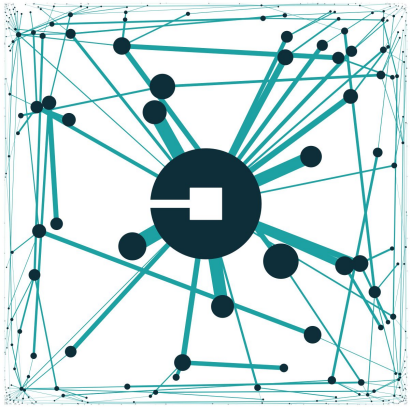
Driving adoption

- xxx number of new schemas / schema changes every day
    - Far more than we can manually review
- Pockets of schema design with idiosyncratic vocab, conventions
- Tips and tricks
    - Style guides
    - Schema review
    - Incentivizing schema developers
    - W3C workshop!

Metadata @Uber

- xxx hundred thousand structured datasets at Uber
- Data protections and user trust
  - GDPR and other regulations, Uber's own data policies
  - What kind of user data? Where is it?
  - Heroic numbers of manual annotations
    - Limited expressivity, limited guarantees
    - *Inference* is required
- Data discovery
  - Where is the source of truth for which data?
  - How to compose microservices?
  - Data lineage
- Combination of manual and automated annotation
  - Datasets, services

Thanks



joshsh@uber.com