Metadata Management in Data Lakes

Rihan Hai

M2U, September 2, 2018
Agenda

- Why use a data lake?
- What is a data lake?
- What are the existing solutions for data lake?
- What are the challenges to facilitate metadata management in a data lake?
Why Use a Data Lake?

- High variety of data
  - Relational data
  - JSON, XML
  - Graphs
  - Stream data
  - Documents
  - ...

- No explicit metadata given in data sources

- Massive amount of data

- Easy exploration of heterogeneous data

- Avoid overhead of data transformation during data ingestion (raw data storage)

- Usage of data unclear in the design phase (store data first, query later)
Why Use a Data Lake?

- No explicit metadata given in data sources
- Usage of data unclear in the design phase
- Store data first, query later

Use case: Building machine data management in a German car manufacturer

What Is a Data Lake?

- Big Data repositories
- Heterogeneous data ingestion: Load-As-Is
- Heterogeneous data store
- Schema-on-read
- Described by metadata

Data maturity increases over time.
Introduction: Basic Concept of Data Lakes

Metadata management is an issue that goes across all layers

- Ingestion
- Storage
- Transformation
- Interaction

Existing Solutions for Data Lakes

- Hadoop and its ecosystem
- Polystore-based data lakes
- Other DL research prototypes/commercial tools
  - Goods: Google’s Datasets
  - Microsoft Azure
  - KAYAK
Hadoop-based Data Lake

- **Hadoop alone ≠ Data Lake**

A basic Hadoop architecture for scalable data lake infrastructure

- Hadoop Distributed File System (HDFS)
  - Input file
  - Output file
  - Map task
    - Input
    - map() partition() combine()
    - Region 1
    - Region 2
    - Region 3
  - Reduce task
    - Input
    - sort() reduce()
    - Output

- With YARN, Hadoop now supports various programming models and both near-real-time and batch outputs.

- The system splits up the jobs, distributes processes, and recombines them via a cluster that can scale to thousands of server nodes.

- Job tracker


Solutions:

- **Hadoop ecosystem**
  - Hadoop ecosystem

- **New research approaches for exploring Hadoop**

**Polystore-based Data Lakes**

**Constance**

- Ingest heterogeneous data
- Data stored in RDBMS & NoSQL
- Metadata managements (SIGMOD’16 Demo)
- Nested schema mapping (ER’18)
- Query rewriting (ADBIS’18)
  - September 4, 2018, Tuesday
  - 14:20-14:50
Microsoft Azure Data Lake

- **Azure Data Lake Store/ Analytics**

- **Multi-tier data storage**

- **Analyze unstructured data with U-SQL**
  - Schema-on-read
  - Custom logic
  - User-defined function


  [Link to Azure Data Lake](https://azure.microsoft.com/en-us/solutions/data-lake/)
Goods: Google’s Datasets

- Metadata gathering of large number of data sets
  - Prioritize and optimize processing of the datasets
  - Identify similar or identical datasets

- Metadata inference using GOODS Catalog
  - Basic metadata
  - Provenance
  - Schema (serialized protocol buffers)
  - Content summary
  - User-provided annotations
  - Semantics

---

Goods: Google’s Datasets

- Metadata gathering of large number of data sets
  - Prioritize and optimize processing of the datasets
  - Identify similar or identical datasets

- Metadata inference using GOODS Catalog
  - Basic metadata
  - Provenance
  - Schema (serialized protocol buffers)
  - Content summary
  - User-provided annotations
  - Semantics

---

KAYAK: Data Preparation in a Data Lake

- **Data preparation**
  - Facilitate data preparation activities in pipelines composed by *primitives*
  - Trade-off between performance and accuracy of primitives’ results

- **Metadata extraction (metadata catalog)**
  - *Intra-dataset* metadata
  - *Inter-dataset* metadata
    - Integrity constraints
    - Joinability
    - Affinity


Challenges: Metadata Management in a Data Lake

- **Schema Extraction**
  - Extracts structural/descriptive metadata from heterogeneous sources
  - Capture implicit metadata properties
  - Metadata modeling
  - Enables the annotation of the metadata with semantic information

- **Schema Matching**

- **Schema Integration**

- **Schema Mapping**
  - Mapping languages with different expressive powers and complexities
  - Translate mappings to executable codes

- **Schema Evolution**

- **Query Answering**

- **Data Quality**
Q&A

Data Lake

- Provide data
- Explore data
- Provide insight

Data maturity increases

Time