Module 2: Data Governance and Privacy

DSC 208R - Data Management for Analytics

1 Data Collection and Governance: Overview

Overview

In this module, you will learn about the sourcing stage of the data science lifecycle and the various "hats" data scientists need to wear when acquiring, reorganizing, preparing, and potentially labeling data. It also covers the principles of data governance and data privacy laws.

1.1 Real-World Data Science Tasks

- Building and training sets
- Cleaning and organizing data
- Collecting data sets
- Mining data for patterns
- Refining algorithms

Data workers spend 90% of their time on data preparation, which includes data collection, cleaning, and organization.

1.2 Data Science Lifecycle: Sourcing

What is Sourcing?

The sourcing stage of the data science lifecycle involves acquiring data from various sources, which can include databases, APIs, web scraping, and more. Data scientists must ensure that the data collected is relevant, accurate, and compliant with privacy regulations. It is the stage where you go from raw datasets to analytics-ready datasets/ML-ready datasets.

1.2.1 Sourcing Challenages

- Heterogeneity of data modalities, file formats, and Sources
- Data access constraints
- Bespoke/diverse kinds of prediction applications
- Unpredictable and continual edits to datasets
- Data quality issues
- Large scale of data
- Poor data governance in organization

1.2.2 Sourcing Lifecycle

1. Raw data
2. Acquiring
3. Reorganizing
4. Cleaning
5. Labeling and Amplification (sometimes)
6. Data Engineering for Analytics
7. Results

2 Data Organization and File Formats

Overview

This study guide covers the fundamental concepts of data organization, file formats, and data models in data science. It explores the relationship between different data structures and their file representations, with a focus on structured, semi-structured, and unstructured data formats. Understanding these concepts is crucial for effective data acquisition, storage, and processing in the data science lifecycle.

2.1 Acquiring Data

Acquiring Data

Acquiring data involves obtaining data from various sources, which can include databases, APIs, web scraping, and more. Data scientists must ensure that the data collected is relevant, accurate, and compliant with privacy regulations.

2.2 Data Organization

Acquiring Data

Data organization refers to the way data is structured and stored in a system. It involves categorizing and arranging data in a manner that makes it easily accessible and usable for analysis. Proper data organization is essential for efficient data retrieval, processing, and analysis.

2.2.1 Data Modalities

- Structured Data: Data that is organized in a predefined format, such as tables or spreadsheets.
  - Examples include relational databases (e.g., SQL).

- Semi-Structured Data: Data that does not have a fixed schema but still contains some organizational properties.
  - Examples include JSON and XML files.
- Sequence Data: Data that is organized in a sequential manner, such as time series data or ordered lists.
  - Examples include time series data, such as stock prices or sensor readings.
- Graph-Structured Data: Data that is represented as a graph, where entities are nodes and relationships are edges.
  - Examples include social networks and recommendation systems.
- Text Data: Data that consists of written language.
  - Examples include documents, emails, and web pages.
- Multimedia Data: Data that includes images, audio, and video files.
  - Examples include PDFs, notebooks, images, audio files, and video files.

2.2.2 Key Terms

- File: A collection of data or information that is stored on a computer or other digital device.
- File Format: The specific way data is encoded and stored in a file.
- Metadata: Data that provides information about other data, such as its structure, format, and context.
- Directory: A cataloging structure with a list of references to files or other directories.
- Database: An organized collection of interrelated data
- Data Model: A conceptual representation of data structures and relationships.

- Logical Level: Data model for higher-level reasoning.
- Physical Level: How bytes are layered on top of files.

2.2.3 Structured Data: Common Forms

Structured Data

Structured data has a regular substructure

- Relational Data
  - Can be used for customer churn prediction
- Data Frame Data
  - Can be used for tabular data
- Matrix/Tensor Data
  - Can be used for statistical computing or scientific computing

2.2.4 Structured Data: Differences

- Ordering: matrix and dataframe have row/columns while relational is orderless.
- Schema Flexibility: matrix has cell numbers, relational conform to schema, and dataframe has no pre-defined schema.
- Transpose: Supported by matrix and dataframe but not relational

2.2.5 Semistructured Data

Semistructured Data

Semistructured data has less regular or more flexible substructure than structured data.

- Typically serialized with JSON or similar formats
- Some data systems offer binary file formats

- It is possible to layer Relations
- Graph-Structured would be an example

2.2.6 Data Lakes

Data Lakes

Loose coupling of data file format for storage and data/query processing stack(vs. RDBM's tight coupling).

2.2.7 Parquet vs Text-Based:Pros and Cons

- Less storage: Parquet stores in compressed form; can be much smaller (even 10x); lowers read latency
- Column pruning: Enables app to read only columns needed to DRAM; even lower query latency
- Schema on file: Rich metadata, stats inside format itself
- Complex types: Can store them in a column
- Human-readability: Cannot open with text apps directly
- Mutability: Parquet is immutable/read-only; no in-place edits
- Decompression/Deserialization overhead: Depends on application tool; can go either way
- Adoption in practice: CSV/JSON support more pervasive but Parquet is catching up