NYU

DS-GA 3001.009 Responsible Data Science Lab 8

Center for Data Science

Data Profiling

NYU

What is Data Profiling?

- Data profiling is the set of activities and processes to determine the metadata about a given dataset.

- Data profiling involves:
    - Collecting descriptive statistics like min, max, count and sum.
    - Collecting data types, length and recurring patterns.
    - Tagging data with keywords, descriptions or categories.
    - Performing data quality assessment, risk of performing joins on the data.
    - Discovering metadata and assessing its accuracy.
    - Identifying distributions, key candidates, foreign-key candidates, functional dependencies, embedded value dependencies, and performing inter-table analysis.

Understand your Data!

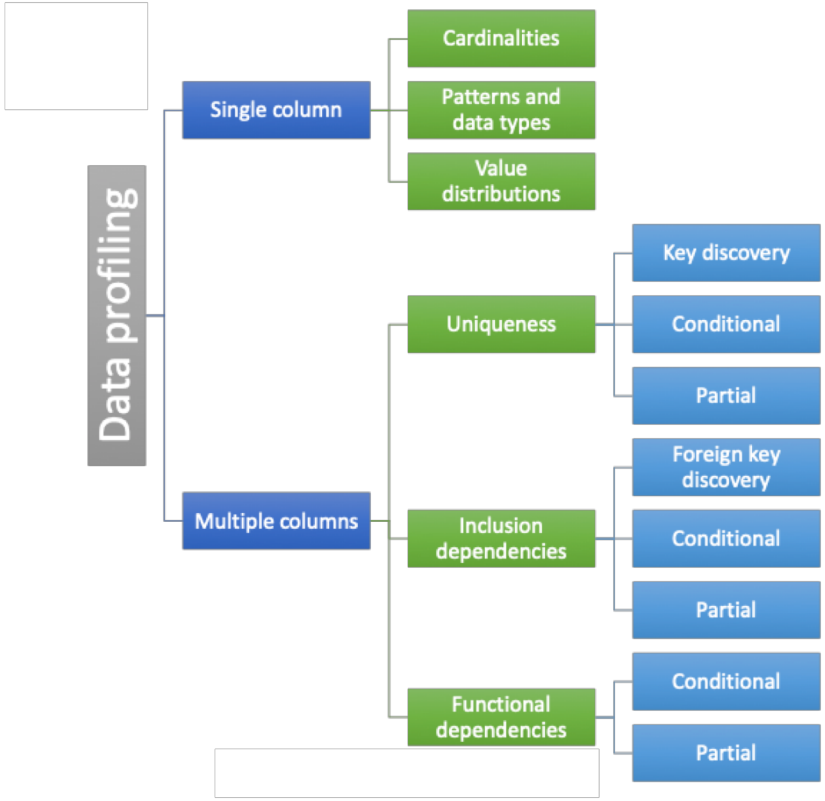
"Given the heterogeneity of the flood of data, **it is not enough merely to record it and throw it into a repository.** Consider, for example, data from a range of scientific experiments. If we just have a bunch of data sets in a repository, it is **unlikely anyone will ever be able to find, let alone reuse,** any of this data. With adequate **metadata**, there is some hope, but even so, challenges will remain due to differences in experimental details and in data record structure."

https://cra.org/ccc/wp-content/uploads/sites/2/2015/05/bigdatawhitepaper.pdf

Understand your Data!

- Need **metadata** to:
  - enable data **re-use** (have to be able to find it!)
  - determine **fitness for use** of a dataset in a task
  - help establish **trust** in the data analysis process and its outcomes
  - A set of activities and processes to determine the metadata about a given dataset
  - Metadata summarizes the data, summaries should be **small** but **informative**.

Classification of Profiling tasks



[Abedjan, Golab, Naumann; SIGMOD 2017]

Classification of Profiling tasks

- Single-column profiling
  - A basic form of data profiling is the analysis of individual columns in a given table.
  - Typically, the generated **metadata** comprise various **counts**, such as the number of values, the number of unique values, and the number of non-null values.
  - More advanced techniques **create histograms of value distributions** and **identify typical patterns in the data values** in the form of regular expressions.

Classification of Profiling tasks

- Multiple-column profiling
  - Multi-column profiling **generalizes profiling tasks on single columns to multiple columns** and also identifies inter-value dependencies and column similarities.
  - Tasks:
    - identify **correlations** between values through frequent patterns or association rules.
    - identify suitable **keys** for a given table. The discovery of unique column combinations, i.e., sets of columns whose values uniquely identify rows, is an important data profiling task.
    - discovery of **foreign keys** with the help of inclusion dependencies.
      - An inclusion dependency states that all values or value combinations from one set of columns also appear in the other set of columns—a prerequisite for a foreign key.
    - Identify **functional dependencies**
      - A functional dependency states that values in one set of columns functionally determine the value of another column.

Research tools for Data Profiling

Tool	Main Goal	Profiling Capabilities
Metanome [Papenbrock et al., 2015a]	Data Profiling	Columns statistics, rule discovery
ProLOD++ [Abedjan et al., 2014a]	LOD profiling and mining	General statistics, pattern discovery, unique discovery
Bellman [Dasu et al., 2002]	Data quality browser	Column statistics, column similarity, candidate key discovery
Potter's Wheel [Raman and Hellerstein, 2001]	Data quality, ETL	Column statistics (including value patterns)
Civilizer [Deng et al., 2017; Fernandez et al., 2016]	Data discovery	Column similarity
Data Auditor [Golab et al., 2010]	Rule discovery	CFD and CIND discovery
RuleMiner [Chu et al., 2014]	Rule discovery	Denial constraint discovery
MADLib [Hellerstein et al., 2012]	Machine learning	Simple column statistics

[Abedjan, Golab, Naumann; SIGMOD 2017]

Questions?

Thank you