Data Processing Pipeline

School of Data Skill Set

schoolofdata.org

Data Sources

- *web pages*
- *text documents*
- *structured documents*
- *databases*
- *scientific data*

Discovery and Acquisition

- Data Relevance (N, 1)
- Searching and Finding (N, 1)
- Completeness and Stopping (N, 1)
- Crowd Sourcing (N, 1)
- Manual Digitisation (1)
- Audit of Existing Resource (2)
- Crawling (2)
- Bulk Digitisation (3)

Extraction

- Loading to Data Store (1)
- Scraping (2)
- Parsing (2)
- Automation (2)
- Natural Language Processing (3)
- Data Pipes (3)

Cleansing, Transformation and Integration

- Data Formats and Standards (1)
- Data Granularity (1)
- Using Reference Data (1, 2)
- Merge/Join (1, 2)
- Mapping (2)
- Handling Manual Corrections (2)
- Normalisation (2)
- Entity Uniqueness (2, 3)
- Treating Duplicates (2)
- Indexing and Optimisation (2)
- Handling Changing Dimensions (2)
- Concept Modelling (2, 3)
- Fuzzy Matching (3)

Analytical Modeling

- Pivoting (1)
- OLAP (2)
- Business Rules (2)
- Regression (2)
- Outliers (2)
- Clustering (2)
- Graph/Network Metrics (2)
- Simulation (3)

Presentation, Analysis, Publishing

- Visualisation and Plotting (1, 3)
- Sorting and Filtering (1)
- Visualisation Method Selection (N, 1)
- Story Telling (N, 2)
- Publishing Online (1)
- Map Geo-Tagging (2)

Governance

- ETL Process Management (2, 3)
- Data Quality Management (2, 3)
- Auditability and Provenance (N, 2)
- Reference Data Management (2)
- Metadata (1, 2)

Tools and Technologies

- SQL (1)
- Programming Basics (2)
- Advanced Programming (3)

Level: N non-technical · 1 beginner · 2 advanced · 3 expert

Stefan Urbanek ▪ @Stiivi ▪ v0.3