

Data Governance & Management Hands-on Workshop

Péter Molnár
pmolnar@gsu.edu

November 6, 2024

Agenda/Activities

2:00 Introductions

2:10 Overview of Activities

Form Teams – Pick Activit[y|ies]

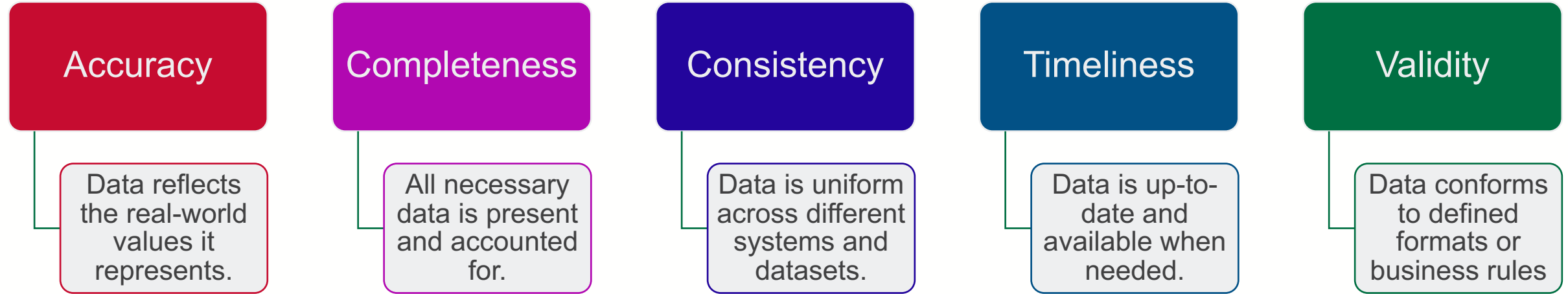
1. Define Data Quality Requirements
2. Analyze Real-world Cases on Failures in Data Governance
3. Explore and Transform Data with OpenRefine (Browser GUI)
4. Implement Data Profiling with GreatExpectations (Python, Jupyter)
5. Analyze Consumer Complaints with pre-trained LLM (Python, Jupyter)

4:00 Groups share their outcomes

Discussion/Questions/Adjourn

Define Data Quality Requirements

Ensuring Data Quality



Data quality refers to the **degree to which data is accurate, complete, consistent, and reliable** for its intended use.

High-quality data is crucial for informed decision-making, operational efficiency, and regulatory compliance

Maintaining **high data quality is an ongoing process** that requires clear standards, continuous monitoring, and the right tools.

Investing in data quality ensures better business outcomes, regulatory compliance, and trust in decision-making processes

Activity: Define Data Quality Requirements

Instructions:

1. Review a data source of your choice with respect to the 5 data quality dimensions: *accuracy, completeness, consistency, timeliness, and validity*
2. Establish thresholds and risks if the quality requirement for the dimension is not met
3. Discuss how to improve and maintain data quality

Notes:

- You may choose a data source in your own organization. Consider data sources that are represented by their owner and consumer in this group.
- You may also use a public data source. Here are some examples:
 - <http://datasource.com>

Analyze Real-world Cases on Failures in Data Governance

Activity: Analyze Real-world Cases on Failures

Instructions:

1. Discuss real-world cases where data governance and data management have failed, particularly in the financial services industry.
2. The goal is to identify what went wrong and derive key lessons that can be applied to improve data governance and management practices in your organization.

Discussion Questions:

- What went wrong? Identify the specific failures in data governance or data management.
- What were the consequences? Consider the financial, operational, regulatory, and reputational impacts.
- What could have been done differently? Propose actions or strategies that could have prevented the failure.
- What lessons can we learn? Summarize key takeaways that your organization can apply to avoid similar issues.

Resources

Cases:

1. Citigroup's repeated failures in risk management and data governance, leading to heavy fines and reputational damage
<https://www.ovaledge.com/blog/citigroup-data-governance-failure>
2. The Bank of America data breach due to third-party vulnerabilities, exposing sensitive customer information
<https://www.securitymagazine.com/articles/100403-security-experts-discuss-recent-bank-of-america-data-breach>
3. Data inaccuracies at financial institutions causing compliance issues and operational inefficiencies
<https://firstlogic.com/insights/data-inaccuracy-is-risky-in-financial-services>

Other Sources (found by Perplexity):

- <https://www.astera.com/type/blog/data-governance-in-financial-services/>
- <https://www.thegoldensource.com/top-data-management-challenges/>
- <https://kpmg.com/au/en/home/insights/2024/06/cyber-security-considerations-financial-services.html>
- <https://www.linkedin.com/pulse/top-10-data-governance-challenges-financial->
- <https://www.ironmountain.com/nl-nl/resources/blogs-and-articles/d/5-data-challenges-facing-financial-services-firms>
- https://www.dfs.ny.gov/Twitter_Report
- <https://www.complianceweek.com/regulatory-enforcement/occ-fed-fine-citi-136m-for-repeated-risk-management-data-governance-failures/35079.article>

MDM Technical Challenges

Data Integration

Integrating MDM with existing systems has complex technical challenges. Institutions often have multiple legacy systems (e.g., CRM, ERP) that store data in different formats. Ensuring seamless integration across these systems without data loss or corruption is difficult. Additionally, moving data between systems can introduce errors, especially when certain fields do not map correctly between applications

Data Standards

Establishing consistent data standards across all departments is critical but challenging. Financial institutions often deal with diverse datasets (e.g., customer, product, transaction data) that may follow **different naming conventions or formats**. Harmonizing these standards across the organization **requires careful planning and coordination** to ensure that all users adhere to them. Failure to standardize can lead to inconsistencies and reduce the effectiveness of MDM.

Data Quality and Cleansing

Maintaining high-quality data is essential for MDM success, but ensuring that all data is accurate, complete, and free from duplicates is a significant challenge. Organizations must **regularly perform data cleansing to eliminate errors like redundant records or outdated information**. Without proper data governance processes in place, poor data quality can undermine the entire MDM initiative

Data Security

Given the sensitive nature of financial data, ensuring robust data security is paramount. **Centralizing master data in an MDM system can create a single point of failure** if not properly secured. Institutions must implement strong encryption, access controls, and monitoring mechanisms to protect against cyberattacks and unauthorized access. Additionally, **compliance with regulations like GDPR adds another layer of complexity**

Data Migration

Migrating data from legacy systems into a new MDM platform is technically challenging due to differences in formats, structures, and quality. During migration, there's a risk of data loss or corruption if not handled carefully. Financial institutions must ensure that the migration process **does not disrupt business operations or compromise the integrity** of critical financial data.

MDM Organizational Challenges

Cross-Departmental Alignment

One of the biggest organizational hurdles is **achieving alignment across different departments and business units on how master data should be managed**. Different teams may have varying priorities or definitions for key data elements (e.g., customer records), making it difficult to agree on a unified approach. This lack of consensus can delay implementation and lead to fragmented data management practices.

Governance and Stewardship

Establishing effective data governance frameworks is crucial for long-term success but can be **difficult to implement consistently across large organizations**. Financial institutions need clear policies for managing master data, including roles for data stewardship—individuals responsible for maintaining data quality over time. Without strong governance, inconsistencies in how data is managed and updated can quickly arise.

Change Management

MDM implementation often requires significant changes to existing workflows and **roles within an organization**. Employees accustomed to siloed systems may resist adopting new processes that require collaboration across departments. Change management strategies are needed to ensure that staff are trained on new tools and understand their role in maintaining master data quality.

Resource Constraints

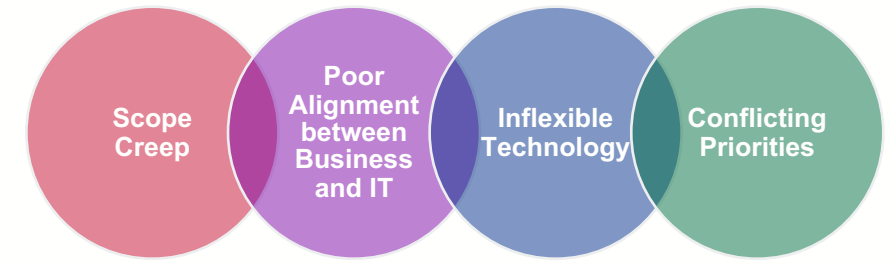
Implementing an MDM system **requires substantial investments in both technology and skilled personnel**. Financial institutions may face resource constraints, including limited budgets for advanced MDM tools or a shortage of qualified staff with expertise in data management. Additionally, ongoing maintenance requires continuous investment in infrastructure and personnel to ensure the system remains effective over time.

Model Agility

The MDM model must be flexible enough to adapt to changing business needs and regulatory requirements over time. However, many organizations struggle with creating an agile model that can accommodate future changes without requiring costly reconfigurations or updates.

Public Failures

<https://www.henricodolfing.com/p/project-failure-case-studies.html>



Vodafone's £59 Million CRM Disaster

In 2016, Vodafone faced a major failure with its customer relationship management (CRM) consolidation project, which was intended to streamline customer data across systems. The company invested £59 million into the initiative, but it resulted in significant disruptions to customer service and billing processes. The failure led to Vodafone receiving the largest fine ever imposed by UK regulators for breaches of consumer protection rules. This case underscores the risks of poor data integration and inadequate testing before going live with MDM or CRM


Lidl's €500 Million Debacle

Lidl, a major European grocery chain, invested around €500 million in a 3rd party MDM system aimed at transforming its inventory management processes. After seven years of development, Lidl abandoned the project because it could not adapt its business processes to fit the rigid data model. Lidl's failure illustrates how critical it is for MDM systems to align with an organization's business processes and operational needs. In this case, Lidl's insistence on using its own unique inventory valuation method clashed with the vendor's standard approach, leading to the project's collapse

LeasePlan's \$100 Million Failure

LeasePlan, an international automotive fleet management company, invested nearly \$100 million in a 3rd party MDM system that was ultimately scrapped before it went live. The company cited the "monolithic nature" of the MDM as incompatible with its need for agility and flexibility in managing fleet data across multiple countries. This case highlights how mismatches between an organization's operational needs and the chosen MDM technology can lead to costly failures.

<https://www.henricodolfing.com/p/project-failure-case-studies.html>



[About](#)[Services +](#)[Insights](#)[Resources +](#)[Contact](#)

Project Failure Case Studies

I research project failures and write case studies about them because it is a great way (for both of us) to learn from others' mistakes. This page is an ever-growing collection of such project failure case studies.

To be notified about new Project Failure Case Studies just sign up for my newsletter by clicking [here](#). You will receive a free copy of my *Project Success Model*™.

- > **Case Study 19: The \$20 Billion Boeing 737 Max Disaster That Shook Aviation**
The Boeing 737 Max, once heralded as a triumph in aviation technology and efficiency, has since become synonymous with one of the most catastrophic failures in modern corporate history.
- > **Case Study 18: How Excel Errors and Risk Oversights Cost JP Morgan \$6 Billion**
In the spring of 2012, JP Morgan Chase & Co. faced one of the most significant financial debacles in recent history, known as the "London Whale" incident. The debacle resulted in losses amounting to approximately \$6 billion, fundamentally shaking the confidence in the bank's risk management practices.
- > **Case Study 17: The Disastrous Launch of Healthcare.gov**
On the first day Healthcare.gov was launched, four million unique users visited the portal, but only six successfully registered. Over the next few days, the site experienced eight million visitors, but according to estimates, around 1% enrolled in a new healthcare plan. Even the users that did sign up experienced errors, including duplicates in enrollment applications submitted to insurers.
- > **Case Study 16: Nike's 100 Million Dollar Supply Chain "Speed bump"**
"This is what you get for 400 million, huh?", Nike President and CEO Phil Knight famously raised the question in a conference call days before announcing the company would miss its third-quarter earnings by at least 28% due to a glitch in the new supply chain management software.
- > **Case Study 15: How the Scottish Police Got £25 Million Back but Lost 3 Years on 16**
The program started with projected savings of £200m to the authority and Police Scotland. However, the program ended in July 2016 with wasted resources, wasted money, wasted time, and

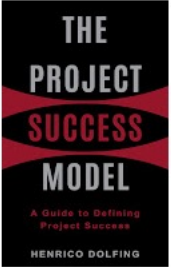
Newsletter

Articles and case studies delivered straight to your inbox.

Trouble Assessment

Is your project headed for trouble? Find out!

Free eBook



I Need Help With ...

Explore and Transform Data with *OpenRefine*

<http://openrefine.org>

OpenRefine is an open-source desktop application designed for cleaning, transforming, and organizing messy data.

It allows users to **import data from various formats** such as CSV, JSON, or Excel, and then explore, filter, and modify it through a **web-based interface**.

Key features include **faceting**, which helps users filter and group data based on specific criteria, and **clustering**, which merges similar values to resolve inconsistencies.

OpenRefine also supports **reconciliation with external databases** and provides an infinite undo/redo function to track and reverse changes.

OpenRefine

OpenRefine is a powerful free, open source tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.



Main features



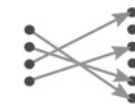
Faceting

Drill through large datasets using facets and apply operations on filtered views of your dataset.



Clustering

Fix inconsistencies by merging similar values thanks to powerful heuristics.



Reconciliation

Match your dataset to external databases via reconciliation services.



Infinite undo/redo

Rewind to any previous state of your dataset and replay your operation history on a new version of it.



Privacy

Your data is cleaned on your machine, not in some dubious data laundering cloud.

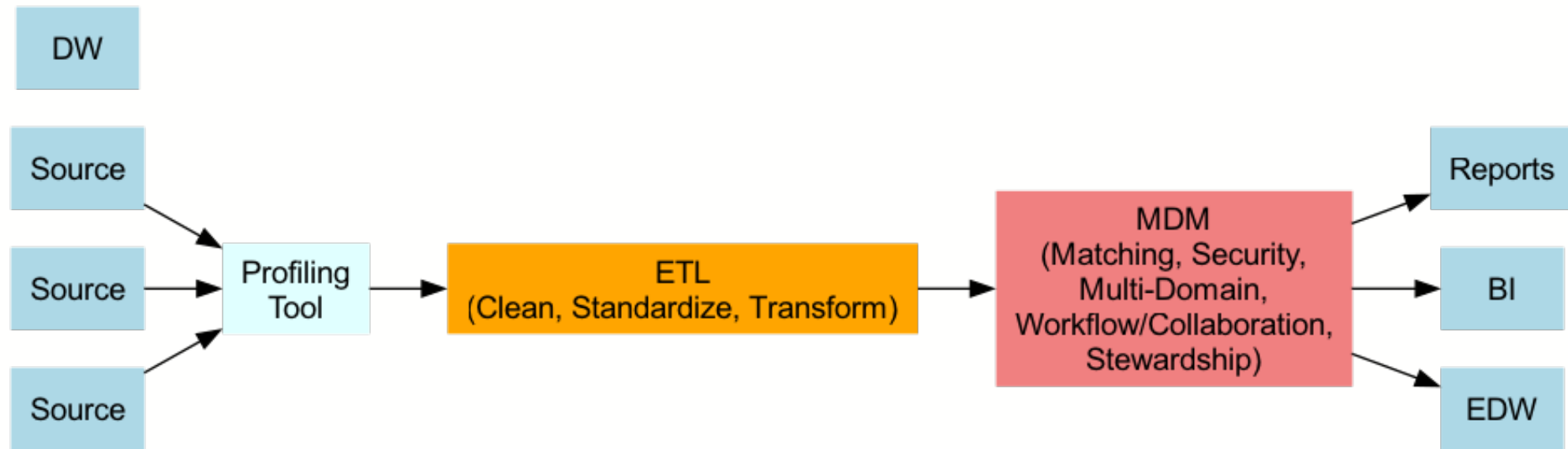


Wikibase

Contribute to Wikidata, the free knowledge base anyone can edit, and other Wikibase instances.

And much more to discover in [our documentation](#).

Master Data Management



Activity: Explore & Transform Data with *OpenRefine*

Instructions:

1. Navigate to the application at <http://10.230.100.212:3333>
2. Establish thresholds and risks if the quality requirement for the dimension is not met
3. Discuss how to improve and maintain data quality

There are five instances using ports 3333 to 3337

Notes:

- You may choose a data source in your own organization. Consider data sources that are represented by their owner and consumer in this group.
- You may also use a public data source. Here are some examples:
 - <http://datasource.com>
 - https://catalog.data.gov/dataset/?q=consumer+finance&sort=views_recent+desc&ext_location=&ext_bbox=&ext_prev_ext=ent=

Links

Tutorial: <https://datacarpentry.org/OpenRefine-ecology-lesson/>

User Manual: <https://openrefine.org/docs>

<https://catalog.data.gov/dataset/?tags=complaint>

Copy the URL for
a CSV file and ...

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are

Get data from

This Computer

Web Addresses (URLs)

Clipboard

Enter one or more web addresses (URLs) pointing to data to

Add another URL

Next »

... paste into
OpenRefine

29 datasets found

Consumer Complaint Database 380 recent views

Consumer Financial Protection Bureau — The Consumer Complaint Database is a collection of complaints about consumer financial products and services that we sent to companies for response. Complaints are...

HTML

Insurance complaints: All data 104 recent views

City of Austin — The Texas Department of Insurance (TDI) handles complaints against people and organizations licensed by TDI, such as companies, agents, and adjusters. To learn...

CSV RDF JSON XML

Patient Advocate Tracking System (PATS)

Department of Veterans Affairs — The Patient Advocate Tracking System (PATS) is a centralized, web based application that records and tracks instances of patient compliments and complaints concerning...

No FEAR Act

Department of Energy — The No FEAR Act requires each Federal agency to post on its public website summary statistical data relating to equal employment opportunity complaints filed against...

HTML

Complaint indexes and policy counts for insurance companies

City of Austin — The Texas Department of Insurance (TDI) regulates the state's insurance industry and oversees the administration of the Texas workers' compensation system. This data...

CSV RDF JSON XML

Antitrust Division Select Case Filings

Department of Justice — Index of select cases and documents filed by the U.S. Department of Justice, Antitrust Division. Cases are listed alphabetically by the last name of individual...

HTML HTML

Collection NHTSA's Office of Defects Investigation (ODI) -

Validation, Documentation, and Profiling with *Great Expectations*

<https://greatexpectations.io/>

Great Expectations is a Python package designed for **data validation, documentation, and profiling.**

It allows users to **define expectations**—verifiable assertions about data quality—and **automatically generate tests** to ensure datasets meet these expectations.

Additionally, it provides tools for **creating detailed reports and documentation**, helping teams maintain data quality and communicate effectively across technical and non-technical stakeholders

great expectations

PRODUCT ▾ COMMUNITY ▾ RESOURCES ▾ PRICING COMPANY ▾ Try GX Cloud Log in

Get started with GX Cloud by joining our bi-weekly hands on workshop.

Have confidence in your data, no matter what

Get everything you need to trust your data with GX Cloud: an end-to-end solution for your data quality process and a unique Expectation-based approach to testing, backed by the world's most popular data quality framework.

Learn more about GX Cloud

Connect to a Data Source

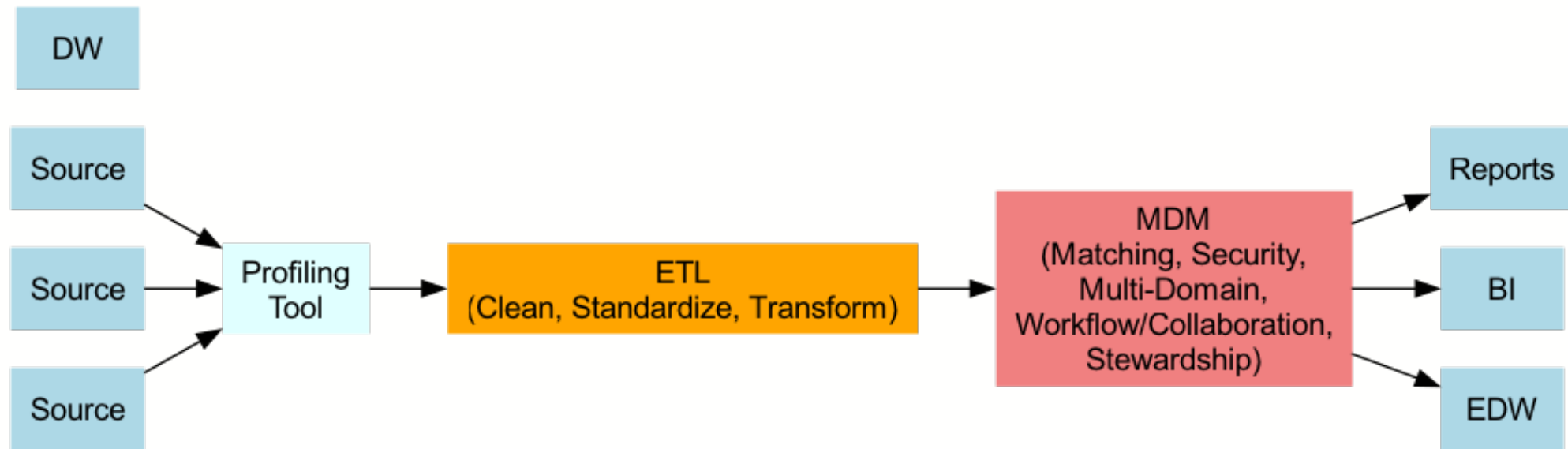
Add a Data Asset

Select tables to import

- ☒ Select all matching tables
- ☒ customers_a
- ☒ customers_b
- ☒ transactions
- ☒ transactions_b

Add 4 Assets

Master Data Management



Activity: Implement Data Quality Monitoring

Instructions:

1. This activity uses Python and JupyterLab on our Analytics Research Cluster. Navigate to <http://10.230.100.236:8000> and login with a guest login.
2. Clone the repository from <https://github.com/molnarai/data-management-and-governance-hands-on>
3. Follow the steps in the tutorial <https://docs.greatexpectations.io/docs/core/connect-to-data/dataframes>

Notes:

- This activity demonstrates how to implement data quality monitoring into the ETL pipeline.
- This example uses CSV files and Pandas. Though, the library also supports production level data sources.
- The tutorial uses a part of the NYC Taxi dataset. Feel free to explore other data.

molnarai / data-management-and-governance-hands-on

Code

Issues

Pull requests

Actions

Projects

Wiki

Security

Insights

Files

main

Go to file

DataHub

GreatExpectations

GE_Tutorial.ipynb

README.md

YdataProfiling.ipynb

OpenRefine

.gitignore

README.md

data-management-and-governance-hands-on / GreatExpectations

institute4insight update

0982b14 · now History

Name	Last commit message	Last commit date
..		
GE_Tutorial.ipynb	initial commit	10 hours ago
README.md	update	now
YdataProfiling.ipynb	initial commit	10 hours ago

README.md

Validation, Documentation, and Profiling with Great Expectations

Great Expectations is a Python package designed for data validation, documentation, and profiling. It allows users to define expectations—verifiable assertions about data quality—and automatically generate tests to ensure datasets meet these expectations.

Additionally, it provides tools for creating detailed reports and documentation, helping teams maintain data quality and communicate effectively across technical and non-technical stakeholders

Instructions

This activity uses Python and JupyterLab on our Analytics Research Cluster.

- Navigate to <http://10.230.100.236:8000> and login with a guest login.
- Clone the repository from <https://github.com/molnarai/data-management-and-governance-hands-on>
- Follow the steps in the tutorial https://docs.greatexpectations.io/docs/core/connect_to_data/dataframes

Notes

Notebook

Tutorial

Analyze Consumer Complaints with pre-trained LLM

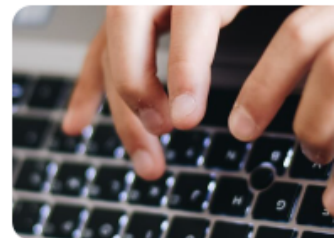
Activity: Analyze Complaints with pre-trained LLM

Instructions:

1. This activity uses Python and JupyterLab on our Analytics Research Cluster. Navigate to <http://10.230.100.236:8000> and login with a guest login.
2. Clone the repository from <https://github.com/molnarai/data-management-and-governance-hands-on>
3. Navigate to  GenAiDataExtraction

Notes:

- This activity demonstrates how to convert (unstructured) text data into categorical features.
- Data source: <https://www.kaggle.com/datasets/selener/consumer-complaint-database>
- The dataset is available ARC in [/data/public/ConsumerComplaintDatabase/](#)



Consumer Complaint Database

Consumer Finance Complaints (Bureau of Consumer Financial Protection)

Data Card Code (15) Discussion (0) Suggestions (0)

About Dataset

Context

These are real world complaints received about financial products and services. Each complaint has been labeled with a specific product; therefore, this is a supervised text classification problem. With the aim to classify future complaints based on its content, we used different machine learning algorithms can make more accurate predictions (i.e., classify the complaint in one of the product categories)

Content

The dataset contains different information of complaints that customers have made about a multiple products and services in the financial sector, such us Credit Reports, Student Loans, Money Transfer, etc.

The date of each complaint ranges from November 2011 to May 2019.

Acknowledgements

This work is considered a U.S. Government Work. The dataset is public dataset and it was downloaded from

<https://catalog.data.gov/dataset/consumer-complaint-database>

on 2019, May 13.

Usability ⓘ

8.24

License

U.S. Government Wo

Update frequency

Unspecified

Tags

Finance

Text

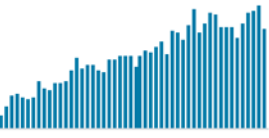
Classification

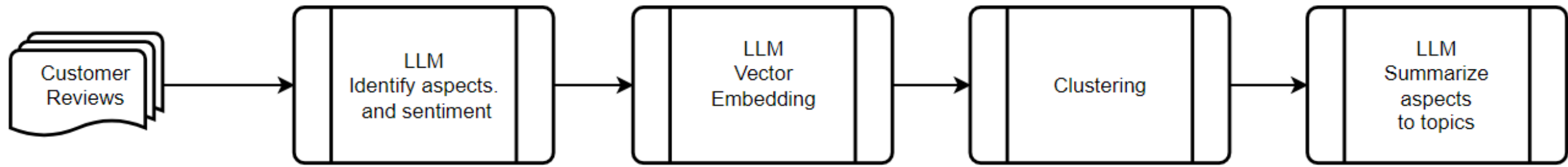
Turn on Kaggl

We heard your fee
Check out Kaggle's

Detail Compact Column

10 of 18 columns

Date received	Product	Sub-product	Issue	Sub-issue	Consumer complain...	Company public res...	Company
 <div>2011-11-302019-11-03</div>	Credit reporting, ... 21%	Credit reporting 21%	Incorrect informa... 13%	[null] 39%	[null] 68%	[null] 63%	EQUIFAX, INC. 10%
	Mortgage 20%	[null] 17%	Loan modification,... 8%	Information belon... 6%	There are many mi... 0%	Company has res... 26%	Experian Informati... 9%
	Other (836407) 59%	Other (896084) 63%	Other (1129584) 79%	Other (785650) 55%	Other (456864) 32%	Other (146032) 10%	Other (1164169) 82%
05/10/2019	Debt collection	Payday loan debt	Communication tactics	Frequent or repeated calls			CURO Intermediate Holdings
05/10/2019	Credit reporting, credit repair services, or other personal consumer reports	Credit reporting	Incorrect information on your report	Old information reappears or never goes away			Ad Astra Recovery Services Inc
05/10/2019	Checking or savings account	Checking account	Managing an account	Banking errors			ALLY FINANCIAL INC.
05/10/2019	Mortgage	Other type of mortgage	Closing on a mortgage				Statebridge Company
05/10/2019	Debt collection	Other debt	Attempts to collect debt not owed	Debt is not yours			Diversified Consultants, Inc.
05/10/2019	Student loan	Federal student loan servicing	Struggling to repay your loan	Problem lowering your monthly payments			Student Loan Direct
05/10/2019	Debt collection	Other debt	Attempts to collect debt not owed	Debt was paid			Diversified Consultants, Inc.
05/10/2019	Credit reporting, credit repair services, or other personal consumer reports	Credit reporting	Incorrect information on your report	Information belongs to someone else			CONTRACT CALLERS INC
05/10/2019	Vehicle loan or lease	Loan	Struggling to pay your loan	Denied request to lower payments			ALLY FINANCIAL INC.





Péter Molnár is an associate professor at Georgia State University's J. Mack Robinson College of Business where he teaches artificial intelligence, focusing on both technical implementation and executive decision-making. His career bridges academia and industry, including work as a data scientist at Amazon Web Services on innovative machine learning applications

pmolnar@gsu.edu

<https://www.linkedin.com/in/petermolnar/>