Data Governance & Management Hands-on Workshop

November 6, 2024

Documents and Code at:
https://github.com/molnarai/data-management-and-governance-hands-on/

Agenda

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

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

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.

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

Activity: Explore & Transform Data with *OpenRefine*

Instructions:

1. Navigate to the application at http://10.230.100.212:3333 (alternatively, use ports 3334 to 3337)
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
    - https://catalog.data.gov/dataset/?q=consumer+finance&sort=views_recent+desc&ext_location=&ext_bbox=&ext_prev_extent=

Links:

- Tutorial: https://datacarpentry.org/OpenRefine-ecology-lesson/
- User Manual: https://openrefine.org/docs

Activity: Validation, Documentation, and Profiling with *Great Expectations*

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.

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 folder 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/**