University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

Electrical & Computer Engineering and Computer Science / Data Science

Department: **Electrical & Computer Engineering and Computer Science**

COURSE:	**DSCI 6007 – Distributed & Scalable Data Engineering**
DSCI 6007 **Final PROJECT**	**Due date: December 8, 2024 (midnight)** **Presentations: December 9 (During Class Time)**
PROJECT TITLE:	1. **Final Project Title**
PROJECT **EXAMPLE:**	**Posted on Canvas: Final Project Module** 2. **Analyze real-time Twitter Sentiment** 3. **Bank Model-CRIPSP-DM (Links provided)**
SEMESTER:	
INFORMATION	Submit project: **12/08/2024 -(midnight)** Present project: **12/09/2024(class time)**
Group Names:	

Table of Contents

Final Project Guidelines & Requirements: Distributed & Scalable Data Engineering

Welcome to the final project phase! This comprehensive guide will take you through the process, expectations, and deliverables required for a successful submission.

Objective: The project should aim to solve a real-world problem or answer a complex question using data analysis. It should encourage students to apply data engineering principles to collect, process, store, and analyze data, and then present the findings in a meaningful way.
Example Project Theme: Analyzing social media trends to understand public sentiment on a current event.

Phase 1: Preparing for Your Project

Team Formation and Planning

1. **Team Leaders** should continue to lead and manage the teams work. Meet early to strategize on your approach to the project.
    - Consider the diverse expertise within your team, from Mathematicians to Computer Scientists and industry experts.
    - Discuss and assign roles based on each member's strengths and skills.

Topic Selection and Initial Research

Engage in brainstorming sessions and conduct preliminary research to decide on your project's focus.

- Refer to the 'Technical Report Template' provided in the 'Final Project Folder' on Canvas (under the Final Project Module).

Phase 2: Mid Term Project Completion (Recap)

Your final project builds on your midterm project. Ensure the following components were completed during the midterm:

Midterm Completion

- **Problem Formulation**: Define the title and scope of your project.
- **Industry Challenge**: Describe the business context and the specific challenge your project addresses.
- **Data Set Preparation**: Outline how you acquired, cleaned, and managed your data.
- **Proposed Technical Solution**: Diagram

Start here-Phase 3: Final Project Execution (Data Engineer Focus)

Technical Implementation [Continue with your MidTerm Project]

1. **Data Engineering Pipeline: your proposed data engineering pipeline for your project (provide your designed schema):**
    a. **Data Ingestion:** Tools and applications
    b. **Data Storage:** Tools and applications
    c. **Data Processing:** Tools and applications
    d. **Data Consumption**: Your App (Tools and applications)
        i. **Model Deployment**: Describe how you've created a deployable environment for your model.
        ii. **Data Visualization:** Showcase the results through comprehensive visualizations.

2. **Operationalization**: Explain how your project creates added value from the end-user perspective.
3. **Pitch Deck Submission**: Submit a summary of your project, including the problem, data, proposed solutions, and tools applied.
4. **Pitch Deck Video Presentation**: A 3-4-minute video summarizing key aspects of your project.

Extended Description of PHASE III

1.Data Ingestion-Data Collection

Requirements:

Define data sources: social media APIs (e.g., Twitter, Reddit), public datasets, or web scraping.
Ensure legality and ethical considerations in data collection.
Implement a method to collect real-time or historical data.
Include instructions on API usage, accessing public datasets, or web scraping tools.

2. Data Storage

Requirements:

Choose appropriate storage solutions (SQL/NoSQL databases, data lakes, or cloud storage).
Design database schema if applicable.
Ensure data integrity and security.
Provide guidelines on setting up the storage solution and importing the collected data.

3. Data Processing and Cleaning

Requirements:

Implement data cleaning methods to handle missing values, duplicates, and incorrect data.
Use data transformation techniques to prepare the data for analysis.
Ensure the scalability of processing scripts for large datasets.
Document the data processing steps and logic used.

4. Data Consumption

a) Analysis and Machine Learning (Optional)

Requirements:

Define specific analysis or machine learning tasks (e.g., sentiment analysis, trend analysis).
Choose appropriate tools and libraries (e.g., pandas for data analysis, scikit-learn or TensorFlow for machine learning).
Implement models or analysis techniques to extract insights from the data.
Validate and evaluate the analysis or model performance.

b) Data Visualization and Reporting

Requirements:

Create visual representations of the data and analysis findings (using libraries like matplotlib, seaborn, or dashboards in tools like Tableau or Power BI).
Ensure visualizations are meaningful and support the project's objective.
Prepare a final report or presentation that summarizes the methodology, analysis, findings, and implications.

Final Project Deliverables

Ensure your submission includes the following:

1. Technical Report

Your report must document

1. **CRISP-DM methodology {Mid Term Project}:**
    - Title of the Project
    - Business Understanding
    - Data Understanding
    - Data Preparation
    - Modeling
    - Evaluation
2. **Data Engineering Pipeline**: Your data engineering pipeline for your project (provide your designed schema):
    - **Data Ingestion:** Tools and applications
    - **Data Storage:** Tools and applications
    - **Data Processing:** Tools and applications
    - **Data Consumption**: Your App (Tools and applications)
        - **Model Deployment**: Describe how you've created a deployable environment for your model.
        - **Data Visualization:** Showcase the results through comprehensive visualizations.
    - **Deployment**
3. **Operationalization:** Explain how your project creates added value from the end-user perspective

2. Product Pitch Deck & Video Presentation

Prepare a compelling pitch deck and a concise video presentation (**max 3-4** minutes) that includes:

- **Introduction – 30 seconds**

    - **Project title**
    - **Team member** introductions and roles
- **Business scenario overview – 30 seconds**
- **Solution overview – 30 seconds**
- **Architecture diagram of the solution – 30 seconds**
- **Lessons learned – 30 seconds**
- **Demo – 45seconds-1 minute**

3. Set Up a Project Repository

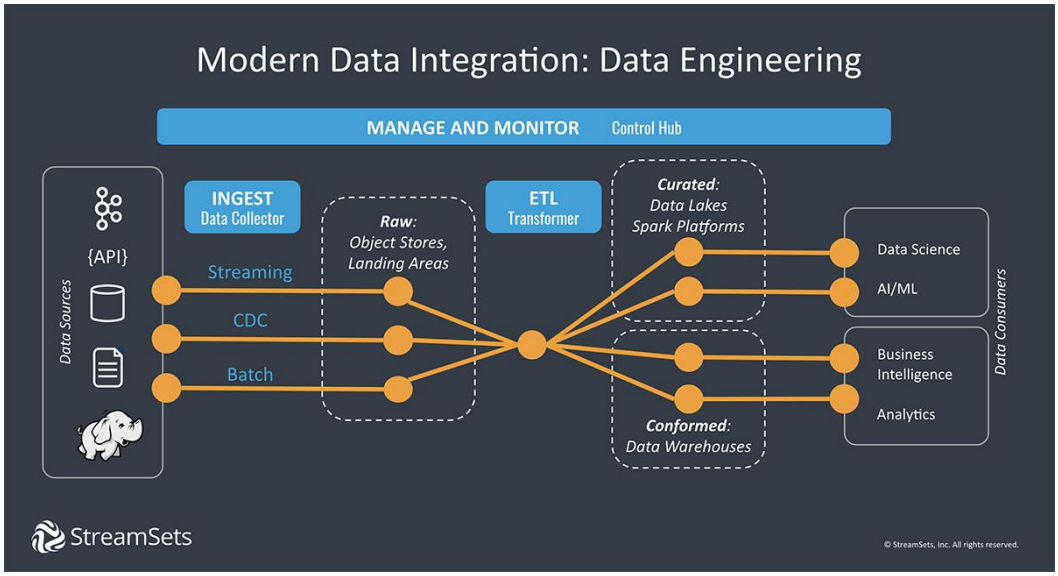
Create a team GitHub repository to manage and track your project work (Learn GitHUb Docs). Your repository should include:

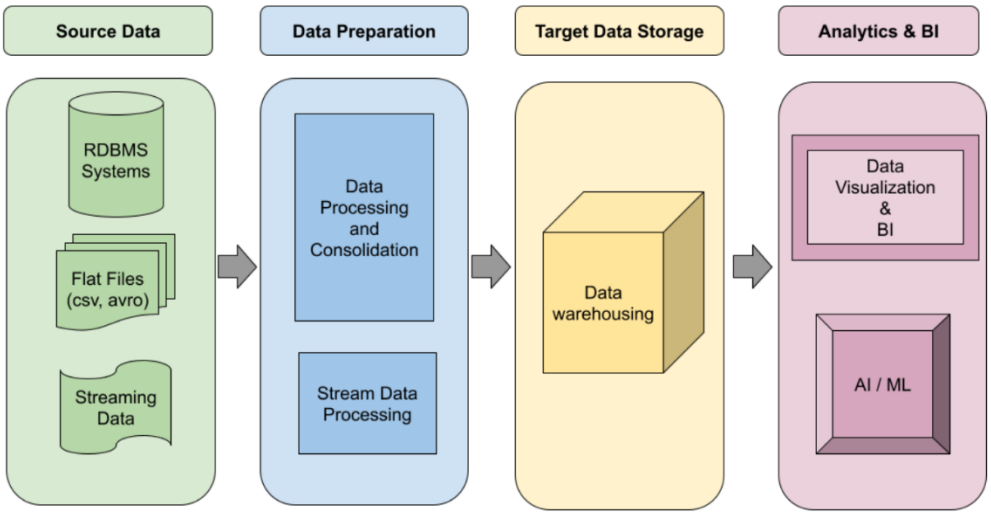
- All project files (data, scripts, models, documentation).
- Organize the repository with clear folder structures: /data, /models, /scripts, /links to your application/docs.
- Ensure all team members have access to the repository and can contribute.

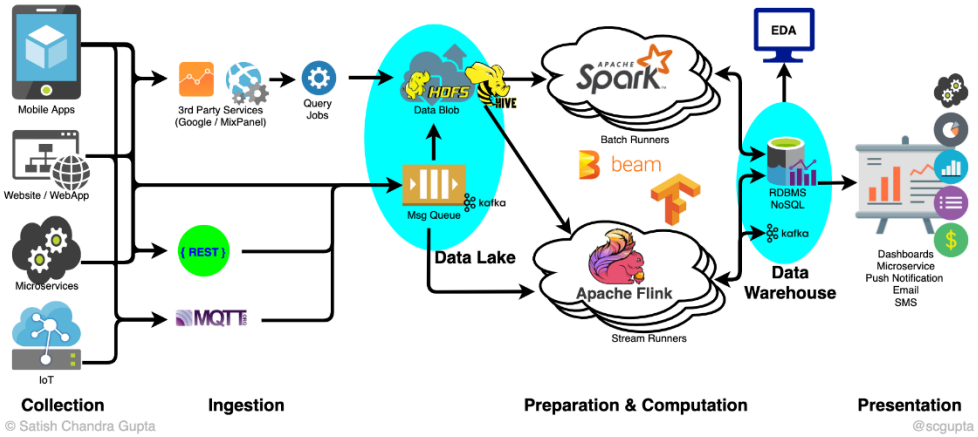
Submission Guidelines

- All team members must submit the same project files via the designated Canvas submission section.
- The GitHub link must be included in your submission.
- **Deadline: 12/08/2024**

Note: Include graphical representations of your Data Engineering Pipeline solution. See the examples:







Assessment Criteria

Your project will be evaluated based on the following:

- **Project Completion**: All Documentation Included
- **Presentation Quality:** Clarity, engagement, and persuasiveness of your pitch.
- **Value Proposition**: The significance and potential impact of your project.
- **Effort Demonstrated**: Comprehensive depth and breadth of work.
- **Originality**: Innovation and uniqueness of your project idea.
- **Relevance**: Suitability of the project for this course and consideration of all data lifecycle stages.
- **Actionability**: Real-world applicability and usefulness.
- **Comprehensiveness** of Deliverables: Inclusion of all required files and documentation.

- **Project Analysis Quality**: Depth of data wrangling, analysis, and clarity of results communication.

Best of luck with your project, we look forward to your innovative solutions!

Project Rubrics

Category I Requirements (Knowledge)	Poor 0.5 pts	Fair 3.5 pts	Good 11 pts	
1.0 Technical Document	Not submitted	Submitted but not suitable for the project	Submitted and suitable for the project	**3 pts**
2.0 Solution-Design: Architecture Diagram	The solution didn't include an architecture diagram.	The solution included an architecture diagram, but the diagram did not show any more detail than the high-level diagram	Completed	**3 pts**
2.1 Microservices/Services Based Architecture	The solution persists the original monolithic application design. There was no evidence of the application running on any containers.	The solution successfully broke up the monolithic application so that it runs on at least two separate containers. However, the solution did not successfully run the containers on Amazon ECS.	The solution properly split the functionality of the monolithic application into two microservices, and the navigation between the customer and employee (/admin) pages worked seamlessly from an end-user experience perspective with no "404 page not found" errors.	**3**
2.2 Scalability/ Resilience	The solution was not successfully built out to use an Amazon ECS cluster.	In the student's solution, the AWS CloudTrail logs did not show any Amazon ECS Update Service events that ran successfully.	The CloudTrail logs showed that Amazon ECS Update Service events ran successfully, and the student's attempt to demonstrate running the aws ecs update-service command with a desired-count parameter also succeeded.	**2**
3.0 Github-Communication of results	Not included	Included but not complete	Completed	**2 pts**
Category II (Presentation)	Poor 0 pts	Fair 0.5 pts	Good 1 pts	
Project Presentation	Not submitted	Submitted but isn't completed	Submitted and presents the project	1
Category III (Originality)	Poor 0 pts	Fair 0.5 pts	Good 1 pts	
Idea	Not original/or innovative	Original/not innovative	Original/and innovative	0.25

Value/effort	Results are not useful	useful	Very useful.	0.50
Actionable	Not actionable	Actionable	Very actionable	0.25
Total Points				15 pts

Final Project Resources

Recommended Courses:

CRISP-DM Methodology Example (Pre-requisite to Final Project Completion)

- 1. CRISP_DM methodology was modeled via the Bank model -link

  Links to an external site.
  - Data Science Methodologies: Making Business Sense
- 2. CRISP-DM methodology used at AWS (https://explore.skillbuilder.aws/learn/course/369/play/1070/process-model-crisp-dm-on-the-aws-stack Links to an external site.)
- **3. End-to-End Data Engineering Project**
  - https://www.linkedin.com/learning/end-to-end-data-engineering-project/what-you-should-know?autoSkip=true&resume=false&u=2359714
  - gitHub

Scientific writing-Resources

1. Marie Davidian:
   http://www4.stat.ncsu.edu/~davidian/st1110a/written_handout.pdf
   Rod Little: http://sitemaker.umich.edu/rlittle/files/styletips.pdf
   Paul Halmos: http://www.matem.unam.mx/ernesto/LIBROS/Halmos-How-To-Write%20Mathematics.pdf
   George Gopen and Judith Swan:
   http://engineering.missouri.edu/civil/files/science-of-writing.pdf

Project Resources

- https://developers.google.com/machine-learning/problem-framing/problem
- https://www.coursera.org/learn/advanced-models-for-decision-making?specialization=analytics-for-decision-making

- https://www.linkedin.com/learning/paths/develop-critical-thinking-decision-making-and-problem-solving-skills?u=2359714
- Set Up a Project GitHub Repository
- Virtual Environments and Packages
- Data Cleaning with Python and Pandas: Detecting Missing Values
- https://seaborn.pydata.org/

Example projects

Example twitter-project topics, to get an idea of how to form the topic yourselves.

- Sentiment analysis - Wikipedia
- https://online.datasciencedojo.com/course/Sentiment-Pipeline-for-Live-Tweets#ccnTab-2
- https://code.datasciencedojo.com/datasciencedojo/tutorials/tree/master/Building%20Real-Time%20Sentiment%20Pipeline%20for%20Live%20Tweets
- Building Real-Time Sentiment Pipeline for Live Tweets · master · Data Science Dojo / tutorials · Code
- Twitter Sentiment Visualization (ncsu.edu)
- Tweet Sentiment Visualization App (ncsu.edu)
- https://careerfoundry.com/en/blog/data-analytics/where-to-find-free-datasets/
- https://developers.google.com/machine-learning/problem-framing/problem
- https://www.coursera.org/learn/advanced-models-for-decision-making?specialization=analytics-for-decision-making
- https://www.linkedin.com/learning/paths/develop-critical-thinking-decision-making-and-problem-solving-skills?u=2359714
- https://online.datasciencedojo.com/course/Sentiment-Pipeline-for-Live-Tweets#ccnTab-2
- https://code.datasciencedojo.com/datasciencedojo/tutorials/tree/master/Building%20Real-Time%20Sentiment%20Pipeline%20for%20Live%20Tweets
- Building Real-Time Sentiment Pipeline for Live Tweets · master · Data Science Dojo / tutorials · Code
- Twitter Sentiment Visualization (ncsu.edu)
- Tweet Sentiment Visualization App (ncsu.edu)
- 2015/02-DataScrapingQuizzes.ipynb at master · cs109/2015 · GitHub
- Justin Blinder
- by Healey and Ramaswamy
    - http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/
- GitHub - bear/python-twitter: A Python wrapper around the Twitter API.
- All Tutorials (datasciencedojo.com)
- Course: What is a Data Engineer (datasciencedojo.com)
- Build a True Data Lake with a Cloud Data Warehouse - Talend | Talend
- https://www.upgrad.com/blog/data-science-vs-data-engineering-difference-between-data-science-data-engineering/

- https://blog.panoply.io/what-is-the-difference-between-a-data-engineer-and-a-data-scientist
- https://www.mastersindatascience.org/careers/data-engineer/
- https://enterprise.2u.com/data-careerkickstarter-mini/
- free student/educator licenses for Alteryx
- https://scikit-learn.org/stable/supervised_learning.html#supervised-learning
- What is Interactive Query in Azure HDInsight? | Microsoft Docs
- Quickstart: Create Spark cluster in HDInsight using Azure portal | Microsoft Docs
- Azure Quickstart Templates (microsoft.com)
- https://developers.google.com/machine-learning/problem-framing/problem
- https://www.coursera.org/learn/advanced-models-for-decision-making?specialization=analytics-for-decision-making
- https://www.linkedin.com/learning/paths/develop-critical-thinking-decision-making-and-problem-solving-skills?u=2359714
- https://online.datasciencedojo.com/course/Sentiment-Pipeline-for-Live-Tweets#ccnTab-2
- https://code.datasciencedojo.com/datasciencedojo/tutorials/tree/master/Building%20Real-Time%20Sentiment%20Pipeline%20for%20Live%20Tweets

GitHub example

- https://github.com/sarahfuchi/Data-Science/blob/main/EDA%20(Exploratory%20data%20Analysis)/README.md

Example data sets:

- https://careerfoundry.com/en/blog/data-analytics/where-to-find-free-datasets/

Useful Resources-Pitch:

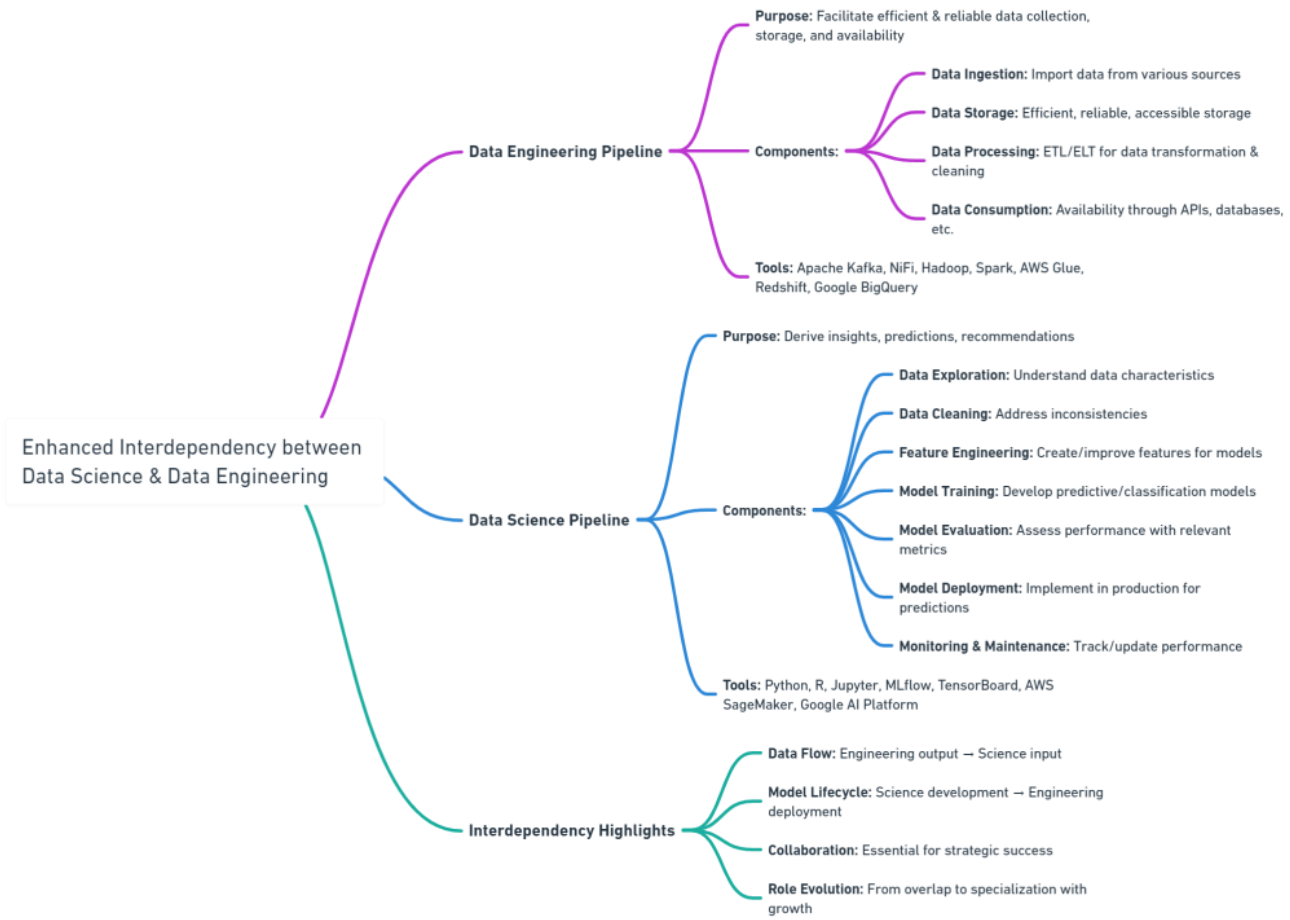
- Here you can find a few useful tips on coming up with a great pitch
- Kevin Hale - How to Pitch Your Startup - YouTube

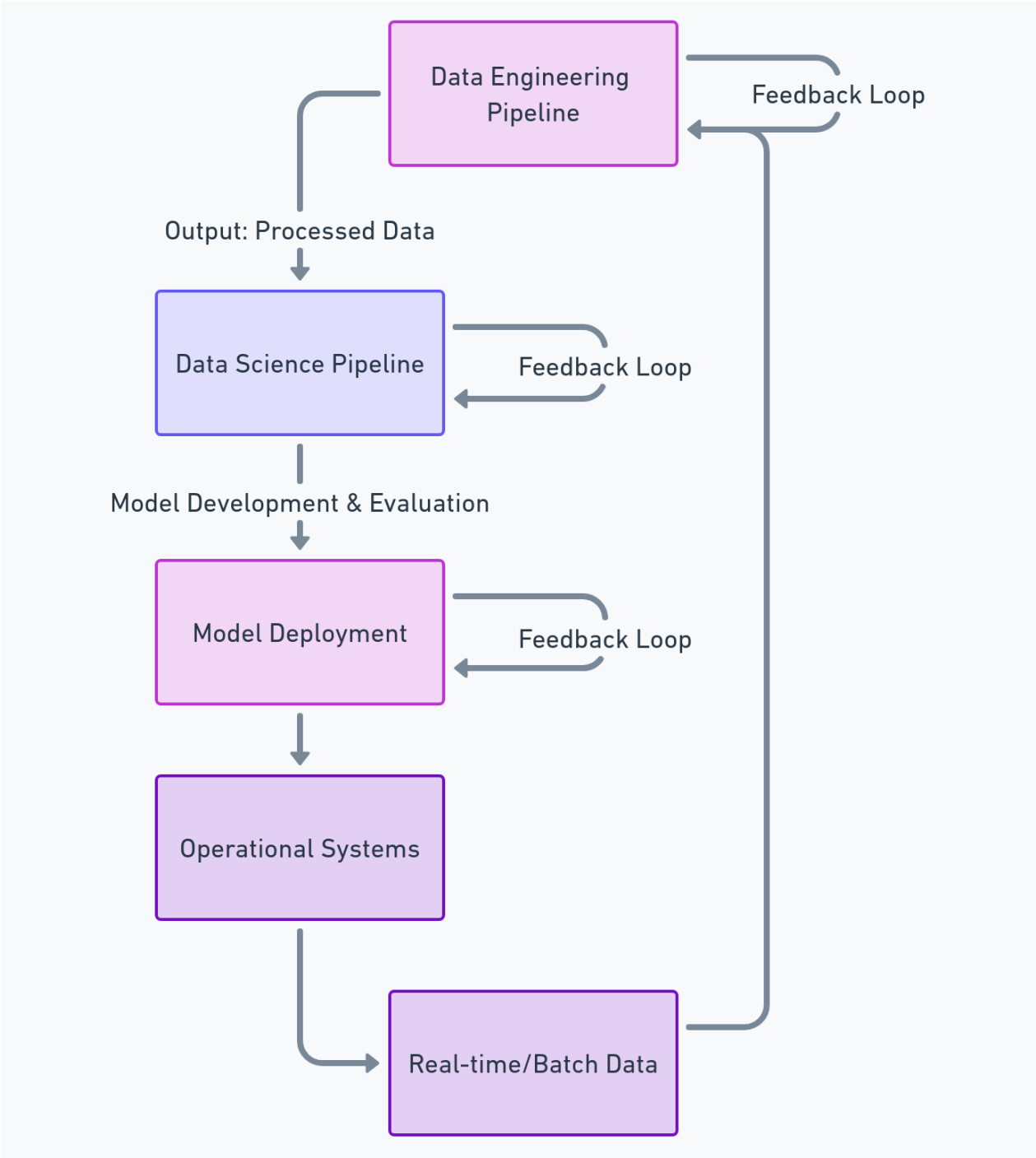
Project Management Tools: Collaborating effectively with your team is key. Platforms like Trello, Asana, or Slack can help keep your team on track.

a. Trello
b. Asana
c. Slack

End-to-End Data Science Project Solution

Interdependency of Data Engineering and Data Science Pipelines




Data Engineering Pipeline
Feedback Loop
Output: Processed Data
Data Science Pipeline
Feedback Loop
Model Development & Evaluation
Model Deployment
Feedback Loop
Operational Systems
Real-time/Batch Data