Building your Intelligent Lakehouse

Business Use Case Scenario

You are a team of consultants for a SaaS (software as a service) company endeavoring to help solve a business use case. Your objective is to create a data lakehouse and a model that provides insights consumed on a dashboard.

Access to the data from Kaggle: (choose any one project from below)

1. https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data
2. https://www.kaggle.com/competitions/predict-energy-behavior-of-prosumers/data
3. https://www.kaggle.com/competitions/nfl-big-data-bowl-2024/data

Please download the data from Kaggle directly

Setting Expectations:

- Use your discretion to choose the relevant tables (minimum four so you can demonstrate a variety of data ingestion, joins, and data wrangling) to meet the requirements listed below and provide valuable insights to reach the business objective of the Kaggle challenge.
- This is a Data Engineering class, not an ML-focused one. From the perspective of this course, we are more interested in the data pipeline than the most performant model. You could use autoML if you choose to. Create a baseline model and describe how you would improve it rather than doing it yourself, as we have limited time and cloud resources. In other words, don’t spend time perfecting and optimizing your model.
- You will be judged for creativity and innovation as well.

Deliverables for Grading:

This assignment will require group and individual grading criteria. The individuals should pick the appropriate roles and submit role assignments as part of the final submission.

Group(20 pts):

Functional and non-functional requirements:

1. Design of various tasks and end-to-end dataflow

        a. Can the reader understand the technical challenges being solved in the diagram?
        b. Was it well articulated?
        c. Will it scale?
        d. What is the data model? (ERD)
        e. Will it be performant?
    2. Is there a final dashboard to present the insights?
    3. Is the design extensible to accommodate new use cases?
        a. Can I get new insights?

Deliverables for Grading:
All students contribute to their role and the group activities, including the presentation. **At most, 20 slides for each section.** You will get **15 minutes** to present and 5 minutes for Q&A. (Time will be kept for you and warned at the 5-minute mark. Please practice beforehand; don't wing it.)

Please use section time to answer any questions about the requirements.

2 Data Engineer (20 pts):

1. Ingesting the various datasets into delta lake tables.
2. Making the pipelines robust enough to be run "daily," assume that you may get data refreshes daily
3. Stream from one delta table to another incrementally using trigger once.
4. Use delta and demonstrate upserts and merges in the gold layer to contain aggregates for reporting in Databricks SQL.

2 Data Scientist (20 pts):

1. Build a model.
2. Use MLFlow for its lifecycle management.
3. Incorporating the ml model into the pipeline to output prediction.
4. Evaluate the difference between your predicted estimate and the actual values.

1 BI Analyst (20 pts):

1. Queries to populate the Databricks SQL dashboard
2. Refresh dashboards daily
3. Relevant visualizations of the insights (5 pts)
4. Create a security model in which you have two groups. Use Grants to do selective access

1 or 2 Data Architects (20 pts):

1. Describe the business use case and the problem that you are solving.
2. Create a detailed ERD with FK and PK relationships. Explain any indexes, cardinality, and scale of all the columns.
3. Explain the impact and performance of various options for partitioning your tables.

4. Business comes back and says they want a streaming solution. How would you design this? (Look at some of the case studies)
5. Refine the provided data flow diagram to add more details of what your data scientists and data engineers have put together as part of the assignment.
6. Articulate the CI/CD deployment process and how you would manage DR for code and data.

Group Requirements:

Choose a leader responsible for coordinating and ensuring progress. The leader and others can choose one of the following roles:

- **1-2** Data Engineers
- **1-2** Data Scientists
- **1-2** BI Analysts
- **1-2** Data Architects

You need to be available in person for the final presentation.

Please evaluate the contribution/effort of each of your teammates (1 per group)
Upload it as part of your deck recording.

Name	Role DE/ML/BI/ architect	Professionalism (1-5)	Timeliness (1-5)	Effort (1-5)	Additional Notes
Abby	Architect	5	5	5	<33333
Kenichi	DE	5	5	5	<33333
Luke	DE	5	5	5	<33333
Liwei	BI/ML	5	5	5	<33333
Peiran	BI	5	5	5	<33333
Selin	ML	5	5	5	<33333
Chijioke	Architect				

Appendix

Rubric

Group(15 pts):
Design of various tasks and end-to-end dataflow (1 pt)
Can the reader understand the technical challenges being solved with the Lakehouse Architecture (1 pt)
Was it well articulated? (deck & the presentation). (2 pts)
Will it scale? (1 pt)
Will it be performant? (1 pt)
Was Delta leveraged appropriately? Medallion architecture? Structured Streaming? (2 pts)
Was MLFlow used for model lifecycle management? (2 pts)
Is there a final dashboard to present the insights? (1 pt)
Is the design extensible to accommodate new use cases? (1 pt)
Are new insights generated using the data? (price, sentiment correlation) (2 pt)
Creativity/Innovation (1 pt)
2 Data Engineer (20 pts):
Ingesting the data source into delta lake tables. (5 pts)
Making the pipelines robust enough to be run “daily” (5 pts)
Stream from one delta table to another incrementally using trigger once. (5 pts)
Use delta and demonstrate upserts and merges in the gold layer to contain aggregates for reporting in Databricks SQL. (5 pts)
2 Data Scientist (20 pts):
Build a model. (5 pts)
Use mlflow for its lifecycle management (5 pts)
Incorporating the model into the pipeline to output prediction. (5 pts)
Evaluate the difference between your predicted estimate and the actuals (5 pts)

1 BI Analyst (20 pts):
Queries to populate the Databricks SQL dashboard (5 pts)
Relevant visualizations of the insights (5 pts)
Refresh dashboards daily (5 pts)
Create a security model with two groups (California and non-California). Only users in the California group can access data in California, while the non-California group cannot. (5 pts)
1 or 2 Data Architect (20 pts):
Create a detailed ERD with FK and PK relationships. Explain any indexes, cardinality, and scale of all the columns. (4 pts)
Explain the impact and performance of various options for partitioning your tables. (4 pts)
Business comes back and says they want a streaming solution. How would you design this? (Look at some of the case studies) (4 pts)
Refine the provided data flow diagram to add more details about what your data scientists and data engineers have put together as part of the assignment. (4 pts)
Articulate the CI/CD deployment process and how you would manage DR for code and data. (4 pts)
Group Dynamics (5 pts)

Ratings

Group Information (Please fill this out one per group)

This is to fill out your group members, how often you have met each other, and what tasks each individual contributed to. Please submit this as part of your final project submission on Canvas.

Please fill out the following:

Name	**Role**	**Tasks**
Abby	Leader, data architect	● Project coordination (setting up meeting times, notes, tracking action items, etc.) ● ERD with PK/FK relationships

		● Data flow diagram ● Table cardinality & scale explanations ● High level overview of ○ Partitioning strategy ○ Streaming solution ○ CI/CD & deployment considerations ○ DR planning
Chijioke	Data architect	
Kenichi	Data engineer	● Implemented Silver Structured Streaming Layer (trigger=once) ● Optimized Gold Layer Performance ● Added Configuration + Data Quality Checks ○ Centralized catalog/schema/volume configuration ○ Table existence checks before streaming ○ Required column validation (lat/long/datetime) ○ Clear error surfacing to prevent silent failures ● End to end lineage digram ● Helper utilities & documentation ○ Developed reusable helper functions used by Data Engineering, Data Science, and BI ● Pipeline Hardening & Documentation ○ Added markdown explanations to notebooks ○ Ensured naming and configuration conventions were consistent ○ Improved maintainability and clarity of the DE pipeline
Liwei	Data scientist; BI analyst	● Creating materialized views for BI & DS personas ● Feature engineering for ML model ● Contributing visualizations to dashboard
Luke	Data engineer	● Creating foundational components for pipeline ● repository & Project Framework ○ Created the GitHub repository and initial notebook structure

		○ Established folder organization used throughout the project ● Bronze Layer Ingestion ○ Converted raw Kaggle CSVs into Delta table ● Batch Silver Layer ○ Joined weather data with county mapping ○ Produced initial Silver weather tables used by downstream consumers ● Gold Aggregation Layer (Batch) ○ Implemented the first Gold-level business table: ○ Built gold_daily_energy_report ○ Performed daily aggregations ○ Added Delta MERGE logic for incremental updates
Peiran	BI analyst	● Driving forward creation of dashboard ● Creating user groups to securely view dashboard
Selin	Data scientist	● Built a model ● Evaluate the difference between predicted estimate and the actuals ● mlflow for lifecycle management ● Incorporating the model into the pipeline to output prediction

Any additional special mentions (any recognition whom you want to share with individuals in the group):

- Luke & Kenichi for setting up the pipeline so quickly, which allowed the rest of the team to move forward with their work in a timely manner
- Liwei for supporting both the BI and DS function - especially re: creating materialized views for folks to work with
- Peiran for creating a beautiful dashboard, with three interesting storylines and really neat visuals
- Selin for driving forward the model creation & validation process
- Kenichi & Selin for keeping clear & open communication with the group while in timezones extremely different to ET
- Abby for getting us off on the right foot, staying organized, and excellent note taking!

Individual Ratings

Please rate your team members' professionalism, timeliness, and group effort. **Then, email your ratings to the teaching staff at Amahapatra@g.harvard.edu**

Please fill out the following (1-5 with 5 being the best):

Name	Role	Professionalism (1-5)	Timeliness (1-5)	Group Effort (1-5)	Additional Notes