KPMG Capstone Meeting Notes, Spring 2024
Team Members: Phillip Kim (ppk2003), David Huang (th3061), Jerry Wang (zw2888), Tom Yu (ty2487)
Mentors: Jim Leach, Ben Carper,
Additional Mentor: Thomas Covella
Advisor: Prof Sining Chen

Monday Jan 29, 2024

Attendees: All Team Members, Mentors, and Advisor

- This kick-off meeting was held to discuss the project.
- Team members were able to ask questions on the project background, scope, expectations.
    - Background for the project is rooted in real-world client needs for model-risk management which includes the understanding of model data lineage, ownership, dependencies, etc.
    - Project team is expected to generate synthetic model meta-data for the project
    - The team will be provided additional details regarding the project later in the week.
    - In the meantime, the team is going to read-up on model meta-data and graph databases, specifically neo4j.

Friday Feb 2, 2024

Meeting with Mentor
Attendees: All Team Members and Thomas Covella

- Thomas agreed to meet us to discuss the additional project details we received the day before.
- The team was able to get further additional details about the project.
    - The entire dataset includes organizational data that leads to reports, in additional to model meta-data.
    - Data Flow: Organization data -> models -> reports
    - A large part of the discussion was around how to create synthetic data.
        - All data must be synthetic and we discussed a couple of ways to generate synthetic data with a combination of chat-gpt and python packages.
        - Aim for around at least 3 different reports and different versions of those reports. Handful of fields of reports and models.
        - Model metadata may be populated using output from ML flow (model management tool) if there is any existing data that we may have. Examples of model metadata include, version, data lineage, location,

lifecycle stage, creation time, last update time, owner. We could also search of model metadata online.
        - One suggestion was to start with the sample dataset provided by neo4jaura (cloud-based graph database) and then add models and reports as additional nodes.
    - We also discussed the orchestrator flow including the integration of the graph database with the LLM. This should be written in python and some of the provided resources provide templates you can use.
    - The chat query will also need to be able to directly query the graph database. In the case of neo4j it would need to translate the natural language into a cypher query. However, Thomas noted that we could have a handful of use cases with few predefined queries written, then use a classification model that will figure out which of the predefined queries to use. Fuzzy matching may be used to find approximate matches. Langchain has various tools related to this.
    - Steamlit is a simple user interface with the LLM. The website provides sample code.
    -

Team Meeting
Attendees: All Team Members

- We discussed the project amongst the team.
- We decided that Phil Kim would be the team lead.
- We set a goal to generate the synthetic data before the next meeting with mentors which is scheduled for Monday Feb 5, 2024.
    - David and Tom decided to look into synthetic data generation
    - Phil and Jerry decided to look for additional datasets.

<u>Sunday Feb 4 2024</u>
Attendees: All Team Members

- Questions for mentors:

    - Meta-data for models (schema). Is data lineage stored in the model metadata or is it implied by the graph db relationships.
    - How to work on neo4j simultaneously or collaboratively
        - Yes
    - Is the northwind dataset complex/large enough?
        - We need more database
    - Our understanding is that we just need to add two new types of nodes: Reports and Models, is that correct?
    - Is a model represented as one node or many as in the presentation.

    - Should we generate data based on sample or schema? I feel like schema one is more similar to the setting of neo4j

Project Milestones

Create Synthetic Data
Build Orchestrator with User Interface
Build connection with LLM
Build connection with graph DB

Monday Feb 5 2024
Attendees: All Team Members and KPMG Ben and Jim

- Ben: We've used the langchain synthetic data generation package to help create real-word data: https://python.langchain.com/docs/use_cases/data_generation
- Ask for recording access
- Query example:
    - Where the data in the report come from?
    - Model difference version comparison
    - Describe the quality of data (avg, min)
    - What are Input, requirement, if change → does it work?
        - Describe the input of the model
        - InputA → model weight(linear regression) → what's the change in variable
    - Explain how the model work, feature importance, data lineage,

Wed Feb 7 2024
Attendees: All Team Members and KPMG Thomas

- Thomas sent over a diagram of what the model nodes should look like and we have some questions about it.
- Tom went into more details surrounding the graph DB schema in terms of all the nodes needed to complete the data pipeline.
    - We can use json files to upload the schema into neo4j
    - Json files should include the relationships as wellj
    - We need to add data elements and data science user
    - Reports should be made up of multiple fields

Mon Feb 12 2024
Attendees: All Team Members and KPMG Ben and Jim

- David led the meeting today. Each team member demonstrated various parts of the project plan so far to get approval/buy-in from the mentors.
- Jerry presented the DB schema.
- Phil presented the overall graph DB schema. Need to add report owner and modify the model version relationships.

- The reports need to have sections which need to have report fields. We need to have at least two reports that are more complex and also the data that feeds into each report need to be consistent with the scenario.
- David presented the API implementation of json -> neo4j. We need to suppress certain nodes to match our graph DB schema.
- Numan asked questions about the orchestrator. We are allowed to use open source or non-commercially licensed code.
- Ben and Jim were generally ok with all the progress we've made so far.

Mon Feb 19 2024

Attendees: All Team Members and KPMG Ben and Jim

- Jerry presented the initial json files representing realistic reports and models for upload into neo4j. The mentors were satisfied with the schema.
- David presented the upload of the initial json files into neo4j.
- Numan presented the initial demo of the orchestrator with the openai and neo4j connections.
- Mentors seemed to be satisfied with the progress so far, but asked to see a project timeline.
- Phil will create a project timeline to keep track of progress.

Mon Feb 26, 2024

Attendees: All Team Members and KPMG Ben and Jim

- Tom presented the additional json files representing realistic reports and models for upload into neo4j. The mentors were mostly satisfied by the completeness of data.
- Davis presented the upload of the complete set of json files into neo4j.
- Numan presented the additional features of the orchestrator including an initial set of standard queries. The orchestrator included the use of langchain.
- Phil presented the initial project timeline which needs to be more completely populated.

Mon Mar 4, 2024

Attendees: All Team Members and KPMG Ben and Jim

- Numan presented the new UI that included an improved workflow
- The mentors suggested more interesting user queries to use in the prompt template.
- They also suggested the use of fuzzy matching to improve accuracy.

Mon Mar 11, 2024

Attendees: All Team Members and KPMG Ben

- This meeting was optional since it is taking place during spring break and the meeting was primarily used to ask outstanding questions.
- Phil asked specifically about the template questions and which questions would be better than the ones we have now.
    - One suggestion by Ben was to ask what changed from one version of the model to the next.

    - The other was to ask how many steps upstream was the data source.
- Jerry specifically asked about how to improve the prompt template and setup a testing regime.

Mon Mar 18, 2024

Attendees: All Team Members and KPMG Jim

- Numan led today's meeting
- The primary discussion was around orchestrator workflow.
- The current configuration was for langchain to handle both common and uncommon questions, but langchain is unpredictable in cypher query generation.
- Jim recommends using a different workflow for common question with pre-written cypher queries.

Mon Mar 25, 2024

Attendees: All Team Members and KPMG Ben & Jim

- Jerry led today's meeting
- The primary discussion of the orchestrator workflow continues.
- Jerry presents a draft diagram that summarizes the orchestrator workflow which includes a different workflow for common, uncommon and irrelevant questions.
- The workflow includes details regarding parameter extraction and entity correction.

Mon Apr 1, 2024

Attendees: All Team Members and KPMG Ben & Jim

- Tom is leading today's meeting.
- Tom presented how he generated test questions for common, uncommon and irrelevant questions. There was some feedback on making sure that the questions ask about specific nodes.
- Jerry and Numan provided an update on the orchestrator.
    - Jerry presented a new diagram about the orchestrator logic workflow
    - Jim suggested a follow-up questions by the orchestrator in order to increase accuracy
    - Numan suggested dividing up the role of the templates into two and also making sure we have a follow-up question to make sure we have all parameters are captured from the user.

Mon Apr 8, 2024

Attendees: All Team Members and KPMG Ben & Jim

- David is leading today's meeting.
- David presented diagrams that will be included with certain common questions as well as an overall graph DB schema that will be included on the left hand side of the streamlit user interface.
- Numan and Jerry provided an update on the orchestrator.

Mon Apr 15, 2024

Attendees: All Team Members and KPMG Ben & Jim

- Phil is leading today's meeting.
- Numan presented an update on the testing script.
- The group also discussed what to include in the posterboard.
- We also discussed the live presentations that need to be prepared for DSI, as well as the two KPMG presentations - one for the graph guild and the other for the executives.

Mon Apr 22, 2024

Attendees: All Team Members and KPMG Ben & Jim

- Phil is leading today's meeting.
- Numan presented an update on the testing script and the difficulty of coming up with a metric to measure final output response accuracy.
- We further discussed the live presentations that need to be prepared for DSI, as well as the two KPMG presentations - one for the graph guild and the other for the executives.

Mon Apr 29, 2024

Attendees: All Team Members and KPMG Ben & Jim

- Phil is leading today's meeting.
- We presented the testing results for intent matching, parameter extraction, and final response accuracy.
    - Common questions scored pretty well in intent matching and parameter extraction.
    - However, it didn't score well in final response accuracy.
    - Uncommon questions did not score well in final response accuracy.
- We also further discussed the live presentations that need to be prepared for DSI, as well as the two KPMG presentations - one for the graph guild and the other for the executives.

Mon May 6, 2024

Attendees: All Team Members and KPMG Ben & Jim

- Phil is leading today's meeting.
- We received specific and detailed feedback from mentors on our final presentation for KPMG executives.