# 1. Problem Statement

We're working on an ambitious project to develop one of the more cutting-edge marketing analytics tools, empowered by Snowflake's leading data platform. Key objectives include a prototype of this tool, which will prove its potential for change in the manner we do marketing analytics. The project covers two critical use cases:

1. Need for dynamic and interactive querying of our marketing data.
2. Demand for a more intuitive, natural language interface in data analysis.

This prototype will help us demonstrate the use of big data analytics, cloud computing, and artificial intelligence in gaining more insights into our marketing efforts and making data-driven decisions more efficiently.

# 2. Task

Our task is to create a prototype of the marketing analytics tool with the following key components:

* A Streamlit application to run complex SQL queries in Snowflake against the TPC-DS dataset by dynamically changing query parameters.
* An AI-powered chatbot interface integrated with the Streamlit application that empowers users to query data on Snowflake using natural language.

These components will enable the tool to fulfill the needs for structured, parameter-based querying and flexible, conversational data exploration.

# 3. Architecture Diagram

A diagram of marketing tools

Description automatically generated

This architecture shows the flow of data and interactions in our prototype:

1. Users interact with the Streamlit frontend.
2. The frontend can directly send SQL queries to Snowflake or pass natural language queries to the LLM chatbot.
3. The chatbot converts natural language to SQL.
4. Snowflake executes queries and returns results.
5. Results are displayed to the user through the Streamlit interface.

# 4. Data Sources

The source of data for this project will be the benchmark dataset TPC-DS, which is also available in the sample data within Snowflake. Note that database names are TPCDS\_SF10TCL and TPCDS\_SF100TCL, together, we have a total of 48 tables. This dataset emulates the decision support system of a retail product supplier. It contains sales data through store, catalog, and web channels, data related to returns and inventory levels, and data about customers. This TPC-DS dataset is very much suitable for this project due to the following reasons:

1. Realistic representation of data that a retail business would have.
2. Wide variety tables and relationships can be used to run complex queries.
3. It is readily available in Snowflake, which is pretty easy to set up and play around with.

# 5. Data Pipelines (If any)

For this prototype, we will not implement complex pipelines for data, since we are going to use the pre-loaded TPC-DS dataset available in Snowflake.

# 6. Business Logic (If any)

Much of the business logic in our prototype is really encapsulated in the SQL queries we chose to implement and in the NLP capabilities of our chatbot. Key aspects include:

1. Selection and testing of 10 relevant queries from the TPC-DS benchmark, useful in providing several marketing insights.
2. Query parameterization will be done to enable flexible analysis.
3. Designing a chatbot instruction prompts freely and efficiently translate marketing-related questions into SQL ones.
4. Interpretation and presentation of the results of the query, enterprise friendly.

# 7. Future Work

While our prototype demonstrates core functionality of the marketing analytics tool, there are several areas for future development:

1. The dataset is large; sometimes it takes a couple of minutes to produce the results.
2. ChatGPT 3.5 has limited intelligence and is likely to produce incorrect SQL queries; we may implement retries or better AI models in the future.
3. Integration with real-time data sources to have up-to-date marketing analytics.
4. More advanced visualizations and dashboards within Streamlit
5. More advanced NLP capability of the chatbot in handling complex queries, and adding explanation of result
6. User Authentication and Role-based Access Control for the security of data
7. Integrate with other marketing tools and platforms for a more comprehensive analytics ecosystem.

# 8. References

1. TPC-DS Benchmark Specification: https://www.tpc.org/tpc\_documents\_current\_versions/pdf/tpc-ds\_v2.5.0.pdf
2. Snowflake SQLAlchemy Toolkit: https://github.com/snowflakedb/snowflake-sqlalchemy
3. Snowflake & Streamlit Tutorial: https://www.youtube.com/watch?v=3WH1DHBf2WQ
4. Frosty LLM Chatbot on Streamlit & Snowflake: https://quickstarts.snowflake.com/guide/frosty\_llm\_chatbot\_on\_streamlit\_snowflake/#0

# 9. Additional Notes (Important)

This assignment provided valuable hands-on experience in building a marketing analytics tool using cutting-edge technologies like Snowflake, Streamlit, and LLM-powered chatbots. However, several challenges and areas for improvement were identified during the development process.

Debugging took quite a long time, especially setting up the configuration in secrets.toml. There were no clear and short documents on the parameters required, so going through all the documentation was quite intensive and time-consuming. One of the very important lessons brought out by the experience is that, for detailed systems, comprehensive yet accessible documentation must also be availed. Although potentially helpful, the video walkthrough did get a bit redundant as I found myself always needing to go deep into the written documentation. Streamlining of the process in future incarnations of this assignment would go a long way. Perhaps a quick-start guide would be very useful in this area besides more detailed documentation.

The TPC-DS dataset specification itself is 138 pages long, thus posing a challenge in itself with regard to understanding and the use of such data. It would therefore help if a set of example queries were included that could be run by students to test the basic functionality of their chatbot. This would help not only as a test tool but also in getting a clearer understanding of the expected outcomes.

Following these suggestions can significantly help improve the learning experience for students so that they focus on the core objectives of the assignment rather than troubleshooting configuration issues.