Article

Not peer-reviewed version

AI-Powered Data Vault 2.0 Modeling for Business Intelligence and Automation

Andreea Vines , Ana-Ramona Bologa * , Andreea-Izabela Bostan

Posted Date: 25 February 2025

doi: 10.20944/preprints202502.2012.v1

Keywords: Data Warehouse; Data Vault; Data Solutions; Prompt Engineering

 doi:10.20944/preprints202502.2012.v1

Article

AI-Powered Data Vault 2.0 Modeling for Business Intelligence and Automation

Andreea Vines, Ana-Ramona Bologa * and Andreea-Izabela Bostan

Department of Computer Science and Cybernetics, Bucharest University of Economic Studies, 010374 Bucharest, Romania
* Correspondence: ramona.bologa@ie.ase.ro;

Abstract: This study explores the innovative application of Artificial Intelligence (AI) in revolutionizing data engineering practices, specifically focusing on the enhancement of the Data Vault modeling process in the context of big data environments. By leveraging the TPC-DS data set, a widely recognized industry benchmark that simulates complex, large-scale data scenarios, the research investigates the capabilities of ChatGPT in automating, accelerating, and refining the creation of Data Vault models. The methodology includes an iterative approach where ChatGPT generates models using various prompt engineering techniques. Comparative analysis is conducted against traditional modeling methods, emphasizing critical factors such as scalability to massive data sets, the speed and efficiency of model creation, precision in handling diverse data formats, and the AI's adaptability to dynamic schema changes. The study also examines ChatGPT's ability to seamlessly integrate new, high-volume data sources into existing models while maintaining performance in big data processing contexts. The findings aim to uncover insights into the practical viability of ChatGPT as a transformation tool for data practitioners, highlighting its potential to ensure higher accuracy and streamline complexities inherent in large-scale Data Vault modeling. This exploration serves as a foundational step toward understanding the broader implications of AI in advancing the state of modern big data warehousing and analytics.

Keywords: Data Warehouse; Data Vault; Data Solutions; Prompt Engineering.

1. Introduction

The rapid expansion of Big Data has fundamentally transformed a wide range of industries, from healthcare and finance to retail and manufacturing, creating both opportunities and challenges in data management. The increased volume, velocity, and variety of data generated by different businesses have made traditional data management practices insufficient. As a result, there is an urgent need for advanced data management and analysis techniques capable of handling these complexities. In response, organizations are increasingly turning to cloud computing platforms, which offer scalable infrastructure and powerful processing capabilities that enable them to store, process, and analyze vast amounts of data with flexibility and efficiency. A study conducted in 2022 was analyzing how different key sectors, Education and Healthcare, were moving their activity to cloud, being able to select the option that fits the best, in terms of their needs – including accessibility, cyber-security and elasticity [1]. Furthermore, a study conducted by one of the current paper's authors regarding technologies used for data products, among multiple data professionals, was showing that only around 7% of the respondents were working with on-premise technologies, rest of them were involved in one or multiple projects that require cloud computing resources [2].

However, data continue to grow exponentially, activity which was driven by digital transformation, IoT (Internet of Things) and smart devices [3]. Noways, companies use data from various sources to find insights, hence selecting the appropriate data modeling approach becomes crucial. The right data modeling framework can greatly impact the efficiency of data storage, the speed of data retrieval, and the precision of analysis. A well-structured data model ensures that information

 doi:10.20944/preprints202502.2012.v1

is organized in a way that allows businesses to quickly access relevant insights and make confident strategic data-driven decisions. Without an optimal model, businesses risk facing data bottlenecks, redundancy, and challenges in scaling their operations effectively. In addition, the ability to integrate diverse data sources, maintain data quality, and adapt to evolving data requirements hinges on the selection of a robust data modeling approach. Therefore, the importance of data modeling frameworks cannot be overstated, as they are fundamental to achieving operational efficiency, supporting advanced analytics, and driving innovation in a data-driven world.

Data Vault modeling is a critical component in modern data warehousing, offering a scalable and flexible approach to managing large data sets. It was firstly introduced by Dan Linstedt, after the 2000s. In 2013, the author released a new version of the model, adapted to today's challenges, and designed to provide long-term historical storage of data coming in from multiple operational systems. It emphasizes flexibility and scalability, which makes it particularly suitable for environments where data is expected to evolve rapidly and unpredictably [4]. Building the data model, using any technique, would require a significant portion of the development time dedicated to tasks like data modeling and mapping between the initial source systems and the desired target system. This phase involves understanding business requirements, designing the model, and structuring data for efficient querying and reporting. In addition, this activity typically involves complex transformations and reconciliations, which further contribute to the time and cost of building a data warehouse. This complexity is further compounded by the need to regularly update the model in response to new data sources or evolving business requirements. As data volumes and complexities increase, these challenges become more pronounced, leading to significant delays and potential errors in the data warehouse process, according to a study conducted to data professionals [5]. ETL (Extract, Transform and Load) processes, including data modeling and mapping, are often the most time-consuming components of data warehouse development, consuming significant amounts of resources and time, particularly when integrating data from disparate sources [6].

Artificial Intelligence (AI) is playing a critical role in the digital transformation of engineering design, particularly by Large Language Models (LLMs) such as GPT, Copilot, or MetaAI. These models facilitate more efficient, data-driven design processes by automating routine tasks, offering innovative solutions, and enhancing creativity in problem solving. Fan et al. discuss the emerging applications of Large Multimodal Models (LMMs) in engineering, emphasizing their potential for accident analysis, human behavior modeling and intervention design, which are able to assess and predict severity [7]. Another study performed in 2024 relates the performance of Chat GPT 4.0 in human engineering tasks, comparing it with 5th-year engineering students in a hackathon. The experiment showed that the LLM can have practices similar to those of the human participants, finishing in the second place. The model was able to provide instructions and create a functional prototype, but faced some limitations such as adding unnecessary complexity and struggling with consistent communication. The study highlights how LLMs can be integrated, using them for innovation and maintaining human oversights [8]. These examples illustrate how AI and LLMs are reshaping engineering design practices, introducing both innovation and challenges in their application, and suggesting a transformative future for these fields.

Although numerous studies have highlighted the effectiveness of AI in various data engineering tasks [9,10], research specifically focused on its application to Data Vault modeling is still limited. This study aims to fill this gap by thoroughly evaluating ChatGPT's capability to generate and refine Data Vault models, focusing on its potential to **enhance the speed and accuracy of the modeling process**. Moreover, the research examines ChatGPT's ability to **manage dynamic changes within the data model,** such as the integration of new data sources, a task that typically requires substantial manual effort. Through this exploration, the study seeks to provide deeper insights into how ChatGPT can streamline and improve the overall process of Data Vault modeling.

This research paper addresses two central research questions that explore the use of ChatGPT in Data Vault modeling.

 doi:10.20944/preprints202502.2012.v1

Q1:The first question is: *How can ChatGPT be leveraged to create efficient Data Vault models*? It focuses on the practical application of ChatGPT in the data modeling process, examining how AI can be utilized to streamline and expedite the creation of Data Vault structures.

Q2: The second question is: *Which are the benefits and limitations of using LLMs in this context?* This question aims to assess the advantages of using Large Language Models, such as ChatGPT, in terms of speed, accuracy, and adaptability, while also identifying potential challenges, such as the AI's ability to understand complex data relationships or manage evolving data requirements. Together, these questions aim to provide a thorough understanding of the role that LLMs can play in modernizing and optimizing Data Vault modeling.

This study is particularly valuable, as it is a starting point for using AI models to significantly accelerate and optimize the design and implementation process of data warehouse solutions. The focus is on Data Vault modeling, a method known for its robustness in handling large and complex data sets [4]. However, the inherent complexity of Data Vault modeling often leads to longer initial modeling and implementation phases, as observed in an interview addressed to multiple Data professionals involved in designing and implementing Data Vault models [5]. This is due to several factors, including the extensive number of table types (Hubs, Links, and Satellites) that must be created and maintained, as well as the stringent requirements for historical data preservation, traceability, and audit-ability. These factors make Data Vault modeling a time-intensive process, necessitating advanced approaches to improve efficiency. By exploring the application of AI, specifically through the use of Large Language Models like ChatGPT, this study seeks to reduce the time and effort required for effective Data Vault design while maintaining or enhancing the model's quality and compliance with best practices. To provide an accurate and comprehensive analysis, the study uses the TPC-DS data set, which contains multiple data structures, providing a robust framework to assess the performance and adaptability of the systems [11].

This paper brings a novel contribution to the field of Data Vault modeling by addressing gaps in the existing literature. Although there is a limited number of specialized studies in the literature that focused on Data Vault, they mainly focus on comparing Data Vault to traditional Data Warehouse methodologies, in terms of ETL implementation, life cycle management, or traceability [12–14]. On the other hand, in terms of AI integration, there is a study provided by Helskyaho that focuses on a limited set of data [15]. This study aims to explore the challenges of modeling large and complex structures and combining data modeling with different prompt engineering techniques. The primary focus of the papers is on providing an overview of Data Vault or evaluating it in comparison to traditional Data Warehouse methodologies. This research goes into greater detail about the challenges involved in modeling complex Data Vault schemas. Specifically, it explores the difficulties of managing a large volume of tables that need to be modeled, dynamic changes in existing models, the unique transformations required by Data Vault, and the potential for human errors in manual model creation.

In addition to addressing the challenges discussed earlier, this research serves as a foundation for creating agile business intelligence and data warehouse solutions by seamlessly integrating Artificial Intelligence (AI) into various stages of the data modeling process. The proposed approach stands out by emphasizing the application of AI to automate various components of data modeling, presenting a more scalable and efficient alternative to traditional modeling methods. This study makes a significant contribution by proposing a practical AI-driven process designed to improve operational efficiency, reduce the complexity of implementing Data Vault models, and ensure consistent adherence to best practices.

The current paper is structured as follows: a short introduction of data modeling warehouse systems and the struggle based on the volume of sources and entities, a comprehensive analysis of the current literature review and the related work, the methodology that is being used, offering a description of the data set used, criteria of evaluating the models, and some insights on how the models were generated. The results obtained by multiple iterations and prompt variations presented in the current paper followed by discussions, conclusions, and future work.

 doi:10.20944/preprints202502.2012.v1

2. Literature Review

Data modeling represents a critical process in the design and management of data warehouses, providing a blueprint of how data are structured, stored, and accessed. Its role is to define and organize the entities and relationships of the data, ensuring that the result is accurate, consistent, and accessible to end users through different applications. There are various approaches to data modeling, each of them tailored to different business needs.

The main methodologies that are being used for data projects are: Dimensional Modeling using Inmon or Kimball approaches and Data Vault Modeling. Bill Inmon is the 'father of data warehousing', proposing a centralized model where data, from multiple sources, are stored in a single data store, which is hosted for all users, using a top-down approach [16]. On the other hand, Ralph Kimball slightly adjusted the perspective of the dimensional modeling, following a bottom-up approach where data are stored in multiple data stores, which was considered the main key of the model, each of them serving specific business needs [17]. It offers a more flexible and responsible solution, by following an iterative agile process. Both approaches follow the same principles and are stored around two core components, fact and dimension tables. Fact tables are being used to store performance metrics, including sales, revenue, stock levels, while dimension tables serve as descriptive attributes for the performance metrics (customers, products, time). The difference between these two approaches is that Kimball's approach is centered on using facts and dimensions to enable analytical insights, whereas in Inmon's approach, they are only using in the final layer, when designing Data Mart for reporting, rest of the model being stored in a normalized format (3NF) [5,18,19] .

The Data Vault model replaces conventional fact and dimension tables with hubs, satellites, and links. To summarize the purpose of these key elements, the hub tables store the core business concepts (such as vendors, products, customers) and only contain the business key and some metadata information, while the attributes and details of the entities are stored in satellite tables, which also conserve the history of the data; link tables are being used to connect between multiple entities which can have different relationships, starting from one to one and until many to many [20]. In addition to the standard key elements described earlier, the complexity of the model can increase by including additional Data Vault components, such as bridge tables or PIT (point-in-time) tables, used to handle large amounts of data and to increase performance [4].

Comparing Data Vault with the traditional Dimension Modeling approach, it has different advantages or challenges. Some of its main advantages are that it represents a support of agile project management, being able to easily split the development phases into sprints. It is also known for its support of integrating multiple sources using satellite tables, relative to Dimensional Model, which requires changing the model to integrate new sources. Another performance is for the loading efficiency, where data can be loaded independently, whereas in the other approach it has multiple dependencies, where only some of the dimension tables can be loaded in parallel. However, this approach also has different challenges, such as query performance issues caused by the large number of tables involved, generating a dependency on the model design as well. Studies showed that it performs well, where some of the challenges mentioned could be addressed depending on the needs [19,20] .

Considering all the challenges and effort involved in designing the data model in Data Vault, the integration of AI and LLM (large language models) might offer a powerful advantage. The AI can speed up the traditionally time-consuming tasks of getting the data definition, by quickly generating data structures, identifying potential relationships. In a recent paper, Helskyaho [21] used LLM models to generate data definition structures for a Data Vault model, providing the initial data structure as input. Using different iterations, the author observed that the AI tool can generate the model but was missing some elements at each iteration. However, the author concluded that AI could be a possible option to automate the design, but further investigations and experiments are still required to validate that. In further research [15], the same author conducted an experiment in which multiple data models were generated using different versions of ChatGPT and compared them using various quality factors, such as correctness, completeness, integrity, flexibility, and simplicity. By analyzing 4 models generated

 doi:10.20944/preprints202502.2012.v1

with AI, it was observed that 2 of them meet more than 80% of the criteria. Both analyses focused on an initial data set with four source tables with information on customers, products, orders, and order details. The current paper aims to analyze how ChatGPT can be used to design a Data Vault model by using a more complex data set that would require the definition of multiple satellites and links between tables. In addition to that, it will also test the performance of AI in terms of defining the changes that are required when additional sources are integrated.

An important element to get the correct and expected results from an LLM is to provide suggestions that can guide the model to the relevant output [22]. A prompt represents a text-based input that is provided to the LLM to see its output. The model interprets the prompt, and it generates a response according to the instruction and the context which were embedded in the prompt, so that the language model has specific instructions to accomplish a particular task. Prompt engineering provides a framework to document patterns that are used to structure prompts and solve various problems and to allow developers to adapt them to different domains [22,23].

The prompt engineering techniques can be used to improve the performance of language models and to allow them to generate more relevant, coherent, and sophisticated responses. Some of the most known and used prompt engineering techniques are:

- Few-shots prompting – which involves providing some examples of the desired input-output pairs to the LLM, so that it can adapt and provide the response based on the pattern observed in the examples provided [24];
- Contextual prompting – which requires providing a context or background information (a few lines describing the area, domain, why it is required to generate the output) to the LLM so that it can base the response on the context provided [25];
- Dynamic / interactive prompting – which refers to the practice of modifying prompts in real time based on previous responses, so that performance improves [26];
- Zero-shot prompting – which requires asking the model to perform the task without providing any examples and relying solely on the instructions within the prompt [27];
- Instruction-Based Prompting – which refers to crafting clear, direct instructions in the prompt to specify the task that needs to be performed; it is usually used in combination with few-shot or zero-shot techniques [28];
- Chain-of-Thought Prompting – which requires the user to encourage the model to think through its reasoning process step by step and to explain its thought process, rather than just providing the final output, which might lead to more accurate and logical outputs [29].

Data solution development can be leveraged with the help of AI which can enhance the efficiency, accuracy, and scalability of various tasks. For example, AI-driven algorithms have been created to detect and correct data anomalies, which can significantly reduce the manual effort required to clean and preprocess the data [30]. Artificial Intelligence can also be used for intelligence data mapping and scheme matching, where AI can align data sources and enable a more seamless data integration across different platforms [31]. All these examples demonstrate the capabilities of AI's potential to transform traditional data tasks and make them more reliable, faster, and adaptable to the growing demands of the market.

3. Methodology

The current paper focuses on responding to two research questions – *How can ChatGPT be used to create efficient Data Vault models?* and *Which are the benefits and limitations of using LLM in Data Vault modeling?* To answer these two questions, an experiment was conducted: to use ChatGPT to generate Data Vault models involving different prompt engineering techniques. Based on an interview conducted with multiple data professionals, it was observed that Data Vault models are efficient, but the design and implementation period requires significant effort [5]. The methodology is structured around iterative prompt engineering techniques, aiming to refine the prompts to achieve an accurate

 doi:10.20944/preprints202502.2012.v1

and comprehensive Data Vault model. The process involves the following steps: preparation of the data set, prompt engineering, and validation and evaluation criteria.

Figure 1 describes the execution flows. The process flow to generate the final output begins with data ingestion, where TPC-DS data is generated according to the provider's instructions. Once the data are ingested, the next step is to extract the table definitions using specific SQL commands, which will serve as input for ChatGPT. These definitions include the table structure, columns, and data types of the source data. For each of the prompt engineering techniques that will be analyzed, one or more requests are sent to ChatGPT. Each request contains specific prompts based on the technique, along with the extracted table definitions. The output from these requests consists of the DDL for the new Data Vault entities, including Hubs, Links, and Satellites. After the DDL is generated, the model is evaluated on the basis of multiple criteria described in the following sections. The final coefficient is calculated by adding the scores for each criterion, providing a total score for each option. In the final step, each model is compared based on the total coefficient, and the results are interpreted to determine the best option, which will then be selected for implementation.

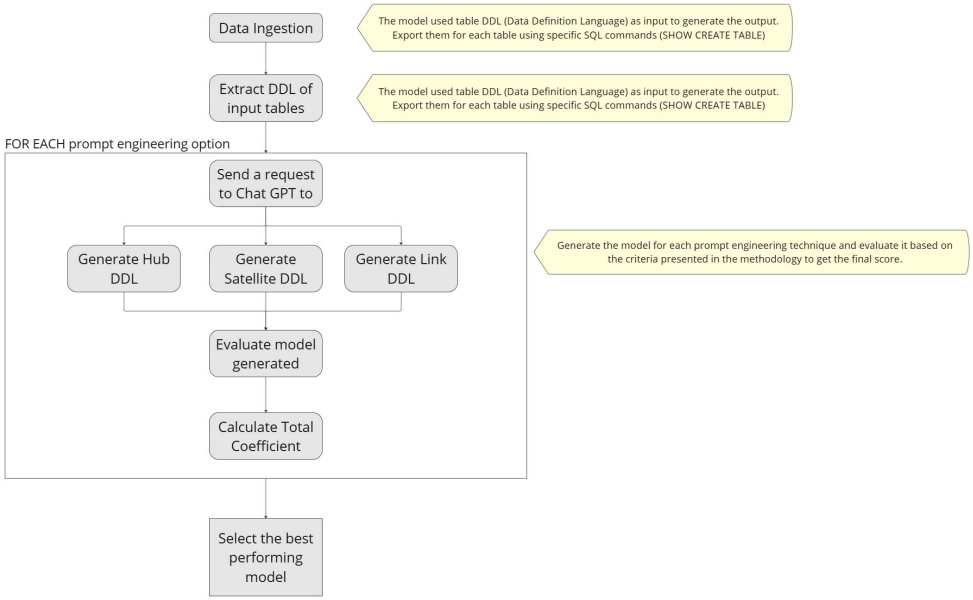

Figure 1. Process execution flow.

3.1. Data Set Preparation

The TPC-DS (Transaction Processing Performance Council – Decision Support) benchmark data set was chosen for the experiment due to its widespread use in the evaluation of decision support systems and its ability to represent complex real-world data scenarios [32]. TPC-DS is a well-established industry-standard benchmark for testing and benchmarking decision support queries and analytics workloads, making it a reliable choice for testing data models in various applications. The data collection is known for its complex schema, which contains around 24 tables with intricate relationships, which makes it ideal for testing advanced data modeling techniques, such as Data Vault. This complexity mirrors the business processes typically encountered in large organizations, ensuring that the model is tested in a realistic environment. In addition, data are also related to commerce, representing a domain in which data warehouse solutions are commonly applied to support strategic business decision making.

In addition to being a robust benchmark for decision support systems, the TPC-DS data set has proven useful in various applications beyond decision support systems, such as in creating dimensional cube models and designing cloud analytics platforms [33,34]. The model represents realistic business scenarios that describe sales and operational data across various channels, including stores, catalogs,

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2025 doi:10.20944/preprints202502.2012.v1

and the Internet, providing a diverse set of data that mirrors the multifaceted nature of modern enterprises. This flexibility allows the data set to be leveraged for multiple modeling approaches.

For the current analysis, a subset of the TPC-DS data model was selected, focusing on sales data from stores, with eight tables covering key business areas such as stores, customers, items, and promotions. This subset provided a rich foundation for generating a Data Vault model. The repository is widely used in research and has been used in numerous studies for benchmarking and performance evaluation, making it a trusted and familiar choice in both academic and industrial contexts. Furthermore, the availability of DDL (Data Definition Language) scripts for creating the tables ensured a structured and consistent data set, which made the model generation process straightforward and reliable. Using this data set, the experiment benefits from a realistic, complex, and standardized test case that allows for meaningful evaluation and an ideal scenario to build various experiments of the model using it.

For the current analysis, a subset of the TPC-DS data set was utilized, focusing specifically on sales data from stores. Nine tables were selected from the benchmark, covering key business areas such as stores, customers, items, and promotions. This subset provided a rich foundation for generating a Data Vault model. To facilitate the process, DDL (Data Definition Language) scripts for creating these tables were extracted from the TPC-DS schema and used as input for the data modeling. This approach ensured a structured and consistent data set, enabling the effective generation and testing of the Data Vault model.

As the benchmark allows data to be extracted using different scale factors, for an accurate analysis, the 1000GB scale factor was used, which contains around 500 million sales records of 12 million customers from 500 stores. The initial model can be observed in Figure 2.

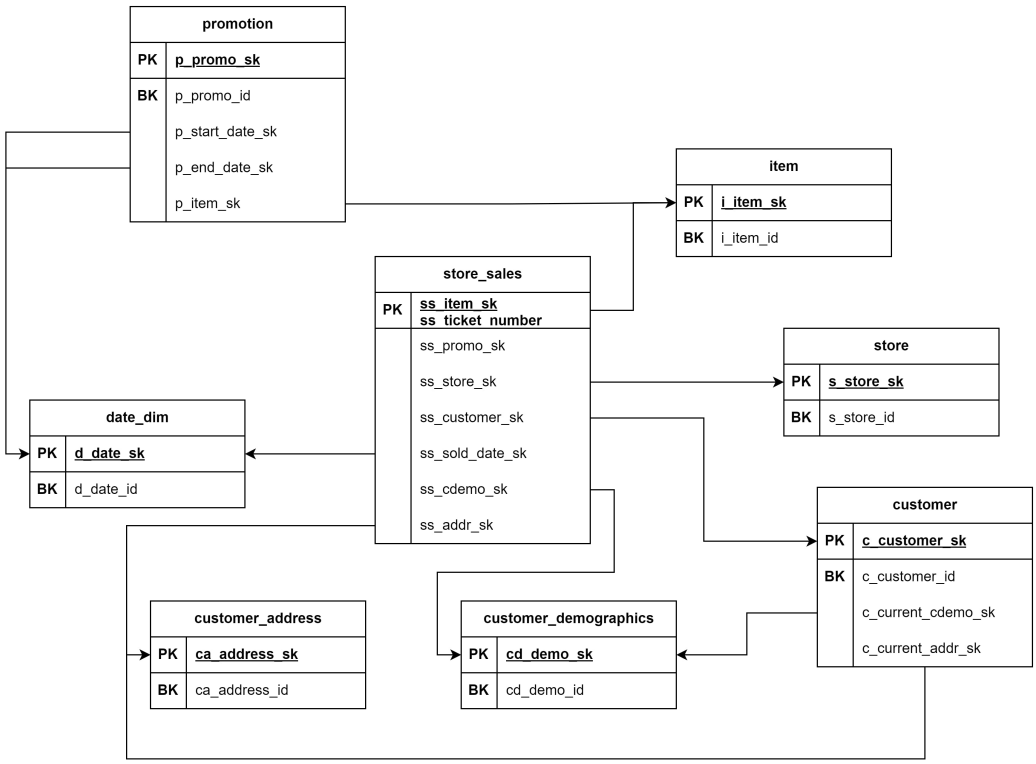

Figure 2. Initial Model

3.2. Data Vault Model Implementation

Data Vault 2.0 represents the latest methodology provided by Dan Linstedt. It was built on top of the foundation of 1.0, adding modern enhancements, such as including the hash key instead of the sequence number for the surrogate keys, and focusing on designs for cloud-based platforms [20].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2025 doi:10.20944/preprints202502.2012.v1

The described data set was modeled following the Data Vault 2.0 principles and presented as part of The 23rd International Conference on Informatics in Economy – IE 2024 [14].

In the model from Figure 3, it can be observed that the following tables were created:

- 4 hub tables for each core business entity - customer, store, promotion, item
- 7 satellite tables - 3 satellites for customer entity (including demographics and address sources), one table for stores, one for promotion, one for item and one for sales data attributes.
- 2 link tables - one between promotion and item and one between all entities which are directly connected with the sales satellite as well.

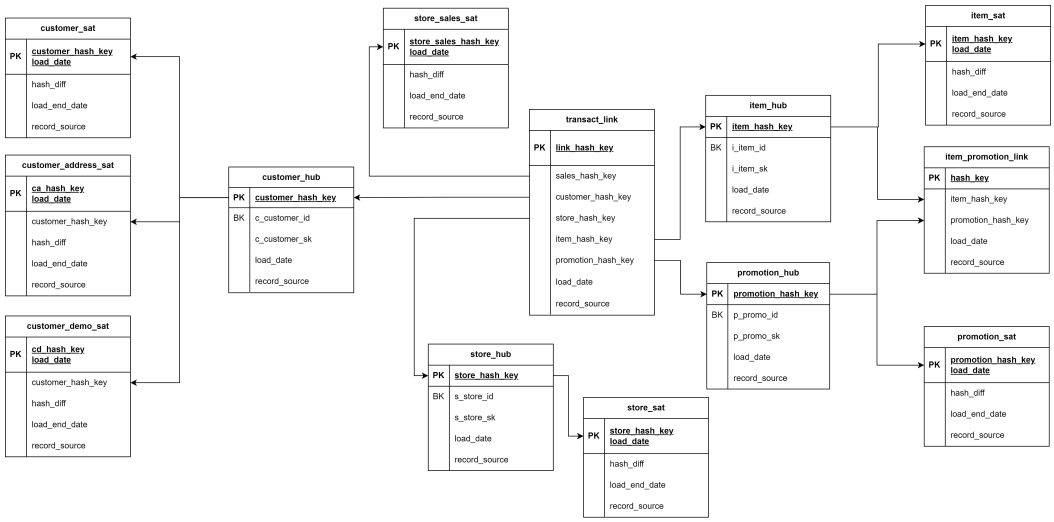

Figure 3. Data Vault Model

This model is used as a reference to validate the designs generated with the LLM.

3.3. Prompt Engineering Techniques

To effectively guide the LLM in generating the responses, different prompt engineering techniques were designed, including few-shot prompting, contextual prompting, dynamic prompting, and instruction-based prompting. Below are the prompts that were used to generate the answers, where each prompt was followed by the initial DDL (Data Definition Language) of the model.

- Option 1 - No Prompt Engineering
  Create a Data Vault 2.0 model by using the below DDL as input.

- Option 2 - Contextual Prompting
  You are an expert in Data Vault modeling. I have a TPC-DS data set represented by DDL scripts. Your task is to generate a Data Vault schema that accurately represents this data set. You need to use Data Vault 2.0 principles to generate the output.

- Option 3 - Few-shot prompting
  Create a Data Vault model by using the below DDL as input. Provide the new model DDLs into a single script.
  Please use this example as reference where there are 2 input tables - bronze_product & bronze_product_details. The output of the provided example will have 1 hub table (hub_product) and 2 satellite tables (sat_product and sat_product_details).
  The output of the provided example will not return 2 hubs (hub_product and hub_product_details) because the bronze_product_details is related to the product (which is the core business entity) using a foreign key from the bronze_product table.

 doi:10.20944/preprints202502.2012.v1

Please provide the DDL for the input model.

- Option 4 - Dynamic prompting

  Prompt 1:

  You are a Data Vault modeling expert. You will need to step-by-step create a Data Vault 2.0 model by using the bronze tables DDL as input.
  The output should be a new set of DDL.
  This is the first step of the iterative process. Please provide the DDL of hub tables, which are the core business entities.
  Be careful to only return the business entities, some tables might not represent business entities, they can be satellites or link or reference tables.

  Prompt 2:

  That is great, thank you for your response!
  Please return the DDL for the satellite tables, which contain all the attributes from the bronze tables (one bronze table will be the source for one satellite).
  Be aware that there can be multiple satellites associated with the same hub entity and satellites can be associated with either hubs, links or reference tables.

  Prompt 3:

  The outputs provided earlier for hubs & satellites are amazing!
  Let's go to the final step of the process – link tables. Based on the hub tables generated earlier, please return the DDL for the link tables associated.
  You do not need to provide links between other tables that are not the hubs or suggest other potential new hub and satellite tables in this step of the process.

- Option 5 - Instruction-based prompting

  You are a Data Vault modeling expert.
  Task: Generate a comprehensive Data Vault schema based on the provided TPC-DS data set DDL statements.
  Instructions:
  1. Identify Hubs:
  - They represent core business entities (e.g. customers, products)
  - Each Hub only includes the business key, the generated hash key, effective_datetime when the record is loaded and record_source.
  2. Identify Satellites:
  - For each Hub, create one or more Satellite tables to store descriptive attributes related to that business entity.
  - There can be more satellites for the same hub and they can come from different tables (if they are related to the same business entity).
  - Each Satellite should include the attributes related to the business entity/hub.
  3. Identify Links:
  - For relationships between the entities created previously, create Link tables.
  - Each Link should include the hash key, foreign keys linked the related tables.

To test the scalability of the model, which is one of the pillars of Data Vault models, a new data source will be introduced. The new data source will contain demographic information about the customer house and will be stored in a different table that has a reference to the customer table. For each of the options presented above, after the response is returned, there will be an additional prompt as follows:

 doi:10.20944/preprints202502.2012.v1

“““

Adjust the model to include a new source – household_demographics. It has a FOREIGN key in the bronze_customer table between hd_demo_sk (from bronze_household_demographics) and c_current_hdemo_sk (from bronze_customer). Please create the required tables so that it follows the pattern of the existing model. Do not forget to respect Data Vault 2.0 principles!

CREATE TABLE bronze_household_demographics (
hd_demo_sk INT,
hd_income_band_sk INT,
hd_buy_potential STRING,
hd_dep_count INT,
hd_vehicle_count INT
);

Please summarize the changes that were made.
“““

3.4. Validation and Evaluation Criteria

The validation and evaluation of the ChatGPT-generated models are structured around five criteria that align with the research objectives by assessing both the benefits (accuracy, efficiency) and limitations (adaptability, stability) of using LLMs for complex data engineering tasks (accuracy, manual review). All criteria were considered to be equally important:

- Model accuracy – this criterion includes comments around adherence to Data Vault principles, checking that the surrogate keys are generated appropriately, and the relationships between entities are captured correctly. This directly addresses the research question regarding ChatGPT’s ability to create efficient and accurate models.
- Efficiency – this criterion includes discussions around the number of tables generated and the effort required to implement the respective solution. It provides insights into how ChatGPT can streamline the traditional labor-intensive process of Data Vault modeling.
- Manual review – this criterion provides a comparison with the manual-crafted model presented in the section above to ensure that automated outputs meet industry standards for accuracy and completeness.
- Adaptability – this criterion includes comments around model scalability to handle changes when a new source is integrated into existing models, a critical requirement for scalable and flexible data warehousing solutions.
- Stability – this criterion is being used to validate if the output provided by the LLM model is consistent and repeatable through multiple iterations, ensuring its reliability for real-world data engineering tasks.

This structured approach to validation and evaluation ensures that the Data Vault models generated by ChatGPT are not only technically accurate but also practically applicable, reliable, and efficient for use in data engineering tasks.

4. Results and Discussions

This section provides an analysis of the results obtained from multiple prompt engineering techniques to generate the output for the Data Vault model, exploring various prompt engineering techniques through LLMs (Large Language Models). Each prompt presented in the previous section was tested using the ChatGPT User Interface to evaluate its effectiveness in providing relevant and accurate responses from the model. All tasks were performed using the GPT-4o model, through

 doi:10.20944/preprints202502.2012.v1

multiple iterations in October 2024. The requests were send using different chat connections, so that they will be evaluated independently and the model will not have any previous context.

It was decided to use ChatGPT, as it is freely available, making it an accessible tool for experimentation without the need for expensive licenses or subscriptions. Its user-friendly interface also played a key role, allowing easy interaction and seamless testing without requiring extensive technical expertise. Additionally, it is hosted by Open AI, which offers an API that can be further integrated into a larger framework for automating data modeling tasks. The combination of cost-effectiveness, ease of use, and potential for integration made Chat GPT an ideal choice for exploring the potential of AI-driven data modeling in the context of Data Vault.

The table below, Table 1, shows all the generated prompts, where the option identifier can be referenced in a section above. It can be seen that the number of tables returned by each prompt varies between 11 and 17 tables.

Table 1. Results of multiple prompt engineering model generations.

Option Identifier	Number of tables	Hub tables	Satellite tables	Link tables
1	17	hub_customer hub_customer_address hub_customer_demographics hub_store hub_promotion hub_item	sat_customer sat_customer_address sat_customer_demographics sat_store sat_promotion sat_item sat_sales	lnk_customer_address lnk_customer_demographics lnk_sales_data lnk_promotion_item
2	16	hub_customer hub_customer_address hub_customer_demographics hub_store hub_promotion hub_item	sat_customer sat_customer_address sat_customer_demographics sat_store sat_promotion sat_item sat_sales	lnk_customer_address lnk_customer_demographics lnk_sales_data
3	11	hub_customer hub_store hub_promotion hub_item	sat_customer sat_customer_address sat_customer_demographics sat_store sat_promotion sat_item	lnk_sales_data
4	13	hub_customer hub_store hub_promotion hub_item	sat_customer sat_customer_address sat_customer_demographics sat_store sat_promotion sat_item sat_sales	lnk_sales_data lnk_promotion_item
5	15	hub_customer hub_customer_address hub_customer_demographics hub_store hub_promotion hub_item	sat_customer sat_customer_address sat_customer_demographics sat_store sat_promotion sat_item	lnk_customer_address lnk_customer_demographics lnk_sales_data

Note: The font size for this table has been adjusted for compactness.

To evaluate the outputs generated based on the specified criteria—model accuracy, efficiency, manual review comparison, adaptability, and stability—each option was rated on a scale of 1 to 5. A score of 5 indicates that all requirements were fully met, while a score of 1 indicates that none were satisfied. Scores varied depending on the irregularities observed for each criterion.

To compare the efficiency of the models, they were implemented, deployed, and tested within Databricks, a unified cloud-based platform for developing data and AI solutions [35]. A Databricks workflow managed the orchestration, running on a D8s_v3 cluster configured with four worker nodes.

As shown in Table 2, Option 4, which utilized dynamic or interactive prompts, performed best in all criteria. It produced the expected output in terms of model accuracy, efficiency, adaptability, and manual review. However, a notable drawback was that its output varied between iterations, raising concerns about the reliability of consistently reproducing results.

 doi:10.20944/preprints202502.2012.v1

Table 2. Model generation results by different criteria.

Option Identifier	Model accuracy	Efficiency	Manual Review	Adaptability	Stability	Total
1	5	3	3	5	5	21
2	4	3	3	5	5	20
3	4	5	4	3	5	21
4	5	5	5	5	2	22
5	4	3	2	5	5	19

Options 1 (without prompt engineering) and 3 (few-shot prompting) also achieved strong scores, outperforming Option 4 in stability. Option 1 received the highest score for model accuracy, correctly linking all tables created, but scored lower in efficiency and manual review due to generating unnecessary business entities for customer address and demographics. Option 3, while less accurate and thorough in manual review, failed to create a link table between promotions and items. It also struggled with adaptability, scoring 3 points, as it treated the new source as a separate business entity instead of integrating it as a satellite table connected to the customer entity.

Options 2 (contextual prompting) and 5 (instruction-based prompting) occupied the fourth and fifth positions, respectively. Option 2 slightly outperformed Option 5 in manual review, as sales attributes were appropriately stored in satellite tables instead of link tables. Both models demonstrated good adaptability, effectively integrating the new source into the existing design and maintaining consistent outputs across multiple iterations.

As shown in Table 3, storing the customer's address and demographics as separate business entities, rather than as satellites associated with the customer entity, required approximately 45% more storage space. This number is based on an initial run, though storage demands could increase rapidly as the model conserves historical data. Furthermore, in terms of execution time, Option 1 took approximately 50% longer to complete compared to Option 4 and the manually created model. In conclusion, the efficiency criterion was reduced to 3 points for entities that considered the customer address and demographics as business entities.

Table 3. Efficiency validation between the models with 6 business entities and Manual model.

	Number of tables	Running time	Total size of data (GB)
6 business entities	17	15m 12s	26.205
Manual model (4 business entities)	13	10m 17s	18.075

5. Conclusions

This study explored the potential of using AI and LLM to derive data models and speed up the design process of a data solution. The experiment aimed to assess the performance of the LLM to generate a Data Vault model through multiple prompt techniques. The key findings of the research indicate that GPT-4o demonstrates the ability to understand and generate components of the Data Vault model, such as hubs, satellites, and links. It can identify different data structures and relationships, responding to complex data model tasks. Although the model showed promise, it was also noted that there were areas where it required further refinement. Specifically, some specific requirements were not fully captured, and iterative prompting was needed to fine-tune the output.

By exploring the results presented in the previous section, it was observed that different prompt techniques such as contextual prompting, few-shot, instruction-based and dynamic prompting were used to generate the outputs. While none of these methods fully replicated the manual design or consistently persisted through multiple iterations, the techniques that performed the best and were also consistent were the few-shot prompt and the first option without any details (no prompting techniques). The difference between these two examples was that the first technique, a few-shot

 doi:10.20944/preprints202502.2012.v1

prompt, was able to use the example and generate the output that combines multiple tables into the same business entity. On the other hand, the option without prompting generated the model in accordance with Data Vault 2.0, but provided a slightly different model with multiple business entities. This version of the model was also implemented and tested, and it was observed that it did not perform as well as the initial manual one in terms of running time and storage, also requiring an additional effort for the development.

Chat GPT demonstrated that it can provide accurate responses with the help of the user through iterative prompting. However, it was not able to independently generate a complete model that fully matched the manual version without additional user guidance. Despite this, when adaptability and integrating an additional data source was tested for the model, the results were highly promising. Most of the prompts performed well, effectively integrating the new sources into the data model initially provided as output. The LLM represents a valid use case for automating repetitive tasks and providing rapid prototyping capabilities, where it significantly reduces the time required for data modeling and design iterations. Its ability to process large amounts of information and generate structured output enhances accuracy by minimizing human errors and inconsistencies. Furthermore, ChatGPT's adaptability allows for seamless integration of new data sources, making it easier to scale and update models in response to evolving business requirements. A key contribution would be its potential to reduce cost and reliance on specialized expertise. Having a model that is trained and adjusted to perform this activity would lower operational costs and enable a business user with limited knowledge to contribute effectively. The model can also detect potential inconsistencies and suggest corrections that help reduce errors early in the development life cycle, improving overall data quality and reliability.

The experiment of using LLMs and different prompt engineering techniques to build a data model following the principles of Data Vault 2.0 can have some limitations. One of the main limitations is the structure of the prompts. If the prompts are structured differently, the model can return different results. This variation emphasizes the challenge of creating an input text based on multiple prompts engineering techniques and defining a version that will consistently meet the needs. Another limitation would be the lack of domain expertise, as the model does not possess the industry-specific knowledge that a human expert would have. Consequently, the model might not fully understand complex business rules or specific data nuances, which could result in a more generic answer that can fit many business cases, but may lack the required depth for particular scenarios. Additionally, there is a potential for inconsistencies across different runs - while the experiment was conducted with a limited number of iterations, results could vary if repeated, due to the probabilistic nature of the model. The inconsistency of the results provided by the current experiment could also be addressed in an extended version of the experiment, which will act as a framework. Using the model API, the results can further be refined using different options, such as adjusting the model temperature to obtain more deterministic and consistent results, avoiding hallucinations. Other options that need to be considered would be the prompt technique used, being required to use a combination of multiple prompts that would generate the required results or ensuring that the output generated matches the expected quality or format. It could also be explored how various models perform this task to select the models that perform the best according to the specific needs. Lastly, the model's output may be influenced by biases inherited from the datasets it was trained on. Since GPT is trained on vast and diverse data, it can reflect inherent biases in the data, leading to skewed or suboptimal results in some cases.

The approach presented in the paper demonstrates the potential of AI-driven tools, such as GPT, in automating complex tasks such as Data Vault model generation. This research could be extended in practice by developing a framework and an automated process where the initial DDL scripts serve as input, and the tool generates the corresponding DDL of the Data Vault model as output, without requiring user intervention. The framework would utilize well-structured prompts, incorporating various prompt engineering techniques to achieve the desired results, with the possibility of further

 doi:10.20944/preprints202502.2012.v1

refinement through RAG (Retrieval-Augmented Generation) or fine-tuning. Moreover, the entire process will be designed using a user interface, where the user can place the initial data, wait for the new model to be generated, perform data quality tests (they are meant to check if the model generated does not contain duplicated tables or columns, respects the pattern in terms of column consistency, has all the mandatory technical columns and also have all the components required), and then it can also generate the source-to-target mapping document and a physical model, with the diagram of relationship between the tables. This would streamline the data modeling process, reducing the manual effort involved in generating the model and schema. It would provide a scalable solution that benefits data engineers and organizations by enabling them to build and maintain data warehouse solutions more efficiently. Moreover, the proposed solution could also contain a versioning control option, which will store the model that were previously generated, an option that will help in terms of reducing the time of generating new results and costs (it will avoid new API calls) and also to review the previously generated models and access them if needed. Another area of expanding this framework could also be with a database integration or with the option to choose, or let the LLM to decide which is the modeling technique that should be used, depending on the scenario. By implementing such a framework, professionals could leverage automation to simplify and accelerate the creation of Data Vault models, ultimately leading to more efficient data warehouse development and management.

This research can also be extended to explore AI in multiple stages of the life cycle of the data solution. This could involve utilizing AI not just for the initial data modeling phase, but also throughout the entire life cycle of a data solution, from data ingestion and transformation to integration and quality assurance. For example, AI could assist in optimizing the ETL (Extract, Transform, Load) processes by automatically generating transformation rules and detecting various data anomalies, ensuring better data quality and consistency. By pursuing these directions, future studies can further bridge the gap between AI capabilities and the evolving needs of data solutions in the digital era.

Author Contributions: Conceptualization, A.V., A.R.B.; methodology, A.V., A.I.B.; software: A.V.; validation: A.R.B.; writing—original draft preparation, A.V.; writing—review and editing, A.R.B., A.O.D.; supervision, A.R.B.. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was co-financed by The Bucharest University of Economic Studies during the PhD program.

Data Availability Statement: All data ws extracted from TPC website, using a free account and following the steps required to generate various data sets. (https://www.tpc.org/tpcds/default5.asp)

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Golightly, L.; Chang, V.; Xu, Q. A.; Gao, X.; Liu, B. S. C. (2022). Adoption of cloud computing as innovation in the organization. In *International Journal of Engineering Business Management, 14*. https://doi.org/10.1177/18479790221093992
2. Vines, A.; Tanasescu, L. (2023). An overview of ETL cloud services: An empirical study based on user's experience. In *Proceedings of the International Conference on Business Excellence, 17*(1), 2085–2098. https://doi.org/10.2478/picbe-2023-0182
3. Clissa, L.; Lassnig, M.; Rinaldi, L. (2023). How Big is Big Data? A comprehensive survey of data production, storage, and streaming in science and industry, In *Frontiers in Big Data*, 6:1271639.
4. Linstedt, D. (2016). Building a Scalable Data Warehouse with Data Vault 2.0, Morgan Kaufmann, 1st ed., Boston, MA, USA.
5. Vines, A.; Tănăsescu, L. (2024). Data Vault Modeling: Insights from Industry Interviews, In *Proceedings of the International Conference on Business Excellence*, vol. 18, no. 1, pp. 3597–3605, 2024.
6. El-Sappagh, S.; Hendawi, A.; El-Bastawissy, A. (2011). A proposed model for data warehouse ETL processes, *Journal of King Saud University - Computer and Information Sciences*, 23, 91–104.
7. Fan, L.; Lee, C.-H.; Su, H.; Feng, S.; Jiang, Z.; Sun, Z. (2024). A New Era in Human Factors Engineering: A Survey of the Applications and Prospects of Large Multimodal Models, In *arXiv*. https://arxiv.org/abs/2405.13426

 doi:10.20944/preprints202502.2012.v1

8. Ege, D. N.; Øvrebø; H. H.; Stubberud, V.; Berg, M. F.; Elverum, C.; Steinert, M.; Vestad, H. (2024). ChatGPT as an inventor: Eliciting the strengths and weaknesses of current large language models against humans in engineering design, In *arXiv*. https://arxiv.org/abs/2404.18479
9. Choi, S.; Gazeley, W. (2024). When Life Gives You LLMs, Make LLM-ADE: Large Language Models with Adaptive Data Engineering, In *arXiv preprint*, arXiv:2404.13028. https://arxiv.org/abs/2404.13028
10. Mantri, A. (2024). Intelligent Automation of ETL Processes for LLM Deployment: A Comparative Study of Dataverse and TPOT, *European Journal of Advances in Engineering and Technology*, 11(4), 154–158.
11. Transaction Processing Performance Council. (2013). TPC-DS Benchmark Specification.
12. Yessad, L.; Labiod, A. (2016). Comparative study of data warehouses modeling approaches: In-mon, Kimball, and Data Vault, In *2016 International Conference on System Reliability and Science (ICSRS)*, Paris, France, pp. 95–99.
13. Naamane, Z.; Jovanovic, V. (2016). Effectiveness of Data Vault compared to Dimensional Data Marts on Overall Performance of a Data Warehouse System, In *International Journal of Computer Science Issues*, 13(1).
14. Vines, A. (2024). Performance Evaluation of Data Vault and Dimensional Modeling: Insights from TPC-DS data set Analysis, In *Proceedings of 23rd International Conference on Informatics in Economy (IE 2024)*.
15. Helskyaho, H.; Ruotsalainen, L.; Männistö, T. (2024). Defining Data Model Quality Metrics for Data Vault 2.0 Model Evaluation, *Inventions*, 9, 21.
16. Inmon, W. H.; Zachman, J. A.; Geiger, J. G. (2008). *Data stores, data warehousing, and the Zachman framework: Managing enterprise knowledge*, McGraw-Hill.
17. Kimball, R.; Ross, M.; Thornthwaite, W.; Mundy, J.; Becker, B. (2008). Data Warehouse Lifecycle Toolkit: Practical Techniques for Building Data Warehouse and Business Intelligence Systems (2nd ed.), *New York: Wiley*.
18. Smith, J.; Elshnoudy, I. A. (2003). *A Comparative Analysis of Data Warehouse Design Methodologies for Enterprise Big Data and Analytics*, Emerging Trends in Machine Intelligence and Big Data, 16–29.
19. Giebler, C.; Gröger, C.; Hoos, E.; Schwarz, H.; Mitschang, B. (2019). Modeling Data Lakes with Data Vault: Practical Experiences, Assessment, and Lessons Learned, In *Proceedings of the 38th Conference on Conceptual Modeling (ER 2019)*, vol. 38, no. 1, pp. 1–10, 2019.
20. Vines, A.; Samoila, A. (2023). An Overview of Data Vault Methodology and Its Benefits, *Informatica Economica*, 27(2), 11–20.
21. Helskyaho, H. (2023). Towards Automating Database Designing, In *Proceedings of the 34th Conference of Open Innovations Association (FRUCT)*.
22. Ggaliwango, M.; Nakayiza, H. R.; Jjingo, D.; Nakatumba-Nabende, J. (2024). Prompt Engineering in Large Language Models.
23. Lo, L. S. (2023). The CLEAR Path: A Framework for Enhancing Information Literacy Through Prompt Engineering. *The Journal of Academic Librarianship*, 49(4), 102720.
24. Ahmed, T.; Pai, K.; Devanbu, P.; Barr, E. (2023). Improving Few-Shot Prompts with Relevant Static Analysis Products, In *Proceedings of the 2023 International Conference on Software Engineering (ICSE)*, 2023.
25. Khattak, M. U.; Rasheed, H., Maaz, M.; Khan, S.; Khan, F. S. (2023). MaPLe: Multi-modal Prompt Learning, In *P*roceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19113–19122. IEEE.
26. Wang, B.; Deng, X.; Sun, H. (2022). Iteratively Prompt Pre-trained Language Models for Chain of Thought, *arXiv preprint*, arXiv:2203.08383. https://arxiv.org/abs/2203.08383.
27. Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. (2023). Large Language Models are Zero-Shot Reasoners, In *arXiv preprint*, arXiv:2205.11916.
28. Alhindi, T.; Chakrabarty, T.; Musi, E.; Muresan, S. (2023). Multitask Instruction-based Prompting for Fallacy Recognition, *arXiv preprint*, arXiv:2301.09992
29. Diao, S.; Wang, P.; Lin, Y.; Pan, R.; Liu, X.; Zhang, T. (2024). Active Prompting with Chain-of-Thought for Large Language Models, In *arXiv preprint*, arXiv:2302.12246.
30. Hegde, C. (2022). Anomaly Detection in Time Series Data using Data-Centric AI, In *Proceedings of the 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1–6.
31. Chai, C.; Tang, N.; Fan, J.; Luo, Y. (2023). Demystifying Artificial Intelligence for Data Preparation, In *Proceedings of the Companion of the 2023 International Conference on Management of Data (SIGMOD '23)*, pp. 13–20.
32. Zhao, H.; Ye, X. (2013). A Practice of TPC-DS Multidimensional Implementation on NoSQL Database Systems, In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 93–108.

 doi:10.20944/preprints202502.2012.v1

33. Al-Kateb, M.; Crolotte, A.; Ghazal, A.; Rose, L. (2013). Adding a Temporal Dimension to the TPC-H Benchmark, In: Nambiar, R., Poess, M. (eds) Selected Topics in Performance Evaluation and Benchmarking. TPCTC 2012. Lecture Notes in Computer Science, vol 7755, Springer, Berlin, Heidelberg.
34. Chen, G.; Johnson, T.; Cilimdzic, M. (2022). Quantifying Cloud Data Analytic Platform Scalability with Extended TPC-DS Benchmark. In: Nambiar, R., Poess, M. (eds) Performance Evaluation and Benchmarking. TPCTC 2021. Lecture Notes in Computer Science(), vol 13169. Springer, Cham.
35. Databricks Documentation (n.d.). https://docs.databricks.com/en/introduction/index.html.