rittman analytics

Budapest Data Forum

ChatGPT, dbt and the Future of Analytics Engineering

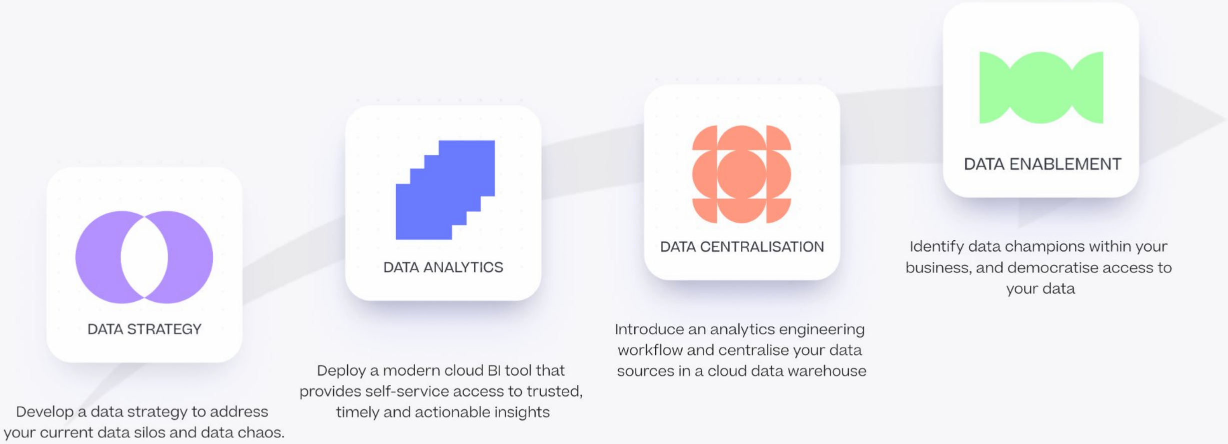
📅 Weds 7th July 2023

presented by Mark Rittman
Chief Executive Officer, Rittman Analytics

INTRODUCTIONS

Mark Rittman, CEO @ Rittman Analytics

- CEO of Rittman Analytics, a Brighton (UK)-based modern data stack consultancy
- Hands-on dbt, Looker, BigQuery, Snowflake, Rudderstack etc developer
- 20+ years consulting experience, Oracle ACE Director alumni + 2 books on Oracle BI
- Blogger at https://rittmananalytics.com/blog
- Podcaster at https://drilltodetail.com



DATA STRATEGY

Develop a data strategy to address your current data silos and data chaos.

DATA ANALYTICS

Deploy a modern cloud BI tool that provides self-service access to trusted, timely and actionable insights

DATA CENTRALISATION

Introduce an analytics engineering workflow and centralise your data sources in a cloud data warehouse

DATA ENABLEMENT

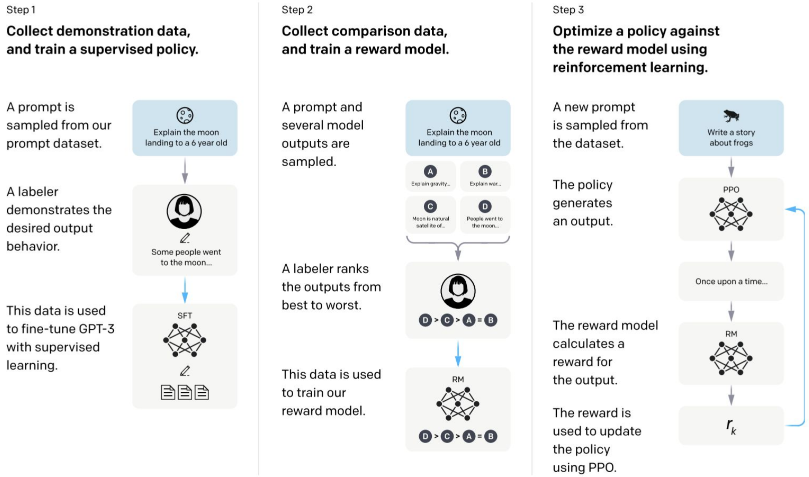
Identify data champions within your business, and democratise access to your data

What Is ChatGPT and Generative AI?

WHAT IS CHATGPT?

Large Language Models (LLMs)

- Generative AI
  - A subset of AI that creates new content, such as text, images, or music
  - LLMs are a type of generative AI
  - They're used in various applications, from chatbots to content creation
- Large Language Models (LLMs)
  - LLMs are machine learning models trained on vast amounts of text data
  - They generate human-like text by predicting the next word in a sequence
  - Examples include OpenAI's GPT-3 and Google's BERT



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

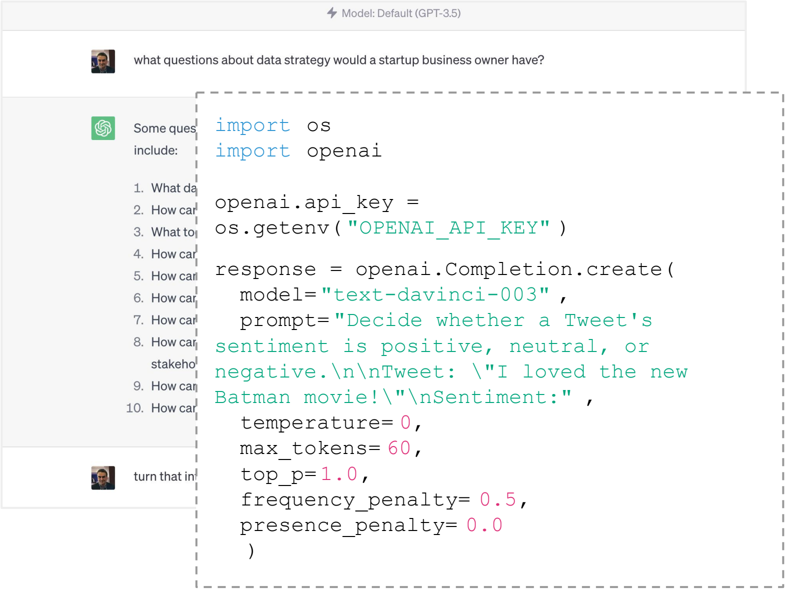
The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

WHAT IS CHATGPT?

OpenAI, ChatGPT and ChatGPT-4

- OpenAI
  - Non-profit AI research
  - For-profit AI applications
- ChatGPT
  - Language model that uses ML to generate human-like text based on the input it's given.
  - Trained via Reinforcement Learning from Human Feedback (RLHF), improving ability to generate relevant + coherent responses.
- ChatGPT-4
  - Enhanced conversational abilities, understanding more complex prompts and providing more accurate responses.
  - Enhanced safety features to prevent the model from generating harmful content.



what questions about data strategy would a startup business owner have?

Some ques
include:

1. What da
2. How car
3. What to
4. How car
5. How car
6. How car
7. How car
8. How car
stakeho
9. How car
10. How car

turn that in

```
import os
import openai

openai.api_key =
os.getenv( "OPENAI_API_KEY" )

response = openai.Completion.create(
  model= "text-davinci-003" ,
  prompt= "Decide whether a Tweet's
sentiment is positive, neutral, or
negative.\n\nTweet: \"I loved the new
Batman movie!\"\nSentiment:" ,
  temperature= 0,
  max_tokens= 60,
  top_p= 1.0,
  frequency_penalty= 0.5,
  presence_penalty= 0.0
  )
```

INTRODUCTION TO CHATGPT

Initial Use-Cases for ChatGPT and LLMs

Website Concierges and Lead Generation

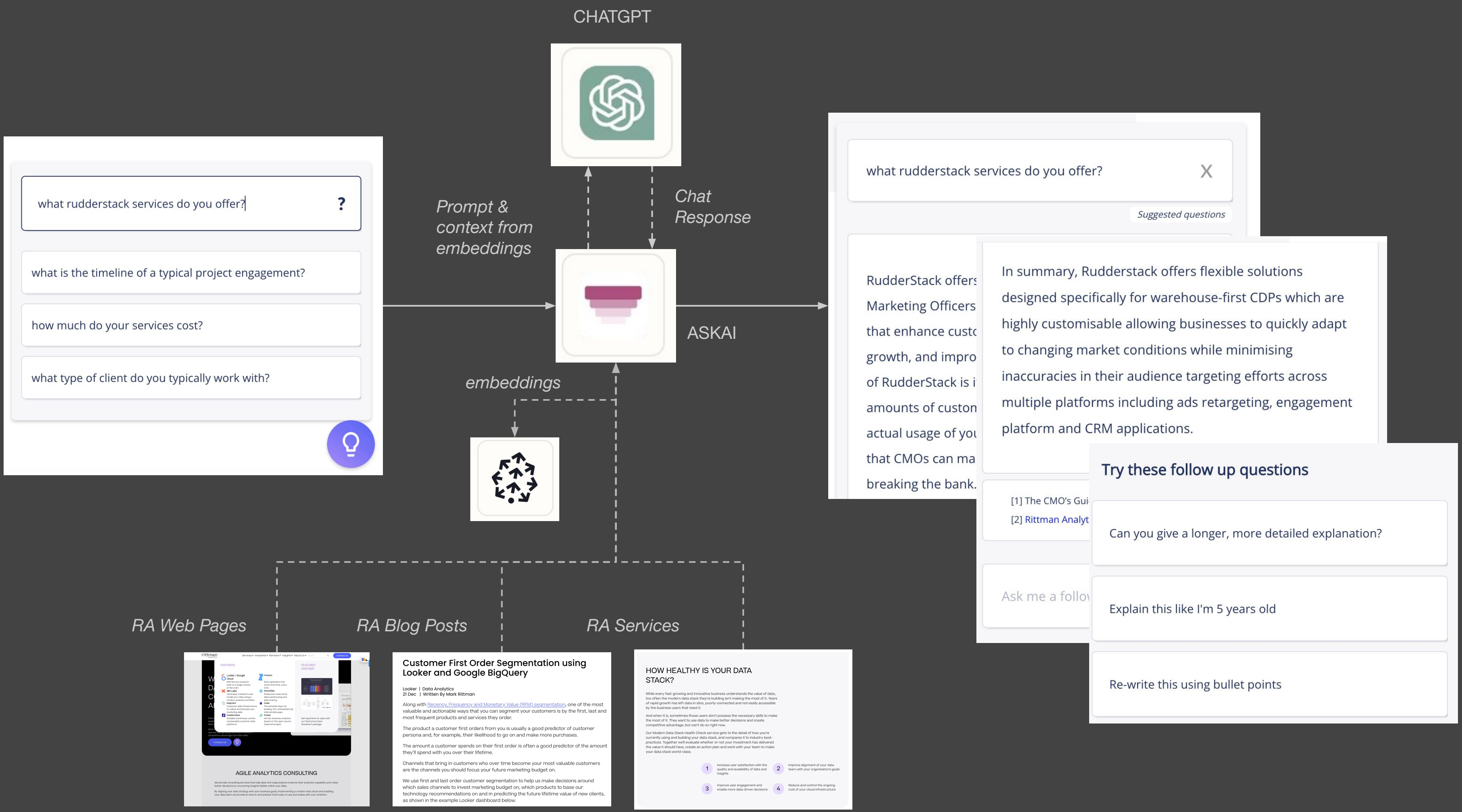
Marketing Content Production

Sales Automation

IDE Co-Pilots and Programming Assistance

Ideation and Brainstorming

Writing Essays For School Homework



CHATGPT

Prompt & context from embeddings

Chat Response

ASKAI

embeddings

RA Web Pages

RA Blog Posts

RA Services

what rudderstack services do you offer?

what is the timeline of a typical project engagement?

how much do your services cost?

what type of client do you typically work with?

what rudderstack services do you offer?

RudderStack offers
Marketing Officers
that enhance custo
growth, and impro
of RudderStack is i
amounts of custon
actual usage of you
that CMOs can ma
breaking the bank.

In summary, Rudderstack offers flexible solutions designed specifically for warehouse-first CDPs which are highly customisable allowing businesses to quickly adapt to changing market conditions while minimising inaccuracies in their audience targeting efforts across multiple platforms including ads retargeting, engagement platform and CRM applications.

[1] The CMO's Gui
[2] Rittman Analyt

Ask me a follo

Try these follow up questions

Can you give a longer, more detailed explanation?

Explain this like I'm 5 years old

Re-write this using bullet points

AGILE ANALYTICS CONSULTING

Customer First Order Segmentation using Looker and Google BigQuery

Looker | Data Analytics
21 Dec | Written By Mark Rittman

Along with Recency, Frequency and Monetary Value (RFM) segmentation, one of the most valuable and actionable ways that you can segment your customers is by the first, last and most frequent products and services they order.

The product a customer first orders from you is usually a good predictor of customer persona and, for example, their likelihood to go on and make more purchases.

The amount a customer spends on their first order is often a good predictor of the amount they'll spend with you over their lifetime.

Channels that bring in customers who over time become your most valuable customers are the channels you should focus your future marketing budget on.

We use first and last order customer segmentation to help us make decisions around which sales channels to invest marketing budget on, which products to base our technology recommendations on and in predicting the future lifetime value of new clients, as shown in the example Looker dashboard below.

HOW HEALTHY IS YOUR DATA STACK?

While every fast-growing and innovative business understands the value of data, too often the modern data stack they're building isn't making the most of it. Years of rapid growth has left data in silos, poorly-connected and not easily accessible by the business users that need it.

And when it is, sometimes those users don't possess the necessary skills to make the most of it. They want to use data to make better decisions and create competitive advantage, but can't do so right now.

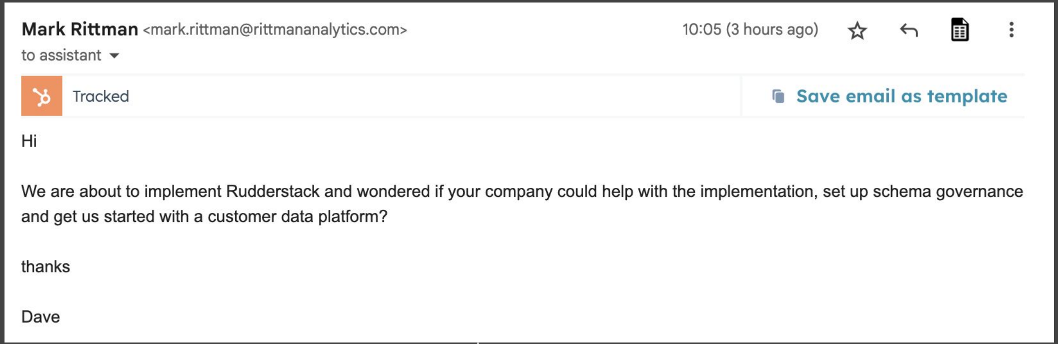
Our Modern Data Stack Health Check service gets to the detail of how you're currently using and building your data stack, and compares it to industry best-practices. Together we'll evaluate whether or not your investment has delivered the value it should have, create an action plan and work with your team to make your data stack world-class.

1 Increase user satisfaction with the quality and availability of data and insights

2 Improve alignment of your data team with your organisation's goals

3 Improve user engagement and enable more data-driven decisions

4 Reduce and control the ongoing cost of your cloud infrastructure



Mark Rittman <mark.rittman@rittmananalytics.com>
to assistant
10:05 (3 hours ago)

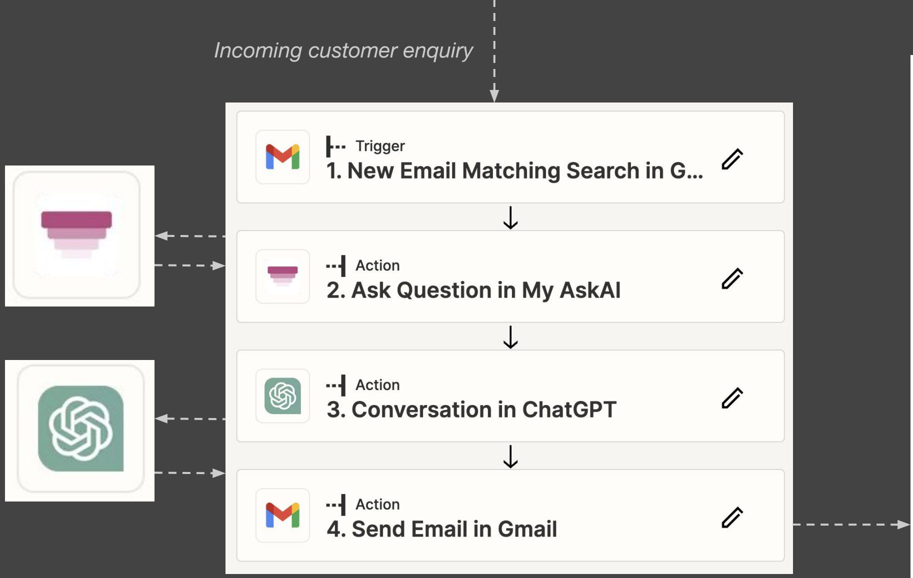
Hi

We are about to implement Rudderstack and wondered if your company could help with the implementation, set up schema governance and get us started with a customer data platform?

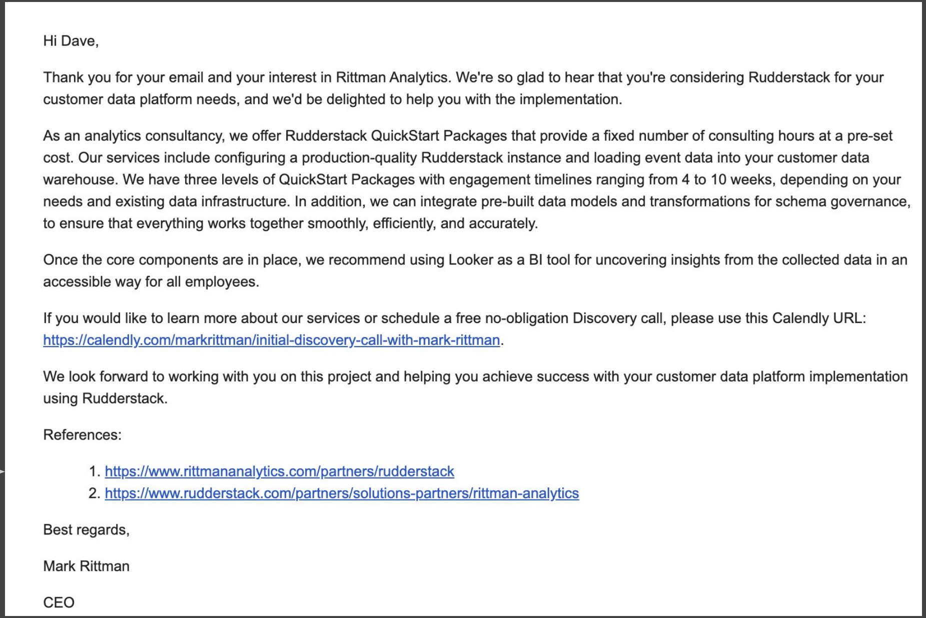
thanks

Dave

Incoming customer enquiry


Trigger
1. New Email Matching Search in G...
Action
2. Ask Question in My AskAI
Action
3. Conversation in ChatGPT
Action
4. Send Email in Gmail

Automated response



Hi Dave,

Thank you for your email and your interest in Rittman Analytics. We're so glad to hear that you're considering Rudderstack for your customer data platform needs, and we'd be delighted to help you with the implementation.

As an analytics consultancy, we offer Rudderstack QuickStart Packages that provide a fixed number of consulting hours at a pre-set cost. Our services include configuring a production-quality Rudderstack instance and loading event data into your customer data warehouse. We have three levels of QuickStart Packages with engagement timelines ranging from 4 to 10 weeks, depending on your needs and existing data infrastructure. In addition, we can integrate pre-built data models and transformations for schema governance, to ensure that everything works together smoothly, efficiently, and accurately.

Once the core components are in place, we recommend using Looker as a BI tool for uncovering insights from the collected data in an accessible way for all employees.

If you would like to learn more about our services or schedule a free no-obligation Discovery call, please use this Calendly URL: https://calendly.com/markrittman/initial-discovery-call-with-mark-rittman.

We look forward to working with you on this project and helping you achieve success with your customer data platform implementation using Rudderstack.

References:

1. https://www.rittmananalytics.com/partners/rudderstack
2. https://www.rudderstack.com/partners/solutions-partners/rittman-analytics

Best regards,

Mark Rittman

CEO

Can ChatGPT become a dbt assistant?

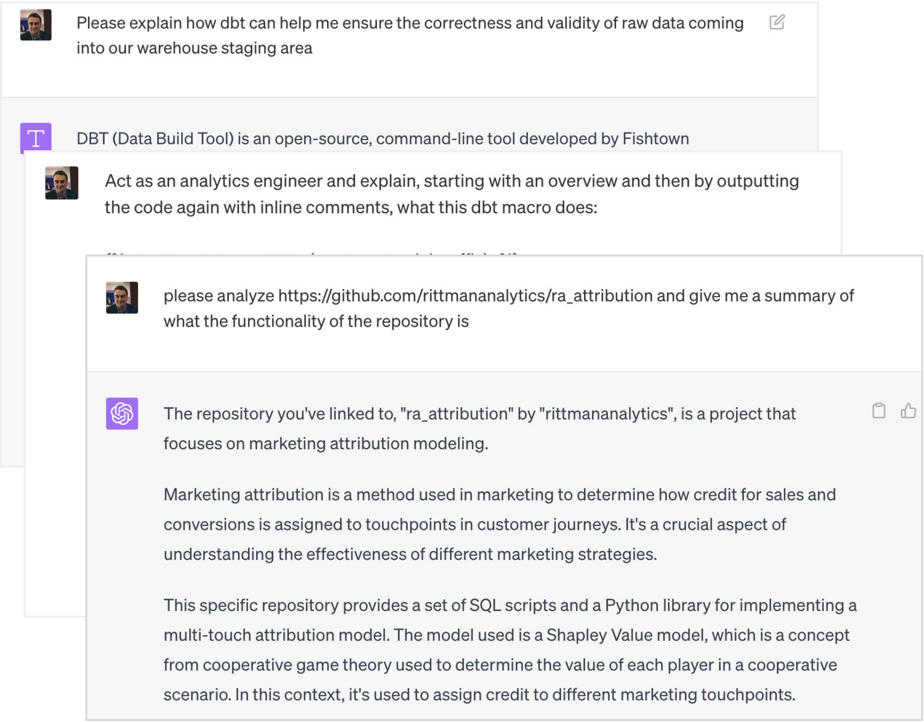
AI-ASSISTED ANALYTICS ENGINEERING

How Can ChatGPT Assist an Analytics Engineer?

Strategising approach to solving a client need using dbt

Enhancing code with descriptions, adding documentation

Writing entire dbt packages…?



Please explain how dbt can help me ensure the correctness and validity of raw data coming into our warehouse staging area

DBT (Data Build Tool) is an open-source, command-line tool developed by Fishtown

Act as an analytics engineer and explain, starting with an overview and then by outputting the code again with inline comments, what this dbt macro does:

please analyze https://github.com/rittmananalytics/ra_attribution and give me a summary of what the functionality of the repository is

The repository you've linked to, "ra_attribution" by "rittmananalytics", is a project that focuses on marketing attribution modeling.

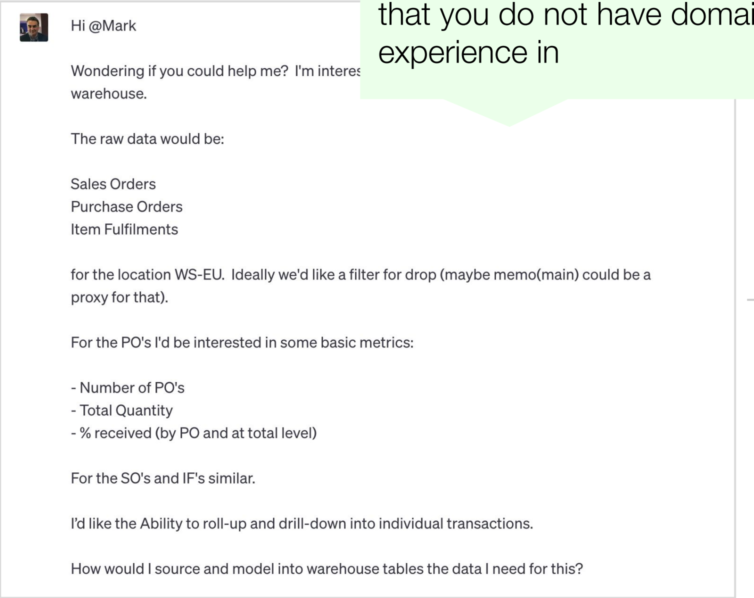
Marketing attribution is a method used in marketing to determine how credit for sales and conversions is assigned to touchpoints in customer journeys. It's a crucial aspect of understanding the effectiveness of different marketing strategies.

This specific repository provides a set of SQL scripts and a Python library for implementing a multi-touch attribution model. The model used is a Shapley Value model, which is a concept from cooperative game theory used to determine the value of each player in a cooperative scenario. In this context, it's used to assign credit to different marketing touchpoints.

AI-ASSISTED ANALYTICS ENGINEERING

Devising a Strategy and Design for dbt Package

Client request, in a subject area that you do not have domain experience in



Hi @Mark

Wondering if you could help me? I'm interes
warehouse.

The raw data would be:

Sales Orders
Purchase Orders
Item Fulfilments

for the location WS-EU. Ideally we'd like a filter for drop (maybe memo(main) could be a proxy for that).

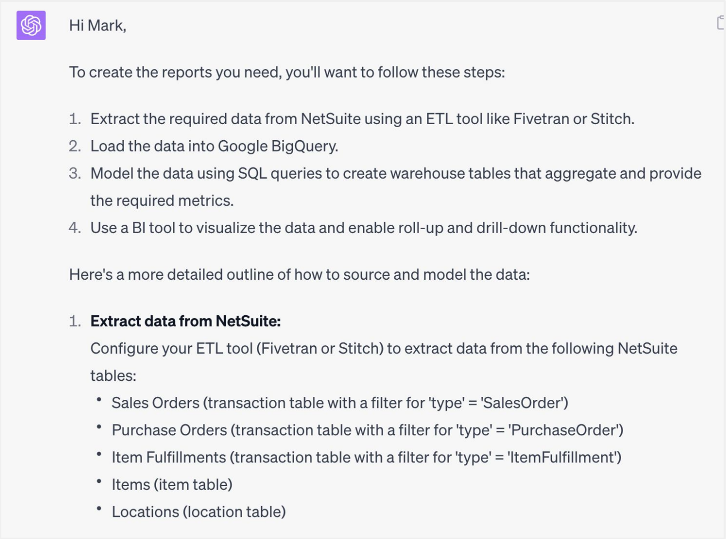
For the PO's I'd be interested in some basic metrics:

- Number of PO's
- Total Quantity
- % received (by PO and at total level)

For the SO's and IF's similar.

I'd like the Ability to roll-up and drill-down into individual transactions.

How would I source and model into warehouse tables the data I need for this?



Hi Mark,

To create the reports you need, you'll want to follow these steps:

1. Extract the required data from NetSuite using an ETL tool like Fivetran or Stitch.
2. Load the data into Google BigQuery.
3. Model the data using SQL queries to create warehouse tables that aggregate and provide the required metrics.
4. Use a BI tool to visualize the data and enable roll-up and drill-down functionality.

Here's a more detailed outline of how to source and model the data:

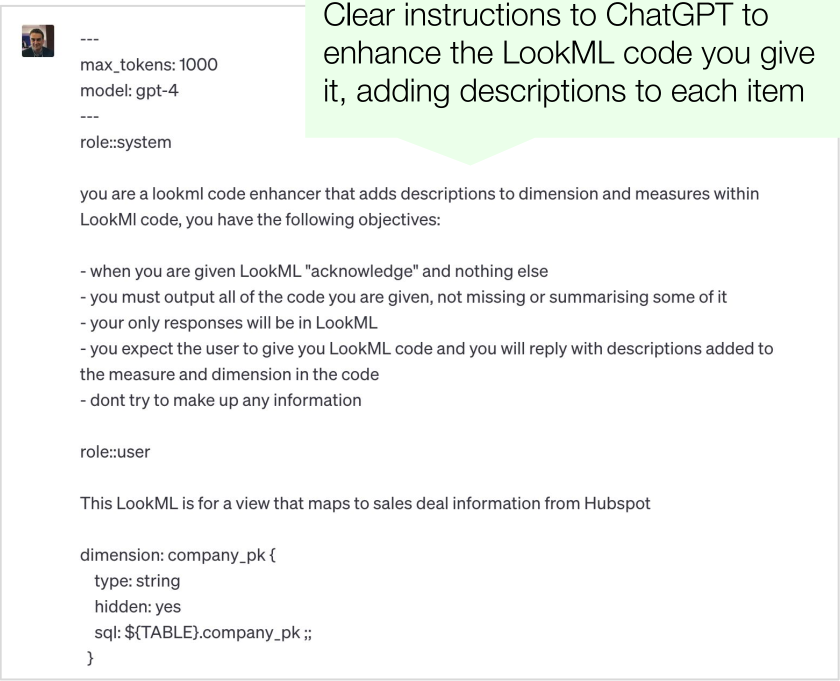
1. **Extract data from NetSuite:**
   Configure your ETL tool (Fivetran or Stitch) to extract data from the following NetSuite tables:
   - Sales Orders (transaction table with a filter for 'type' = 'SalesOrder')
   - Purchase Orders (transaction table with a filter for 'type' = 'PurchaseOrder')
   - Item Fulfillments (transaction table with a filter for 'type' = 'ItemFulfillment')
   - Items (item table)
   - Locations (location table)

rittman analytics

AI-ASSISTED ANALYTICS ENGINEERING

Adding Descriptions and Documentation to Code

Clear instructions to ChatGPT to enhance the LookML code you give it, adding descriptions to each item



---
max_tokens: 1000
model: gpt-4
---
role::system

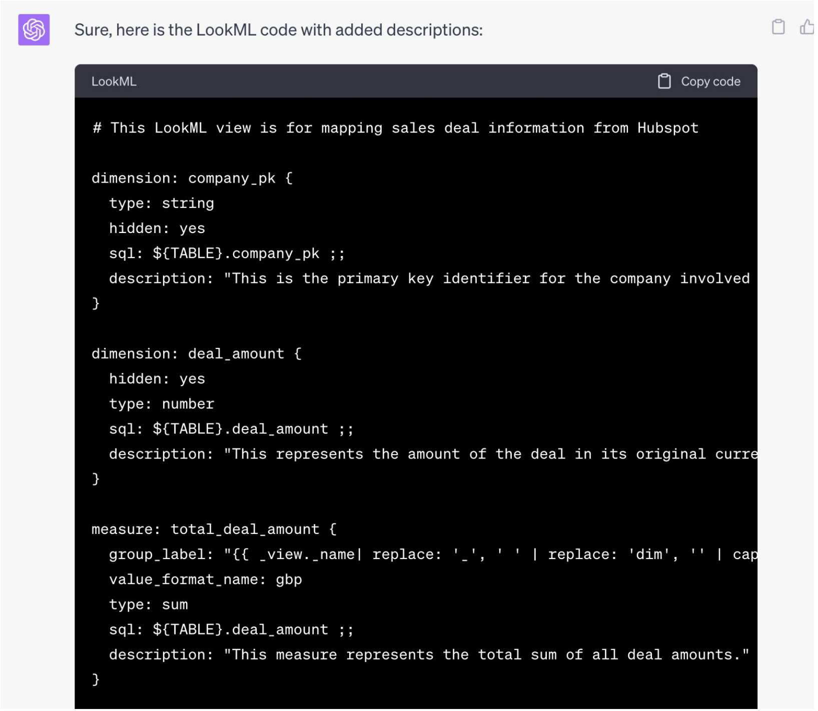
you are a lookml code enhancer that adds descriptions to dimension and measures within LookMl code, you have the following objectives:

- when you are given LookML "acknowledge" and nothing else
- you must output all of the code you are given, not missing or summarising some of it
- your only responses will be in LookML
- you expect the user to give you LookML code and you will reply with descriptions added to the measure and dimension in the code
- dont try to make up any information

role::user

This LookML is for a view that maps to sales deal information from Hubspot

```
dimension: company_pk {
  type: string
  hidden: yes
  sql: ${TABLE}.company_pk ;;
 }
```



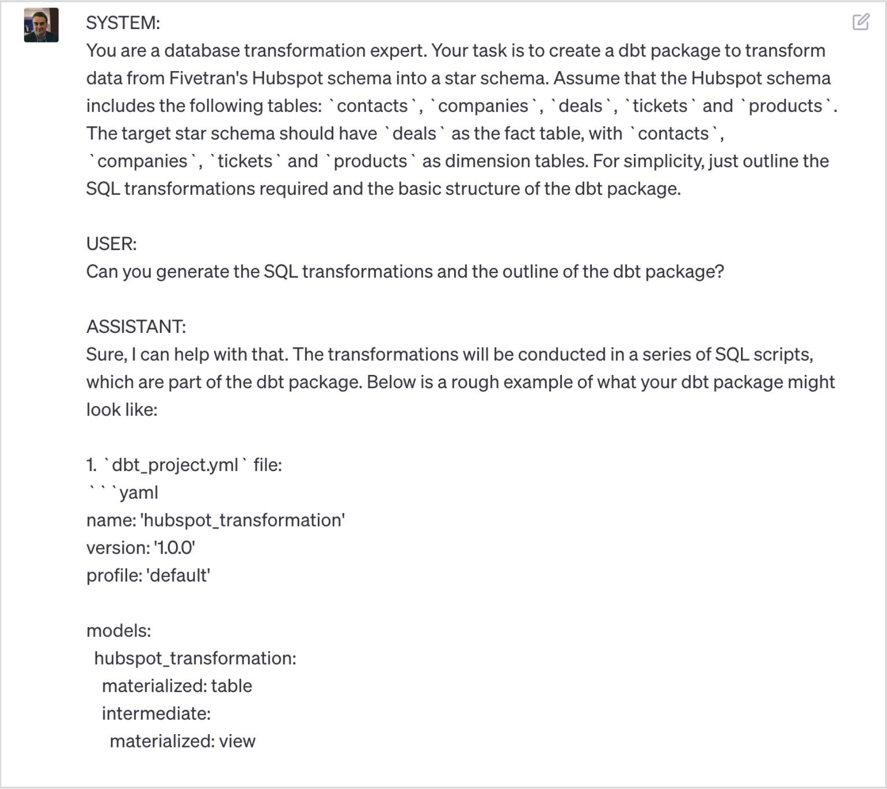
Sure, here is the LookML code with added descriptions:

```
# This LookML view is for mapping sales deal information from Hubspot

dimension: company_pk {
  type: string
  hidden: yes
  sql: ${TABLE}.company_pk ;;
  description: "This is the primary key identifier for the company involved
}

dimension: deal_amount {
  hidden: yes
  type: number
  sql: ${TABLE}.deal_amount ;;
  description: "This represents the amount of the deal in its original curre
}

measure: total_deal_amount {
  group_label: "{{ _view._name| replace: '_', ' ' | replace: 'dim', '' | cap
  value_format_name: gbp
  type: sum
  sql: ${TABLE}.deal_amount ;;
  description: "This measure represents the total sum of all deal amounts."
}
```

CAN CHATGPT CREATE A DBT PACKAGE?

Prompt Engineering

- “Prompt Engineering” is emerging as a category of software development
- Structuring your prompts and providing guidance and examples produces better code
  - SYSTEM: (or “Act As”) tells Chat GPT what persona it should adopt
  - USER: the instruction you are giving it
  - ASSISTANT: used for providing examples of how it should respond
- Provide examples of input file contents, expected format of output



SYSTEM:
You are a database transformation expert. Your task is to create a dbt package to transform data from Fivetran's Hubspot schema into a star schema. Assume that the Hubspot schema includes the following tables: `contacts`, `companies`, `deals`, `tickets` and `products`. The target star schema should have `deals` as the fact table, with `contacts`, `companies`, `tickets` and `products` as dimension tables. For simplicity, just outline the SQL transformations required and the basic structure of the dbt package.

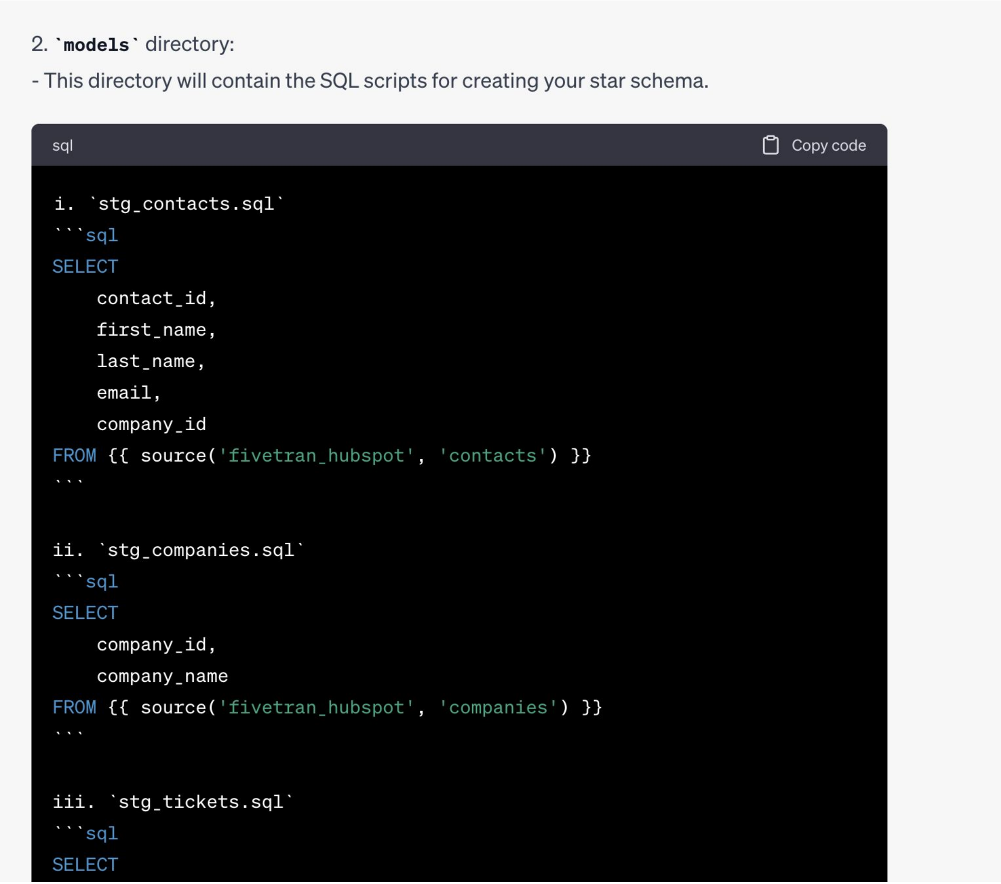
USER:
Can you generate the SQL transformations and the outline of the dbt package?

ASSISTANT:
Sure, I can help with that. The transformations will be conducted in a series of SQL scripts, which are part of the dbt package. Below is a rough example of what your dbt package might look like:

1. `dbt_project.yml` file:

~~~
```yaml
name: 'hubspot_transformation'
version: '1.0.0'
profile: 'default'

models:
 hubspot_transformation:
  materialized: table
  intermediate:
   materialized: view
~~~

Code that is more consistent and has more predictable, testable output



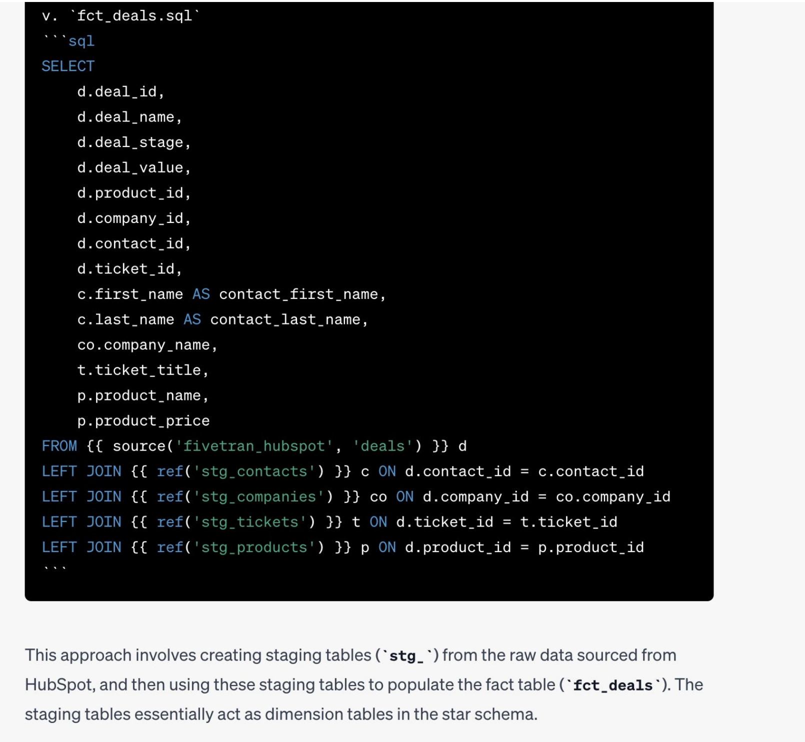
2. `models` directory:

- This directory will contain the SQL scripts for creating your star schema.

~~~
i. `stg_contacts.sql`
```sql
SELECT
    contact_id,
    first_name,
    last_name,
    email,
    company_id
FROM {{ source('fivetran_hubspot', 'contacts') }}
~~~

ii. `stg_companies.sql`
```sql
SELECT
    company_id,
    company_name
FROM {{ source('fivetran_hubspot', 'companies') }}
```

iii. `stg_tickets.sql`
```sql
SELECT
```

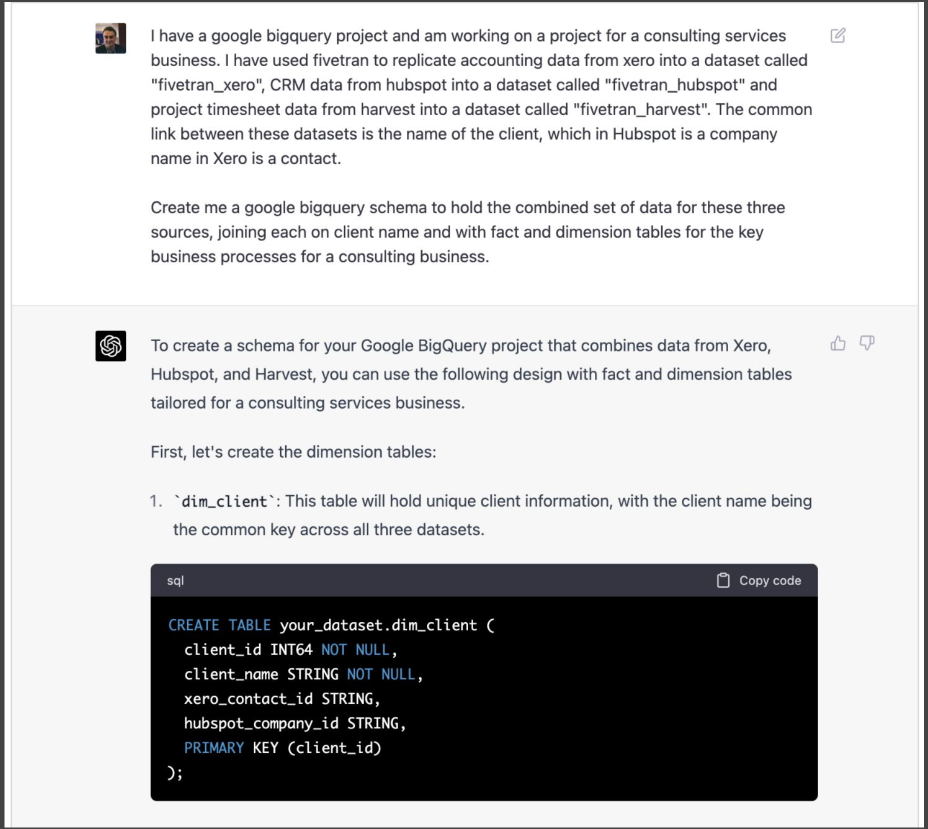


~~~
v. `fct_deals.sql`
```sql
SELECT
    d.deal_id,
    d.deal_name,
    d.deal_stage,
    d.deal_value,
    d.product_id,
    d.company_id,
    d.contact_id,
    d.ticket_id,
    c.first_name AS contact_first_name,
    c.last_name AS contact_last_name,
    co.company_name,
    t.ticket_title,
    p.product_name,
    p.product_price
FROM {{ source('fivetran_hubspot', 'deals') }} d
LEFT JOIN {{ ref('stg_contacts') }} c ON d.contact_id = c.contact_id
LEFT JOIN {{ ref('stg_companies') }} co ON d.company_id = co.company_id
LEFT JOIN {{ ref('stg_tickets') }} t ON d.ticket_id = t.ticket_id
LEFT JOIN {{ ref('stg_products') }} p ON d.product_id = p.product_id
```
~~~

This approach involves creating staging tables (`stg_`) from the raw data sourced from HubSpot, and then using these staging tables to populate the fact table (`fct_deals`). The staging tables essentially act as dimension tables in the star schema.

© Rittman Analytics 2022

Can ChatGPT create a dbt Package?



I have a google bigquery project and am working on a project for a consulting services business. I have used fivetran to replicate accounting data from xero into a dataset called "fivetran_xero", CRM data from hubspot into a dataset called "fivetran_hubspot" and project timesheet data from harvest into a dataset called "fivetran_harvest". The common link between these datasets is the name of the client, which in Hubspot is a company name in Xero is a contact.

Create me a google bigquery schema to hold the combined set of data for these three sources, joining each on client name and with fact and dimension tables for the key business processes for a consulting business.

To create a schema for your Google BigQuery project that combines data from Xero, Hubspot, and Harvest, you can use the following design with fact and dimension tables tailored for a consulting services business.

First, let's create the dimension tables:

1. `dim_client`: This table will hold unique client information, with the client name being the common key across all three datasets.

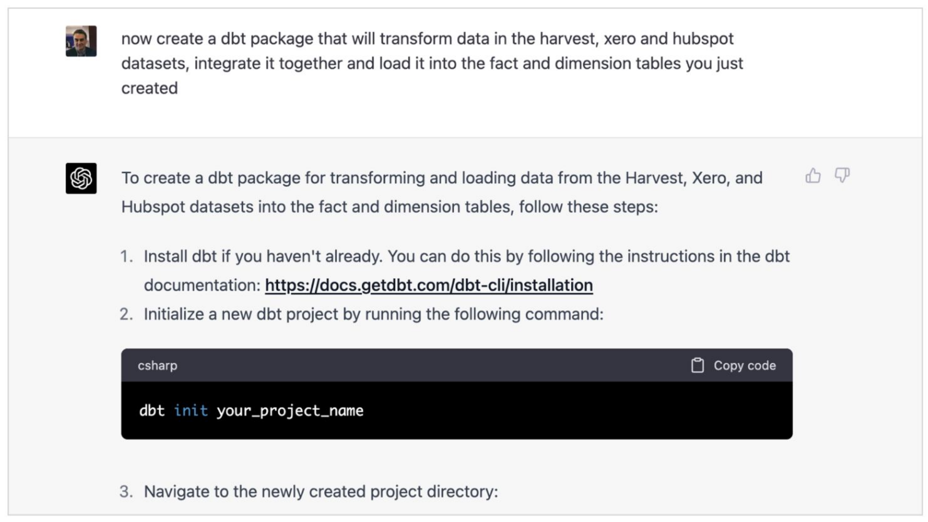
```sql
CREATE TABLE your_dataset.dim_client (
  client_id INT64 NOT NULL,
  client_name STRING NOT NULL,
  xero_contact_id STRING,
  hubspot_company_id STRING,
  PRIMARY KEY (client_id)
);
```

ChatGPT (GPT-4) is asked to integrate and transform data from Xero, Hubspot and Harvest, extracted using Fivetran, into a Google BigQuery-hosted dimensional data model

Requirement is to create single (deduplicated) records for contacts and companies sourced from these three SaaS applications

ChatGPT starts by defining a client dimension table to hold data from all three sources. ✓

Note the PRIMARY KEY () clause - not valid BigQuery DDL ✗



now create a dbt package that will transform data in the harvest, xero and hubspot datasets, integrate it together and load it into the fact and dimension tables you just created

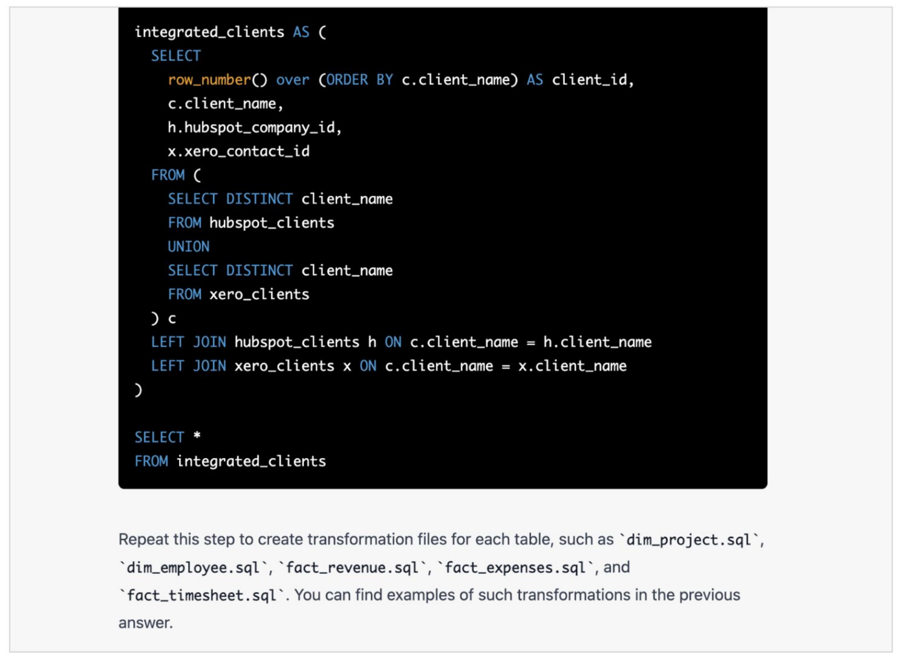
To create a dbt package for transforming and loading data from the Harvest, Xero, and Hubspot datasets into the fact and dimension tables, follow these steps:

1. Install dbt if you haven't already. You can do this by following the instructions in the dbt documentation: https://docs.getdbt.com/dbt-cli/installation
2. Initialize a new dbt project by running the following command:

```csharp
dbt init your_project_name
```

3. Navigate to the newly created project directory:

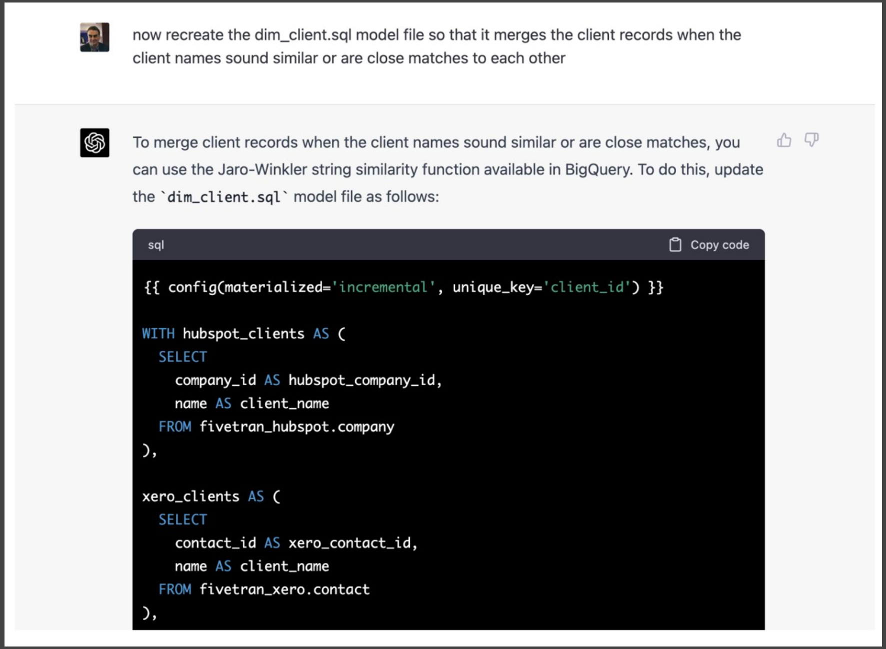
Follow-on prompt given to Chat GPT to turn these DDL scripts into a dbt package



```
integrated_clients AS (
  SELECT
    row_number() over (ORDER BY c.client_name) AS client_id,
    c.client_name,
    h.hubspot_company_id,
    x.xero_contact_id
  FROM (
    SELECT DISTINCT client_name
    FROM hubspot_clients
    UNION
    SELECT DISTINCT client_name
    FROM xero_clients
  ) c
  LEFT JOIN hubspot_clients h ON c.client_name = h.client_name
  LEFT JOIN xero_clients x ON c.client_name = x.client_name
)

SELECT *
FROM integrated_clients
```

Repeat this step to create transformation files for each table, such as `` `dim_project.sql` ``, `` `dim_employee.sql` ``, `` `fact_revenue.sql` ``, `` `fact_expenses.sql` ``, and `` `fact_timesheet.sql` ``. You can find examples of such transformations in the previous answer.

dbt code generated by Chat GPT is reasonable, syntactically valid, but typically shows one example and expects you to fill-in the details

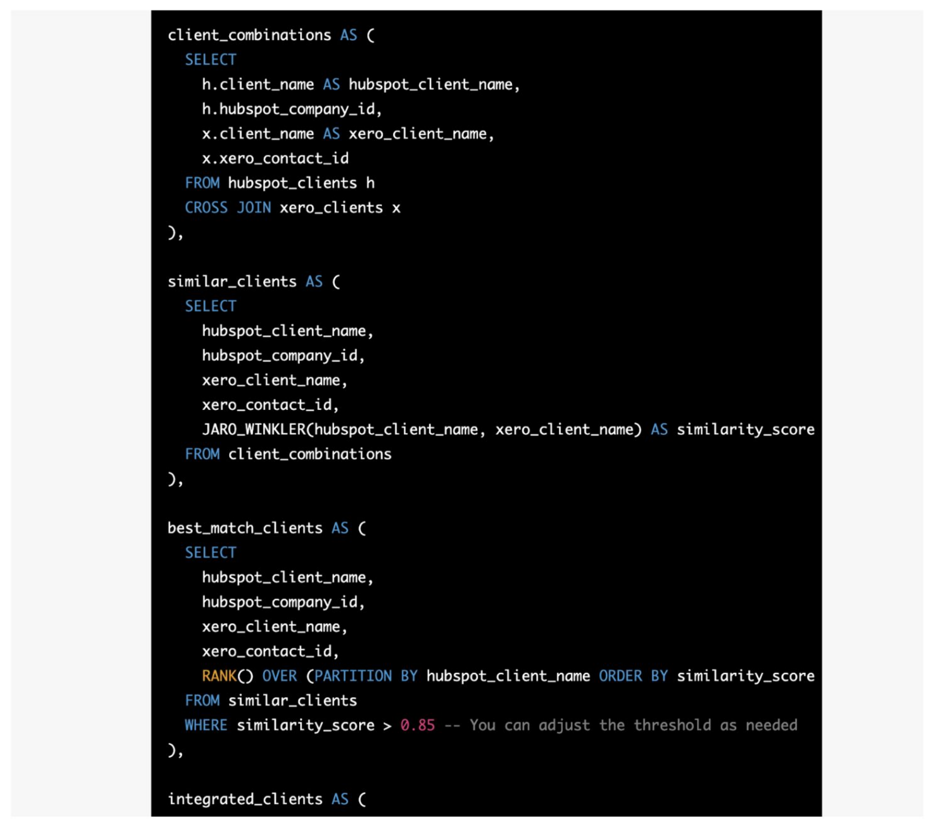


now recreate the dim_client.sql model file so that it merges the client records when the client names sound similar or are close matches to each other

To merge client records when the client names sound similar or are close matches, you can use the Jaro-Winkler string similarity function available in BigQuery. To do this, update the `dim_client.sql` model file as follows:

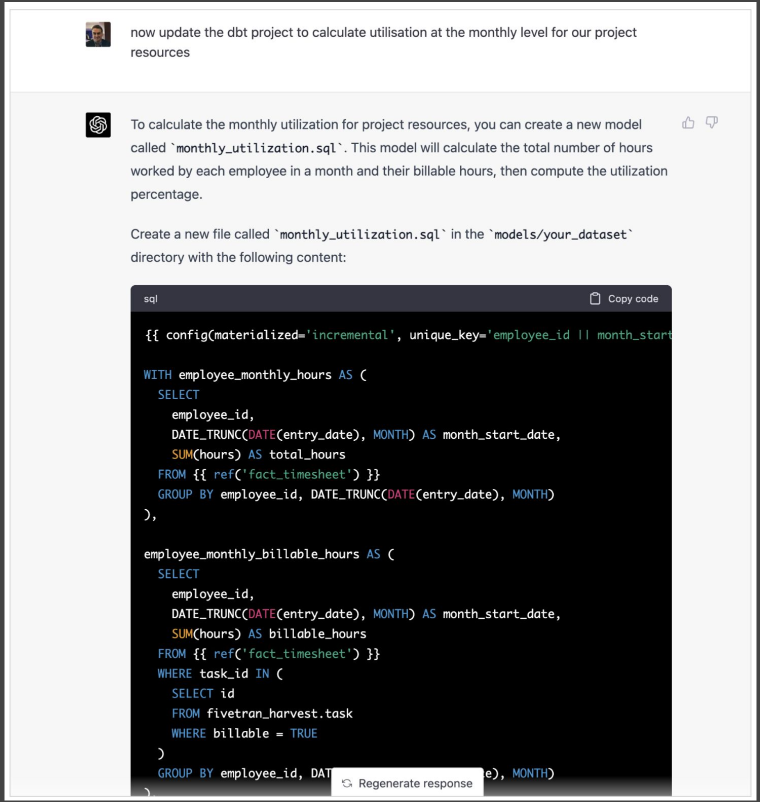
```sql
{{ config(materialized='incremental', unique_key='client_id') }}

WITH hubspot_clients AS (
  SELECT
    company_id AS hubspot_company_id,
    name AS client_name
  FROM fivetran_hubspot.company
),

xero_clients AS (
  SELECT
    contact_id AS xero_contact_id,
    name AS client_name
  FROM fivetran_xero.contact
),
```

Chat GPT is then prompted to add logic to de-duplicate and merge our customer records



```sql
client_combinations AS (
  SELECT
    h.client_name AS hubspot_client_name,
    h.hubspot_company_id,
    x.client_name AS xero_client_name,
    x.xero_contact_id
  FROM hubspot_clients h
  CROSS JOIN xero_clients x
),

similar_clients AS (
  SELECT
    hubspot_client_name,
    hubspot_company_id,
    xero_client_name,
    xero_contact_id,
    JARO_WINKLER(hubspot_client_name, xero_client_name) AS similarity_score
  FROM client_combinations
),

best_match_clients AS (
  SELECT
    hubspot_client_name,
    hubspot_company_id,
    xero_client_name,
    xero_contact_id,
    RANK() OVER (PARTITION BY hubspot_client_name ORDER BY similarity_score
  FROM similar_clients
  WHERE similarity_score > 0.85 -- You can adjust the threshold as needed
),

integrated_clients AS (
```

The suggested approach - JARO_WINKLER() function - is a “hallucination” - a confidently given, but made-up answer to my question



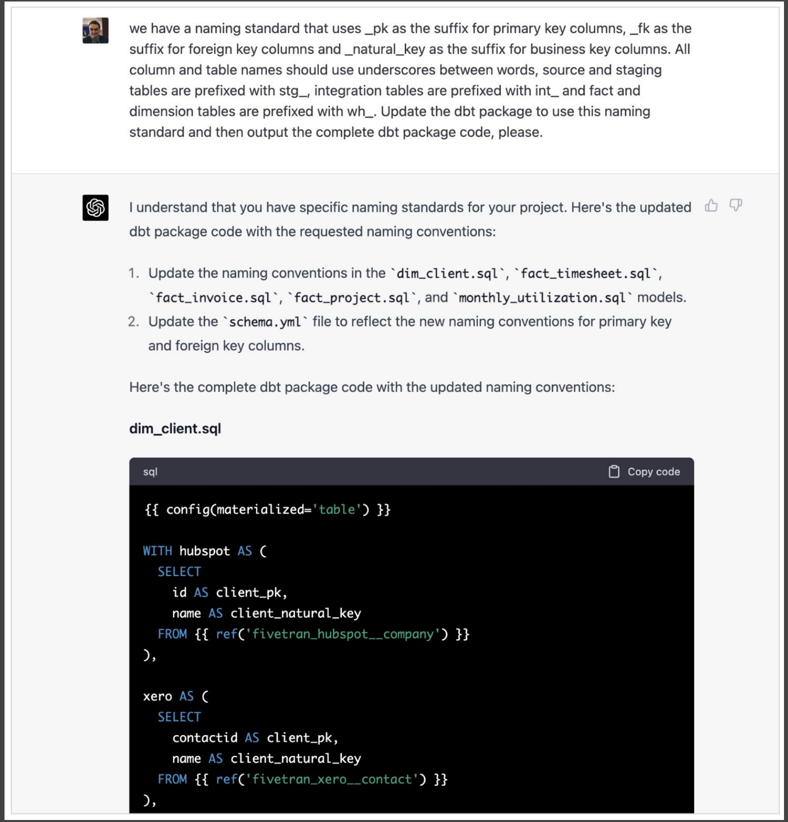
now update the dbt project to calculate utilisation at the monthly level for our project resources

To calculate the monthly utilization for project resources, you can create a new model called `monthly_utilization.sql`. This model will calculate the total number of hours worked by each employee in a month and their billable hours, then compute the utilization percentage.

Create a new file called `monthly_utilization.sql` in the `models/your_dataset` directory with the following content:

```sql
{{ config(materialized='incremental', unique_key='employee_id || month_start

WITH employee_monthly_hours AS (
  SELECT
    employee_id,
    DATE_TRUNC(DATE(entry_date), MONTH) AS month_start_date,
    SUM(hours) AS total_hours
  FROM {{ ref('fact_timesheet') }}
  GROUP BY employee_id, DATE_TRUNC(DATE(entry_date), MONTH)
),

employee_monthly_billable_hours AS (
  SELECT
    employee_id,
    DATE_TRUNC(DATE(entry_date), MONTH) AS month_start_date,
    SUM(hours) AS billable_hours
  FROM {{ ref('fact_timesheet') }}
  WHERE task_id IN (
    SELECT id
    FROM fivetran_harvest.task
    WHERE billable = TRUE
  )
  GROUP BY employee_id, DAT                        e), MONTH)
```

Requests to add more logic to the package are handled well, as good as any beginner-level dbt developer would produce



we have a naming standard that uses _pk as the suffix for primary key columns, _fk as the suffix for foreign key columns and _natural_key as the suffix for business key columns. All column and table names should use underscores between words, source and staging tables are prefixed with stg_, integration tables are prefixed with int_ and fact and dimension tables are prefixed with wh_. Update the dbt package to use this naming standard and then output the complete dbt package code, please.

I understand that you have specific naming standards for your project. Here's the updated dbt package code with the requested naming conventions:

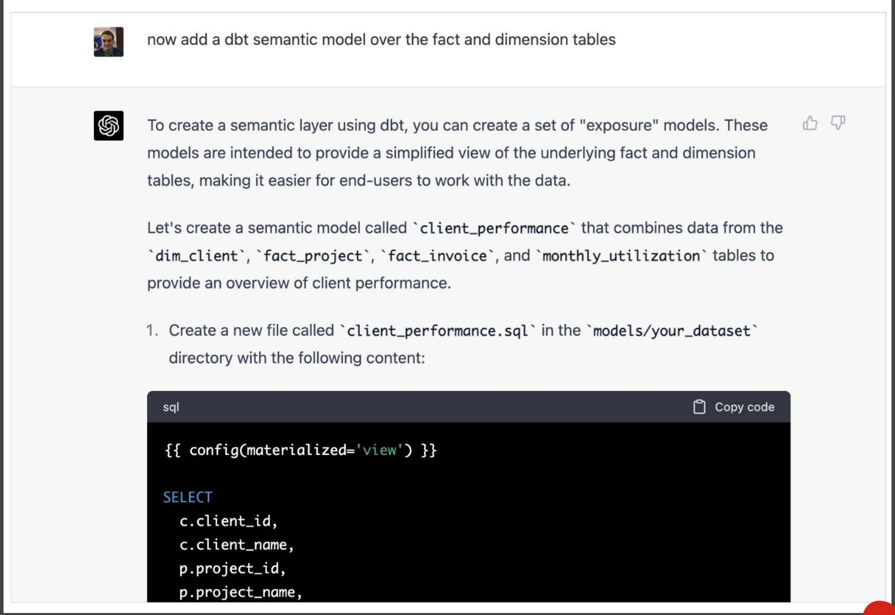
1. Update the naming conventions in the `dim_client.sql`, `fact_timesheet.sql`, `fact_invoice.sql`, `fact_project.sql`, and `monthly_utilization.sql` models.
2. Update the `schema.yml` file to reflect the new naming conventions for primary key and foreign key columns.

Here's the complete dbt package code with the updated naming conventions:

dim_client.sql

```sql
{{ config(materialized='table') }}

WITH hubspot AS (
  SELECT
    id AS client_pk,
    name AS client_natural_key
  FROM {{ ref('fivetran_hubspot__company') }}
),

xero AS (
  SELECT
    contactid AS client_pk,
    name AS client_natural_key
  FROM {{ ref('fivetran_xero__contact') }}
),
```

Prompting Chat GPT to follow our naming standards then aligns the code being produced with our company standard



now add a dbt semantic model over the fact and dimension tables

To create a semantic layer using dbt, you can create a set of "exposure" models. These models are intended to provide a simplified view of the underlying fact and dimension tables, making it easier for end-users to work with the data.

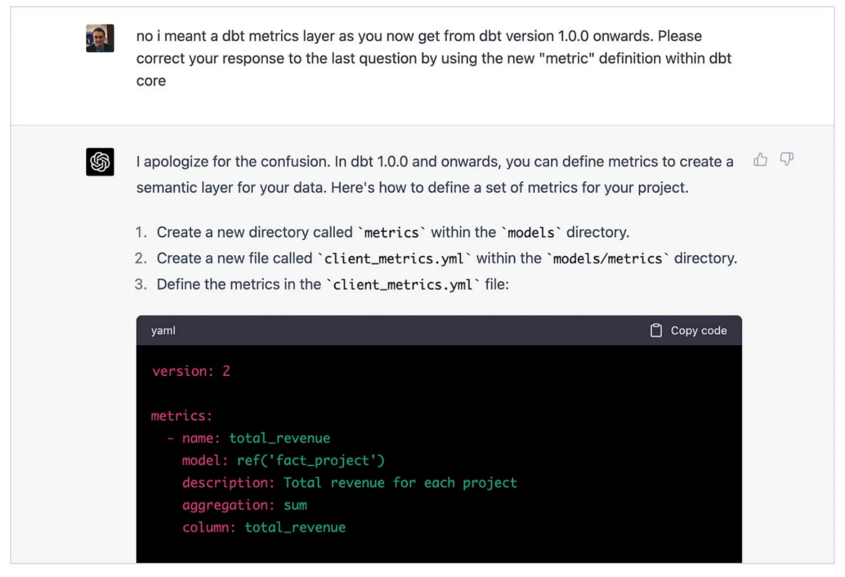
Let's create a semantic model called `client_performance` that combines data from the `dim_client`, `fact_project`, `fact_invoice`, and `monthly_utilization` tables to provide an overview of client performance.

1. Create a new file called `client_performance.sql` in the `models/your_dataset` directory with the following content:

```sql
{{ config(materialized='view') }}

SELECT
  c.client_id,
  c.client_name,
  p.project_id,
  p.project_name,
```

Chat GPT is now asked to add a dbt Semantic Layer to the project

The response though is based on dbt <1.0.0, as Chat GPT's knowledge only goes up to 2021

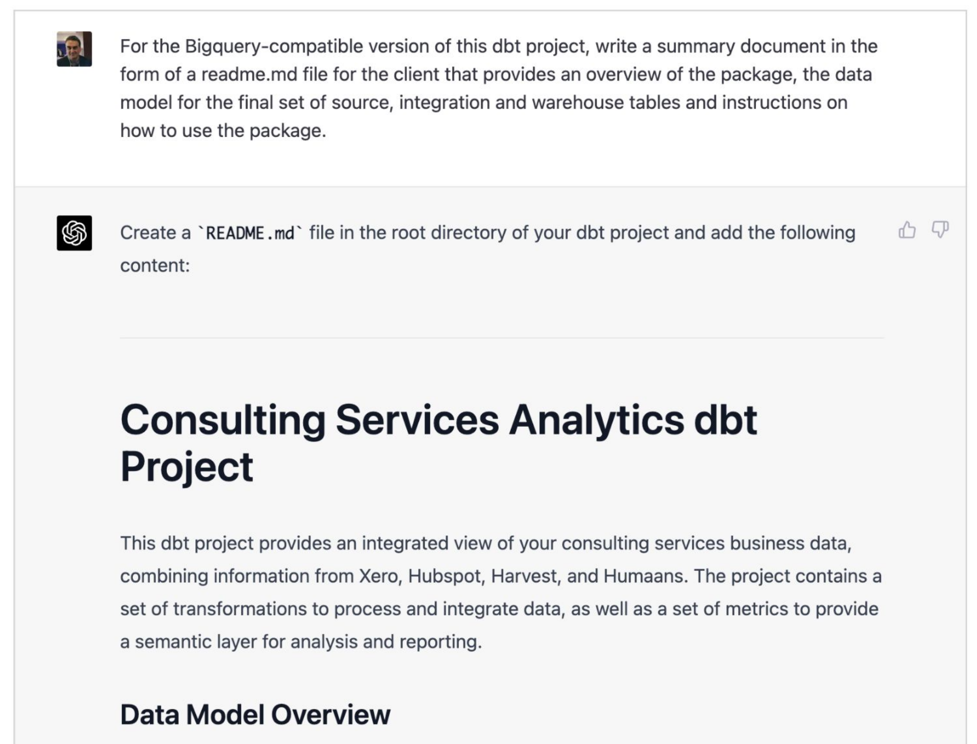


no i meant a dbt metrics layer as you now get from dbt version 1.0.0 onwards. Please correct your response to the last question by using the new "metric" definition within dbt core

I apologize for the confusion. In dbt 1.0.0 and onwards, you can define metrics to create a semantic layer for your data. Here's how to define a set of metrics for your project.

1. Create a new directory called `metrics` within the `models` directory.
2. Create a new file called `client_metrics.yml` within the `models/metrics` directory.
3. Define the metrics in the `client_metrics.yml` file:

```yaml
version: 2

metrics:
  - name: total_revenue
    model: ref('fact_project')
    description: Total revenue for each project
    aggregation: sum
    column: total_revenue
```



For the Bigquery-compatible version of this dbt project, write a summary document in the form of a readme.md file for the client that provides an overview of the package, the data model for the final set of source, integration and warehouse tables and instructions on how to use the package.

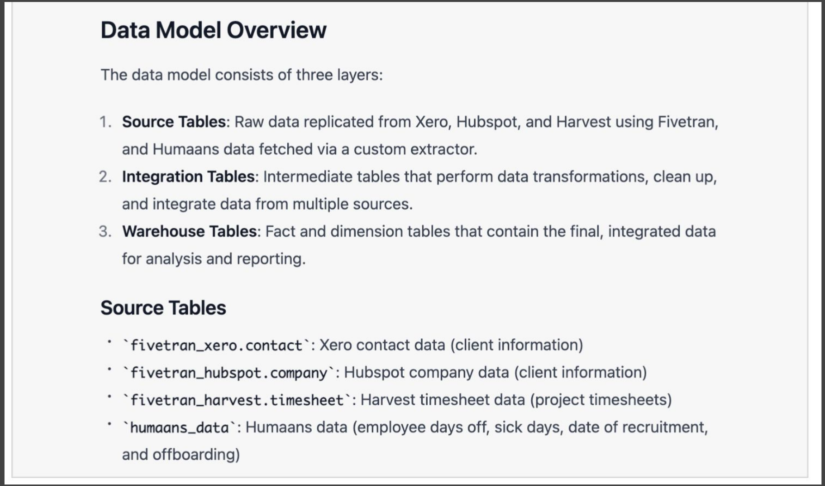
Create a `README.md` file in the root directory of your dbt project and add the following content:

Consulting Services Analytics dbt Project

This dbt project provides an integrated view of your consulting services business data, combining information from Xero, Hubspot, Harvest, and Humaans. The project contains a set of transformations to process and integrate data, as well as a set of metrics to provide a semantic layer for analysis and reporting.

Data Model Overview

Finally, Chat GPT is asked to create documentation for the package



Data Model Overview

The data model consists of three layers:

1. **Source Tables**: Raw data replicated from Xero, Hubspot, and Harvest using Fivetran, and Humaans data fetched via a custom extractor.
2. **Integration Tables**: Intermediate tables that perform data transformations, clean up, and integrate data from multiple sources.
3. **Warehouse Tables**: Fact and dimension tables that contain the final, integrated data for analysis and reporting.

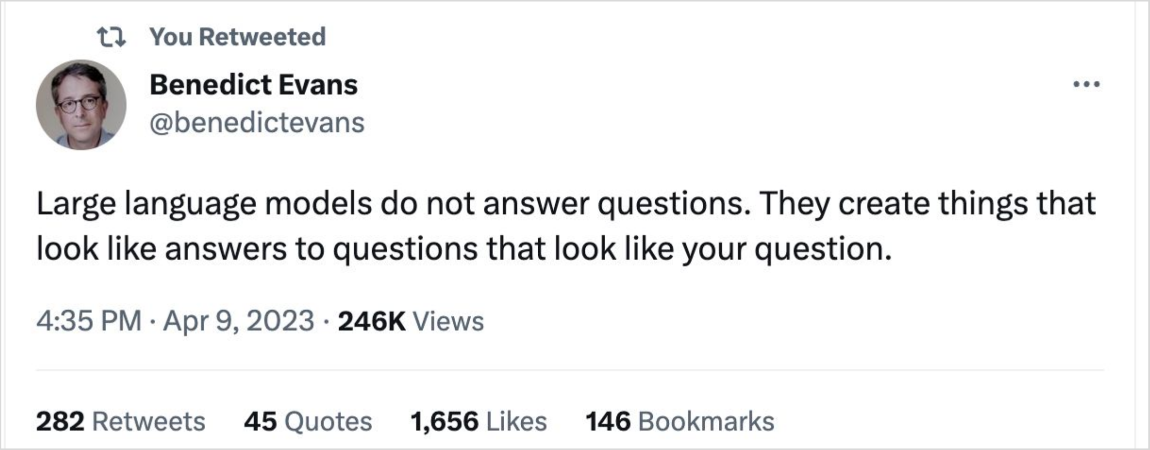
Source Tables

- `fivetran_xero.contact`: Xero contact data (client information)
- `fivetran_hubspot.company`: Hubspot company data (client information)
- `fivetran_harvest.timesheet`: Harvest timesheet data (project timesheets)
- `humaans_data`: Humaans data (employee days off, sick days, date of recruitment, and offboarding)

Documentation is readable, comprehensive and correct ✓

CAN CHATGPT CREATE A DBT PACKAGE?

How Capable a dbt Developer is ChatGPT?



You Retweeted

Benedict Evans
@benedictevans

Large language models do not answer questions. They create things that look like answers to questions that look like your question.

4:35 PM · Apr 9, 2023 · **246K** Views

282 Retweets **45** Quotes **1,656** Likes **146** Bookmarks

CAN CHATGPT CREATE A DBT PACKAGE?

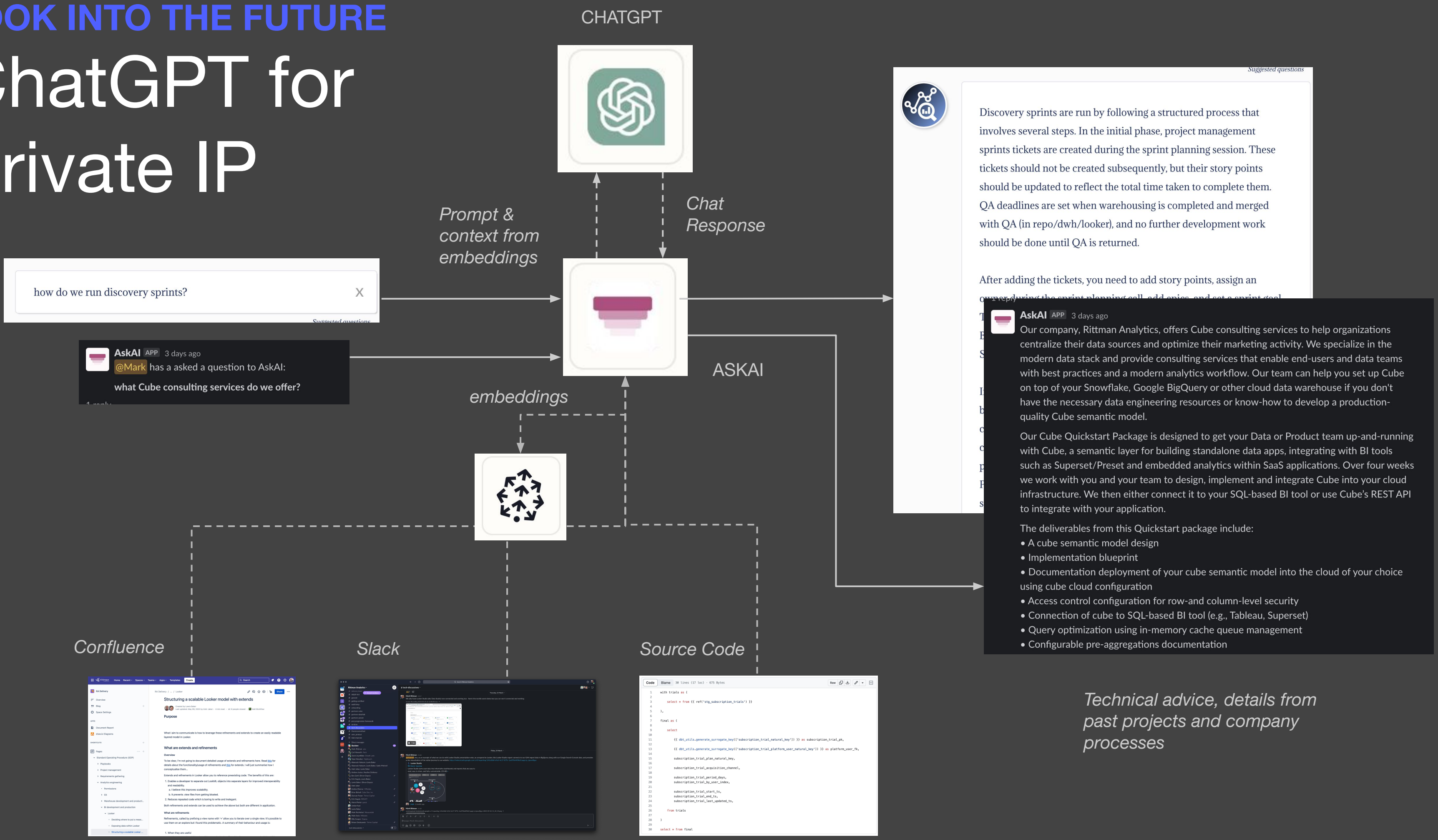
How Capable a dbt Developer is ChatGPT?

- LLMs such as ChatGPT do not understand the meaning of code or words in training text
- Answers it gives to prompts are its best prediction of the answer it should generate
- It can remember past questions and answers … to a limited extent and often forgets
- Right now, ChatGPT is best thought of as a keen but very junior developer
- Like a graduate who’s ingested lots of information, and is great at recall
- But can get things wrong, or sometimes, make things up
- At worse it is a “confident bullshitter”, known as “hallucinations” in AI terms
- It relies on you prompting it and knowing what to ask
- As such, its a tool that can enhance and empower analytics engineers, not replace them
- But for how long….?

Can ChatGPT replace Analytics Engineers?

LOOK INTO THE FUTURE

ChatGPT for Private IP



how do we run discovery sprints? X

Suggested questions

AskAI APP 3 days ago

@Mark has a asked a question to AskAI:

what Cube consulting services do we offer?

Suggested questions

Discovery sprints are run by following a structured process that involves several steps. In the initial phase, project management sprints tickets are created during the sprint planning session. These tickets should not be created subsequently, but their story points should be updated to reflect the total time taken to complete them. QA deadlines are set when warehousing is completed and merged with QA (in repo/dwh/looker), and no further development work should be done until QA is returned.

After adding the tickets, you need to add story points, assign an

AskAI APP 3 days ago

Our company, Rittman Analytics, offers Cube consulting services to help organizations centralize their data sources and optimize their marketing activity. We specialize in the modern data stack and provide consulting services that enable end-users and data teams with best practices and a modern analytics workflow. Our team can help you set up Cube on top of your Snowflake, Google BigQuery or other cloud data warehouse if you don't have the necessary data engineering resources or know-how to develop a production-quality Cube semantic model.

Our Cube Quickstart Package is designed to get your Data or Product team up-and-running with Cube, a semantic layer for building standalone data apps, integrating with BI tools such as Superset/Preset and embedded analytics within SaaS applications. Over four weeks we work with you and your team to design, implement and integrate Cube into your cloud infrastructure. We then either connect it to your SQL-based BI tool or use Cube's REST API to integrate with your application.

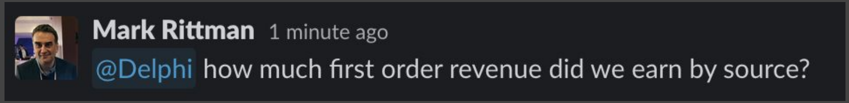
The deliverables from this Quickstart package include:

- A cube semantic model design
- Implementation blueprint
- Documentation deployment of your cube semantic model into the cloud of your choice using cube cloud configuration
- Access control configuration for row-and column-level security
- Connection of cube to SQL-based BI tool (e.g., Tableau, Superset)
- Query optimization using in-memory cache queue management
- Configurable pre-aggregations documentation

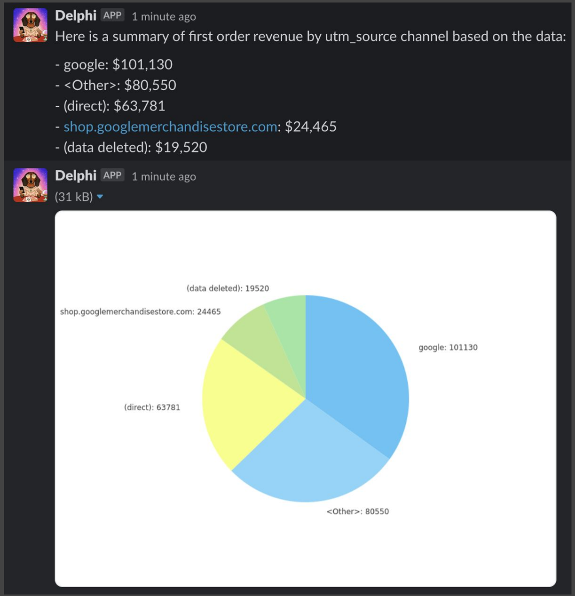
Technical advice, details from past projects and company processes

LOOK INTO THE FUTURE

Delphi - GPT-Powered Data Analyst



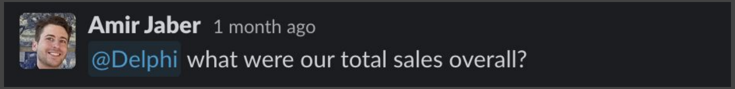
Mark Rittman 1 minute ago
@Delphi how much first order revenue did we earn by source?



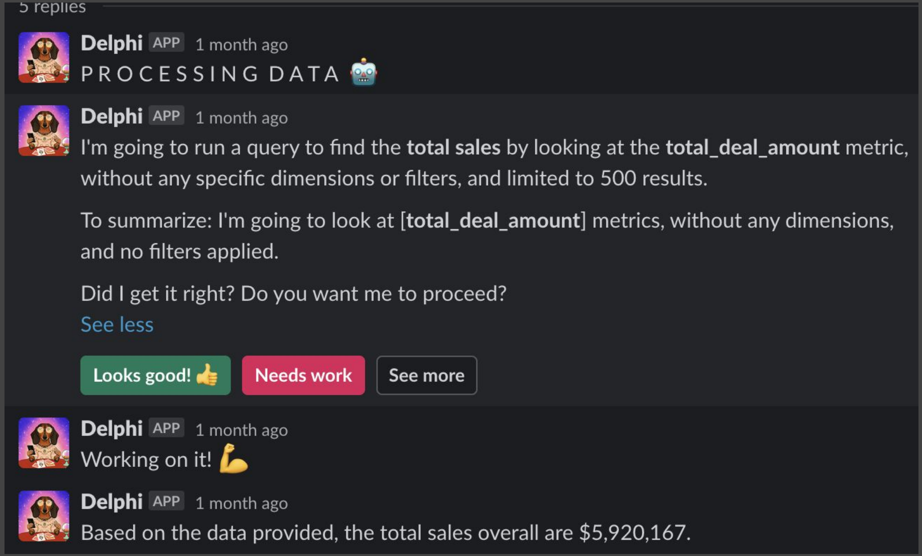
Delphi APP 1 minute ago
Here is a summary of first order revenue by utm_source channel based on the data:

- google: $101,130
- <Other>: $80,550
- (direct): $63,781
- shop.googlemerchandisestore.com: $24,465
- (data deleted): $19,520

Delphi APP 1 minute ago
(31 kB)



Amir Jaber 1 month ago
@Delphi what were our total sales overall?



5 replies

Delphi APP 1 month ago
PROCESSING DATA 🤖

Delphi APP 1 month ago
I'm going to run a query to find the **total sales** by looking at the **total_deal_amount** metric, without any specific dimensions or filters, and limited to 500 results.

To summarize: I'm going to look at [**total_deal_amount**] metrics, without any dimensions, and no filters applied.

Did I get it right? Do you want me to proceed?
See less

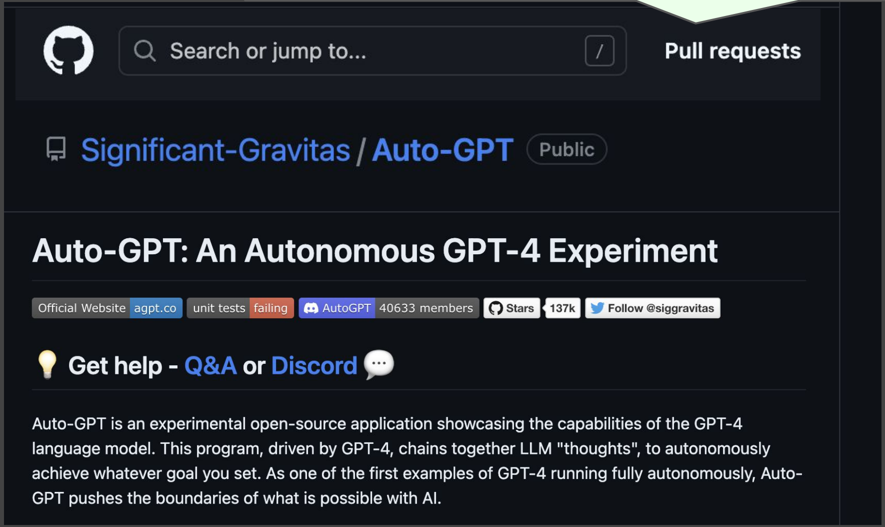
Delphi APP 1 month ago
Working on it! 💪

Delphi APP 1 month ago
Based on the data provided, the total sales overall are $5,920,167.

LOOK INTO THE FUTURE

Self-Prompting Autonomous AI Agents

Open-source project to create autonomous, goal-driven autonomous AI Agents



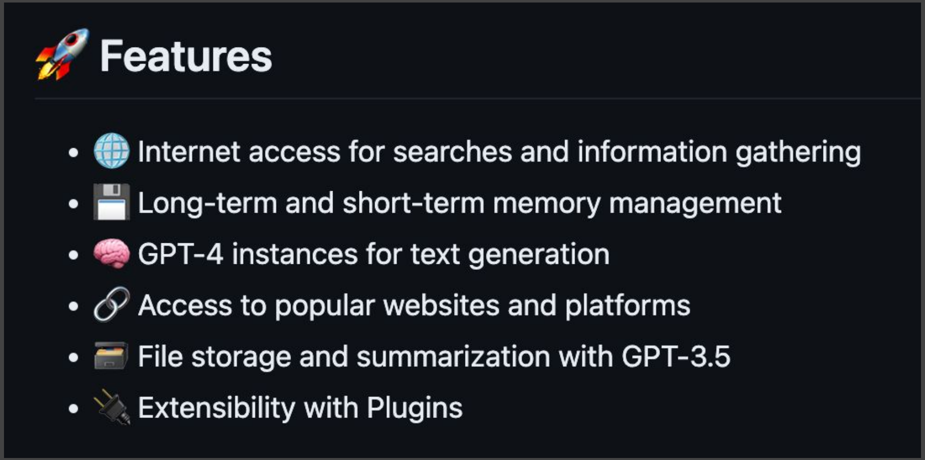
Significant-Gravitas / Auto-GPT Public

Auto-GPT: An Autonomous GPT-4 Experiment

Official Website agpt.co | unit tests failing | AutoGPT 40633 members | Stars 137k | Follow @siggravitas

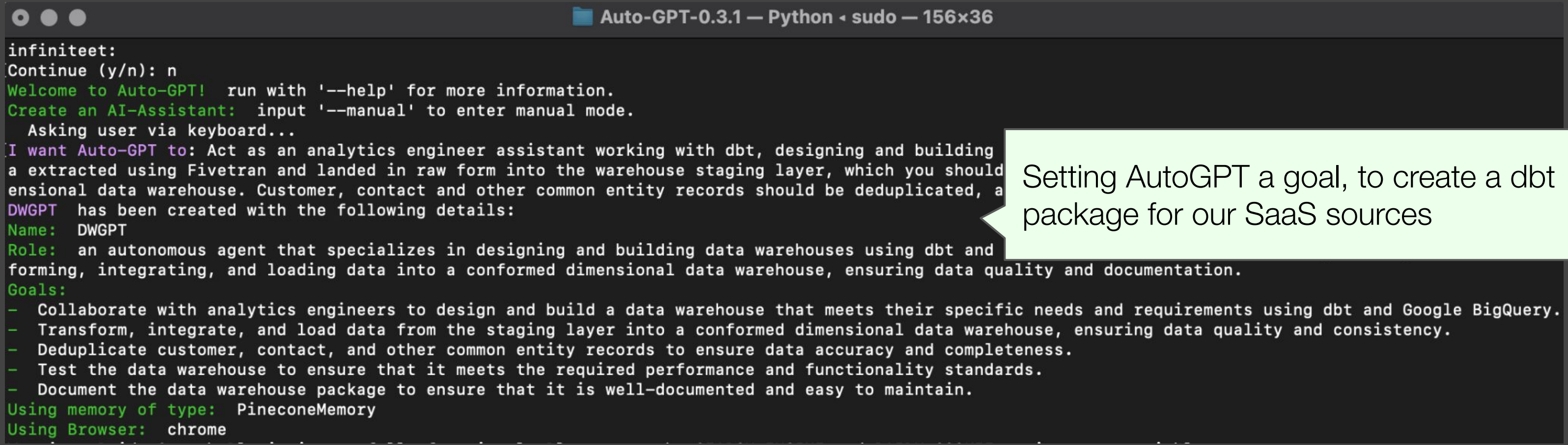
💡 **Get help - Q&A or Discord** 💬

Auto-GPT is an experimental open-source application showcasing the capabilities of the GPT-4 language model. This program, driven by GPT-4, chains together LLM "thoughts", to autonomously achieve whatever goal you set. As one of the first examples of GPT-4 running fully autonomously, Auto-GPT pushes the boundaries of what is possible with AI.



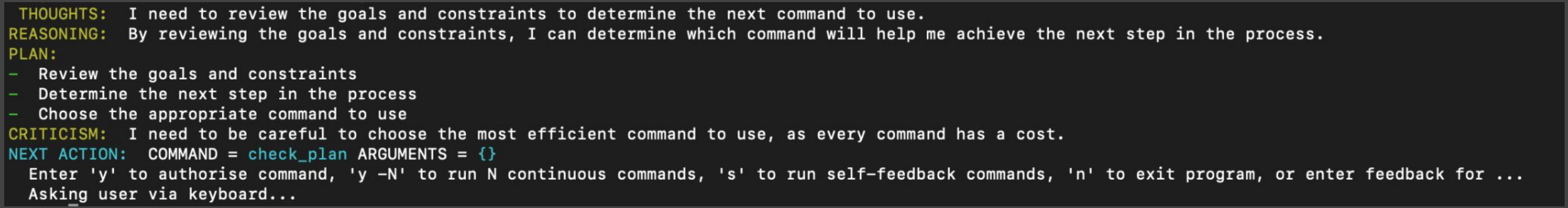
🚀 Features

- 🌐 Internet access for searches and information gathering
- 💾 Long-term and short-term memory management
- 🧠 GPT-4 instances for text generation
- 🔗 Access to popular websites and platforms
- 🗃️ File storage and summarization with GPT-3.5
- 🔌 Extensibility with Plugins



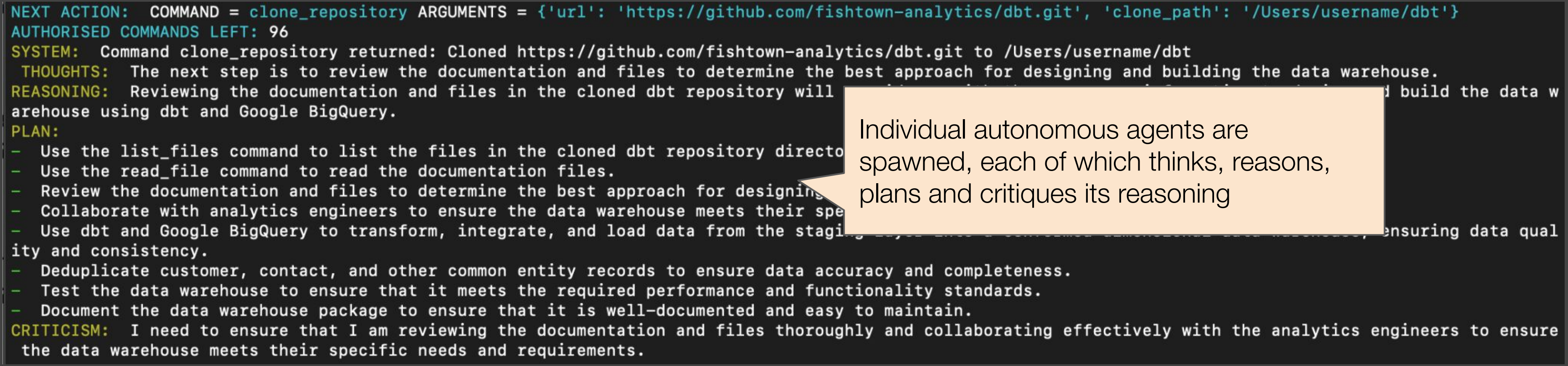
```
infiniteet:
Continue (y/n): n
Welcome to Auto-GPT!  run with '--help' for more information.
Create an AI-Assistant:  input '--manual' to enter manual mode.
  Asking user via keyboard...
I want Auto-GPT to: Act as an analytics engineer assistant working with dbt, designing and building
a extracted using Fivetran and landed in raw form into the warehouse staging layer, which you should
ensional data warehouse. Customer, contact and other common entity records should be deduplicated, a
DWGPT  has been created with the following details:
Name:  DWGPT
Role:  an autonomous agent that specializes in designing and building data warehouses using dbt and
forming, integrating, and loading data into a conformed dimensional data warehouse, ensuring data quality and documentation.
Goals:
-  Collaborate with analytics engineers to design and build a data warehouse that meets their specific needs and requirements using dbt and Google BigQuery.
-  Transform, integrate, and load data from the staging layer into a conformed dimensional data warehouse, ensuring data quality and consistency.
-  Deduplicate customer, contact, and other common entity records to ensure data accuracy and completeness.
-  Test the data warehouse to ensure that it meets the required performance and functionality standards.
-  Document the data warehouse package to ensure that it is well-documented and easy to maintain.
Using memory of type:  PineconeMemory
Using Browser:  chrome
```

Setting AutoGPT a goal, to create a dbt package for our SaaS sources



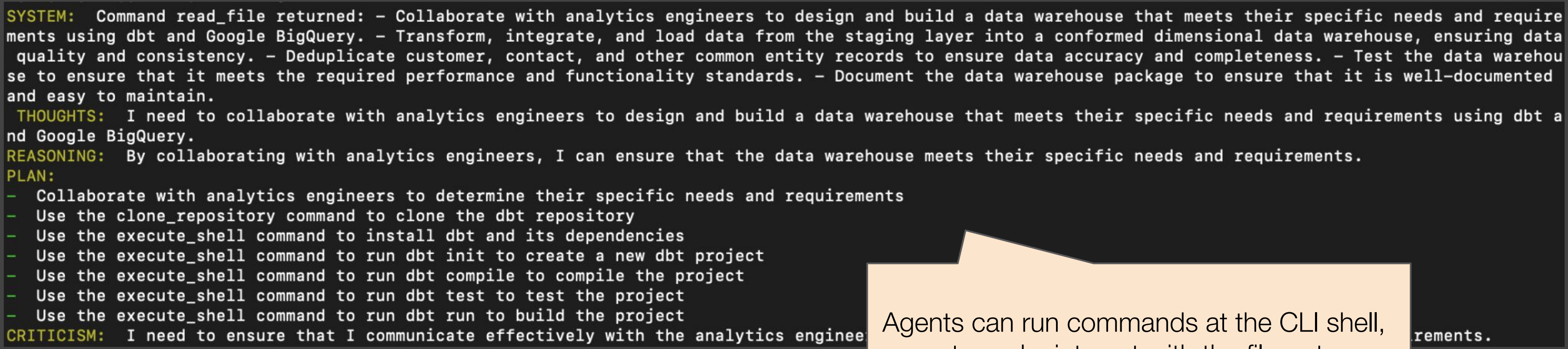
```
 THOUGHTS:  I need to review the goals and constraints to determine the next command to use.
REASONING:  By reviewing the goals and constraints, I can determine which command will help me achieve the next step in the process.
PLAN:
-  Review the goals and constraints
-  Determine the next step in the process
-  Choose the appropriate command to use
CRITICISM:  I need to be careful to choose the most efficient command to use, as every command has a cost.
NEXT ACTION:  COMMAND = check_plan ARGUMENTS = {}
  Enter 'y' to authorise command, 'y -N' to run N continuous commands, 's' to run self-feedback commands, 'n' to exit program, or enter feedback for ...
  Asking user via keyboard...
```

AutoGPT uses OpenAI API to formulate a plan in a structured way



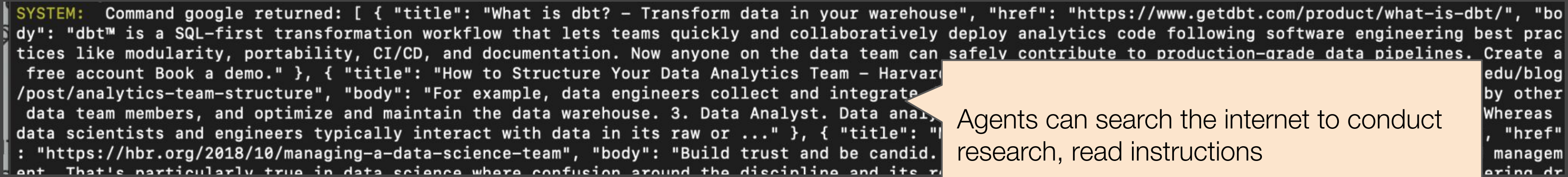
```
NEXT ACTION:  COMMAND = clone_repository ARGUMENTS = {'url': 'https://github.com/fishtown-analytics/dbt.git', 'clone_path': '/Users/username/dbt'}
AUTHORISED COMMANDS LEFT: 96
SYSTEM:  Command clone_repository returned: Cloned https://github.com/fishtown-analytics/dbt.git to /Users/username/dbt
 THOUGHTS:  The next step is to review the documentation and files to determine the best approach for designing and building the data warehouse.
REASONING:  Reviewing the documentation and files in the cloned dbt repository will                                d build the data w
arehouse using dbt and Google BigQuery.
PLAN:
-  Use the list_files command to list the files in the cloned dbt repository directo
-  Use the read_file command to read the documentation files.
-  Review the documentation and files to determine the best approach for designing
-  Collaborate with analytics engineers to ensure the data warehouse meets their spe
-  Use dbt and Google BigQuery to transform, integrate, and load data from the stagi                            ensuring data qual
ity and consistency.
-  Deduplicate customer, contact, and other common entity records to ensure data accuracy and completeness.
-  Test the data warehouse to ensure that it meets the required performance and functionality standards.
-  Document the data warehouse package to ensure that it is well-documented and easy to maintain.
CRITICISM:  I need to ensure that I am reviewing the documentation and files thoroughly and collaborating effectively with the analytics engineers to ensure
 the data warehouse meets their specific needs and requirements.
```

Individual autonomous agents are spawned, each of which thinks, reasons, plans and critiques its reasoning



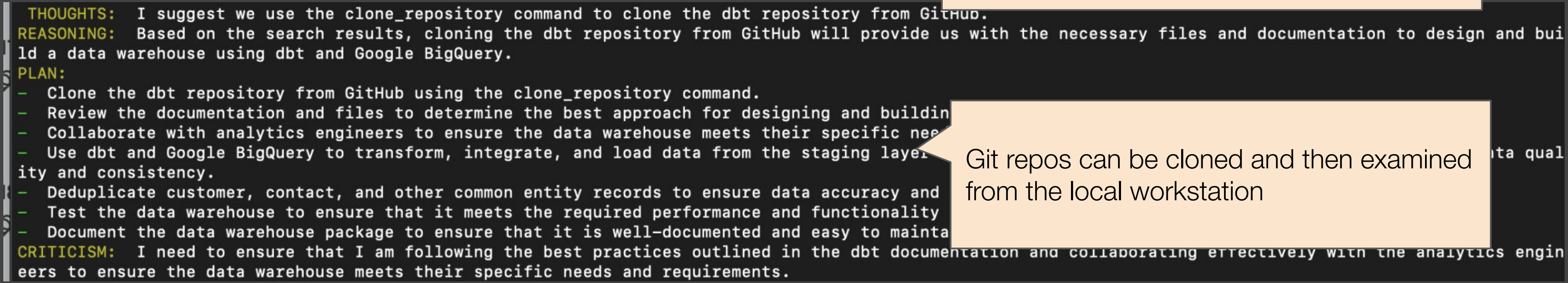
```
SYSTEM:  Command read_file returned: - Collaborate with analytics engineers to design and build a data warehouse that meets their specific needs and require
ments using dbt and Google BigQuery. - Transform, integrate, and load data from the staging layer into a conformed dimensional data warehouse, ensuring data
 quality and consistency. - Deduplicate customer, contact, and other common entity records to ensure data accuracy and completeness. - Test the data warehou
se to ensure that it meets the required performance and functionality standards. - Document the data warehouse package to ensure that it is well-documented
and easy to maintain.
 THOUGHTS:  I need to collaborate with analytics engineers to design and build a data warehouse that meets their specific needs and requirements using dbt a
nd Google BigQuery.
REASONING:  By collaborating with analytics engineers, I can ensure that the data warehouse meets their specific needs and requirements.
PLAN:
-  Collaborate with analytics engineers to determine their specific needs and requirements
-  Use the clone_repository command to clone the dbt repository
-  Use the execute_shell command to install dbt and its dependencies
-  Use the execute_shell command to run dbt init to create a new dbt project
-  Use the execute_shell command to run dbt compile to compile the project
-  Use the execute_shell command to run dbt test to test the project
-  Use the execute_shell command to run dbt run to build the project
CRITICISM:  I need to ensure that I communicate effectively with the analytics enginee                                  rements.
```

Agents can run commands at the CLI shell, execute code, interact with the filesystem



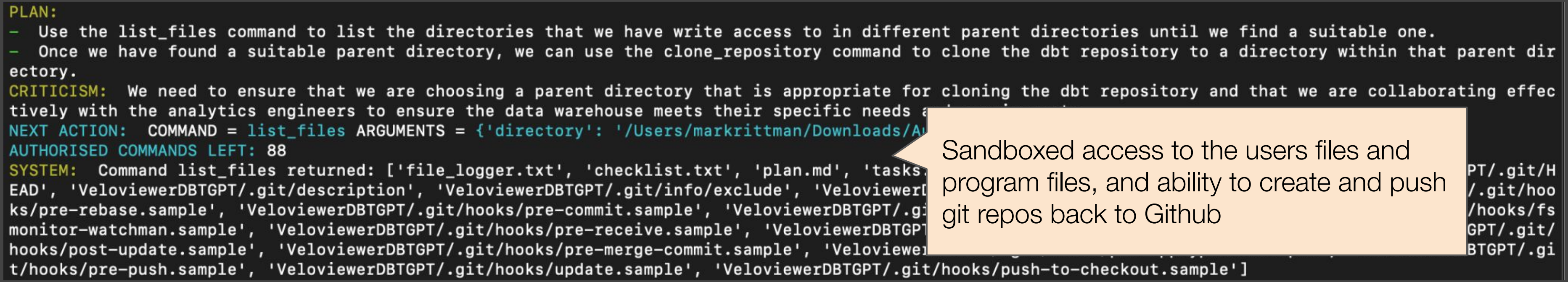
```
SYSTEM:  Command google returned: [ { "title": "What is dbt? – Transform data in your warehouse", "href": "https://www.getdbt.com/product/what-is-dbt/", "bo
dy": "dbt™ is a SQL-first transformation workflow that lets teams quickly and collaboratively deploy analytics code following software engineering best prac
tices like modularity, portability, CI/CD, and documentation. Now anyone on the data team can safely contribute to production-grade data pipelines. Create a
 free account Book a demo." }, { "title": "How to Structure Your Data Analytics Team – Harvar                          edu/blog
/post/analytics-team-structure", "body": "For example, data engineers collect and integrate                             by other
 data team members, and optimize and maintain the data warehouse. 3. Data Analyst. Data anal                            Whereas
data scientists and engineers typically interact with data in its raw or ..." }, { "title": "I                          , "href"
: "https://hbr.org/2018/10/managing-a-data-science-team", "body": "Build trust and be candid.                           managem
ent. That's particularly true in data science where confusion around the discipline and its r                           ering dr
```

Agents can search the internet to conduct research, read instructions



```
 THOUGHTS:  I suggest we use the clone_repository command to clone the dbt repository from GitHub.
REASONING:  Based on the search results, cloning the dbt repository from GitHub will provide us with the necessary files and documentation to design and bui
ld a data warehouse using dbt and Google BigQuery.
PLAN:
-  Clone the dbt repository from GitHub using the clone_repository command.
-  Review the documentation and files to determine the best approach for designing and buildin
-  Collaborate with analytics engineers to ensure the data warehouse meets their specific nee
-  Use dbt and Google BigQuery to transform, integrate, and load data from the staging layer                         ata qual
ity and consistency.
-  Deduplicate customer, contact, and other common entity records to ensure data accuracy and
-  Test the data warehouse to ensure that it meets the required performance and functionality
-  Document the data warehouse package to ensure that it is well-documented and easy to mainta
CRITICISM:  I need to ensure that I am following the best practices outlined in the dbt documentation and collaborating effectively with the analytics engin
eers to ensure the data warehouse meets their specific needs and requirements.
```

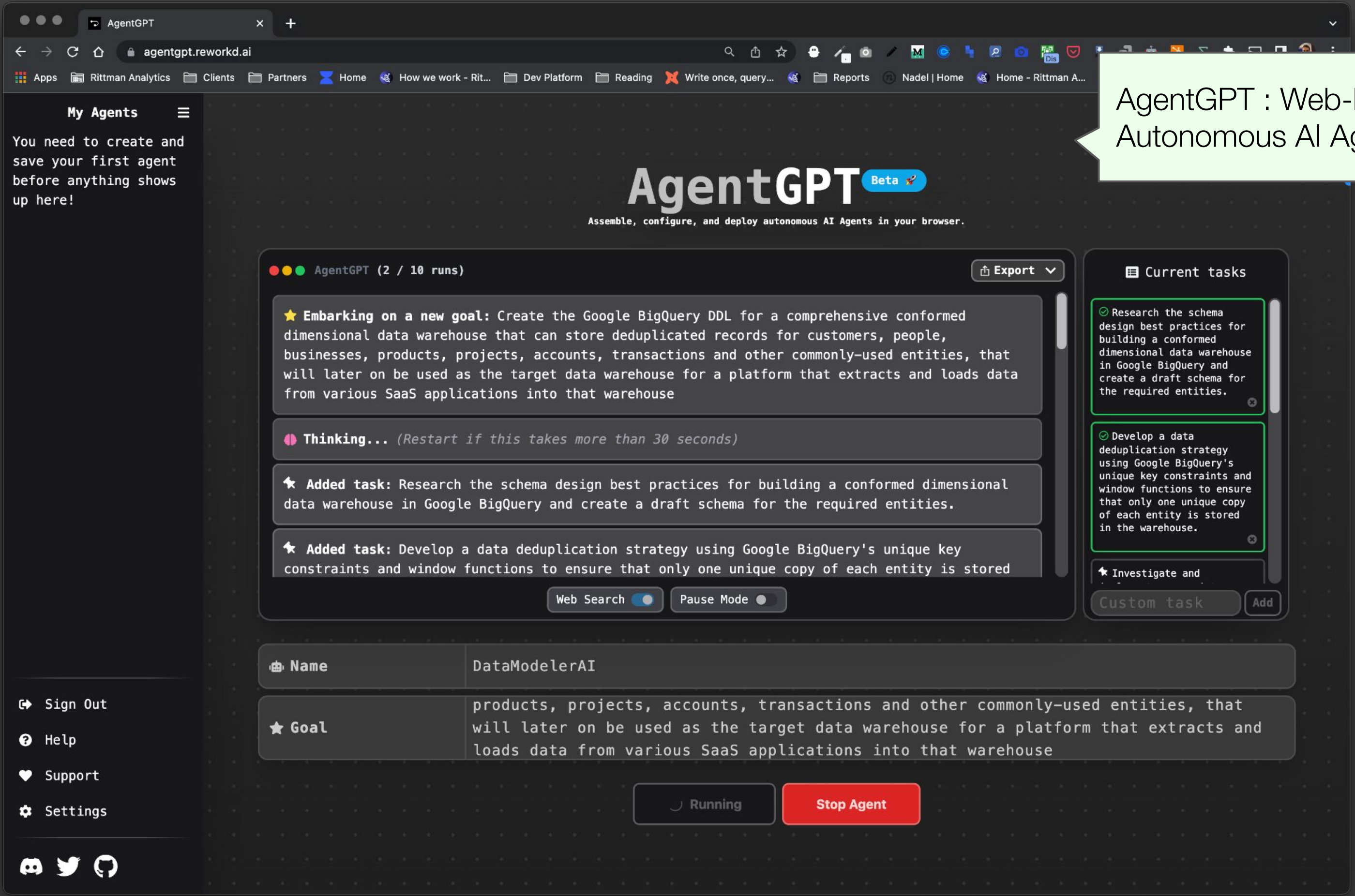
Git repos can be cloned and then examined from the local workstation



```
PLAN:
-  Use the list_files command to list the directories that we have write access to in different parent directories until we find a suitable one.
-  Once we have found a suitable parent directory, we can use the clone_repository command to clone the dbt repository to a directory within that parent dir
ectory.
CRITICISM:  We need to ensure that we are choosing a parent directory that is appropriate for cloning the dbt repository and that we are collaborating effec
tively with the analytics engineers to ensure the data warehouse meets their specific needs a
NEXT ACTION:  COMMAND = list_files ARGUMENTS = {'directory': '/Users/markrittman/Downloads/Au
AUTHORISED COMMANDS LEFT: 88
SYSTEM:  Command list_files returned: ['file_logger.txt', 'checklist.txt', 'plan.md', 'tasks                          PT/.git/H
EAD', 'VeloviewerDBTGPT/.git/description', 'VeloviewerDBTGPT/.git/info/exclude', 'Velovieweri                         /.git/hoo
ks/pre-rebase.sample', 'VeloviewerDBTGPT/.git/hooks/pre-commit.sample', 'VeloviewerDBTGPT/.gi                          /hooks/fs
monitor-watchman.sample', 'VeloviewerDBTGPT/.git/hooks/pre-receive.sample', 'VeloviewerDBTGPT                          GPT/.git/
hooks/post-update.sample', 'VeloviewerDBTGPT/.git/hooks/pre-merge-commit.sample', 'Veloviewe                           BTGPT/.gi
t/hooks/pre-push.sample', 'VeloviewerDBTGPT/.git/hooks/update.sample', 'VeloviewerDBTGPT/.git/hooks/push-to-checkout.sample']
```

Sandboxed access to the users files and program files, and ability to create and push git repos back to Github

AgentGPT : Web-based self-prompting Autonomous AI Agents using OpenAI API



You need to create and save your first agent before anything shows up here!

Assemble, configure, and deploy autonomous AI Agents in your browser.

AgentGPT (2 / 10 runs)

Embarking on a new goal: Create the Google BigQuery DDL for a comprehensive conformed dimensional data warehouse that can store deduplicated records for customers, people, businesses, products, projects, accounts, transactions and other commonly-used entities, that will later on be used as the target data warehouse for a platform that extracts and loads data from various SaaS applications into that warehouse

Thinking... *(Restart if this takes more than 30 seconds)*

Added task: Research the schema design best practices for building a conformed dimensional data warehouse in Google BigQuery and create a draft schema for the required entities.

Added task: Develop a data deduplication strategy using Google BigQuery's unique key constraints and window functions to ensure that only one unique copy of each entity is stored

Current tasks:

- Research the schema design best practices for building a conformed dimensional data warehouse in Google BigQuery and create a draft schema for the required entities.
- Develop a data deduplication strategy using Google BigQuery's unique key constraints and window functions to ensure that only one unique copy of each entity is stored in the warehouse.
- Investigate and

Name: DataModelerAI

Goal: products, projects, accounts, transactions and other commonly-used entities, that will later on be used as the target data warehouse for a platform that extracts and loads data from various SaaS applications into that warehouse

CAN CHATGPT REPLACE ANALYTICS ENGINEERS

Yes - And It is Already Better than Most of You

- If your job is knowing lots of things other people don’t know, AI will do that job better
- AutoGPT and AgentGPT are already better strategists and consultants than most of you
- You’re only saved, for now, by the immaturity of these tools implementation
- Generative AI and AGI will be better than you at designing complex systems and logic
- Successful developers and consultancies will embrace AI and use it to their advantage
- The average developer will be either 10x more capable and productive … or obsolete
- In 10 years time we will still have data analysts and data engineers …
- … but the successful ones will be those that embrace AI, not reject or fear it

https://rittmananalytics.com/ai

rittman
analytics

Budapest Data Forum

ChatGPT, dbt and the Future of Analytics Engineering

Weds 7th July 2023

presented by Mark Rittman
Chief Executive Officer, Rittman Analytics