

Wednesday 30-Minute Webinar Series

# ChatGPT, dbt and the Future of Analytics Engineering

📅 Weds 26th July 4pm UK / 11am ET / 8am PT

Register Now

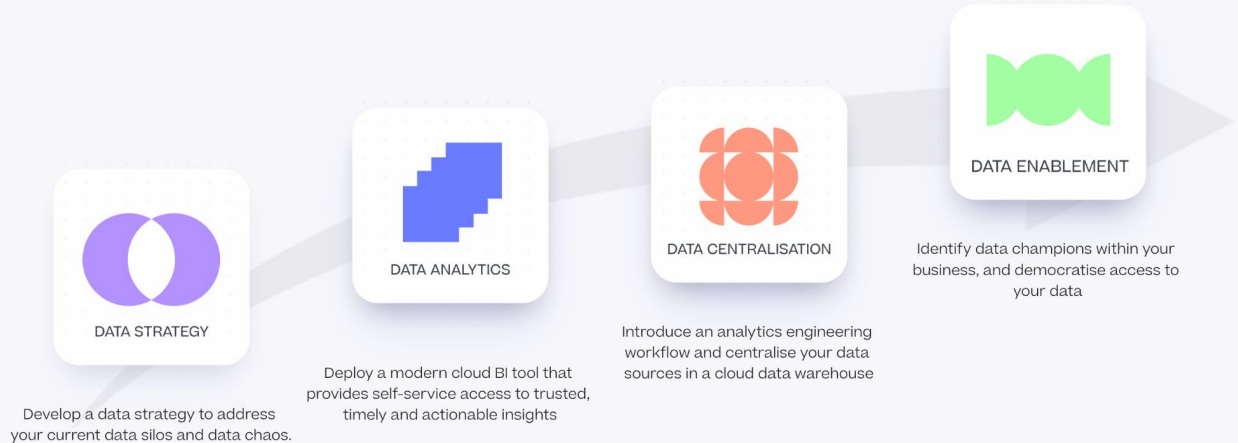
presented by Mark Rittman  
Chief Executive Officer, Rittman Analytics



## 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>

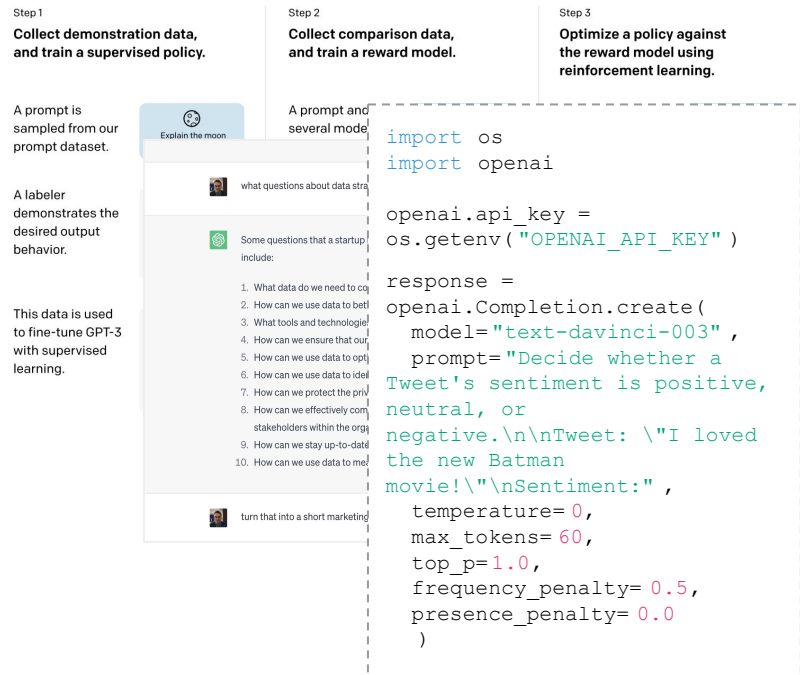


What Is ChatGPT and who are Open AI?

## WHAT IS CHATGPT?

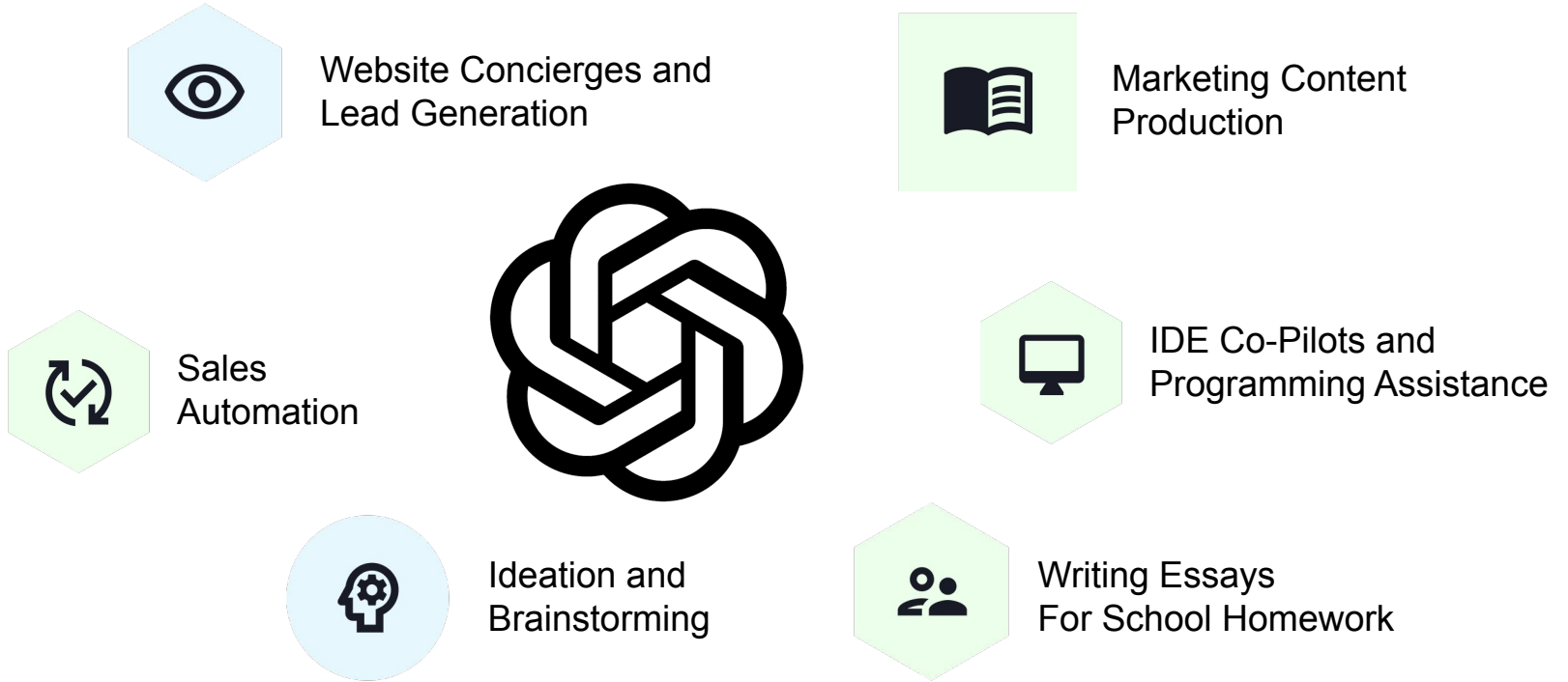
# Large Language Models (LLMs)

- 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
- 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.



## INTRODUCTION TO CHATGPT

# Initial Use-Cases for ChatGPT and LLMs



CHATGPT



Prompt & context from embeddings

Chat Response



ASKAI

embeddings




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? X

Suggested questions

RudderStack offers Marketing Officers that enhance customer growth, and improve RudderStack is i amounts of customer actual usage of you that CMOs can make 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.

Try these follow up questions

[1] The CMO's Guide  
[2] Rittman Analytics

Can you give a longer, more detailed explanation?

Ask me a follow up question

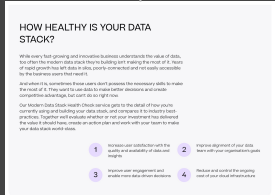
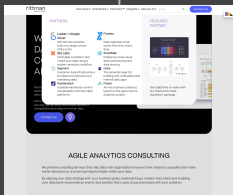
Explain this like I'm 5 years old

Re-write this using bullet points

RA Web Pages

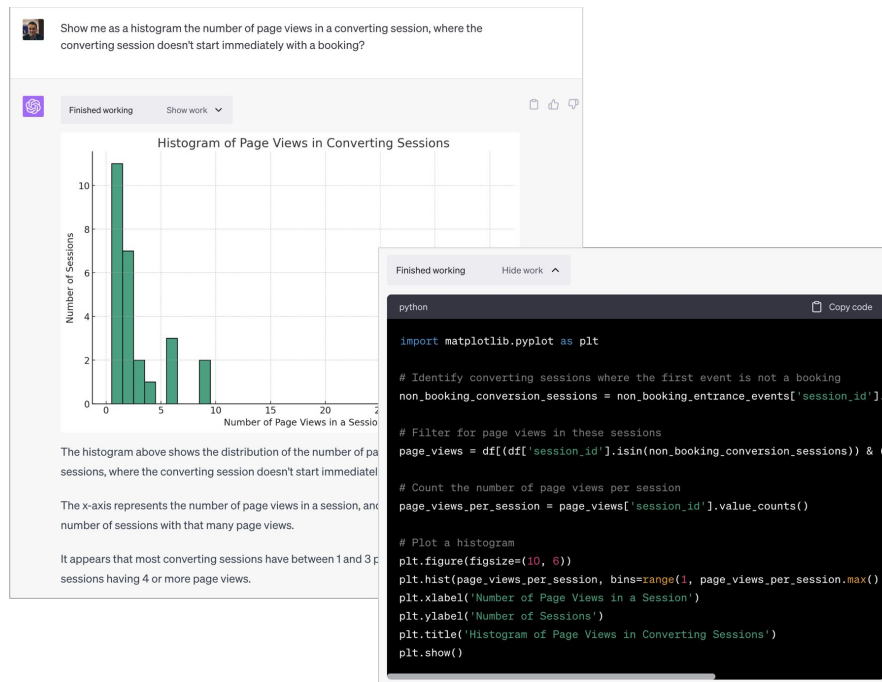
RA Blog Posts

RA Services



# ChatGPT & Code Interpreter Plugin

- AI-Powered Data Analysis
- Conversational UI from ChatGPT
- Writes and execute Python code
- Suits ChatGPT paradigm because:
  - Bounded problem space (data model)
  - Conversational UI suits user persona
  - Scope of code required is limited
  - Prompts naturally provide clarity
- See “Chat with Your Data” plugin



## ChatGPT as a dbt Coding Assistant



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:

```
{%- macro r  
(  
  {% set rel  
  {% for sou  
  {% do rel  
  {% endfor  
  
  {{ dbt_util  
  relations  
  }}  
  )  
}  
{%- endmac
```



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)



# Can ChatGPT actually code a dbt Package?

<https://chat.openai.com/share/914df714-bb36-44c8-b3fa-883aef3744c5>



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 Copy code  
  
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

Copy code

```
dbt init your_project_name
```

3. Navigate to the newly created project directory:

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

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.



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 Copy code
{{ 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
),
```

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

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

```
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 (
```



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
Copy code

{{ config(materialized='incremental', unique_key='employee_id || month_start_date') }}

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

employee_monthly_billable_hours AS (
  SELECT
    employee_id,
    DATE_TRUNC(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, DATE_TRUNC(ENTRY_DATE, 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
Copy code

{{ 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  Copy code  
  
{{ config(materialized='view') }}  
  
SELECT  
  c.client_id,  
  c.client_name,  
  p.project_id,  
  p.project_name,
```



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

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




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  Copy code  
  
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?

# 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
```

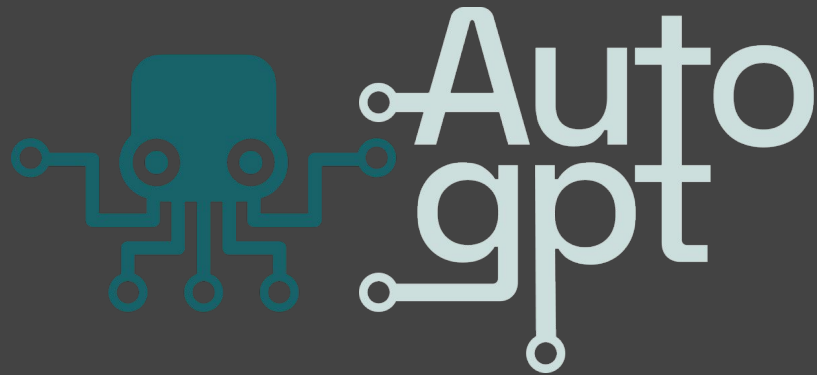
```
    intermediate:
```

```
      materialized: view
```



LOOK INTO THE FUTURE

# Self-Prompting Autonomous AI Agents









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

A screenshot of the GitHub repository page for Auto-GPT. The page shows the repository name 'Significant-Gravitas / Auto-GPT' with a 'Public' badge. Below the repository name, there is a section titled 'Auto-GPT: An Autonomous GPT-4 Experiment'. Underneath this title, there are several badges: 'Official Website agpt.co', 'unit tests failing', 'AutoGPT 40633 members', 'Stars 137k', and 'Follow @siggravitas'. A lightbulb icon is followed by the text 'Get help - Q&amp;A or Discord'. At the bottom of the screenshot, there is a paragraph of text describing the project as an experimental open-source application showcasing the capabilities of the GPT-4 language model.



## 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

```
Auto-GPT-0.3.1 — Python - sudo — 156x36
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 data warehouse to run on Google BigQuery that uses dat
a extracted using Fivetran and landed in raw form into the warehouse staging layer, which you should then transform, integrate and load into a conformed dim
ensional data warehouse. Customer, contact and other common entity records should be deduplicated, and all data should be tested and the package documented.
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 Google BigQuery. It assists analytics engineers in trans
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 help me determine the best approach to build the data warehouse using dbt and Google BigQuery.
PLAN:
- Use the list_files command to list the files in the cloned dbt repository directory.
- Use the read_file command to read the documentation files.
- Review the documentation and files to determine the best approach for designing and building the data warehouse.
- Collaborate with analytics engineers to ensure the data warehouse meets their specific needs and requirements.
- Use dbt and Google BigQuery to 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.
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 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.
THOUGHTS: I need to collaborate with analytics engineers to design and build a data warehouse that meets their specific needs and requirements using dbt and 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 engineers to ensure the data warehouse meets their specific needs and requirements.
```

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

So can ChatGPT replace  
analytics engineers?

## CAN CHATGPT CREATE A DBT PACKAGE?

# How Capable a dbt Developer is ChatGPT?



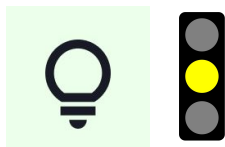
## CAN CHATGPT CREATE A DBT PACKAGE?

# Lifecycle of a dbt / Data Engineering Project

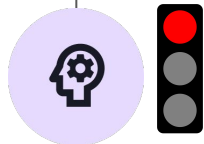
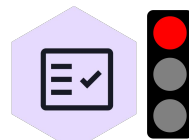
Gather Requirements



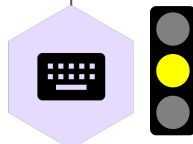
Design Package



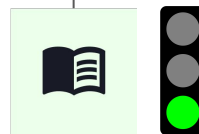
Test & QA Package



Understand Context



Code Package



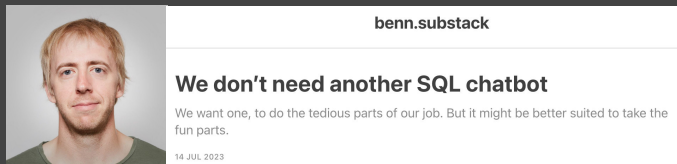
Document Package

## CAN CHATGPT CREATE A DBT PACKAGE?

# No - At best its an assistant, at worst a Stuck Dalek

- Right now, ChatGPT is best thought of as a keen but very junior developer
- Like a graduate who's ingested lots of (now out-of-date) information, and is great at recall
- 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
- But can get things wrong, or sometimes, make things up
- At worse it is a "confident bullshitter", known as "hallucinations" in AI terms
- But...





Because so far, one of the most striking things about LLMs is that they're much better at the creative parts of analysis than they are at the mechanical parts.<sup>8</sup> Ask ChatGPT to write a SQL query against an artificially simple schema; *it's a junior analyst*, at best. But ask it to come up with possible hypotheses to explain why there's some anomaly in a metric, and it does better than I would.<sup>9</sup>



## CAN CHATGPT REPLACE ANALYTICS ENGINEERS

# Yes - For to Those Who Embrace It

- 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
- 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, like Davros from Doctor Who, will be those that embrace AI (Daleks), not reject or fear them





**rittman**  
analytics

Wednesday 30-Minute Webinar Series

# How Rittman Analytics does Product Analytics with Snowplow, dbt and BigQuery



Weds 9th August 4pm UK / 11am ET / 8am PT

[Register Now](#)

presented by Jordan Ilyat  
Analytics Consultant, Rittman Analytics

