This document implements few NLP tasks using GenAI plugins step using Pentaho Data Integration (PDI) which calls OpenAI based LLM. The tests are done via "gpt-4o-mini" model.

Below is a snapshot of various Natural Language Processing (NLP) tasks are categorized as below. Few of these NLP tasks we will address using PDI and LLM (OpenAI).

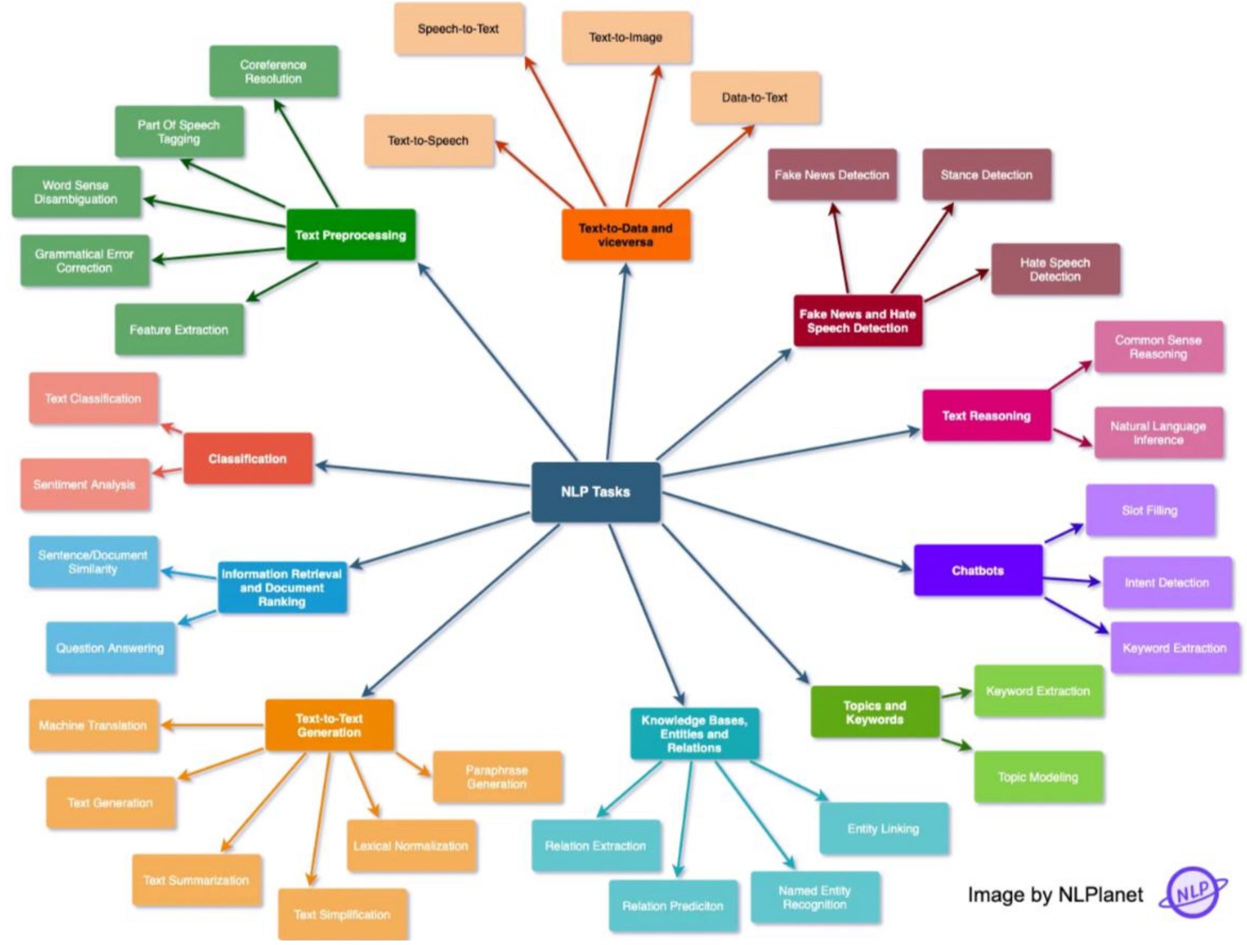
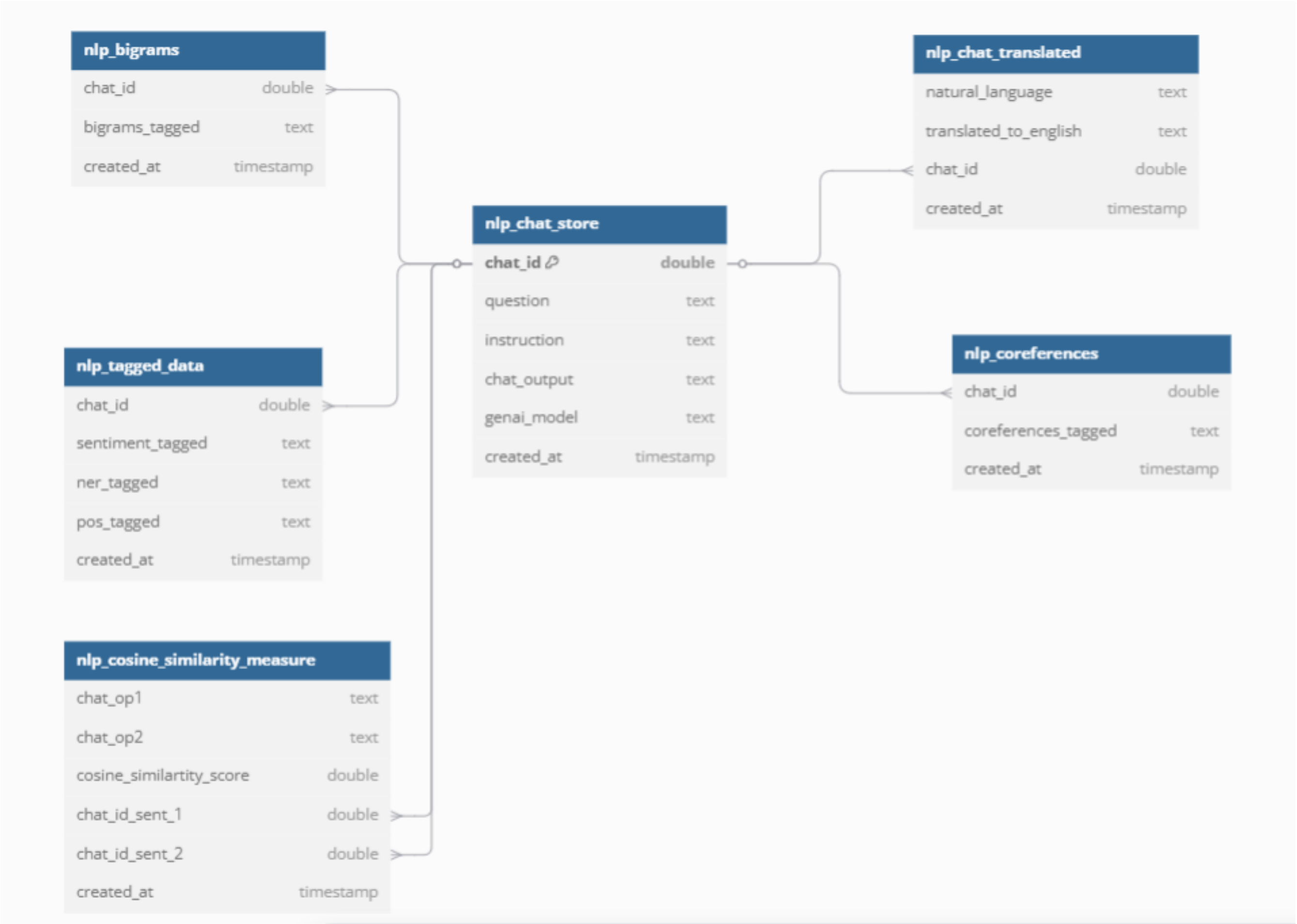


Image by NLPlanet

Data Model:

ERD: https://dbdiagram.io/d/ChatGpt_PDI-67c71518263d6cf9a030e12c



nlp_bigrams	
chat_id	double
bigrams_tagged	text
created_at	timestamp

nlp_chat_translated	
natural_language	text
translated_to_english	text
chat_id	double
created_at	timestamp

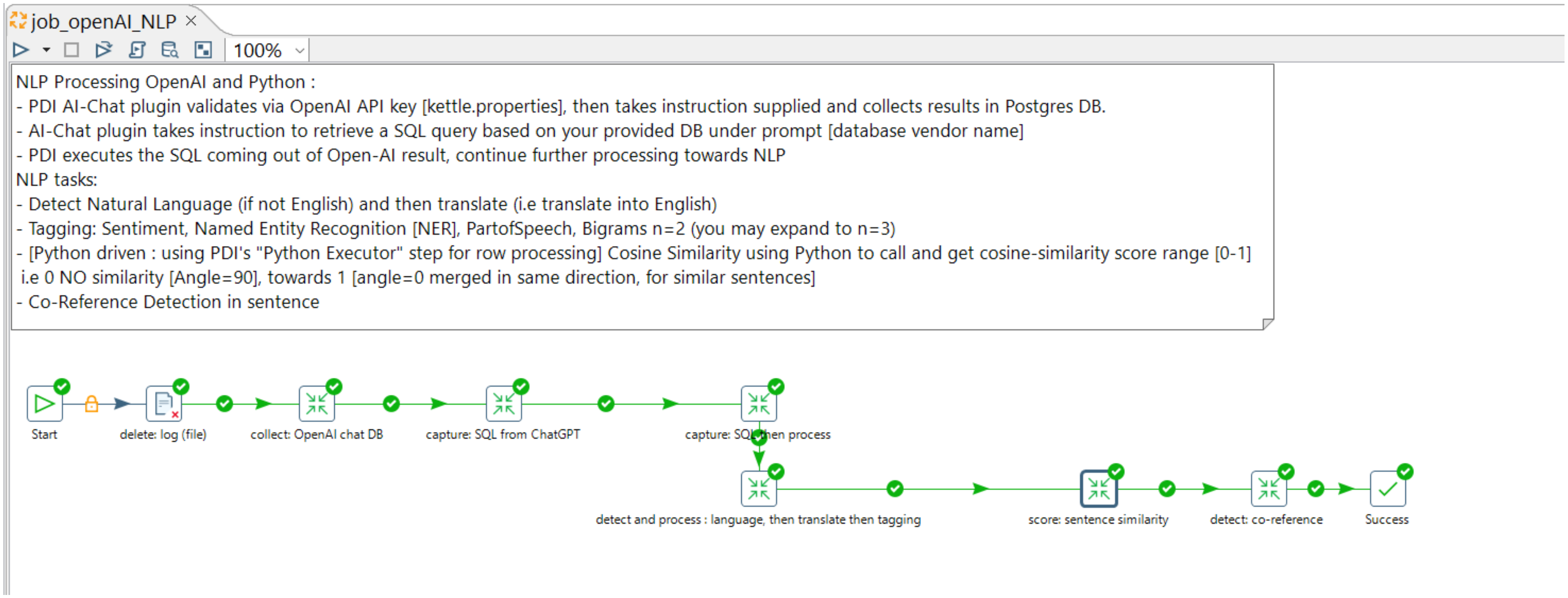
nlp_chat_store	
chat_id	double
question	text
instruction	text
chat_output	text
genai_model	text
created_at	timestamp

nlp_tagged_data	
chat_id	double
sentiment_tagged	text
ner_tagged	text
pos_tagged	text
created_at	timestamp

nlp_coreferences	
chat_id	double
coreferences_tagged	text
created_at	timestamp

nlp_cosine_similarity_measure	
chat_op1	text
chat_op2	text
cosine_similartity_score	double
chat_id_sent_1	double
chat_id_sent_2	double
created_at	timestamp

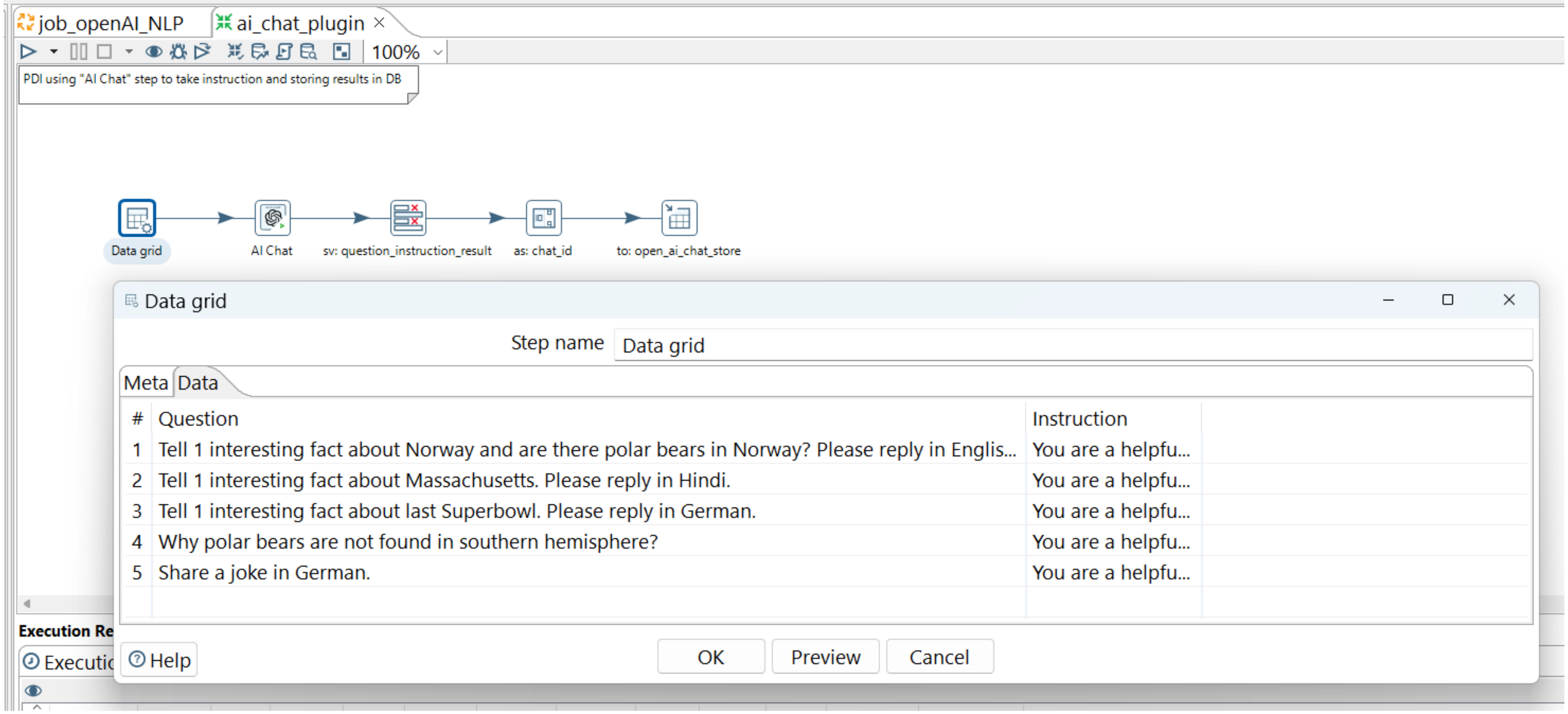
Pentaho Data Integration jobs and transformtations:


job_openAI_NLP
100%
Start
delete: log (file)
collect: OpenAI chat DB
capture: SQL from ChatGPT
capture: SQL then process
detect and process : language, then translate then tagging
score: sentence similarity
detect: co-reference
Success

NLP Processing OpenAI and Python :
- PDI AI-Chat plugin validates via OpenAI API key [kettle.properties], then takes instruction supplied and collects results in Postgres DB.
- AI-Chat plugin takes instruction to retrieve a SQL query based on your provided DB under prompt [database vendor name]
- PDI executes the SQL coming out of Open-AI result, continue further processing towards NLP

NLP tasks:
- Detect Natural Language (if not English) and then translate (i.e translate into English)
- Tagging: Sentiment, Named Entity Recognition [NER], PartofSpeech, Bigrams n=2 (you may expand to n=3)
- [Python driven : using PDI's "Python Executor" step for row processing] Cosine Similarity using Python to call and get cosine-similarity score range [0-1] i.e 0 NO similarity [Angle=90], towards 1 [angle=0 merged in same direction, for similar sentences]
- Co-Reference Detection in sentence

#Generate Data (multiple natural languages) using OpenAI and store in a database for processing NLP

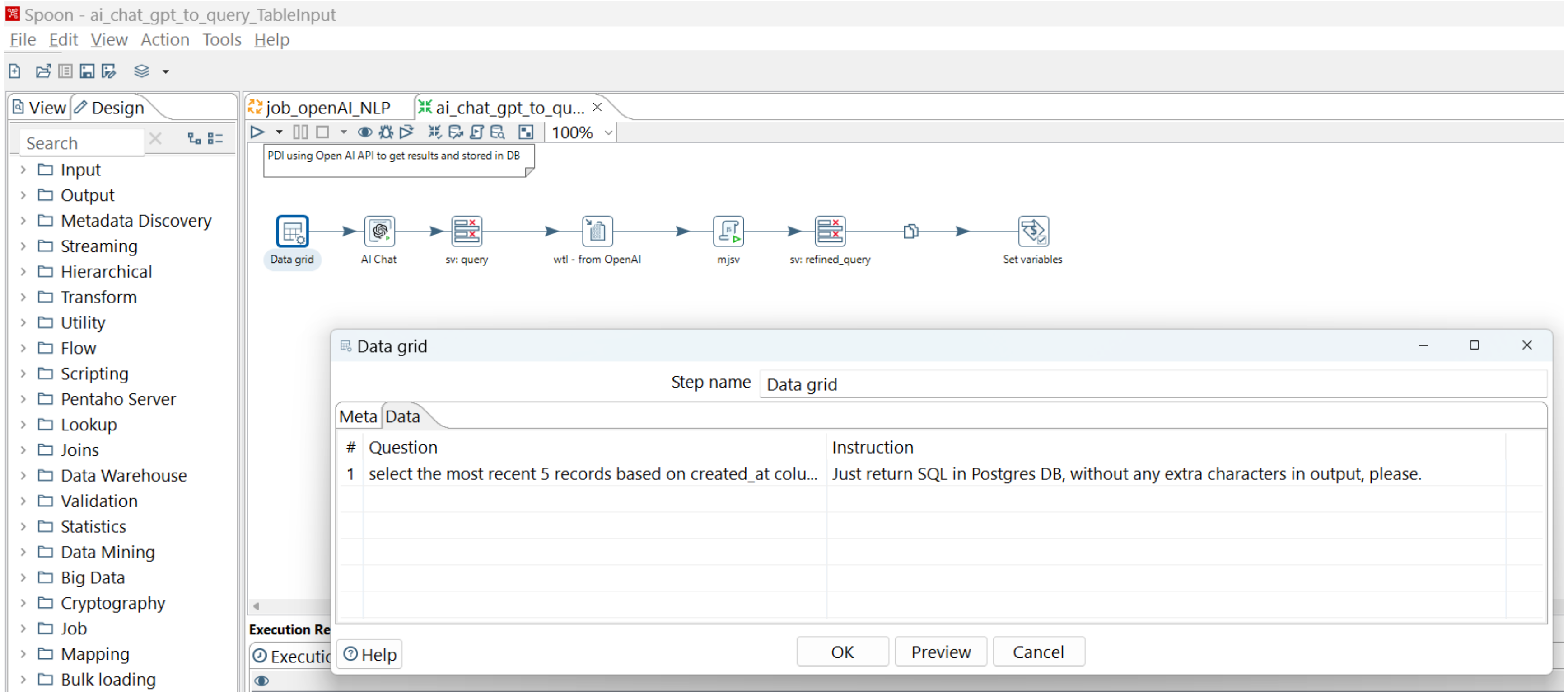


job_openAI_NLP
ai_chat_plugin
100%
PDI using "AI Chat" step to take instruction and storing results in DB
Data grid
AI Chat
sv: question_instruction_result
as: chat_id
to: open_ai_chat_store
Execution Re
Executic

Data grid

Step name: Data grid

#	Question	Instruction
1	Tell 1 interesting fact about Norway and are there polar bears in Norway? Please reply in Englis...	You are a helpfu...
2	Tell 1 interesting fact about Massachusetts. Please reply in Hindi.	You are a helpfu...
3	Tell 1 interesting fact about last Superbowl. Please reply in German.	You are a helpfu...
4	Why polar bears are not found in southern hemisphere?	You are a helpfu...
5	Share a joke in German.	You are a helpfu...

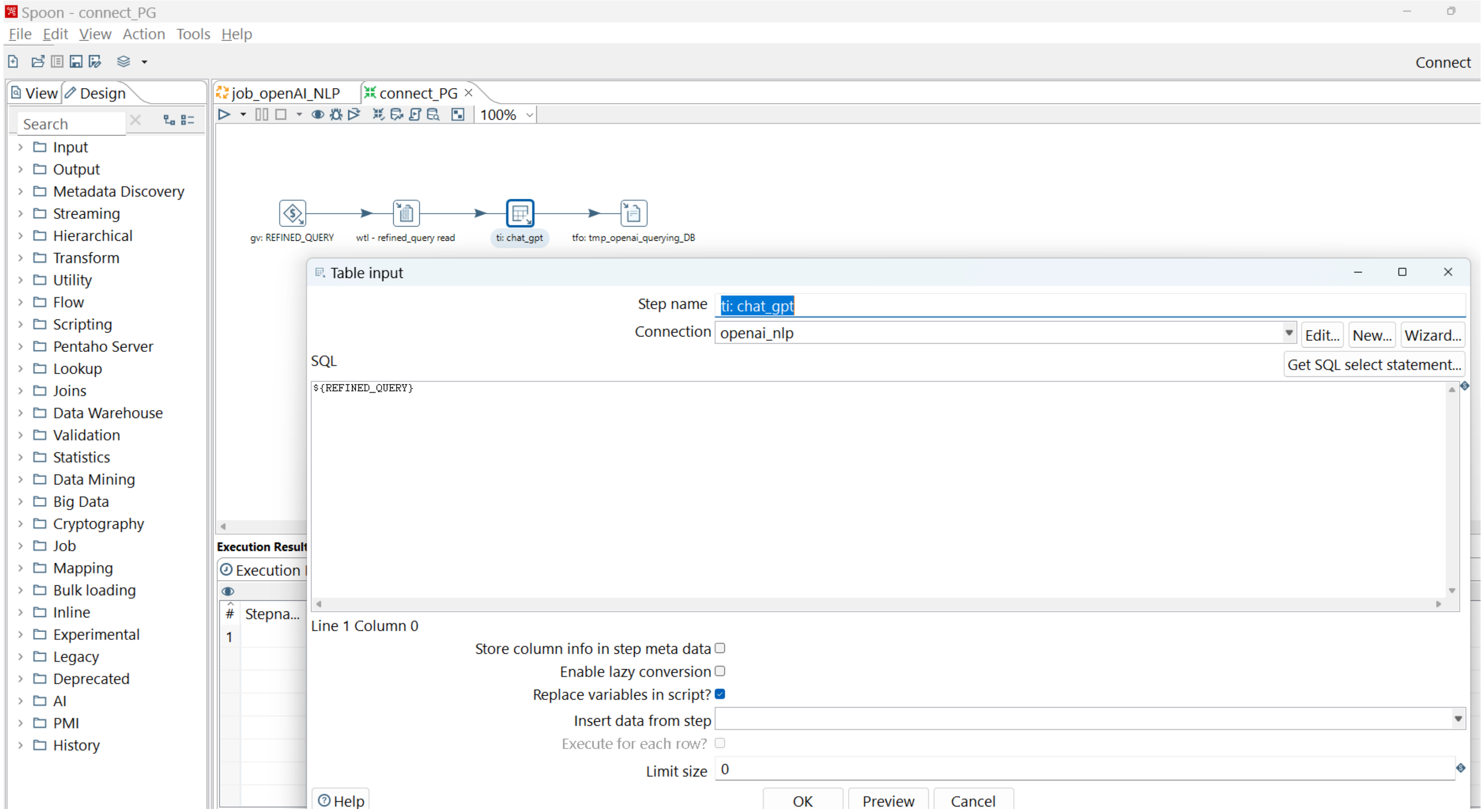


PDI using Open AI API to get results and stored in DB

Step name: Data grid

#	Question	Instruction
1	select the most recent 5 records based on created_at colu...	Just return SQL in Postgres DB, without any extra characters in output, please.

#Testing to see SQL works



Step name: ti: chat_gpt

Connection: openai_nlp

SQL

```
${REFINED_QUERY}
```

Line 1 Column 0

Store column info in step meta data □

Enable lazy conversion □

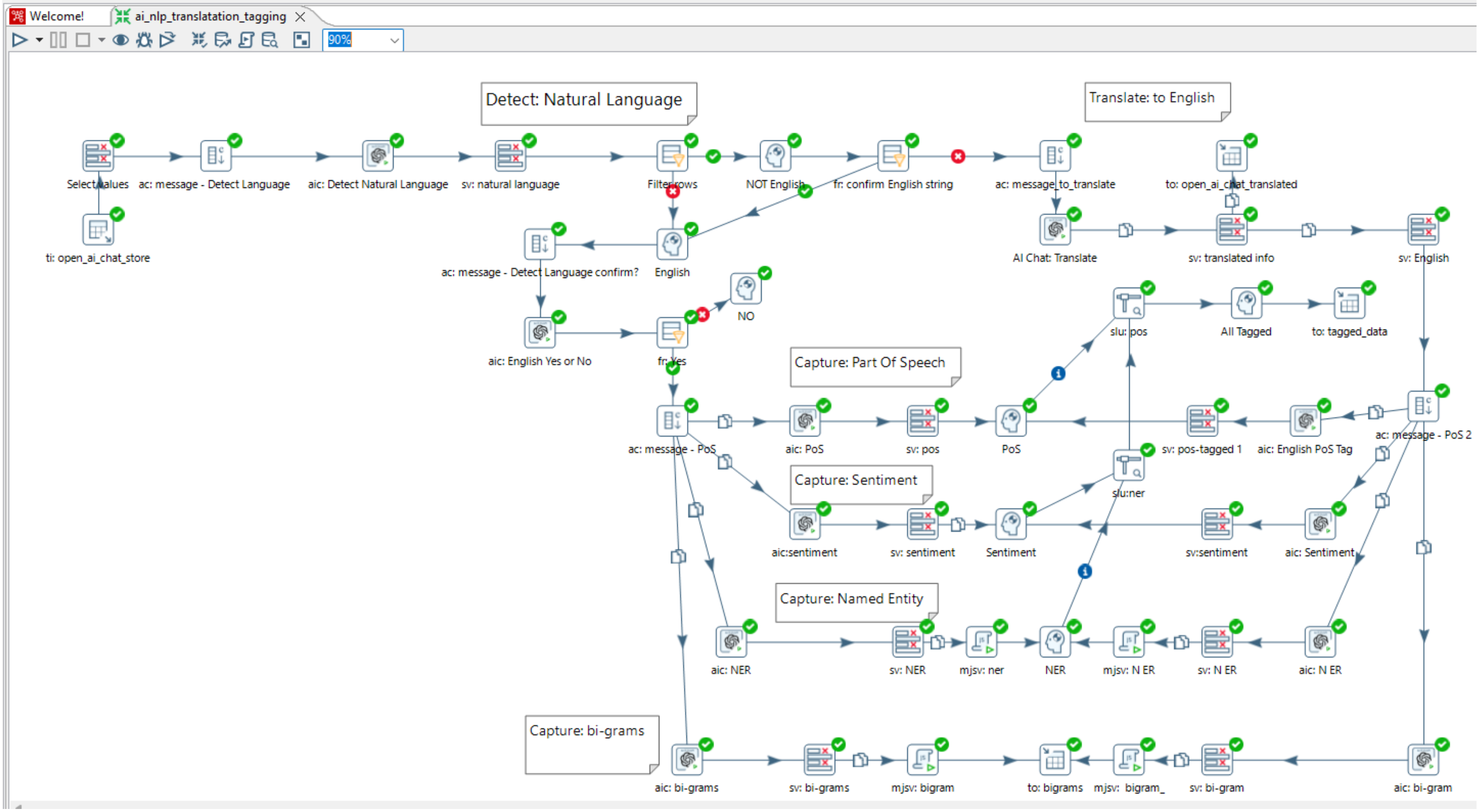
Replace variables in script? ☑

Insert data from step

Execute for each row? □

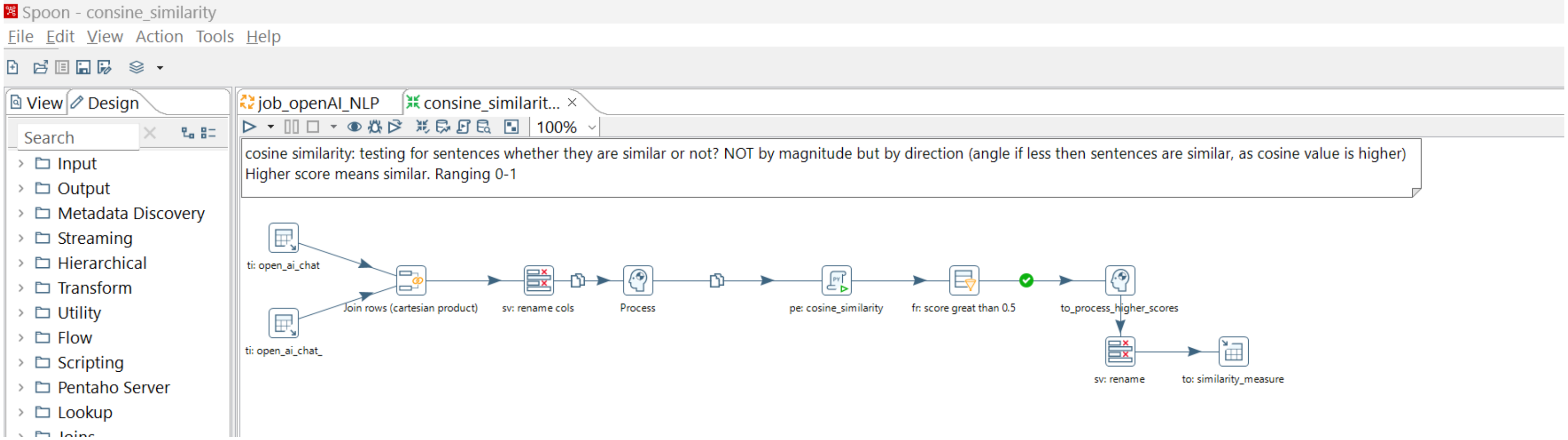
Limit size 0

#Translation, Sentiment, NER, bigrams, PoS



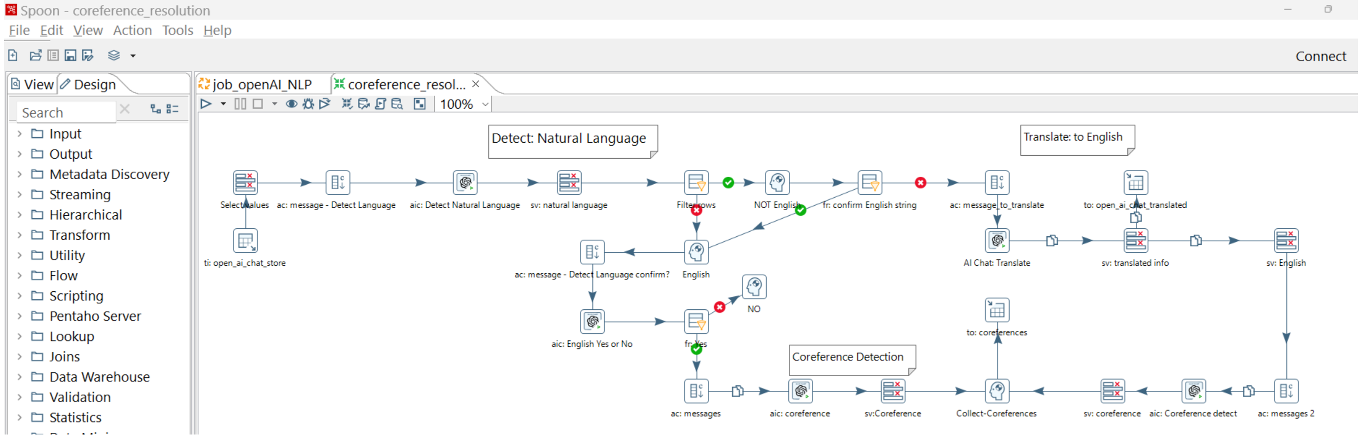
#Sentence Similarity (How similar are two sentences?)

ChatGPT couldn’t provide a score through prompt, so having this via Python while using data processing in PDI transformation (data in flight).



cosine similarity: testing for sentences whether they are similar or not? NOT by magnitude but by direction (angle if less then sentences are similar, as cosine value is higher)
Higher score means similar. Ranging 0-1

--Co-Reference


Spoon - coreference_resolution
File Edit View Action Tools Help
Connect
View
Design
Search
Input
Output
Metadata Discovery
Streaming
Hierarchical
Transform
Utility
Flow
Scripting
Pentaho Server
Lookup
Joins
Data Warehouse
Validation
Statistics
job_openAI_NLP
coreference_resol...
100%
Detect: Natural Language
Translate: to English
Coreference Detection
Select values
ti: open_ai_chat_store
ac: message - Detect Language
aic: Detect Natural Language
sv: natural language
Filter rows
NOT English
fr: confirm English string
ac: message_to_translate
to: open_ai_chat_translated
AI Chat: Translate
sv: translated info
sv: English
English
ac: message - Detect Language confirm?
NO
aic: English Yes or No
fr: Yes
to: coreferences
ac: messages
aic: coreference
sv:Coreference
Collect-Coreferences
sv: coreference
aic: Coreference detect
ac: messages 2