The Pneuma Project:
Reifying Information Needs as Relational Schemas to Automate Discovery, Guide Preparation, and Align Data with Intent

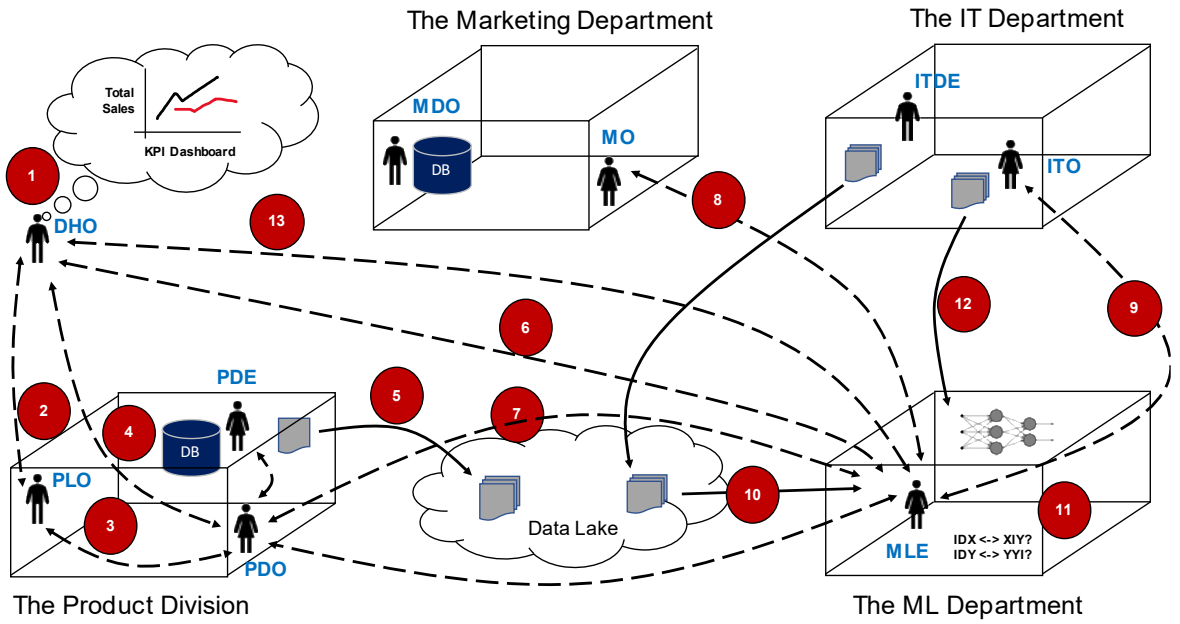
Luthfi Balaka, Raul Castro Fernandez

(and Danni Liu, John Hillestand, Kemal Badur, and many more)

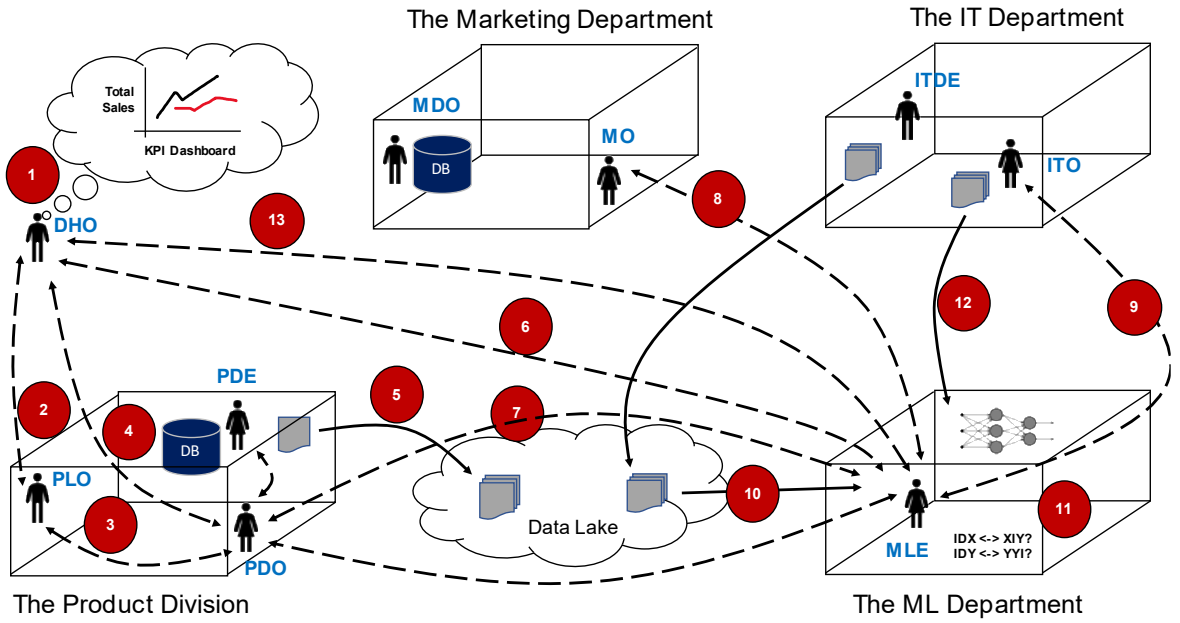
THE UNIVERSITY OF CHICAGO

Data Management's Persistent Problem

Data Management's Persistent Problem

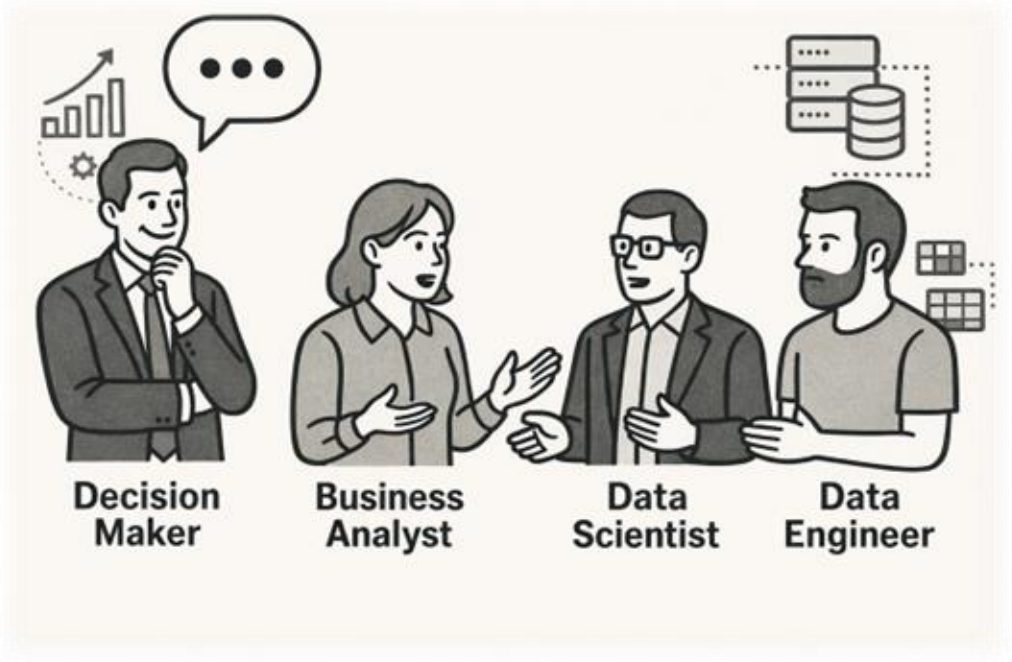


Data Management's Persistent Problem

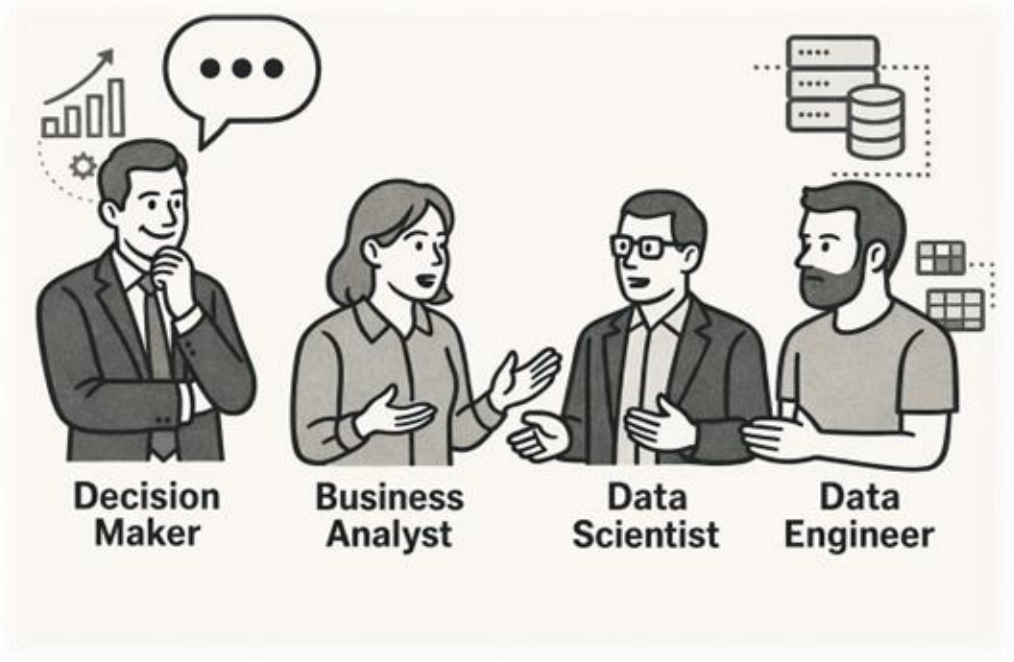


Answer questions over large volumes of data with unfamiliar data models

From Questions to Answers (Today)



From Questions to Answers (Today)

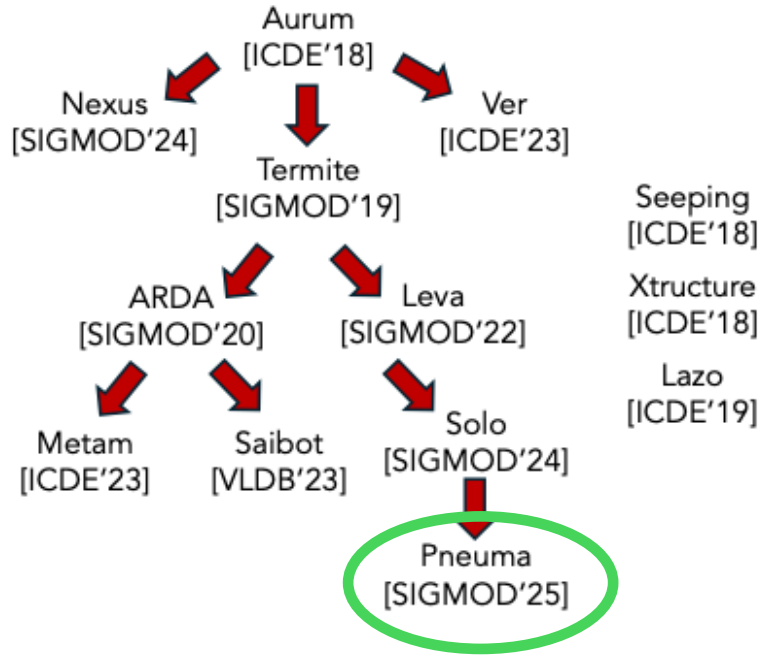


- Distributed “tribal” knowledge calls for collaboration across multiple parties…
- …that’s time-consuming, non-scalable, error-prone…

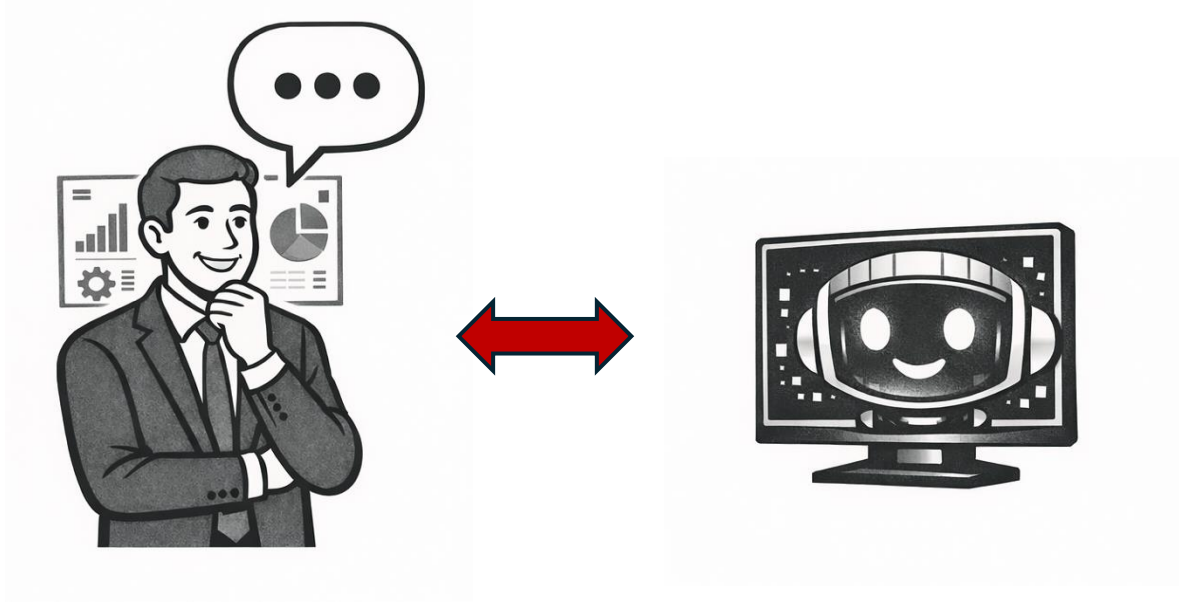
Helping Users Navigate Data

- Data catalogs, search systems, better BI tools…

Helping Users Navigate Data



The North Star: Task Delegation



- Delegate data task to an AI agent
- Interact via text-based interfaces

The Main Roadblocks

1. **Questions are moving targets** and difficult to describe
   - The *question* changes as users explore the data

The Main Roadblocks

1. **Questions are moving targets** and difficult to describe
   - The question changes as users explore the data
2. **Unclear system's objective**
   - Without a goalpost, agents cannot progress reliably

The Main Roadblocks

1. **Questions are moving targets** and difficult to describe
   - The question changes as users explore the data
2. **Unclear system's objective**
   - Without a goalpost, agents cannot progress reliably
3. **People are useful! They possess "tribal knowledge"**
   - Disambiguate meaning
   - Reconcile data models
   - Guide users in the right direction

The Main Roadblocks

1. **Questions are moving targets** and difficult to describe
   - The question changes as users explore the data
2. **Unclear system's objective**
   - Without a goalpost, agents cannot progress reliably
3. **People are useful! They possess "tribal knowledge"**
   - Disambiguate meaning
   - Reconcile data models
   - Guide users in the right direction

Pneuma’s Approach: Reifying Info Needs

Pneuma's Approach: Reifying Info Needs

What's the *ideal* table?

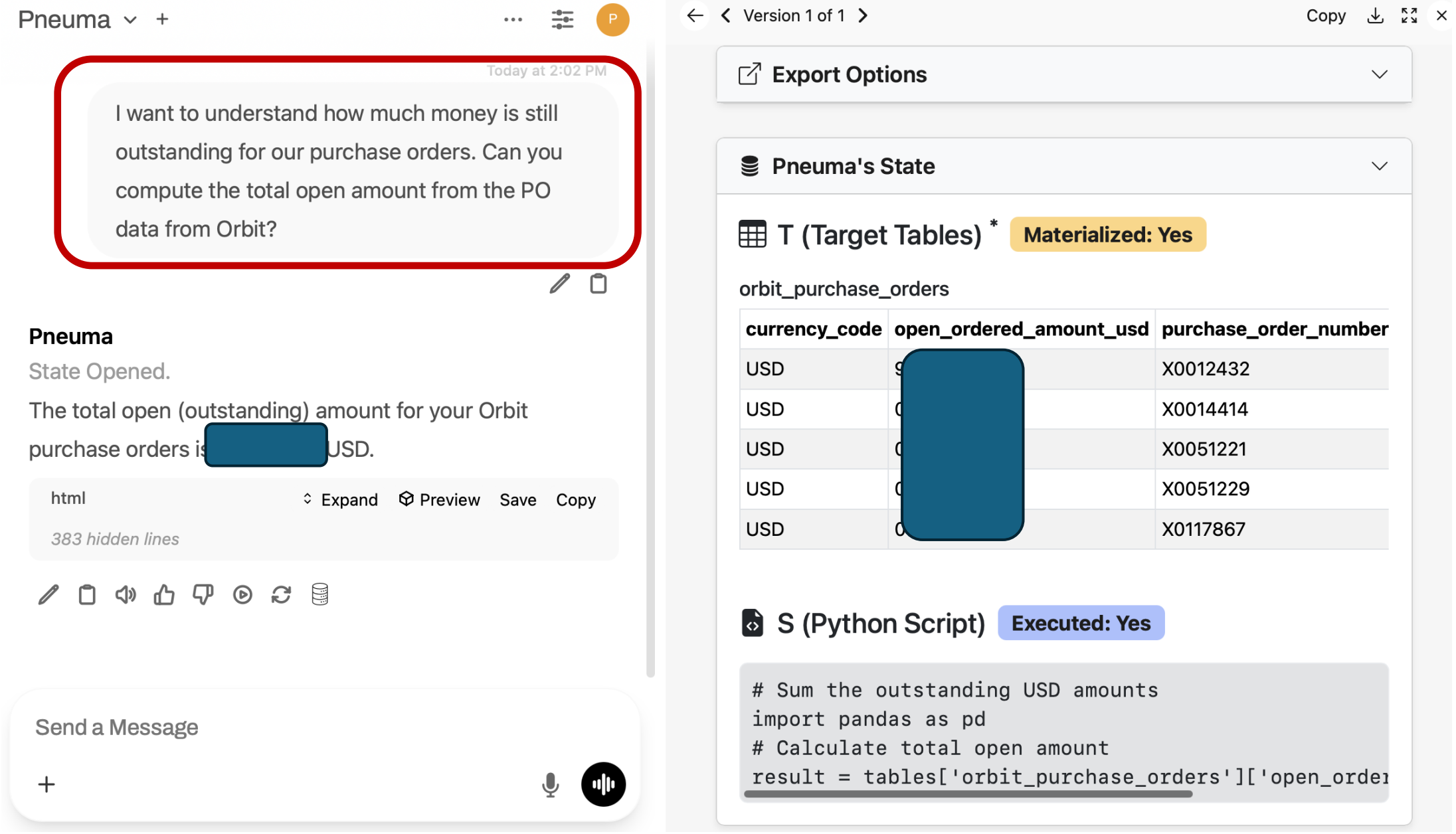
Construct the *ideal* table

Iterate on ***ideal representation*** until it satisfies the user's information need

Example: Procurement Office at UChicago

Procurement Database Use Case

- Jaggaer data model
  - ~50 tables
- Extracts from a variety of other related databases


Pneuma
Today at 2:02 PM
I want to understand how much money is still outstanding for our purchase orders. Can you compute the total open amount from the PO data from Orbit?
Pneuma
State Opened.
The total open (outstanding) amount for your Orbit purchase orders is [redacted] USD.
html
Expand
Preview
Save
Copy
383 hidden lines
Send a Message
Version 1 of 1
Copy
Export Options
Pneuma's State
T (Target Tables) *
Materialized: Yes
orbit_purchase_orders
currency_code
open_ordered_amount_usd
purchase_order_number
USD
X0012432
USD
X0014414
USD
X0051221
USD
X0051229
USD
X0117867
S (Python Script)
Executed: Yes
# Sum the outstanding USD amounts
import pandas as pd
# Calculate total open amount
result = tables['orbit_purchase_orders']['open_order

Pneuma

Today at 2:02 PM

I want to understand how much money is still outstanding for our purchase orders. Can you compute the total open amount from the PO data from Orbit?

Pneuma

State Opened.

The total open (outstanding) amount for your Orbit purchase orders is [redacted] USD.

html — Expand · Preview · Save · Copy

383 hidden lines

Send a Message

Version 1 of 1 — Copy

Export Options

Pneuma's State

T (Target Tables) * — Materialized: Yes

orbit_purchase_orders

currency_code	open_ordered_amount_usd	purchase_order_number
USD	[redacted]	X0012432
USD	[redacted]	X0014414
USD	[redacted]	X0051221
USD	[redacted]	X0051229
USD	[redacted]	X0117867

S (Python Script) — Executed: Yes

```
# Sum the outstanding USD amounts
import pandas as pd
# Calculate total open amount
result = tables['orbit_purchase_orders']['open_order
```

Pneuma

Today at 2:02 PM

I want to understand how much money is still outstanding for our purchase orders. Can you compute the total open amount from the PO data from Orbit?

Pneuma

State Opened.

The total open (outstanding) amount for your Orbit purchase orders is [redacted] USD.

html Expand Preview Save Copy

383 hidden lines

Send a Message

Version 1 of 1 Copy

Export Options

Pneuma's State

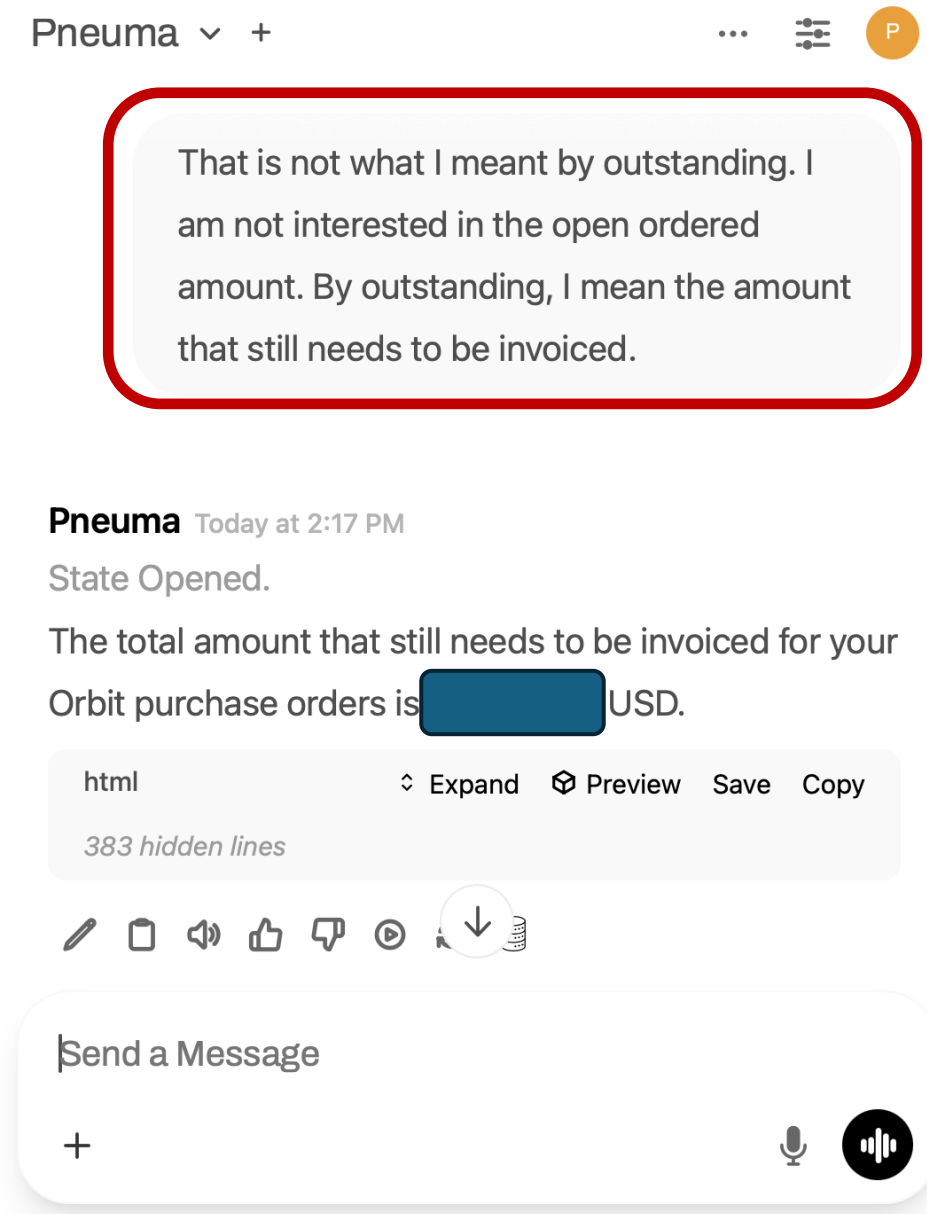
T (Target Tables) * Materialized: Yes

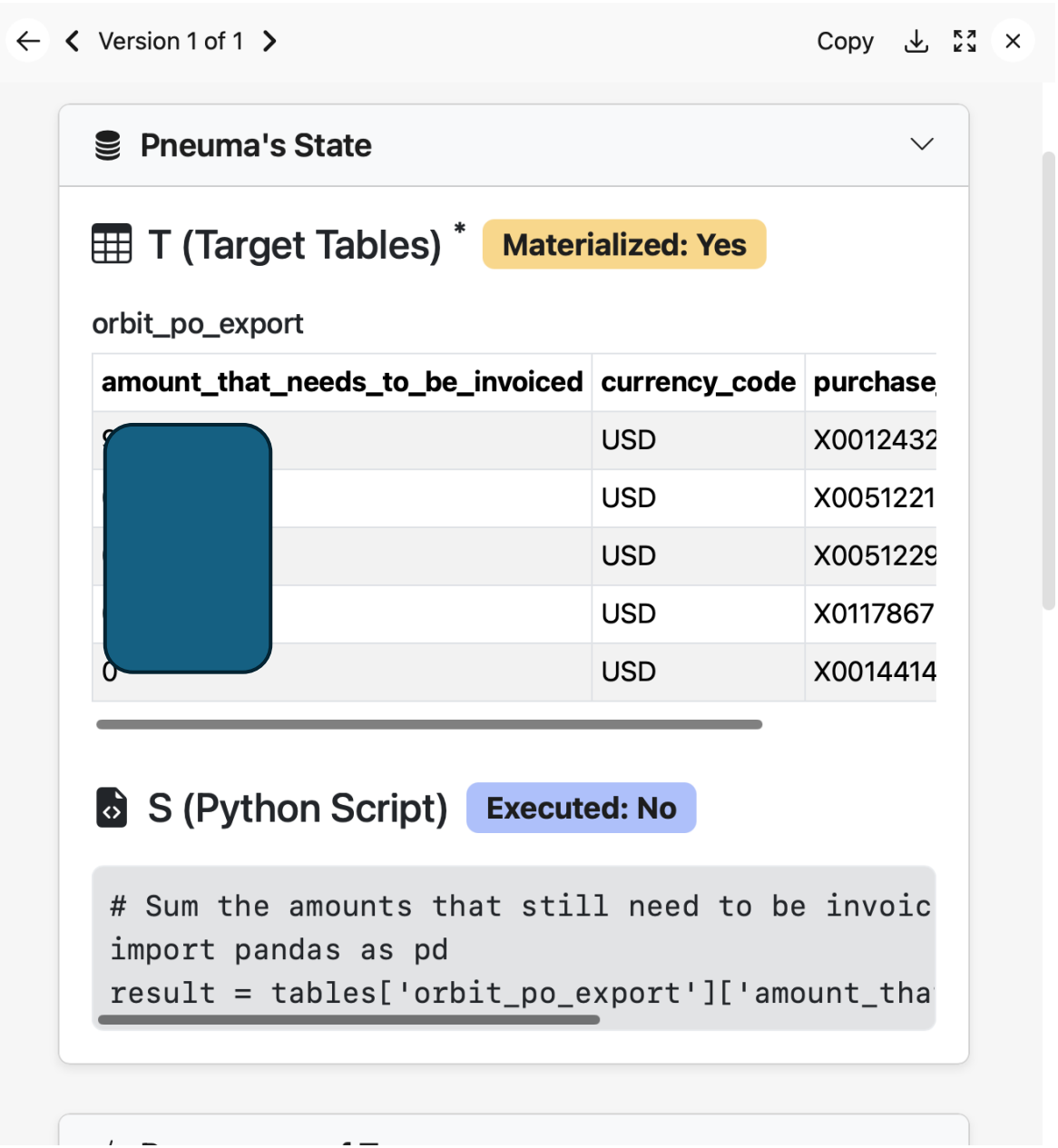
orbit_purchase_orders

currency_code	open_ordered_amount_usd	purchase_order_number
USD	[redacted]	X0012432
USD		X0014414
USD		X0051221
USD		X0051229
USD		X0117867

S (Python Script) Executed: Yes

```
# Sum the outstanding USD amounts
import pandas as pd
# Calculate total open amount
result = tables['orbit_purchase_orders']['open_order
```


Pneuma
That is not what I meant by outstanding. I am not interested in the open ordered amount. By outstanding, I mean the amount that still needs to be invoiced.
Pneuma Today at 2:17 PM
State Opened.
The total amount that still needs to be invoiced for your Orbit purchase orders is [redacted] USD.
html
Expand
Preview
Save
Copy
383 hidden lines
Send a Message


Version 1 of 1
Copy
Pneuma's State
T (Target Tables) *
Materialized: Yes
orbit_po_export
amount_that_needs_to_be_invoiced
currency_code
purchase
[redacted]
USD
X0012432
USD
X0051221
USD
X0051229
USD
X0117867
0
USD
X0014414
S (Python Script)
Executed: No
# Sum the amounts that still need to be invoic
import pandas as pd
result = tables['orbit_po_export']['amount_tha

Pneuma

That is not what I meant by outstanding. I am not interested in the open ordered amount. By outstanding, I mean the amount that still needs to be invoiced.

Pneuma Today at 2:17 PM

State Opened.

The total amount that still needs to be invoiced for your Orbit purchase orders is [redacted] USD.

html Expand Preview Save Copy

383 hidden lines

Send a Message

Version 1 of 1 Copy

Pneuma's State

T (Target Tables) * Materialized: Yes

orbit_po_export

amount_that_needs_to_be_invoiced	currency_code	purchase
[redacted]	USD	X0012432
	USD	X0051221
	USD	X0051229
	USD	X0117867
0	USD	X0014414

S (Python Script) Executed: No

```
# Sum the amounts that still need to be invoic
import pandas as pd
result = tables['orbit_po_export']['amount_tha
```

Content

That is not what I meant by outstanding. I am not interested in the open ordered amount. By outstanding, I mean the amount that still needs to be invoiced.

Pneuma

State Opened.

The total amount that still needs to be invoiced for your Orbit purchase orders is [redacted] SD.

html Expand Preview Save Copy

383 hidden lines

Send a Message

Version 1 of 1 Copy

Provenance of T

This document describes how **T** was generated, including all source data and processing steps. You can download the script here.

Used Data

The following source data was used:

```
tables["fy25_oracle_po_export_from_orbit"] = pd.read_cs
```

Data Processing Steps

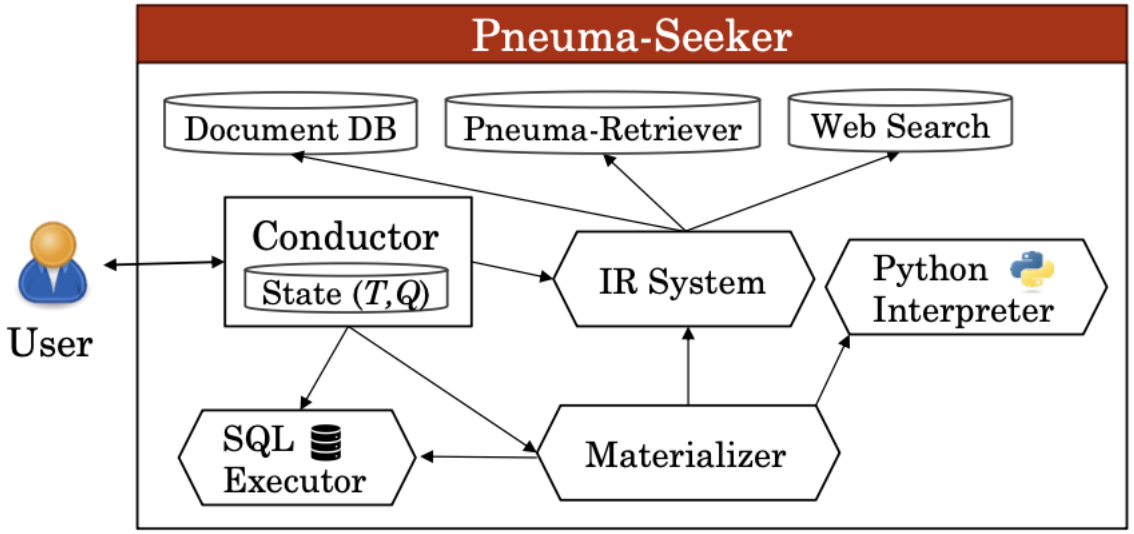
The following processing steps were applied to the source data to generate **T**:

Step 1: Directly selects a table (ID: `fy25_oracle_po_export_from_orbit`; columns: ['purchase_order_number', 'currency_code', 'amount_that_needs_to_be_invoiced']) to form a target table: `orbit_po_export`.

```
orbit_po_export = tables["fy25_oracle_po_export_from_o
```

Technical Contributions

- Reified information need kills two birds with one stone
- Context specialization
- Dynamic Planning with Conductor



The Main Roadblocks

1. **Questions are moving targets** and difficult to describe
   - The question changes as users explore the data
2. **Unclear system's objective**
   - Without a goalpost, agents cannot progress reliably
3. **People are useful! They possess "tribal knowledge"**
   - Disambiguate meaning
   - Reconcile data models
   - Guide users in the right direction

Internal Data Markets

The central argument of this paper is that sharing, dis- covering, and integrating data is hard because data owners lack information and incentives to make their data available in a way that increases consumers' utility.

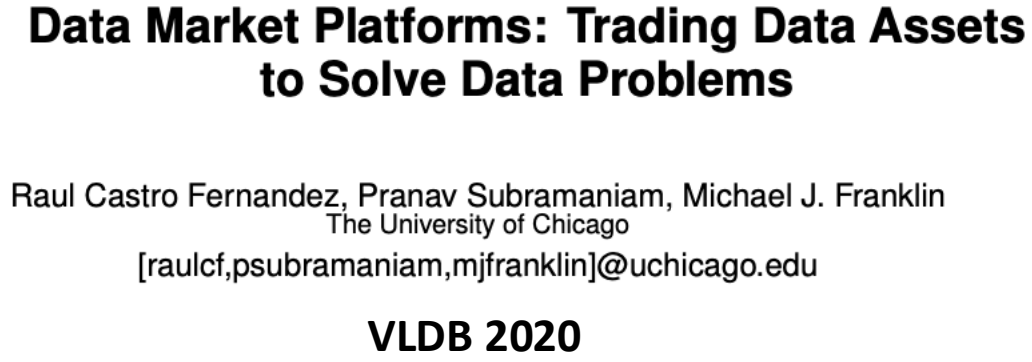
Data Market Platforms: Trading Data Assets to Solve Data Problems

Raul Castro Fernandez, Pranav Subramaniam, Michael J. Franklin
The University of Chicago
[raulcf,psubramaniam,mjfranklin]@uchicago.edu

VLDB 2020

Internal Data Markets

- *Supply* of tribal knowledge
- *Demand* for it
- *Match* supply with demand
  - Allocate data where it's most useful



Data Market Platforms: Trading Data Assets to Solve Data Problems

Raul Castro Fernandez, Pranav Subramaniam, Michael J. Franklin
The University of Chicago
[raulcf,psubramaniam,mjfranklin]@uchicago.edu

VLDB 2020

Internal Data Markets

- *Supply* of tribal knowledge
- *Demand* for it
- *Match* supply with demand
  - Allocate data where it's most useful
- Developed techniques to design internal markets
- Socio-technical bottleneck: how do we get people to "supply" tribal knowledge?
  - Standards? Internal documentation? Associate it with bonus?

Data Market Platforms: Trading Data Assets to Solve Data Problems

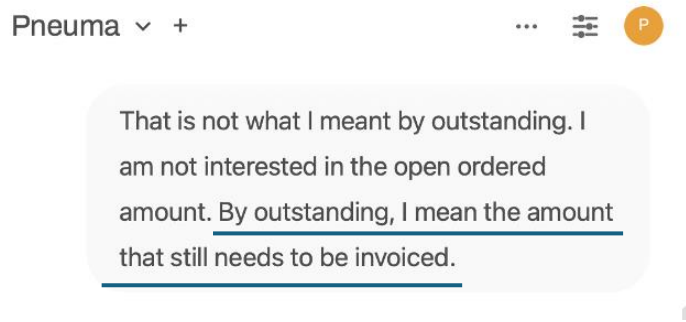
Raul Castro Fernandez, Pranav Subramaniam, Michael J. Franklin
The University of Chicago
[raulcf,psubramaniam,mjfranklin]@uchicago.edu

VLDB 2020

Pneuma enables Internal Data Markets

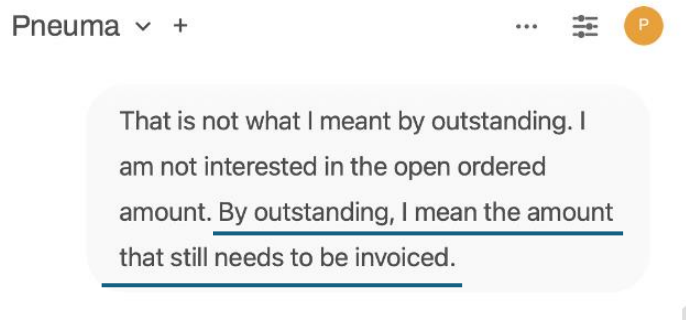
Pneuma enables Internal Data Markets

- Users have an **incentive to describe what data they need**, for what purpose, and to spell out carefully their **assumptions** about their question, data, and **business domain**.



Pneuma enables Internal Data Markets

- Users have an **incentive to describe what data they need**, for what purpose, and to spell out carefully their **assumptions** about their question, data, and **business domain**.



- Users **transfer their tribal knowledge** to a system that's learning from everyone, capturing the organization's ***knowhow*** and **allocating it where valuable**.

The Pneuma Project at UChicago

- Pneuma as an enabler of internal data markets
  - *Capture* and *deploy* tribal knowledge
- Pneuma system in use by procurement
  - Onboarding “advancement”, preparing focus groups
- ***We are onboarding interested users, reach out!***

Raul Castro Fernandez

raulcf@uchicago.edu

The University of Chicago