vectara

Vectara Workshop for TI Hackathon

Feb 28, 2025

Workshop Agenda

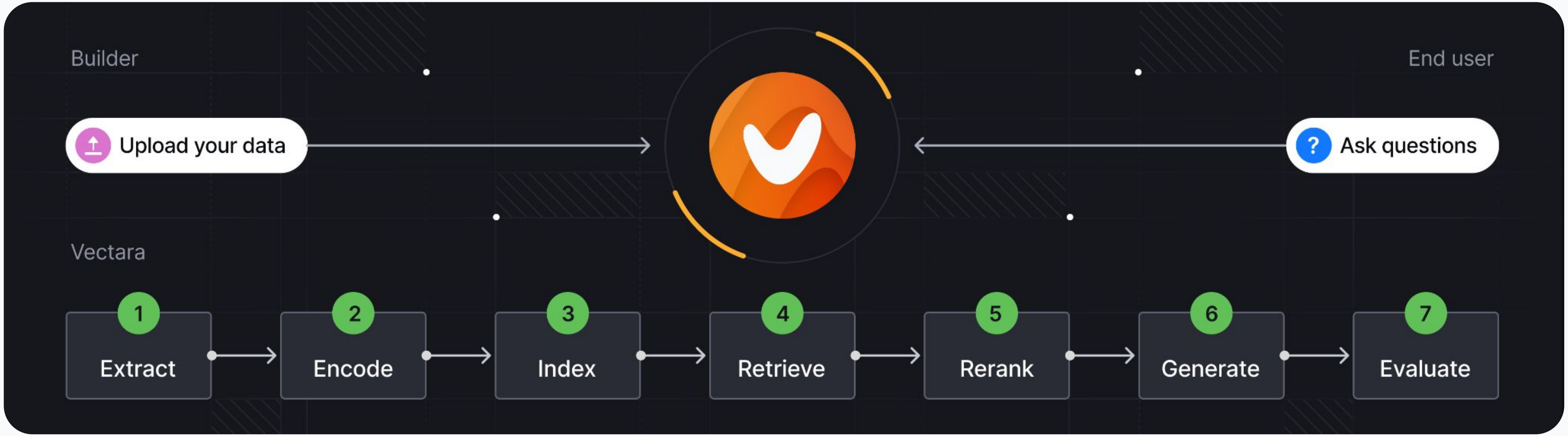
1. **Vectara Concepts:** Presentation, 20 mins.
2. **Demo Apps & Console Tour:** Presentation, 15 mins.
3. **Exercise 1 - Account Setup, Q&A App Prototype:** Hands On, 30 mins.
4. ***Break:** Chai/Coffee, 5 mins.*
5. **App Development:** Presentation, 30 mins.
6. **Exercise 2 - SDK/API:** Hands On, 30 mins.
7. ***Break:** Chai/Coffee, 5 mins.*
8. **Agentic RAG:** Presentation, 15 mins.
9. **Exercise 3 - Agentic App:** Hands On, 20 mins.

Vectara Concepts

(20 mins.)

End-to-end platform, with proprietary AI models

*Platform to **rapidly** create **trustworthy** AI Assistants & AI Agents that are **grounded on your data***



- **Retrieval model** builds vectors representing data’s meaning, to match with a query’s intent, regardless of language.
- **Reranker models** evaluate search results and reorder them based on relevance, diversity, or custom rules.
- **RAG-optimized generative LLM** (or an OpenAI LLM) generates output using retrieved facts.
- **Guardrails model** scores output for risk of hallucination, to enable blocking non-factual responses from users.

** Top hallucination detection model on Hugging Face*

Vectara Use Cases

AI Assistants

Question Answering • Conversational AI

Vectara's best-in-class retrieval, superior cross-language operation, chat history, and multi-turn generation enable AI Assistants to go many layers deep with high accuracy and minimal hallucinations.

AI Agents

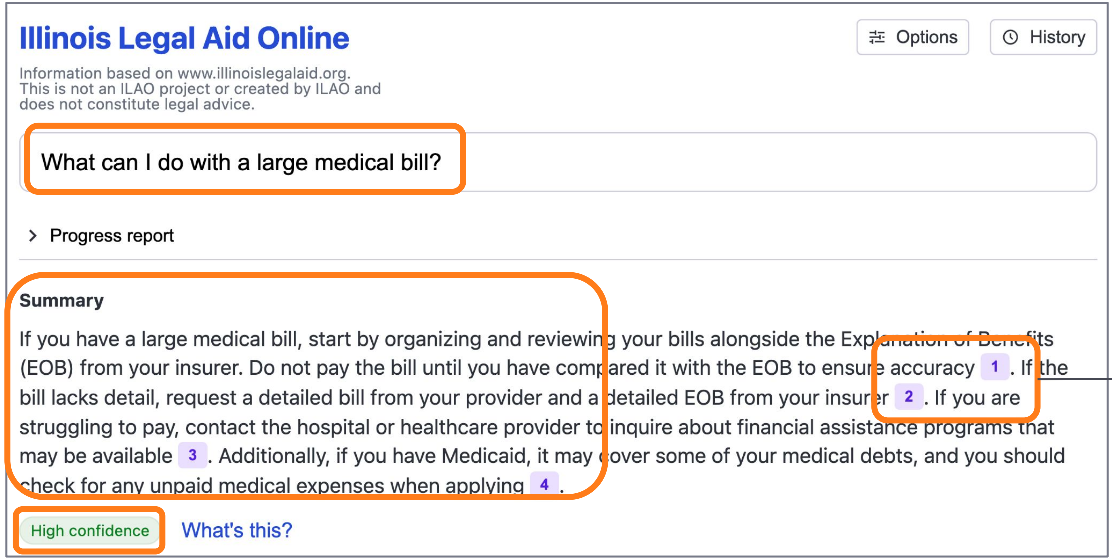
Semantic App Search • Action Engines

Vectara's best in class retrieval, enhanced configurability of generation, and API extensibility and scaffolding help turn answers into actions directly in your application with minimal effort.

Vectara Use Cases

Question Answering AI Assistant

User Input - Expressed Naturally, in Any Language



Illinois Legal Aid Online

Information based on www.illinoislegalaid.org.
This is not an ILAO project or created by ILAO and does not constitute legal advice.

What can I do with a large medical bill?

Summary

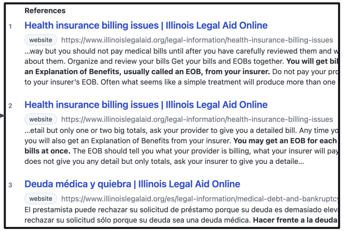
If you have a large medical bill, start by organizing and reviewing your bills alongside the Explanation of Benefits (EOB) from your insurer. Do not pay the bill until you have compared it with the EOB to ensure accuracy 1. If the bill lacks detail, request a detailed bill from your provider and a detailed EOB from your insurer 2. If you are struggling to pay, contact the hospital or healthcare provider to inquire about financial assistance programs that may be available 3. Additionally, if you have Medicaid, it may cover some of your medical debts, and you should check for any unpaid medical expenses when applying 4.

High confidence

Summarized Answer

Score indicating trustworthiness

References Provide Explainability



References

1 Health insurance billing issues | Illinois Legal Aid Online

website https://www.illinoislegalaid.org/legal-information/health-insurance-billing-issues

...way but you should not pay medical bills until after you have carefully reviewed them and w about them. Organize and review your bills Get your bills and EOBs together. **You will get bil an Explanation of Benefits, usually called an EOB, from your insurer.** Do not pay your pro to your insurer's EOB. Often what seems like a simple treatment will produce more than one

2 Health insurance billing issues | Illinois Legal Aid Online

website https://www.illinoislegalaid.org/legal-information/health-insurance-billing-issues

...etail but only one or two big totals, ask your provider to give you a detailed bill. Any time yo you will also get an Explanation of Benefits from your insurer. **You may get an EOB for each bills at once.** The EOB should tell you what your provider is billing, what your insurer will pay does not give you any detail but only totals, ask your insurer to give you a detaile...

3 Deuda médica y quiebra | Illinois Legal Aid Online

website https://www.illinoislegalaid.org/es/legal-information/medical-debt-and-bankruptcy

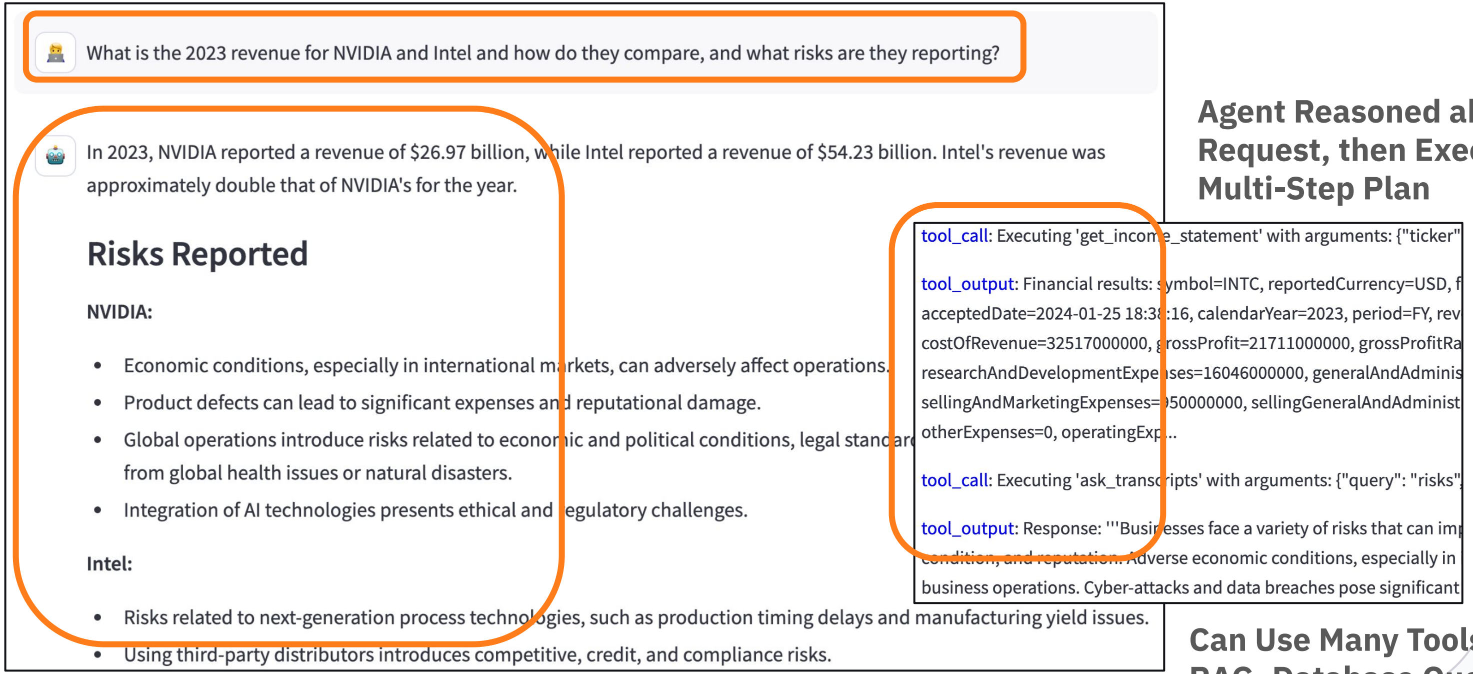
El prestamista puede rechazar su solicitud de préstamo porque su deuda es demasiado elev rechazar su solicitud sólo porque su deuda sea una deuda médica. **Hacer frente a la deuda**

Vectara Use Cases

Conversational AI Assistant

Complex Request or Task, in Any Language

Complete Response / Action



What is the 2023 revenue for NVIDIA and Intel and how do they compare, and what risks are they reporting?

In 2023, NVIDIA reported a revenue of $26.97 billion, while Intel reported a revenue of $54.23 billion. Intel's revenue was approximately double that of NVIDIA's for the year.

Risks Reported

NVIDIA:

- Economic conditions, especially in international markets, can adversely affect operations.
- Product defects can lead to significant expenses and reputational damage.
- Global operations introduce risks related to economic and political conditions, legal standar from global health issues or natural disasters.
- Integration of AI technologies presents ethical and regulatory challenges.

Intel:

- Risks related to next-generation process technologies, such as production timing delays and manufacturing yield issues.
- Using third-party distributors introduces competitive, credit, and compliance risks.

```
tool_call: Executing 'get_income_statement' with arguments: {"ticker"

tool_output: Financial results: symbol=INTC, reportedCurrency=USD, f
acceptedDate=2024-01-25 18:38:16, calendarYear=2023, period=FY, rev
costOfRevenue=32517000000, grossProfit=21711000000, grossProfitRa
researchAndDevelopmentExpenses=16046000000, generalAndAdminis
sellingAndMarketingExpenses=950000000, sellingGeneralAndAdminist
otherExpenses=0, operatingExp...

tool_call: Executing 'ask_transcripts' with arguments: {"query": "risks",

tool_output: Response: '''Businesses face a variety of risks that can imp
condition, and reputation. Adverse economic conditions, especially in
business operations. Cyber-attacks and data breaches pose significant
```

Agent Reasoned about Request, then Executed Multi-Step Plan

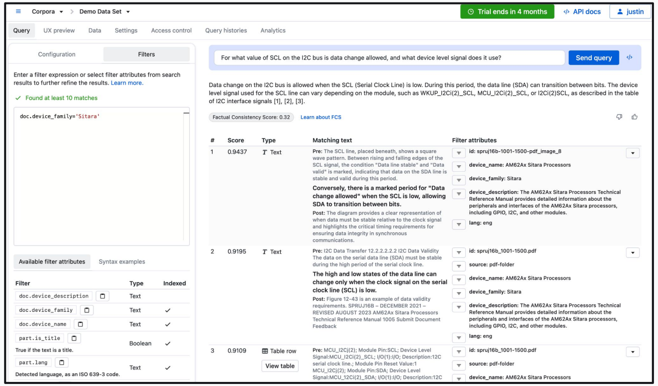
Can Use Many Tools: Vectara RAG, Database Query, API to take Action, ...

How to Use Vectara

Web Console

For learning & app development assistance

- Account management
- App prototyping
- Query config tuning
- API exploration



Enter a filter expression or select filter attributes from search results to further refine the results. Learn more.

✓ Found at least 10 matches

```
doc.device_family='Sitara'
```

Filter	Type	Indexed
`doc.device_description`	Text	
`doc.device_family`	Text	✓
`doc.device_name`	Text	✓
`part.is_title` True if the text is a title.	Boolean	✓
`part.lang` Detected language, as an ISO 639-3 code.	Text	✓

For what value of SCL on the I2C bus is data change allowed, and what device level signal does it use?

Data change on the I2C bus is allowed when the SCL (Serial Clock Line) is low. During this period, the data line (SDA) can transition between bits. The device level signal used for the SCL line can vary depending on the module, such as WKUP_I2Ci(2)_SCL, MCU_I2Ci(2)_SCL, or I2Ci(2)SCL, as described in the table of I2C interface signals [1], [2], [3].

Factual Consistency Score: 0.32

#	Score	Type	Matching text	Filter attributes
1	0.9437	Text	**Pre:** The SCL line, placed beneath, shows a square wave pattern. Between rising and falling edges of the SCL signal, the condition "Data line stable" and "Data valid" is marked, indicating that data on the SDA line is stable and valid during this period. **Conversely, there is a marked period for "Data change allowed" when the SCL is low, allowing SDA to transition between bits.** **Post:** The diagram provides a clear representation of when data must be stable relative to the clock signal and highlights the critical timing requirements for ensuring data integrity in synchronous communications.	**id:** spruj16b-1001-1500-pdf_image_8 **device_name:** AM62Ax Sitara Processors **device_family:** Sitara **device_description:** The AM62Ax Sitara Processors Technical Reference Manual provides detailed information about the peripherals and interfaces of the AM62Ax Sitara processors, including GPIO, I2C, and other modules. **lang:** eng
2	0.9195	Text	**Pre:** I2C Data Transfer 12.2.2.2.2.2 I2C Data Validity The data on the serial data line (SDA) must be stable during the high period of the serial clock line. **The high and low states of the data line can change only when the clock signal on the serial clock line (SCL) is low.** **Post:** Figure 12-43 is an example of data validity requirements. SPRUJ16B – DECEMBER 2021 – REVISED AUGUST 2023 AM62Ax Sitara Processors Technical Reference Manual 1005 Submit Document Feedback	**id:** spruj16b_1001-1500.pdf **source:** pdf-folder **device_name:** AM62Ax Sitara Processors **device_family:** Sitara **device_description:** The AM62Ax Sitara Processors Technical Reference Manual provides detailed information about the peripherals and interfaces of the AM62Ax Sitara processors, including GPIO, I2C, and other modules. **lang:** eng
3	0.9109	Table row	**Pre:** MCU_I2Cj(2); Module Pin:SCL; Device Level Signal:MCU_I2Ci(2)_SCL; I/O(1):I/O; Description:12C serial clock line.; Module Pin Reset Value:1 MCU_I2Cj(2); Module Pin:SDA; Device Level Signal:MCU_12Ci(2)_SDA; I/O(1):I/O; Description:12C	**id:** spruj16b_1001-1500.pdf **source:** pdf-folder **device_name:** AM62Ax Sitara Processors

Developer Tools

For building production apps for end users

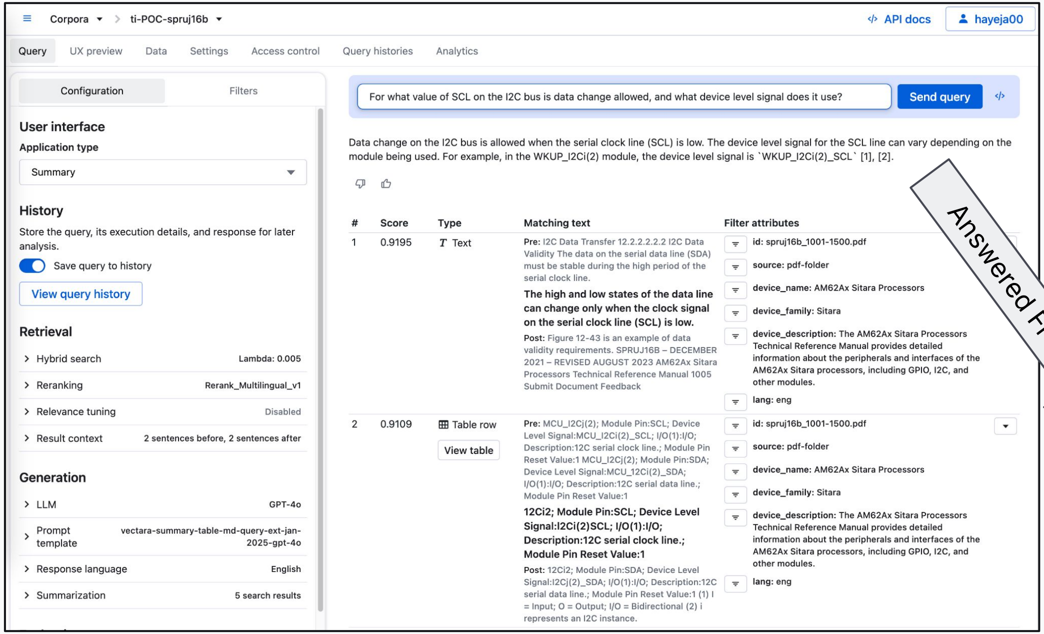
- API (REST) - call from your apps to ingest, query, and administer
- SDK (Python) - same
- Utilities - accelerate app development
  - **vectara-ingest** - crawlers (folder, web, confluence, servicenow, ...); table/image extraction; large file splitting; and more
  - **vectara-agentic** - agentic app framework for complex queries, conversations, multi-tool workflows, text2sql, actions; built on top of llama-index
  - **create-ui** - create simple Q&A/chat app
  - Others at github.com/vectara

Demo Apps & Console Tour

(15 mins.)

Demo: Question Answering AI Assistant

For what value of SCL on the I2C bus is data change allowed, and what device level signal does it use?



User interface

Application type: Summary

History

Store the query, its execution details, and response for later analysis.

● Save query to history

Retrieval

- Hybrid search: Lambda: 0.005
- Reranking: Rerank_Multilingual_v1
- Relevance tuning: Disabled
- Result context: 2 sentences before, 2 sentences after

Generation

- LLM: GPT-4o
- Prompt template: vectara-summary-table-md-query-ext-jan-2025-gpt-4o
- Response language: English
- Summarization: 5 search results

For what value of SCL on the I2C bus is data change allowed, and what device level signal does it use?

Data change on the I2C bus is allowed when the serial clock line (SCL) is low. The device level signal for the SCL line can vary depending on the module being used. For example, in the WKUP_I2Ci(2) module, the device level signal is `WKUP_I2Ci(2)_SCL` [1], [2].

#	Score	Type	Matching text	Filter attributes
1	0.9195	Text	Pre: I2C Data Transfer 12.2.2.2.2.2 I2C Data Validity The data on the serial data line (SDA) must be stable during the high period of the serial clock line. **The high and low states of the data line can change only when the clock signal on the serial clock line (SCL) is low.** Post: Figure 12-43 is an example of data validity requirements. SPRUJ16B – DECEMBER 2021 – REVISED AUGUST 2023 AM62Ax Sitara Processors Technical Reference Manual 1005 Submit Document Feedback	id: spruj16b_1001-1500.pdf source: pdf-folder device_name: AM62Ax Sitara Processors device_family: Sitara device_description: The AM62Ax Sitara Processors Technical Reference Manual provides detailed information about the peripherals and interfaces of the AM62Ax Sitara processors, including GPIO, I2C, and other modules. lang: eng
2	0.9109	Table row	Pre: MCU_I2Cj(2); Module Pin:SCL; Device Level Signal:MCU_I2Ci(2)_SCL; I/O(1):I/O; Description:12C serial clock line.; Module Pin Reset Value:1 MCU_I2Cj(2); Module Pin:SDA; Device Level Signal:MCU_12Ci(2)_SDA; I/O(1):I/O; Description:12C serial data line.; Module Pin Reset Value:1 **12Ci2; Module Pin:SCL; Device Level Signal:I2Ci(2)SCL; I/O(1):I/O; Description:12C serial clock line.; Module Pin Reset Value:1** Post: 12Ci2; Module Pin:SDA; Device Level Signal:I2Cj(2)_SDA; I/O(1):I/O; Description:12C serial data line.; Module Pin Reset Value:1 (1) I = Input; O = Output; I/O = Bidirectional (2) i represents an I2C instance.	id: spruj16b_1001-1500.pdf source: pdf-folder device_name: AM62Ax Sitara Processors device_family: Sitara device_description: The AM62Ax Sitara Processors Technical Reference Manual provides detailed information about the peripherals and interfaces of the AM62Ax Sitara processors, including GPIO, I2C, and other modules. lang: eng

Answered From

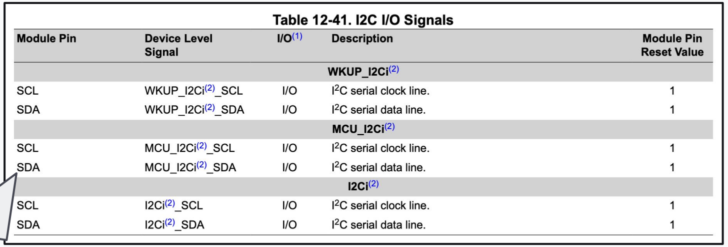
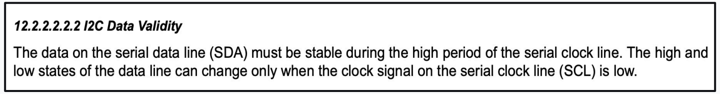


Table 12-41. I2C I/O Signals

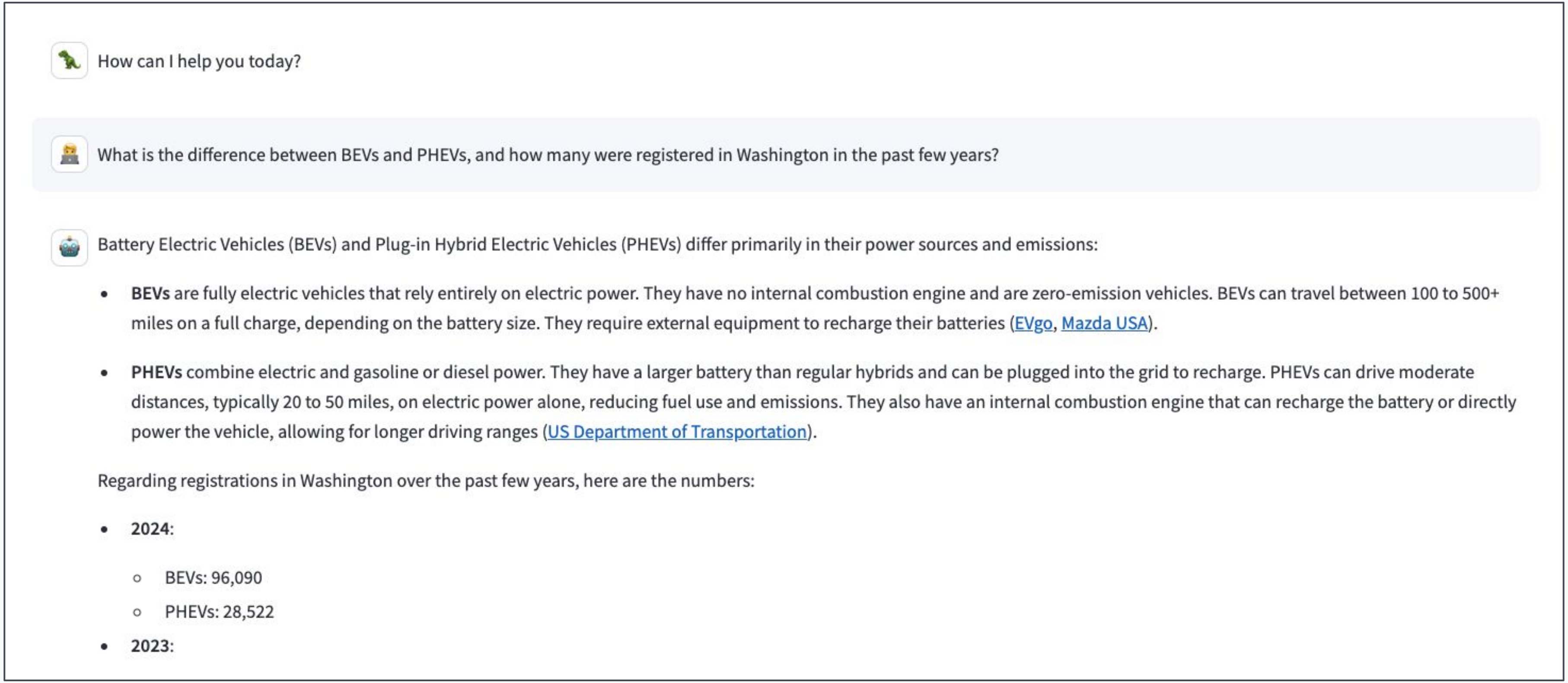
Module Pin	Device Level Signal	I/O$^{(1)}$	Description	Module Pin Reset Value
			WKUP_I2Ci$^{(2)}$	
SCL	WKUP_I2Ci$^{(2)}$_SCL	I/O	I^{2}C serial clock line.	1
SDA	WKUP_I2Ci$^{(2)}$_SDA	I/O	I^{2}C serial data line.	1
			MCU_I2Ci$^{(2)}$	
SCL	MCU_I2Ci$^{(2)}$_SCL	I/O	I^{2}C serial clock line.	1
SDA	MCU_I2Ci$^{(2)}$_SDA	I/O	I^{2}C serial data line.	1
			I2Ci$^{(2)}$	
SCL	I2Ci$^{(2)}$_SCL	I/O	I^{2}C serial clock line.	1
SDA	I2Ci$^{(2)}$_SDA	I/O	I^{2}C serial data line.	1



12.2.2.2.2.2 I2C Data Validity

The data on the serial data line (SDA) must be stable during the high period of the serial clock line. The high and low states of the data line can change only when the clock signal on the serial clock line (SCL) is low.

Demo: Electric Vehicle Assistant (Agentic App)



🦖 How can I help you today?

🧑‍💻 What is the difference between BEVs and PHEVs, and how many were registered in Washington in the past few years?

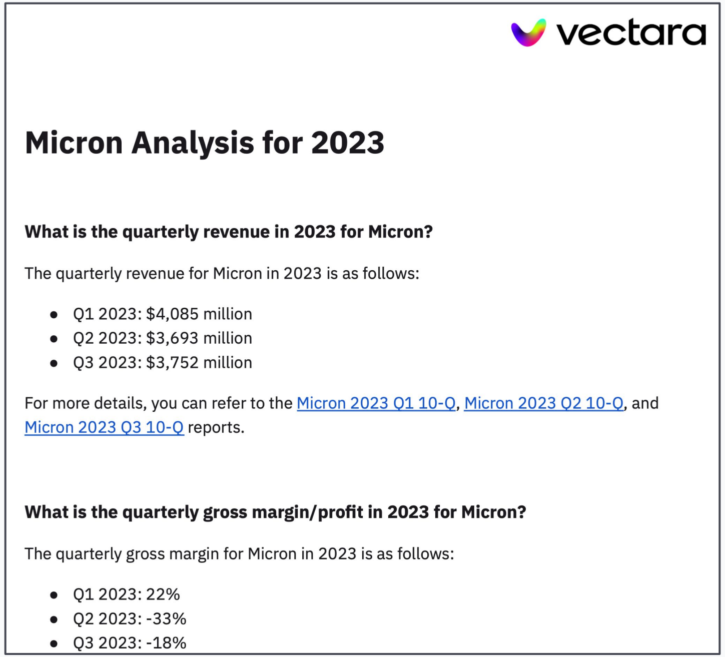
🤖 Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) differ primarily in their power sources and emissions:

- **BEVs** are fully electric vehicles that rely entirely on electric power. They have no internal combustion engine and are zero-emission vehicles. BEVs can travel between 100 to 500+ miles on a full charge, depending on the battery size. They require external equipment to recharge their batteries (EVgo, Mazda USA).
- **PHEVs** combine electric and gasoline or diesel power. They have a larger battery than regular hybrids and can be plugged into the grid to recharge. PHEVs can drive moderate distances, typically 20 to 50 miles, on electric power alone, reducing fuel use and emissions. They also have an internal combustion engine that can recharge the battery or directly power the vehicle, allowing for longer driving ranges (US Department of Transportation).

Regarding registrations in Washington over the past few years, here are the numbers:

- **2024**:
  - BEVs: 96,090
  - PHEVs: 28,522
- **2023**:

Demo: Document Generation



Micron Analysis for 2023

What is the quarterly revenue in 2023 for Micron?

The quarterly revenue for Micron in 2023 is as follows:

- Q1 2023: $4,085 million
- Q2 2023: $3,693 million
- Q3 2023: $3,752 million

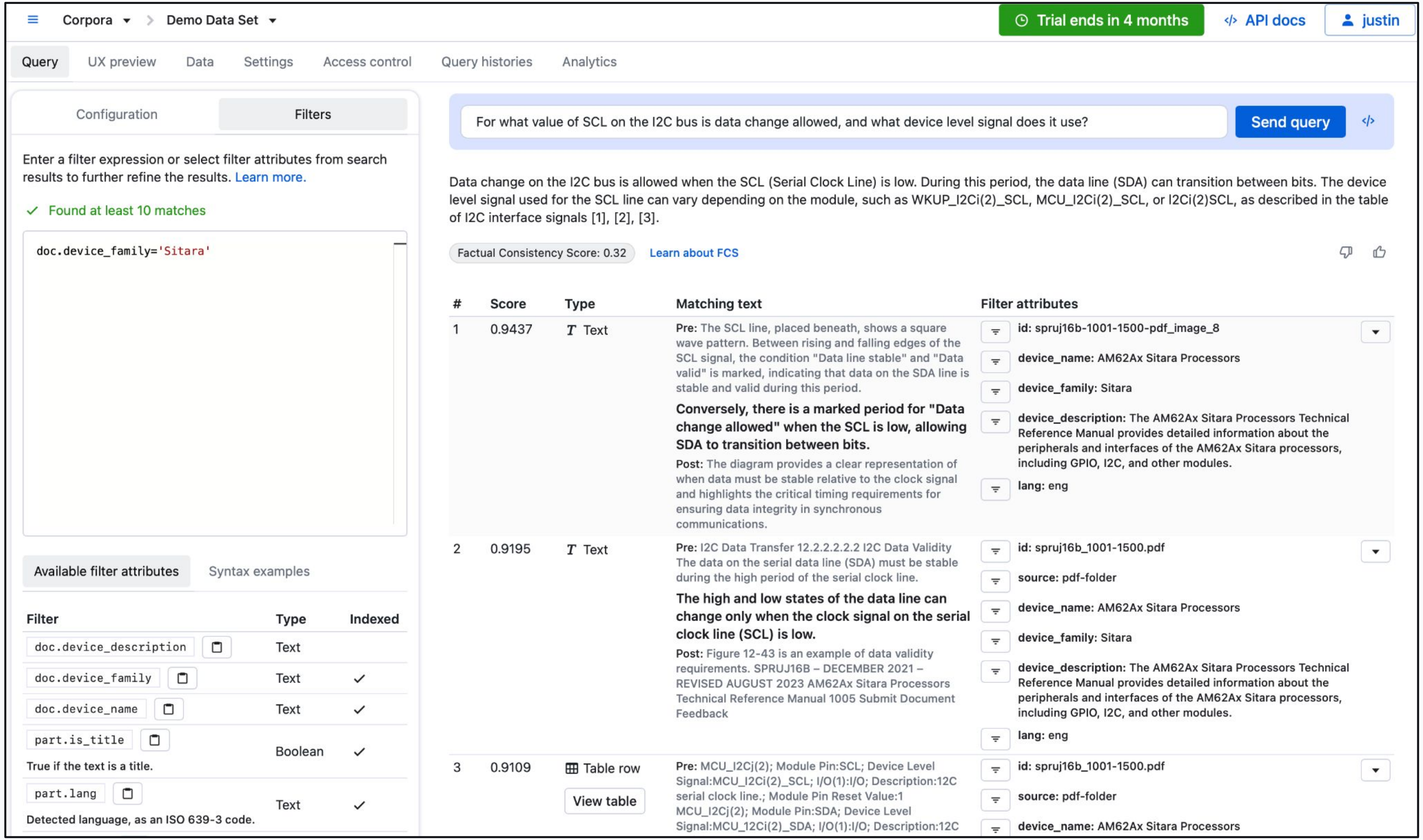
For more details, you can refer to the Micron 2023 Q1 10-Q, Micron 2023 Q2 10-Q, and Micron 2023 Q3 10-Q reports.

What is the quarterly gross margin/profit in 2023 for Micron?

The quarterly gross margin for Micron in 2023 is as follows:

- Q1 2023: 22%
- Q2 2023: -33%
- Q3 2023: -18%

Demo: Console Tour



Enter a filter expression or select filter attributes from search results to further refine the results. Learn more.

✓ Found at least 10 matches

```
doc.device_family='Sitara'
```

Filter	Type	Indexed
`doc.device_description`	Text	
`doc.device_family`	Text	✓
`doc.device_name`	Text	✓
`part.is_title` True if the text is a title.	Boolean	✓
`part.lang` Detected language, as an ISO 639-3 code.	Text	✓

For what value of SCL on the I2C bus is data change allowed, and what device level signal does it use?

Data change on the I2C bus is allowed when the SCL (Serial Clock Line) is low. During this period, the data line (SDA) can transition between bits. The device level signal used for the SCL line can vary depending on the module, such as WKUP_I2Ci(2)_SCL, MCU_I2Ci(2)_SCL, or I2Ci(2)SCL, as described in the table of I2C interface signals [1], [2], [3].

Factual Consistency Score: 0.32 Learn about FCS

#	Score	Type	Matching text	Filter attributes
1	0.9437	Text	**Pre:** The SCL line, placed beneath, shows a square wave pattern. Between rising and falling edges of the SCL signal, the condition "Data line stable" and "Data valid" is marked, indicating that data on the SDA line is stable and valid during this period. **Conversely, there is a marked period for "Data change allowed" when the SCL is low, allowing SDA to transition between bits.** **Post:** The diagram provides a clear representation of when data must be stable relative to the clock signal and highlights the critical timing requirements for ensuring data integrity in synchronous communications.	**id:** spruj16b-1001-1500-pdf_image_8 **device_name:** AM62Ax Sitara Processors **device_family:** Sitara **device_description:** The AM62Ax Sitara Processors Technical Reference Manual provides detailed information about the peripherals and interfaces of the AM62Ax Sitara processors, including GPIO, I2C, and other modules. **lang:** eng
2	0.9195	Text	**Pre:** I2C Data Transfer 12.2.2.2.2.2 I2C Data Validity The data on the serial data line (SDA) must be stable during the high period of the serial clock line. **The high and low states of the data line can change only when the clock signal on the serial clock line (SCL) is low.** **Post:** Figure 12-43 is an example of data validity requirements. SPRUJ16B – DECEMBER 2021 – REVISED AUGUST 2023 AM62Ax Sitara Processors Technical Reference Manual 1005 Submit Document Feedback	**id:** spruj16b_1001-1500.pdf **source:** pdf-folder **device_name:** AM62Ax Sitara Processors **device_family:** Sitara **device_description:** The AM62Ax Sitara Processors Technical Reference Manual provides detailed information about the peripherals and interfaces of the AM62Ax Sitara processors, including GPIO, I2C, and other modules. **lang:** eng
3	0.9109	Table row View table	**Pre:** MCU_I2Cj(2); Module Pin:SCL; Device Level Signal:MCU_I2Ci(2)_SCL; I/O(1):I/O; Description:12C serial clock line.; Module Pin Reset Value:1 MCU_I2Cj(2); Module Pin:SDA; Device Level Signal:MCU_12Ci(2)_SDA; I/O(1):I/O; Description:12C	**id:** spruj16b_1001-1500.pdf **source:** pdf-folder **device_name:** AM62Ax Sitara Processors

Exercise 1 - Account Setup, Q&A App Prototype

(30 mins.)

Account Setup

To create an account for yourself:

1. https://vectaradev.itg.it.com
2. Click on `Sign up`
3. Enter email and password
4. Can use web console, API, SDK

^^ We will do this today ^^

To add another person to your account:

1. Log on to web console
2. Go to “Team” page
3. Add team member
4. Enter other person’s info
5. Copy link/code & send to other person
6. Other person logs in with link/code and customer ID, and sets password
7. Other person can use web console, API, SDK

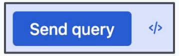
^^ Only including this for reference ^^

Notes:

- Client applications authenticate via API Key or OAuth2 identity
- Access control permissions can be granted to users, API Keys, and OAuth2 identities

Q&A App Prototype

1. Create corpus
    a. "Summarized semantic search", enter name and key, keep the other defaults
2. Upload file(s)
    a. Start with taming_shrew.txt
    b. Do not select "table extraction"
    c. Can also upload other files: PDF/DOC/PPT ok
    d. Files using different languages ok
3. View uploaded documents on Data tab
    a. Document, metadata, parts
4. Test simple queries on Query tab
    a. Vectara expects RAG query to be a question, not a prompt or instruction
    b. Use "Summary" for application type
    c. Query in different language ok - it will respond in the language of the data
        i. Respond in language of query coming soon
    d. Enable "Save query to history"
5. Inspect API details via </> icon



    a. Request, response, compiled prompt, errors
6. Test query config options
    a. Hybrid search
    b. Rerankers
    c. Sentences before/after
    d. Number of results to summarize
        i. If you see FCS error, try setting this lower
    e. Filters - can only filter on automatically extracted metadata fields (e.g. doc.id, part.is_title)
7. View query history

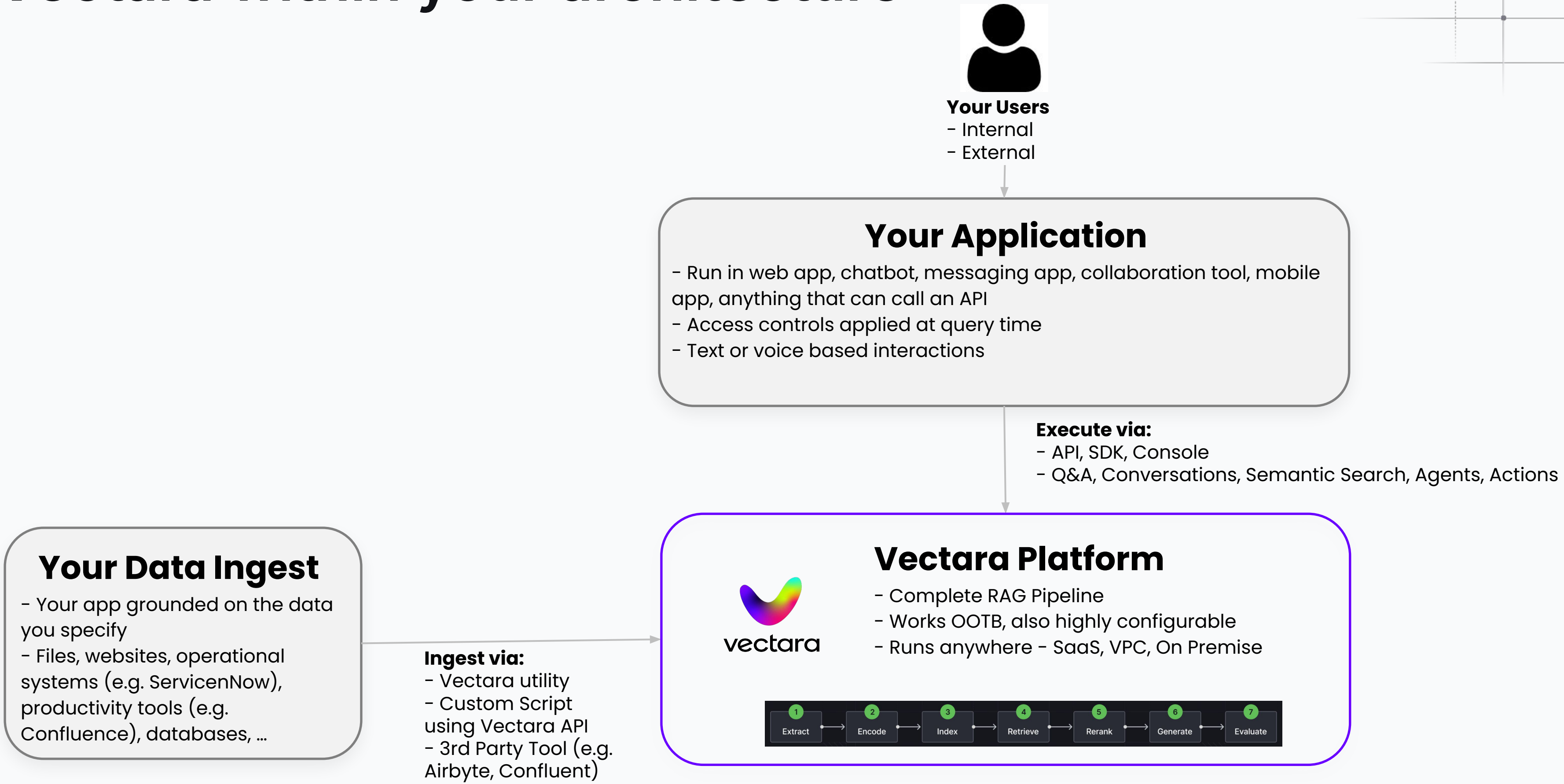
Break

(5 mins.)

App Development

(30 mins.)

Vectara within your architecture



Your Users
- Internal
- External

Your Application

- Run in web app, chatbot, messaging app, collaboration tool, mobile app, anything that can call an API
- Access controls applied at query time
- Text or voice based interactions

Execute via:
- API, SDK, Console
- Q&A, Conversations, Semantic Search, Agents, Actions

Your Data Ingest

- Your app grounded on the data you specify
- Files, websites, operational systems (e.g. ServiceNow), productivity tools (e.g. Confluence), databases, …

Ingest via:
- Vectara utility
- Custom Script using Vectara API
- 3rd Party Tool (e.g. Airbyte, Confluent)

Vectara Platform

- Complete RAG Pipeline
- Works OOTB, also highly configurable
- Runs anywhere - SaaS, VPC, On Premise

Common Vectara Application Patterns

Best practice is to crawl... walk... run through these from left to right

Q&A Assistant

Data: manuals, policy docs, wiki pages, website

UI: search bar in existing app, collab tool (Slack, Teams, WhatsApp, etc)

Dev Tools: vectara-ingest or custom script, Query API

Chat Assistant

Data: manuals, policy docs, wiki pages, website

UI: chat bot in existing app, custom agentic app

Dev Tools: vectara-ingest or custom script, Query API, vectara-agentic

Research & Analysis

Data: all sources related to a topic (e.g. manuals, data sheets, support tickets, emails), other operational systems (e.g. database, API, other app, etc)

UI: search bar or chat bot in existing app, custom agentic app

Dev Tools: vectara-ingest or custom script, Query API, vectara-agentic, create-ui

AI Agent

Data: all sources related to a topic (e.g. manuals, data sheets, support tickets, emails), other operational systems (e.g. database, API, other app, etc)

UI: search bar or chat bot in existing app, custom agentic app, standalone app triggered or run on demand

Dev Tools: vectara-ingest or custom script, Query API, vectara-agentic, custom tool development

Vectara App Development Process

Design → Ingest Dev → UI Dev → Query Config Tuning → Usage Monitoring & Analytics

Design

1. Data structure
    a. Files or raw data
    b. If raw data: pre-chunk or not (recommend to not pre-chunk b/c Vectara does that for you)
2. Corpus modeling
    a. Filters - purpose; choose document level or part level, and data type
    b. Other metadata - to use in UDF reranker, custom prompt, response rendering (e.g. `source_url` can be used to add a link to the original source in the UI)
    c. One corpus or many
3. Client type
    a. Standalone app or embedded into existing app
    b. Mode: direct RAG or agentic
    c. Stateless Q&A or stateful chat
4. Access control (see next slide)

Vectara App Development Process

Design → Ingest Dev → UI Dev → Query Config Tuning → Usage Monitoring & Analytics

ABAC

- Involves leveraging attributes (e.g., user roles, document tags, or contextual metadata) to control access to documents and retrieval results.

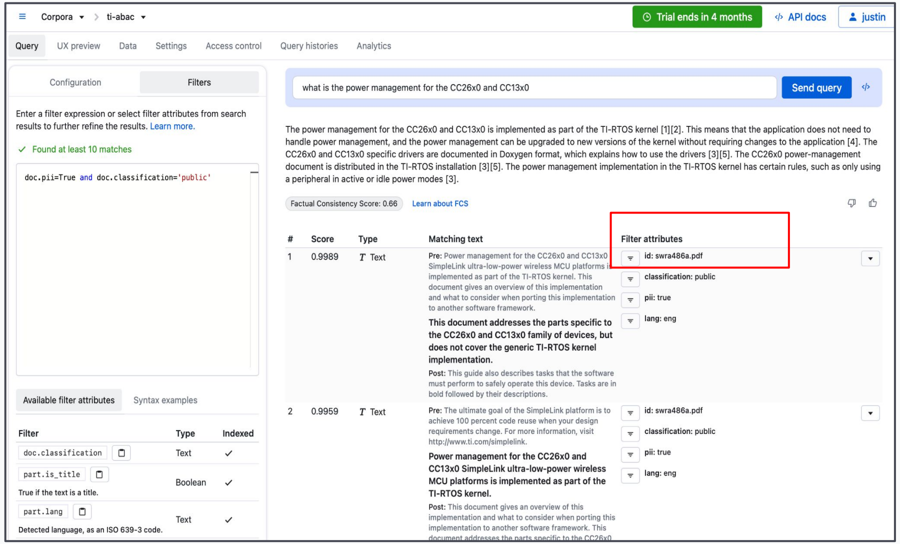
Key Components

- **Resource Attributes:** Metadata attached to data at ingest time (e.g. classification level, project tags).
- **Query-Time Filtering:** During a query, the application applies access control filters using the user’s attributes, and Vectara evaluates the user's attributes against the document metadata to enforce access controls, filtering data that the user is not permitted to see.

Vectara App Development Process

Design → Ingest Dev → UI Dev → Query Config Tuning → Usage Monitoring & Analytics

Returns different results based on the attributes of a user (e.g. group they belong to, job level, permissions granted) as enforced by the filter expression.



Enter a filter expression or select filter attributes from search results to further refine the results. Learn more.

✓ Found at least 10 matches

```
doc.pii=True and doc.classification='public'
```

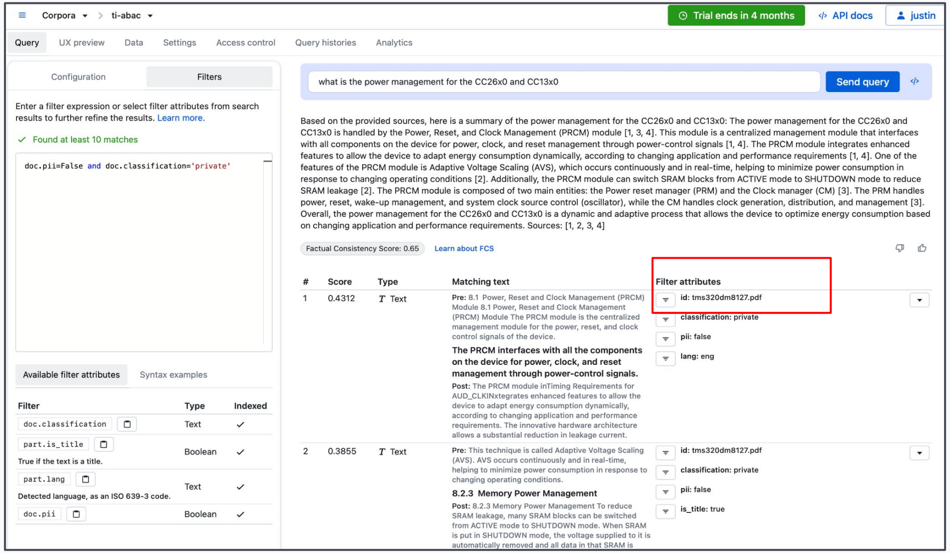
Filter	Type	Indexed
`doc.classification`	Text	✓
`part.is_title` True if the text is a title.	Boolean	✓
`part.lang` Detected language, as an ISO 639-3 code.	Text	✓

what is the power management for the CC26x0 and CC13x0

The power management for the CC26x0 and CC13x0 is implemented as part of the TI-RTOS kernel [1][2]. This means that the application does not need to handle power management, and the power management can be upgraded to new versions of the kernel without requiring changes to the application [4]. The CC26x0 and CC13x0 specific drivers are documented in Doxygen format, which explains how to use the drivers [3][5]. The CC26x0 power-management document is distributed in the TI-RTOS installation [3][5]. The power management implementation in the TI-RTOS kernel has certain rules, such as only using a peripheral in active or idle power modes [3].

Factual Consistency Score: 0.66

#	Score	Type	Matching text	Filter attributes
1	0.9989	T Text	**Pre:** Power management for the CC26x0 and CC13x0 SimpleLink ultra-low-power wireless MCU platforms is implemented as part of the TI-RTOS kernel. This document gives an overview of this implementation and what to consider when porting this implementation to another software framework. **This document addresses the parts specific to the CC26x0 and CC13x0 family of devices, but does not cover the generic TI-RTOS kernel implementation.** **Post:** This guide also describes tasks that the software must perform to safely operate this device. Tasks are in bold followed by their descriptions.	id: swra486a.pdf classification: public pii: true lang: eng
2	0.9959	T Text	**Pre:** The ultimate goal of the SimpleLink platform is to achieve 100 percent code reuse when your design requirements change. For more information, visit http://www.ti.com/simplelink. **Power management for the CC26x0 and CC13x0 SimpleLink ultra-low-power wireless MCU platforms is implemented as part of the TI-RTOS kernel.** **Post:** This document gives an overview of this implementation and what to consider when porting this implementation to another software framework. This document addresses the parts specific to the CC26x0	id: swra486a.pdf classification: public pii: true lang: eng



Enter a filter expression or select filter attributes from search results to further refine the results. Learn more.

✓ Found at least 10 matches

```
doc.pii=False and doc.classification='private'
```

Filter	Type	Indexed
`doc.classification`	Text	✓
`part.is_title` True if the text is a title.	Boolean	✓
`part.lang` Detected language, as an ISO 639-3 code.	Text	✓
`doc.pii`	Boolean	✓

what is the power management for the CC26x0 and CC13x0

Based on the provided sources, here is a summary of the power management for the CC26x0 and CC13x0: The power management for the CC26x0 and CC13x0 is handled by the Power, Reset, and Clock Management (PRCM) module [1, 3, 4]. This module is a centralized management module that interfaces with all components on the device for power, clock, and reset management through power-control signals [1, 4]. The PRCM module integrates enhanced features to allow the device to adapt energy consumption dynamically, according to changing application and performance requirements [1, 4]. One of the features of the PRCM module is Adaptive Voltage Scaling (AVS), which occurs continuously and in real-time, helping to minimize power consumption in response to changing operating conditions [2]. Additionally, the PRCM module can switch SRAM blocks from ACTIVE mode to SHUTDOWN mode to reduce SRAM leakage [2]. The PRCM module is composed of two main entities: the Power reset manager (PRM) and the Clock manager (CM) [3]. The PRM handles power, reset, wake-up management, and system clock source control (oscillator), while the CM handles clock generation, distribution, and management [3]. Overall, the power management for the CC26x0 and CC13x0 is a dynamic and adaptive process that allows the device to optimize energy consumption based on changing application and performance requirements. Sources: [1, 2, 3, 4]

Factual Consistency Score: 0.65

#	Score	Type	Matching text	Filter attributes
1	0.4312	T Text	**Pre:** 8.1 Power, Reset and Clock Management (PRCM) Module 8.1 Power, Reset and Clock Management (PRCM) Module The PRCM module is the centralized management module for the power, reset, and clock control signals of the device. **The PRCM interfaces with all the components on the device for power, clock, and reset management through power-control signals.** **Post:** The PRCM module inTiming Requirements for AUD_CLKINxtegrates enhanced features to allow the device to adapt energy consumption dynamically, according to changing application and performance requirements. The innovative hardware architecture allows a substantial reduction in leakage current.	id: tms320dm8127.pdf classification: private pii: false lang: eng
2	0.3855	T Text	**Pre:** This technique is called Adaptive Voltage Scaling (AVS). AVS occurs continuously and in real-time, helping to minimize power consumption in response to changing operating conditions. **8.2.3 Memory Power Management** **Post:** 8.2.3 Memory Power Management To reduce SRAM leakage, many SRAM blocks can be switched from ACTIVE mode to SHUTDOWN mode. When SRAM is put in SHUTDOWN mode, the voltage supplied to it is automatically removed and all data in that SRAM is	id: tms320dm8127.pdf classification: private pii: false is_title: true

Sample filter expressions:

- `doc.pii=False`
- `doc.classification!='private'`
- `doc.pii=True and doc.classification='private'`

Vectara App Development Process

Design → **Ingest Dev** → UI Dev → Query Config Tuning → Usage Monitoring & Analytics

Ingest Dev

1. Approach
    1. Custom script - uses API/SDK directly; most flexibility
    2. vectara-ingest - utility works OOTB and is also extensible; fast time to production
    3. Third party tool - if already in use or other features required; e.g. Airbyte, Unstructured
2. Integrate with source systems - get access, API key, etc
3. Pre-process data (if necessary)
    1. Structure into sections, or chunks/parts (if pre-chunking)
    2. Entity recognition/extraction
    3. Enrich, transform, normalize
    4. Pre-summarize table/images
4. Attach metadata

Vectara App Development Process

Design → Ingest Dev → **UI Dev** → Query Config Tuning → Usage Monitoring & Analytics

UI Dev

1. Filters - Restrict data set, apply access controls
2. Develop client app for end users
   a. Let users enter only their query, or also specify filters?
      i. *Note: automatic generation of filter expressions based on natural language request coming soon to Vectara API; can do today in agentic client app*
   b. Decide how to show citations and links to original source?
3. Custom prompt development
   a. To change behavior of LLM - style/format, guardrails, apply basic reasoning
4. Integrate into existing app via API or SDK calls (if applicable)
5. Develop agentic application - instructions, tools, etc (if applicable)

Vectara App Development Process

Design → Ingest Dev → UI Dev → **Query Config Tuning** → Usage Monitoring & Analytics

Query Config Tuning

1. Common config options
   a. Hybrid search
   b. Rerankers - relevance, UDF, maximum marginal relevance (diversity), knee
   c. Sentences before/after
   d. Number of results to summarize
   e. Filters
   f. Custom prompt
2. Response quality evaluation process
   a. Repeatable process to quantify point-in-time response quality (ideally automated script)
   b. Test data+query+response set
   c. Automated scoring of responses or human expert scoring or both
   d. Run during development, and periodically once in production

Vectara App Development Process

Design → Ingest Dev → UI Dev → Query Config Tuning → **Usage Monitoring & Analytics**

Usage Monitoring & Analytics

1. Train users to set expectations and/or reset habits
2. Enable saving query history
    a. Review occasionally to learn about usage patterns, gaps in data, etc
3. Capture user feedback - via feedback form, thumbs up/down, etc

Temporary Constraints for TI's Vectara Instance

Features in Vectara SaaS not yet available in the on premise Vectara instance for the Hackathon

***General Note**: This is a dev instance on undersized hardware (esp GPU capacity), so expect ingest and query latencies that are slower than what will be on the prod instance. This is the case especially with a large corpus (e.g. > 50,000 PDF pages) and if there are multiple parallel indexing requests.*

1. Cannot use Chat API feature
   a. Workaround: use vectara-agentic
2. Cannot use Query Intelligence feature (automatic filter generation from natural language query)
   a. Workaround: use vectara-agentic
3. Cannot use Tabular Extraction & Pre-Summarization within File Upload API
   a. Workaround: use vectara-ingest `parse_tables` option (for PDF, DOC, PPT, HTML files)
4. No Option for Image Extraction & Pre-Summarization

All are coming in phase 2 of the TI Vectara instance.

Exercise 2 - SDK/API

(30 mins.)

Basic API Usage - Query

1. Scan the Query API docs at https://docs.vectara.com/docs/api-reference/search-apis/search and https://docs.vectara.com/docs/rest-api/query-corpus
2. Take the following CURL sample, replace `:corpus_key` and `<API_KEY_VALUE>`, run on cmd line:

```
curl -L -X POST 'https://vectaradev.itg.it.com/api/v2/corpora/:corpus_key/query' \
-H 'Content-Type: application/json' \
-H 'Accept: application/json' \
-H 'x-api-key: <API_KEY_VALUE>' \
--data-raw '{
  "query": "What is this about?",
  "search": {
    "lexical_interpolation": 0.005,
    "context_configuration": {"sentences_before": 1, "sentences_after": 1},
    "reranker": {"type": "customer_reranker", "reranker_name": "Rerank_Multilingual_v1", "limit": 25}
  },
  "generation": {"generation_preset_name": "mockingbird-1.0-2024-07-16", "max_used_search_results": 5,
"response_language": "auto", "citations": {"style": "numeric"}, "enable_factual_consistency_score": true},
  "stream_response": false, "save_history": false
}'
```

3. Examine the response payload. It might help to copy it into a JSON formatter to see it better.
4. Experiment with different queries and different API payload configs - e.g. lexical_interpolation (hybrid search), context_configuration, reranker, max_used_search_results
   a. Note: click on the “Inspect” button for a query in the console to see the corresponding CURL version (this one uses OAuth though)

Basic API Usage - Document Summarization

1. Read the “Document Summarization” docs at https://docs.vectara.com/docs/api-reference/documents-apis/summarize-document and https://docs.vectara.com/docs/rest-api/summarize-corpus-document
2. Take the following CURL sample, replace `:corpus_key` and `:document_id` and `<API_KEY_VALUE>`, run on cmd line:

```
curl -L -X POST 'https://vectaradev.itg.it.com/api/v2/corpora/:corpus_key/documents/:document_id/summarize' \
-H 'Content-Type: application/json' \
-H 'Accept: application/json' \
-H 'x-api-key: <API_KEY_VALUE>' \
--data-raw '{
  "llm_name": "llm_custom_llama3_70B",
  "prompt_template": "[{\"role\":\"user\",\"content\":\"Summarize the document: #foreach ($part in
$vectaraDocument.parts()) $part.text() #end\"}]",
  "stream_response": false
}'
```

3. Examine the response payload. It might help to copy it into a JSON formatter to see it better.
4. Experiment with different documents and different versions of prompt_template
    a. You can use `$vectaraDocument.parts().subList(0,1000)`to use a subset of the parts (i.e. chunks) if the document is too large.
    b. You can use `$vectaraDocument.parts().size()`to get the total number of parts, to give an upper bounds, e.g. `$vectaraDocument.parts().subList(0,$vectaraDocument.parts().size())`

Prerequisites for SDK Exercise

1. Tested OS: Windows, Mac, Ubuntu
2. Python Distro: 3.11+, preferably 3.12
3. Copy code repo locally, or copy zipped version from TI network

```
> git clone https://github.com/vectara/python-sdk.git
```

4. Install Jupyter Lab, or install from TI network

```
> pip install jupyterlab
```

Basic SDK Usage

1. Copy details from console

   Personal API Key (from Authorization page), Customer ID (top right username button)
2. Access sample notebooks in python-sdk repo
   ```
   > cd python-sdk/examples/01_getting_started
   ```
3. Create a virtual env
   ```
   > python -m venv venv
   ```
4. Start Jupyter Lab
   ```
   > jupyter lab
   ```
5. Run `00_setup_authentication.ipynb` notebook
   1. In the code block with `LabHelper.setup_authentication()`: enter customer ID, select API Key for auth, and enter personal API Key
   2. Before running the `Credential Test` section, edit the `lab` profile in `.vec_auth.yaml` in your home directory
      ```
      api_enpdoint: https://vectaradev.itg.ti.com/api
      auth_enpdoint: https://vectaradev.itg.ti.com/ory
      verify_ssl: false
      ```
   3. Run the `Credential Test` code block
6. Run other notebooks: `01_query_api.ipynb`, `02_index_api.ipynb`, `03_upload_api.ipynb`

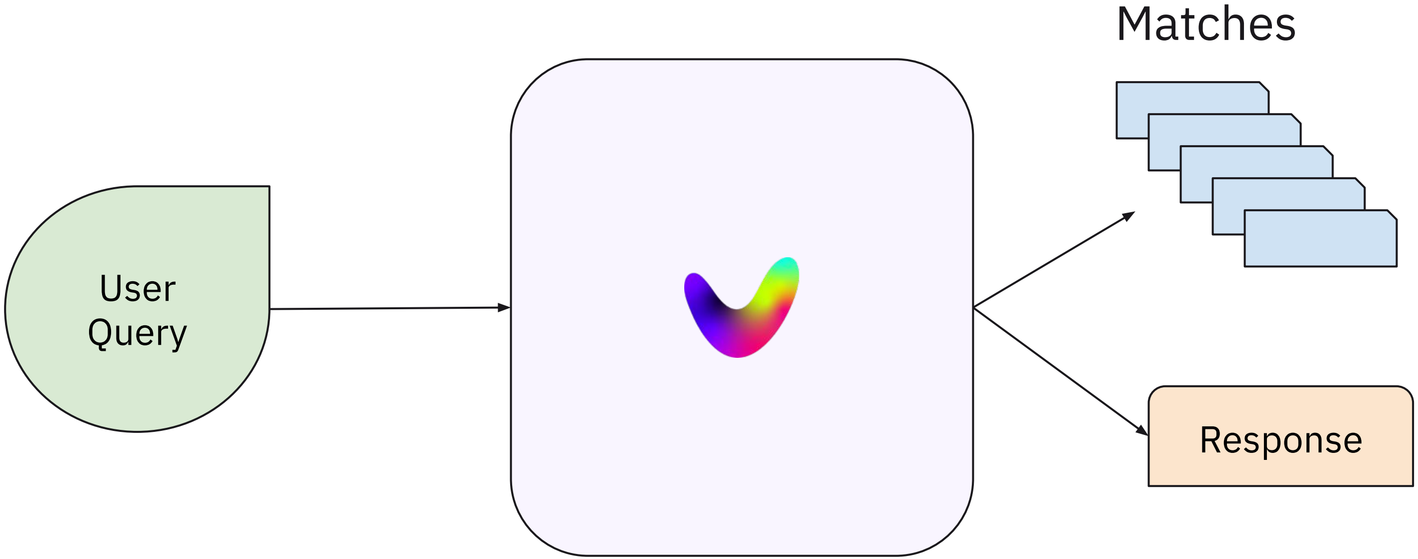
Break

(5 mins.)

Agentic RAG

(15 mins.)

What does Vectara RAG provide today?



User query is used **verbatim** for retrieving relevant facts from the index and for LLM response generation

More Complex Scenarios…
Example 1: Complex Queries

"What is morality and what is justice? Compare and contrast and provide a summary of the pros and cons"

Vectara today would send the full query against the index; best matches are retrieved and might not get all the right details

More Complex Scenarios…
Example 2: Query with Filtering

"What was the revenue for Uber in 2022?"

LLM pre-processing can rephrase this as:

Query = "revenues"

Filter = "doc.company = 'Uber' and doc.year = 2022"

Note: this capability is being added to the core Vectara query API shortly

More Complex Scenarios…
Example 3: Using Additional Tools

“Please summarize the interaction we had so far with WidgetCo in terms of the recent POC and when should we follow up?”

Even if we capture all slack data and all CRM data, this query sent directly to Vectara’s Query API may not provide a satisfactory response.

Instead, we can provide a “CRM tool” that can pull the latest (real time) information (e.g. including the emails 2 seconds ago) and include that when generating the response.

Proprietary and Confidential, © 2024, Vectara Inc.

More Complex Scenarios…
Example 4: Analytical Queries

“What is a technical default and how many of our customers are in technical default?”

Traditional RAG does not support “counting” operations, and those are more easily solved via SQL queries against operational datastores of the enterprise.

With a “customer SQL” tool an agent can call that tool to answer that part of the user query.

Proprietary and Confidential, © 2024, Vectara Inc.

More Complex Scenarios…
Example 5: Complex Workflows with Actions

"How do I delete a user from Unix?"

Agentic application can understand a process and step the user through it, gathering info from user, looking up info from RAG tool or other tools, triggering actions, etc

Proprietary and Confidential, © 2024, Vectara Inc.

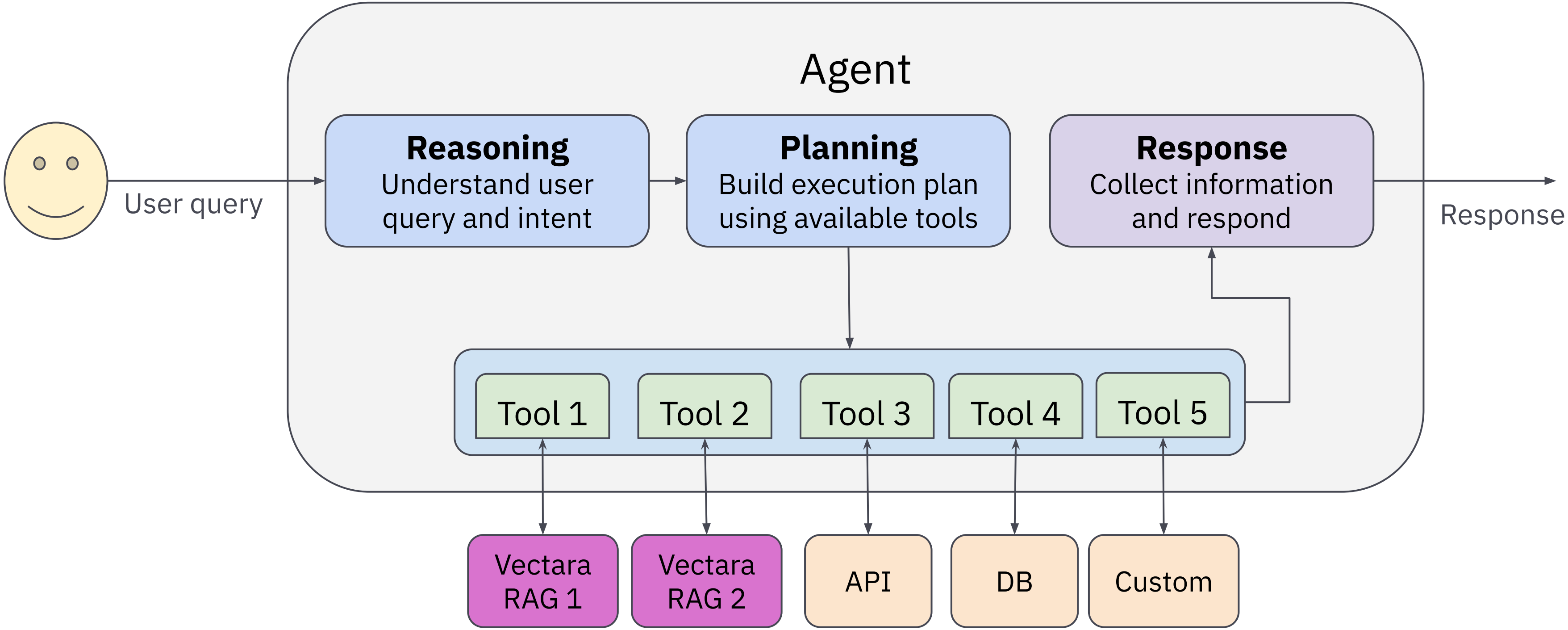
Agentic Applications Handle these Scenarios

1. Complex, multi-part queries
2. Queries with filtering
3. Calling multiple tools - RAG, SQL Database, Operational Tools (ticketing, CRM, etc), Emails, etc
4. Analytical queries and queries requiring deduction
5. Complex workflows that trigger actions

Note: developing agentic apps is different than developing other apps. Start simple and add capabilities as you learn more.

Proprietary and Confidential, © 2024, Vectara Inc.

Agentic Applications - How they Work



Agent

Reasoning
Understand user query and intent

Planning
Build execution plan using available tools

Response
Collect information and respond

Proprietary and Confidential, © 2024, Vectara Inc.

Exercise 3 - Agentic App

(20 mins.)

Build an Agentic Application

1. Get copy of `agentic_app.py` from TI network, and copy locally

```
> cd ../../..
> mkdir agentic
> cp /path/to/agentic_app.py agentic/
> cd agentic
```

2. Install latest vectara SDK and vectara-agentic libraries

```
> pip install vectara vectara-agentic
```

3. Update the following variables in the agentic_app.py file

```
vectara_api_key='VECTARA_API_KEY',
vectara_corpus_key='VECTARA_CORPUS_KEY' #e.g. justin_01-getting-started-query-api
private_llm_api_key="f0wshQtAGMRUGA10jlkeoQMTdEAlNVPH"
```

4. Run agentic app

```
> python agentic_app.py "What is the Wager about in Taming of the Shrew?"
> python agentic_app.py "Who made the Wager in the shrew play?"
> python agentic_app.py "Who died in Romeo and Juliet?"
> python agentic_app.py "What is the Wager about in romeo?"
```

Workshop Cleanup

1. Run notebook in Jupyter Lab to clean up Vectara account (optional)

   99_cleanup.ipynb

Hackathon Resources

How to Get Help

- **TI Support**
  - Who:
  - How:
- **Vectara Support**
  - Product docs & API Playground: https://docs.vectara.com
  - Sample API client apps: https://github.com/vectara/getting-started/tree/main/language-examples/python/vectara-rest
  - Community https://discuss.vectara.com
  - Ask question or submit feature request ti-hackathon@vectara.com

Thank You !!!

Admin Notes

Admin Notes

- Aerospike data storage currently has rate limits in place
- All the vectara services are in the vectara namespace.
    - When required to share log files during troubleshooting, getting the logs from everything in the vectara namespace is required.
- The service endpoints are:
    - vectaradev.itg.it.com/api
    - vectaradev.itg.it.com/ory
    - Everything is on that one endpoint with different mount points