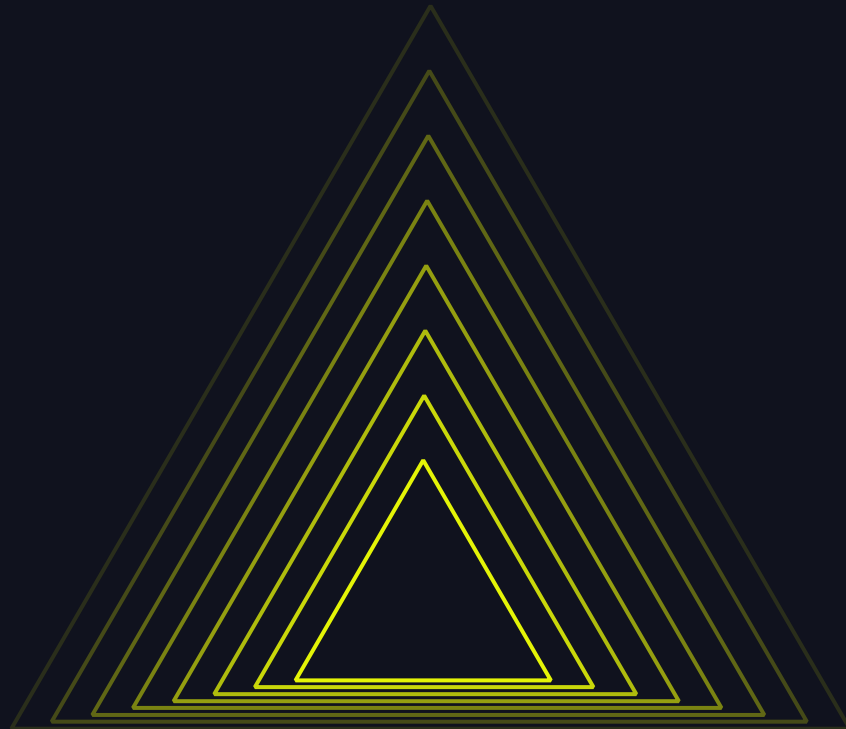


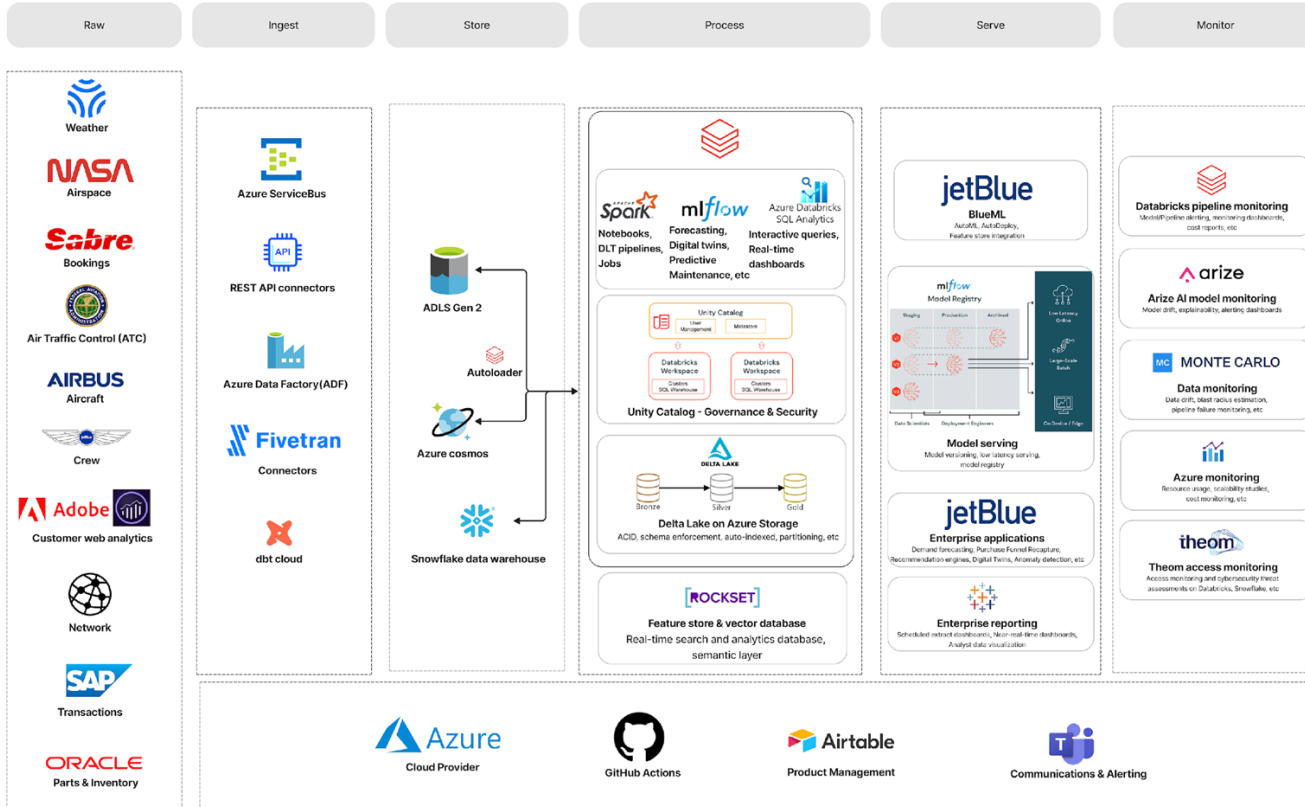
Building a secure & scalable LLM framework

Sai Ravuru - JetBlue, Navindra Yadav - Theom
Jun 12, 2024



Technology Stack

jetBlue Data, Analytics and Machine Learning Architecture



Pillars of a successful LLM stack

Efficient and prompting

Organizational knowledge repository

Data and access governance

- Training
- User feedback
- Internet access

- Curated data
- Accuracy testing
- Role based access

- Regular audits
- Logging
- Active Directory integration

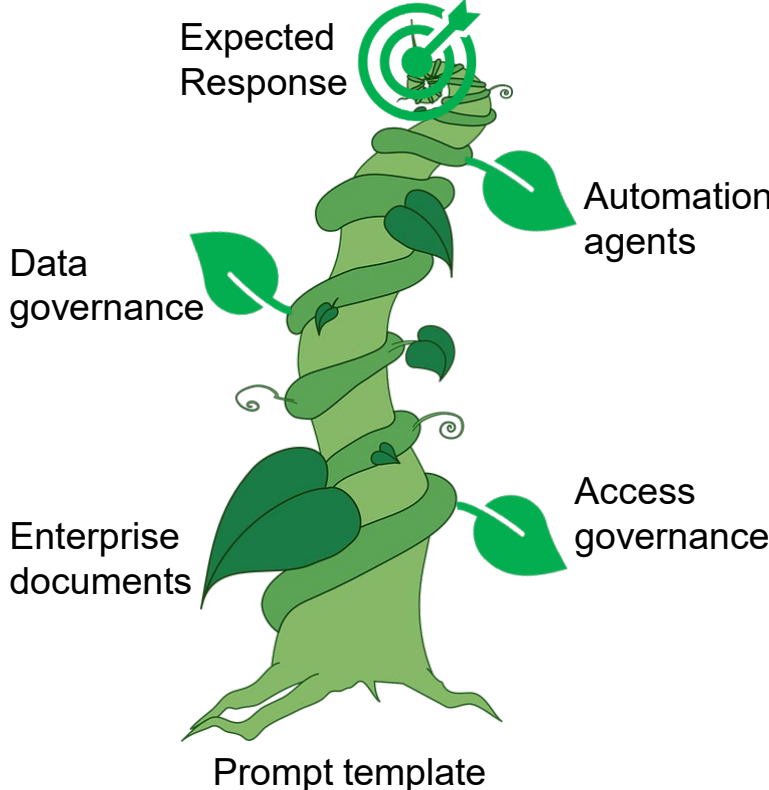


Generative AI maturity curve

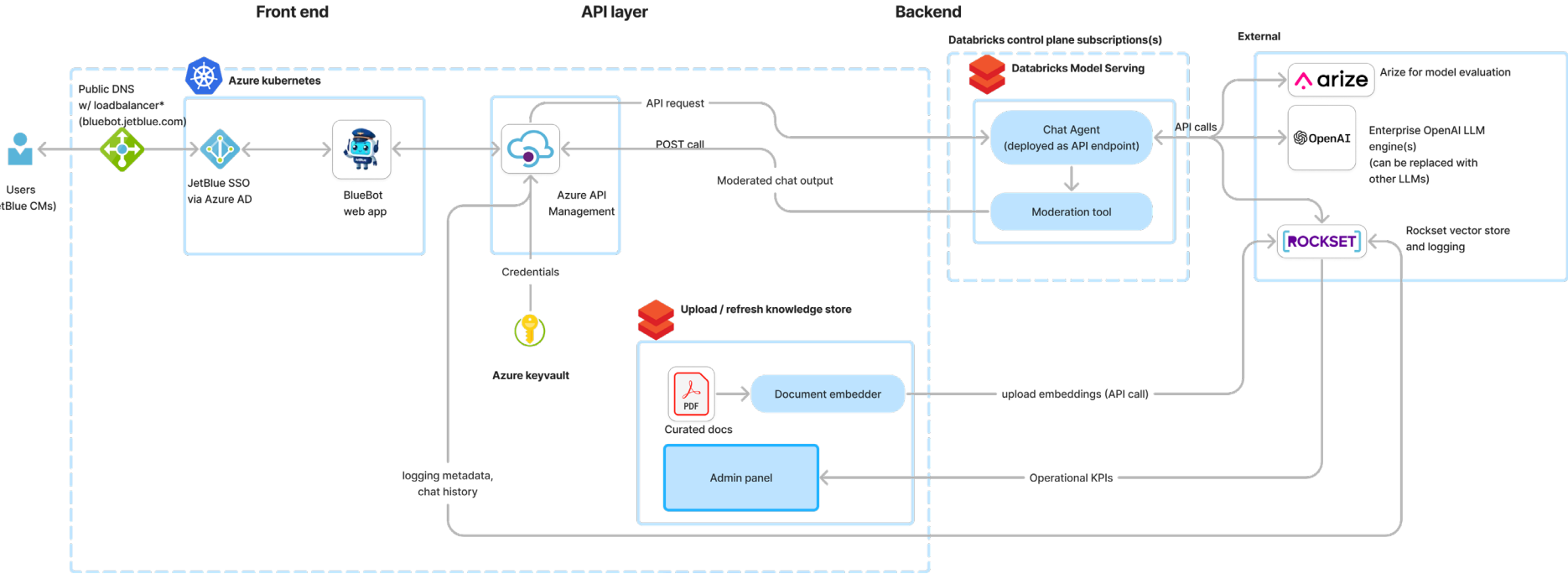
Expectation



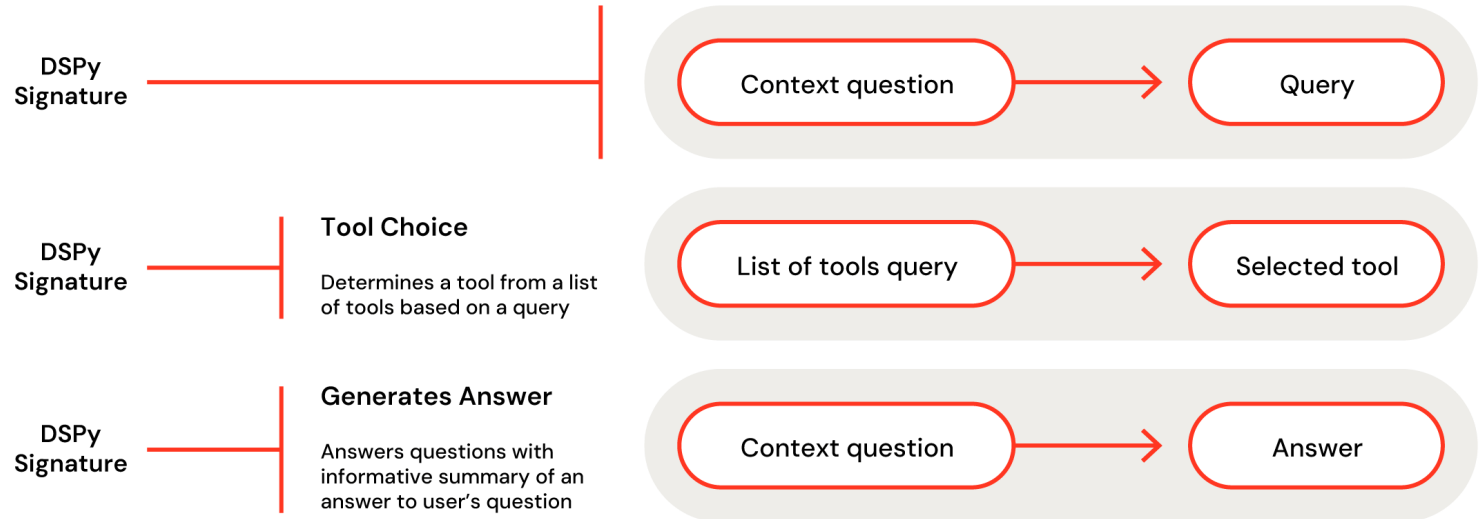
Reality



Generative AI @ scale



DsPy to the rescue



DsPy retriever initialization

Sample Python Code

PYTHON

```
class ToolRetriever(dspy.Module):
    def __init__(self):
        self.generate_query = dspy.ChainOfThought("context, question -> query")
        self.choose_tool = dspy.ChainOfThought(ToolChoice)
        self.generate_answer = dspy.ChainOfThought(GenerateAnswer)
        self.tools = "[answer_payroll_faq, irrelevant_content]"

    def irrelevant_content(self):
        return "Ask something else."
```

DsPy PyTorch-like assembly

Sample Python Code

PYTHON

```
def forward(self, question):
    client = OpenAI(api_key=openai_api_key)
    retrieve = DatabricksRM()
    context = []
    query_output = self.generate_query(context = context, question=question)
    tool_choice = self.choose_tool(list_of_tools=self.tools, query=query_output.query)

    if tool_choice.selected_tool == "irrelevant_content":
        return self.irrelevant_content()
    else:
        search_query_embedding = client.embeddings.create(model="text-embedding-ada-002",
        input=[question]).data[0].embedding
        retrieved_context = retrieve(search_query_embedding, 1)
        context += retrieved_context
        return self.generate_answer(context=context, question=question)
```


DsPy deployment

Sample Python Code

PYTHON

```
def predict(self, context, model_input):
    outputs = []
    outputs = self.run(model_input.values[0][0])
    return outputs

def run(self, prompt):
    output = self.dspy_lm(prompt)
    return pd.DataFrame([output.answer])
```

BlueBot @ JetBlue - Endpoint

Serving endpoints >

dspy-agent-new

Serving endpoint state: 🟢 Ready
Created by: dp15886@jetblue.com
URL: https://adb-742956373569...
Tags:

Active configuration

Entity

- payroll_agent_dspy

Metrics Events Logs

Query endpoint ✕

Browser Curl Python SQL


Request ?

```
{"inputs":["How is drug testing pay paid out?"]}
```

Response ?
from payroll_agent_dspy-38

```
{  "predictions": [    {      "0": "Drug testing pay is paid out through payroll on the 8th paycheck after the month it was completed. This means that if the drug test was completed in January, it would be included on the February 8th paycheck."    }  ]}
```

BlueBot @ JetBlue

jetBlue  Bluebot Ravuru, Sai Pradhan

JetBlue's Internal FAQ chatbot

Before you begin please read below guidelines and responsible usage guidelines

Sample Questions :

- ✓ Do I have to return my uniform after employment ends ?
- ✓ What is the maximum number of weeks of FML can I request to take off in a year ?
- ✓ What does CASM stand for ?
- ✓ What kinds of time off programs does jetblue have ?
- ✓ How many days of unpaid time off can crewmembers request ?
- ✓ What is jetblue's PTO policy for Personal Leave ?

Important*

Please read our Privacy Policy, Terms And Conditions by clicking the below button

[Jetblue Terms Of Use](#)

[Jetblue Data Camp](#)

✓ By Continuing you agree to the privacy policy and terms and conditions

[Start New Conversation](#)



BlueBot @ JetBlue

File Explorer Bluebot RS Ravuru, Sai Pradhan Bluebot Admin

Organization Files Showing 9 Files Upload File + FAQ Bot v Hide Drag Files

Drag & Drop Files To Upload

SI No	File Name	Size	Chat Bot	Status	Modified	Actions
1	2022_JBLU_annualreport.pdf	977.95 KB	FAQ Bot	Published	5/22/2024 10:59 PDT	
2	2023_JBLU_annualreport.pdf	1.36 MB	FAQ Bot	Uploaded Failed Indexing	5/22/2024 10:59 PDT	
3	AI_Acceptable_Usage_Policy_AAUP 1.pdf	3.85 MB	FAQ Bot	Published	5/24/2024 07:09 PDT	
4	Business_Case_Policy_BCP.pdf	1.7 MB	FAQ Bot	Uploaded Failed Indexing	5/22/2024 10:59 PDT	
5	Crewmember_Blue_Book_CBB.pdf	4.95 MB	FAQ Bot	Published	5/22/2024 10:59 PDT	
6	Document.pdf	3.85 MB	FAQ Bot	Published	5/28/2024 07:45 PDT	
7	Seat Map - Copy - Copy (2).pdf	1.64 KB	FAQ Bot	Published	5/28/2024 10:29 PDT	
8	Seat Map.pdf	1.62 KB	FAQ Bot	Published	5/28/2024 07:43 PDT	



BlueBot @ JetBlue

The screenshot displays a chat window titled "Messages" with a header for "Bluebot" and a user profile for "Ravuru, Sai Pradhan". The chat history shows a conversation where the user asks "What is CASM?" and the bot provides a detailed explanation of Cost per Available Seat Mile, accompanied by a PDF link to the 2022 JBLU annual report. The user then asks "What is the boarding checklist for pilots?" and the bot provides a list of tasks including fuel level checks, flight plan reviews, and pre-flight inspections, with a link to the JetBlue sitemap.

Messages

FAQ Bot New Chat

Search

General

W What is CASM? 6

FAQ Bot

What is CASM? FAQ Bot

Wed May 22, 2024 at 12:04 PDT

Ravuru, Sai Pradhan

What is CASM? RS

Wed May 22, 2024 at 12:04 PDT

FAQ Bot

CASM stands for Cost per Available Seat Mile, which is a measure used by JetBlue Airways to calculate the cost of operating an available seat mile. It is a non-GAAP measure that excludes items beyond the company's control, such as fuel costs, and is believed to be more indicative of the company's ability to manage airline costs.

[2022_JBLU_annualreport.pdf](#) Open

Wed May 22, 2024 at 12:04 PDT

Ravuru, Sai Pradhan

What is the boarding checklist for pilots? RS

Wed May 22, 2024 at 12:04 PDT

FAQ Bot

The boarding checklist for pilots includes checking the aircraft's fuel levels, reviewing the flight plan and weather conditions, conducting a pre-flight inspection of the aircraft, and communicating with the flight attendants and ground crew. They also perform a final check of all systems and equipment before takeoff.

[jetblue-sitemap-.html](#) Open

Wed May 22, 2024 at 12:04 PDT

Type a message or type '/' to select a prompt Send



Securing with Theom

 Unique Theom Capabilities

1

Data Access Governance

Principle of Least Privilege; Data Access Compliance
Ideal State Discovery & Remediation

Discover Data, Policies &
Compliance Issues

Automate Remediation

2

Detect & Stop Data Breaches

Insider Threats & Data Leaks

Detect Insider Risks
& Active Data Attacks

Stop Data Attacks
(via ITSM/SIEM/SOAR)

3

Data Mesh & Contract Governance

License Discovery; Usage Tracking; Data Lineage;
Simulation; Financial Liability Visibility

Govern Vendor/Market
Data & Internal Data

Monetize Trust –
Data Products

4

AI Governance & Security

RAG/LLM/ML Governance and Security; Learning,
Inference/Serving Stack

Govern AI Models and
Vectors

Runtime Protection

THEOM FOUNDATION 5 W'S:



WHAT

What
Data?



WHERE

Where does it
flow? Data
Topology?



WHO

Who has access?
Do they need
access?



WHEN

When was it
accessed?

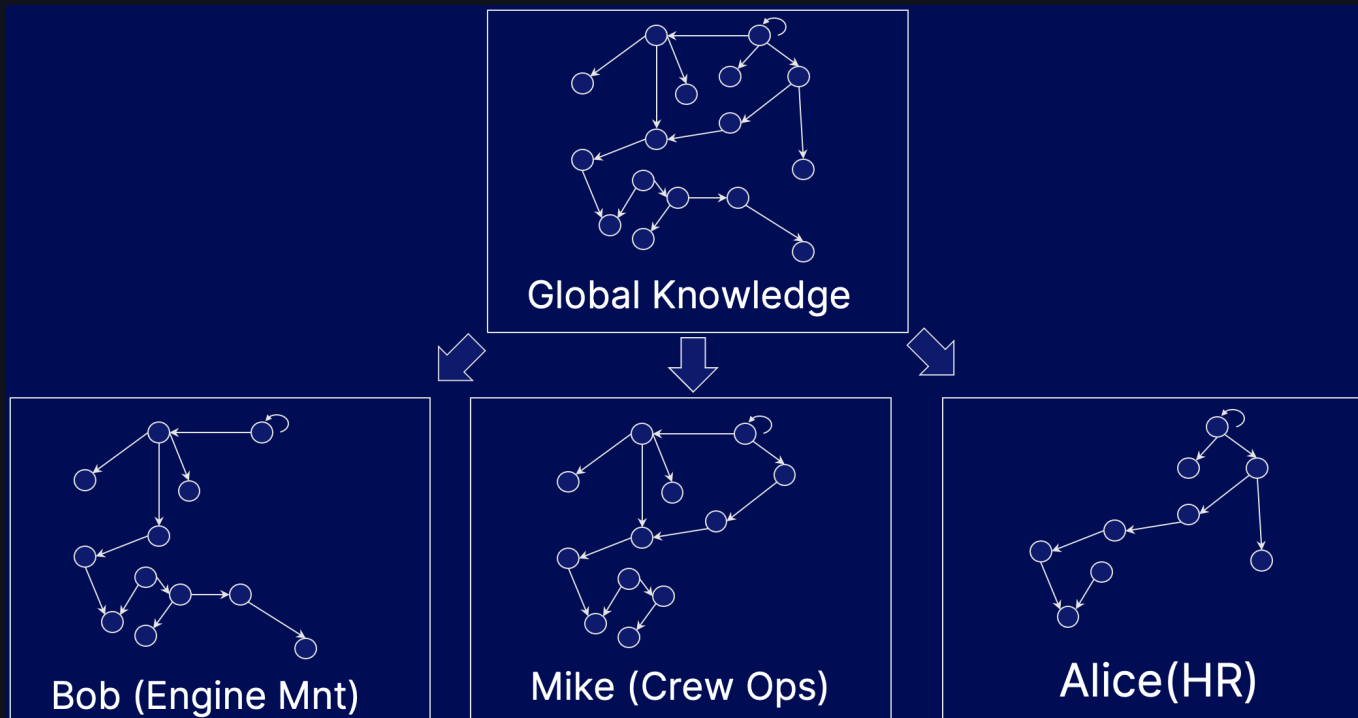


WHY

What's the purpose
of access?



Where Theom adds value



Theom RAG & Inference protection

