

DS684
Cloud Computing
Week 10

Regarding Labs and Assignments

- Class participation means more than Zoom attendance. You must actively participate in the discussion and labs, and answer questions.
- Must hit Submit button, otherwise no grade
- If you need extension in time, must send written request (**email**). Otherwise no grade and no makeup. Requests sent over Zoom chat do not count.
- For any technical difficulty (installation, Azure access, etc), you must send written explanation (**email**) before the deadline. Otherwise no grade and no makeup.

Regarding Labs and Assignments

Please submit all your labs and assignments **BEFORE** final presentation begins!

No make-up submissions after this point

Teaching Schedule

Week 7: Azure Synapse Analytics Part I: Data Warehouse

Week 8: Azure Synapse Analytics Part II: Data Engineering

Week 9: Visualization using Power BI

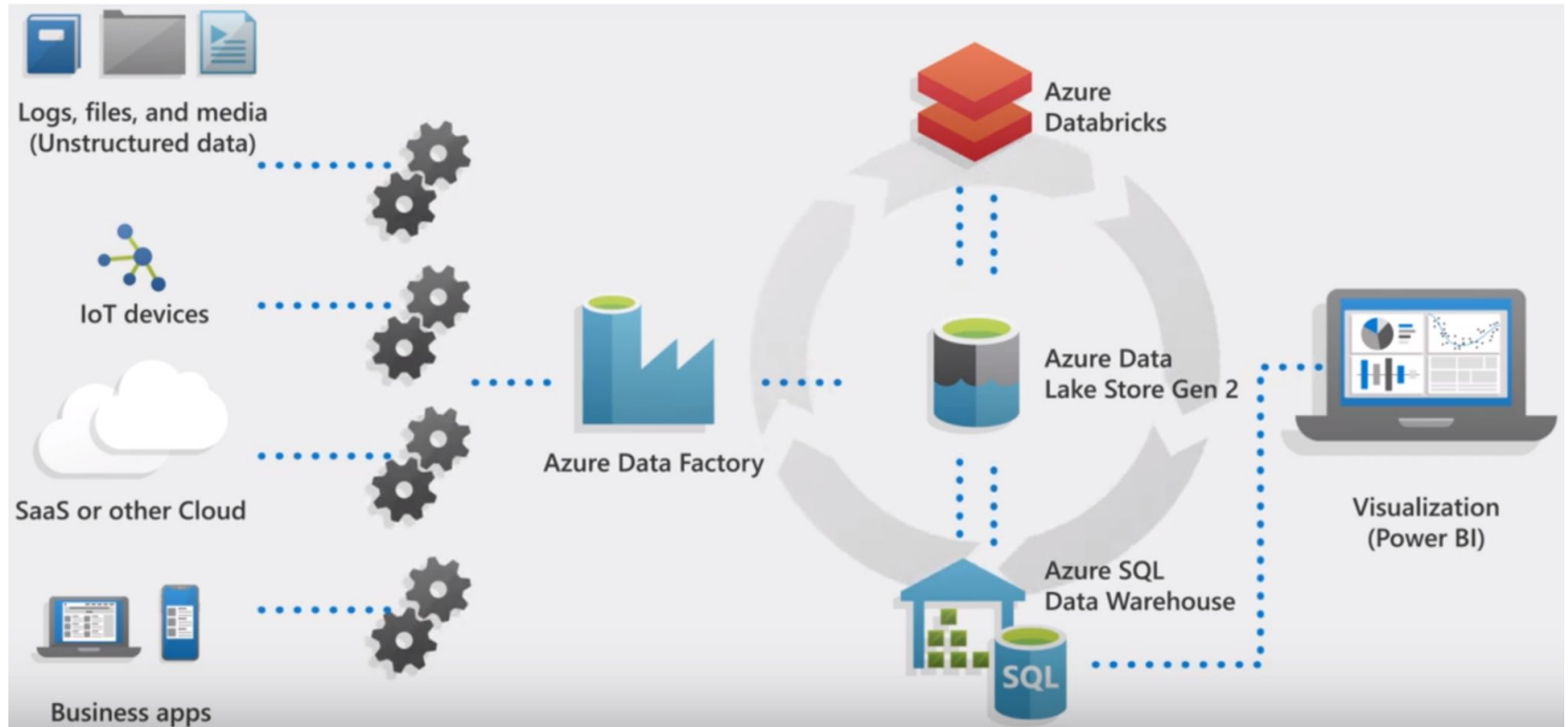
Week 10: Machine Learning and Artificial Intelligence

Week 11: Final project presentation

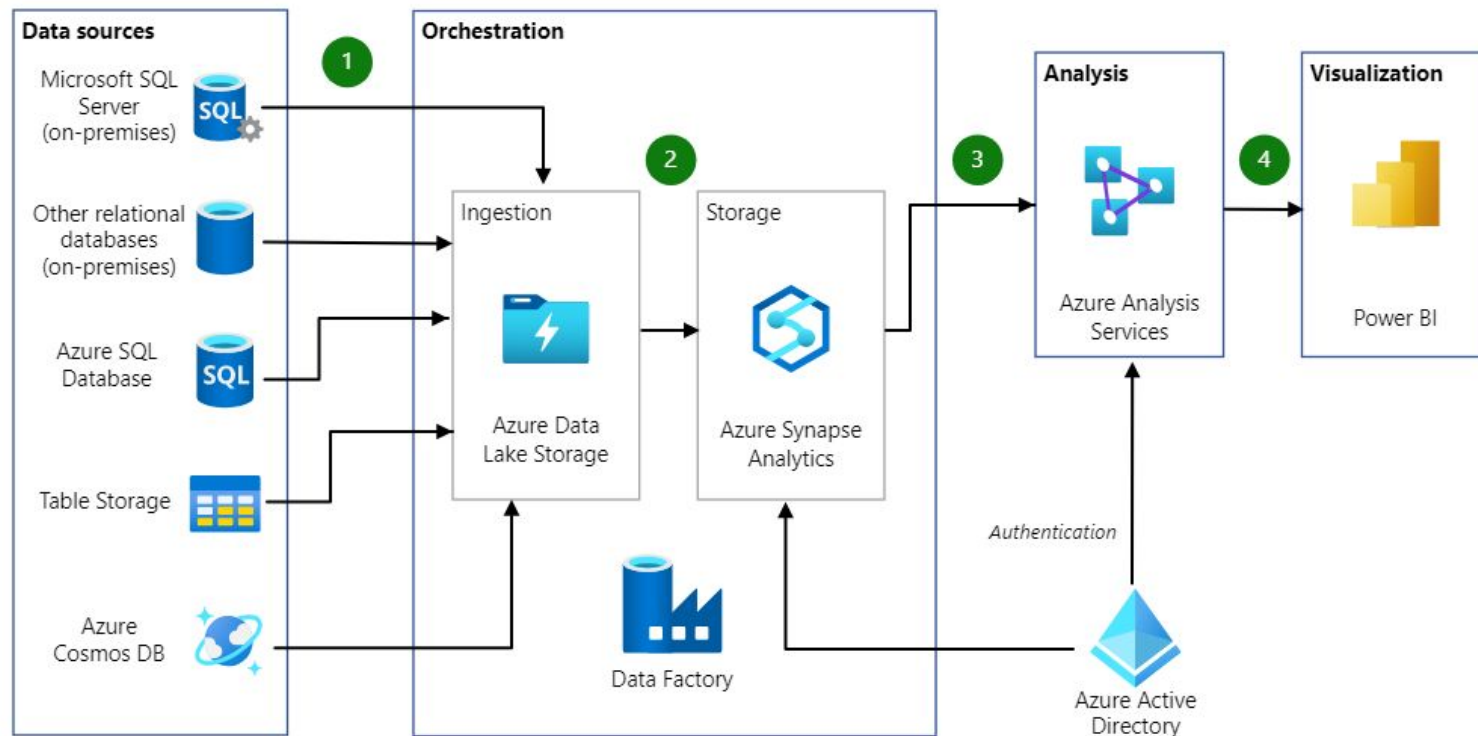
Agenda

- Surviving Different Microsoft Data Services
- Microsoft AI
 - Large Language Model (LLM) and Retrieval Augmented Generation (RAG)
 - Copilot
 - Azure AI Services
 - Azure Infrastructure
 - Azure Machine Learning
- Fabric Data Science
- Cloud Certifications

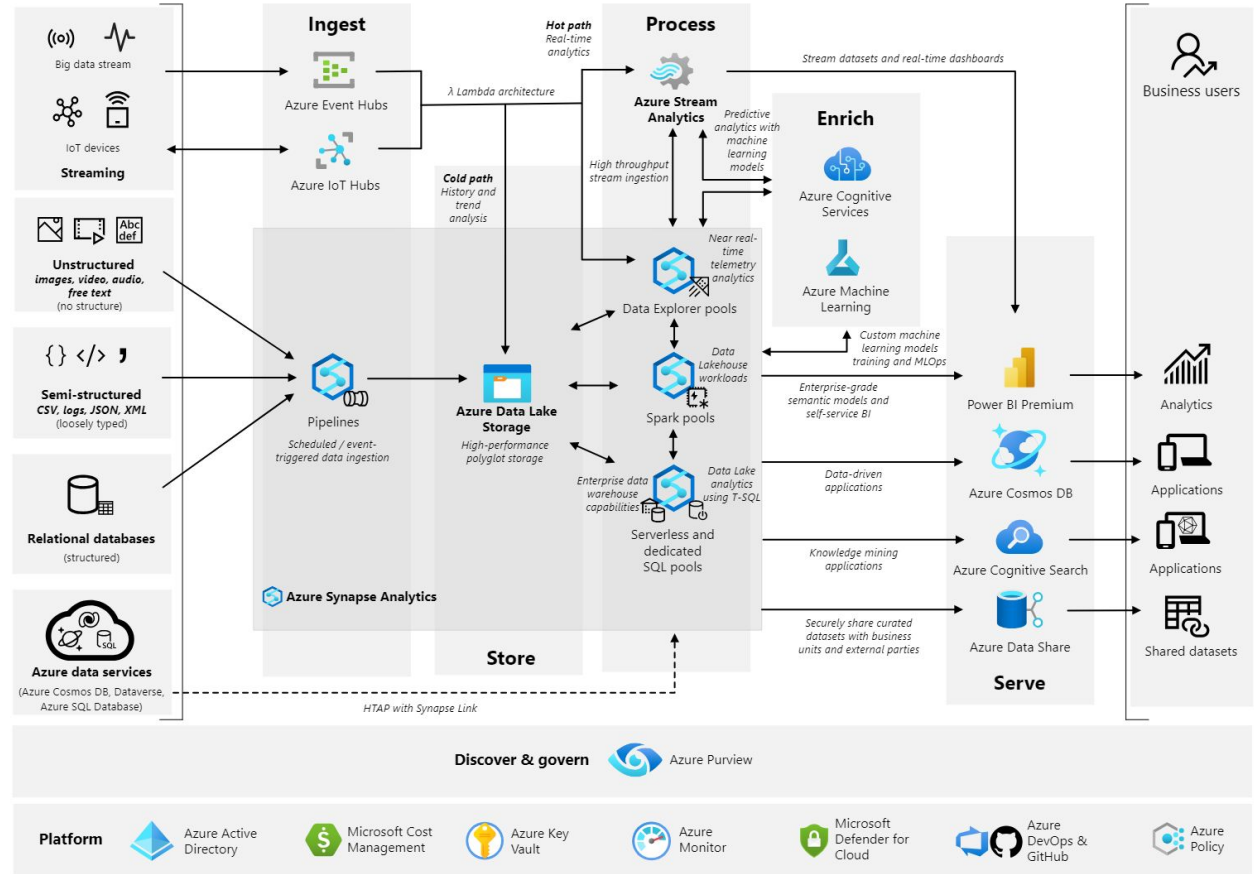
Back in 2015



How It Looked Like In 2021



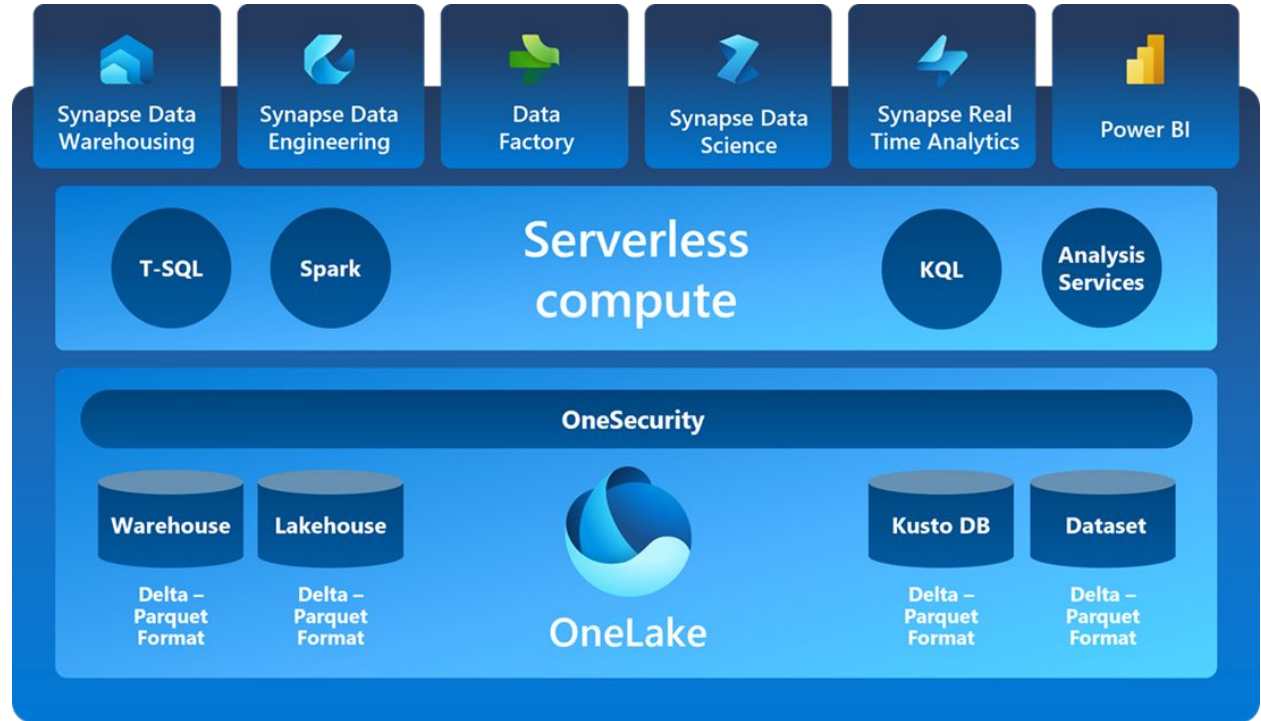
Synapse Analytics as of 2023



Microsoft Fabric

A new product that is still not fully matured (released in Nov 2023).

A combination of some most popular Synapse features and PowerBI features



Azure Data Factory

- Azure Data Factory: SSIS -> ADF -> Synapse Pipeline -> ADF -> Fabric Data Factory
 - 95% similar among these versions, with some differences

There are some mixed marketing messages from Microsoft. At one point it seems that Microsoft will phase out Azure Data Factory. But then it seems that Microsoft will phase out Synapse Pipelines instead. Then yet again Fabric comes with its own version of Data Factory.

However since they pretty much share the same core, learning either one is fine.

We will only use Spark notebooks in this class

But fundamentally...

Compute and Storage

- What do you want to do to the data, extracting (reading), transforming, loading (storing)?
 - Virtual machines
 - Azure data processing services
 - Your own SQL/python code
- Where are the datasets stored at, in what format?
 - Relational databases
 - Data lake files
 - NoSQL databases

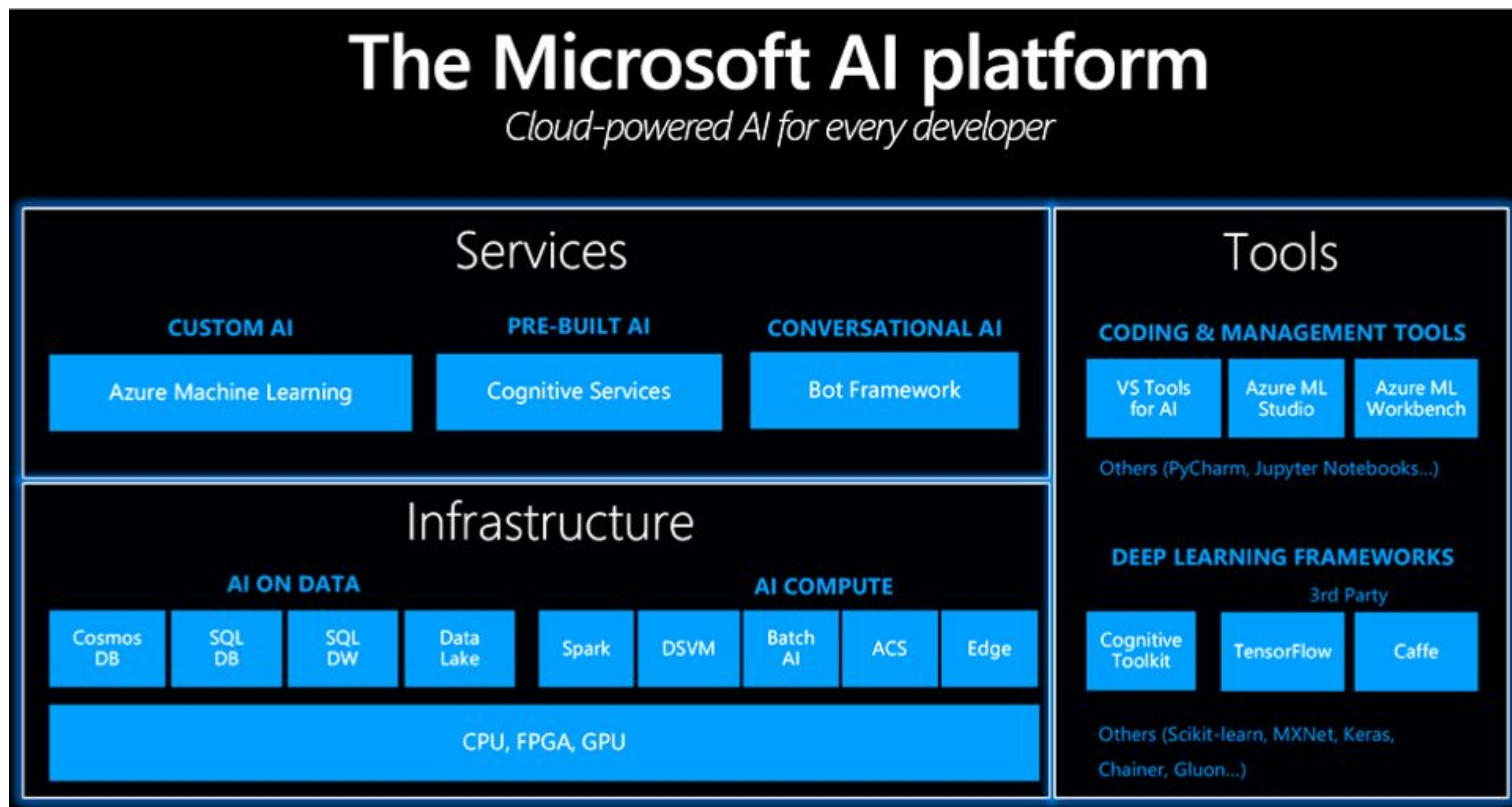
Agenda

- Surviving Different Microsoft Data Services
- Microsoft AI
 - Large Language Model (LLM) and Retrieval Augmented Generation (RAG)
 - Copilot
 - Azure AI Services
 - Azure Infrastructure
 - Azure Machine Learning
- Fabric Data Science
- Cloud Certifications

Agenda

- Surviving Different Microsoft Data Services
 - Fabric demo
- Microsoft AI
 - Large Language Model (LLM) and Retrieval Augmented Generation (RAG)
 - Copilot
 - Azure AI Services
 - Azure Infrastructure
 - Azure Machine Learning
- Fabric Data Science
- Cloud Certifications

Microsoft AI Offerings



Microsoft AI Offerings

Not only Azure, but the whole Microsoft

Microsoft AI: <https://www.microsoft.com/en-us/ai> > Solutions

- Large Language Model (LLM): OpenAI, CoPilot
- Many different AI products offered by different Microsoft products: Copilot
- Azure AI: <https://azure.microsoft.com/en-us/solutions/ai/>
 - Azure AI Services
 - Azure Machine Learning
 - Azure AI infrastructure

Agenda

- Surviving Different Microsoft Data Services
 - Fabric demo
- Microsoft AI
 - Large Language Model (LLM) and Retrieval Augmented Generation (RAG)
 - Copilot
 - Azure AI Services
 - Azure Infrastructure
 - Azure Machine Learning
- Fabric Data Science
- Cloud Certifications

LLM and RAG

- Large Language Model (LLM)
 - Machine learning models that can comprehend and generate human language text. They work by analyzing massive data sets of language.
- Retrieval Augmented Generation (RAG)
 - Optimizing the output of a large language model, so it references an authoritative knowledge base outside of its training data sources before generating a response

LLM AI as a Microsoft Offerings

LLM and OpenAI: <https://platform.openai.com/>

- Integrating LLM into existing frameworks
 - LLMs' capability and limit: Hallucination
 - Chunking and Embedding: customization for LLM answers
 - Similarity processing and vector database

LLM Limitations

JU

what does this course teach me



I'd be happy to help you understand what a course teaches, but I need more specific information about the course you're referring to. Courses can cover a wide range of subjects and topics, so if you could provide the name of the course or some details about its content, I can give you a more accurate answer.

Enhancing LLM with Customized Information

Find related information in your knowledge base, and add them as a part of your question to LLMs.



what does this course teach me The desired outcome of the course is the student's ability to put conceptual knowledge to practical use. Whether you are taking this course for future academic research, for work in industry, or for an innovative startup idea, this course should help you master the fundamentals of cloud computing.



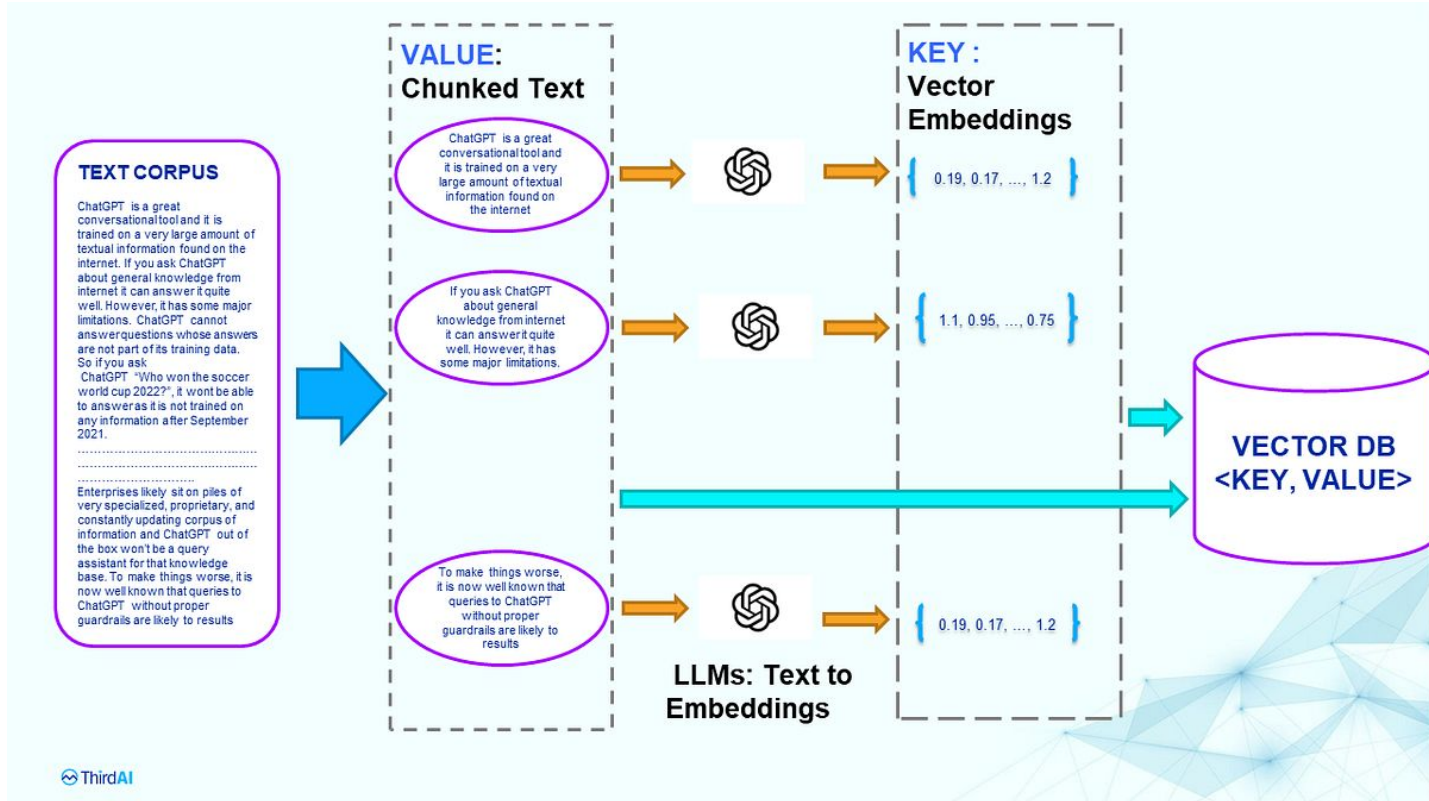
The course you described aims to teach students how to apply conceptual knowledge of cloud computing to practical situations. It has the following goals and intended outcomes:



1. **Conceptual Knowledge:** The course will provide you with a solid understanding of the theoretical and conceptual aspects of cloud computing. This might include topics like cloud architecture, virtualization, cloud service models (IaaS, PaaS, SaaS), security, scalability, and more.
2. **Practical Application:** The primary objective of the course is to equip you with the skills and knowledge necessary to apply what you've learned in real-world scenarios. This could involve tasks such as setting up and managing cloud-based infrastructure, deploying applications in the cloud, and optimizing cloud resources.

Regener

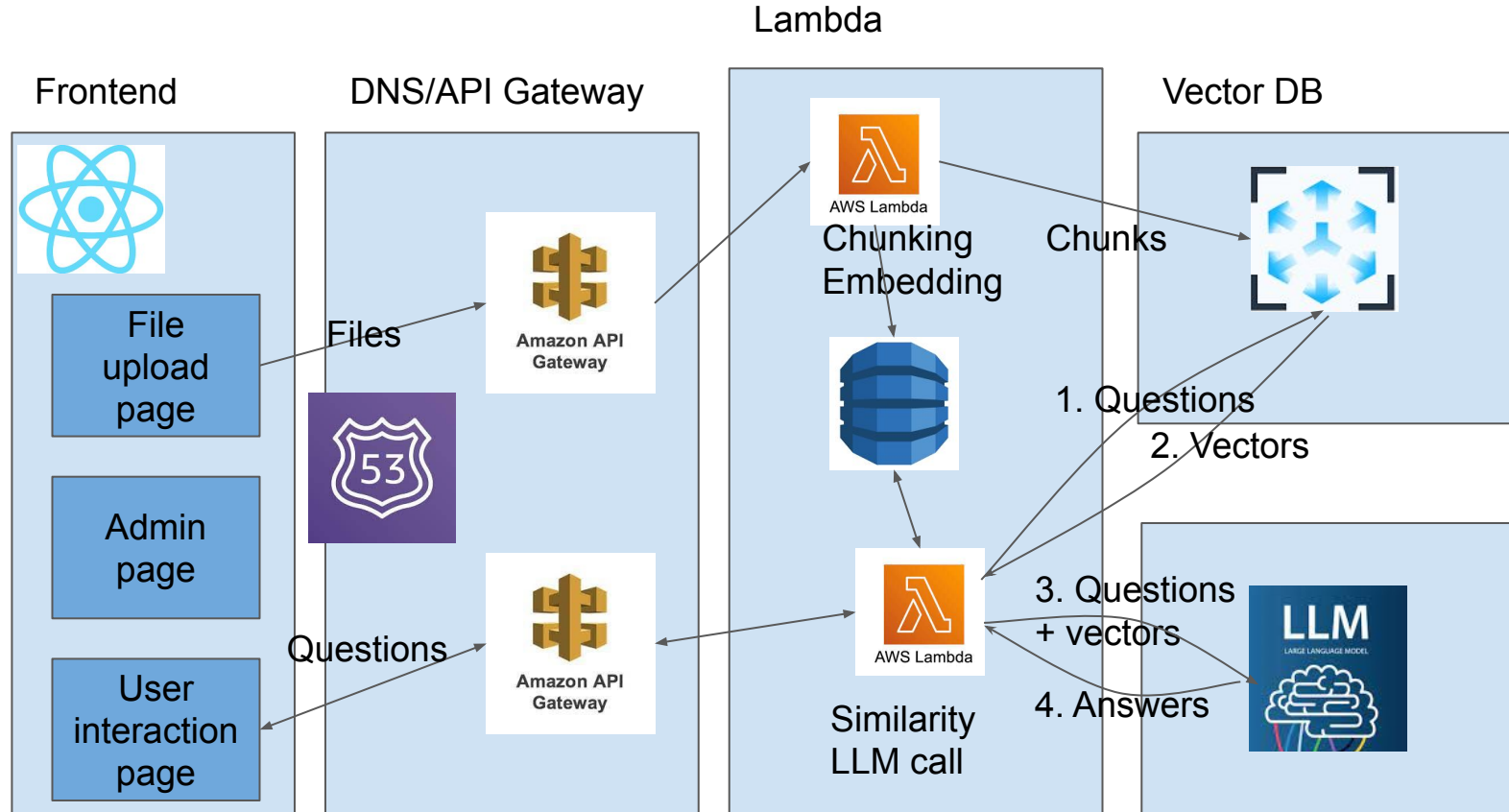
Chunking and Embedding



Similarity Processing and LLM Invoking

- Similarity processing: Finding the Top n (3) vectors that are closest to your question
- Invoking LLM: The original question + the most similar vectors

System Diagram Example



Agenda

- Surviving Different Microsoft Data Services
 - Fabric demo
- Microsoft AI
 - Large Language Model (LLM) and Retrieval Augmented Generation (RAG)
 - Copilot
 - Azure AI Services
 - Azure Infrastructure
 - Azure Machine Learning
- Fabric Data Science
- Cloud Certifications

Copilot

<https://adoption.microsoft.com/en-us/copilot/>

Microsoft's LLM AI powered chatbot, aiming at helping users using different MSFT products

Example:

- Prepare a meeting by researching on previous emails
- Audio to text conversion
- Summarize meeting minutes, and create action items
- Summarize action results (Excel spreadsheet, Word doc, Github code) and
- Write a follow-up email

Office Copilot

Use CoPilot with Excel or Word or other office products

Streamline workflow
(multiple meetings, e.g.)

The screenshot displays an Excel spreadsheet with three columns: Manufacturing Price, Sale Price, and Gross Sales. The data is as follows:

| Manufacturing Price | Sale Price | Gross Sales |
|---------------------|------------|-------------|
| \$10 | \$352 | \$1,500 |
| \$10 | \$352 | \$1,400 |
| \$10 | \$352 | \$1,400 |
| \$10 | \$352 | \$1,300 |
| \$10 | \$127 | \$620 |
| \$5 | \$127 | \$620 |
| \$270 | \$302 | \$590 |
| \$220 | \$302 | \$600 |
| \$3 | \$380 | \$470 |
| \$224 | \$302 | \$440 |
| \$120 | \$302 | \$420 |
| \$224 | \$302 | \$410 |
| \$3 | \$380 | \$390 |
| \$210 | \$380 | \$600 |

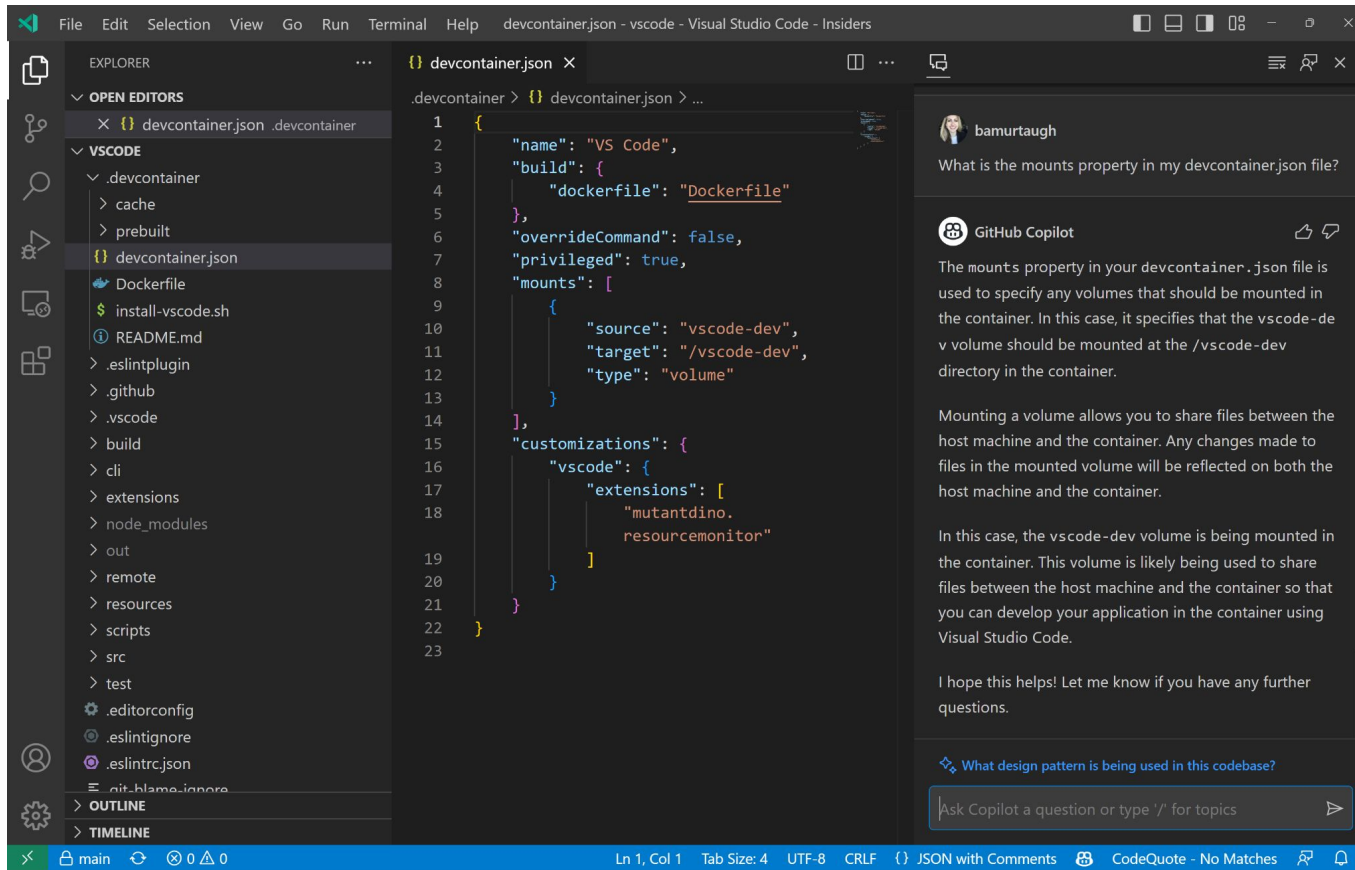
On the right, the Office Copilot sidebar is open, showing a green header bar with the Copilot logo and a close button. Below it, a green box contains the prompt: "results and summarize three key trends". A large green-bordered box highlights the "Key trends:" section, which lists:

- Sales are trending up for all customers, except Proseware which is down this quarter.
- Manufacturing costs are down this quarter after an increase in Q3.
- Low discount products made up nearly half of total sales last month.

Below the trends, there is an "Explain" button with a magnifying glass icon, and two thumbs up/down icons. At the bottom of the sidebar, there are two buttons: "Filter Discount Band to Low" and "Add a margin column" with a refresh icon.

Copilot

Provide help



The screenshot shows the Visual Studio Code editor with the `devcontainer.json` file open. The file content is as follows:

```
1 {
2   "name": "VS Code",
3   "build": {
4     "dockerfile": "Dockerfile"
5   },
6   "overrideCommand": false,
7   "privileged": true,
8   "mounts": [
9     {
10      "source": "vscode-dev",
11      "target": "/vscode-dev",
12      "type": "volume"
13    }
14  ],
15  "customizations": {
16    "vscode": {
17      "extensions": [
18        "mutantdino.resourcemonitor"
19      ]
20    }
21  }
22 }
23
```

On the right side, the GitHub Copilot chat interface is visible. It shows a conversation with a user named `bamurtaugh` asking: "What is the mounts property in my devcontainer.json file?". The Copilot response explains that the `mounts` property is used to specify volumes to be mounted in the container, specifically mentioning the `vscode-dev` volume and the `/vscode-dev` directory. It also notes that mounting a volume allows for sharing files between the host machine and the container.


At the bottom of the chat interface, there is a prompt: "What design pattern is being used in this codebase?". Below this, a text input field contains the text: "Ask Copilot a question or type '/' for topics".

Copilot

Provide coding support

```
3
4 function bubbleSort(arr: number[]): number[] {
5     const n = arr.length;
6     for (let i = 0; i < n - 1; i++) {
7         for (let j = 0; j < n - i - 1; j++) {
8             if (arr[j] > arr[j + 1]) {
9                 const temp = arr[j];
10                arr[j] = arr[j + 1];
11                arr[j + 1] = temp;
12            }
13        }
14    }
15    return arr;
16 }
```

convert this to use quick sort

 Copilot generated code may be incorrect

Azure AI Studio

Latest product. Still in the making.

Agenda

- Surviving Different Microsoft Data Services
 - Fabric demo
- Microsoft AI
 - Large Language Model (LLM) and Retrieval Augmented Generation (RAG)
 - Copilot
 - Azure AI Services
 - Azure Infrastructure
 - Azure Machine Learning
- Fabric Data Science
- Cloud Certifications

Microsoft AI Offerings

Not only Azure, but the whole Microsoft

Microsoft AI: <https://www.microsoft.com/en-us/ai> > Solutions

- Many different AI products offered by different Microsoft products: Copilot
- Azure AI: <https://azure.microsoft.com/en-us/solutions/ai/>
 - Azure AI Services
 - Azure Machine Learning
 - Azure AI infrastructure

Azure AI Services

Prepackaged AI Services that you can utilize: model trained by Microsoft

<https://azure.microsoft.com/en-us/products/ai-services/>

- Vision
- Speech
- Language
- Translator
- Documents
- Audio and video
- Decision
- Metrics Advisor

AI Services

Usage example 1: Online customer support (real estate management)

- Image recognition (text labeling)
- Large Language Model text processing

Usage example 2: Online monitoring of customer support quality (healthcare)

- Speech to text
- Text sentiment analysis

Agenda

- Surviving Different Microsoft Data Services
 - Fabric demo
- Microsoft AI
 - Large Language Model (LLM) and Retrieval Augmented Generation (RAG)
 - Copilot
 - Azure AI Services
 - Azure AI Infrastructure
 - Azure Machine Learning
- Fabric Data Science
- Cloud Certifications

Azure AI Infrastructure

You run your own ML/AI, with Azure infrastructure

<https://azure.microsoft.com/en-us/solutions/high-performance-computing/ai-infrastructure/>

- VMs, with pre-configured images and GPUs

Agenda

- Surviving Different Microsoft Data Services
 - Fabric demo
- Microsoft AI
 - Large Language Model (LLM) and Retrieval Augmented Generation (RAG)
 - Copilot
 - Azure AI Services
 - Azure Infrastructure
 - Azure Machine Learning
- Fabric Data Science
- Cloud Certifications

Azure Machine Learning

You run your own AI/ML, with tools from Microsoft

<https://azure.microsoft.com/en-us/products/machine-learning/>

- Prepare data
- Build and train AI and ML models
- Validate and deploy
- Manage and monitor

Azure Machine Learning Workspace

<https://ml.azure.com/>

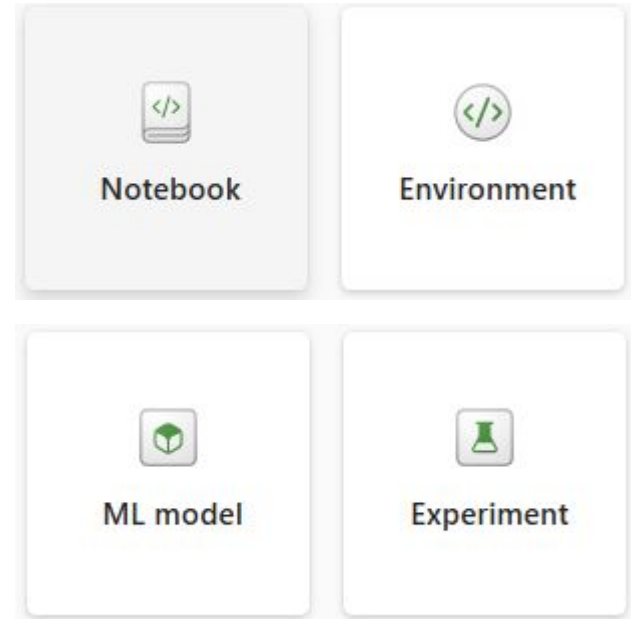
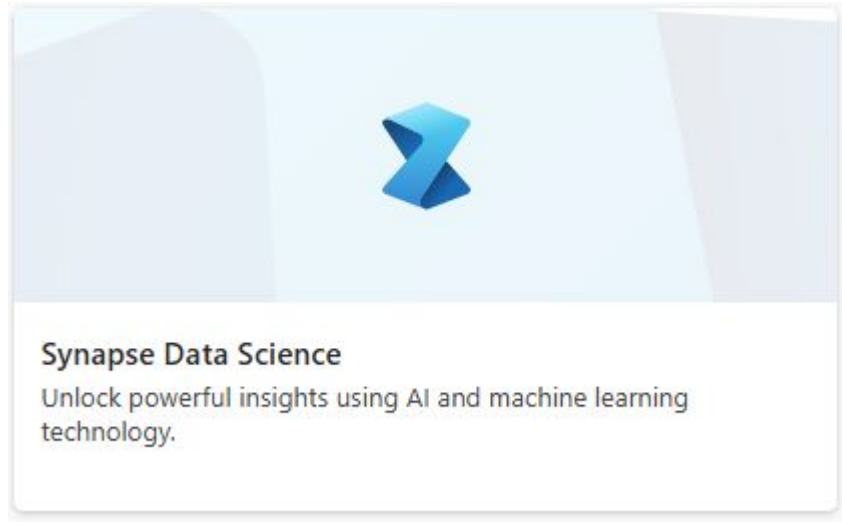
- Data: Data assets, Datastores
- Notebook
- Pipeline:
 - Data
 - Components: Data, Transformation, etc.
- Models and Jobs

Agenda

- Surviving Different Microsoft Data Services
 - Fabric demo
- Microsoft AI
 - Large Language Model (LLM) and Retrieval Augmented Generation (RAG)
 - Copilot
 - Azure AI Services
 - Azure Infrastructure
 - Azure Machine Learning
- Fabric Data Science
- Cloud Certifications

Fabric Data Science

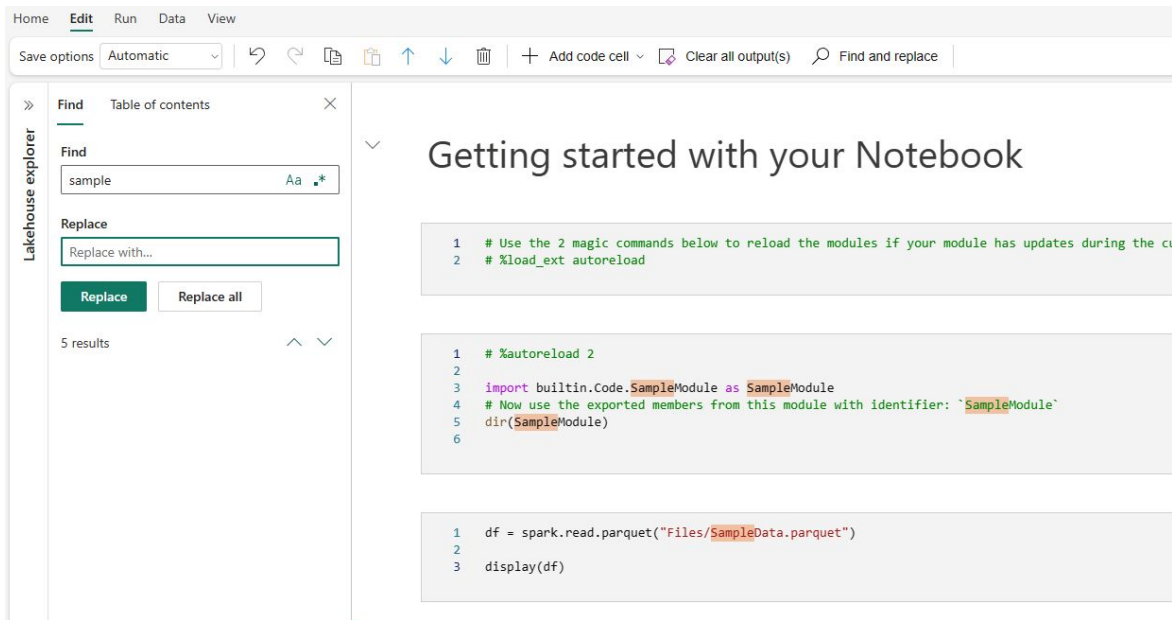
- Direct usage of Synapse Data Science
- Allows normal data science operations



Fabric Data Science Notebook

Same notebook as what we have in Data Engineer

- More interactions
- More customization of environment: package, compute power
- Same need for python/SQL power
- Less orchestration



Environment

Use environment to ensure the setup is consistent and can collaborate with other team members

The screenshot shows the DWH Setup Notebook interface. The top bar includes the notebook name 'DWH Setup Notebook', a 'Saved' status, and a search bar. Below the top bar, there's a toolbar with icons for file operations, a 'Run all' button, and a 'Connect' dropdown. The main interface is divided into three sections: 'Explorer' on the left, a central workspace, and a right sidebar. The 'Explorer' section has a 'Data sources' button and three categories: 'Resources' (Uploaded data and files), 'Lakehouses' (2 item(s) added), and 'Warehouses' (0 item(s) added). The central workspace shows a 'Source data location' with a URL 'https://www.kaggl...' and a 'select' button. Below this, there's a table with 4 rows and 5 columns. The right sidebar has an 'Environment' dropdown menu open, showing 'Workspace default' and 'Available environments' including 'MyEnvironment' and 'pallavi'. A 'New environment' button is at the bottom of the dropdown.

DWH Setup Notebook | Saved

Find detailed information about this artifact

Run all

Connect

Spark SQL

Environment

Workspace default

Data Wrangler

Filter by keyword

Workspace default

✓ Workspace Settings
Runtime: 1.2 (Spark 3.4, Delta 2.4), Compute: Medium, 1-3 nodes

Available environments

MyEnvironment
Runtime: 1.2 (Spark 3.4, Delta 2.4), Compute: Medium, 1-3 nodes

pallavi
Runtime: 1.2 (Spark 3.4, Delta 2.4), Compute: Medium, 1-3 nodes

New environment

| Table | ABC Customer | 1-2-333 | 2-3-444 | 3-4-555 | 4-5-666 |
|-------|--------------|---------|---------|---------|---------|
| 1 | 1-2-333 | | | | |
| 2 | 2-3-444 | | | | |
| 3 | 3-4-555 | | | | |
| 4 | 4-5-666 | | | | |

Environment

Defines the software (library)
and hardware (compute)
configurations of the notebook

The screenshot shows the 'MyEnvironment' configuration page in Databricks. The interface includes a top navigation bar with a search bar and a 'Home' tab. Below the navigation bar, there are tabs for 'Libraries', 'Publish', 'Runtime', and '1.2 (Spark 3.4, Delta 2.4)'. The left sidebar contains a list of configuration categories: 'Libraries' (with sub-items: Built-in Libraries, Public libraries, Custom libraries), 'Spark compute' (with sub-items: Compute, Spark properties), and 'Storage' (with sub-item: Resources). The 'Compute' item is currently selected. The main content area is titled 'Spark compute configuration' and includes a description: 'This configuration applies to all notebooks and and Spark job definitions in this env customize. [Learn more](#)'. Below this, there is a section for 'Environment pool' with a dropdown menu set to 'Default pool'. A 'Pool details' table is displayed, showing 'Node family' as 'Auto (Memory optimized)', 'Node size' as 'Medium', and 'Number of nodes' as '1 - 3'. At the bottom, there are configuration options for 'Spark driver core' (set to 8) and 'Spark driver memory' (set to 56GB). The 'Spark executor core' option is partially visible at the bottom.

MyEnvironment ▾

Search

Home

⌨️ Publish Runtime 1.2 (Spark 3.4, Delta 2.4) ▾

Libraries

- Built-in Libraries
- Public libraries
- Custom libraries

Spark compute

- Compute**
- Spark properties

Storage

- Resources

Spark compute configuration

This configuration applies to all notebooks and and Spark job definitions in this env customize. [Learn more](#)

Environment pool

Default pool ▾

| Pool details ⓘ | | |
|-------------------------|-----------|-----------------|
| Node family | Node size | Number of nodes |
| Auto (Memory optimized) | Medium | 1 - 3 |

Compute

Spark driver core

8 ▾

Spark driver memory

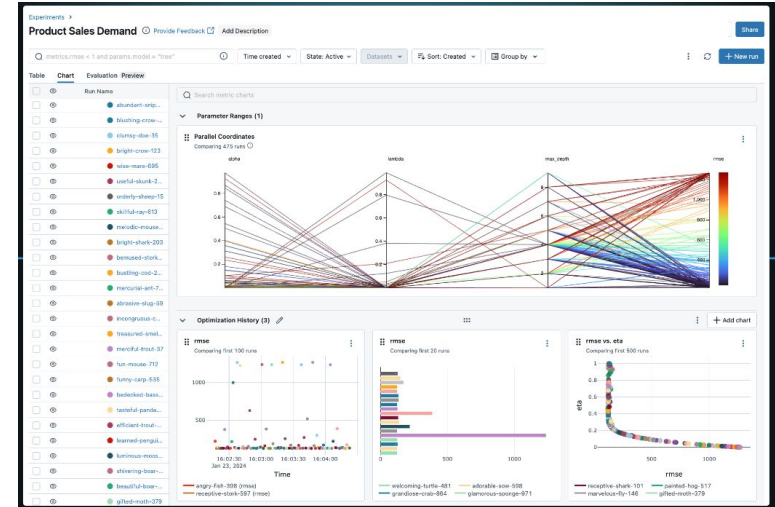
56GB ▾

Spark executor core

Machine Learning Framework

MLFlow <https://mlflow.org/>: an open-source platform that helps manage the machine learning lifecycle.

- Tracking: Allows users to record and compare experiment parameters and results
- Model packaging: Provides a generic way to package ML models
- Code structuring: Helps improve reproducibility
- Dependency management: Helps manage dependencies



MLFlow Experiment

- Define a notebook with ML model code
- Run your experiment from notebook
- Record run results (metrics)
- Publish runs as models

The screenshot displays the MLFlow web interface. On the left, the 'Explorer' sidebar shows 'Data sources', 'Resources' (Uploaded data and files), 'Lakehouses' (0 item(s) added), and 'Warehouses' (0 item(s) added). The main area shows a code notebook with the following Python code:

```
1 import mlflow
2 mlflow.set_experiment("MyExp")
3
4 # Start a new run
5 with mlflow.start_run():
6     # Log parameters, metrics, and artifacts within the run
7     mlflow.log_param("learning_rate", 0.01)
8     mlflow.log_metric("accuracy", 0.89)
9     # Train your model and log any output files as artifacts
10
```

Below the code, the execution status is shown: [1] ✓ - Session ready in 9 sec 893 ms. Command executed in 7 sec 362 ms by Jun Shan on 8:23:17 PM, 10/4/2024. Below this, there are tabs for 'Run list' and 'Run comparison'. The 'Run list' tab is active, showing a table with the following data:

| Run name | Start time |
|-----------------------|-------------------|
| modest_feast_d91cgv72 | 10/4/2024 8:23 PM |

The screenshot displays the MLFlow web interface for a specific run. The left sidebar shows the 'MyExp' experiment with a search bar and a list of runs. The selected run is 'modest_feast...' with a green checkmark and the date '10/4/2024'. Below the list, it says 'No files were logged for viewing.' The main area shows the 'Properties' section for the selected run. It includes a 'Description' field, a 'Run name' field, a 'Start date' field, a 'Duration' field, and a 'Status' field. The 'Run ID' is 'f11567e3-1ed9-...' and the 'Created by' is 'Jun Shan'. The 'Source' is 'MyExp-2607'. The 'Run details' section is expanded, showing 'Run metrics (1)' and 'Run Parameters (1)'. The metrics table shows 'accuracy' with a value of '0.89'. The parameters table shows 'learning_rate' with a value of '0.01'.

| Run name | Start date | Duration | Status |
|----------------------|-------------------|----------|-----------|
| modest_feast_d91c... | 10/4/2024 8:23 PM | 3s | Completed |

| Run ID | Created by | Source |
|-------------------|------------|------------|
| f11567e3-1ed9-... | Jun Shan | MyExp-2607 |

| Run metrics (1) | |
|-----------------|------|
| accuracy | 0.89 |

| Run Parameters (1) | |
|--------------------|------|
| learning_rate | 0.01 |

MLFlow Model

- Model can be published and everyone else can use
- With parameters and metrics, user can pick the best fitting one

The screenshot displays the MLFlow Model interface for 'Model1'. The left sidebar shows 'Version 1' selected, with a timestamp of '10/4/2024'. The main panel is divided into two sections: 'Properties' and 'Version details'.

Properties

Description

Version name: Version 1 Created time: 10/4/2024 8:46 PM Last modified: 10/4/2024 8:46 PM Created by: Jun Shan

Experiment name: MyExp Run name: modest_feast_d91...

Version details

ML model version metrics (1)

| | |
|----------|------|
| accuracy | 0.89 |
|----------|------|

ML model Parameters (1)

| | |
|---------------|------|
| learning_rate | 0.01 |
|---------------|------|

MLFlow Model

Sample code to invoke
the model

Copy code to apply ML model predictions



Copy this code into a new or existing notebook to use this ML model for scoring.

Preview code snippet

Pyspark

Copy to clipboard

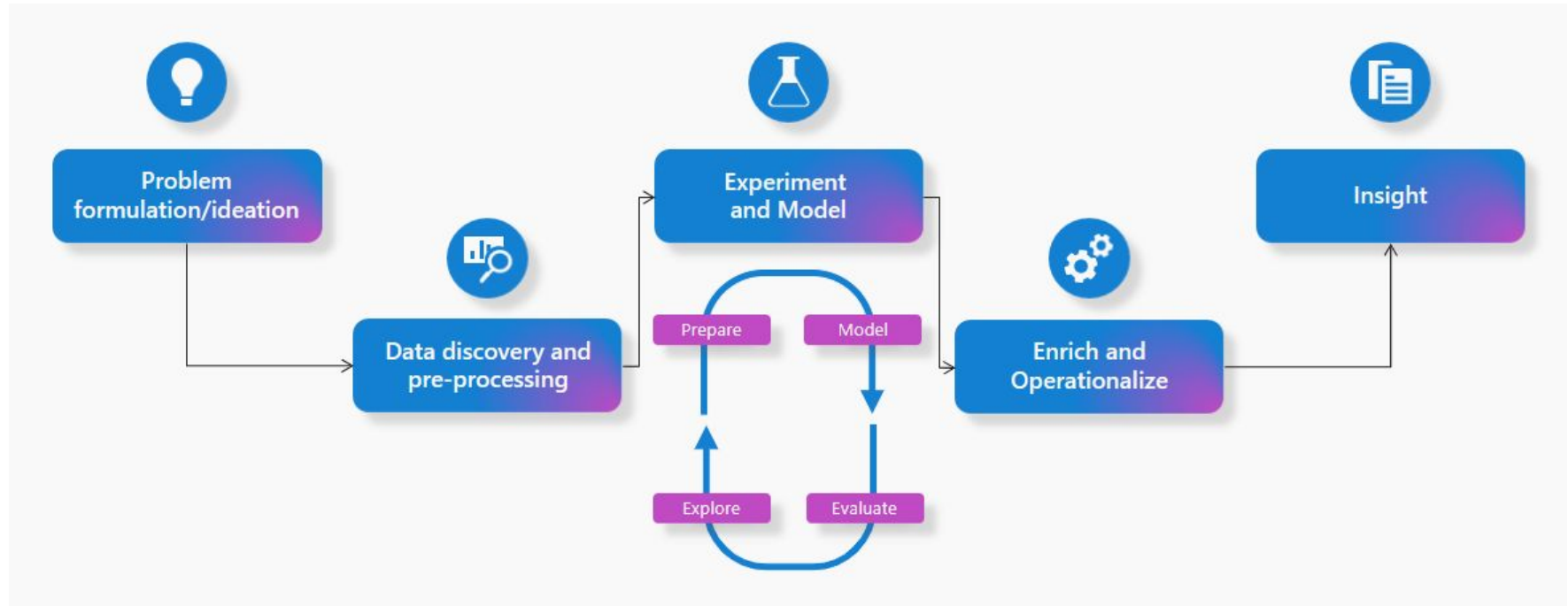
```
import mlflow
from synapse.ml.predict import MLFlowTransformer

df = spark.read.format("delta").load(
    <INPUT_TABLE> # Your input table filepath here
)

model = MLFlowTransformer(
    inputCols=[], # Your input columns here
    outputCol="predictions", # Your new column name here
    modelName=<MODEL_NAME>, # Your model name here
    modelVersion= # Your model version here
)
df = model.transform(df)

df.write.format('delta').mode("overwrite").save(
    <OUTPUT_TABLE> # Your output table filepath here
)
```

MLFlow Model



AI Startups

- Traditional industry
- Labor intensive

Insurance sign up process

Security monitoring

Fashion design

Agenda

- Surviving Different Microsoft Data Services
 - Fabric demo
- Microsoft AI
 - Large Language Model (LLM) and Retrieval Augmented Generation (RAG)
 - Copilot
 - Azure AI Services
 - Azure Infrastructure
 - Azure Machine Learning
- Fabric Data Science
- Cloud Certifications

Microsoft Certifications

There are many certifications. The following slides only listed what I feel can be useful for your career.

Each certification costs time and money. So choose the path carefully

Microsoft Certifications

Fundamentals

| |
|--|
| Exam PL-900: Microsoft Power Platform Fundamentals |
| Exam AI-900: Microsoft Azure AI Fundamentals |
| Exam DP-900: Microsoft Azure Data Fundamentals |
| Exam AZ-900: Microsoft Azure Fundamentals |

Microsoft Certifications

Intermediate (data track) - Not official Microsoft term

| |
|--|
| Exam PL-300: Microsoft Power BI Data Analyst |
| Exam DP-100: Designing and Implementing a Data Science Solution on Azure |
| Exam DP-203: Data Engineering on Microsoft Azure |
| Exam DP-300: Administering Microsoft Azure SQL Solutions |

Microsoft Certifications

Intermediate (AI track) - Not official Microsoft term

Exam AI-102: Designing and Implementing a Microsoft Azure AI Solution

Microsoft Certifications

Intermediate (Azure engineer track) - Not official Microsoft term

| |
|---|
| Exam AZ-104: Microsoft Azure Administrator |
| Exam AZ-204: Developing Solutions for Microsoft Azure |

Microsoft Certifications

Intermediate (Specialty track) - Not official Microsoft term

| |
|---|
| Exam DP-420: Designing and Implementing Cloud-Native Applications Using Microsoft Azure Cosmos DB |
| Exam AZ-500: Microsoft Azure Security Technologies |
| Exam AZ-700: Designing and Implementing Microsoft Azure Networking Solutions |

Microsoft Certifications

Analytics focused

Exam DP-500: Designing and Implementing Enterprise-Scale Analytics Solutions Using Microsoft Azure and Microsoft Power BI (retired)

Exam DP-600: Implementing Analytics Solutions Using Microsoft Fabric

Exam DP-700: Implementing Data Engineering Solutions Using Microsoft Fabric (beta)

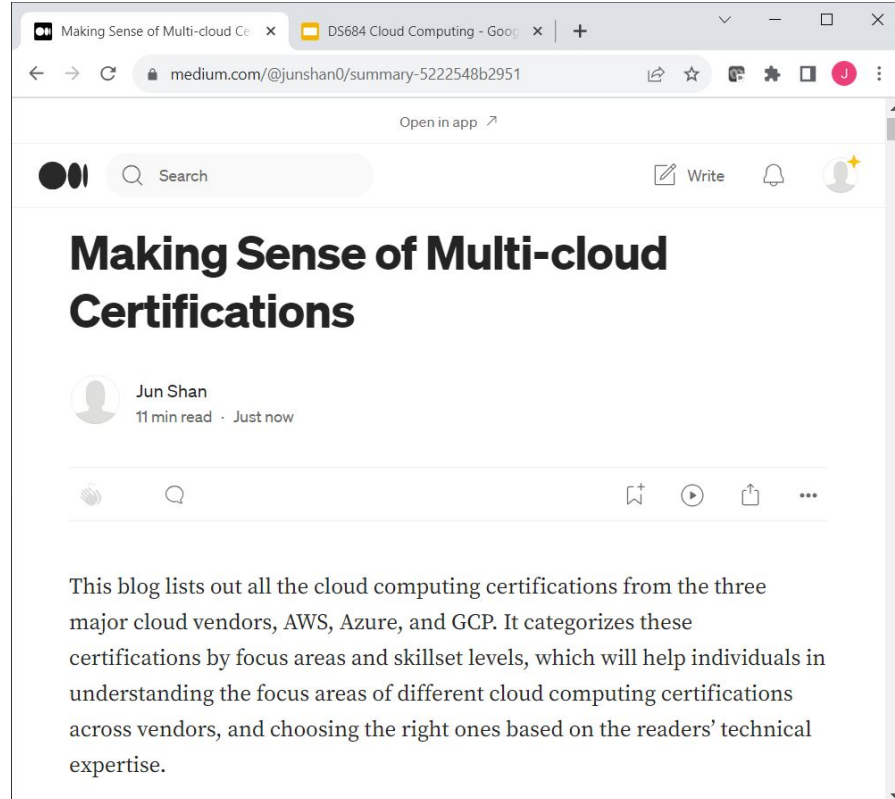
Microsoft Certifications

Advanced

| |
|--|
| Exam AZ-305: Designing Microsoft Azure Infrastructure Solutions |
| Exam AZ-400: Designing and Implementing Microsoft DevOps Solutions |
| Exam PL-600: Microsoft Power Platform Solution Architect |

Multi-cloud Certifications

[https://medium.com/
@junshan0/summary
-5222548b2951](https://medium.com/@junshan0/summary-5222548b2951)



Plan for Final Project Presentation

Expectations

Time slots