

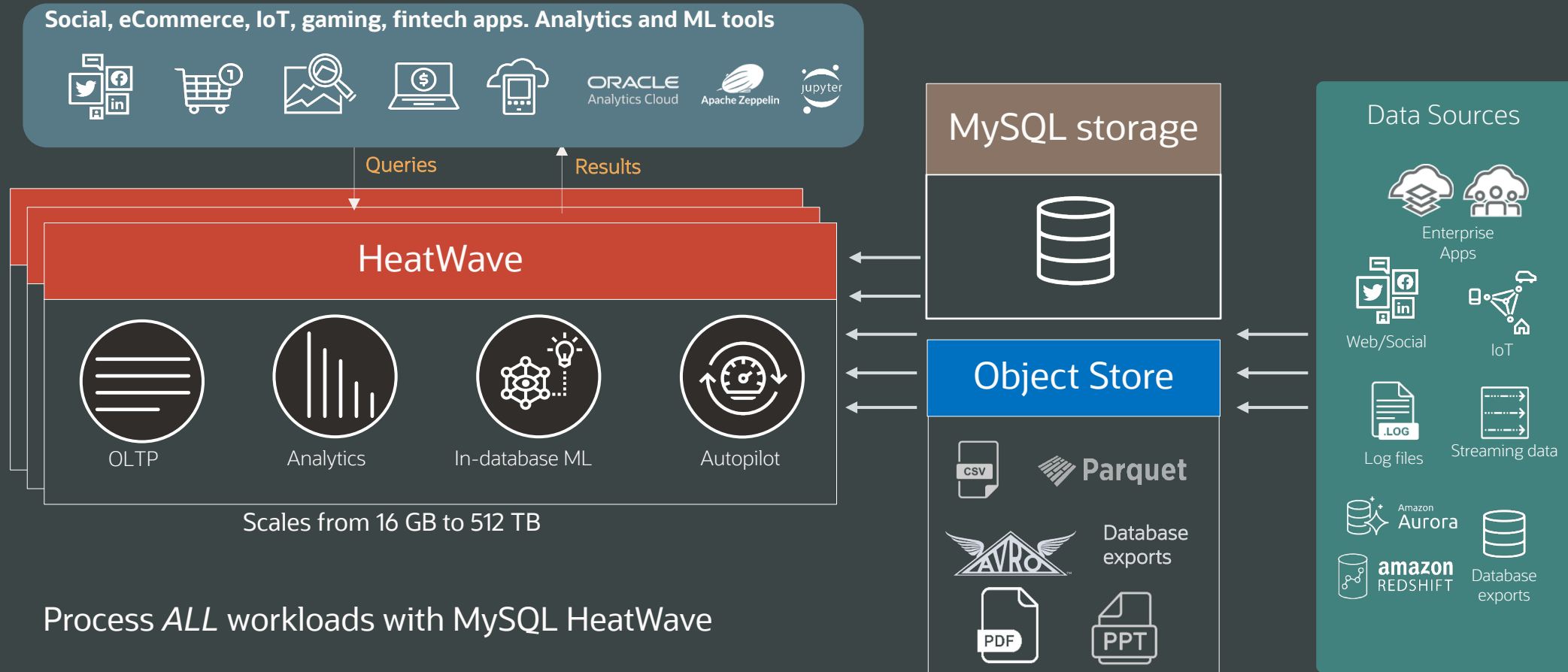
Building New Applications using HeatWave GenAI & Vector Store

Nipun Agarwal

Senior Vice President, MySQL HeatWave

MySQL HeatWave

LAKEHOUSE, REAL-TIME ANALYTICS, MACHINE LEARNING, TRANSACTION PROCESSING IN ONE DATABASE SERVICE

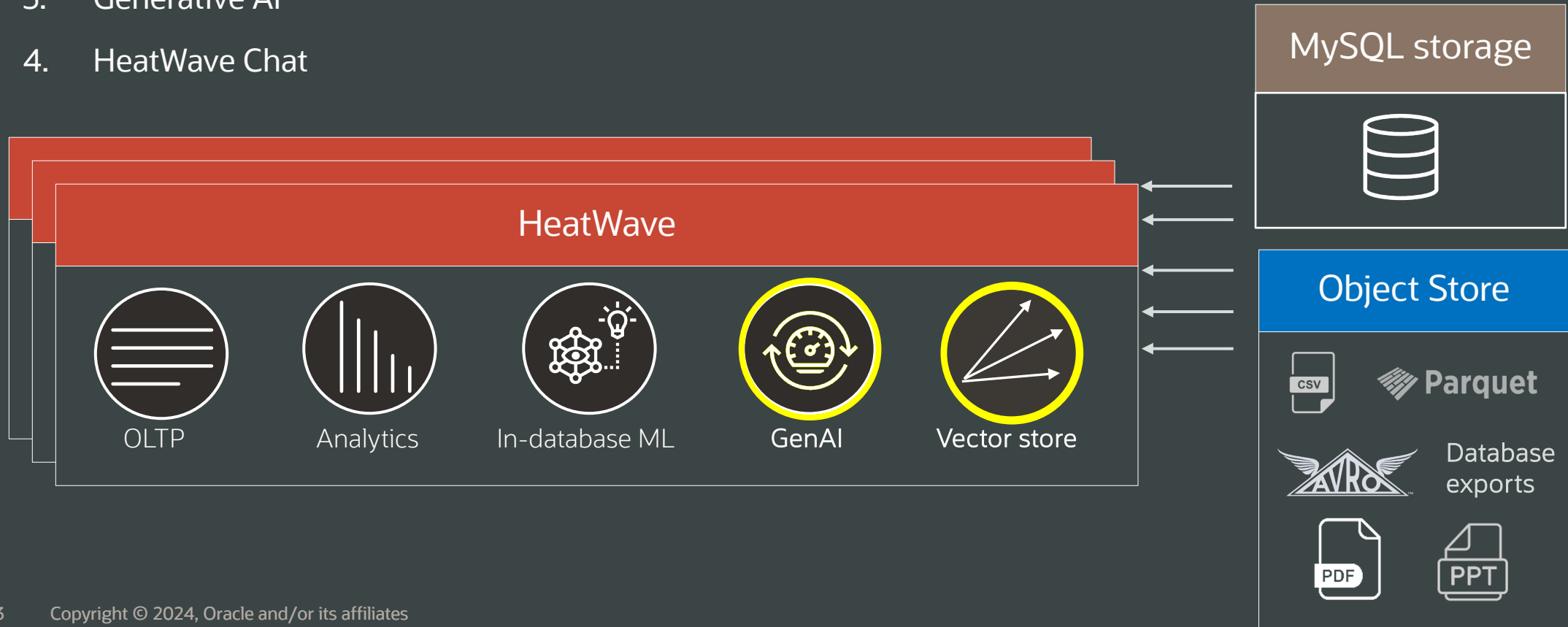


Process ALL workloads with MySQL HeatWave



HeatWave Generative AI and Vector Store

1. Vector store
2. Vector processing
3. Generative AI
4. HeatWave Chat

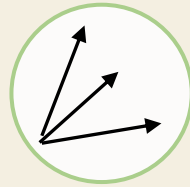


Generative AI in HeatWave enables new use cases



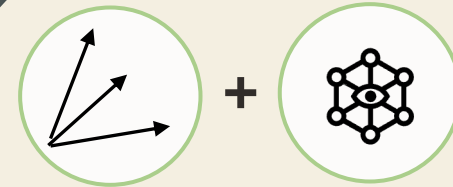
Content generation & summarization

- Generate insights from enterprise documents
- Generate blogs from pdf instruction manuals
- Summarize logs for root cause analysis



Retrieval Augmented Generation

- Search on public and private enterprise data
- Search on unstructured data in vector store



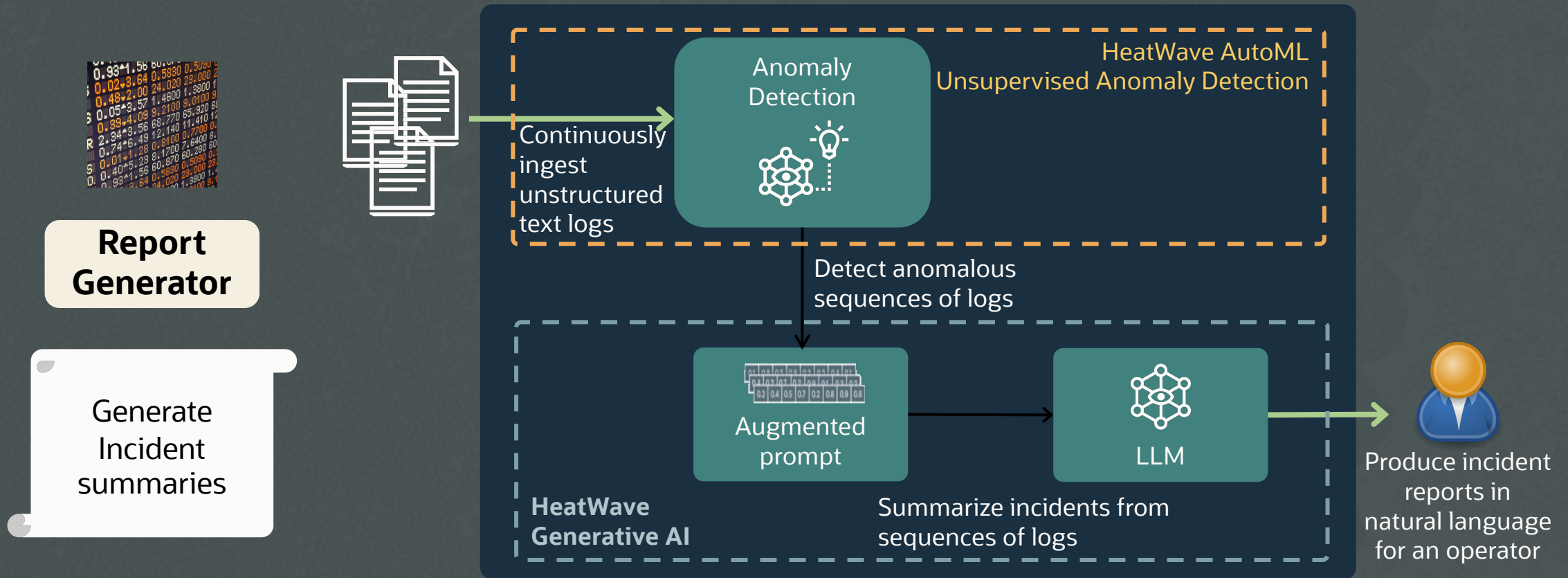
Natural language interaction

- Natural language interaction with proprietary unstructured data
- Personalized content retrieval and response back in natural language

Usecase 1: Predictive maintenance

CONTENT GENERATION/SUMMARY

Report Generation: HW AutoML + Summarization of anomalous logs



Usecase 1: Predictive maintenance

CONTENT GENERATION / SUMMARIZATION

Report Generation: Summarization of anomalous logs

```
[{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777883620,"msg":"[8145:140643603412800][fit_transform:140][ 1.939] fit_transform: Remove non-categorical columns with > 20% missing values. Dropping 0 columns"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777883805,"msg":"[8145:140643603412800][fit_transform:147][ 1.939] fit_transform: Remove constant columns. Dropping 0 columns"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777964868,"msg":"[8145:140643603412800][fit_transform:227][ 173.397] fit_transform: After feature engineering and transformations. Updated shape : (157776, 72976)"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777964868,"msg":"[8145:140643603412800][fit_transform:231][ 173.397] fit_transform: col_types_ : ['text', 'text']"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777964879,"msg":"[8145:140643603412800][fit_transform:247][ 173.397] fit_transform: class counts: 0 157494\n1 282\n\nName: label, dtype: int64"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777975619,"msg":"[8145:140641670100736][run:272][ 249.292] (249.29190063476562, 249.29190063476562) (current,max) GB for process 8145 is NOT within the threshold 248.94811630249023 GB memory usage"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777975619,"msg":"[8145:140641670100736][run:279][ 249.292] Sending SIGTERM to the main thread"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777975725,"msg":"[8145:140641670100736][run:272][ 250.038] (250.03799438476562, 250.03799438476562) (current,max) GB for process 8145 is NOT within the threshold 248.19862747192383 GB memory usage"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777975725,"msg":"[8145:140641670100736][run:279][ 250.038] Sending SIGTERM to the main thread"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777975830,"msg":"[8145:140641670100736][run:272][ 250.782] (250.78213500976562, 250.78213500976562) (current,max) GB for process 8145 is NOT within the threshold 247.45304489135742 GB memory usage"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777975830,"msg":"[8145:140641670100736][run:279][ 250.782] Sending SIGTERM to the main thread"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777975936,"msg":"[8145:140641670100736][run:272][ 251.528] (251.52822875976562, 251.52822875976562) (current,max) GB for process 8145 is NOT within the threshold 246.70550918579102 GB memory usage"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777975936,"msg":"[8145:140641670100736][run:279][ 251.528] Sending SIGTERM to the main thread"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777976042,"msg":"[8145:140641670100736][run:272][ 252.274] (252.27432250976562, 252.27432250976562) (current,max) GB for process 8145 is NOT within the threshold 245.95797729492188 GB memory usage"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777976042,"msg":"[8145:140641670100736][run:279][ 252.274] Sending SIGTERM to the main thread"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777976148,"msg":"[8145:140641670100736][run:272][ 253.02] (253.02041625976562, 253.02041625976562) (current,max) GB for process 8145 is NOT within the threshold 245.21044158935547 GB memory usage"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777976148,"msg":"[8145:140641670100736][run:279][ 253.02] Sending SIGTERM to the main thread"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777976253,"msg":"[8145:140641670100736][run:272][ 253.767] (253.76651000976562, 253.76651000976562) (current,max) GB for process 8145 is NOT within the threshold 244.46290588378906 GB memory usage"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777976253,"msg":"[8145:140641670100736][run:279][ 253.767] Sending SIGTERM to the main thread"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777976359,"msg":"[8145:140641670100736][run:272][ 254.513] (254.51260375976562, 254.51260375976562) (current,max) GB for process 8145 is NOT within the threshold 243.71537017822266 GB memory usage"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777976359,"msg":"[8145:140641670100736][run:279][ 254.513] Sending SIGTERM to the main thread"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777976465,"msg":"[8145:140641670100736][run:272][ 255.261] (255.26065063476562, 255.26065063476562) (current,max) GB for process 8145 is NOT within the threshold 242.96588134765625 GB memory usage"}
{"analyticsClusterId":"","component":"ml_driver","level":"INFO","ts":1688777976465,"msg":"[8145:140641670100736][run:279][ 255.261] Sending SIGTERM to the main thread"}

```

Ask the question (“What is the main problem in the following collection of logs. Provide a two-sentence summary.”)

“The main problem in this collection of logs is that the memory usage of a particular process (with ID 8145) is consistently exceeding the defined memory threshold. This issue is repeatedly highlighted in the logs, and SIGTERM signals are sent to the main thread to terminate the process when it exceeds the memory threshold. Would you like me to help you with anything else regarding this collection of logs?”

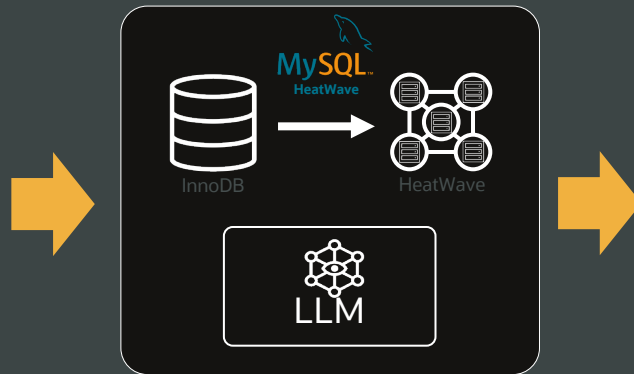


Usecase 2: Financial fraud detection

CONTENT GENERATION / SUMMARIZATION

DAT_TXN		3/31/12
COD_CC_BRN_TXN		160
COD_USERNO		0
CTR_BATCH_NO		0
REF_SYS_TR_AUD_NO		0
REF_SUB_SEQ_NO		
COD_ACCT_NO		1601000010862
TXN_TXN_DESC	Balance brought forward	
DAT_VALUE		3/31/12
REF_CHQ_NO		
COD_DRCR	C	
COD_TXN_MNEMONIC		5010
COD_TXN_LITERAL	PBC	
AMT_TXN		0
FLG_PASBKUPD	N	
RAT_CCY		1
COD_MSG_TYP		200
COD_PROC		251000
CTR_UPDAT_SRLNO		
DAT_POST		3/31/12
AMT_TXN_TCY		0
COD_TXN_CCY		
RAT_CONV_TCLCY		1
COD_SC		
COD_AUTH_ID		
REF_CARD_NO		
REF_TXN_NO		
REF_USR_NO		
TXN_ADDL_INFO		1
DAT_TS_TXN_INIT		2015/02/18 15:38:58
DAT_TS_SENT_FOR_AUTH		2015/02/18 15:38:58
DAT_TS_AUTH		2015/02/18 15:38:58
DAT_TS_SUB_FOR_PROCESSING		2015/02/18 15:38:58
DAT_TS_SENT_TO_HOST		2015/02/18 15:38:58
DAT_TS_HOST_PROCESSING		2015/02/18 15:38:58
COD_ENTITY_VPD		240

Transaction to Natural Language Description



The provided bank transaction information indicates that on March 31, 2012, a balance was brought forward from the previous accounting period to the current one. The account number associated with this transaction is 1601000010862, and the amount of the transaction is zero. The flag indicating whether the transaction has been successfully processed is set to N, which means that the transaction was not successful. The currency code for the transaction is 1, and the message type is 200, which indicates an error message. The processing code is 251000, which may indicate a specific error or issue with the transaction. The transaction has been sent to the host for processing, but it has not yet been processed. The entity VPD associated with this transaction is 240.

There are several suspicious aspects of this transaction that require further investigation. Firstly, the fact that the amount of the transaction is zero raises questions about the purpose and validity of the transaction. Secondly, the flag indicating that the transaction was not successful suggests that there may have been an issue with the processing or authorization of the transaction.

```
SET @context = (SELECT txn_description FROM open_transaction WHERE txn_id = mule_txn_id);
SET @query = "Provided is a bank transaction information with field names and field values. Each field is separated by '\n\n'. Please provide a concise summary of this bank transaction. What is suspicious about this transaction? Your audience is a bank officer.";
SELECT sys.ML_GENERATE(@query, JSON_OBJECT("context", @context));
```



Usecase 3 : Personalization of Online Delivery

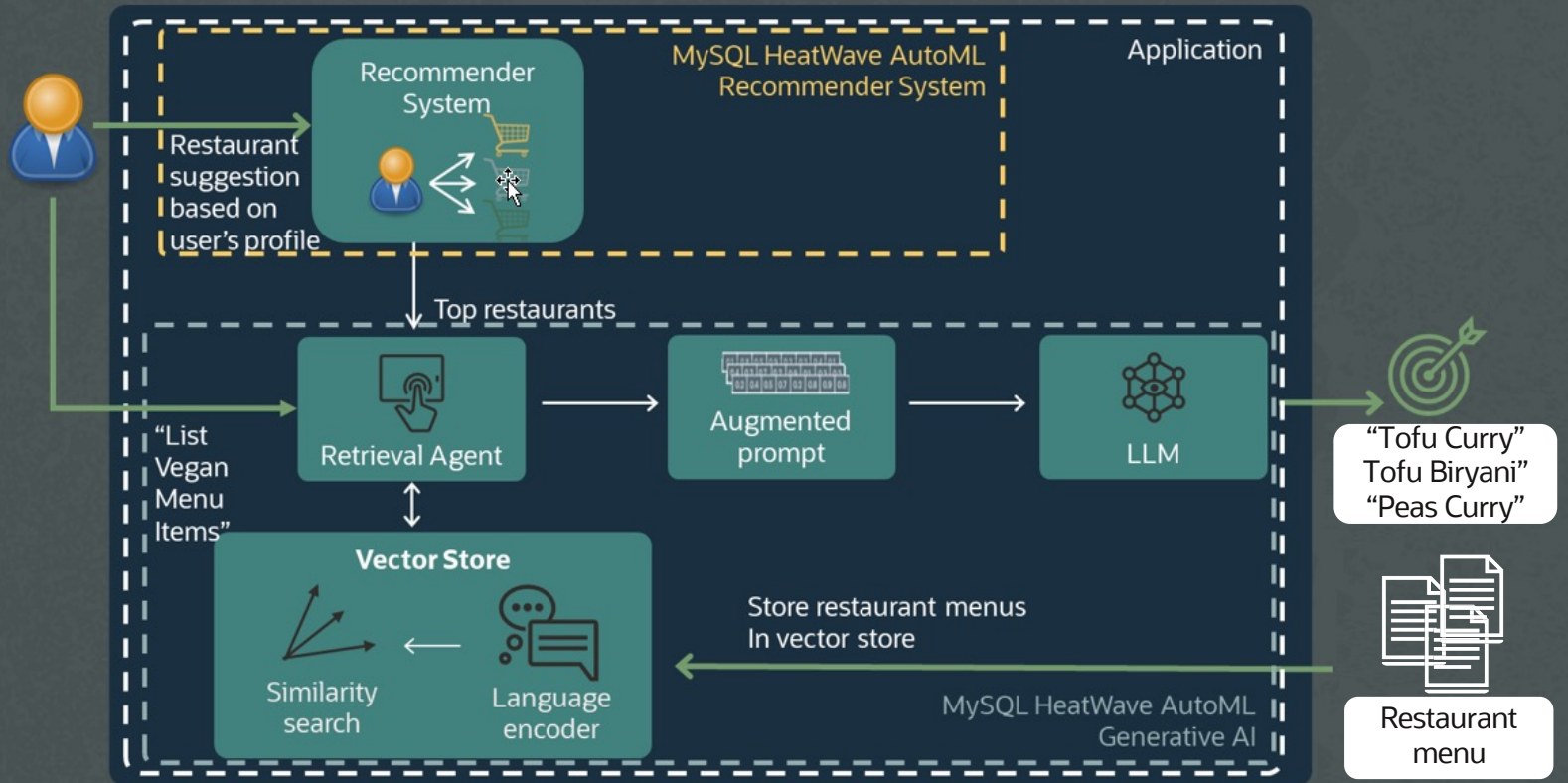
RETRIEVAL AUGMENTED GENERATION



Online food delivery

Recommend, Retrieve and Generate description of dishes based on user preference

Personalized Menu: HW AutoML + Retrieval Augmented Generation



Usecase 3: Personalization of Online Delivery

RETRIEVAL AUGMENTED GENERATION

Home > Dubai > Al Barsha 1 Restaurants > China Chowk

Sweet Spicy Chili Beef
Mains
"Fiery Beef Fusion: Succulent beef stir-fried with a tantalizing blend of sweet chili sauce, aromatic garlic, and punch of dry chili, creating a harmonious balance of sweetness and heat"
48 AED

Tofu Stir Fry
Mains
Indulge in our delightful **Sweet Chili Tofu Stir-Fry!** This vibrant dish features tender tofu cubes stir-fried with green beans, crunchy bean sprouts, and julienne-cut carrots, all perfectly coated in a tantalizing sweet chili sauce. The harmony of textures and flavors creates a savory-sweet sensation that's both satisfying and wholesome.
44 AED

Sweet And Sour Prawns
Mains
"Golden tempura-coated shrimp tossed with a trio of colorful capsicums, sweet pineapple, and white onion—a fusion of crispy, savory, and sweet flavors in every bite."
45 AED

Home > Dubai > Al Barsha 1 Restaurants > Geppetto Pizzeria

Margherita Vegan Pizza
Pizza Vegan
Italian Tomato, Vegan Mozzarella, Basil And Oregano
49 AED
Customize

Pepperoni Vegan Pizza
Pizza Vegan
Italian Tomato, Vegan Mozzarella, Vegan Sausage, Basil And Oregano
96 AED
Customize

Quattro Formaggi Vegann Pizza
Pizza Vegan
Italian Tomato, Vegan Mozzarella, Vegan Parmesan, Vegan Cheddar, Vegan Gouda And Basil
75 AED
Customize

Verde Vegan Pizza
Pizza Vegan
Italian Tomato, Fresh Vegan Pesto, Vegan Mozzarella, Red Onion, Pepper, Zucchini, Eggplant... more
65 AED
Customize

Funghi Vegan Pizza
Pizza Vegan
Italian Tomato, Vegan Mozzarella, Mushrooms, Basil And Oregano
53 AED
Customize

Tartufo Vegan Pizza
Pizza Vegan
Italian Vegan Truffle Cream, Vegan Cream Cheese, Mushrooms And Fresh Italian Summer Truffl... more
99 AED
Customize



```
ask_question('Recommend three vegan dishes with tofu', color='b')
```

1. Sweet Chili Tofu Stir-Fry - a savory and sweet dish featuring tender tofu cubes stir-fried with vegetables and coated in a delicious sweet chili sauce. 2. CustomizeTofu Stir Fry - a customizable dish where you can choose various vegetables and sauces to accompany your tofu stir-fry. 3. Vegan Pizza - Italian Tomato, Vegan Mozzarella, Tofu, and Basil - a delicious and satisfying pizza option by customizing with tofu as your protein choice.

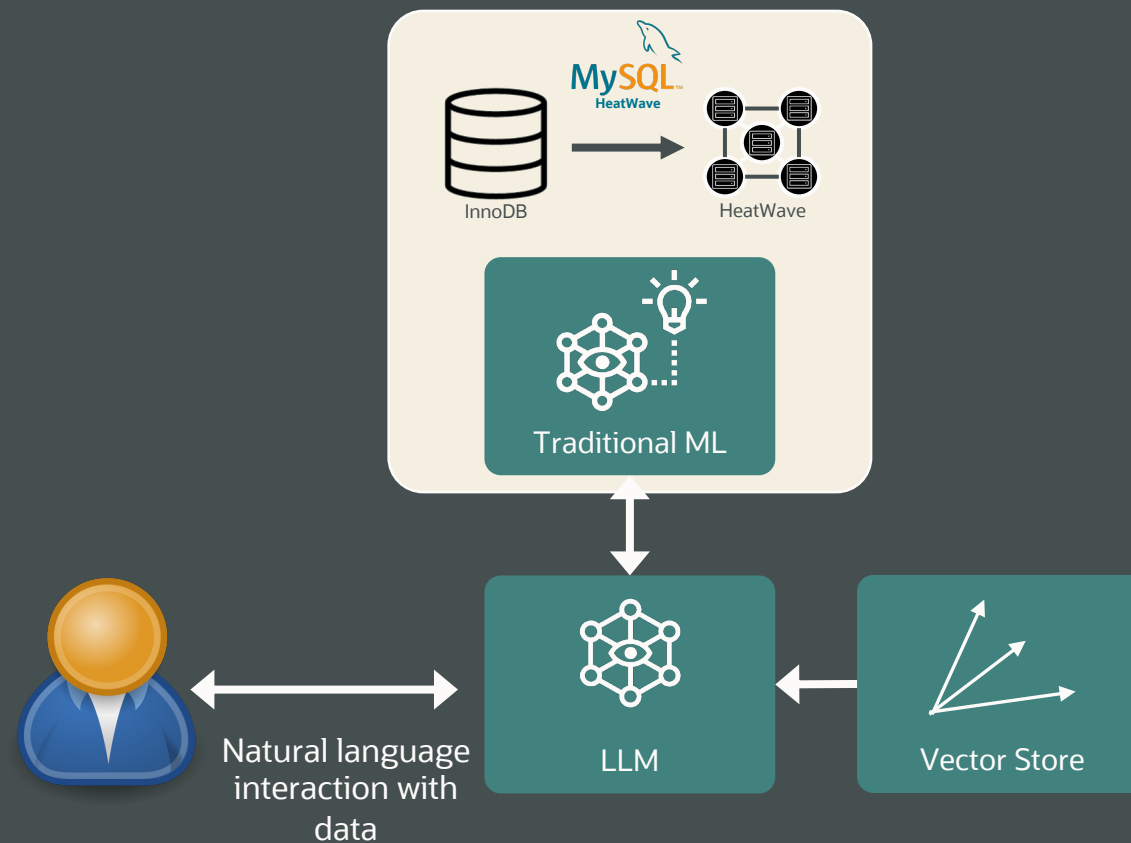
Would you like me to assist you with more vegan dish recommendations that incorporate tofu?



Synergy between HeatWave AutoML and Generative AI

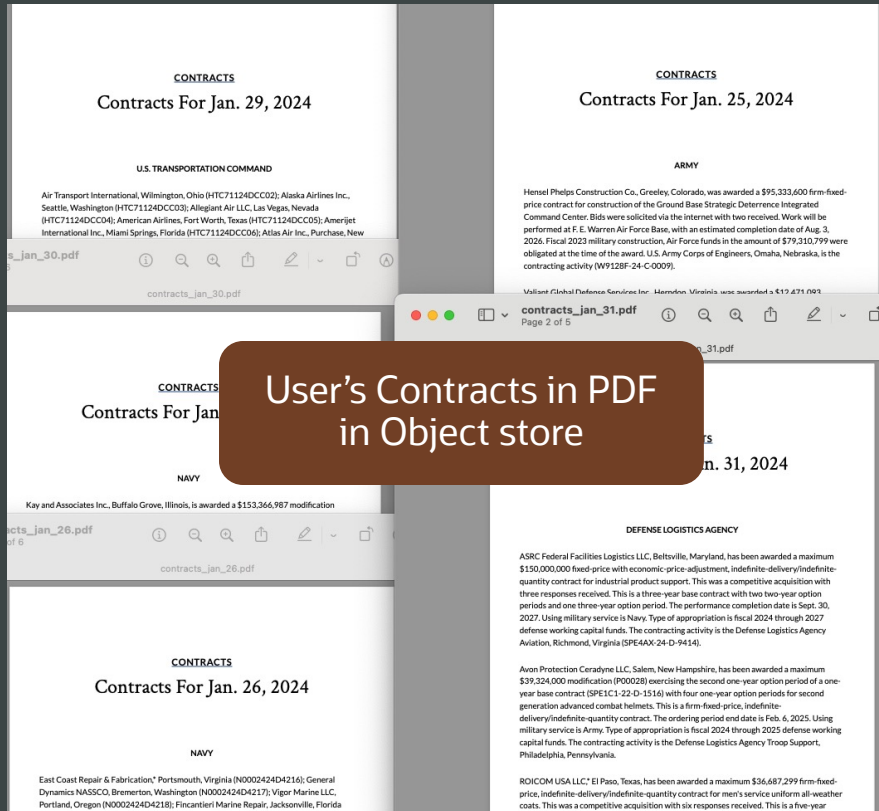
A DIFFERENTIATOR

- More accurate LLM results by filtering irrelevant data
- Faster LLM inference due to smaller search space



Usecase 4 : Natural language interaction

EMPLOYEE ASSISTANT



Employee Assistant: Improve employee productivity

```
ask_question(query='Which firm was awarded a contract for Dell laptops?', color='blue')
```

✓ 15.7s Python

MCP Computer Products was awarded a contract for Dell laptops. Would you like to know more about this company?

```
ask_question(query='Were there any contracts for the Rolling Airframe Missile?', color='blue')
```

✓ 10.3s Python

Yes, Raytheon Co. was awarded a \$13.4 million cost-plus-fixed-fee modification contract for design and engineering support services for the Rolling Airframe Missile. This was awarded on January 25, 2024, and work is expected to be completed by March 2028. Would you like to know more about this contract?

```
ask_question(query='Were there any contracts mentioning Lassa Fever?', color='blue')
```

✓ 14.2s Python

Yes, Primmune Therapeutics Inc. based in San Diego, California, was awarded a contract worth \$22,480,552 for the development of an oral broad-spectrum antiviral TLR7 agonist for the treatment of Lassa Fever. The contract is identified as HDTRA1-24-C-0014 and was awarded by the Defense Threat Reduction Agency (DTRA). Would you like to know more about this contract?

Retrieval Augmented Generation

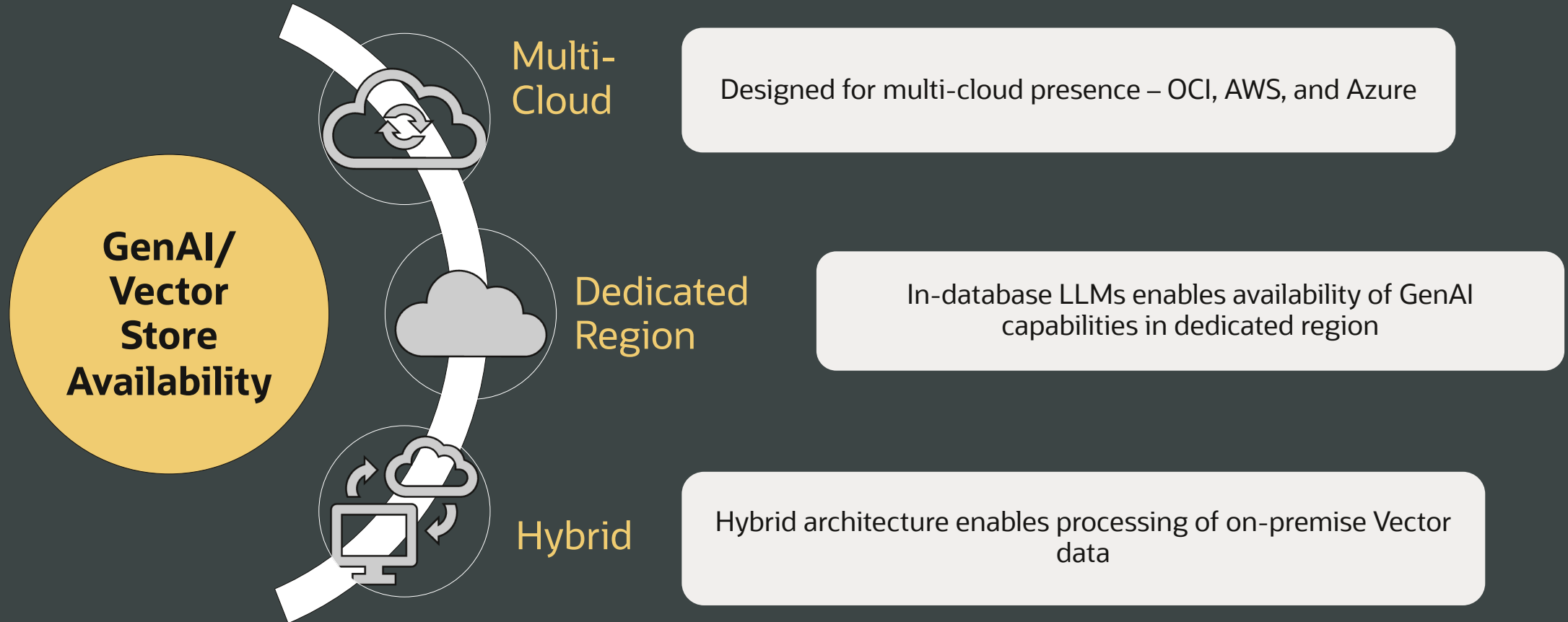


HeatWave offers choice for running LLMs

- **In-HeatWave LLMs**
 - Native execution within the HeatWave database
 - Run smaller LLMs like Llama2-7B and Mistral-7B
 - Secure, lower cost, guaranteed availability
- **OCI Generative AI service LLMs**
 - Support larger models like Cohere-command and Llama2-70B and run on GPUs
 - Higher quality, better performance

Available in Public Cloud, Dedicated Region, Hybrid Environment

VECTOR STORE AND ABILITY TO RUN LLMS INSIDE HEATWAVE PROVIDES FLEXIBILITY OF DEPLOYMENT



Vector Support

Vector Datatype

- Introducing new Vector data type
- In-memory hybrid-columnar storage format for vector columns

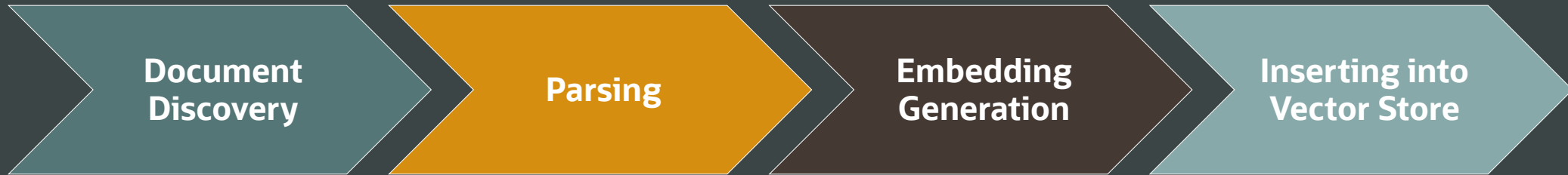
Vector Processing

- Parallelize vector processing across the HeatWave nodes
- Processes at near memory bandwidth

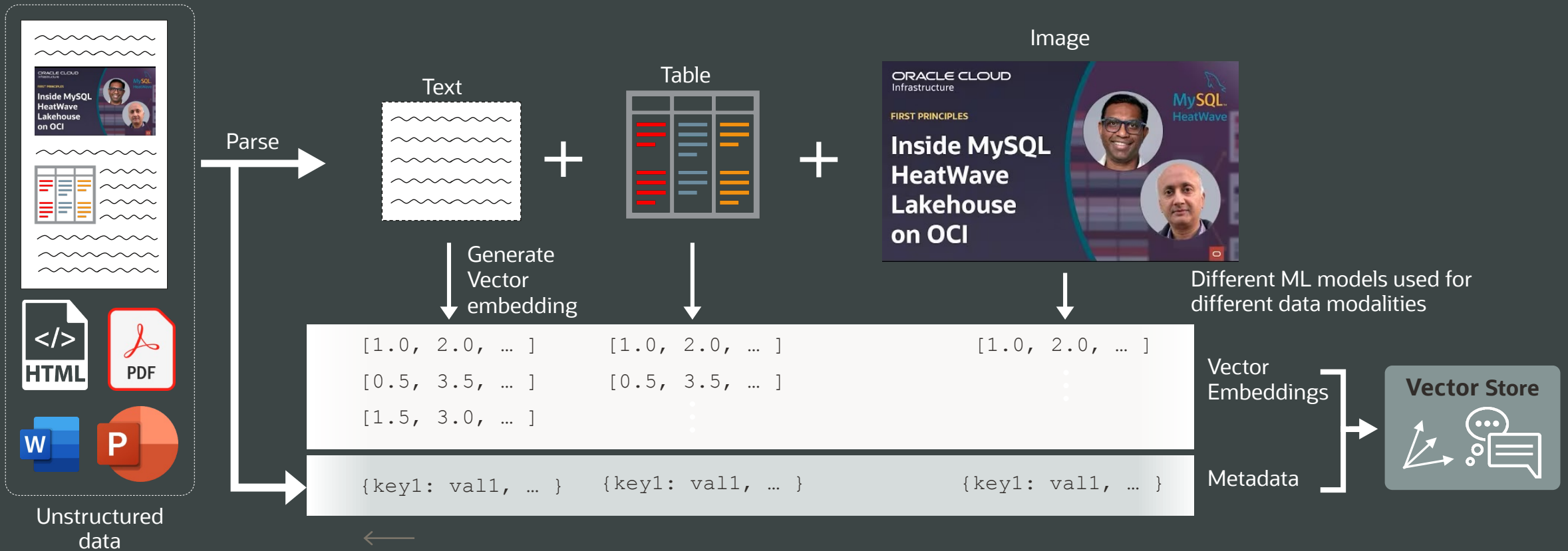
Vector Store

- In-database parsing, parallel embedding generation
- Vector stored in the object store

All phases of creating a vector store done inside HeatWave



HeatWave parses and automatically creates embeddings for documents

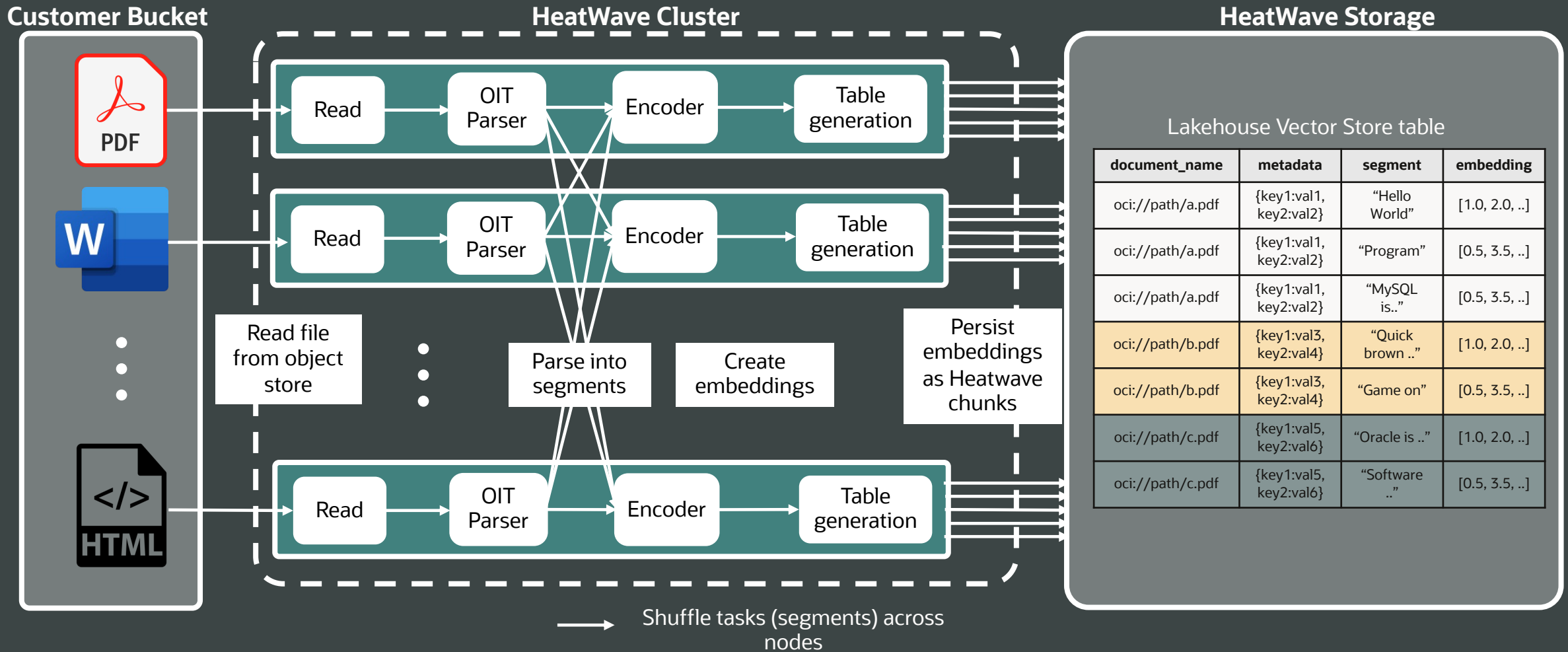


Automatically generate embedding for text from multiple file formats



Vector Store creation with HeatWave is parallelized

PARSE SOURCE FILES AND CONCURRENT EMBEDDING GENERATION ACROSS NODES

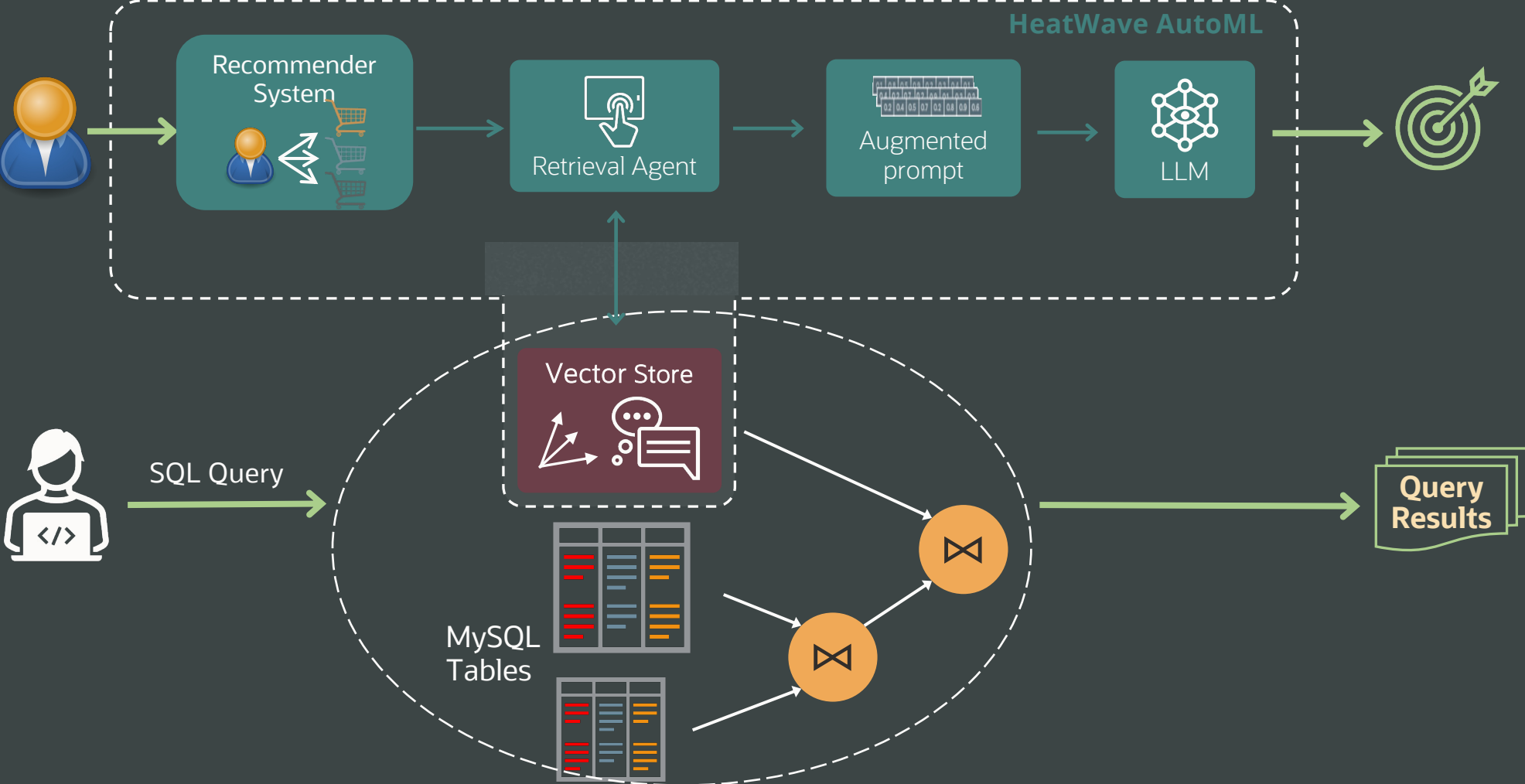


Vector Store creation in HeatWave scales out

	File	Size	Pages	Parsing Time	HW threads or nodes	Encoding Time (sec)	Vector Store creation time (s)	Speedup
Intra-document	Single PDF	44 MB	6963	16 sec	1 thread	8467	8588	1
					25 threads	738	840	10.2
					50 threads	370	481	17.8
Inter-document	Multiple PDFs	4 x 44 MB	27852	17 sec	1 node	2733	2882	1
					2 nodes	1471	1691	1.7
					4 nodes	807	952	3.0



Vector Store can be queried by natural language or SQL



Example of using HeatWave vector store

Create Vector Store

```
# Ingest documents from Object Store like any Lakehouse table  
CALL sys.heatwave_load("vector_store", @load_params);
```

Query Vector Store Native SQL syntax

```
# Example: Find books semantically most similar to input and are in print  
SELECT id, title  
FROM books b, books_in_print ip  
WHERE b.title = ip.title  
ORDER BY DISTANCE(b.segment_embedding, @query_embedding, "DOT") as distance  
DESC LIMIT 10;
```

Query Vector Store ML_RAG

```
# Example: Answer questions using data in documents ingested into Vector Store  
CALL sys.ML_RAG("Which state has maximum carbon?", @output);
```

Vector data type support

- Standard SQL interface to create tables with vector columns
- Vector data storage
 - HeatWave: In-memory columnar format
 - InnoDB: BLOB

```
mysql> CREATE TABLE wikipedia (  
    id INT,  
    title VARCHAR(1024)  
    page_data TEXT  
    page_list TEXT,  
    page_url TEXT,  
    page_embedding VECTOR(1024)  
ENGINE_ATTRIBUTE='{ "model": "cohere" }'  
    ENGINE=lakehouse,  
SECONDARY_ENGINE=rapid;
```

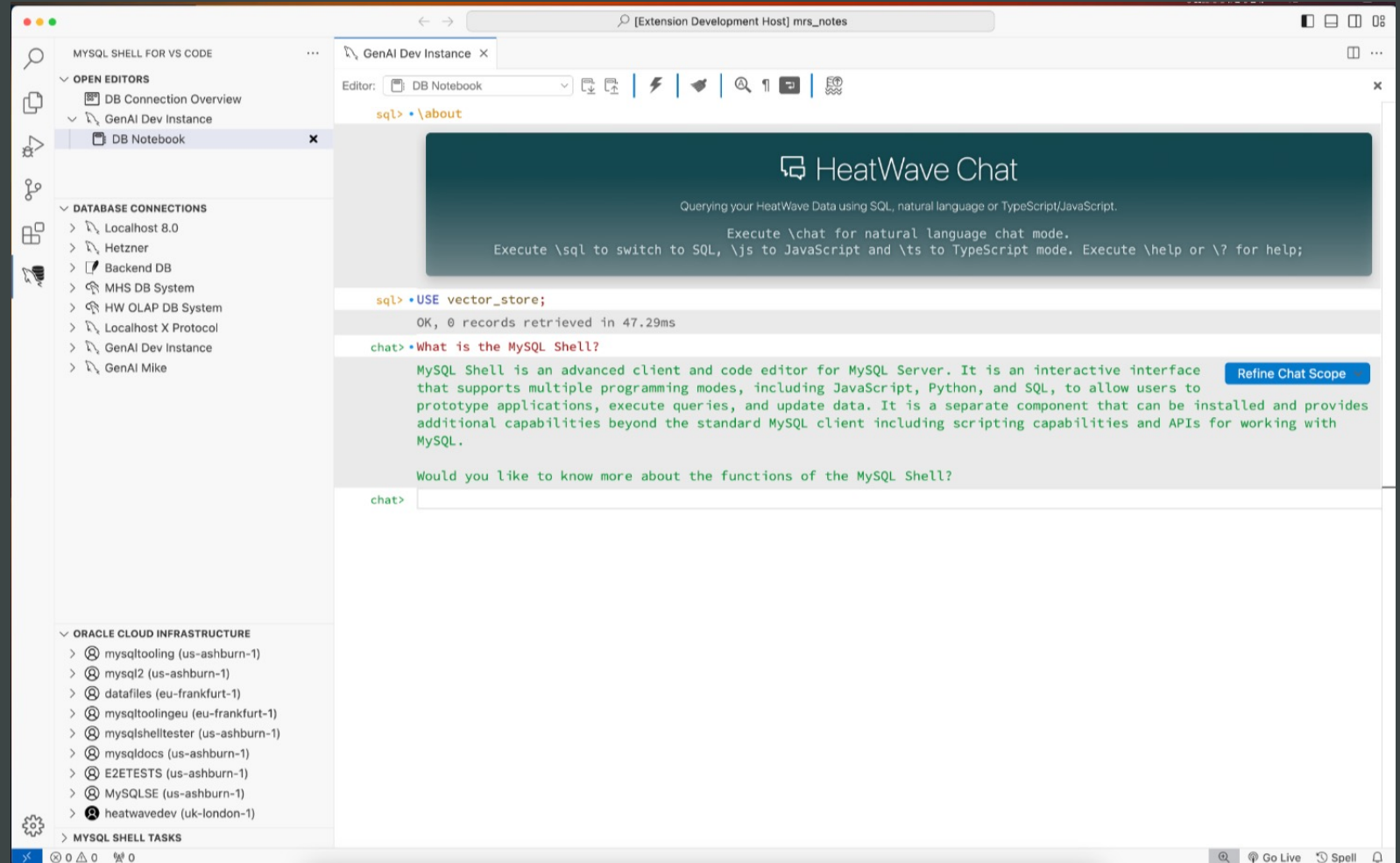
- Example distance functions

- L1/MANHATAN
- L2/EUCLIDIAN
- L1^2/MANHATAN_SQUARED
- L2^2/EUCLIDIAN_SQUARED
- COSINE
- DOT
- HAMMING

HeatWave Chat

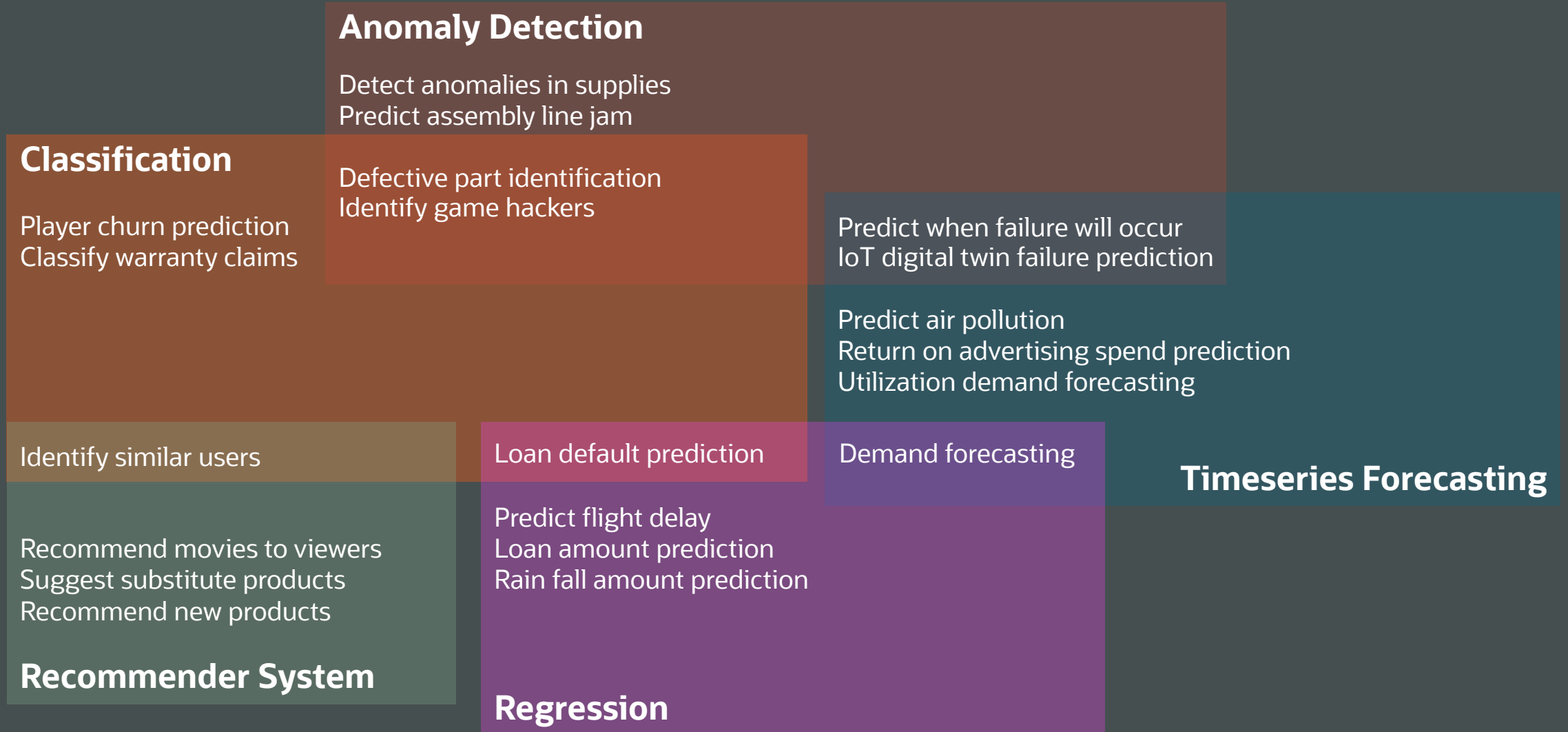
NATURAL LANGUAGE INTERACTION

- For querying unstructured text documents in Lakehouse
- Allows for refinement of chat scope (querying documents in a specific folder, using different ML models)

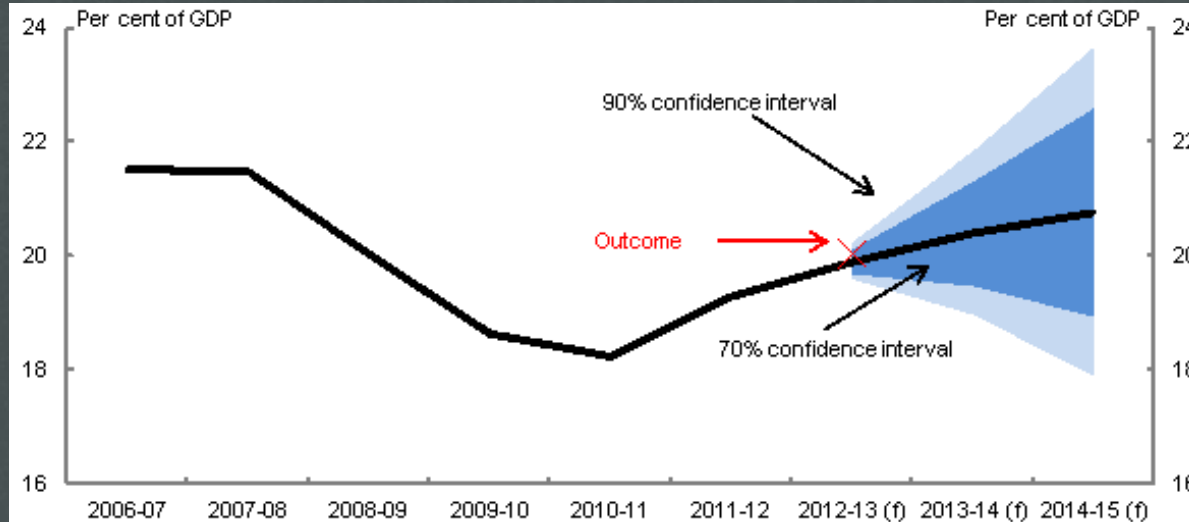


HeatWave AutoML

IN DATABASE MACHINE LEARNING, FULLY AUTOMATED TRAINING, EXPLANATIONS, 25X FASTER THAN REDSHIFT ML



Confidence Interval for forecasting in HeatWave AutoML



- Represents a range where future values are likely to fall, based on a certain confidence level (e.g., 95%)
- Helps users assess risk, make informed decisions, and understand the uncertainty in model predictions

```
MySQL> CALL sys.ML_PREDICT_TABLE('mlcorpus.`date_test_positive`', @model4, 'mlcorpus.predictions', JSON_OBJECT('prediction_interval', 0.95));

MySQL> SELECT ml_results FROM mlcorpus.predictions limit 5;

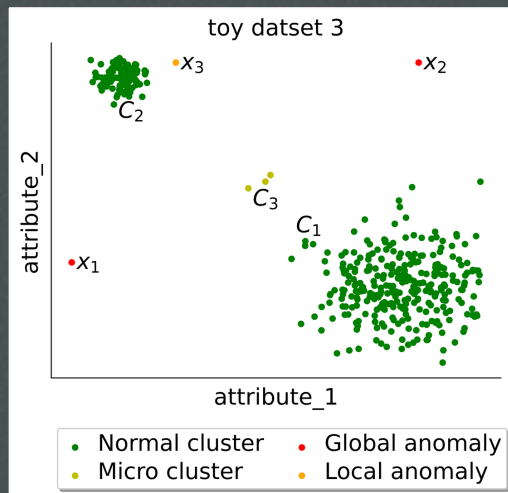
+-----+
| ml_results |
+-----+
| {"predictions": {"C1": 616250423.9112742, "C2": 456582262.8511299, "prediction_interval_C1": [-184191809.507195, 1416692657.3297439], "prediction_interval_C2": [-196645000.461776, 1109809526.164036]}} |
| {"predictions": {"C1": 1636307524.452425, "C2": 1172929787.002929, "prediction_interval_C1": [724636104.7043328, 2547978944.2005186], "prediction_interval_C2": [425149558.2772062, 1920710015.7286525]}} |
| {"predictions": {"C1": 1113801796.688651, "C2": 1296259009.615075, "prediction_interval_C1": [176442284.4147286, 2051161308.9625735], "prediction_interval_C2": [521088058.2289947, 2071429961.0011563]}} |
| {"predictions": {"C1": 997977894.4861443, "C2": 1445240945.087429, "prediction_interval_C1": [6166579.356579423, 1989789209.6157093], "prediction_interval_C2": [602453234.1985078, 2288028655.976351]}} |
| {"predictions": {"C1": 973596300.6006613, "C2": 1112061057.088149, "prediction_interval_C1": [-84395450.6604342, 2031588051.8617566], "prediction_interval_C2": [247888576.1548234, 1976233538.0214758]}} |
+-----+
```

New models for anomaly detection : GLOF, PCA

AUGMENTS THE EXISTING GKNN ALGORITHM

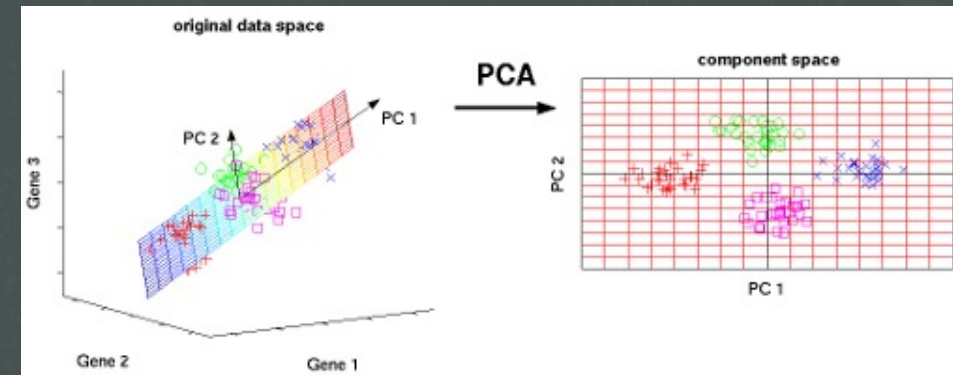
Generalized Local Outlier Factor (GLOF) :

An internally developed model, it detects global and clustered anomalies like GkNN, while also detecting local anomalies.



Principal Component Analysis (PCA) :

The model reduces complex data into its main components to more easily identify outliers

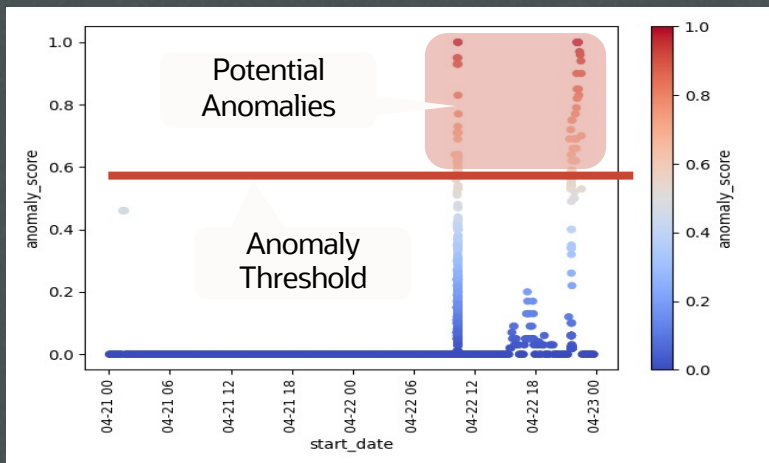


Anomaly detection for logs

NOW TRAINED FOR MYSQL LOGS

Heatwave processes and generalizes incoming machine logs, then builds a tailored anomaly detection model

This model helps in identifying anomalies in logs, enabling effective preventative maintenance and root cause analysis



- [2024-03-05 13:28:59 - 2024-03-05 13:29:20] Group replication-related failure
[GCS] Failure reading from fd=<:NUM:> <:*:> from <:IP:>:<:NUM:>
- [2024-03-05 13:29:25 - 2024-03-05 13:29:27]
'This server is not able to reach a majority of members in the group. This server will now block all updates. The server will remain blocked until contact with the majority is restored. It is possible to use group replication force members to force a new group membership.'
- [2024-03-05 13:40:59 - 2024-03-05 13:41:21] Potential connection leak in group replication
[GCS] Old incarnation found while trying to add node
- [2024-03-05 17:26:31]
Database was not shutdown normally! Starting crash recovery. Starting to parse redo log at lsn = <:NUM:>

Summary

- In-database vector store for querying unstructured text content
- In-database Generative AI in HeatWave brings power of LLMs to enterprise content
- Vector processing in HeatWave can be combined with other SQL operators
- Continued innovation in HeatWave AutoML
- Best performance, price performance, scalability in the industry for querying data
- Single service for machine learning, GenAI, analytics, Lakehouse and OLTP

ORACLE