



LLMs in Customer Support

AI Data Cloud; Welcome to the Era of Enterprise AI

Agenda

- > Introduction
- > Initial Response (LLMs)
- > Incident Detection (LLMs, Embeddings, Clustering)
- > University Recruiting Program
- > Current Opportunities

Who We Are



Kamil Anikiej
Eng Manager,
Snowflake



Marcin Wawerka
Senior Software
Engineer, Snowflake



Paweł Karwacki
Senior Software
Engineer, Snowflake



Kasia Dziewonska-Jucha
Students/Grads Programs
Manager, Snowflake

History of Snowflake



**FOUNDED IN 2012
BY INDUSTRY
VETERANS**



**FIRST CUSTOMERS
IN 2014, GENERAL
AVAILABILITY
IN 2015**



**EMPLOYEES
& OFFICES
ON 5 CONTINENTS**



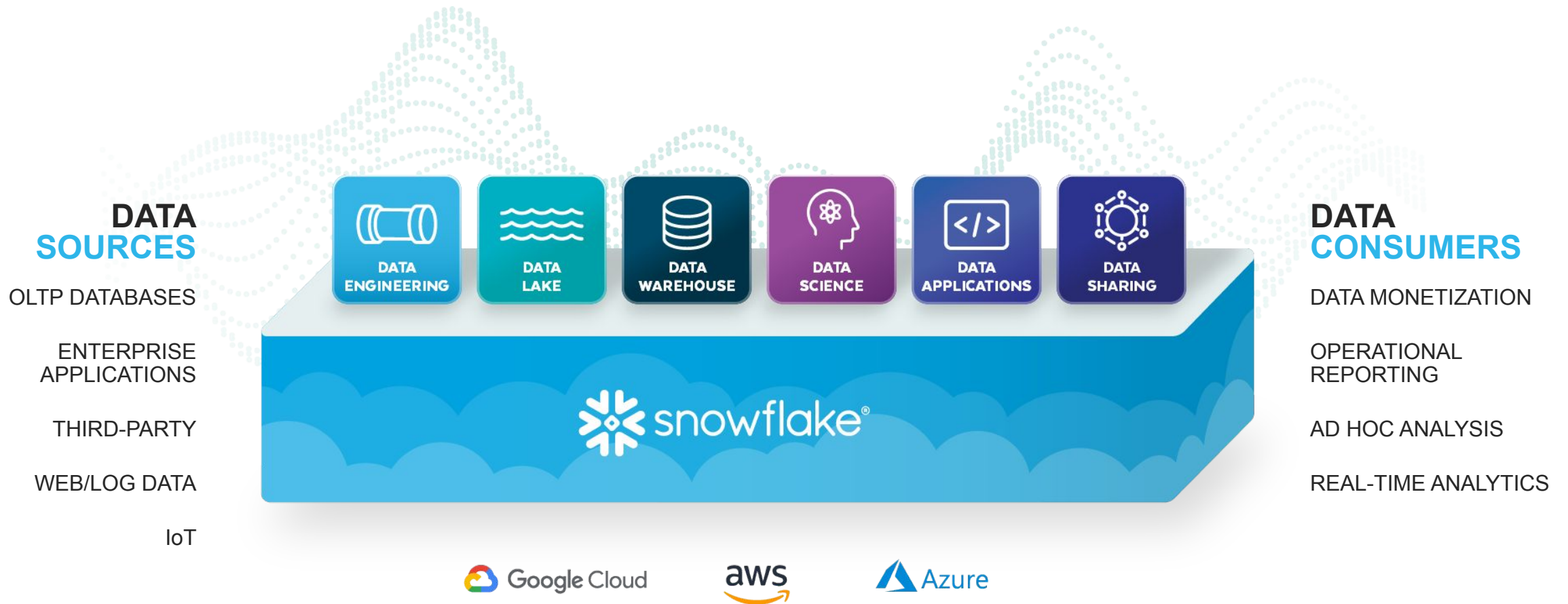
**SNOWFLAKE GOES PUBLIC
IN 2020 WITH THE BIGGEST
SOFTWARE IPO EVER**



**OVER 5900
ACTIVE CUSTOMERS**



Snowflake's Platform



INITIAL RESPONSE ASSISTANCE




What is our goal


Help customer support engineers craft a meaningful initial response to customer's case

What is a case

Subject

See Timeout issue while running External
Access APIs 

Description

We are trying to utilize Snowflake's new
feature to run external API's. However I
ended up with below timeout error. Error:
"100357 (P0000): Handler execution
timed out in 21m36.461195168s vs. max
of 10m in function
OKTA_DEVICES_DATA_INGESTION with
handler fetch_okta_devices" So, I'm not
sure if we have any timeout limits in
place, if yes, can we change these
predefined timeout limits ? 

Meaningful initial response

- Introduction
- Issue clarification/summary
- Acknowledging/requesting additional information
- Plan of action

How we did it

Case summary

- Summarize the case through an LLM
- Limited usefulness
- Introduces more noise for vague problem descriptions

How we do it

Initial response RAG (Retrieval Augmented Generation)

- Retrieve context based on historical cases and official documentation/knowledge base articles
 - Snowflake's [Cortex Search](#)
- Generate initial response using llama3.1-405b
 - Snowflake's [Cortex Complete](#)
 - Never to be used without human in the loop

Hello Philippe,

Thank you for contacting Snowflake Support. My name is Ellappan Natarajan, and I will be assisting you with this case.

Regarding the Snowflake Notification about the OCSP and TLS impacts from the Snowflake load balancer (Envoy) migration, I understand that you are looking for guidance on the impact of switching to providing TLS termination via Digicert certificates for your internal reporting product.

Based on the information provided, I can confirm that:

1. For your AWS Lambda functions using Python runtime to connect and ingest data into Snowflake databases, you will not be impacted by the switch to Digicert certificates. Snowflake's Python driver includes updated truststores which are not impacted by this change.
2. For reporting purposes, when using 3rd party software such as Power BI or Tableau desktop to query data in Snowflake, the impact depends on the specific software and its configuration. If the software relies on the operating system's truststore, you may need to update the truststore to include the Digicert Global Root G2 certificate authority (CA). However, if the software uses its own truststore, you may not need to take any action.

To confirm, I recommend checking the documentation for the specific software you are using to determine if any action is required.

Additionally, I would like to provide you with some general information about the Envoy migration and the impact on Snowflake customers. You can find more details in our Knowledge Base article as below,

[1] <https://docs.snowflake.com/en/release-notes/bcr-bundles/un-bundled/bcr-1657?css>

[2] <https://community.snowflake.com/s/article/FAQ-Updates-on-Migration-of-Traffic-Serving-Proxy-Load-Balancing-Infrastructure?css>

Please let me know if you have any further questions or concerns.

Thanks & Regards,

Hello Philippe,

Thank you for contacting Snowflake Support. My name is [Your Name], and I will be assisting you with this case.

Regarding the Snowflake Notification about the OCSP and TLS impacts from Snowflake load balancer (Envoy) migration, I understand that you are looking for guidance on the impact of switching to providing TLS termination via Digicert certificates for your internal reporting product.

Based on the information provided, I can confirm that:

1. For your AWS Lambda functions using Python runtime to connect and ingest data into Snowflake databases, you will not be impacted by the switch to Digicert certificates. Snowflake's Python driver includes updated truststores which are not impacted by this change.
2. For reporting purposes, when using 3rd party software such as Power BI or Tableau desktop to query data in Snowflake, the impact depends on the specific software and its configuration. If the software relies on the operating system's truststore, you may need to update the truststore to include the Digicert Global Root G2 certificate authority (CA). However, if the software uses its own truststore, you may not need to take any action.

To confirm, I recommend checking the documentation for the specific software you are using to determine if any action is required.

Additionally, I would like to provide you with some general information about the Envoy migration and the impact on Snowflake customers. You can find more details in our Knowledge Base article: FAQ: Envoy migration updates (Migration of Traffic-Serving Proxy/Load Balancing Infrastructure).

Please let me know if you have any further questions or concerns.

Best regards,

[Your Name]

Snowflake Support Engineer

[00851466]

This was awesome, I was unaware of this BCR change and the LLM suggestion pointed me right to it, saved me some time.

One thing tho, the link provided in the generated response doesn't actually include the text:

I would like to inform you that this behavior is expected due to a recent change in Snowflake's behavior, as mentioned in our documentation: "Masking policy: Comply with the scale and precision of a column" (<https://docs.snowflake.com/en/sql-reference/sql/create-masking-policy>). This change ensures that the masked value complies with the scale and precision of the column.

Instead, that text is in a BCR article: https://docs.snowflake.com/en/release-notes/bcr-bundles/2024_04/bcr-1355



LLM generated response:

```
1 Hi there,  
2  
3 Thank you for reaching out to Snowflake Support. I'm happy to help  
you with your question about creating a notification integration for  
email.  
4  
5 To create a notification integration, you'll need to navigate to the  
"Account" page and click on the "Integrations" tab. However, it  
seems like you're not seeing this option.  
6  
7 Can you please confirm if you have the necessary permissions to view  
and manage integrations in your Snowflake account? Additionally, you  
can try checking the Snowflake documentation on notification  
integrations for more information on the steps to follow:  
https://docs.snowflake.com/en/user-guide/email-stored-procedures  
8  
9 If you're still having trouble, please let me know and I'll be happy  
to guide you through the process.  
10  
11 Best regards,  
12 [Your Name]
```

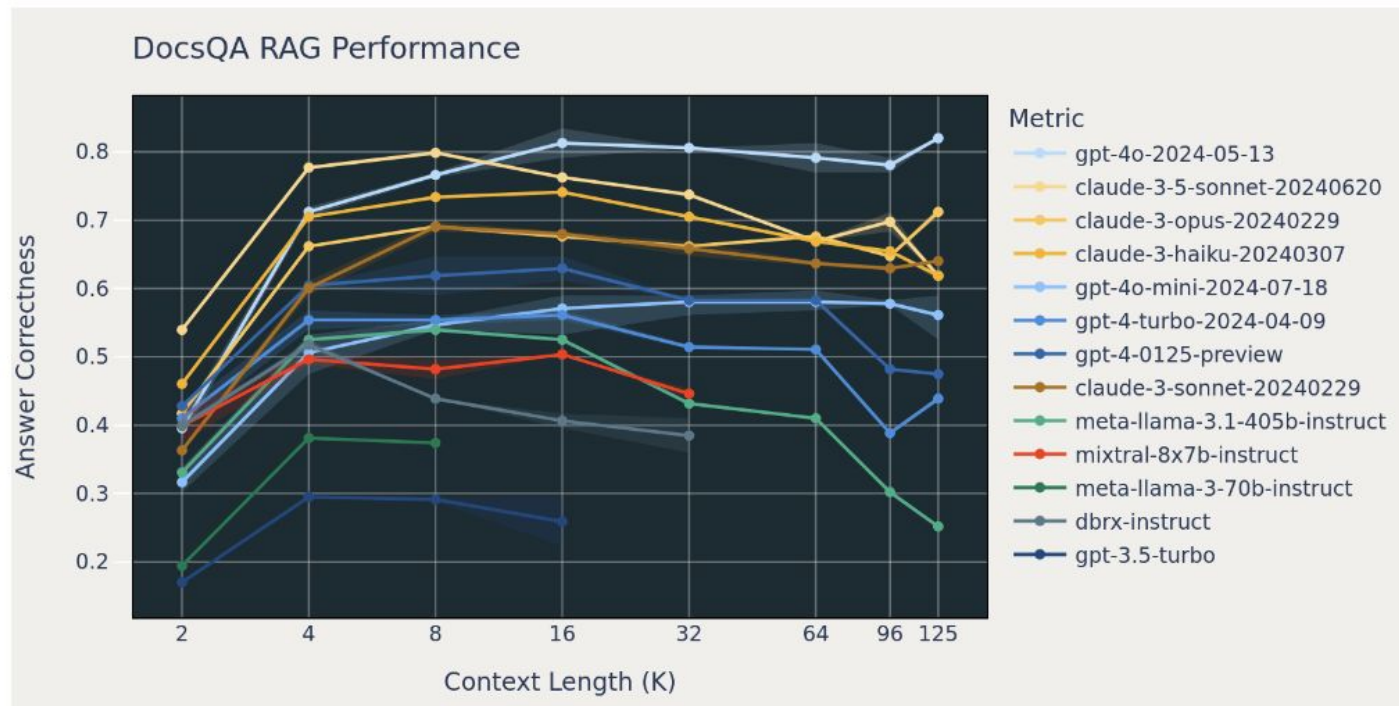
#Initialresponse

- The statement "To create a notification integration, you'll need to navigate to the "Account" page and click on the "Integrations" tab." is not correct. There is no tab like 'integrations' under 'account' in Snowsight UI.
- The customer question is about 'create notification integration for email'. So the product documentation link should be <https://docs.snowflake.com/user-guide/notifications/email-notifications> or <https://docs.snowflake.com/user-guide/notifications/email-notifications#label-create-email-notification-integration>



LLM limitations

- Hallucinations
- 128K tokens context window
- Performance degradation



Metrics

RAG metrics

metrics

- Answer relevance
 - Is the answer relevant to a question?
- Context relevance
 - Is the context provided relevant to the question?
- Groundedness
 - Is the answer based on context?

Similarity

- ROUGE metrics
 - Based on n-gram repetitions
- Embedding vector distance
 - Semantic similarity
- Edit distance
 - Levenshtein/Jaro-Winkler

Calculating RAG metrics

- LLM judges LLM responses
- Trulens framework

```
"""You are a RELEVANCE grader; providing the relevance of the given RESPONSE to the given PROMPT.  
Respond only as a number from 0 to 10 where 0 is the least relevant and 10 is the most relevant.  
  
A few additional scoring guidelines:  
  
- Long RESPONSES should score equally well as short RESPONSES.  
  
- RESPONSE must be relevant to the entire PROMPT to get a score of 10.  
  
- RELEVANCE score should increase as the RESPONSE provides RELEVANT context to more parts of the PROMPT.  
  
- RESPONSE that is RELEVANT to none of the PROMPT should get a score of 0.  
  
- RESPONSE that is RELEVANT to some of the PROMPT should get a score of 2, 3, or 4. Higher score indicates more RELEVANCE.  
  
- RESPONSE that is RELEVANT to most of the PROMPT should get a score between a 5, 6, 7 or 8. Higher score indicates more RELEVANCE.  
  
- RESPONSE that is RELEVANT to the entire PROMPT should get a score of 9 or 10.
```

QUESTIONS?



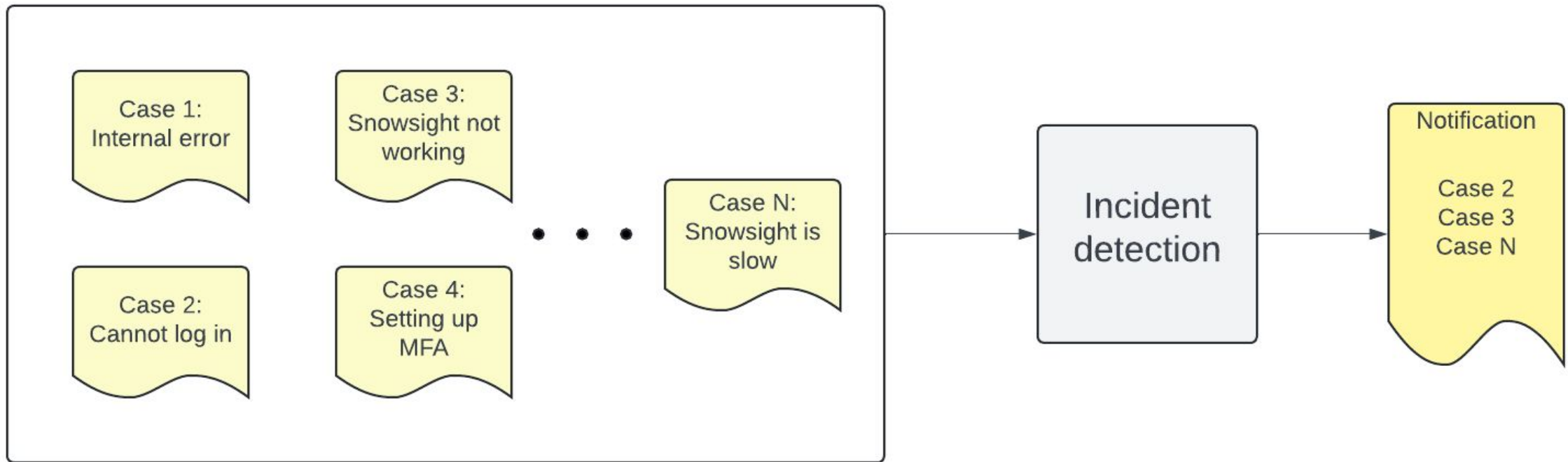
INCIDENT DETECTION



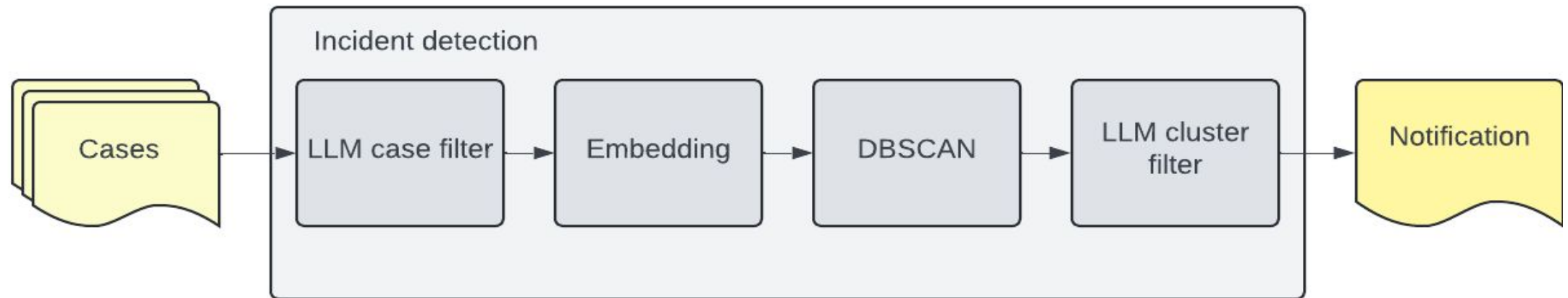
Incident detection - our goal

- General goal: we want to reduce the time it takes us to detect major incidents in Snowflake product.
- Our focus: using support case data for detecting problems affecting multiple customers, potentially indicating a major incident.

General idea



Algorithm



Algorithm - LLM case filter

- We ask an LLM if a given case describes a problem with Snowflake product or is a different type of request, e.g. request to explain how something works.
- Filters out ~60% of all cases.

```
System
Given a Snowflake support case, you need to judge whether its main goal is to report that the Snowflake system is not working
the right way and the problem is affecting the customer operations in a significant way. Answer starting with YES or NO,
providing a very concise, single sentence explanation.

User
Read only access Hi,

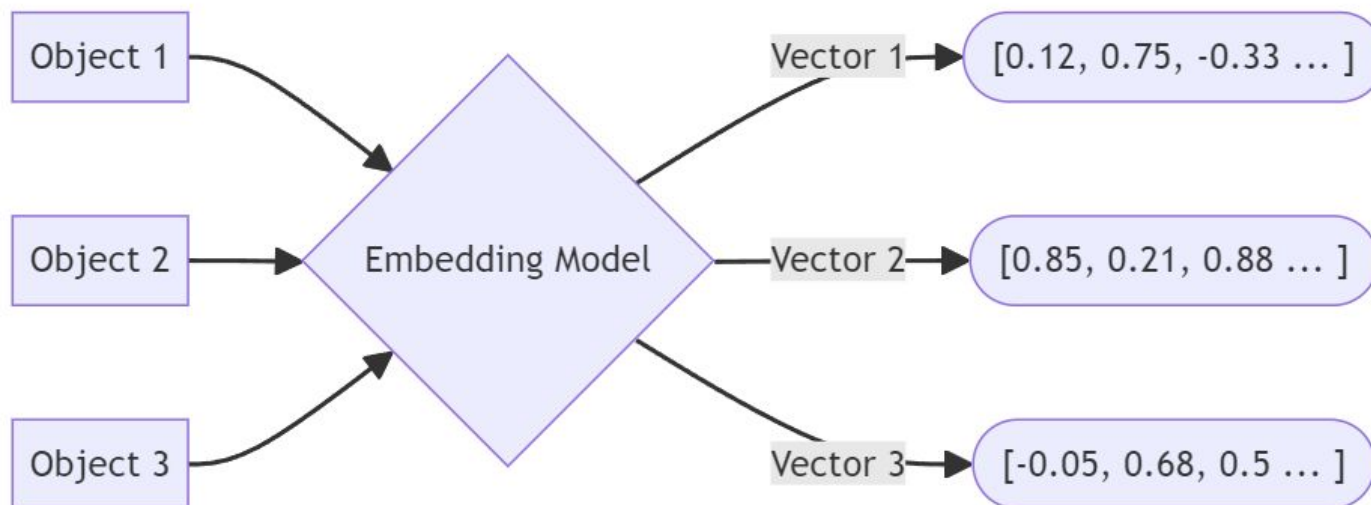
I wanted to know how would I go about granting a user just Read Access to our system?...

Show all ▾

mistral-large
No, the main goal of this Snowflake support case is to request instructions on how to grant a user read-only access to the
system, not to report a significant issue affecting customer operations.
```

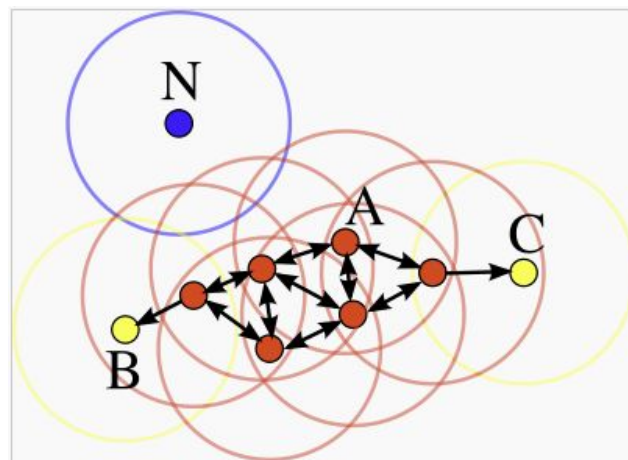
Algorithm - embedding model

- Converts support cases into high-dimensional vector space.
- Vectors are located on a unit hypersphere.
- For well-behaving embedding model we can expect similar support cases to be closer to each other with respect to their cosine similarity.



Algorithm - DBSCAN

- Commonly used density-based clustering algorithm.
- Does not require specifying the number of clusters.
- Very good for disregarding noise.
- Easily tunable:
 - eps - max distance between cases to be considered similar.
 - min samples - minimal number of cases that can form a cluster.



Algorithm - LLM cluster filter

- We ask an LLM if a given case cluster seems to be related to a Snowflake major incident.
- Filters out ~70% of all clusters.

System

You are a Snowflake support engineer. You are given several support cases that are similar to each other. You need to judge if a given set of cases might indicate that there is a major incident happening on the Snowflake side. Answer with just YES or NO.

User

<case>

Intermittent Issues Logging into Snowflake While logging into our account URLs users have received a "Connection is not private error". This has prevented users from logging in, it usually resolves itself after several reattempts...

Show all ▾

mistral-large

Yes, these cases might indicate a major incident happening on the Snowflake side. The issues reported include intermittent problems with logging in, slow performance, and glitches in the UI, which could suggest a widespread problem affecting multiple users and functions.



Results so far

Clusters are being recorded since 13.07.2024. From that time there were:

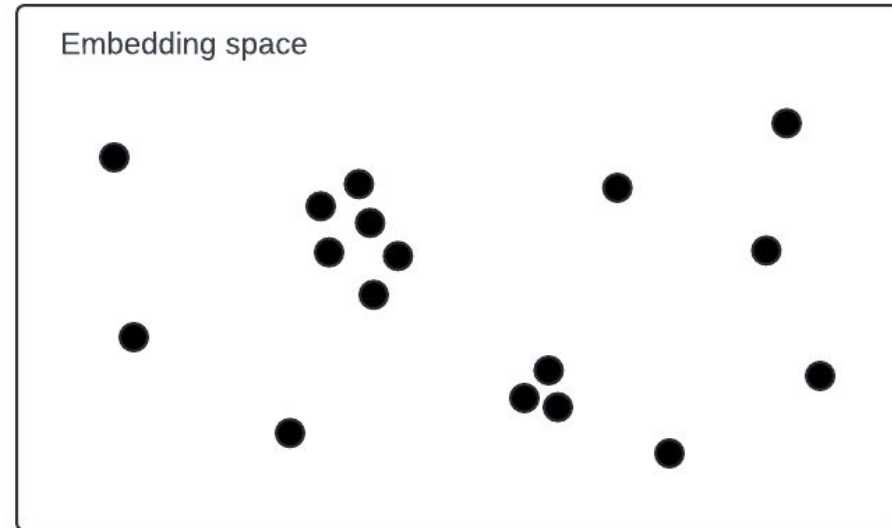
- 25 incidents with at least 1 support case.
- 19 incidents with at least 1 support case before declaration.
- 17 incidents with at least 3 support cases before declaration.
- 6 incidents detected by our service...
- ... and at least 1 prevented 😊

Problem with this algorithm

LLM judgement on whether a given cluster may be related to a major incident or not is unreliable.

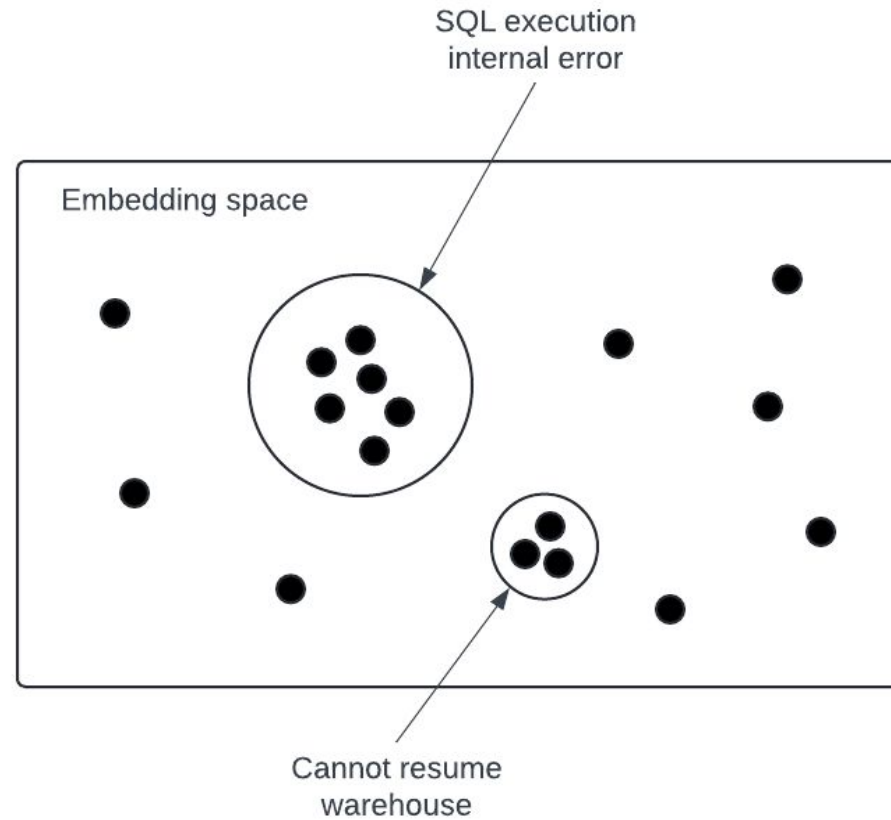
Incorporating anomaly detection

Observation: Not every cluster is equally surprising.

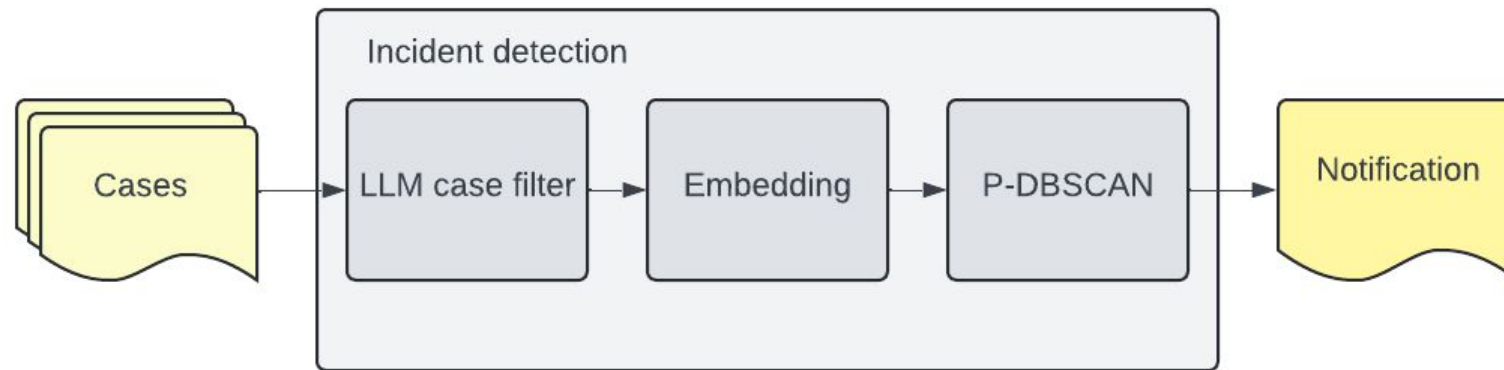


Incorporating anomaly detection

Observation: Not every cluster is equally surprising.



New algorithm



P-DBSCAN - algorithm

1. For each data point:
 - 1.1. Count the number of data points in its eps proximity in the last 24h.
 - 1.2. Count the number of data points in its eps proximity in historical data and divide by the number of days in historical data.
 - 1.3. Calculate the probability of current surroundings, modelling it using a Poisson distribution from historical data.
2. Reject all data points with too high probability (tunable).
3. Apply DBSCAN on all remaining data points.

P-DBSCAN - advantages

- Tunable - we can make it more or less permissive by tuning the probability parameter.
- Explainable - for each cluster we can understand why it was rejected or not.
- Deterministic - we can repeat experiments with complete certainty as there is no randomness in the algorithm.

QUESTIONS?



UNIVERSITY RECRUITING PROGRAM



Life in the SnowClub



PROFESSIONAL DEVELOPMENT

Events to help you learn and grow as a professional



BUILDING CONNECTIONS

Events to foster relationships and embrace diversity



SNOWFLAKE FOR GOOD

Events centered around giving back to our community



SNOW AND TELL



Deliver a project presentation to one's manager, team, or greater org!

Intern project recordings will be shared globally for all to view!



The Full SnowClub Experience!

CURRENT OPPORTUNITIES



We're Hiring!

Join our team and help us mobilize the world's data

- Discover opportunities across our different teams
- Explore our University Recruiting program
- Join our Talent Community
- Check out our employee blog

careers.snowflake.com/university



QUESTIONS?



