

SYSTÈME RAG AVANCÉ

Conception et déploiement d'un système de recommandation d'événements culturels à partir d'OpenAgenda

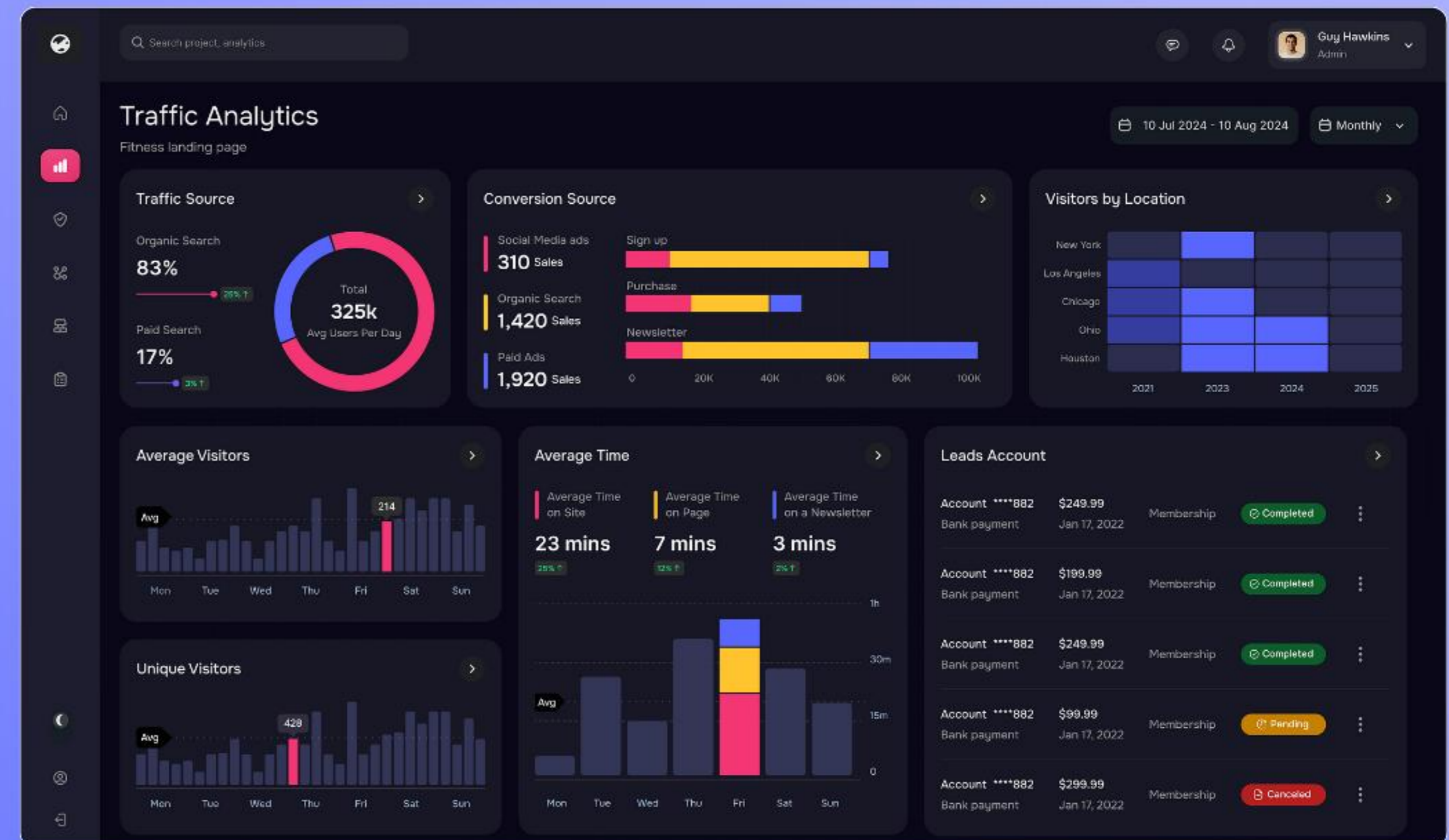
Rudy Desplan | Projet 11 — Master Data Engineer
Encadrant : Jérémie | OpenClassrooms

CONTEXTE & PROBLÉMATIQUE

ASSISTANT PULS-EVENTS

Besoin d'un assistant intelligent pour guider les utilisateurs à travers un corpus d'événements culturels massifs.

- 🧠 Un LLM seul hallucine et ignore les données locales.
- 🕒 Données dynamiques (dates, lieux, statuts).
- ⚙️ Nécessité d'un système RAG contrôlé.



OBJECTIFS DU PROJET



PIPELINE ROBUSTE

Construire une architecture modulaire couvrant tout le cycle : données → retrieval → évaluation.



RÉDUCTION HALLUCINATION

Garantir la fiabilité des réponses par un contrôle strict du contexte injecté.



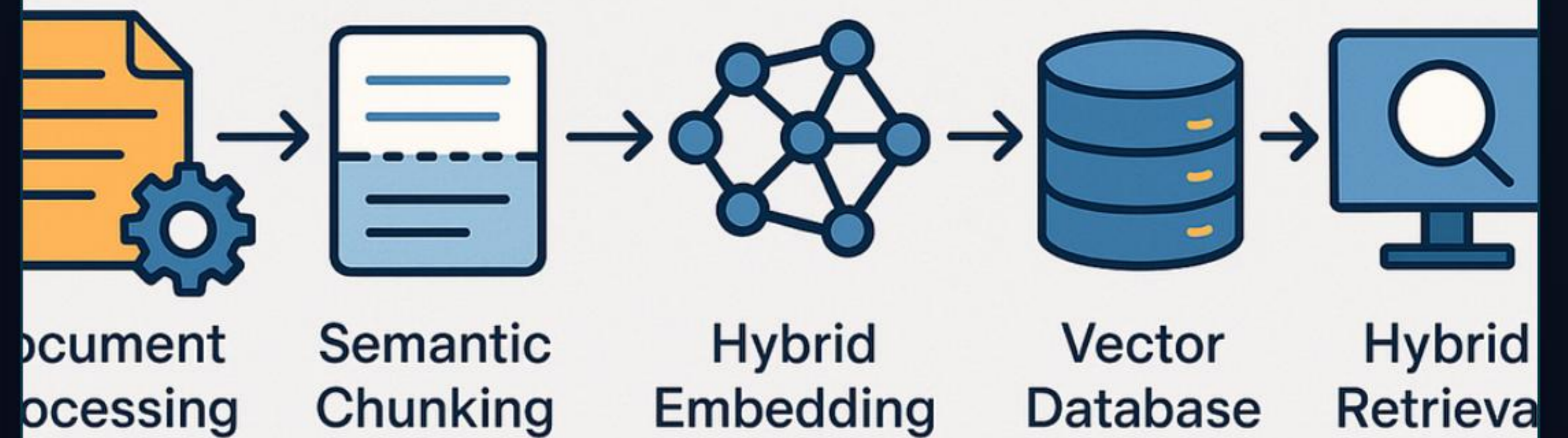
ÉVALUATION RIGOUREUSE

Justification de chaque brique via des mesures quantitatives (Recall, NDCG, Scores Juge).

ARCHITECTURE DU PIPELINE RAG

- 🔍 **Retriever** : BGE-M3 + FAISS IndexFlatIP.
- 🔽 **Reranker** : Cross-encoder BGE (Hybrid Adaptive).
- 📊 **Builder** : Format TOON tabulaire compact.
- 🤖 **Générateur** : Gemini 2.5 Flash via LCEL.




I Rebuilt My RAG Pipeline Times - Here's the Architecture



| LE CORPUS OPENAGENDA

GRANULARITÉ ÉLEVÉE

56 champs structurés permettant une recommandation personnalisée (âge, ville, accessibilité).

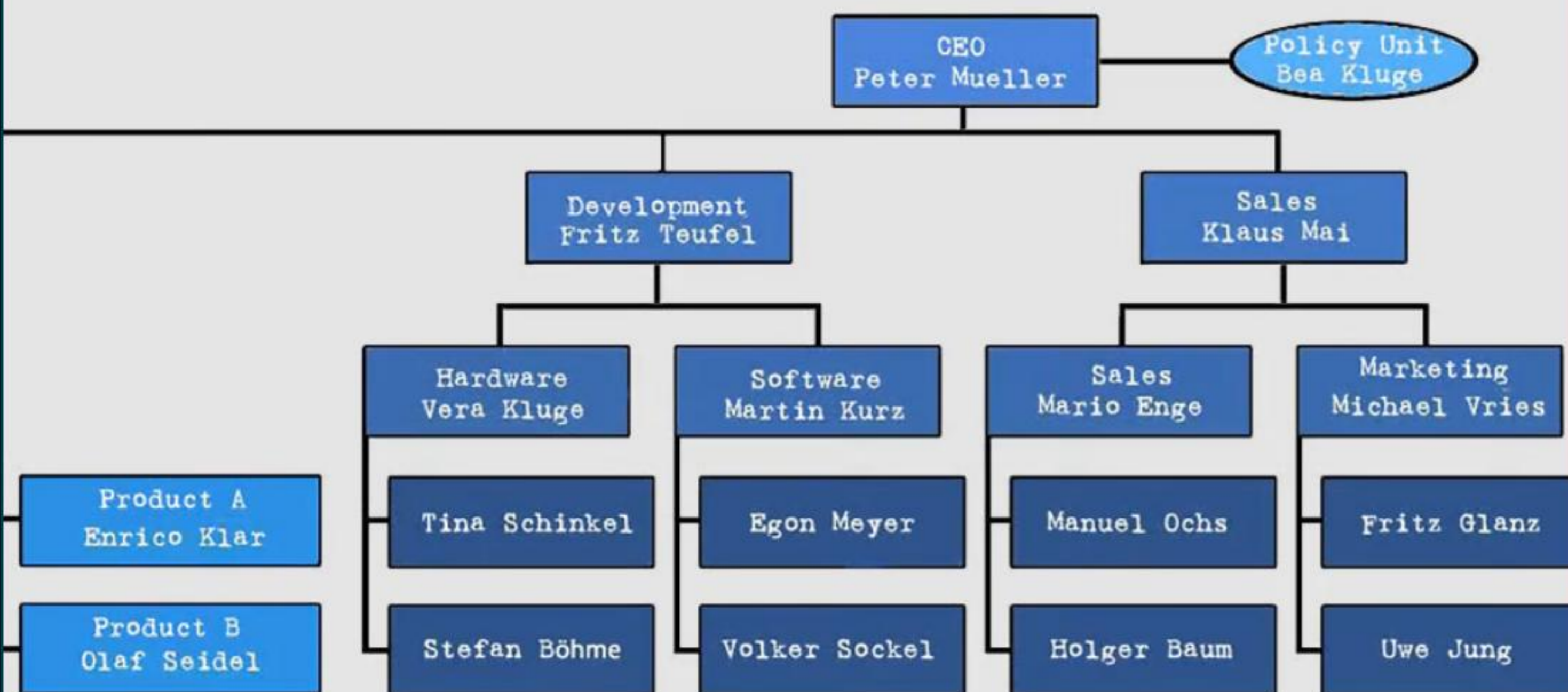
-  Texte riche : Titres & Descriptions longues.
-  Données temporelles et géographiques précises.
-  Documents atomiques : 1 événement = 1 chunk.

ADAPTÉ AU RAG

Combinaison optimale de texte et métadonnées pour une recherche hybride (sémantique + filtres).

INGESTION & PRÉPROCESSING

ORGANISATION CHART



⚡ **Chargement** : Utilisation de orjson (performance).

🔍 **Filtrage** : 1 an d'historique + événements futurs.

🧹 **Nettoyage** : Désérialisation JSON et gestion des NaN.

📦 **Export** : JSONL optimisé pour l'indexation FAISS.

| QUALITÉ DES DONNÉES

100%

Dépendance du LLM au contexte injecté

"Un document textuel propre et structuré augmente drastiquement la pertinence du retrieval." — Rothman

| INDEXATION VECTORIELLE

EMBEDDINGS BGE-M3

Multilingue et support natif des contextes longs (8192 tokens).
Modèle SOTA pour le retrieval dense.

INDEX FAISS INDEXFLATIP

Choix de la simplicité : reproductibilité parfaite, scores 1-to-1 via
normalisation L2.



| CHOIX TECHNIQUE : FAISS CPU

VITESSE

Écrit en C++, offre des performances exceptionnelles en environnement local.

SCALABILITÉ POC

FlatIP est optimal pour ~72k docs.
Le passage à HNSW n'est requis qu'au delà de 1M docs.

INTÉGRATION

Large adoption industrielle et intégration native avec LangChain.

ÉVALUATION DU RETRIEVER

| Métrique | Résultat | Interprétation |
|-----------------|----------|--|
| Recall @ 5 | 99.99 % | Le document pertinent est quasi toujours trouvé. |
| Rank Médian | 1 | Majorité des résultats en première position. |
| Score Gap (P95) | 0.033 | Les échecs sont associés à une faible confiance. |

| LE BESOIN DE RERANKING

RETRIEVER DENSE

High Recall / Low Precision. Rapide mais peut être imprécis sur les détails fins.

RERANKER CROSS-ENCODER

High Precision / Expensive. Analyse simultanée (question, doc) pour un tri optimal.

Huyen (Chap. 6) recommande cette architecture hybride pour les RAG professionnels.

OBSERVATIONS DU RERANKER

📈 **Impact** : Amélioration du rang dans 18.2 % des cas.

⚠️ **Risque** : Dégradation dans 16.3 % des cas.

📉 **Corrélation Confiance/Erreur** : Les dégradations sont souvent associées à des scores de confiance élevés (1.23).

Conclusion : Un reranker ne doit jamais être utilisé aveuglément en production.

| SÉCURITÉ : CONFIDENCE GATING

PRINCIPE DE NEUTRALISATION

Le reranker est bypassé si deux conditions sont réunies :




- Forte confiance du reranker.

- Désaccord flagrant avec le retriever dense.



Réduction massive du taux d'injection dangereuse (de 27.4% à 4.7%).

| FORMAT DE CONTEXTE TOON

-  **Tabulaire** : Compact et explicite pour le LLM.
-  **Optimisé** : Plus de données dans le même budget token.
-  **Déterministe** : Nettoyage, déduplication et budget tokens intelligent.

```
toon
documents[N]{uid,dense_score,rank_score,content}:
1234,0.91,0.87,"Contenu nettoyé..."
5678,0.83,0.80,"Contenu nettoyé..."
```

KPI DU CONTEXT BUILDER



RECALL RATE

100 %

Le bon document est toujours conservé après construction.



POSITION MOYENNE

1.55

L'information clé est priorisée pour l'attention du LLM.






CONTEXT COST

0.11

Efficiency optimale du budget tokens via TOON.

| ORCHESTRATION LANGCHAIN (LCEL)

POURQUOI LANGCHAIN ?

-  **Modularité** : Wrappers personnalisés pour chaque brique.
-  **Mémoire** : ConversationBufferMemory pour le suivi naturel.
-  **Observabilité** : Traçabilité via callbacks.

LangChain n'est introduit qu'après stabilisation des briques fondamentales pour garantir une architecture propre.

ARCHITECTURE LCEL FINALE

RETRIEVER

Dense search

CONTEXT

TOON Builder

RERANKER

Cross-encoder

L M

Gemini 2.5

`RunnableWithMessageHistory` wrap la chaîne RAG pour la mémoire conversationnelle.

| GÉNÉRATION : GEMINI 2.5 FLASH

CONFIGURATION

Température basse (0.0) pour un comportement déterministe.

Prompt Strict : Pas de connaissances externes. Hallucination = 1 si info absente du TOON.

ROBUSTESSE

Parsing JSON robuste pour extraire scores et explications sans rupture de pipeline.

| ÉVALUATION MULTI-NIVEAUX

COMPOSANTS

Recall@k, Rank, Score Gap sur le retriever et reranker.

CONTEXTE

Recall Rate, NDCG et Position moyenne sur le builder.

END-TO-END

LLM-as-a-Judge pour la qualité de la réponse utilisateur.

| LLM-AS-A-JUDGE

LE JUGE

Modèle : gemini-3-flash-preview (stricte, non créative).

Faithfulness : Fidélité au contexte.

Relevance : Réponse à la question.

Hallucination : Signal strict 0/1.

Évaluation sémantique de bout en bout simulant le regard de l'utilisateur final.

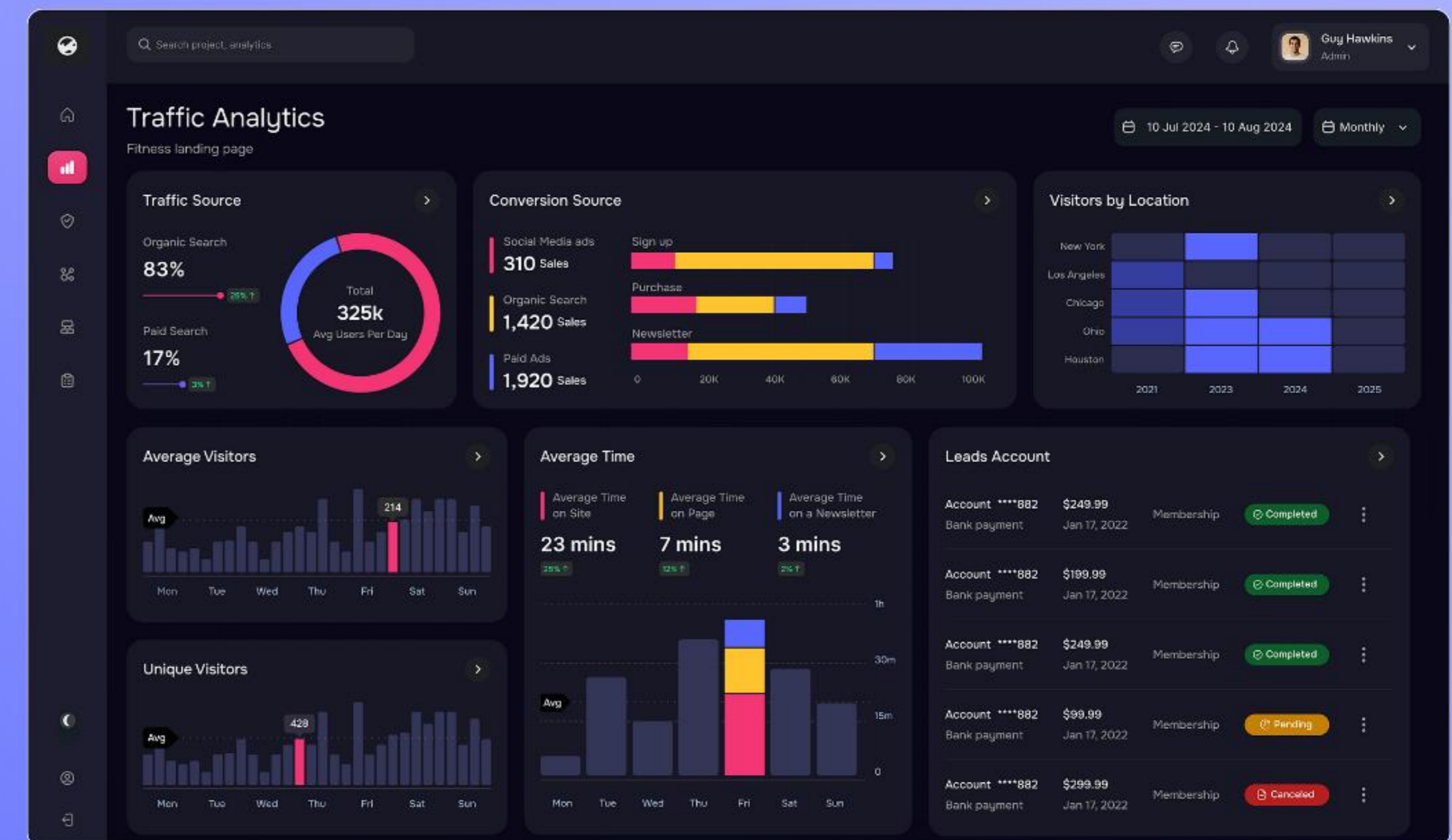
SYNTHÈSE DES RÉSULTATS

| Métrique Finale | Score | Analyse |
|----------------------|---------|---|
| Faithfulness | 4.4 / 5 | Très haute fidélité factuelle. |
| Relevance | 5.0 / 5 | Pipeline comprend parfaitement les questions. |
| Hallucination Métier | 0 % | Aucune invention factuelle observée. |

Note : Un taux brut de 50% d'hallucinations système a été requalifié (références à TOON).

DÉMONSTRATION LIVE

- ▶ Lancement de l'interface Chainlit.
- 💬 Recommandation par thèmes (ex: "événements jazz").
- 📍 Filtres géo-temporels.
- ⊘ Comportement en cas de contexte insuffisant.



| LIMITES & DÉFIS

SCALABILITÉ

FAISS Flat Index est dépendant de la RAM et non managé.





RERANKER

Encore perfectible sur certaines nuances sémantiques fines.

STATIQUE

Manque de RAG agentique pour des appels API dynamiques.

| PERSPECTIVES D'ÉVOLUTION

-  Migration vers base managée (Pinecone) + Index HNSW.
-  Transition vers un RAG Agentique avec guardrails (NeMo).
-  Intégration de boucles de feedback utilisateur réelles.
-  Déploiement Cloud pour production scale-out.

| CONCLUSION

POC

Validé : Architecture robuste et factuellement fiable

Le projet démontre la faisabilité d'un assistant de recommandation contrôlé, aligné sur les bonnes pratiques de Rothman et Huyen, prêt pour une phase de déploiement cloud.

MERCI POUR VOTRE ATTENTION

Avez-vous des questions techniques sur le pipeline ?

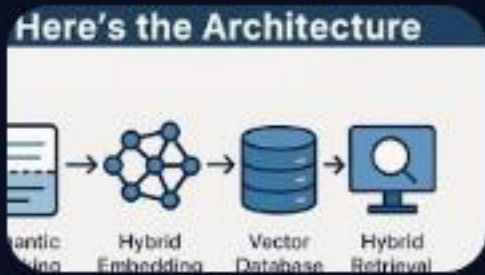
Rudy Desplan | Master Data Engineer

IMAGE SOURCES



<https://cdn.dribbble.com/userupload/43387716/file/original-9dc9eca58aac68b6b25ef5be9f82e78b.png?resize=2048x1536&vertical=center>

Source: dribbble.com



https://miro.medium.com/v2/resize:fit:1200/1*NybUdowBAV76fcQKbmOrNA.png

Source: medium.com



<https://t2informatik.de/en/wp-content/uploads/sites/2/2024/01/organisation-chart.png>

Source: t2informatik.de



<https://static.vecteezy.com/system/resources/thumbnails/025/255/749/small/abstract-digital-futuristic-with-plexus-background-big-data-visualization-global-connection-networking-science-and-technology-design-concept-illustration-vector.jpg>

Source: www.vecteezy.com



<https://static.vecteezy.com/system/resources/thumbnails/029/920/098/small/abstract-digital-technology-futuristic-secure-key-lock-safe-blue-background-cyber-security-science-tech-innovation-future-ai-big-data-internet-network-connection-cloud-hi-tech-illustration-vector.jpg>

Source: www.vecteezy.com