

PASSEZ UN SYSTÈME IA DU POC AU MVP

Passage du POC au MVP – Chatbot RAG

Projet Data Engineer – KELLENI Antoine

ROADMAP DE LA MISSION

1. **Contexte & Objectifs**
2. Réalisation et limites du POC
3. Enjeux et phases du MVP
4. Architecture globale de la solution
5. Choix du cloud provider
6. Macro Backlog
7. Estimations des coûts
8. Conclusion

Contexte & objectifs

Contexte :

- Besoin de recommandation d'événements culturels pertinents
- Utilisation de données publiques issues d'OpenAgenda
- Recherche conversationnelle plus naturelle via IA
- Approche basée sur un chatbot RAG (Retrieval Augmented Generation)

Objectifs :

- Valider la faisabilité d'un chatbot RAG (POC)
- Transformer le POC en MVP exploitable
- Concevoir une architecture scalable et modulaire
- Maîtriser les coûts et les risques techniques
- Préparer une mise en production cloud



ROADMAP DE LA MISSION

1. Contexte & Objectifs
2. **Réalisation et limites du POC**
3. Enjeux et phases du MVP
4. Architecture globale de la solution
5. Choix du cloud provider
6. Macro Backlog
7. Estimations des coûts
8. Conclusion

Réalisation du POC

Réalisation clés effectués dans le passé :

Pipeline Python local

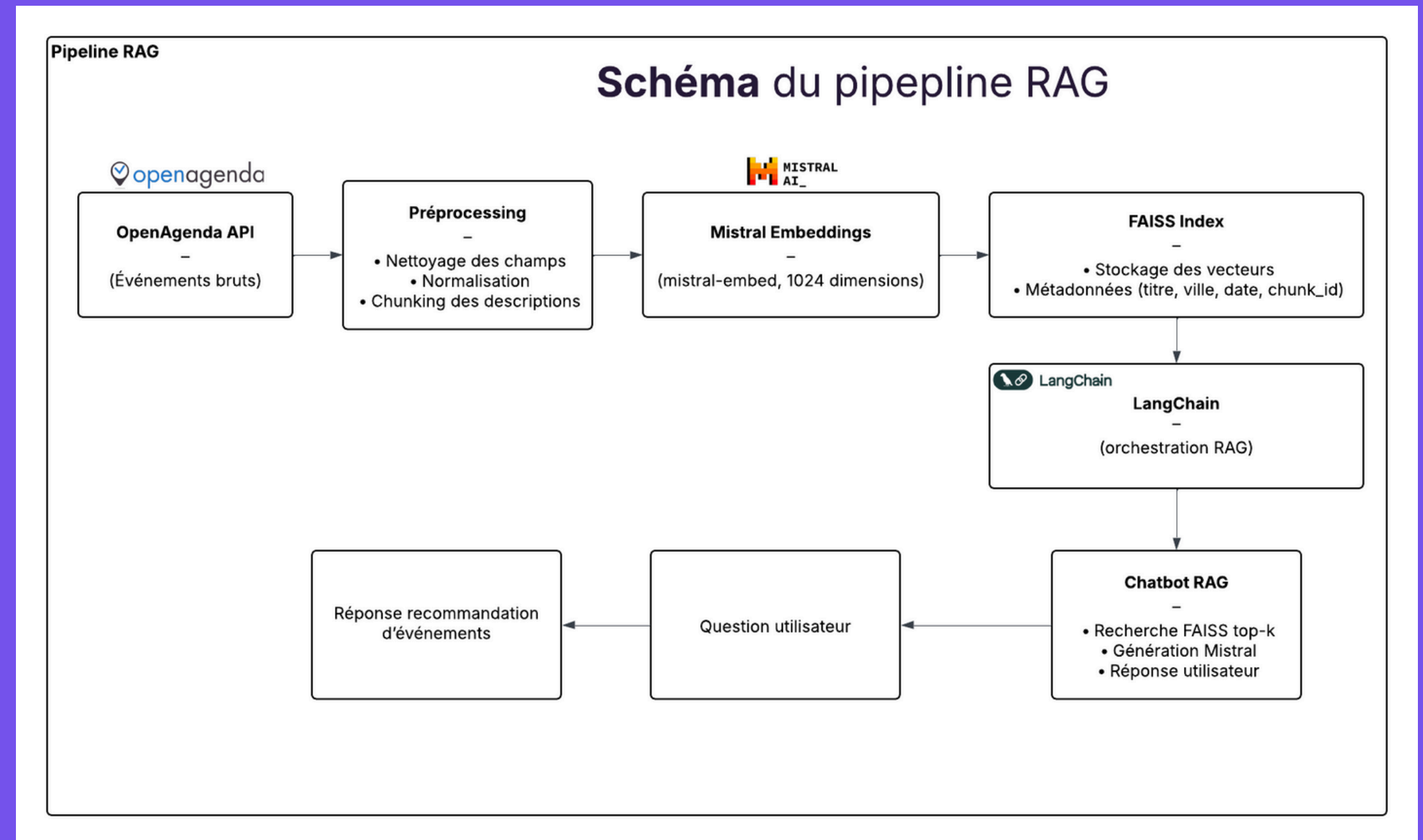
Ingestion des données OpenAgenda

Nettoyage et normalisation

Vectorisation NLP (embeddings Mistral)

Indexation FAISS

Chatbot conversationnel fonctionnel



Limites identifiées du POC

- Quotas et crédits API Mistral limitant la génération d'embeddings
- Erreurs API lors de volumes plus importants
- Architecture locale non scalable
- Absence de déploiement cloud
- Pas de monitoring ni d'observabilité
- Interface utilisateur minimale (terminal)

Embeddings Mistral

Modèle utilisé : mistral-embed
Dimension : 1024
Calcul par batchs de 64
Résultat obtenu → openagenda_events_embeddings.npy

Contraintes rencontrées :

- Quota Mistral → limitation à 200 embeddings
- Gestion des erreurs API (code 429)

```
OK Batch 0 -> 64
OK Batch 64 -> 128
OK Batch 128 -> 148
OK Embeddings générés
OK Données sauvegardées dans data\openagenda_events_preprocessed.csv
OK Embeddings sauvegardés dans data\openagenda_events_embeddings.npy
OK Préprocessing terminé
```

```
raise models.SDKError("API error occurred", http_res, http_res.text)
mistralai.models.sdkerror.SDKError: API error occurred: Status 429. Body: {"object": "error", "message": "Service tier capacity exceeded for this model."}
```

12

openagenda Mistral AI LangChain

Chatbot intelligent

CHATBOT RAG (FAISS + MISTRAL)

Pipeline RAG :

- 1.Embedding de la requête utilisateur
- 2.Recherche FAISS top-k
- 3.Sélection du contexte pertinent
- 4.Génération Mistral
- 5.Réponse naturelle et contextualisée

Fonctionnalités :

- Recommandation d'événements
- Filtrage implicite par ville / type
- Reformulation intelligente

```
PS C:\Users\anto\Documents>python chatbot_rag.py
OK Chatbot prêt (FAISS + Mistral). Tape 'quit' pour sortir.

User: Combien avez vous en possession d'événements ? Peux tu me détailler le nombre par ville aussi ?

Bot:
- Juppé-en-Deux : 2 événements
- Paris : 17 événements
- Placide : 2 événements
- Région de France : 3 événements
- Versailles : 55 événements
- Virefay : 5 événements
Total général d'événements : 100

User: J'aimerais sortir le weekend à versailles, as tu des recommandations à me proposer pour une sortie en amoureux ?

Bot: Pour une sortie en amoureux à Versailles ce week-end, je te recommande les événements suivants :
1. "Journée découverte au Centre de musique baroque de Versailles" (le 6 décembre)
   "Pourquoi ?" : Cette journée propose une variété d'activités comme des ateliers, des spectacles, des visites guidées et un jeu de piste ludique. C'est une excellente occasion de passer un moment culturel et ludique ensemble. Le jeu de piste "La saga Papillon à Versailles" est particulièrement adapté pour une sortie en couple, car il permet de découvrir Versailles de manière interactive et amusante.
2. "Soirée danseuse ANNELEIGH - Versailles Indigo Dance" (le 6 décembre)
   "Pourquoi ?" : Si vous aimez danser, cette soirée organisée par Versailles Indigo Dance pour le Mithos est une excellente option. L'ambiance festive et solidaire peut rendre votre sortie encore plus spéciale. De plus, le prix d'entrée est accessible (14 € minimum), ce qui permet de profiter de la soirée sans se ruiner. Malheureusement, la braderie de jouets ne semble pas adaptée pour une sortie en amoureux.

User: pour cette fin d'année, je compte passer 2 jours à Paris, peux tu me proposer 3 activités à faire en famille avec mes enfants ?

Bot: Voici trois activités adaptées pour une sortie en famille à Paris pendant les fêtes de fin d'année :
1. "Contes et histoires : Noël en famille"
   "Lieu" : Hôtel de la Marine, Paris
   "Dates" : 5 décembre 2023 à 18h00
   "Description" : Cet événement propose des contes et des histoires pour célébrer Noël en famille dans un cadre scintillant. C'est une activité idéale pour les enfants, qui pourront profiter de l'ambiance festive tout en écoutant des récits captivants.
2. "Sortie ludique : Escape Game"
   "Lieu" : Paris
   "Dates" : 11 décembre 2023 à 10h00
   "Description" : Une sortie en escape game pour les jeunes de 11 à 17 ans. Les enfants pourront résoudre des énigmes, travailler leur esprit logique et vivre une aventure ludique. Cette activité est parfaite pour une journée en famille, où les enfants pourront tester leurs capacités de résolution de problèmes et de travail d'équipe.
3. "Sortie ludique : Escape Game"
   "Lieu" : Paris
   "Dates" : 11 décembre 2023 à 10h00
   "Description" : Une autre option d'escape game, cette fois dans une salle de 400m². Les enfants pourront vivre la frisson d'une chasse au trésor et résoudre des énigmes dans un environnement immersif. Cette activité est idéale pour une journée en famille, où les enfants pourront développer leurs compétences en communication et en gestion du temps.
```

openagenda Mistral AI LangChain

ROADMAP DE LA MISSION

1. Contexte & Objectifs
2. Réalisation et limites du POC
3. **Enjeux et phases du MVP**
4. Architecture globale de la solution
5. Choix du cloud provider
6. Macro Backlog
7. Estimations des coûts
8. Conclusion

Enjeux du passage au MVP

Pourquoi passer au MVP ?

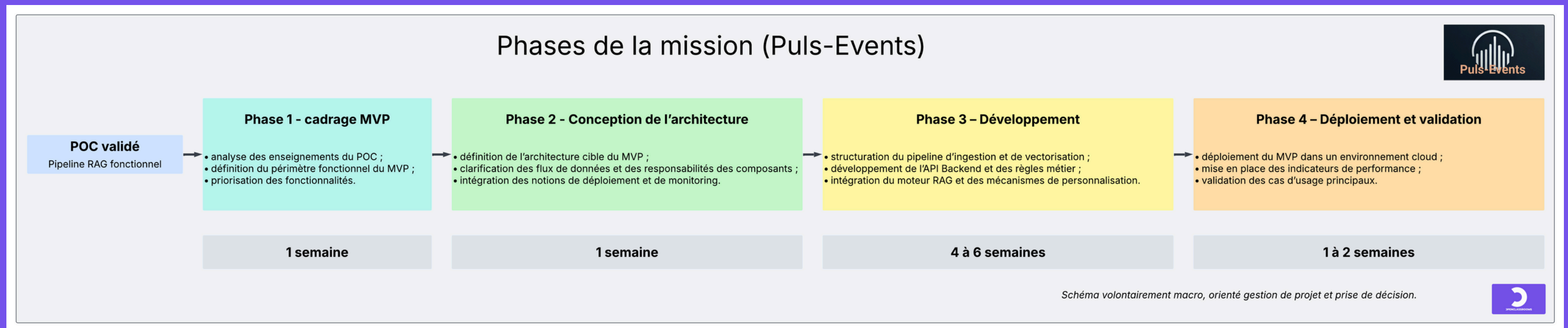
- Industrialiser le pipeline existant
- Séparer clairement les responsabilités
- Améliorer la maintenabilité
- Anticiper la montée en charge
- Intégrer déploiement, monitoring et maîtrise des coûts

Transition vers le MVP

- Analyse critique des limites du POC
- Définition d'un périmètre fonctionnel réaliste
- Priorisation des fonctionnalités (Must / Nice)
- Choix d'une architecture cible orientée production



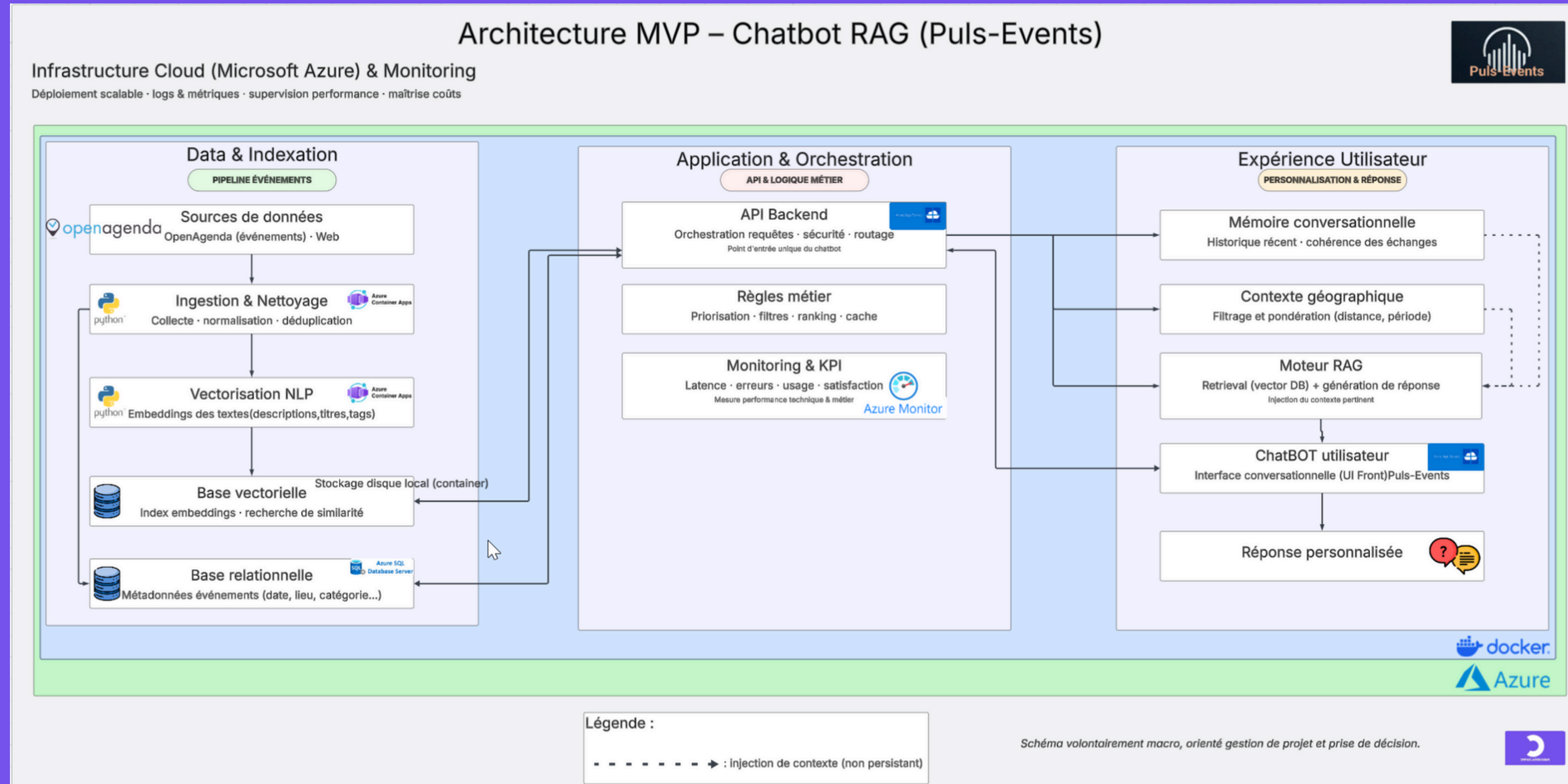
Les différentes phases du MVP



ROADMAP DE LA MISSION

1. Contexte & Objectifs
2. Réalisation et limites du POC
3. Enjeux et phases du MVP
4. **Architecture globale de la solution**
5. Choix du cloud provider
6. Macro Backlog
7. Estimations des coûts
8. Conclusion

Architecture globale de la solution

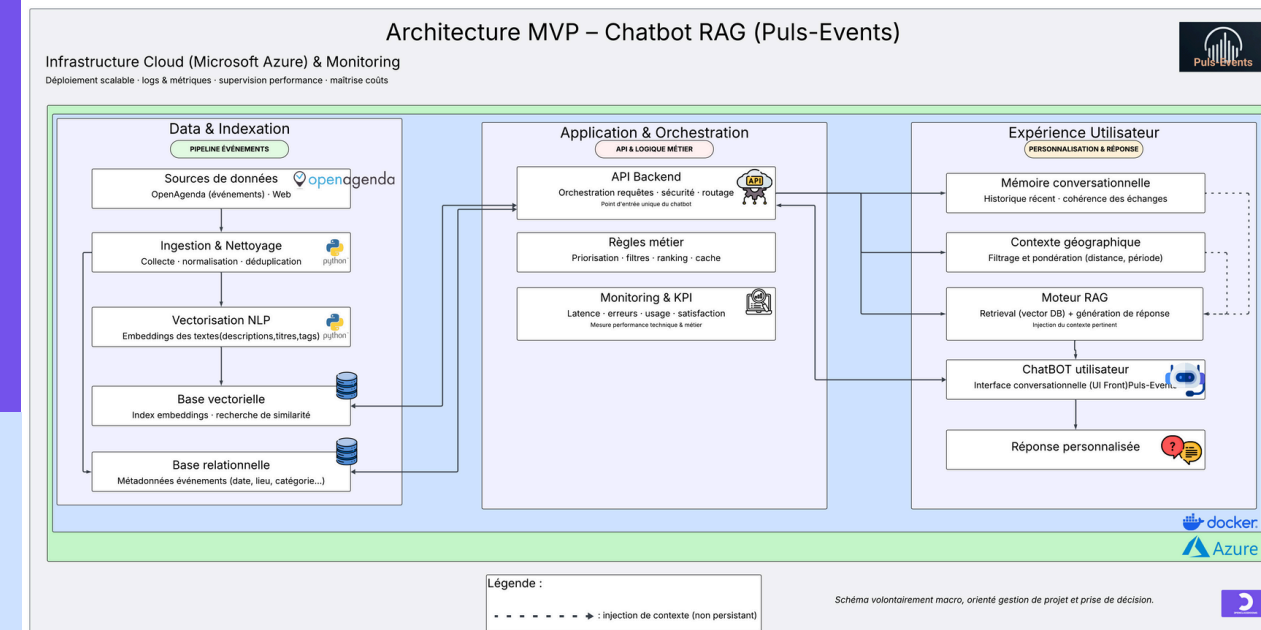
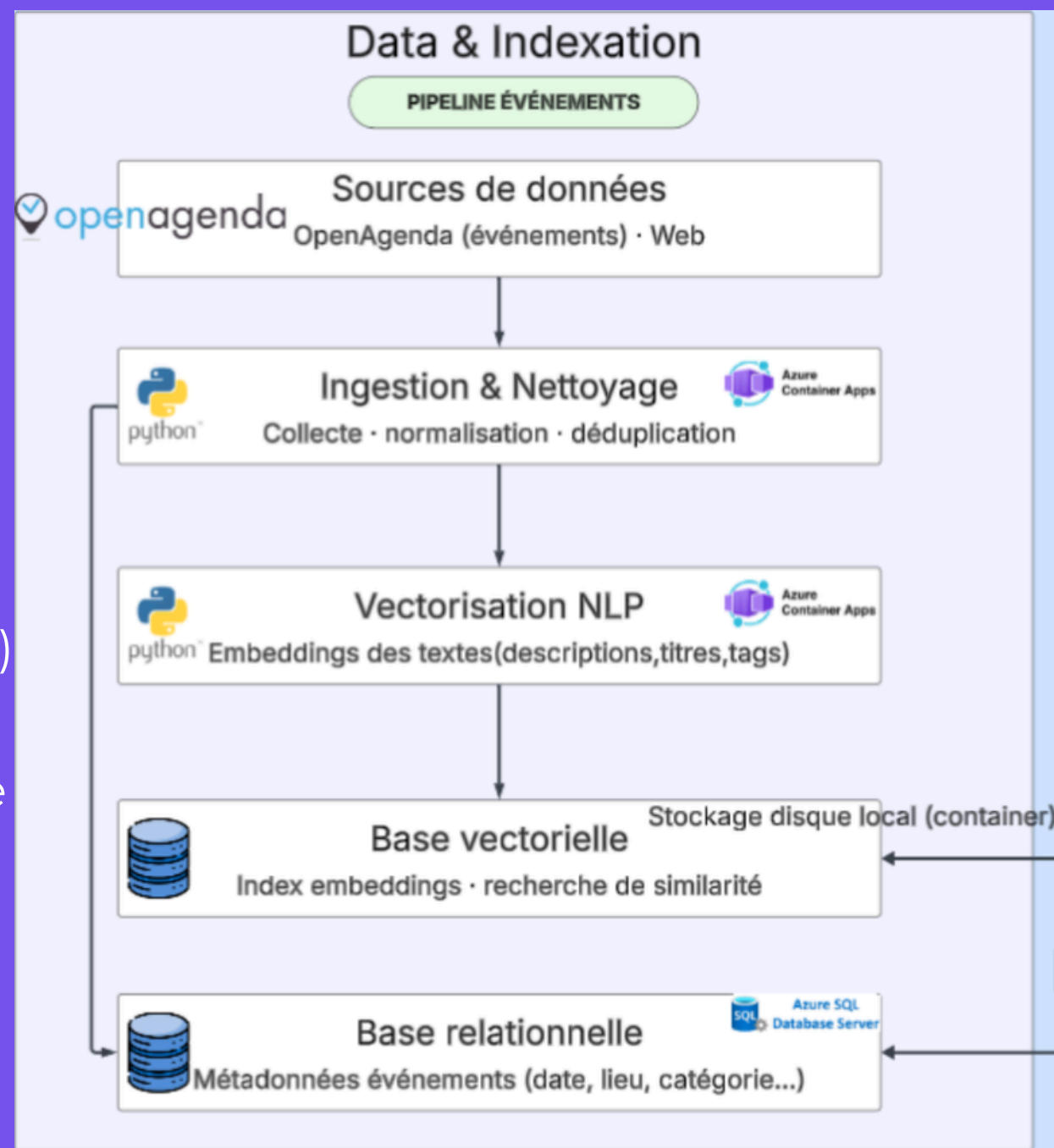


Séparation en trois couches :
Data & Indexation
Application & Orchestration
Expérience Utilisateur

API Backend comme point d'entrée unique
Intégration d'un moteur RAG
Architecture modulaire et évolutive

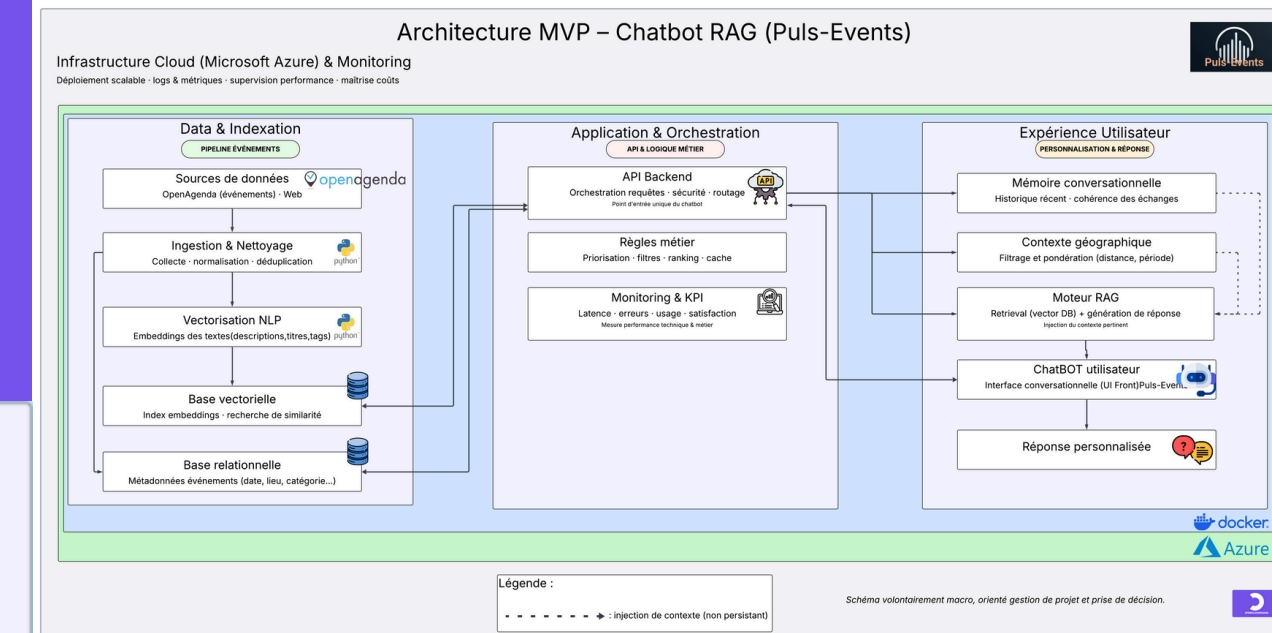
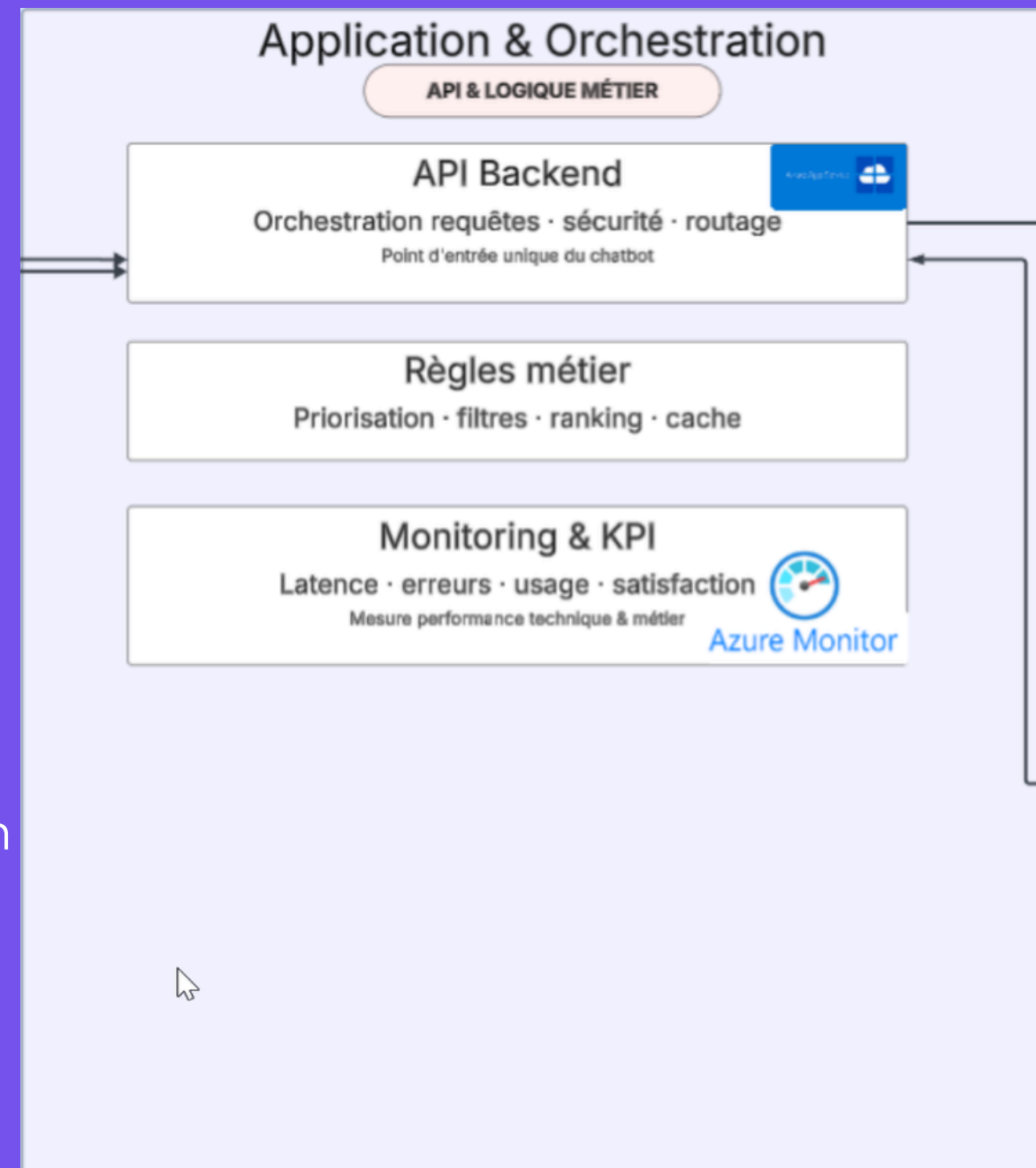
1er couche : Data & Indexation

Sources : OpenAgenda (événements)
Ingestion et nettoyage automatisés
Vectorisation NLP
Base relationnelle :
 métadonnées (date, lieu, catégorie...)
Base vectorielle :
 recherche sémantique par similarité



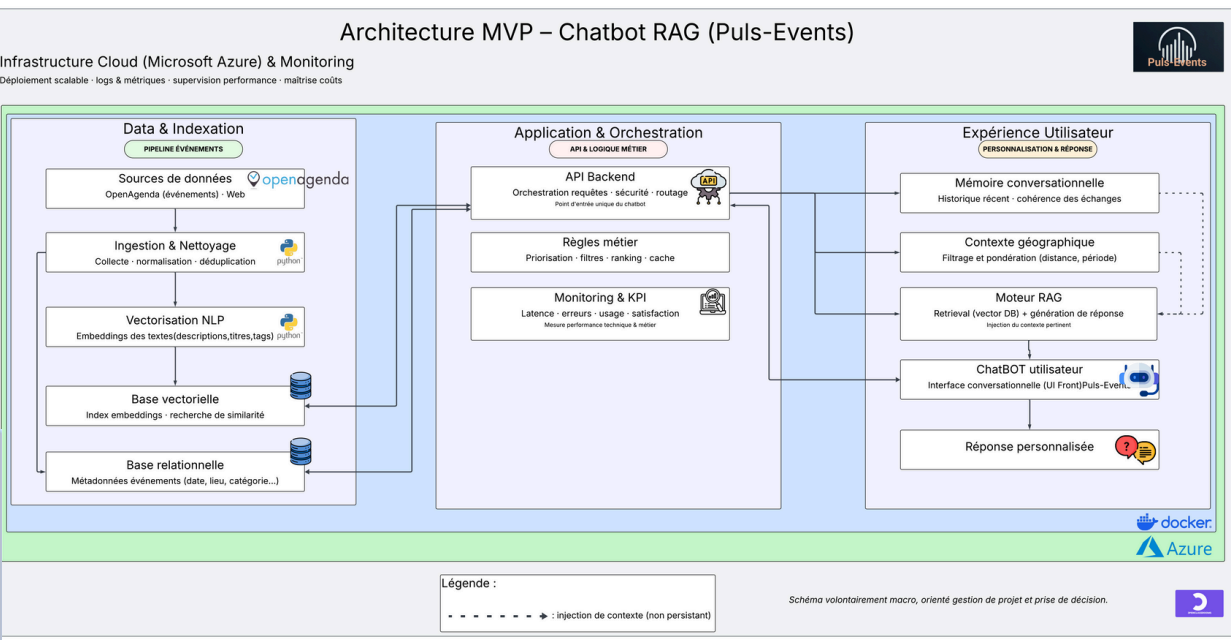
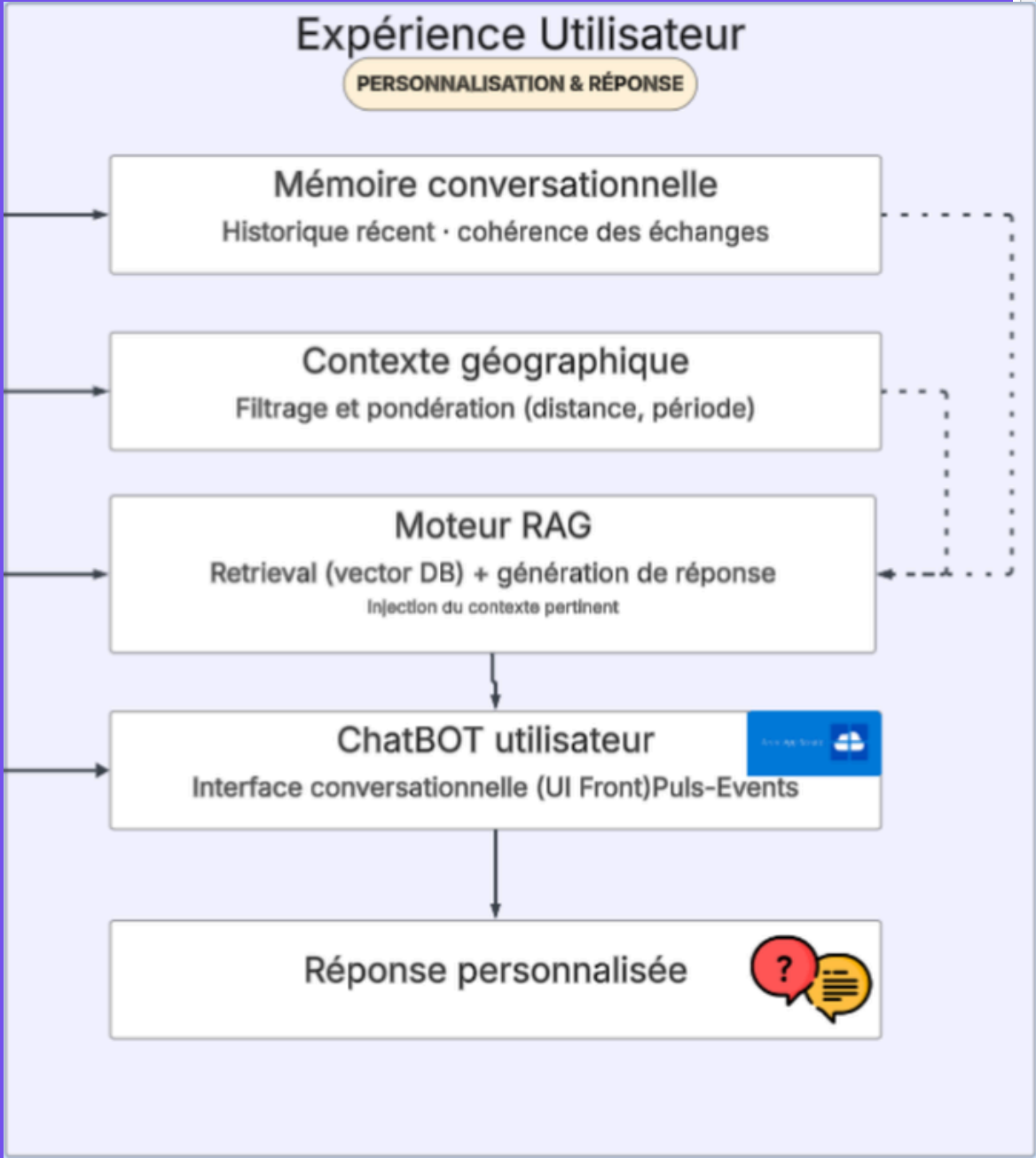
2e couche : Application & Orchestration

- API Backend :
 - orchestration des requêtes
 - sécurité et routage
- Règles métier :
 - filtres, priorisation, ranking, cache
- Monitoring & KPI :
 - latence, erreurs, usage, satisfaction



3eme couche : Expérience Utilisateur

Interface chatbot conversationnelle
Mémoire conversationnelle (court terme)
Contexte géographique injecté
Moteur RAG :
 retrieval depuis la base vectorielle
 génération de réponses contextualisées
Réponse personnalisée à l'utilisateur



ROADMAP DE LA MISSION

1. Contexte & Objectifs
2. Réalisation et limites du POC
3. Enjeux et phases du MVP
4. Architecture globale de la solution
5. **Choix du cloud provider**
6. Macro Backlog
7. Estimations des coûts
8. Conclusion

Choix du Cloud Provider

- Veille menée sur :
 - AWS
 - GCP
 - Microsoft Azure
- Choix retenu : Microsoft Azure

Tableau – Comparaison des cloud providers pour le MVP RAG

Critère	Microsoft Azure	Amazon Web Services	Google Cloud Platform
Positionnement	Généraliste, orienté entreprise	Très large, très granulaire	Data & ML <u>centric</u>
Simplicité de mise en œuvre (MVP)	Élevée	Moyenne à faible	Moyenne
Services managés conteneurs	App Service / Container Apps simples	ECS / EKS plus complexes	Cloud Run efficace
Observabilité native	Azure Monitor intégré	<u>CloudWatch</u> à configurer	Cloud Monitoring
Lisibilité des coûts (faible charge)	Bonne	Plus complexe	Correcte
Adaptation à une démarche MVP	Très bonne	Moyenne	Bonne
Scalabilité long terme	Élevée	Très élevée	Élevée
Temps de <u>delivery</u>	Court	Plus long	Moyen

Justifications

- Azure offre le meilleur compromis pour un MVP : simplicité, services managés et coûts maîtrisés à faible charge.
- AWS est plus complexe et surdimensionné pour cette phase.
- GCP est très orienté data/ML avancé, moins adapté à un MVP applicatif.



ROADMAP DE LA MISSION

1. Contexte & Objectifs
2. Réalisation et limites du POC
3. Enjeux et phases du MVP
4. Architecture globale de la solution
5. Choix du cloud provider
6. **Macro Backlog**
7. Estimations des coûts
8. Conclusion

Macro Backlog

Fonctionnalité	Description	Priorité	Complexité	Risques identifiés	Mitigation
Ingestion OpenAgenda	Collecte automatisée des événements via API OpenAgenda	Must-Have	Moyenne	Volume important	Pagination, filtres
Nettoyage & normalisation	Nettoyage texte, normalisation champs, déduplication	Must-Have	Moyenne	Données hétérogènes	Règles de nettoyage
Stockage relationnel	Base relationnelle pour métadonnées (date, lieu, catégorie...)	Must-Have	Moyenne	Modèle mal adapté	Schéma simple & évolutif
Chunking des descriptions	Découpage des textes longs avec overlap	Must-Have	Faible	Perte de contexte	Overlap contrôlé
Génération embeddings	Embeddings NLP avec Mistral	Must-Have	Moyenne	Quota API	Limitation volume
Base vectorielle	Index FAISS pour recherche sémantique	Must-Have	Moyenne	Performance	Index simple
API Backend	Point d'entrée unique du chatbot	Must-Have	Élevée	Couplage fort	Architecture modulaire
Moteur RAG	Retrieval + génération de réponse	Must-Have	Élevée	Hallucinations	Contexte contrôlé
Chatbot utilisateur	Interface conversationnelle (CLI / UI simple)	Must-Have	Faible	UX limitée	MVP assumé
Règles métier	Filtres date / ville / catégorie / ranking	Nice-to-Have	Moyenne	Complexité logique	Règles simples
Contexte géographique	Pondération par distance	Nice-to-Have	Moyenne	Données imprécises	Géoloc simplifiée
Monitoring & KPI	Logs, erreurs, latence	Nice-to-Have	Faible	Manque visibilité	Logs centralisés



ROADMAP DE LA MISSION

1. Contexte & Objectifs
2. Réalisation et limites du POC
3. Enjeux et phases du MVP
4. Architecture globale de la solution
5. Choix du cloud provider
6. Macro Backlog
7. **Estimations des coûts**
8. Conclusion

Estimation des coûts - BUILD

Data Engineer freelance : TJM de 500 €

Poste de coût ▼	Description ▼	Nature du coût ▼	Charge estimé ▼	Coût ▼
Cadrage & conception	Analyse POC, définition MVP, architecture cible	Ponctuel	3 jours	1 500 €
Pipeline data	Ingestion OpenAgenda, nettoyage, normalisation	Ponctuel	5 jours	2 500 €
Base relationnelle	Modélisation et mise en place des métadonnées	Ponctuel	3 jours	1 500 €
Vectorisation NLP	Génération des embeddings et indexation	Ponctuel	3 jours	1 500 €
Base vectorielle	Mise en place FAISS et tests de performance	Ponctuel	2 jours	1 000 €
API Backend	Orchestration, sécurité, règles métier	Ponctuel	6 joiurs	3 000 €
Moteur RAG	Retrieval + génération contrôlée	Ponctuel	5 jours	2 500 €
Interface chatbot	Interface utilisateur simple (CLI / UI web)	Ponctuel	2 jours	1 000 €
Tests & validation	Tests fonctionnels et cas d'usage	Ponctuel	2 jours	1 000 €
Documentation	Documentation technique & projet	Ponctuel	2 jours	1 000 €
Total			33 jours	16 500 €

16 500€ d'investissement initial



Estimation des coûts - OPEX

Hypothèses d'usage

- 1 000 utilisateurs / mois
- 5 requêtes chatbot / utilisateur
- 5 000 requêtes RAG / mois
- 50 000 événements OpenAgenda

Outil de calcul de tarification | Microsoft Azure

Votre estimation		Calculatrice de prix		Microsoft	
		Estimate better. Build smarter. Decide faster			
		Prise en main d'Azure			
App Service	Niveau Basic ; 1 B1 (1 Cœur(s), 1.75 Go de RAM, 10 Go de s...	Frais initiaux : 0,00 €	Frais mensuels : 11,17 €		
Azure SQL Database	Base de données unique, modèle d'achat DTU, niveau Basic...	Frais initiaux : 0,00 €	Frais mensuels : 5,20 €		
Azure Monitor	Log analytics : Log Data Ingestion: 0 GB Daily Auxiliary Logs...	Frais initiaux : 0,00 €	Frais mensuels : 9,43 €		
Frais initiaux estimés		0,00 €			
Coût mensuel estimé		25,80 €			

Poste de coût	Détail du périmètre couvert	Hypothèses retenues	Type de coût	Coût mensuel estimé
Hébergement cloud	Hébergement de l'API Backend, du moteur RAG et de l'interface utilisateur via des conteneurs (Azure App Service / Container Apps)	Charge faible à modérée (phase MVP, trafic limité)	Récurrent	12 €
Stockage relationnel	Base de données pour les métadonnées structurées des événements (dates, lieux, catégories, descriptions courtes)	Volume < 1 Go, faible nombre d'écritures	Récurrent	6 €
Stockage vectoriel	Stockage local de l'index FAISS et des embeddings vectoriels	FAISS embarqué, stockage disque inclus dans l'hébergement	Récurrent	0 €
Appels NLP / LLM	Génération des réponses du chatbot (RAG) et appels aux modèles de langage	~5 000 requêtes/mois, ~1 000 tokens par requête	Variable	60 €
Réseau	Trafic entrant/sortant, appels API externes (OpenAgenda, LLM)	Volumes faibles, inclus dans l'offre cloud	Récurrent	Inclus
Monitoring & logs	Collecte des métriques de performance, logs applicatifs, erreurs et supervision	Azure Monitor / Application Insights, faible volumétrie	Récurrent	10 €
Maintenance opérationnelle	Ajustements mineurs, surveillance, correctifs légers	Maintenance minimale en phase MVP	Récurrent	Inclus
Total				88 €

85-90€/mois

Optimisation budgétaire

Levier	Description	Effet
Limitation du périmètre MVP	Fonctionnalités essentielles uniquement	Réduction BUILD
Réduction volume embeddings	Vectorisation ciblée	Réduction OPEX
Cache des réponses	Réutilisation résultats fréquents	Réduction appels LLM
Batch processing	Traitements groupés	Réduction coûts NLP
Montée en charge progressive	Ajustement à l'usage réel	Maîtrise OPEX
Architecture modulaire	Remplacement facile des briques	Pérennité



ROADMAP DE LA MISSION

1. Contexte & Objectifs
2. Réalisation et limites du POC
3. Enjeux et phases du MVP
4. Architecture globale de la solution
5. Choix du cloud provider
6. Macro Backlog
7. Estimations des coûts
8. **Conclusion**

Conclusion

Bilan

Limites techniques identifiées et traitées
MVP structuré et orienté production
Architecture scalable et maintenable
Maîtrise des coûts et des risques

Perspectives

Amélioration de l'expérience utilisateur
Optimisation des coûts LLM
Montée en charge progressive
Déploiement à plus grande échelle

Conclusion

Projet mené de bout en bout : POC → MVP



MERCI