Architecture Microservices pour DeepSeek

Système de Traitement d'Intelligence Artificielle Distribuée

Architecture Technique

22 septembre 2025

Table des matières

1 Introduction

Cette architecture microservices pour DeepSeek présente un système distribué modulaire conçu pour traiter les requêtes d'intelligence artificielle de manière scalable et résiliente. L'architecture est basée sur des principes de séparation des responsabilités, d'indépendance des services, et de communication asynchrone.

2 Vue d'ensemble de l'Architecture

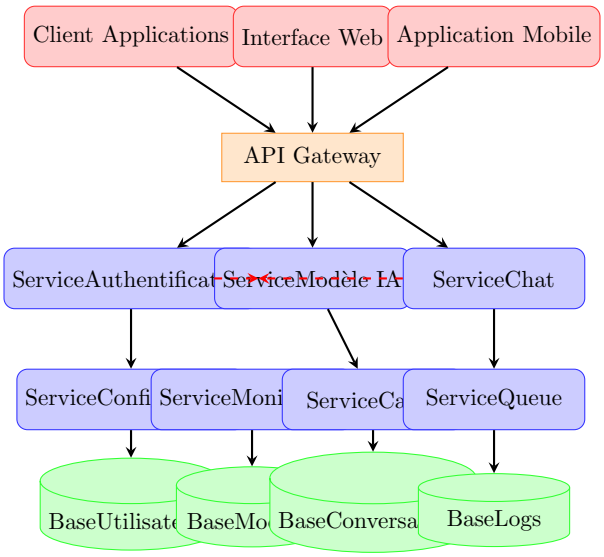

FIGURE 1 – Vue d'ensemble de l'architecture microservices DeepSeek

3 Composants Principaux

3.1 API Gateway

L'API Gateway sert de point d'entrée unique pour toutes les requêtes externes. Il assure :

- Routage intelligent des requêtes
- Authentification et autorisation
- Limitation du taux de requêtes
- Transformation des protocoles
- Agrégation des réponses

3.2 Service d'Authentification

Gère l'identité et l'accès des utilisateurs :

- Authentification JWT
- Gestion des sessions
- Contrôle d'accès basé sur les rôles (RBAC)
- Intégration avec des fournisseurs externes (OAuth)

3.3 Service de Modèle IA

Cœur du système, responsable du traitement des modèles DeepSeek :
— Chargement dynamique des modèles
— Optimisation des requêtes
— Gestion de la mémoire GPU
— Mise à l'échelle automatique

3.4 Service de Chat

Gère les conversations et interactions :
— Historique des conversations
— Gestion des contextes
— Streaming des réponses
— Personnalisation des interactions

4 Architecture de Déploiement

5 Flux de Données

5.1 Traitement d'une Requête IA

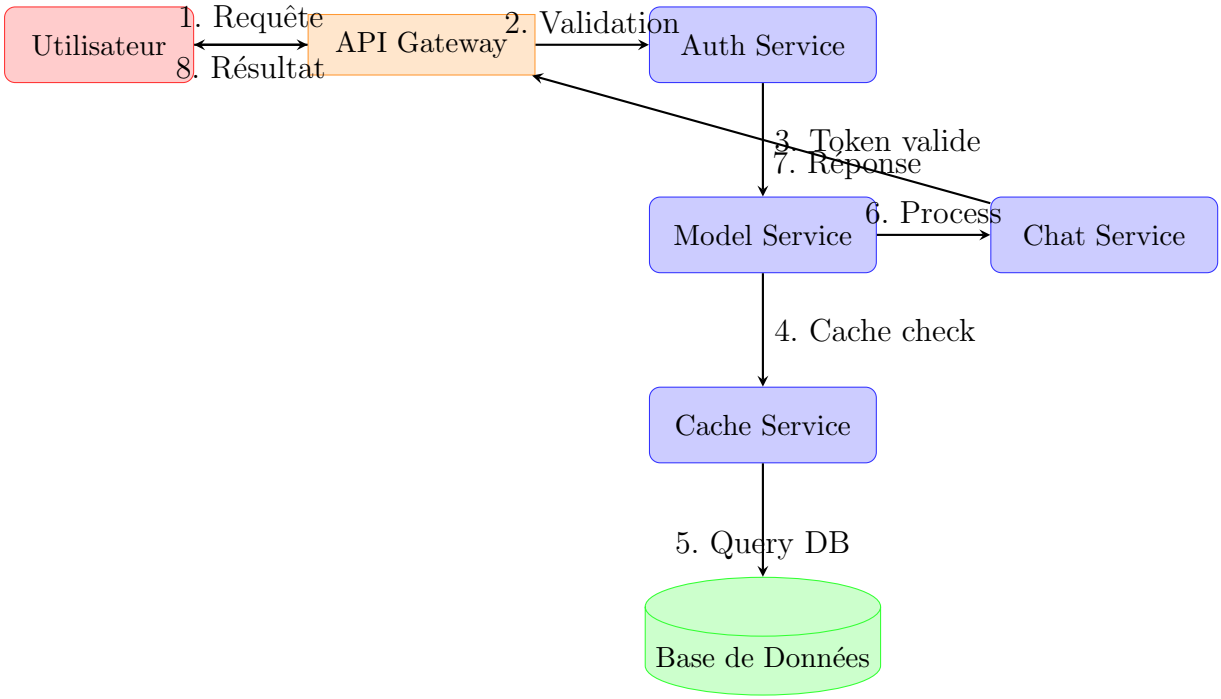

FIGURE 2 – Flux de traitement d'une requête IA

6 Patterns et Technologies

6.1 Patterns Architecturaux Utilisés

— **Circuit Breaker** : Protection contre les défaillances en cascade
— **Bulkhead** : Isolation des ressources critiques
— **Saga** : Gestion des transactions distribuées
— **CQRS** : Séparation commande/requête pour les opérations complexes
— **Event Sourcing** : Traçabilité complète des événements

6.2 Stack Technologique

Composant	Technologie
Orchestration	Kubernetes
Service Mesh	Istio
API Gateway	Kong / Envoy
Messagerie	Apache Kafka
Cache	Redis Cluster
Base de données	PostgreSQL / MongoDB
Monitoring	Prometheus + Grafana
Logging	ELK Stack
CI/CD	GitLab CI / Jenkins

TABLE 1 – Stack technologique recommandée

7 Scalabilité et Performance

7.1 Stratégies de Mise à l'Échelle

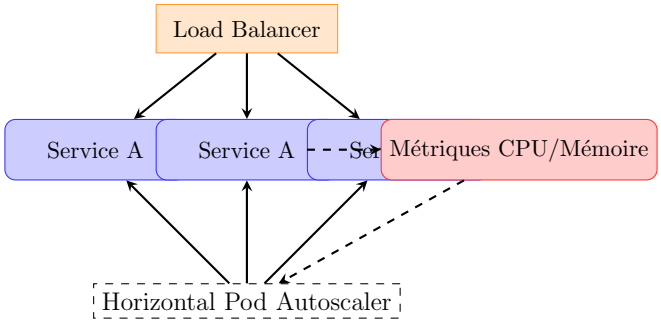

FIGURE 3 – Stratégie de mise à l'échelle horizontale automatique

8 Sécurité

8.1 Mesures de Sécurité Implémentées

— **Authentification forte** : OAuth 2.0 + OpenID Connect
— **Chiffrement** : TLS 1.3 pour toutes les communications
— **Isolation réseau** : Segmentation par namespace Kubernetes

— **Secrets management** : Vault ou Kubernetes Secrets
— **Scanning de vulnérabilités** : Intégration continue des images
— **Politique de réseau** : Firewall applicatif (WAF)

9 Monitoring et Observabilité

9.1 Architecture de Monitoring

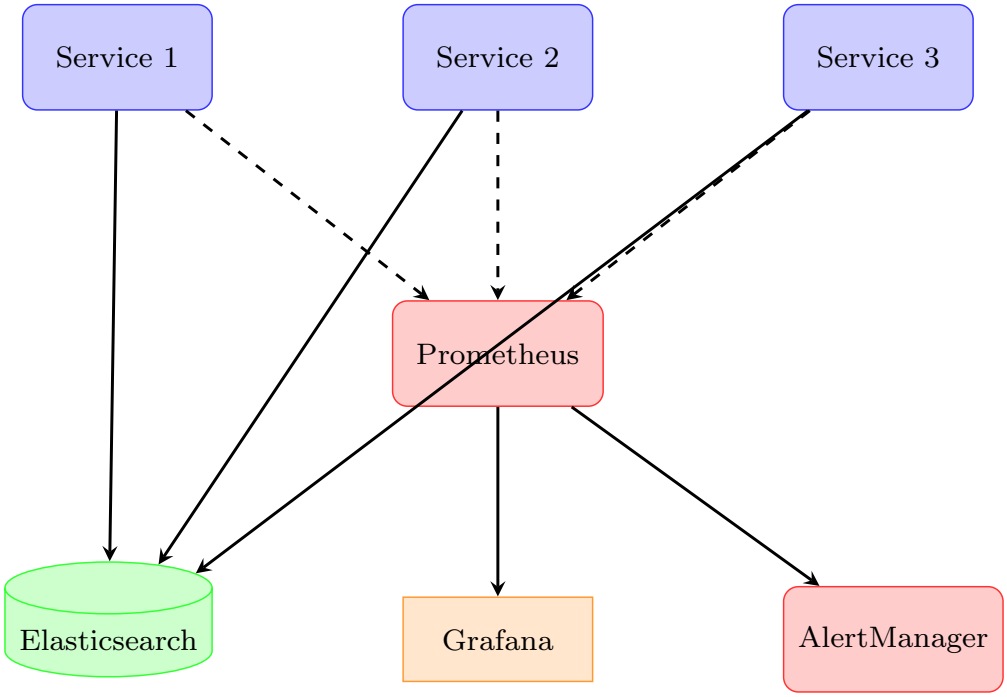

FIGURE 4 – Architecture de monitoring et observabilité

10 Plan de Déploiement

10.1 Phases de Déploiement

1. **Phase 1** : Infrastructure de base (Kubernetes, Service Mesh)
2. **Phase 2** : Services fondamentaux (Auth, API Gateway)
3. **Phase 3** : Services métier (Model, Chat)
4. **Phase 4** : Services de support (Cache, Queue, Monitoring)
5. **Phase 5** : Optimisation et mise à l'échelle

11 Conclusion

Cette architecture microservices pour DeepSeek offre une solution robuste, scalable et maintenir pour le traitement distribué de l'intelligence artificielle. L'approche modulaire permet une évolution indépendante de chaque composant tout en maintenant la cohérence globale du système.

Les patterns architecturaux implémentés assurent la résilience et la performance, tandis que l'utilisation de technologies cloud-native garantit une intégration optimale avec les environnements modernes de déploiement.