PRÉSENTATION

My Content

MVP SOLUTION DE RECOMMANDATION DE CONTENU



Problématique : Face à la surcharge informationnelle, comment aider les utilisateurs à découvrir des articles et livres qui correspondent réellement à leurs centres d'intérêt pour favoriser l'engagement et la lecture ?

Grégoire Mureau

Septembre 2025

Sommaire



01 Contexte et Analyse:

- Problématique My Content
- Défis identifiés
- Dataset



02 Modélisation et Algortihmes:

- Content-Based Filtering
- Collaborative Filtering
- Solution MVP retenue



03 Architecture et Déploiement:

- Cahier des charges d’une solution serverless:
  - fonction azure
  - interface de test
- Architecture MVP
- Démonstration



04 Évolution et scalabilité:

- Architecture cible
- Roadmap et perspectives

CONTEXTE ET ANALYSE



- My Consent : Startup dédiée à l'encouragement de la lecture

- Objectif : créer une solution de recommandation personnalisée

- User story : 5 articles personnalisés par utilisateur



- Absence de Données utilisateurs initiales

- Prise en charge des évolutions de la solution et de la base d'utilisateurs

- Contraintes techniques :
  - architecture serverless (Azure Fonctions)
  - quel algorithme de recommandation choisir ?
  - application minimale viable (MVP)

01: Contexte et Analyse - Dataset

Fichier	Description	Contenu	Dimensions
clicks.zip	Interactions complètes	Sessions utilisateurs (fichiers par heure)	~3 millions de clics
articles_metadata.csv	Métadonnées articles	ID, catégorie, éditeur, nb mots	364,047 articles (catalogue complet)
articles_embeddings.pickle	Embeddings CHAMELEON*	Vecteurs 250D pré-calculés	364,047 × 250D

Statistiques Dataset Global	Valeur
Sessions utilisateur	~1 million
Clics totaux	~3 millions
Articles avec interactions	46,033 (sur 364K du catalogue)
Couverture temporelle	2006 - 2018 (12 ans)

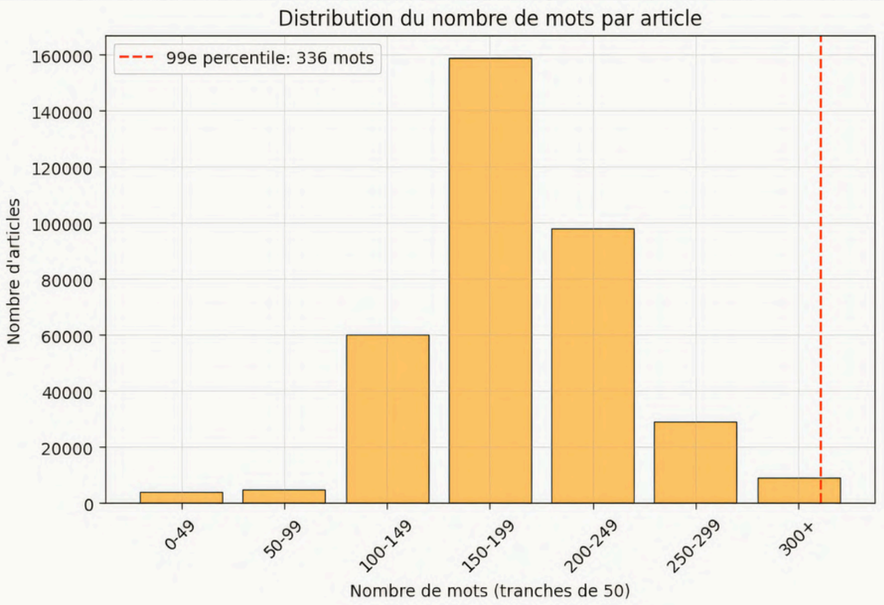
Dataset de base : globo.com (Kaggle)

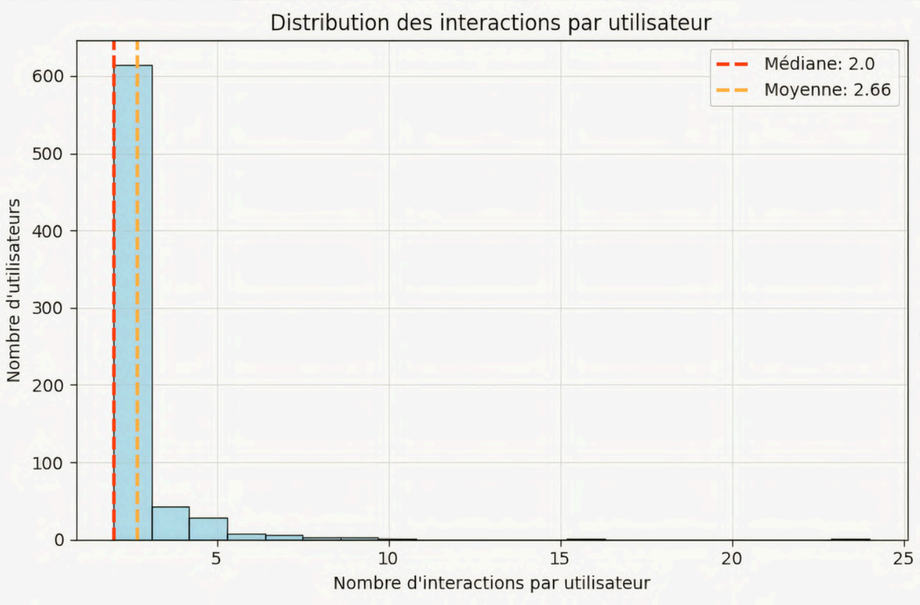
Dataset réduit MVP :

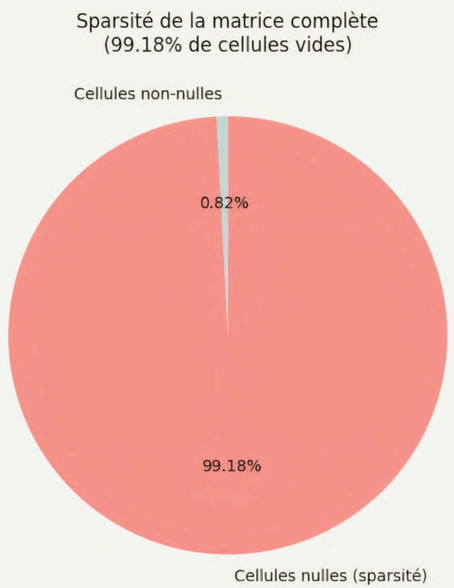
Composant	Valeur MVP	Restriction vs Global
Fichier interactions	**clicks_sample.csv**	**Échantillon 48h vs dataset complet**
Fichier embeddings	**articles_embeddings.pickle**	**Utilisé intégralement (364K articles)**
Utilisateurs uniques	707	vs ~millions dans le global
Articles cliqués	323	vs 46,033 dans le global
Interactions totales	1,883	vs 3 millions dans le global
Période couverte	47h35min (1-3 oct 2017)	Échantillon temporel réduit
Sparsité matrice	99.18%	Très élevée (données limitées)

01: Contexte et Analyse - Dataset réduit

- Articles courts
- Peu de clics par utilisateurs (habitudes dures à définir)
- Majorité d'articles non lus = sparsité extrême



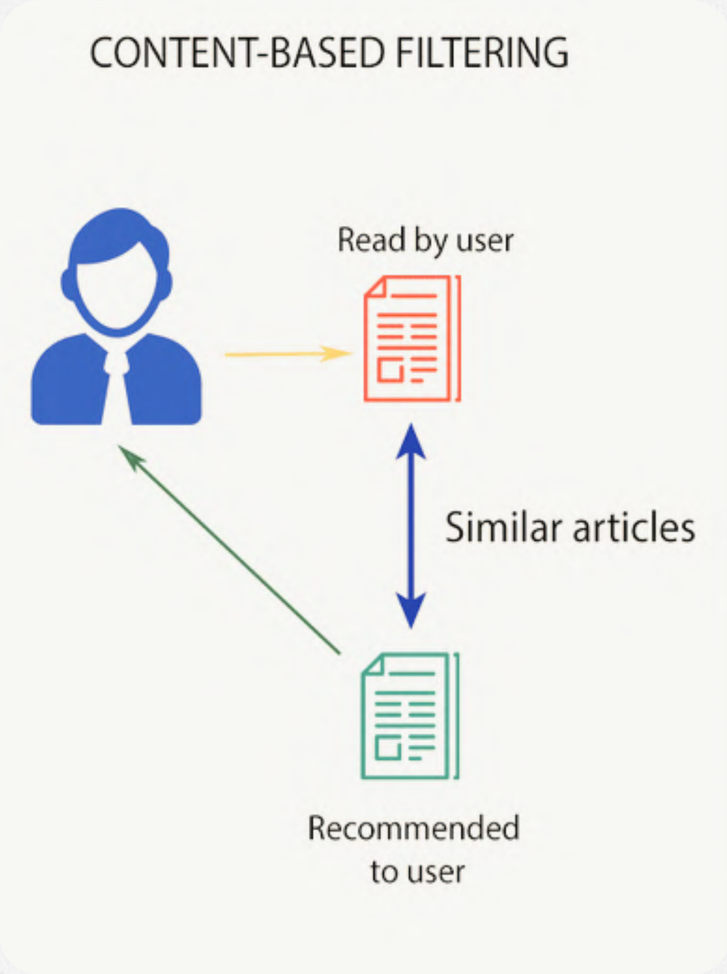




MODÉLISATION ET ALGORITHMES



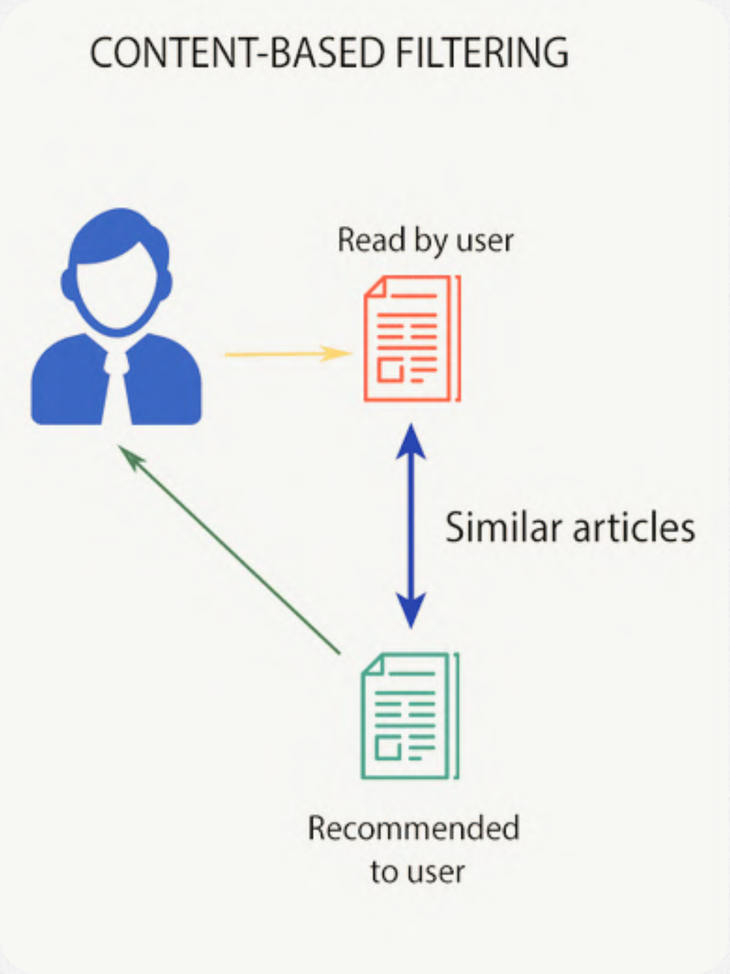
02: Modélisation et Algorithmes - Content-Based filtering



- Entrées:
  - Historique utilisateur
  - Représentations vectorielles articles

- Traitement:
  - Calcul du profil utilisateur (reference embedding : last, mean, random)
  - Calcul de similarité cosinus avec corpus complet
  - Filtrage : exclusion articles déjà consultés
  - Classement : tri par ordre décroissant

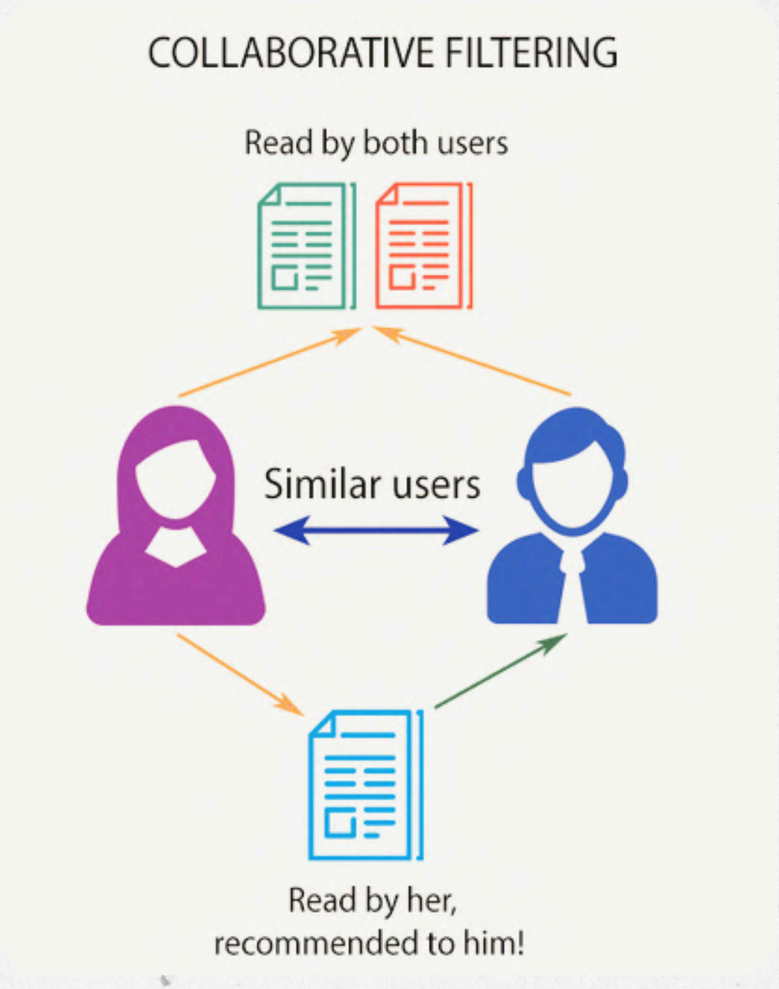
02: Modélisation et Algorithmes - Content-Based filtering



- Avantages:
  - Pas de cold start articles
  - Indépendance utilisateurs
  - Transparence
  - ~~Biais de popularité~~

- Inconvénients:
  - cold start utilisateurs
  - dépendance embeddings (actualisation)
  - bulle de filtres
  - surspécialisation

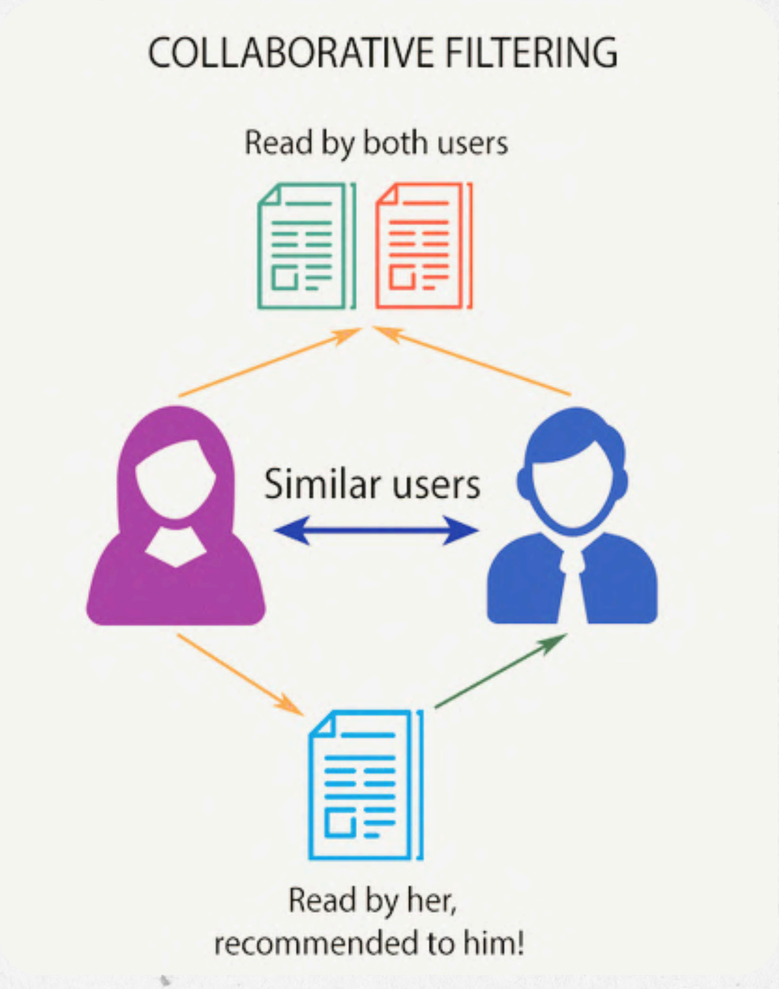
02: Modélisation et Algorithmes - Collaborative filtering



- Entrées:
  - Matrice utilisateur-article
  - Utilisateur cible

- Traitement SVD (machine learning):
  - Factorisation matricielle (via Scikit-Surprise)
  - Décomposition en facteurs latents (utilisateurs × facteurs cachés)
  - Apprentissage des patterns d'interaction cachés
  - Prédiction des ratings manquants par reconstruction matricielle
  - Filtrage : exclusion articles déjà consultés
  - Classement : tri par scores prédits décroissants

02: Modélisation et Algorithmes - Collaborative filtering



- Avantages:
  - exploitation des patterns comportementaux “cachés”
  - sérendipité: découvrir par hasard contenu inattendu
  - amélioration avec le temps
  - Indépendance du contenu
- Inconvénients:
  - explicabilité faible (facteurs latents abstraits)
  - biais de popularité : articles globalement les plus cliqués
  - cold start users
  - cold start articles
  - dépendance à la sparsité des données

02: Modélisation et Algorithmes - Solution MVP retenue

Catégorie	Critères d'Évaluation	Algorithmes de Recommandation	
		Content-Based Filtering	Collaborative Filtering (SVD)
Contraintes Dataset MVP		**Sparsité 99.18% • 707 utilisateurs • 1,883 interactions • Période 48h**	
Robustesse MVP	Sparsité des données (99.18%)	✓ Fonctionne indépendamment	✗ Performance très dégradée
	Nouveaux articles	✓ Immédiatement recommandables	✗ Jamais recommandés
	Échantillon réduit (48h)	✓ Utilise les embeddings complets	✗ Patterns insuffisants
Ressources	Données disponibles	✓ Embeddings CHAMELEON prêts	Matrice user-item à construire
	Explicabilité	✓ "Similaire à vos lectures"	✗ Facteurs latents abstraits
DÉCISION MVP		**✓ SÉLECTIONNÉ** **Optimal pour contraintes MVP**	**PLANIFIÉ V2** **Avec accumulation données**

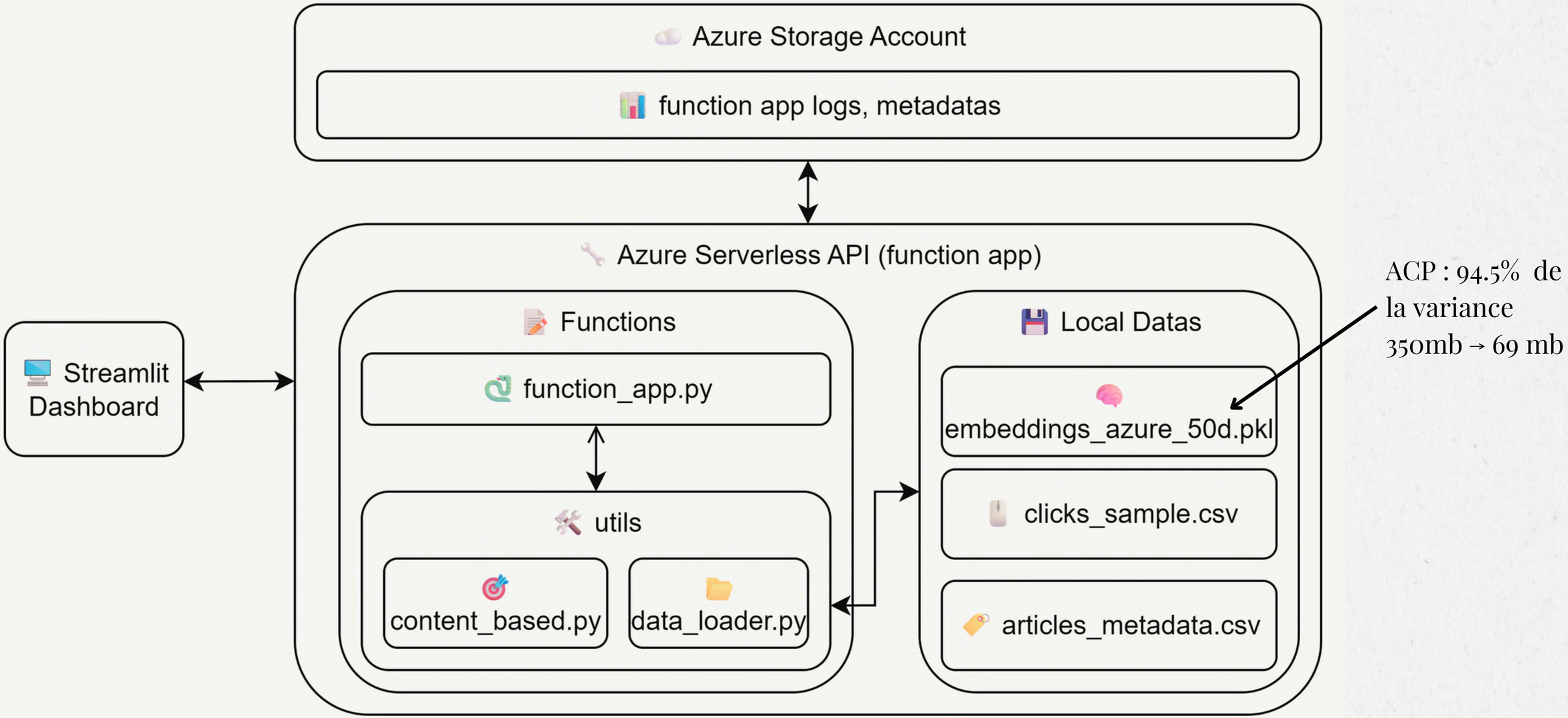
ARCHITECTURE ET DÉPLOIEMENT



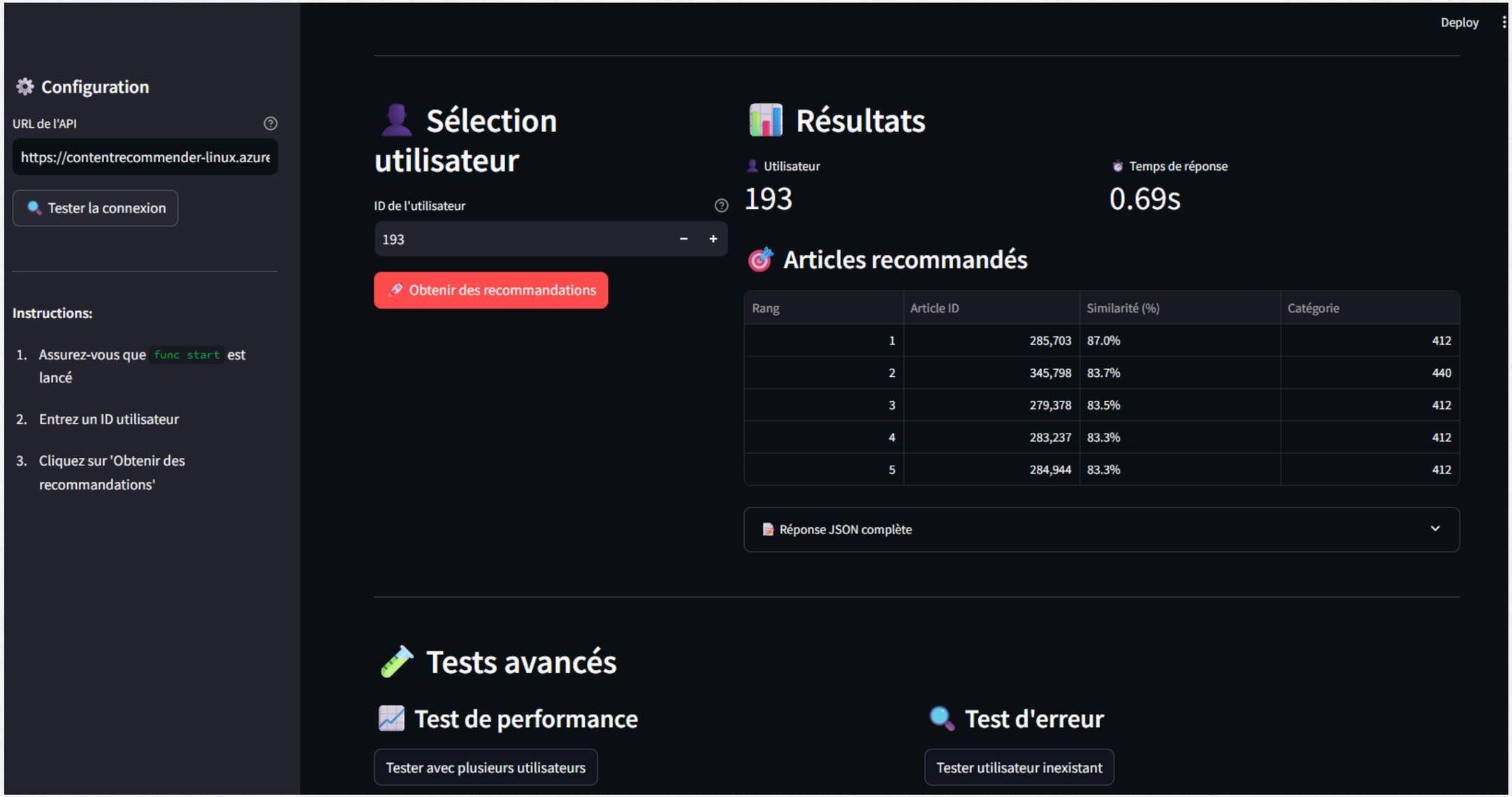
03: Architecture et déploiement - Cahier des charges

- Fonctionnalité critique attendue : suggestion de 5 articles
- Contraintes techniques imposées:
  - Architecture servereless obligatoire : Azure Functions
  - Données externes : Utilisation dataset Kaggle (globo.com), travail sur sous échantillon (MVP)
  - Plan gratuit : contrainte de ressource Azure
- Livrables MVP :
  - API fonctionnelle: Azure Function retournant 5 articles
  - Interface de test : Application Streamlit pour validation
  - Code déployé : Solution en production Azure
  - Démonstration : preuve de concept opérationnelle

03: Architecture et déploiement - Architecture MVP Technique



03: Architecture et déploiement - Démonstration



EVOLUTION ET SCALABILITÉ



04: Évolution et scalabilité - Architecture Cible

Base Technique

Stockage persistant

Base de données SQL

Données réelles

Clicks, temps d'attention, interactions My Content

Catalogue étendu

Articles • Livres • Contenus éditoriaux de qualité

Solution Algorithme Hybride

Content-Based

Base robuste (comme MVP)

Collaborative Filtering

Découverte avec données enrichies

Pondération intelligente

Équilibrage selon contexte utilisateur

Gestion Cold Start

Nouveaux utilisateurs :

Onboarding questionnaire → catégories/profils similaires • Articles populaires par défaut • Apprentissage progressif

Nouveaux articles/livres :

Pipeline automatisé embeddings • Recommandation immédiate par similarité • Intégration catalogue temps réel

Infrastructure Production

Base données persistante

Profils + historiques + catalogue

Monitoring qualité

Métriques engagement utilisateur

Maintenance embeddings

Adaptation vocabulaire émergent

04: Évolution et scalabilité - Roadmap

- **Phase 1 (3-6 mois) - Fondations**
  - **Embeddings adaptés** : acquisition corpus littéraire (Project Gutenberg, OpenLibrary, corpus académiques)
  - **Migration base de données** : stockage persistant profils et catalogue
  - **Pipeline nouveaux articles** : automatisation calculs embeddings
- **Phase 2 (6-12 mois) - Enrichissement**
  - **Collaborative Filtering basique** : SVD avec données clics/interactions
  - **Onboarding hybride (nouveaux utilisateurs)** : Exemples de lectures passées ↠ profil Content-Based immédiat + fallback par centres d'intérêt
  - **Collecte métriques engagement** : Temps lecture, completion rate
- **Phase 3 (12+ mois) - Optimisation**
  - **Collaborative Filtering avancé** : Intégration métriques engagement enrichies
  - **Algorithme hybride** : Pondération intelligente Content-Based + Collaborative
  - **Expansion catalogue** : Livres + contenus éditoriaux diversifiés
  - (Maintenance embeddings : Adaptation éventuel vocabulaire émergent)

CONCLUSION



Conclusion et Perspectives

- **Objectifs MVP Atteints**
  - **Système fonctionnel :** API Azure Functions + Interface Streamlit opérationnelles
  - **Choix technique justifié :** Content-Based optimal pour contraintes données (sparsité 99.18%)
  - **Architecture évolutive :** Base solide pour transition vers production My Content
- **Apports Techniques**
  - **Optimisation ACP** : Réduction 250D ↠ 50D (94.5% variance conservée)
  - **Solution serverless** : Déploiement cloud auto-scalable
  - **Approche comparative** : Content-Based vs Collaborative Filtering analysés
- **Vision My Content**
  - **Différenciation métier :** Encouragement lecture vs consommation rapide d'actualités
  - **Roadmap réaliste :** 18 mois pour passage MVP ↠ Production complète
  - **Innovation algorithme :** Pipeline nouveaux articles + onboarding hybride
- **Prochaines Étapes Critiques**
  - **Phase 1 prioritaire** : Acquisition embeddings littéraires (Project Gutenberg)
  - **Validation utilisateur** : Tests avec premiers utilisateurs My Content
  - **Collecte données** : Métriques engagement adaptées à la lecture
- My Content : Transformer la découverte de contenu en parcours de lecture enrichissant

PRÉSENTATION

My Content

MERCI · MERCI · MERCI · MERCI · MERCI · MERCI · MERCI ·

Avez-vous des questions ?
N’hésitez pas !



Grégoire Mureau

Septembre 2025