Databricks, Azure AI Search et Snowflake ne sont pas des frameworks à proprement parler, mais plutôt des plateformes ou des services qui intègrent des capacités pour faciliter l'implémentation du RAG. Cependant, ils offrent des fonctionnalités et des outils qui encapsulent les différentes étapes du RAG.

Voici comment chacun d'eux intègre le RAG :

Databricks

Databricks propose une approche complète pour le RAG via sa plateforme Data Intelligence, notamment avec le **Mosaic AI Agent Framework**.

- **Mosaic AI Agent Framework** : C'est une suite d'outils conçue pour aider les développeurs à construire et déployer des applications d'IA générative de haute qualité en utilisant le RAG. Il simplifie l'évaluation de la qualité des applications RAG, permet une itération rapide et assure la gouvernance et les garde-fous.
    - **Vector Search** : Databricks propose un service de recherche vectorielle évolutif (Databricks Vector Search) qui automatise l'indexation et la récupération des données, simplifiant ainsi le workflow RAG traditionnellement complexe.
    - **Data Pipelines (Delta Lake, Lakeflow Declarative Pipelines)** : Pour l'ingestion et la préparation des données, essentielles au RAG.
    - **Mosaic AI Gateway** : Fournit un accès transparent à divers modèles (Meta Llama, Amazon Bedrock, Hugging Face) via une API unique, facilitant l'expérimentation et le changement de modèle.
    - **Mosaic AI Agent Evaluation** : Permet d'évaluer la qualité des réponses générées par les applications RAG, avec des vérifications basées sur des règles, des juges LLM et des retours humains.
    - **Intégration avec des outils tiers** : Databricks peut s'intégrer avec des outils comme Tonic Textual pour la préparation de données non structurées (parsing, chunking, enrichment de métadonnées).

Databricks fournit une plateforme unifiée pour toutes les étapes du RAG, de l'ingestion des données à l'évaluation du modèle, avec un accent sur la production et la gouvernance.

Azure AI Search

Azure AI Search est un service de recherche cloud qui est un composant clé pour construire des solutions RAG sur Azure.

- **Capacités de recherche hybride** : Azure AI Search combine la recherche vectorielle (avec des algorithmes comme HNSW ou brute-force KNN) avec la recherche en texte intégral, ce qui est crucial pour un RAG performant. Il supporte également la recherche floue, géospatiale et les filtres.
- **Indexation et vectorisation intégrées** : Il permet de créer des pipelines d'indexation qui peuvent charger, découper (chunk), intégrer (embed) et ingérer du contenu. La vectorisation peut être effectuée à l'ingestion ou à la requête.
- **Intégration avec Azure OpenAI et Azure AI Foundry** : Azure AI Search est conçu pour fonctionner de manière transparente avec les modèles d'embedding (pour la vectorisation) et les modèles de complétion de chat (pour la génération des réponses) d'Azure OpenAI Service ou d'Azure AI Foundry.
- **Optimisation de la pertinence** : Il offre des fonctionnalités pour maximiser la pertinence des résultats de recherche avant de les passer au LLM, ce qui est essentiel pour minimiser les "hallucinations".
- **Orchestrateurs** : Bien qu'Azure AI Search ne soit pas un orchestrateur en soi, il est souvent utilisé en combinaison avec des frameworks comme Semantic Kernel, Azure AI Agent service ou LangChain, qui se chargent de l'orchestration globale de la solution RAG.

Azure AI Search se positionne comme le moteur de recherche et de récupération d'informations essentiel pour les solutions RAG sur Azure, fournissant une base solide pour la gestion des données et la pertinence de la recherche.

Snowflake

Snowflake a également renforcé ses capacités pour le RAG, en se concentrant sur la simplicité et la sécurité au sein de sa plateforme.

- **Snowflake Cortex** : C'est le service d'IA de Snowflake qui inclut des fonctions LLM serverless pour l'embedding et la complétion de texte. Il permet aux équipes d'intégrer des questions et du contexte pour trouver les informations les plus pertinentes et générer des réponses contextualisées.
    - **Cortex Search Service** : Un service d'indexation et de récupération entièrement géré qui simplifie les applications RAG en créant automatiquement des embeddings et des index, et en fournissant un service de recherche d'entreprise accessible via des API.

- **Support natif du type de données VECTOR** : Snowflake prend en charge le type de données VECTOR nativement, éliminant ainsi le besoin d'intégrer et de gérer un magasin ou un service distinct pour les vecteurs.
- **Snowpark User Defined Functions (UDFs)** : Permet d'intégrer des frameworks comme LangChain pour des tâches spécifiques comme le "chunking" (découpage) des documents, directement dans l'environnement Snowflake.
- **Fonctions de traitement de documents** : Snowflake Cortex propose des fonctions comme PARSE_DOCUMENT pour lire les documents PDF directement à partir des zones de transit et SPLIT_TEXT_RECURSIVE_CHARACTER pour découper le texte en chaînes plus courtes, simplifiant la phase de préparation des données pour le RAG.
- **Streamlit in Snowflake** : Permet de construire des interfaces utilisateur (UI) pour les applications RAG directement dans Snowflake, assurant la sécurité des données et les politiques d'accès.

Snowflake vise à fournir une solution RAG complète et sécurisée, sans nécessiter d'intégrations complexes, de gestion d'infrastructure ou de déplacement de données, en tirant parti de ses capacités de gouvernance des données et de son moteur de calcul.