Journal of Information Systems Engineering and Management
2024, 9(4s)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Federated RAG Architectures for Distributed Enterprise Knowledge Systems

Mahesh Kumar Gaddam

Federal Express Corporation

ARTICLE INFO

Received: 02 Nov 2024

Accepted: 28 Dec 2024

ABSTRACT

Retrieval-augmented generation (RAG) has become a practical pattern for grounding large language models on enterprise data, but most production deployments still assume a largely centralized corpus, uniform access model, and single retrieval stack. That assumption breaks down in real enterprises, where knowledge is fragmented across document repositories, search appliances, data lakes, knowledge graphs, SaaS systems, and regional boundaries. This paper defines federated RAG as an architectural synthesis of RAG, federated search, federated query processing, enterprise knowledge graphs, and policy-aware access control. The goal is not merely to answer questions from more sources, but to do so while preserving source autonomy, respecting fine-grained authorization, containing latency, and maintaining provenance. Building from work on RAG, dense retrieval, late interaction, federated search, query federation, enterprise knowledge graphs, and attribute-based access control, the paper proposes a brokered multi-stage architecture for distributed enterprise knowledge systems. It argues that the strongest enterprise designs are neither dense-only nor centralized-only. Instead, they combine source selection, hybrid sparse-dense retrieval, hierarchical re-ranking, policy enforcement, and evidence fusion. A comparative analysis shows that federated RAG is most valuable when organizations face data sovereignty, business-unit autonomy, heterogeneous schemas, and rapidly changing private corpora. The paper concludes with design patterns, evaluation criteria, and practical trade-offs for deploying fail-safe, cost-aware, and governance-ready enterprise assistants.

Keywords: retrieval-augmented generation, federated search, enterprise knowledge systems, distributed retrieval, hybrid retrieval, knowledge graphs, access control

1. Introduction

Enterprise knowledge systems have moved from static search portals to conversational assistants, but the underlying knowledge problem remains distributed. Business documents live in content management systems, tickets in service platforms, policies in collaboration tools, metadata in operational databases, and structured semantics in knowledge graphs. Classical federated search literature describes this problem as one of aggregating multiple searchable sources through collection description, source selection, results merging, and presentation. Modern federated query processing extends the same idea to heterogeneous and autonomous sources, where the

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

system must identify relevant sources and build efficient execution plans rather than simply fan out every query. RAG adds a generative layer on top of retrieval, but by itself it does not solve the enterprise realities of fragmented ownership, heterogeneous schemas, or access-control asymmetry.

The term **federated RAG** is therefore used here as a synthesized architectural concept rather than a single canonical model from the 2014-2023 literature. In this paper, it means a RAG system in which retrieval is brokered across multiple autonomous enterprise sources, grounded evidence is normalized and policy-checked, and only then supplied to the generator. This framing matters because centralized indexing is often operationally attractive yet legally, organizationally, or economically impossible. Multinational enterprises, regulated sectors, and post-merger organizations often need source autonomy and selective exposure, not a full raw-data consolidation layer.

Table 1. Drivers for federated RAG in enterprises

Enterprise condition	Why centralized RAG struggles	Why federation helps
Autonomous business units	Conflicting schema, ownership, and update cycles	Lets each unit keep its own index and governance
Data sovereignty or residency rules	Full replication may violate policy	Keeps data local while exposing controlled retrieval
Mixed unstructured and structured knowledge	One index often under-serves graphs, tables, and documents	Broker can route by source type
High corpus churn	Central pipelines lag or become costly	Incremental per-source indexing reduces blast radius
Fine-grained authorization	Central stores often flatten permissions	Source-aware enforcement preserves native controls

2. Architectural Foundations

RAG, as introduced by Lewis et al., couples a generator with non-parametric memory accessed through retrieval, explicitly addressing the limits of purely parametric knowledge and improving provenance and updateability. That makes it naturally attractive for enterprise settings, where factual freshness and traceability matter. Yet enterprise deployment quality is shaped less by generation alone and more by the design of the retrieval layer: corpus organization, candidate recall, ranking latency, index size, and evidence fusion. Dense Passage Retrieval showed that dense dual-encoder retrieval can materially outperform strong sparse baselines on open-domain QA, with reported gains of 9 to 19 percentage points in top-20 passage retrieval accuracy over a Lucene-BM25 baseline. ColBERT then demonstrated that late interaction can recover stronger token-level matching at far lower query cost than full cross-encoders, making it appealing for high-quality second-stage ranking.

The federated side of the problem has its own mature foundations. Han et al. characterize distributed information retrieval as a pipeline of resource description, selection, merging, and presentation. Garba et al. later review federated search techniques as a still-active area spanning resource selection, rank fusion, evaluation, and personalization. Endris et al. show that federated query processing depends on source descriptions both to identify relevant sources and to construct execution plans that balance completeness and execution cost. These ideas map directly into enterprise RAG: a broker needs source descriptors, a routing policy, a query plan, and a principled merge strategy before any large language model can safely answer.

A third foundation is enterprise semantics. Reinanda et al. show that knowledge graphs can support information retrieval in multiple ways, whether public or proprietary. Duan and Xiao argue that enterprise knowledge graphs have value beyond narrow business tasks and can evolve into broader enterprise knowledge management assets. In federated RAG, this means graphs should not be seen as competing with document retrieval. Instead, they serve

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

as schema alignment, entity resolution, relationship expansion, and provenance infrastructure across otherwise disconnected repositories.

Table 2. Foundational building blocks behind federated RAG

Building block	Core idea	Relevance to federated RAG
RAG	Combine generation with retrieved external evidence	Grounds answers on current enterprise content
Distributed IR / federated search	Select sources, retrieve locally, merge globally	Enables multi-repository retrieval without full consolidation
Federated query processing	Optimize subqueries over heterogeneous sources	Reduces unnecessary fan-out and latency
Dense retrieval	Semantic recall over paraphrased or weakly lexical queries	Helps retrieve meaning, not just keywords
Late interaction / hierarchical retrieval	Improve ranking quality without full cross-encoder cost	Useful for multi-stage broker pipelines
Enterprise KG	Normalize entities, relations, and metadata across silos	Supports routing, fusion, and Explainability
ABAC	Evaluate attributes of user, object, action, environment	Enforces policy before evidence reaches the generator

3. Reference Architecture for Federated RAG

A practical federated RAG architecture has six layers. First, a **policy-aware query broker** receives the user request, identity, session context, and task objective. Second, a **source routing layer** uses source descriptions, metadata, and learned heuristics to select candidate repositories. Third, each repository performs **local retrieval** using its preferred native method, such as BM25, dense vector search, SQL, SPARQL, or graph traversal. Fourth, a **global evidence fusion layer** normalizes scores, de-duplicates semantically overlapping evidence, and attaches provenance and policy labels. Fifth, a **generator and answer controller** composes the answer only from authorized evidence. Sixth, an **observability loop** records source selection decision, retrieval traces, and answer grounding for evaluation. This architecture is the natural composition of RAG with federated search phases and federated query-planning principles.

The most important design decision is whether the broker retrieves from every source or first performs source selection. In enterprises, selective routing is usually superior. Source selection reduces network calls, avoids policy violations from unnecessary probing, and narrows later-stage ranking cost. Han et al.'s work on collection selection remains relevant here because the central problem is unchanged: predicting which sources are likely to contain relevant answers. In modern federated RAG, that prediction can be informed not only by query terms but also by entities, source freshness, document type, jurisdiction, and user entitlements.

Another critical decision is evidence representation. Simple document chunks are often insufficient in enterprise settings because the same concept may appear as a policy PDF, a service ticket, a table row, and a graph relation. Knowledge graphs help unify these fragments through entity identity and relationship structure, while documents preserve the full evidential wording. A strong federated design therefore uses a **dual evidence model**: graph metadata for routing and linking, documents for quoted grounding, and structured results for exact values. This is more robust than forcing everything into a single vector index.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

From an operational standpoint, federation also improves fault isolation. A broken connector, stale index, or region-specific outage does not need to collapse the whole assistant. The broker can degrade gracefully by excluding failed sources, lowering confidence, or returning a scoped answer with explicit provenance. That makes federated RAG not just a retrieval strategy but a resilience strategy. The price is higher control-plane complexity: more routing logic, more score calibration, and more observability requirements.

4. Retrieval and Ranking Strategies: Comparative Analysis

The retrieval layer is where most federated RAG systems succeed or fail. Dense-only pipelines are good at semantic matching, but sparse retrieval remains valuable for exact identifiers, policy codes, part numbers, and legal phrases. This is especially true in enterprises, where users ask for "the Q4 India travel policy version," ticket IDs, or product SKUs that should not be semantically paraphrased away. DPR established the strength of dense retrieval for semantic recall, RocketQA improved dense retriever training with cross-batch negatives, denoised hard negatives, and augmentation, and ART later showed that strong dense retrievers can also be trained without labeled data. At the same time, hybrid evidence remains practically necessary because lexical precision and semantic generalization solve different failure modes.

The literature also supports multi-stage ranking rather than single-shot retrieval. ColBERT's late interaction offers stronger token-level matching than plain bi-encoders with manageable cost. HHR shows that combining sparse and dense retrieval in hierarchical stages improves robustness and zero-shot generalization, reporting an average 4.69% improvement in recall@100 over dense hierarchical retrieval on zero-shot TriviaQA and Web Questions. This is highly relevant to enterprise assistants, where out-of-domain drift is constant because new business terms, products, and project names appear faster than retrievers can be fully retrained.

Cost and memory are equally important. Yamada et al.'s Binary Passage Retriever reduces passage-index memory from 65 GB to 2 GB without losing accuracy on standard QA benchmarks, showing that retrieval quality and infrastructure efficiency can be co-optimized. That matters in federated deployments because each source may maintain its own index footprint. Similarly, Zhou et al.'s hyperlink-induced pre-training improves passage retrieval under zero-shot and few-shot conditions, which suggests a broader lesson: federation benefits when local retrievers are pretrained from structure already present in enterprise systems, such as links, tickets, citations, or workflow edges.

A more enterprise-facing signal comes from industrial FAQ retrieval. Seo et al. show that a hybrid dense-plus-sparse retriever for industrial FAQs outperforms single-retriever settings, with hybrid configurations reaching at least 0.8 Hit@1 while single models remained at or below about 0.65 in their comparison. Although that study is narrower than enterprise knowledge systems, it supports a general design principle: hybrid retrieval is usually the most reliable default for operational domains containing both semantic variation and critical exact-match terminology.

Table 4. Comparative analysis of retrieval options for federated RAG

Retrieval design	Strengths	Weaknesses	Best enterprise use
Sparse only	Excellent exact matching, cheap, interpretable	Weak paraphrase handling	Policies, IDs, product codes, legal clauses
Dense bi-encoder	Strong semantic recall, fast ANN retrieval	Misses exact lexical constraints, score drift across sources	General document search and FAQ retrieval
Late interaction	Better ranking quality than bi-encoders	Higher storage and compute cost	High-value second-stage ranking
Hybrid sparse+dense	Balances lexical precision and semantic recall	Requires score calibration	Best default for heterogeneous enterprise corpora

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Retrieval design	Strengths	Weaknesses	Best enterprise use
Hierarchical hybrid	Better zero-shot recall and controllable latency	Highest orchestration complexity	Large federated deployments with many sources
Graph-augmented retrieval	Strong multi-hop reasoning and entity linking	Graph upkeep cost, schema work	Expert finding, dependency analysis, policy lineage

A useful rule follows from this comparison: in distributed enterprise knowledge systems, **federation should happen before final ranking, but not before coarse filtering**. In other words, first select sources, then do local hybrid retrieval, then global fusion, then a stronger cross-source rerank. That ordering keeps cost under control while preserving answer quality. Dense-only global retrieval over all sources sounds elegant, but in practice it wastes compute, weakens policy boundaries, and amplifies score incomparability across repositories.

5. Security, Governance, and Evaluation

In enterprise assistants, the hardest problem is often not relevance but **authorization**. RAG systems are unsafe if they retrieve evidence the user is not entitled to see, even if the final text never quotes it directly. ABAC is important here because it evaluates subject, object, operation, and environment attributes rather than relying only on static roles. NIST's guide explicitly frames ABAC as a way to support enterprise information sharing while maintaining protection requirements. For federated RAG, that implies policy checks must be applied at retrieval time and evidence-fusion time, not only as an after-the-fact output filter.

Governance also requires provenance. The generator should receive evidence with source identifiers, timestamps, policy tags, and confidence signals. This allows the system to abstain when sources disagree, explain where claims came from, and support audit trails. Knowledge graphs can help here by storing entity lineage and source relationships, while federated brokers can record which repositories were queried and why. Without provenance, the enterprise cannot distinguish a grounded answer from a plausible synthesis.

Evaluation must therefore go beyond answer correctness. Classical federated search work contributed benchmarks such as FedWeb Greatest Hits for studying source selection and merging. Modern enterprise RAG needs a broader scorecard: source-selection precision, authorized recall, citation fidelity, latency percentiles, abstention quality, and cost per answer. A system that gives a correct answer from an unauthorized source is a failure, not a success. Similarly, a system that queries too many repositories may be accurate yet economically unfit for production.

Table 5. Governance and evaluation framework

Dimension	What to measure	Why it matters
Relevance	Recall@k, MRR, grounded answer accuracy	Captures retrieval and final utility
Federation quality	Source-selection precision, merge quality	Tests whether routing is efficient and sound
Security	Authorized recall, leakage rate, policy-denial correctness	Prevents unsafe evidence exposure
Provenance	Citation fidelity, source completeness, freshness	Supports trust and auditability
Operations	p50/p95 latency, connector failure tolerance, cache hit rate	Determines production viability

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(4s)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Dimension	What to measure	Why it matters
Economics	Cost per answer, tokens per answer, index footprint	Keeps enterprise deployment sustainable

6. Deployment Patterns and Trade-offs

Three deployment patterns stand out. The first is **brokered federation without replication**, used when governance or sovereignty dominates. The second is **metadata-central, content-local federation**, where lightweight source descriptors, embeddings, or graph metadata are centralized but documents remain local. The third is **selective replication**, where only a subset of low-risk or high-value corpora are centralized, while the rest stay federated. The second pattern is often the most balanced because it supports efficient routing without demanding full content movement.

The trade-off is simple: centralization tends to optimize latency and ranking consistency, while federation optimizes autonomy, compliance, and resilience. Enterprises should therefore choose architecture by constraint, not fashion. If the main problem is pure search speed over homogeneous documents, centralized RAG may be sufficient. If the problem is cross-unit knowledge access under heterogeneous ownership, federated RAG is the more realistic design. If the problem requires multi-hop reasoning over people, assets, policies, and incidents, a graph-augmented federated variant is likely best.

7. Conclusion

Federated RAG is best understood as the convergence of four mature ideas: externalized memory for language models, federated retrieval, query planning over heterogeneous sources, and enterprise-grade governance. The literature from 2014 to 2023 already provides most of the ingredients. What changes in the enterprise setting is the objective function. The target is not only relevance, but relevance under autonomy, authorization, provenance, latency, and cost constraints. Under those conditions, the most effective architectures are policy-aware, source-selective, hybrid in retrieval, hierarchical in ranking, and explicit about provenance.

A final practical takeaway is that enterprises should not ask whether federated RAG is “better” than centralized RAG in the abstract. They should ask which constraints dominate their environment. Where data ownership is fragmented and governance is strict, federated RAG is not a luxury feature. It is the architecture that aligns the assistant with the enterprise itself.

References

[1] Hu, V. C., Ferraiolo, D., Kuhn, D. R., Schnitzer, A., Sandlin, K., Miller, R., and Scarfone, K. (2014). *Guide to Attribute Based Access Control (ABAC) Definition and Considerations*. NIST SP 800-162. DOI: **10.6028/NIST.SP.800-162**.

[2] Demeester, T., Trieschnigg, D., Nguyen, D., Hiemstra, D., and Zhou, K. (2015). *FedWeb Greatest Hits: Presenting the New Test Collection for Federated Web Search*. DOI: **10.1145/2740908.2742755**.

[3] Han, B., Chen, L., and Tian, X. (2018). *Knowledge based collection selection for distributed information retrieval*. *Information Processing & Management*, 54(1), 116-128. DOI: **10.1016/j.ipm.2017.10.002**.

[4] Duan, R., and Xiao, Y. (2019). *Enterprise Knowledge Graph From Specific Business Task to Enterprise Knowledge Management*. DOI: **10.1145/3357384.3360314**.

[5] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. DOI: **10.48550/arXiv.2005.11401**.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[6] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). *Dense Passage Retrieval for Open-Domain Question Answering*. DOI: **10.18653/v1/2020.emnlp-main.550**.

[7] Khattab, O., and Zaharia, M. (2020). *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT*. DOI: **10.1145/3397271.3401075**.

[8] Endris, K. M., Acosta, M., and Vidal, M.-E. (2020). *Federated Query Processing*. DOI: **10.1007/978-3-030-53199-7_5**.

[9] Reinanda, R., Meij, E., and de Rijke, M. (2020). *Knowledge Graphs: An Information Retrieval Perspective*. *Foundations and Trends in Information Retrieval*, 14(4), 289-444. DOI: **10.1561/1500000063**.

[10] Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., Dong, D., Wu, H., and Wang, H. (2021). *RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering*. DOI: **10.18653/v1/2021.naacl-main.466**.

[11] Yamada, I., Asai, A., and Hajishirzi, H. (2021). *Efficient Passage Retrieval with Hashing for Open-domain Question Answering*. DOI: **10.18653/v1/2021.acl-short.123**.

[12] Seo, J., Lee, T., Moon, H., Park, C., Eo, S., Aiyanyo, I. D., Park, K., So, A., Ahn, S., and Park, J. (2022). Dense-to-Question and Sparse-to-Answer: Hybrid Retriever System for Industrial Frequently Asked Questions. Mathematics, 10(8), 1335. DOI: 10.3390/math10081335.

[13] Zhou, J., Li, X., Shang, L., Luo, L., Zhan, K., Hu, E., Zhang, X., Jiang, H., Cao, Z., Yu, F., Jiang, X., Liu, Q., and Chen, L. (2022). Hyperlink-induced Pre-training for Passage Retrieval in Open-domain Question Answering. DOI: 10.18653/v1/2022.acl-long.493.

[14] Garba, A., Wu, S., and Khalid, S. (2023). Federated search techniques: an overview of the trends and state of the art. Knowledge and Information Systems, 65, 5065-5095. DOI: 10.1007/s10115-023-01922-6.

[15] Sachan, D. S., Lewis, M., Yogatama, D., Zettlemoyer, L., Pineau, J., and Zaheer, M. (2023). Questions Are All You Need to Train a Dense Passage Retriever. Transactions of the Association for Computational Linguistics, 11, 600-616. DOI: 10.1162/tacl_a_00564.

[16] Arivazhagan, M. G., Liu, L., Qi, P., Chen, X., Wang, W. Y., and Huang, Z. (2023). Hybrid Hierarchical Retrieval for Open-Domain Question Answering. DOI: 10.18653/v1/2023.findings-acl.679.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.