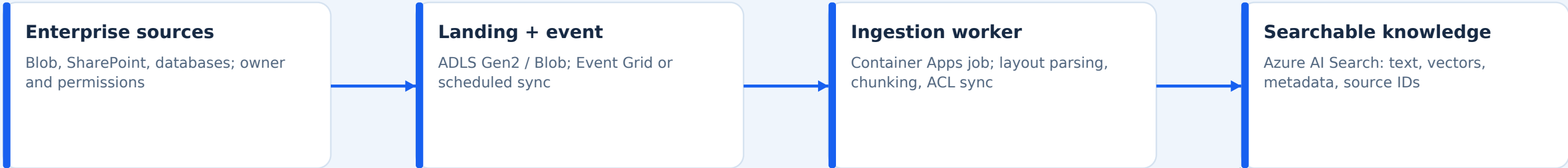


Azure RAG: governed, cited answers

Baseline pattern: controlled ingestion + hybrid retrieval + grounded generation. Optional agentic retrieval is a measured extension.

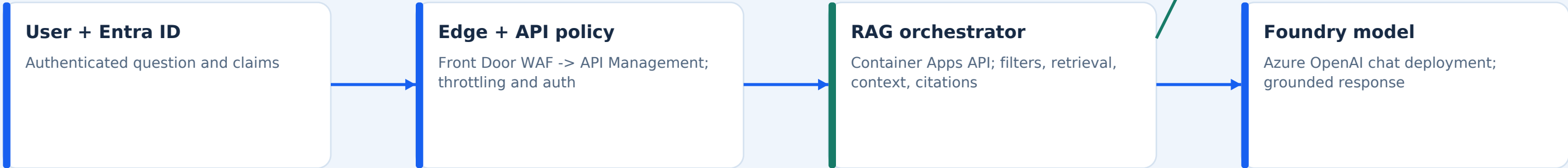
KNOWLEDGE PLANE

Asynchronous ingestion; sources remain authoritative



EXPERIENCE PLANE

Synchronous request; identity and entitlement enforced before generation



retrieve

Cross-cutting controls
Private endpoints + managed identity + Key Vault; Azure Monitor / App Insights; policy-as-code; dev/test/prod separation.

Two pipelines, one knowledge contract

Persist provenance and permissions with every chunk. Only authorized, current evidence may reach the language model.

A. INGEST AND RECONCILE

Use indexers for straightforward Blob content; use a worker for mixed sources, layout, permissions and delete propagation.



B. RETRIEVE AND ANSWER

Application constructs filters from trusted identity; the user cannot supply authorization predicates.



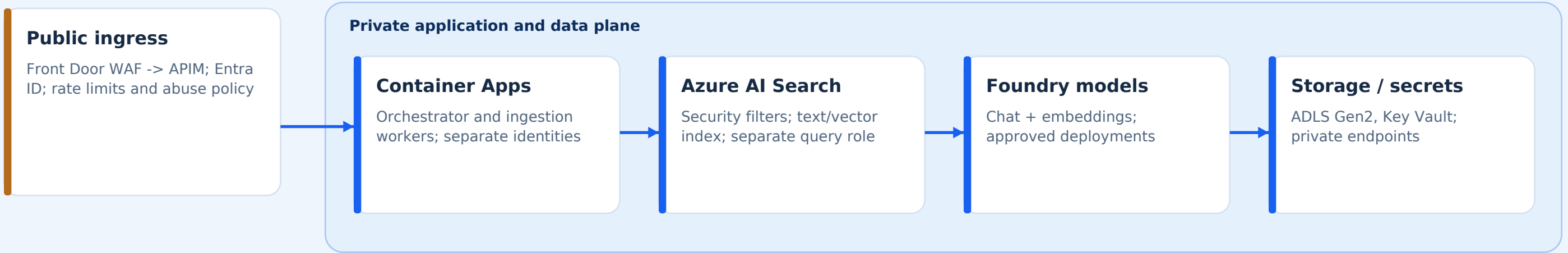
Chunk contract chunk_id, parent_id, tenant_id, title, section_path, content, vector, source_uri, page, version, effective_from/to, acl_ids, hash, deleted_at.

Security boundary and quality loop

The search index is a derivative store. Authorization, version reconciliation and evaluation are product responsibilities.

PRODUCTION BOUNDARY

Public ingress ends at the edge. Data plane services use private connectivity where supported.



Identity

- Access**
Fail closed on missing ACL, unknown tenant or stale entitlement.
- Privacy**
Minimize prompt logs; classify source data; set retention and deletion SLAs.
- Network**
Use managed identity, private endpoints/DNS, egress policy and CMK where required.

Quality

- Retrieval**
Gold queries: Recall@k, nDCG@10, exact-ID and temporal slices.
- Answer**
Groundedness, citation coverage, refusal on missing evidence.
- Service**
p95 latency, token cost, indexing lag, failures and drift.

Implementation decisions and release gates

Use the baseline to prove relevance and security before adding query planning, agents or model adaptation.

Workstream	Build actions	Acceptance evidence
Phase 1: corpus	Inventory sources, owners, ACLs, update cadence; create 100+ judged questions across query classes.	Source inventory + versioned evaluation set.
Phase 2: index	Implement extraction, chunk contract, permission propagation, deletion and index alias rollout.	No orphaned chunks; ACL and delete tests pass.
Phase 3: retrieval	Hybrid query with semantic reranking; tune k, text pool, field boosts and filters on held-out set.	Recall/nDCG improves over BM25 baseline; p95 within budget.
Phase 4: answer	Ground only in retrieved passages; include source title/version/page; test unsupported requests.	Claims trace to sources; abstention and citation checks pass.
Phase 5: operate	IaC + OIDC deployments, observability, rollback, reindex runbook, owner sign-off.	Staging soak and production rollback rehearsed.

Optional: agentic retrieval / LoRA after measured gains

Preview ACL features require separate review

BM25 k1/b are index-level settings

Microsoft documentation used for service choices

[1] Hybrid search

<https://learn.microsoft.com/azure/search/hybrid-search-overview>

[3] Integrated vectorization

<https://learn.microsoft.com/azure/search/vector-search-integrated-vectorization>

[2] Security filters

<https://learn.microsoft.com/azure/search/search-security-trimming-for-azure-search>

[4] RAG evaluation

<https://learn.microsoft.com/azure/foundry/concepts/evaluation-evaluators/rag-evaluators>