White paper

Couchbase as the AI Data Plane for Banking and Telecommunications

Use cases, reference architecture, regulatory alignment and roadmap for GCC institutions

Prepared by
Interkey

Date
September 2026

From Vision to Reality

Executive summary

Banks and telecom operators do not lack AI models. What holds them back is getting the right customer, network and transaction data to those models in milliseconds, under strict regulation. This paper argues that the operational database should become the AI data plane: the single layer that serves live context, vectors, features and agent memory to every AI workload.

Couchbase fits this role because one distributed platform combines:

- **Operational speed:** a memory-first key-value and JSON engine that answers in single-digit milliseconds at very high throughput.
- **AI-native retrieval:** vector search alongside SQL++, full-text and geospatial search, so a single query can mix semantic similarity with hard business filters.
- **Real-time analytics:** a columnar engine for aggregates and features, kept apart from the transactional path.
- **Reach to the edge:** mobile and edge sync for branches, field devices and network sites that must work offline.
- **Sovereign deployment:** self-managed, Kubernetes or managed (Capella) options, with multi-region replication (XDCR) to meet data-residency rules.

For banking, this supports real-time fraud scoring, RAG-grounded advisory and service assistants, KYC/AML investigation copilots and hyper-personalised offers. For telecommunications, it supports network-operations copilots, churn and next-best-action engines, AI-driven customer care and edge inference close to RAN and MEC sites.

The paper is for CIOs, CDOs, enterprise architects and AI platform leads in regulated industries, with a focus on the GCC. It covers the capabilities, a reference architecture, detailed use cases, regulatory alignment, alternatives and a phased roadmap.

The AI data plane: why data, not the model, is the bottleneck

An AI data plane is the runtime layer that feeds live, governed context to models and agents at the moment of decision. It sits between systems of record (core banking, BSS/OSS, network inventory) and the AI layer (LLMs, ML models, agents). It is distinct from the control plane, which trains, evaluates and governs models.

Foundation models are now a commodity that any bank or operator can call. The advantage comes from what surrounds the model: the customer's last ten transactions, the cell site's current alarms, the product terms that apply today, and what the agent already said in this conversation. Three trends make this urgent:

1. **From batch to real time.** Fraud, network assurance and customer care all need decisions in under 100 ms, while analytics platforms still refresh in hours.
2. **From prediction to generation and agency.** Retrieval-augmented generation (RAG) and agents need vector retrieval, conversational memory and tool state, not just a feature table.
3. **From central cloud to edge and sovereign.** GCC regulators require in-country data residency, and telecom AI increasingly runs at the network edge.

What an AI data plane must provide

Capability	What it means in practice
Low-latency context serving	Customer 360, session and device profiles read in single-digit milliseconds
Hybrid retrieval	Vector similarity combined with SQL filters, full-text and geospatial search in one query
Online feature serving	Fresh features for ML scoring, consistent with what the model was trained on
Agent memory and state	Conversation history, tool results and preferences, kept per user and per session
Event-driven freshness	Changes from source systems and streams reflected in seconds (CDC, Kafka)
Governance built in	Encryption, role-based access, audit, masking and residency applied to every AI read
Resilience	Multi-zone and multi-region availability with no single point of failure
Edge reach	The same data model available offline on devices, branches and network sites

The cost of a fragmented stack

Most first-generation GenAI projects bolt together a cache, an operational database, a separate vector database, a search engine and a feature store. Each copy adds sync lag, a new security perimeter, another audit scope and another on-call team. In a regulated bank or operator, every extra system that holds customer data also widens the compliance footprint. A converged data plane reduces this to one platform with one security model.

Banking: pressures and data challenges

Banks must make AI decisions in real time on data spread across decades of systems, while regulators watch every decision. The main pressures:

- **Instant payments and real-time fraud.** Instant payment rails (such as sarie in Saudi Arabia and Aani in the UAE) leave no window to recall a payment. Fraud and mule detection must score each transaction before it is authorised.
- **Digital-first competition.** Digital banks and fintechs licensed under open-banking regimes set expectations for personalised, conversational service.
- **Financial-crime cost.** KYC, AML and sanctions teams spend most of their time gathering evidence across systems. They need investigation copilots that can explain their reasoning.
- **Regulatory scrutiny of AI.** Supervisors expect explainability, model risk management, data lineage and human oversight for automated decisions such as credit and fraud blocks.
- **Legacy cores.** Core banking platforms are the system of record but cannot serve high-volume, low-latency reads for digital and AI channels without risk to stability.

Data challenges that block banking AI

Challenge	Impact on AI
Customer data split across core, cards, CRM, digital and loan systems	Assistants and models see a partial customer, so answers and scores are weaker
Batch-fed data warehouses	Features are hours old; fraud models miss what happened in the last minutes
Separate vector store outside the bank's security perimeter	New audit scope, duplicate PII, harder residency evidence

Challenge	Impact on AI
Unstructured knowledge (policies, product terms, circulars) not indexed	RAG answers are ungrounded or out of date
Peak loads at salary days, Ramadan, White Friday and Hajj seasons	Systems must scale out without re-architecture
Residency and outsourcing rules	Data and models must often stay in-country, on-premises or in a sovereign cloud

The answer is not another silo. It is an operational layer that mirrors the core in real time, serves AI-ready context, and applies the bank's controls to every read.

Telecommunications: pressures and data challenges

Operators run the largest real-time data estates of any industry, but little of that data reaches AI at decision speed. The main pressures:

- **Flat core revenue and 5G payback.** Operators need new revenue from enterprise, 5G standalone, network slicing and digital services to pay back spectrum and network investment.
- **Autonomous networks.** The industry is moving toward higher levels of network autonomy (the TM Forum autonomous networks model). That requires AI that reads live topology, alarms and performance counters.
- **Customer experience and churn.** In prepaid-heavy GCC markets, customers switch easily. Next-best-action and proactive care must react within the session.
- **Digital channels and super-apps.** Operator apps now carry payments, content and lifestyle services, which adds fintech-grade requirements.
- **Edge computing.** Multi-access edge computing (MEC) and on-premises enterprise 5G bring AI inference close to the radio network, often with intermittent backhaul.

Data challenges that block telecom AI

Challenge	Impact on AI
Very high event volumes (CDRs, xDRs, probes, counters, alarms)	Warehouses aggregate too late for live assurance or offers

Challenge	Impact on AI
BSS and OSS in separate silos	Care agents and copilots cannot link a complaint to a network fault
Network knowledge in manuals, runbooks and tickets	NOC copilots cannot answer without RAG over this unstructured content
Subscriber context needed in under 10 ms for policy, charging and offers	General-purpose databases and warehouses cannot serve this path
Thousands of sites, edge nodes and field devices	Data must sync to and from the edge and work offline
Lawful intercept, subscriber privacy and residency obligations	Data movement to public AI services is restricted

Operators need a data plane that absorbs event streams, joins BSS and OSS context, supports semantic search over network knowledge, and extends to the edge.

Couchbase capabilities for the AI data plane

Couchbase is a distributed, memory-first JSON database. Its services can be scaled independently: Data, Query, Index, Search, Eventing, Analytics and Backup. Version 8.0, generally available since 21 October 2025, added billion-scale vector indexing and native encryption at rest (Couchbase 8.0 blog). The capabilities below map to the requirements in the previous section.

1. Operational core: key-value and JSON

- A built-in managed cache serves reads from memory, with sub-millisecond to single-digit-millisecond latency for key lookups.
- A flexible JSON model holds a customer 360, a subscriber profile or an agent's session in one document, without schema migrations.
- ACID transactions across documents cover workflows such as limit updates or order capture.
- The Magma storage engine supports data sets much larger than memory. Version 8.0 lowered its minimum memory quota to 100 MB per bucket.

2. SQL++ query

SQL++ is SQL for JSON: joins, subqueries, window functions and common table expressions over documents. Developers and data teams use familiar SQL skills. Global secondary indexes (GSI) keep queries fast at scale.

3. Vector search: three index types

Index type	Best for	How it works
Hyperscale Vector Index (HVI)	Very large corpora, from millions to over a billion vectors	Hybrid graph and cluster algorithm (Vamana + IVF), partitioned on disk, quantization (PQ, SQ) to reduce memory
Composite Vector Index (CVI)	Vector search that must obey hard filters (customer ID, product, region, entitlement)	Vector column inside a GSI, with scalar filters applied before similarity ranking
Search Vector Index	Hybrid search: semantic plus keyword, geospatial and range in one request	Vectors inside the Search (full-text) service

Vector search is available in SQL++ (for example APPROX_VECTOR_DISTANCE), SDKs and REST. It supports cosine, dot product and Euclidean (L2) similarity. Couchbase reports about 19,000 queries per second at 28 ms latency (66% recall), and 700+ queries per second at 93% recall, on its own billion-scale benchmark (press release). These are vendor figures; validate them with your own data during a proof of concept.

4. Capella AI Services

In the managed Capella service, AI Services add the building blocks for RAG and agents next to the data (Couchbase docs):

Component	Purpose	Status (per docs)
Model Service	Hosts LLMs and embedding models inside Capella, close to the data	Enterprise support
Data Processing Service	Chunks and vectorizes structured and unstructured data	Enterprise support

Component	Purpose	Status (per docs)
Agent Catalog	Governs the tools and prompts that agents use, across LLM frameworks	Generally available
Agent Memory	Persistent, per-user memory for agentic applications	Enterprise support
AI Functions	SQL++ functions for summarization, classification and sentiment	Generally available
MCP Server	Lets LLMs and agents query Couchbase data through the Model Context Protocol	Generally available

Version 8.0 also adds natural-language querying (USING AI) through Capella iQ.

5. Eventing

Server-side JavaScript functions react to every document change. Typical uses: re-embed a document when it changes, recompute a feature, publish an alert, or call a scoring endpoint. Freshness is measured in seconds, not batch cycles.

6. Real-time analytics

The Analytics service (self-managed) and Capella Analytics (formerly Capella Columnar) run aggregate queries on a columnar copy of the data, isolated from the transactional workload. They compute features, cohort statistics and monitoring views without slowing the operational path.

7. Mobile and edge

Couchbase Lite is an embedded database for mobile, desktop and edge devices, including vector search on the device. App Services and Sync Gateway sync data both ways with the central cluster. Apps keep working offline and sync when the connection returns.

8. Replication and resilience

- Automatic sharding (vBuckets), replicas and auto-failover within a cluster; rack and zone awareness.
- Cross Data Center Replication (XDCR) between clusters for active-active or active-passive topologies, disaster recovery and data locality, with filtering to keep some data in-region.

9. Security

- Role-based access control down to bucket, scope and collection; LDAP and SSO integration.
- TLS in transit, including node-to-node encryption.
- Native encryption at rest (8.0) with customer-managed keys through KMS and KMIP, including automatic key rotation.
- Auditing of administrative and data access events.

10. Deployment choice

Couchbase runs self-managed on bare metal or VMs, on Kubernetes (Autonomous Operator), or as Capella on AWS, Azure and Google Cloud. For GCC institutions, self-managed deployment in an in-country data center or sovereign cloud is often required.

Reference architecture

Every AI consumer reaches data through one gateway and one governed Couchbase layer, fed in real time from the systems of record.

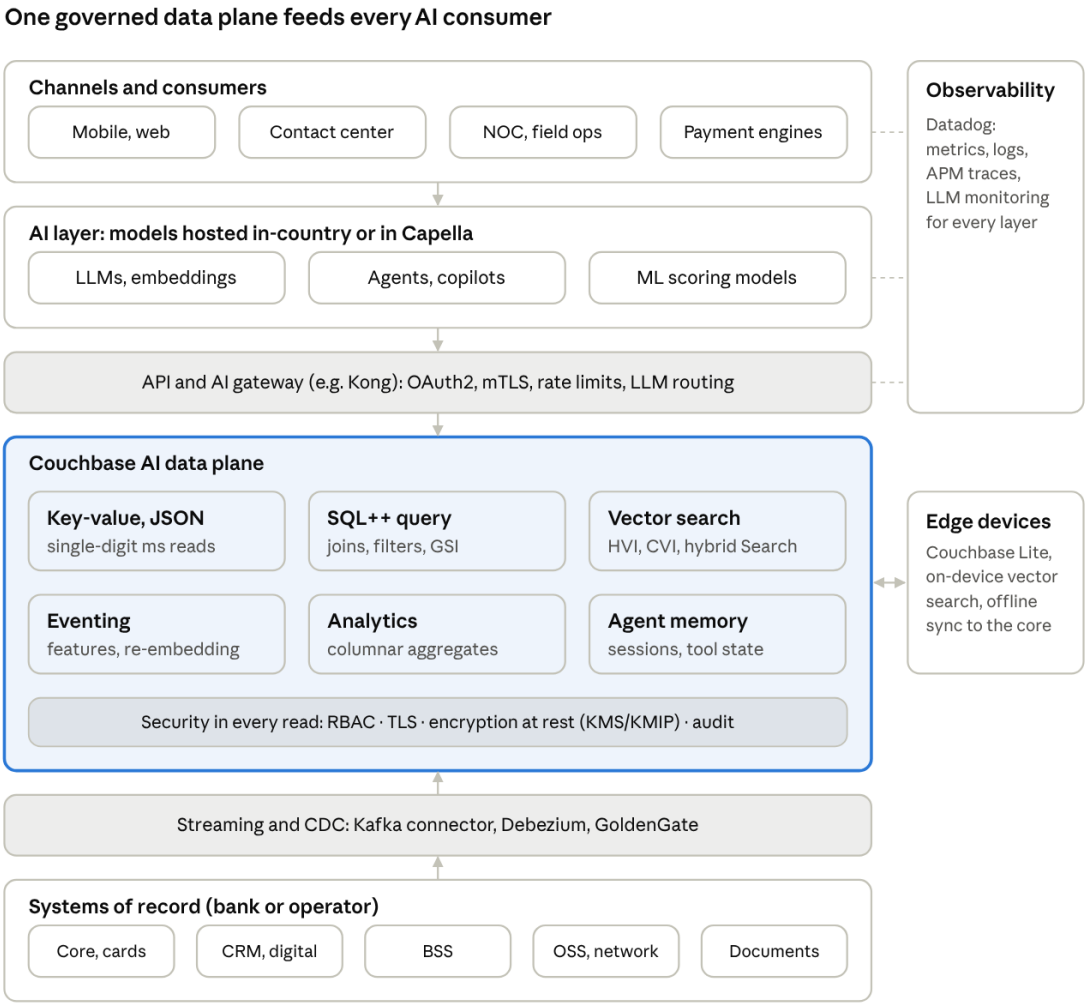


Figure 1. Reference architecture · 6 layers, edge and observability

Read it from the bottom up. Kafka and CDC stream changes from core, BSS and OSS systems into Couchbase within seconds. Eventing keeps features and embeddings current. Models, agents and scoring services call Couchbase through the gateway for context, vectors and memory, and every read passes RBAC, encryption and audit. Edge sites sync a scoped copy both ways. Datadog observes each layer end to end.

Request path for a grounded AI answer

1. The channel sends the question and the authenticated customer or subscriber ID through the gateway.
2. The agent reads the customer 360 or subscriber document by key.

3. It runs one SQL++ query that combines vector similarity with entitlement and date filters.
4. It adds conversation history from agent memory and calls the in-country LLM.
5. The answer, cited sources and model version are written back as an audit record.

Banking use cases

Each use case below names the problem, how the data plane serves it, and the Couchbase features involved. Target values are typical design goals to confirm during a proof of concept, not measured results.

B1. Real-time fraud and mule-account detection

Problem: Instant payments settle in seconds. A fraud model scoring on hour-old features misses account takeover and mule chains.

How it works:

1. Card, payment and login events stream in through Kafka and update a per-customer and per-device profile document.
2. Eventing keeps rolling features current, such as transaction velocity, new-payee count and device changes in the last 24 hours.
3. The authorisation service reads the profile by key and calls the fraud model inside the payment's time budget.
4. A vector index of behaviour embeddings finds accounts that act like known mule accounts, filtered by region and segment.
5. The decision and its reasons are written back for audit and for the case-management system.

Couchbase features: KV reads, Eventing, Kafka connector, Composite Vector Index, ACID transactions, XDCR for active-active across data centers.

Design target: profile read under 5 ms; full scoring path within the authorisation budget (often under 100 ms).

B2. GenAI customer-service and relationship-manager assistant

Problem: Contact-center agents and RMs search many systems and PDFs to answer one question. Public chatbots cannot see account context and risk ungrounded answers.

How it works: Product terms, fee schedules, SAMA or central-bank circulars and internal procedures are chunked and embedded. At question time, the assistant retrieves the customer's profile by key, then runs one hybrid query: vector similarity over the knowledge base, filtered to products the customer holds and to documents in force today. The LLM answers with citations. Conversation state is kept as agent memory.

Couchbase features: Search Vector Index (semantic plus keyword for product codes), SQL++ filters, Agent Memory, AI Functions for summarization of the call, RBAC so that an agent sees only permitted data.

B3. KYC, AML and sanctions investigation copilot

Problem: Analysts spend hours collecting evidence for each alert across core, cards, trade finance and screening systems.

How it works: Alerts arrive from the monitoring system. The copilot, using tools registered in Agent Catalog, pulls the customer's KYC file, related parties and transaction history through SQL++. It uses vector search to find similar past cases and their outcomes, and drafts a case narrative. The analyst reviews and decides; every tool call and retrieved document is logged.

Couchbase features: SQL++ joins across collections, vector similarity on case narratives, Agent Catalog, audit logging, Analytics for network and aggregate views.

B4. Hyper-personalised offers and next-best-action

Problem: Campaign tools segment customers weekly. Offers arrive after the moment has passed.

How it works: Real-time events (salary credit, travel booking, large card spend) trigger Eventing functions that update the customer's features. The decision service retrieves eligible offers with a filtered vector search that matches customer embedding to offer embedding, and applies business rules and consent. The offer is shown in the app within the session.

Couchbase features: Eventing, Composite Vector Index, KV, Couchbase Lite for offline-capable mobile banking app content.

B5. Credit decisioning and early-warning signals

Problem: Loan and SME credit decisions need current data and must be explainable to the regulator and the customer.

How it works: Application data, bureau results, open-banking account data and internal behaviour are assembled into one JSON application document. The scoring model reads features from it; the data plane stores the model version, inputs and reason codes with the decision. For existing borrowers, Analytics runs daily and intraday early-warning queries.

Couchbase features: JSON documents with flexible schema, ACID transactions, Analytics, retention through TTL and backup.

B6. Branch, ATM and field-sales edge

Problem: Field sales staff and remote branches work with poor connectivity but need customer context and product guidance.

How it works: Couchbase Lite on tablets holds the assigned customer portfolio, product catalogue and embedded knowledge. On-device vector search answers product questions offline. Sync Gateway enforces which documents each user may receive and syncs updates when back online.

Couchbase features: Couchbase Lite with vector search, App Services or Sync Gateway access control, encryption on device.

Telecommunications use cases

The same pattern applies: live context by key, hybrid retrieval, event-driven freshness and governed access. As with banking, target values are design goals to confirm in a proof of concept.

T1. NOC and network-operations copilot

Problem: When an alarm storm hits, engineers search vendor manuals, runbooks and old tickets while the outage grows.

How it works:

1. Vendor documentation, method-of-procedure documents, runbooks and resolved trouble tickets are chunked and embedded into a Hyperscale Vector Index.
2. Alarms and KPIs stream in through Kafka and update a live document per network element and per site.
3. The copilot reads the affected elements and their topology neighbours by key, then runs a hybrid search: similar past incidents, filtered to the same vendor, technology (4G/5G, transport, core) and software release.

4. It proposes a probable root cause and the next steps, with citations. Actions that change the network go through the existing change-management tools and a human approval.

Couchbase features: Hyperscale and Composite Vector Indexes, Kafka connector, SQL++ for topology joins, Agent Catalog for tool governance, MCP Server.

T2. AI-driven customer care and digital assistant

Problem: A customer calls about slow data. The agent sees billing but not the network; the chatbot gives generic advice.

How it works: A subscriber 360 document joins BSS data (plan, usage, bills, open orders) with OSS signals (serving cells' health, recent outages in the home area). The assistant retrieves this by MSISDN, adds hybrid search over FAQs, tariff terms and known-issue notices, and answers with the specific cause and fix. Summaries and dispositions are written back by AI Functions.

Couchbase features: KV, Search Vector Index, geospatial filters for area outages, AI Functions, Agent Memory.

T3. Churn prediction and next-best-action in the session

Problem: Churn models run monthly; the retention offer arrives after the port-out request.

How it works: Eventing functions update churn features as events arrive: dropped calls, recharge gaps, complaint tickets, competitor-number calls. The decision engine scores in real time when the customer opens the app, calls care or visits a store. It retrieves eligible offers through filtered vector search and applies consent and margin rules.

Couchbase features: Eventing, KV, Composite Vector Index, Analytics for cohort and campaign measurement.

T4. Real-time subscriber context for policy, charging and offers

Problem: Policy and online-charging functions, and in-app offers, need subscriber state within a few milliseconds at millions of operations per second.

How it works: Couchbase serves as the high-throughput session and subscriber-profile store behind these functions and the digital channels. The same documents feed AI features, so the model and the network function see the same state.

Couchbase features: memory-first KV, horizontal scale-out, XDCR across core sites.

Design target: single-digit-millisecond reads at peak busy-hour load.

T5. Edge AI at MEC, enterprise 5G and field operations

Problem: Private 5G sites (ports, oil and gas, smart cities) and field technicians need AI when the backhaul is slow or down.

How it works: Couchbase Lite or a small edge cluster runs at the site or on technician devices. It holds the local inventory, procedures and embeddings for on-device vector search. Results and new data sync to the central cluster when the link is available; sync rules keep each site's data scoped.

Couchbase features: Couchbase Lite with vector search, Sync Gateway or App Services, XDCR from edge clusters to core.

T6. Fraud and revenue assurance

Problem: SIM-swap fraud, international revenue share fraud (IRSF), Wangiri and subscription fraud cost operators revenue and harm customers.

How it works: Call and event records update per-subscriber and per-number behaviour profiles. Behaviour embeddings and filtered similarity search surface numbers acting like known fraud patterns. SIM-swap events are checked in real time and shared as a signal to partner banks through an API, which links the banking and telecom use cases.

Couchbase features: Kafka connector, Eventing, Composite Vector Index, Analytics, API exposure through the gateway layer.

Security, governance and regulatory alignment

Keeping vectors, context and agent memory inside the same governed platform as the operational data shrinks the compliance surface. No database makes an institution compliant by itself; the table shows which platform controls support each obligation.

Key frameworks in the GCC and beyond

Framework	Applies to	What it asks of the AI data plane	Supporting Couchbase controls
Saudi Personal Data Protection Law (PDPL) and its implementing and transfer regulations (SDAIA)	Banks and operators processing personal data in KSA	Lawful purpose, minimisation, data-subject rights, controlled cross-border transfer	Collection-level RBAC, field-level redaction in the app tier, TTL for retention, XDCR filters to keep data in-country, audit logs
SDAIA AI Ethics Principles and Generative AI Guidelines	All AI systems in KSA (guidelines are non-binding)	Fairness, transparency, privacy, accountability, human oversight	Stored decision records with inputs and model version; Agent Catalog for governed tools and prompts
NCA Essential Cybersecurity Controls (ECC) and Cloud Cybersecurity Controls (CCC)	Critical sectors including banking and telecom	Encryption, identity and access, logging, resilience, in-country hosting for sensitive data	Encryption at rest with KMS/KMIP, TLS, LDAP/SSO, auditing, self-managed in-country deployment
SAMA Cyber Security Framework, Business Continuity and outsourcing rules	Saudi banks	Protection of customer data, tested DR, control over outsourced and cloud services	Multi-zone clusters, XDCR for DR, backup and point-in-time recovery (preview in 8.0), self-managed option
CST regulations (cloud and ICT)	Saudi operators and cloud providers	Data classification, cloud hosting and subscriber-data handling	In-country deployment, access control and audit
UAE PDPL, CBUAE and TDRA rules	UAE banks and operators	Similar residency, privacy and resilience expectations	Same controls as above
PCI DSS v4.0	Anyone storing card data	Protect stored cardholder data, restrict and log access	Encryption at rest, RBAC, audit; keep PANs tokenised outside AI prompts
EU DORA and EU AI Act	GCC groups with EU operations or EU customers	ICT resilience testing; high-risk AI (such as credit scoring) needs data governance and logging	Resilience features, decision logging, lineage metadata in documents

This summary draws on the CMS AI regulation guide for Saudi Arabia. It is not legal advice; confirm obligations with your compliance and legal teams.

Governance patterns for AI on Couchbase

1. **Scope data per AI use case.** Put AI-facing data in dedicated scopes and collections. Give each assistant or agent a service account that can read only those collections.
2. **Minimise before embedding.** Mask or tokenise national IDs, IBANs, card numbers and MSISDNs before text is embedded or sent to an LLM. Store the mapping in a restricted collection.
3. **Filter retrieval by entitlement.** Use Composite Vector Index filters (customer ID, branch, product, clearance level) so a retrieval can never return a document the caller may not see.
4. **Keep models close to the data.** Host embedding and LLM models in-country (self-hosted or Capella Model Service in a permitted region) to avoid sending personal data abroad.
5. **Log every AI decision.** Write the prompt, retrieved document IDs, model version, output and human decision as a record, with a retention period that matches the regulator's.
6. **Govern agent tools.** Register tools and prompts in Agent Catalog. Require human approval for any tool that moves money, changes a customer's account or alters the network.
7. **Monitor for drift and misuse.** Feed query, latency and access metrics to the observability platform, and alert on unusual retrieval patterns.

Performance, scale, resilience and deployment

The data plane must hold up at the busiest hour of the year, not the average one. Couchbase's multi-dimensional scaling lets each service grow on its own: for example, add Index and Search nodes for vector load without over-sizing the Data nodes.

Sizing principles

Workload	Main service	Sizing drivers
Profile and session reads (fraud, policy, care)	Data (KV)	Active working set in memory, ops per second at peak, replica count
Filtered vector retrieval (RAG, offers)	Index (Composite or Hyperscale)	Vector count × dimensions, quantization choice, target recall, queries per second
Hybrid semantic plus keyword search	Search	Corpus size, text analyzers (Arabic and English), vector dimensions

Workload	Main service	Sizing drivers
Feature and event processing	Eventing	Mutation rate, function complexity
Aggregates and early-warning	Analytics	Data volume, query concurrency

Memory for vectors is the largest variable. As a rough guide, one million 1,536-dimension float32 vectors take about 6 GB before index overhead; product or scalar quantization reduces this several-fold. Choose embedding dimensions and quantization during the proof of concept by measuring recall against a labelled set of real questions.

Resilience topologies

Topology	When to use	Notes
Single cluster across three zones or racks	Every production deployment	Replicas plus server-group awareness survive the loss of a zone
Active-passive across two data centers (XDCR)	Standard DR, typical for SAMA BCM testing	Promote the passive site on failover; plan RPO in seconds
Active-active across two regions (XDCR)	Always-on channels such as payments authorisation or subscriber context	Design keys and conflict resolution so each record has a clear owner
Hub and edge (XDCR or Sync Gateway)	Branches, MEC sites, private 5G, field devices	Filter replication so each site gets only its data

Deployment options

Option	Fit for banks and operators
Self-managed on bare metal or VMs in the institution's data center	Strictest residency; full control of keys and network
Kubernetes with the Couchbase Autonomous Operator (OpenShift, Tanzu, upstream)	Private cloud platforms; GitOps and automated scaling

Option	Fit for banks and operators
Capella on a hyperscaler region inside the country	Fastest to start; managed AI Services; confirm region and outsourcing approval with the regulator
Hybrid	Core data self-managed; development, sandbox or non-personal AI workloads on Capella

Version 8.0 also improved operations: faster node warm-up options, auto-failover for unresponsive disks and ephemeral buckets, and a query workload repository for troubleshooting (Couchbase 8.0 blog).

Ecosystem integration

The data plane is only as useful as its connections. Couchbase plugs into the streaming, API, observability and AI tooling that banks and operators already run.

Layer	Typical tools	Integration with Couchbase
Change data capture and streaming	Apache Kafka, Confluent, Debezium, IBM MQ, Oracle GoldenGate	Couchbase Kafka connector (source and sink); CDC from core banking, BSS and OSS into collections; Eventing publishes changes back out
API and AI gateway	Kong Gateway and Kong AI Gateway, Apigee, Azure API Management	Expose data-plane services and AI endpoints with OAuth2, mTLS, rate limits and audit; route and govern LLM calls; open-banking and CAMARA-style network APIs
Observability	Datadog, Prometheus and Grafana, Dynatrace	Built-in Prometheus metrics; Datadog Couchbase integration for cluster health; trace RAG and agent flows end to end with APM and LLM observability
AI frameworks	LangChain, LlamaIndex, Semantic Kernel, CrewAI and other agent frameworks	Couchbase vector-store integrations; Agent Catalog works across frameworks; MCP Server gives agents governed data access
Models	Self-hosted open models (such as Llama, Mistral, Qwen or Arabic-capable models), Capella Model Service, cloud LLM APIs	Store embeddings from any model; keep regulated data with in-country models
Analytics and BI	Spark, Databricks, Snowflake, Tableau, Power BI	Spark connector, Analytics service links, SQL++ and JDBC/ODBC access for reporting

Layer	Typical tools	Integration with Couchbase
Identity and security	Active Directory or LDAP, SSO, HSM and KMS, SIEM	LDAP and SSO authentication, KMIP key management, audit log export to the SIEM

A note on the Couchbase, Kong and Datadog pattern

A common enterprise pattern pairs Couchbase as the data plane, Kong as the API and AI gateway in front of it, and Datadog as the single view across both. Kong enforces who may call which AI or data service. Couchbase supplies the context and applies its own access control. Datadog links a slow customer answer to the gateway, the query and the model call. A companion paper can cover this pattern in depth.

Comparison with alternative architectures

A converged operational data plane is the strongest fit when AI must act in real time on operational data under strict residency rules. Other patterns remain valid for other needs, and most institutions will run a mix.

Criterion	Converged data plane (Couchbase)	Cache + RDBMS + standalone vector DB	Lakehouse or warehouse with vector features	Relational DB with vector extension (such as PostgreSQL + pgvector)	Other document DB with vector search
Latency for live context	Single-digit ms from memory	Fast, but data spread across three systems	Seconds; built for analytics	Good at moderate scale; vertical limits	Good; depends on memory architecture
Vector scale	Billion-scale (HVI), filtered (CVI), hybrid (Search)	Strong in the vector DB	Good for batch and offline RAG	Good to tens of millions; harder beyond	Varies by product
Filtered and hybrid retrieval in one query	Yes (SQL++, CVI, Search)	Needs app-side joins across systems	Yes, but not at operational latency	Yes, in SQL	Usually yes
Data freshness	Seconds (Kafka, Eventing)	Sync lag between systems	Minutes to hours	Seconds	Seconds

Criterion	Converged data plane (Couchbase)	Cache + RDBMS + standalone vector DB	Lakehouse or warehouse with vector features	Relational DB with vector extension (such as PostgreSQL + pgvector)	Other document DB with vector search
Offline edge and mobile	Yes (Couchbase Lite with vector search)	No	No	No	Limited
Multi-region active-active	XDCR	Different per component	Platform-dependent	Complex	Varies
Security and audit perimeter	One	Three or more	One (analytics side)	One	One
Self-managed, in-country option	Yes	Depends on each vendor	Often cloud-only	Yes	Varies
Main trade-off	New platform to learn if not already in use	Operational overhead and duplicate PII	Not for the transactional path	Scale and HA effort at high volume	Check vector scale, edge and residency fit

When a different choice is better

- **Model training and large-scale historical analysis** belong in the lakehouse or warehouse. The data plane serves the online side and feeds curated data back.
- **Small, low-traffic internal RAG** over a few thousand documents may not justify a distributed platform.
- **Strict relational core workloads** such as the general ledger stay in the core banking or billing RDBMS. Couchbase sits beside them as the serving layer, not a replacement.

Implementation roadmap and success metrics

Start with one measurable use case, prove it on real data, then turn it into a shared platform that later use cases reuse.

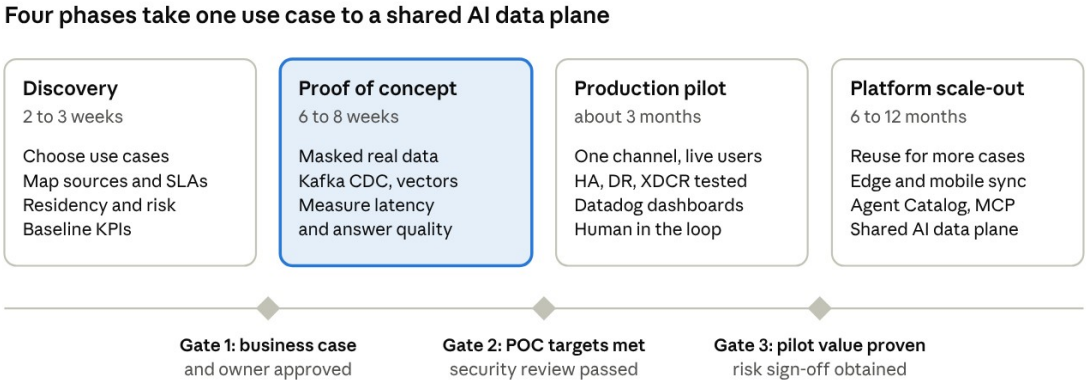


Figure 2. Implementation roadmap · 4 phases, 3 gates

Durations are typical for a bank or operator with Kafka or CDC already in place; they grow if source integration must be built first.

Success metrics to agree at Gate 1

Area	Banking example	Telecom example
Latency	Fraud profile read p99 under 5 ms	Subscriber context read p99 under 10 ms
Retrieval quality	Recall@10 on a labelled set of real customer questions	Correct runbook in top 5 results for past incidents
Business outcome	Fraud loss rate, false-positive rate, analyst hours per alert	Mean time to repair, first-contact resolution, churn rate
Freshness	Transaction reflected in features within seconds	Alarm reflected in site document within seconds
Resilience	Zone failure with no data loss; DR drill within the BCM target	Site failover without service interruption
Governance	100% of AI answers logged with sources and model version	Same, plus no subscriber data leaving the country
Cost	Systems retired or avoided (cache, standalone vector DB)	Same

Conclusion and next steps

Banks and operators that win with AI will be the ones that give every model and agent live, governed context, not the ones with the largest model. Couchbase brings operational speed, filtered and hybrid vector retrieval, event-driven freshness, edge sync and sovereign deployment into one platform. That makes it a strong foundation for the AI data plane in regulated GCC industries.

Recommended next steps:

- Pick one high-value use case per business line (for example B1 fraud or T1 NOC copilot) with a named business owner.
- Run a two-day discovery workshop to map sources, latency targets, residency constraints and success metrics.
- Build a 6 to 8 week proof of concept on real, masked data, measuring latency, recall and answer quality against the current process.
- Agree the target topology (self-managed, Kubernetes or Capella) with security, risk and the regulator-facing team.
- Plan the step from pilot to platform so the second and third use cases reuse the same data plane.

Glossary

Term	Meaning
AI data plane	The runtime layer that serves live, governed context to models and agents
RAG	Retrieval-augmented generation: retrieving relevant documents to ground an LLM's answer
HVI / CVI	Hyperscale Vector Index / Composite Vector Index in Couchbase 8.0
SQL++	Couchbase's SQL dialect for JSON
XDCR	Cross Data Center Replication between Couchbase clusters
MCP	Model Context Protocol, a standard way for LLM agents to call tools and data
MEC	Multi-access edge computing at the telecom network edge
BSS / OSS	Business and operations support systems of a telecom operator

Sources

- Couchbase 8.0: Unified Data Platform for Hyperscale AI Applications (Couchbase blog)
- Couchbase 8.0 press release, 21 October 2025
- Capella AI Services documentation
- Couchbase vector search across all products
- Capella Analytics SDK documentation
- CMS expert guide: AI regulation in Saudi Arabia
- Haveli Investments completes acquisition of Couchbase (Couchbase has been privately held since September 2025)