Data Storage Architecture

SQL Database vs Vector Database in Compliance AI Platform (POC)

---

1. Purpose of This Document

This document explains **how data is stored, separated, and used** in the Compliance AI Platform.

It focuses on: - Why we use **two different databases** - What data goes into a **traditional (SQL) database** - What data goes into a **vector database** - Why this separation is critical for **fintech compliance, auditability, and scalability**

This module should be read after: - Master Overview Document - Content Generation Module

---

2. Why One Database Is Not Enough

In regulated AI systems, data has **two very different natures**:

1. **Authoritative, structured, auditable data**
2. **Semantic, similarity-based, contextual data**

Trying to store both in one database leads to: - Poor compliance auditability - Inefficient retrieval - High cost and complexity

Therefore, the platform intentionally uses:

- **SQL Database (Amazon RDS)** → Source of truth
- **Vector Database (Amazon OpenSearch)** → Semantic retrieval engine

---

3. SQL Database (Amazon RDS – PostgreSQL)

3.1 Role of the SQL Database

The SQL database is the **authoritative system of record**.

It answers questions like: - What rules are active? - Who generated this content? - What decision was made? - When was it made?

These questions require **strong consistency and audit guarantees**.

3.2 What Is Stored in the SQL Database

3.2.1 User & Role Data

- user_id
- role (agent / admin / super_admin)
- timestamps

3.2.2 Compliance Rules (Source of Truth)

- rule_id
- rule_text (legal wording)
- category (IRDAI / Brand / SEO)
- severity (LOW / MEDIUM / HIGH)
- is_active flag
- version number
- created_by
- created_at / updated_at

Rules stored here: - Are enforceable - Are versioned - Are auditable

---

3.2.3 Content Generation Records

- request_id
- user_id
- input_reference
- final_content_reference
- compliance_status
- rules_triggered
- timestamps

3.2.4 Audit Logs

- action_type (generate / validate / update_rule)
- actor
- decision_summary
- rule_version_used
- timestamp

3.3 Why SQL Is Critical for Fintech

- ACID guarantees
- Referential integrity
- Easy audits
- Historical traceability
- Legal defensibility

The SQL database is the **final authority**.

---

4. Vector Database (Amazon OpenSearch)

4.1 Role of the Vector Database

The vector database is a **semantic retrieval system**, not a decision engine.

It answers questions like: - Which regulatory clauses are relevant to this prompt? - Which sections are semantically similar?

It does **not** answer: - Is content compliant? - Should content be approved?

4.2 What Is Stored in the Vector Database

Only **semantic representations** are stored.

4.2.1 Regulatory Clause Embeddings

- IRDAI circular clauses
- Regulatory explanations
- Brand guideline paragraphs

Each vector entry includes metadata:

```
{
  "embedding": [ ... ],
  "rule_id": "IRDAI-3.4",
  "source_doc": "IRDAI_Circular_2023.pdf",
  "jurisdiction": "India",
  "risk_category": "High",
  "version": "v2"
}
```

4.3 What Is Explicitly NOT Stored in Vector DB

- User information
- Generated content
- Compliance decisions
- Rule enforcement logic
- Audit logs

This keeps the vector DB **non-authoritative** and low risk.

---

5. How SQL DB and Vector DB Work Together

5.1 Combined Flow

```
User Request
   ↓
Load Active Rules (SQL DB)
   ↓
Retrieve Relevant Regulatory Context (Vector DB)
   ↓
LLM Generation & Review
   ↓
Final Decision Using SQL Rules
   ↓
Persist Results (SQL DB)
```

5.2 Authority Hierarchy

1. SQL Database (Rules & Decisions)
2. Vector Database (Context Only)
3. LLMs (Language Only)

This hierarchy ensures **compliance safety**.

---

6. Why Vector DB Is More Relevant Than Keyword Search

6.1 Limitations of Keyword Search

- Misses semantic matches
- Requires exact wording
- Fails on paraphrased regulations

6.2 Advantages of Vector Search

- Semantic similarity
- Context-aware retrieval
- Clause-level grounding
- Smaller prompts
- Better LLM accuracy

Example: - Prompt mentions “assured returns” - Vector DB retrieves clause about “guaranteed benefits”

Keyword search would fail; vector search succeeds.

---

7. Why SQL DB Cannot Replace Vector DB

SQL excels at: - Exact queries - Structured filters

SQL fails at: - Semantic similarity - Natural language matching

Therefore: - SQL is the law book - Vector DB is the research assistant

---

8. Versioning & Updates

8.1 Rule Updates

- Rule text updated in SQL DB
- Version incremented
- Related regulatory text re-embedded if needed

Old compliance results remain intact.

8.2 Vector Refresh Strategy

- Re-embedding triggered only on:
    - Rule text change
    - Regulatory document update

Avoids unnecessary cost.

---

9. Security & Compliance Considerations

- SQL DB protected via IAM, VPC, backups

- Vector DB stores no PII
- No decision logic stored in vector DB
- Clear blast-radius separation

---

10. Key Takeaway

The SQL database provides legal authority and auditability, while the vector database provides semantic understanding. Both are required, but they serve fundamentally different purposes.

This document defines the complete data storage strategy for the Compliance AI POC.