# Data Architecture

## 1. DynamoDB Details

Table Name: document\_metadata

|  |  |  |
| --- | --- | --- |
| Attribute | Type | Description |
| aws\_guid | STRING (PK) | Unique identifier generated for each document. |
| file\_name | STRING | Original name of the uploaded file. |
| s3\_url | STRING | S3 path where the document is stored. |
| upload\_date | TIMESTAMP | Date when the file was uploaded. |
| user\_id | STRING | ID of the user who uploaded the file. |
| checksum | STRING | Used for data integrity validation. |
| retention\_policy | STRING | Defines how long the document is retained. |
| metadata | MAP | Additional metadata (file type, tags, source system). |

DynamoDB Design Decisions:

- Billing Mode: On-Demand (Auto-scales based on demand).

- Partition Key: aws\_guid ensures fast lookups and avoids hot partitions.

- Global Secondary Index (GSI): user\_id for user-specific document retrieval.

## 2. Data Backup Strategy

### S3 Backup & Retention

- Versioning: Enabled to keep multiple versions of a document.

- Lifecycle Policies:

- Archive old versions to Glacier after X days.

- Delete non-accessed documents after Y months (as per compliance).

- Cross-Region Replication: Enabled for disaster recovery.

### DynamoDB Backup Strategy

- Point-in-Time Recovery (PITR): Enables restoring data up to 35 days back.

- Daily Snapshots: Stored in Amazon S3 for long-term backup.

## 3. S3 Partitioning Strategy

|  |  |
| --- | --- |
| Partitioning Key | Example |
| Bucket Name | company-docs-bucket |
| Year/Month/Day | 2025/02/14/ |
| User ID-Based | user-1234/ |
| Document Type | invoices/, contracts/, reports/ |

Why This Matters?

- Faster retrieval using structured prefixes.

- Efficient query performance in AWS Athena or S3 Select.

- Access control at the folder level (e.g., restrict access by user\_id).

## 4. Binary Metadata Handling

Since documents are stored as binaries in S3, metadata is stored in DynamoDB, including:

- MIME Type (PDF, DOCX, PNG, etc.)

- File Size

- Upload Source (System, User, API, etc.)

- Last Accessed Timestamp

- Retention & Compliance Policies

This approach decouples the metadata from S3 storage, allowing faster queries on document properties without scanning large S3 files.

## Next Steps

📌 Review & Approve: Ensure alignment with compliance and performance requirements.

📌 Implementation: Update CloudFormation templates and migration scripts accordingly.

📌 Monitoring Setup: Define CloudWatch alarms and logs for backup, partitioning, and metadata validation.