TherapyAI Database Schema Documentation

Overview

This document describes the database schema for the TherapyAI application, which uses Azure Cosmos DB as the primary database. The system is designed to support therapy sessions, user management, and audio processing workflows.

Database Configuration

- **Database Name**: `therapy_ai_db` (configurable via `COSMOS_DATABASE_NAME`)
- **Database Type**: Azure Cosmos DB (NoSQL)
- **Partition Strategy**: Different containers use different partition keys optimized for query patterns

Containers (Tables)

1. users_container

Purpose: Stores user profiles for both patients and therapists.

Partition Key: `/id` (user ID)

Schema:

```
{
  "id": "string",                     // 9-digit random string (Primary Key)
  "email": "string",                  // Unique email address
  "password": "string",               // bcrypt hashed password
  "userFullName": "string",           // Full name of the user
  "fullName": "string",               // Duplicate of userFullName for
compatibility
  "type": "string",                   // "PATIENT" or "THERAPIST"
  "dateOfBirth": "string",            // Format: "YYYY-MM-DD"
  "pictureUrl": "string"              // Profile picture URL
}
```

Indexes:

- Case-insensitive email search
- User type filtering
- Name search capabilities

2. user_tokens_container

Purpose: Manages refresh tokens and password reset tokens.

Partition Key: `/id` (token ID)

TTL: Configured with automatic expiration

Schema:

```
{
  "id": "string",                    // Token ID (Primary Key)
  "user_id": "string",               // Reference to users_container
  "type": "string",                  // "refresh" or "password_reset"
  "token_string": "string",          // JWT or reset token string
  "issued_at": "string",             // ISO 8601 timestamp
  "expires_at": "string",            // ISO 8601 timestamp
  "ttl": "number",                   // Time-to-live in seconds
  "email": "string"                  // Optional: for password reset tokens
}
```

3. session_headers_container

Purpose: Stores session metadata and status information.

Partition Key: `/therapist_id` (optimized for therapist queries)

Schema:

```
{
  "id": "string",                    // Session UUID (Primary Key)
  "sessionId": "string",             // Duplicate of id for compatibility
  "therapist_id": "string",          // Reference to users_container
  "therapist_name": "string",        // Therapist's full name
  "patient_id": "string",            // Reference to users_container
  "patient_name": "string",          // Patient's full name
  "session_date": "string",          // Format: "YYYY-MM-DD"
  "title": "string",                 // Human-readable session title
  "status": "string",                // Session workflow status (see Status
Values)
  "summary_preview": "string",       // Brief summary (max 150 chars)
  "overall_sentiment_positive": "number",  // 0.0 to 1.0
  "overall_sentiment_neutral": "number",   // 0.0 to 1.0
  "overall_sentiment_negative": "number",  // 0.0 to 1.0
  "uploaded_at": "string",           // ISO 8601 timestamp
  "last_updated_at": "string",       // ISO 8601 timestamp
  "processing_id_link": "string",    // Reference to
session_processing_data_container
  "finalized_session_details_id": "string", // Reference to
session_details_container
  "blob_url": "string"               // Optional: Direct blob URL
}
```

Status Values:

- `QUEUED_FOR_CELERY_STT`: Just uploaded, waiting for speech-to-text
- `COMPLETED_PENDING_REVIEW`: STT complete, awaiting therapist review
- `PENDING_FINAL_PROCESSING`: Therapist submitted edits, processing sentiment
- `FINALIZED`: Session complete and available for viewing
- `UPLOAD_FAILED_STT_DISPATCH`: Failed to queue for STT processing
- `ARCHIVED_FINALIZED`: Finalized and archived

4. session_processing_data_container

Purpose: Stores session processing data, transcripts, and workflow information.

Partition Key: `/id` (session ID)

Schema:

```
{
  "id": "string",                    // Session UUID (Primary Key)
  "header_id": "string",             // Reference to session_headers_container
  "therapist_id": "string",          // Reference to users_container
  "patient_id": "string",            // Reference to users_container
  "patient_name": "string",          // Patient's full name
  "patient_email": "string",         // Patient's email
  "session_date": "string",          // Format: "YYYY-MM-DD"
  "original_filename": "string",     // Original audio file name
  "blob_container_name": "string",   // Azure Blob Storage container
  "blob_name": "string",             // Blob path within container
  "blob_storage_path": "string",     // Full blob URL
  "status": "string",                // Processing status (see Status Values)
  "uploaded_at": "string",           // ISO 8601 timestamp
  "summary_initial": "string",       // Initial summary from therapist
  "general_notes": "array",          // Initial general notes
  "timed_notes": "array",            // Initial timed notes
  "transcript_processed": "array",   // STT-generated transcript
  "therapist_speech_stt": "string",  // Optional: therapist speech text
  "patient_speech_stt": "string",    // Optional: patient speech text
  "stt_processed_at": "string",      // ISO 8601 timestamp
  "final_transcript_edited_payload": "array", // Therapist-edited transcript
  "therapist_final_edit_at": "string",        // ISO 8601 timestamp
  "pending_review": "boolean",       // Whether awaiting therapist review
  "archived_at": "string"            // ISO 8601 timestamp when archived
}
```

Transcript Structure:

```
[
  {
    "speaker": "string",             // "Therapist" or "Patient"
    "text": "string",               // Spoken text
    "timestamp": "string"           // Format: "HH:MM:SS.mmm"
  }
]
```

Timed Notes Structure:

```
[
  {
    "timestamp": "string",          // Format: "HH:MM:SS.mmm"
    "content": "string"             // Note content
  }
]
```

5. session_details_container

Purpose: Stores finalized session data with sentiment analysis.

Partition Key: `/id` (session ID)

Schema:

```
{
  "id": "string",                    // Session UUID (Primary Key)
  "sessionId": "string",             // Duplicate of id for compatibility
  "therapist_id": "string",          // Reference to users_container
  "therapist_name": "string",        // Therapist's full name
  "therapist_email": "string",       // Therapist's email
  "patient_id": "string",            // Reference to users_container
  "patient_name": "string",          // Patient's full name
  "patient_email": "string",         // Patient's email
  "patient_date_of_birth": "string", // Format: "YYYY-MM-DD"
  "session_date": "string",          // Format: "YYYY-MM-DD"
  "summary": "string",               // Final session summary
  "timed_notes": "array",            // Final timed notes
  "general_notes": "array",           // Final general notes
  "positive": "number",              // Overall positive sentiment (0.0-1.0)
  "neutral": "number",               // Overall neutral sentiment (0.0-1.0)
  "negative": "number",              // Overall negative sentiment (0.0-1.0)
  "sentiment_scores": "array",       // Detailed sentiment per sentence
  "finalized_at_timestamp": "string" // ISO 8601 timestamp
}
```

Sentiment Scores Structure:

```
[
  {
    "speaker": "string",             // "Therapist" or "Patient"
    "text": "string",               // Sentence text
    "timestamp": "string",          // Format: "HH:MM:SS.mmm"
    "positive": "number",           // Positive sentiment (0.0-1.0)
    "neutral": "number",            // Neutral sentiment (0.0-1.0)
    "negative": "number"            // Negative sentiment (0.0-1.0)
  }
]
```

6. devices_container

Purpose: Manages device registrations for push notifications.

Partition Key: `/user_id` (user ID)

Schema:

```
{
  "id": "string",                    // User ID (Primary Key)
  "user_id": "string",               // Reference to users_container
  "fcm_token": "string",             // Firebase Cloud Messaging token
  "platform": "string",             // "android" or "ios"
  "user_type": "string",             // "PATIENT" or "THERAPIST"
  "last_registered_at": "string"     // ISO 8601 timestamp
}
```

Data Flow and Relationships

User Management Flow

1. **User Registration**: Users created in `users_container`
2. **Authentication**: JWT tokens managed via `user_tokens_container`
3. **Device Registration**: FCM tokens stored in `devices_container`

Session Processing Flow

1. **Upload**: Audio uploaded to blob storage, metadata in `session_headers_container` and `session_processing_data_container`
2. **STT Processing**: Speech-to-text results stored in `session_processing_data_container`
3. **Review**: Therapist reviews and edits transcript via `session_processing_data_container`
4. **Finalization**: Final data with sentiment analysis stored in `session_details_container`

Sentiment Analysis

- Applied only to patient speech
- Calculated per sentence and aggregated for overall session sentiment
- Stored in both `session_headers_container` (overall) and `session_details_container` (detailed)

Query Patterns

User Queries

- Find user by email (case-insensitive)
- Find user by ID
- Search patients by name/email/ID (therapist access only)

Session Queries

- Get sessions by therapist ID (partition key optimization)
- Get sessions by patient ID (cross-partition query)
- Search sessions by patient name, date, or content
- Get pending review sessions by therapist

Security and Access Control

- JWT-based authentication with refresh tokens
- Role-based access (PATIENT vs THERAPIST)
- Session data access restricted to involved parties
- Therapist-only access to patient search and pending reviews

Data Types and Constraints

Date Formats

- **Database Storage**: "YYYY-MM-DD" (ISO 8601 date only)
- **API Responses**: "dd-MM-YYYY" (client-friendly format)
- **Timestamps**: ISO 8601 with timezone ("2023-01-01T12:00:00.000Z")

Sentiment Scores

- All sentiment values are floats between 0.0 and 1.0
- Positive + Neutral + Negative should sum to 1.0
- Default sentiment is neutral (1.0) for non-patient speech

Status Transitions

Sessions progress through these states:

```
QUEUED_FOR_CELERY_STT → COMPLETED_PENDING_REVIEW → PENDING_FINAL_PROCESSING →
FINALIZED
```

Backup and Recovery

- TTL configured for token expiration
- Archived sessions maintained for historical data
- Blob storage contains original audio files

Configuration

Key configuration values (from `config.py`):

- `COSMOS_URI`: Cosmos DB endpoint
- `COSMOS_KEY`: Cosmos DB access key
- `COSMOS_DATABASE_NAME`: Database name
- `AZURE_STORAGE_CONNECTION_STRING`: Blob storage connection
- `AUDIO_UPLOAD_BLOB_CONTAINER_NAME`: Audio storage container