BRD Analysis - 2025-08-08

Okay, here's a summary of the key requirements extracted from the provided BRD content, focusing on clarity and actionable information: **Core Functionality & API Overview:** * **Purpose:** The IDFC Gen-AI App Store API enables document analysis across various formats (PDF, Excel, Images, Videos) utilizing AI models for tasks like RAG, document classification, and data extraction. * **API Endpoints:** * `/health`: Returns service health metrics. * `/file`: The primary endpoint for all document processing requests. * **Supported Document Types:** PDF, Excel (XLS/XLSX), Images (JPG/PNG/JPEG), Videos (MP4). **AI Model & RAG Processes:** * **Model Flexibility:** Supports several models (gemma3:12b, gemma3:4b, GPT-4.1-mini, simplyfillm, llama3.2 - vision:latest, mistral:latest) allowing users to select the best model based on the task and desired performance. * **RAG Process:** Employs chunking, relevance identification, and context-based answer generation. * **Simplyfillm:** A default model running all models to select the best response for critical queries. **Request Parameters & Data Handling:** * **`content`:** Required string parameter for the user query or message. * **`pdf`, `excel`, `image`, `video`:** File upload fields indicating the type of document being processed. * **`model`:** Optional string parameter specifying the AI model to use. * **File Guidelines:** PDF (no size limit), Excel (XLS/XLSX), Images (300 DPI recommended), Videos (MP4, max 5 minutes). **Response & Audit Trail:** * **Response Format:** Structured JSON with: * `description`: Text content extracted from the document. * `audit_trail`: Detailed timestamped event log including: * `timestamp` * `request_id` * `metrics`: Request counts, error rates, response times. * `events`: Detailed log of each action taken (input received, guardrail check, etc.). * **Guardrail System:** Filters potentially harmful content and generates a 403 error if detected. **Error Handling:** * **Common Errors:** * `400 No input provided`: Missing `content` parameter. * `403 Input flagged by guardrail`: Harmful content detected. * `500 Internal server error`: Technical issue - check file format. Do you want me to delve deeper into a specific aspect of these requirements (e.g., the guardrail system, the RAG process, or the error handling) or perhaps address a particular question you have about the BRD?