Product Requirements Document (PRD)

Feature: Synthetic Data Generator

Platform: RagaAI Catalyst

Role: APM Candidate Submission

1. Problem Statement & Opportunity

RagaAI Catalyst is a developer-first platform for building, testing, and governing AI agents and RAG-style pipelines. However, new users often stall early in the journey because they lack domain-specific datasets to begin testing prompt chains, tuning retrieval strategies, or detecting hallucinations. This friction discourages experimentation and reduces onboarding success.

To solve this, I have proposed a "One-Click Synthetic Data Generator" feature. It will allow users to instantly generate domain-specific datasets, including a realistic knowledge base and question-answer (Q&A) pairs with ground-truth answers, tailored to verticals like e-commerce, healthcare, or fintech.

2. Goals & Success Criteria

Product Goals:

- Allow users to generate synthetic domain datasets within seconds.
- Provide high-quality, domain-specific knowledge bases and Q&A pairs.
- Integrate seamlessly with Catalyst's current project and pipeline setup flow.

Success Metrics:

Metric	Target
% of users generating synthetic data in first session	≥ 70%
Avg time to first working pipeline	< 5 minutes
User satisfaction score post-generation	≥ 4.5/5
Reduction in support tickets tagged "no data to start"	≥ 60%

3. User Personas & Primary Use Cases

Persona 1: Early-stage Agent Developer

- **Goals:-**
  - Wants to test pipelines quickly
  - Blocked by lack of sample data

- **Pain Points:-**

  - Doesn't have a domain dataset.
  - Needs something to test the agent end-to-end right away.
  - Doesn't want to write 100 Q&A examples manually.

- **How They Use This Feature:-**

  - Select a domain (e.g., "Fintech Support Chat").
  - Generate synthetic knowledge base and Q&A pairs.
  - Plug into RAG pipeline for immediate testing.

Persona 2: Data Scientist / AI Researcher

- **Goals:**

  - Benchmark different retrieval strategies.
  - Stress-test RAG systems with large-scale synthetic data.
  - Validate hallucination handling with ground truth.

- **Pain Points:**

  - Manually generating test data is time-consuming.
  - Needs structured data with metadata (e.g., answer accuracy, categories).

- **How They Use This Feature:**

  - Choose large dataset size, tweak schema settings.
  - Export structured dataset to CSV/JSON.
  - Compare multiple agents with controlled test inputs.

Persona 3: Startup Founder / Prototyper

- **Goals:**
    - Validate a product concept (e.g., AI doctor, AI shopping assistant).
    - Build a working demo with minimal data engineering effort.
- **Pain Points:**
    - Doesn't have labeled data.
    - Doesn't understand how to generate good Q&A pairs manually.
- **How They Use This Feature:**
    - Use pre-filled templates for verticals like healthcare or e-commerce.
    - Generate 10–20 sample Q&As with dummy data.
    - Demo prototype to stakeholders or investors.

Key Use Cases:

- Generate knowledge base and Q&A data for RAG pipelines
- Benchmark hallucination detection
- Prototype an AI product with believable dummy data

4. Assumptions & Out-of-Scope

Assumptions:

- Users will select a domain/vertical.
- Speed of generation matters more than real-world accuracy.
- Synthetic data will not include real or sensitive PII.
- Users can set dataset size and toggle PII filtering.
- Users expect export functionality.

Out-of-Scope (V1):

- Uploading real datasets as templates
- Fine-tuning LLMs
- Multi-language support
- Custom schema creation
- Saving datasets as templates

5. Solution Overview

Problem with Existing Flow: Users must upload a CSV to generate data. New users without data are blocked.

New Feature Flow:
1. User clicks "Generate Synthetic Data" on the Knowledge Base tab.
2. User selects a domain (e.g., healthcare).
3. Chooses dataset size (e.g., 20 Q&As).
4. Toggles PII filter.
5. Clicks Generate.
6. System returns:
   - Synthetic knowledge base
   - Q&A pairs with ground-truth
7. User can preview, export, or use the data in the RAG pipeline setup.

System Flow:
- UI collects inputs (domain, size, filter)
- Backend calls generation API with domain-specific prompts
- System filters out PII (if selected)
- Output formatted as structured knowledge base + Q&As
- Data is stored and previewed in Catalyst

6. Detailed Functional Requirements

6.1 Data- set Specification Workflow

The workflow enables users to define the structure and scope of the synthetic dataset before generation:
1. Entry: User selects “Generate Synthetic Data” from the Knowledge Base tab.
2. Parameter Form: Domain selector, dataset size input, PII filter toggle.
3. Validation: System validates parameters and displays estimated generation time.
4. Generation Trigger: User clicks “Generate”; an async API call is fired.
5. Preview & Confirm: Generated data appears with tabs for Knowledge Base and Q&A pairs.
6. Commit: User chooses “Use in Pipeline” to store the dataset inside the project.

6.2 Generation Parameters

Parameter	Description	Options / Range
Domain Schema	Pre-defined schema templates per vertical.	E-commerce, Healthcare, Fintech, Edtech
Dataset Size	Number of Q&A pairs plus supporting docs.	10 – 100 (step 10)
Language	Content language of generated text.	English (v1)
PII Controls	Toggle to strip personally identifiable info.	On / Off

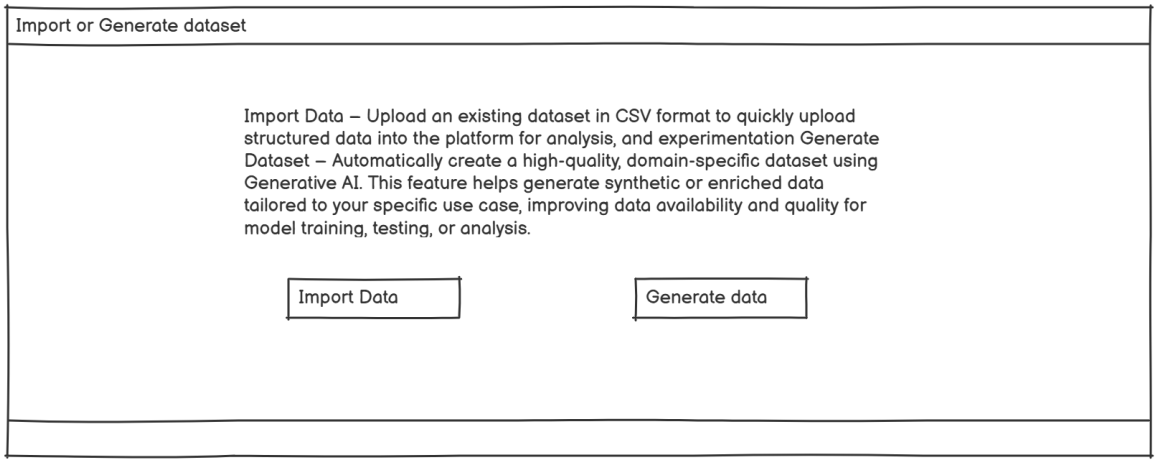
6.3 UI Entry Points

- Knowledge Base tab – primary "Generate Synthetic Data" button next to "Upload CSV".
- Project Onboarding wizard – optional step to jump-start with synthetic data.
- Empty State CTA – visible when no data sources exist in a new project.

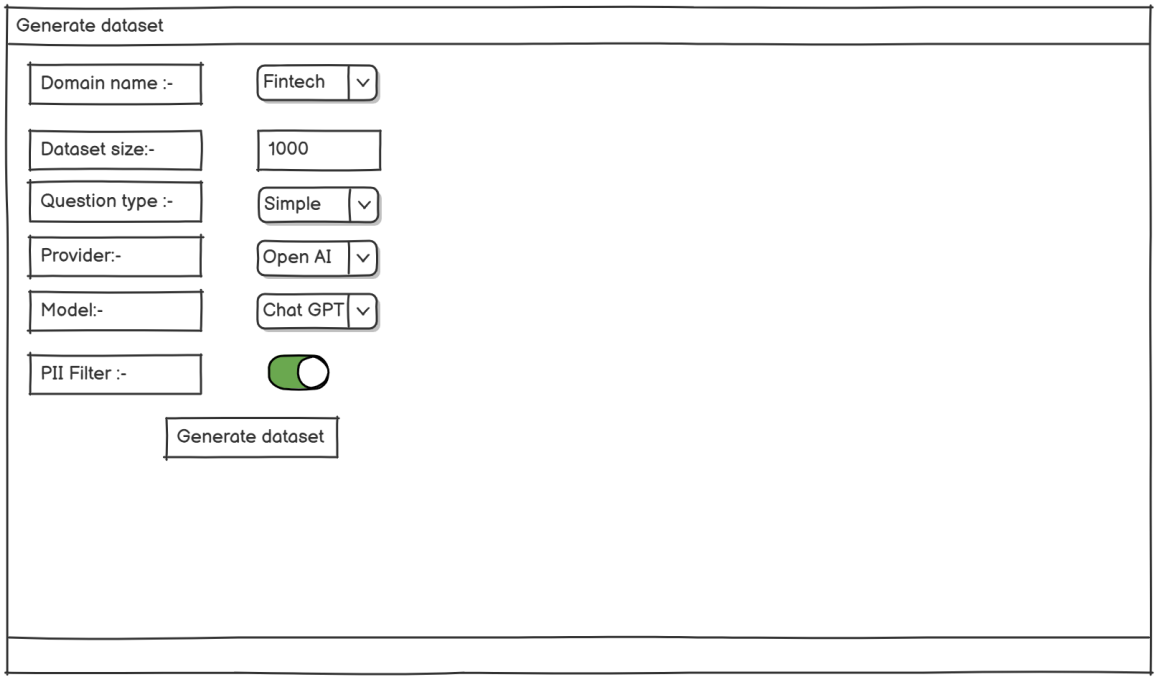
7. Wireframes & User Journey

The following wireframes illustrate key screens and states:-

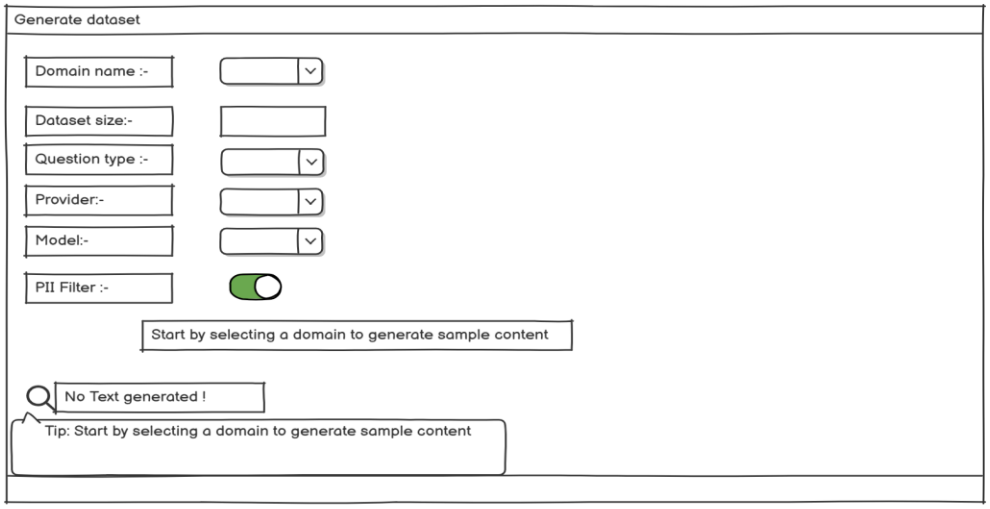
1. WF-1 Entry Point: Choice between Upload CSV vs Generate Data.



2. WF-2 Generation Form: Domain selector, size slider, PII toggle, Generate button.

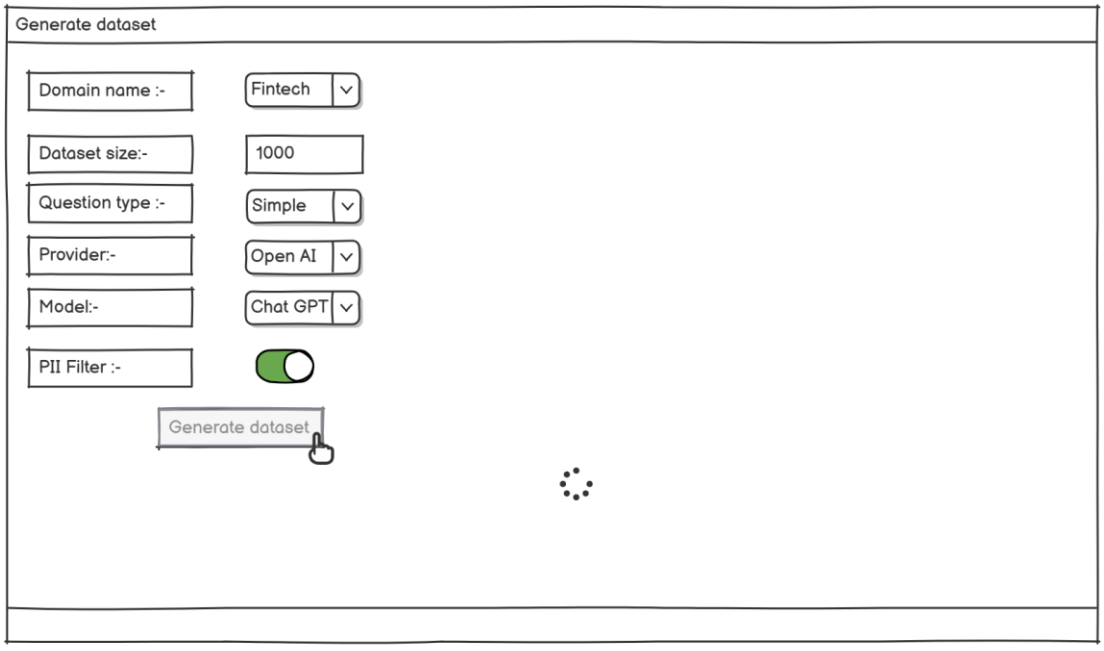


3. WF-3 Empty State: No data generated yet with CTA.



Explaination:-
This window appears When the user opens the "Generate Synthetic Data" tab or model for the first time.

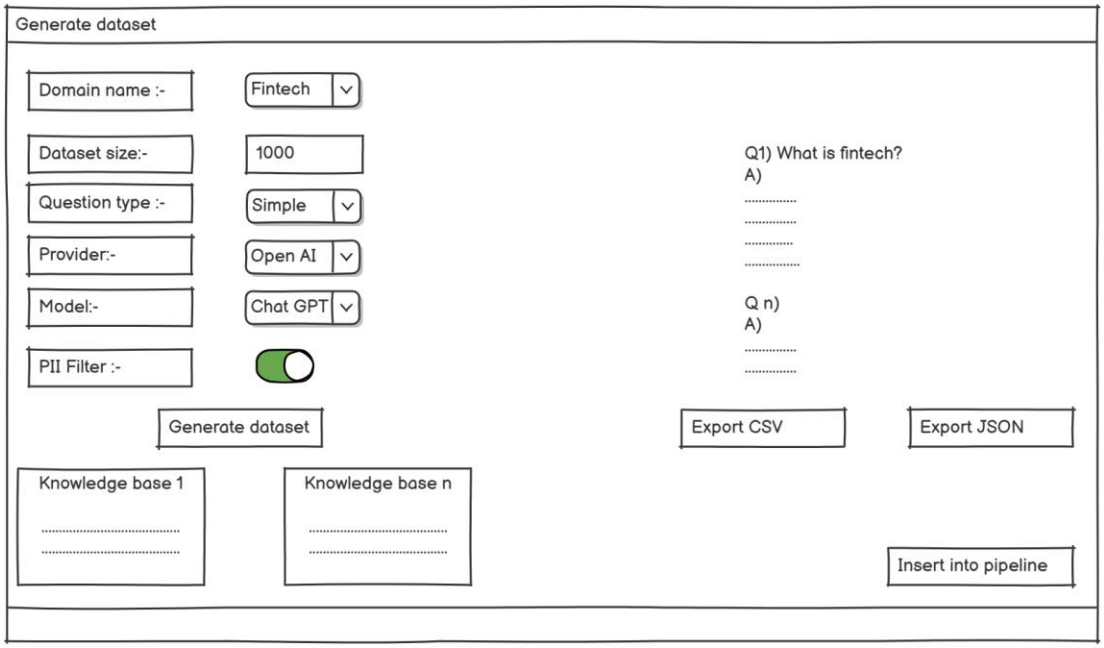
4. WF-4 Loading State: progress bar post-click.



Explaination:-

After clicking "Generate dataset button, the loading animation appears and the data starts generating

5. WF-5 Preview Screen: Tabs for Knowledge Base & Q&A pairs, Export/Use buttons.

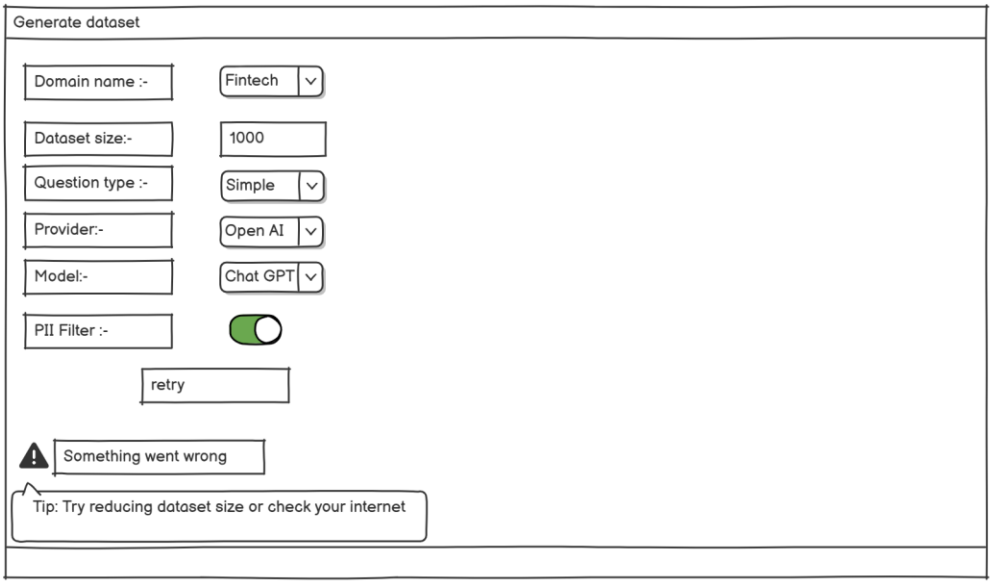


Explaination:-

After the Dataset is generated, the user will able to see two sections :-
1) Knowledge Base Preview (articles, content)
2) Q&A Pairs Preview

And then
Buttons: Export CSV I Export JSON I Use in Pipeline

6. WF-6 Error State: Generation failed notice with Retry option.



Explaination:-

If an error exists during the data generation, a retry button will appear along with an error message and a tip.

Prepared by: Maheedar Balivada
Role: APM Candidate, RagaAI Catalyst
Gmail:- maheedarb@gmail.com
Linkedin:- https://www.linkedin.com/in/maheedarbalivada/