SMART INDIA HACKATHON 2025

SMART INDIA HACKATHON 2025

- **PROBLEM STATEMENT ID :** 25040
- **PROBLEM STATEMENT TITLE :** FloatChat - AI-Powered Conversational Interface for ARGO Ocean Data Discovery and Visualization
- **THEME :** Miscellaneous
- **PS CATEGORY** : Software
- **TEAM ID** : 86270
- **TEAM NAME** : Nautilus Crew

"The ocean of data is not silent. Our mission is to translate its depths into conversation."

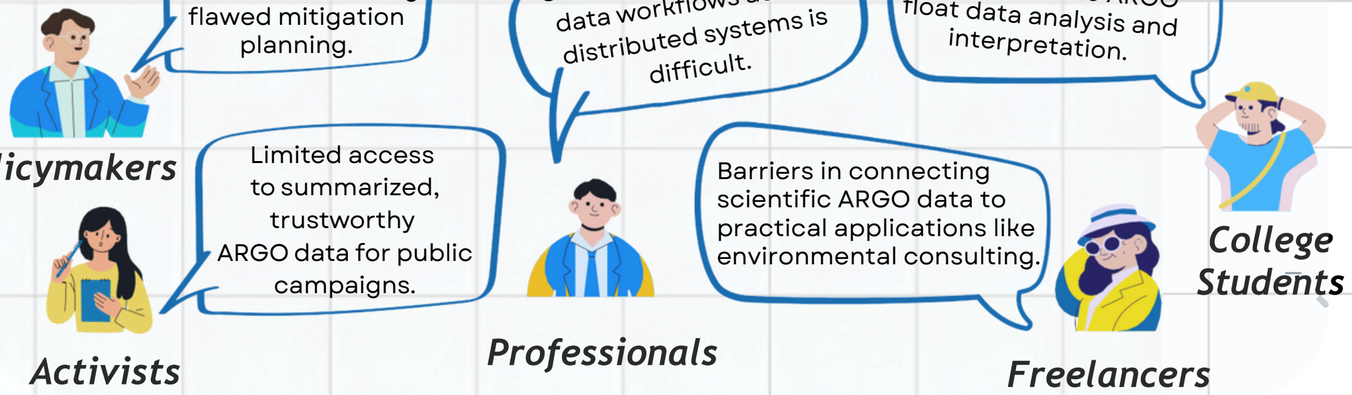
Nautilus Crew

FloatChat - AI-Powered Conversational Interface for ARGO Ocean Data Discovery and Visualization

SMART INDIA HACKATHON 2025

"AI-driven platform which democratizes complex ocean data, designed for rapid, accurate disaster prediction and informed policy action."

STAKEHOLDER ANALYSIS: CHALLENGES AND BARRIERS



Data Acquisition

Fetching the ARGO DAC datasets using python scripts, using the netCDF4 module.

Data Processing

Clean, filter, and standardize the ingested ARGO netCDF datasets. Remove missing or unreliable points, correct errors, and create uniform, analysis-ready tables.

Infrastructure

Hosting backend services and APIs on a scalable cloud platform. To improve data ingestion.

CORE IDEA

Embedding & Storage

Vectorize float metadata and oceanographic parameters.

Visualization & Interface

Display geospatial maps, time series graphs, and summary statistics.

AI Query Engine

Generate the corresponding SQL or vector search query using LangChain and Rasa.

INNOVATION & DIFFERENTIATION

Conversational Filter with Intelligent Suggestion

Utilizes RASA and LangChain to analyze user prompts, detect missing information, and proactively suggest enhancements.

Query Intent Analysis

LangChain adds a verification step to check and correct semantic queries, ensuring they are logically accurate before use.

Contextual Data Retrieval

Using API calls we access language models for generating and processing text. This generates assistive text, which helps users get context aware answers.

Nautilus Crew

TECHNICAL APPROACH

SMART INDIA HACKATHON 2025

ARGO Data Acquisition

Data Processing

AI Query Engine

Cloud Infrastructure

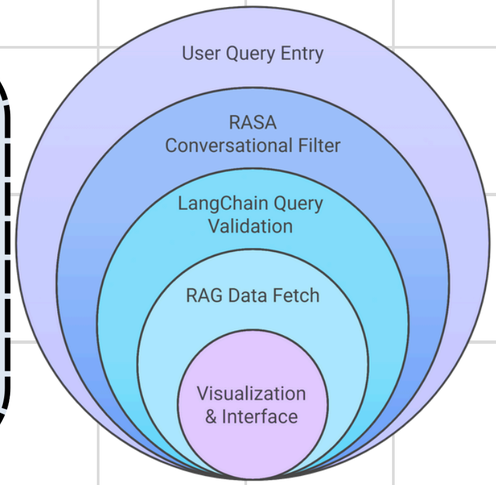
Visualization & Interface

Outlier Detection Approach

- **Goal** : Improve ARGO data accuracy by removing outliers.
- **Key Features:** Uses α convex hull and point-in-polygon methods on temperature-salinity profiles to identify erroneous points deviating from climatological patterns.
- **Outlier Detection:** Flags points that deviate significantly from expected physical relationships or physical constraints, including sensor glitches and spikes.
- **Outcome**: Cleaner, high-quality ARGO datasets

Leveraging RASA for Conversational Filtering

- **Filters** and **interprets** user queries to identify **true intent**.
- Suggests **clarifications** or **better query options** for **vague inputs.**
- **Maintains context** across multi-turn conversations.
- Enhances **accuracy** and user experience by **reducing misinterpretations**.



Technologies to be Used

- **Python & Shell:** Automate ARGO data retrieval and preprocessing.
- **Pandas, GeoPandas:** Clean and standardize ocean dataset files.
- **PostgreSQL, PostGIS:** Store and query vectorized and spatial data efficiently.
- **LangChain:** Convert natural language queries to database searches, manage conversational flows.
- **RASA:** Filter and clarify user queries via multi-turn dialogue.
- **React.js, Tailwind CSS, Chart.js, Django:** Interactive dashboards and geospatial visualizations.
- **Google Cloud Platform:** Deploy scalable cloud backend services.

FRAMEWORKS

Google Cloud, PostgreSQL, LangGraph, PyTorch

PROGRAMMING LANGUAGES

pandas, NumPy

LangChain Query Validation

Suggestion Model

- Checks queries for logic errors and unsafe commands.
- Refines and optimizes queries for accuracy.
- Prevents invalid or risky queries from reaching the data pipeline.
- Improves reliability and data integrity.

Vector Database Engine

Database

- Stores high-dimensional vector embeddings for efficient semantic search and retrieval.
- Enables similarity-based matching between user queries and ARGO dataset profiles.
- Supports fast, scalable queries across massive oceanographic datasets.
- Scales efficiently for large, high-dimensional ocean datasets without sacrificing query speed.

RAG Models for Data Fetching

RAG & Retrieval

- Combines large language models with real-time semantic search on ARGO datasets.
- Retrieves relevant metadata and profiles based on user natural language queries.
- Synthesizes fetched data into coherent, context-aware responses.
- Ensures up-to-date, precise information for accurate scientific analysis and user queries.

Visualization & Interface

Interface

- Builds interactive dashboards and geospatial maps for data exploration.
- Uses React.js, Tailwind CSS, Chart.js, and MapBox GL for rich user experience.
- Presents complex ocean data intuitively for both experts and lay users.
- Enables real-time updates and seamless interaction with scientific visualizations.

FEASIBILITY & VIABILITY

Nautilus Crew

SMART INDIA HACKATHON 2025

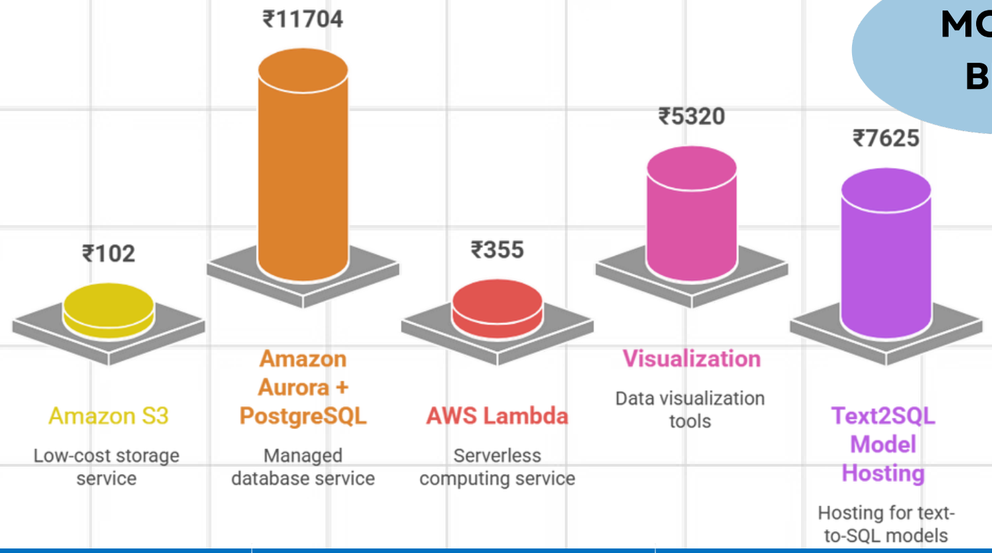
TECHNICAL FEASIBILITY

Sustaining AI Performance: Mitigating long-term model drift through a continuous monitoring and automated retraining pipeline, ensuring the AI's accuracy against evolving datasets and user behavior.

Production-Grade Security: To support a multi-user environment, the architecture incorporates robust authentication protocols and role-based access control (RBAC), guaranteeing data integrity and user privacy.

Scientific Reproducibility: The system architecture mandates rigorous versioning for both datasets and AI models. This ensures every generated insight is auditable and can be precisely reproduced for validation.

MONTHLY COST BREAKDOWN



Estimated Monthly Cost: ~₹25,000

NOTE: All cost estimates are projected for an initial scale of 1,000 daily active users.

Current Prototype on GCP

Data Acquisition: Python+xArray
Data Processing: Pandas, GeoPanadas
Embedding and Processing: PostgreSQL, PostGIS
AI Query Engine: Rasa+Langchain+HuggingFace
Visualization+Interface: Reactjs+TailwindCSS+Leaflet
Infrastructure: GCP+API Gateway

Our current GCP setup will struggle under higher load and global traffic, leading to latency spikes and increased costs.

To avoid these scaling bottlenecks, we will migrate core services to AWS (S3, Aurora, Lambda) for elastic capacity and optimized performance.

Scaled Production

Data Acquisition: Python+xArray
Data Processing: Pandas, GeoPandas
Embedding and Processing: PostgreSQL, PostGIS
AI Query Engine: Rasa+Langchain+HuggingFace
Visualization+Interface: Reactjs+TailwindCSS+Leaflet
Infrastructure: Amazon S3, Amazon Aurora, AWS Lambda

Nautilus Crew

IMPACT AND BENEFITS

SMART INDIA HACKATHON 2025

Scientific Impact

- **Democratizes Argo oceanographic data** for researchers, students, and policymakers.
- Encourages **interdisciplinary research** (AI+ Oceanography + Climate Science).
- Supports **high-quality publications** through easy access to QC/adjusted data.

Societal Impact

- **Policy & Governance:** Enables evidence-based decision-making for fisheries, coastal planning, and climate adaptation.
- **Education:** Provides students with an intuitive tool to learn from real-world ocean data.
- **Awareness:** Increases public understanding of ocean health and climate change.

IMPACT

BENEFITS

Economic & Industrial Benefits

- **Fisheries & Marine Economy:** Enhanced ocean monitoring → accurate yield forecasts & sustainable practices.
- **Shipping & Navigation:** Real-time float data → optimized routing, safety, and fuel efficiency.
- **Disaster Preparedness:** Early anomaly detection (temperature, salinity) → better prediction of cyclones, floods, algal blooms.

Environmental Benefits

- Supports **climate change monitoring** by analyzing long-term ocean temperature and salinity trends.
- Helps track **ocean deoxygenation and pollution impacts**.
- Contributes to **sustainable ocean resource management**.

Nautilus Crew

RESEARCH AND REFERENCES

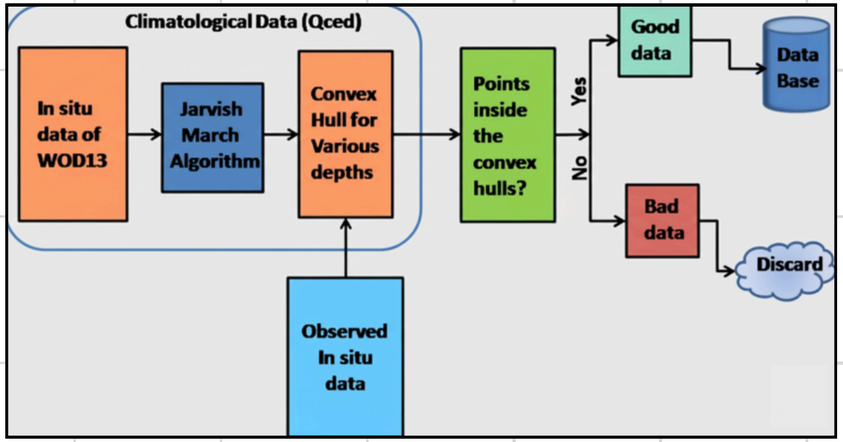
SMART INDIA HACKATHON 2025

PROOF OF WORK

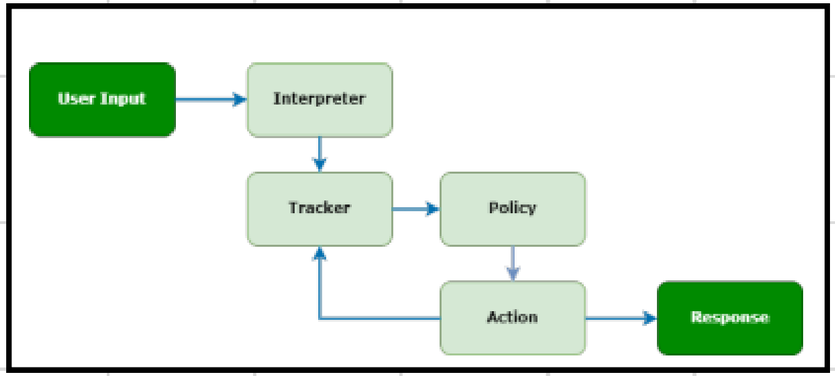
PROTOTYPE

YouTube

For a live demonstration of our working prototype, here is the YouTube link



Convex hulls for outlier detection



RASA Core Architecture

LINKS

PROTOTYPE APPROACH 1

- Retrieval-Augmented Generation for Large Language Models
- Application of retrieval-augmented generation for interactive industrial knowledge management via a large language model

PROTOTYPE APPROACH 2

- Quality control of oceanographic in situ data from Argo floats
- Deep Argo Improves the Accuracy and Resolution of Ocean Bathymetry
- A Complete Survey on LLM-based AI Chatbots

REFERRED RESEARCHES

- AI Powered Legal Querying System using NLP
- DBCopilot: Natural Language Querying over Massive Databases via Schema Routing

Two Aspects

Handling Incomplete and Inaccurate User Queries:

Users may not always know exactly what they want, resulting in incomplete or inaccurate prompts. To handle this, we will use RASA to implement a suggestion model that processes user input and provides relevant query suggestions or clarification options, helping users refine their commands.

Ensuring Logical and Efficient SQL Queries:

Generated SQL queries may contain logical errors or inefficiencies, such as infinite loops or incorrect syntax. To prevent this, we will perform validation checks on each SQL query after generation. LangChain will be used to automate the validation and refinement process, ensuring queries are correct and efficient before execution.