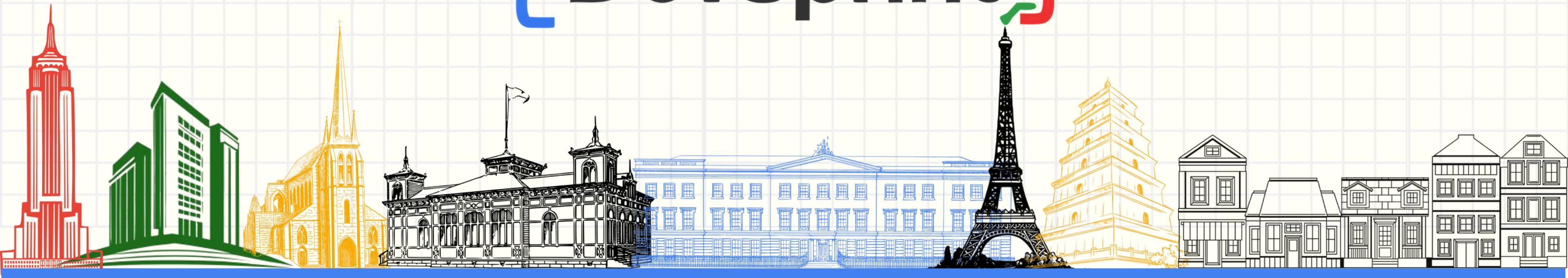


Google Developer Groups
On Campus • MITS-DU Gwalior



Team name: Mind Mesh

Team leader name: Pulatsya Bhagwat

Problem Statement: Open Innovation “Intelligent Data Science Copilot”

Brief about solution and problem statement

Problem Statement

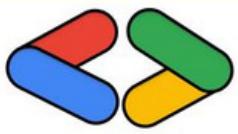
Data science workflows are complex and manual, requiring multiple tools, high expertise, and significant time.

This makes it difficult for beginners to start and forces professionals to repeat routine tasks

Solution – Data Science Agent

Data Science Agent is an autonomous AI system that automates the entire data science pipeline using natural language.

Users upload a dataset, describe the goal, and the agent handles data preparation, modeling, and visualization—making data science faster, simpler, and accessible.



Opportunities

a. How different is it from any of the other existing ideas?

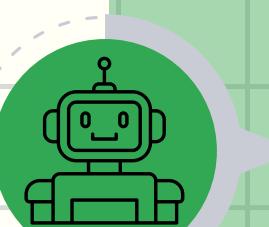
| Existing Tools | Data Science Agent |
|-----------------------------|--|
| Chat-based suggestions only | Autonomous execution |
| Manual ML pipelines | Fully automated workflows |
| No memory across steps | Session memory & context awareness |
| Fixed AutoML pipelines | Dynamic reasoning + tool chaining |
| Limited explainability | Dedicated reasoning & Business summaries |



b. How will it be able to solve the problem?



01
Natural Language Input
Users describe goals in plain English



02
End-to-End Automation
Automates the complete data science workflow



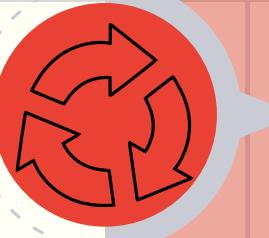
03
Context Awareness
Maintains session memory across steps



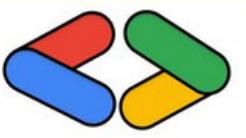
04
Smart Data Processing
Handles data cleaning and feature engineering



05
Intelligent Tool Orchestration
Automatically selects and executes the right ML tools



06
Self-Correcting Execution
Detects errors and retries with fixes



List of features offered by the solution

Intelligent Agent Capabilities

- Dual LLM support (Gemini + Groq)
- Natural language task understanding
- Autonomous tool execution
- Session memory across conversations
- Automatic error recovery
- Built-in code interpreter

Multiple Interfaces

- Web UI (React.js)
- Command Line Interface (CLI)
- REST API (Cloud Run ready)
- Python SDK

Complete Data Science Pipeline

- Dataset profiling & quality checks
- Intelligent data cleaning
- Advanced feature engineering
- Model training & ensemble learning
- Hyperparameter tuning (Optuna)
- Explainable AI (SHAP, feature importance)
- Interactive & static visualizations
- Automated EDA reports

Performance & Scalability

- Polars & DuckDB for fast processing
- Token-optimized LLM communication
- SQLite caching for efficiency
- BigQuery integration for large datasets

Google Technologies Used

Google Cloud Run

Deploys the agent as a scalable REST API

Google AI Studio

Used to integrate and experiment with Gemini models for AI reasoning.

Google BigQuery

Enables efficient querying, profiling, and analysis of large datasets

Google Container Registry (GCR)

Stores Docker images for deployment

Gemini 2.0 Flash

Provides fast, low-latency LLM reasoning for decision-making.

01

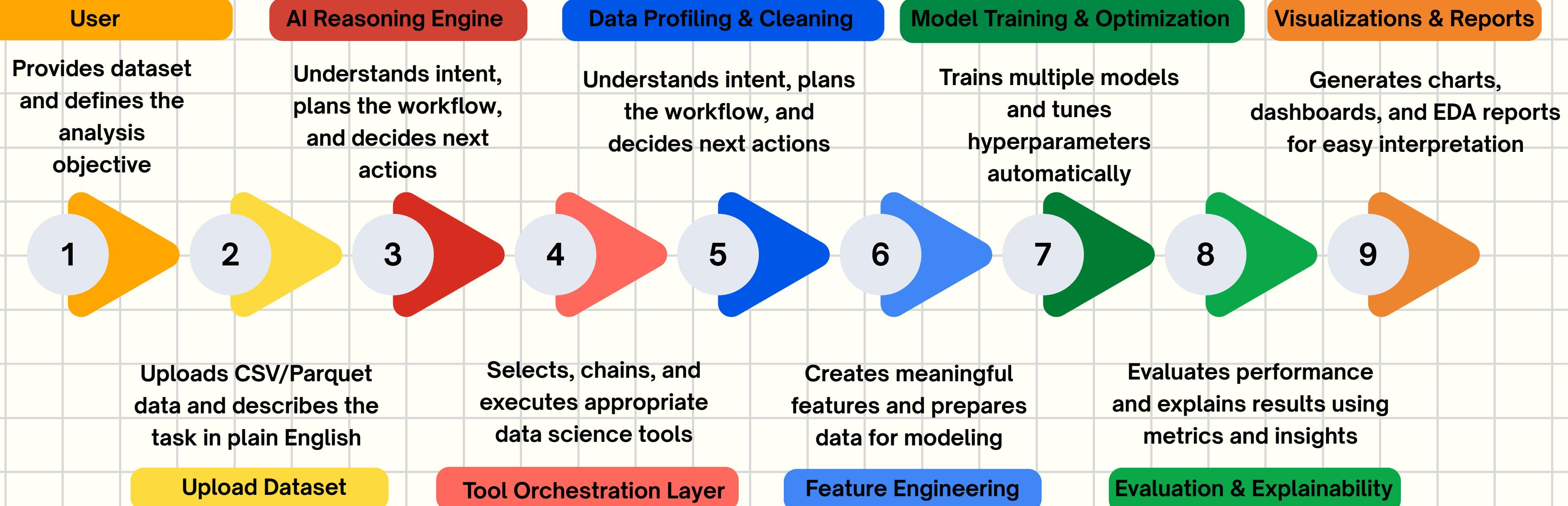
02

03

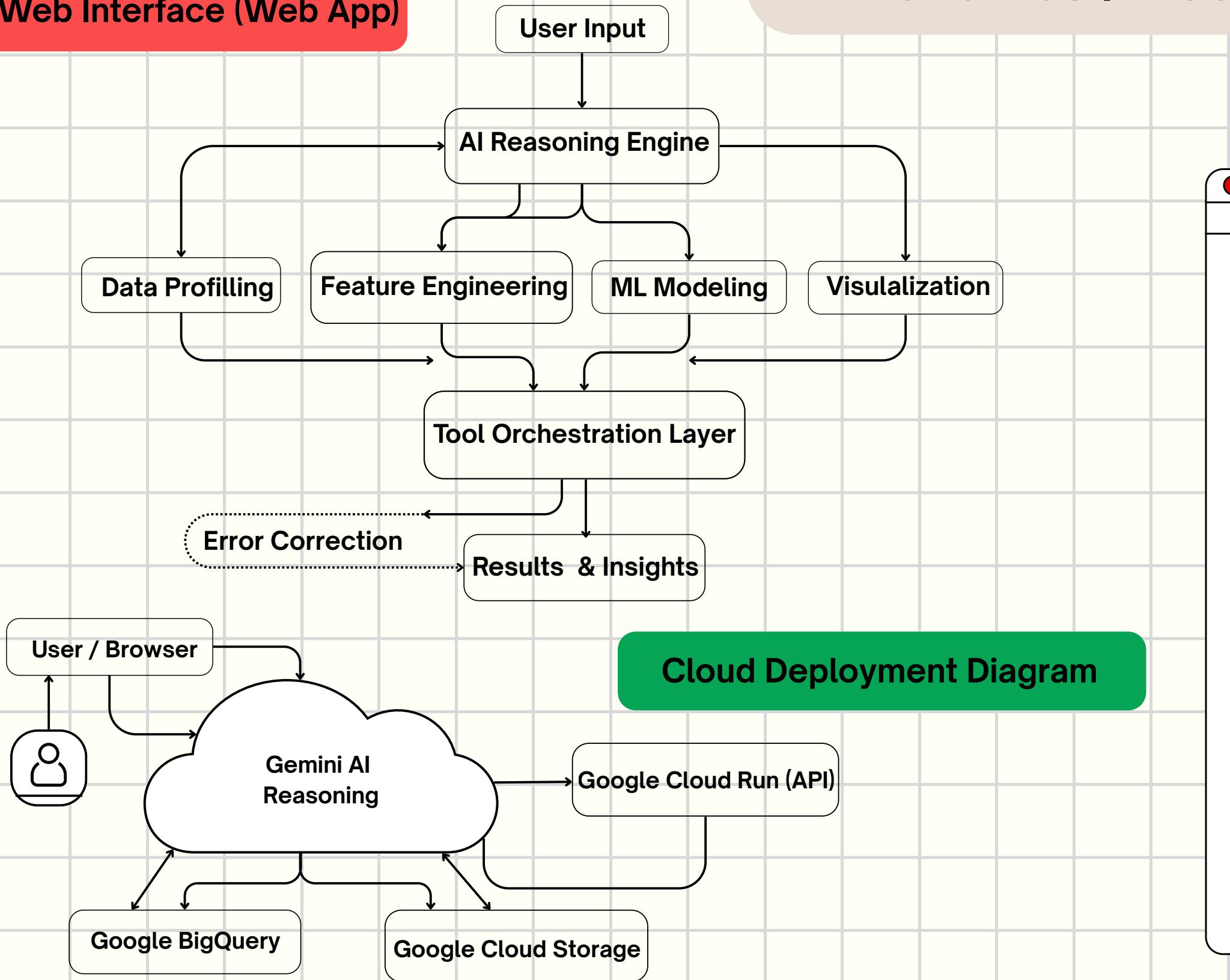
04

05

Process Flow Diagram

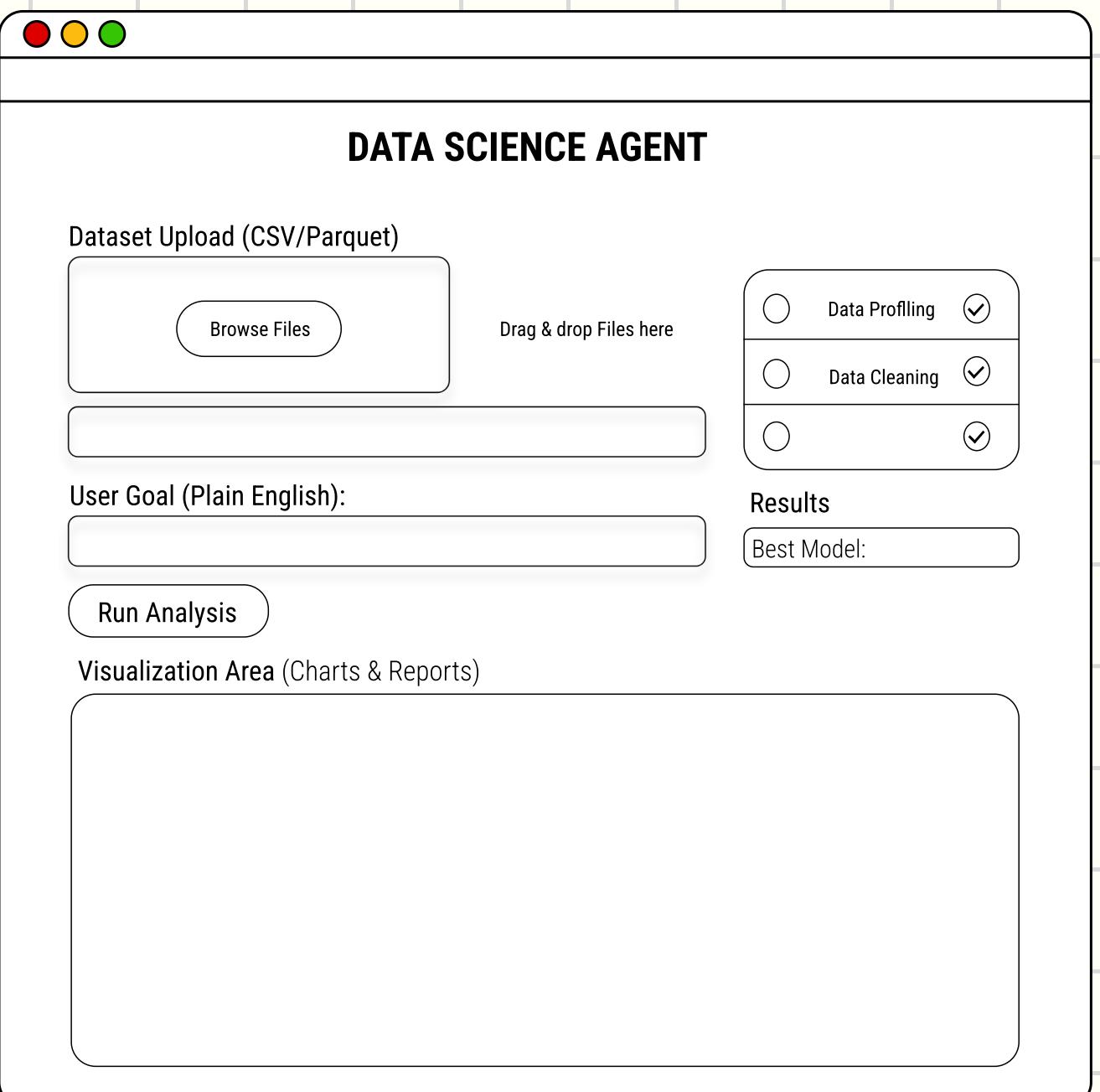


Main Web Interface (Web App)

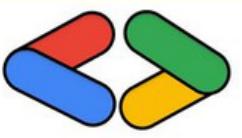


Wireframes / Mock Diagram

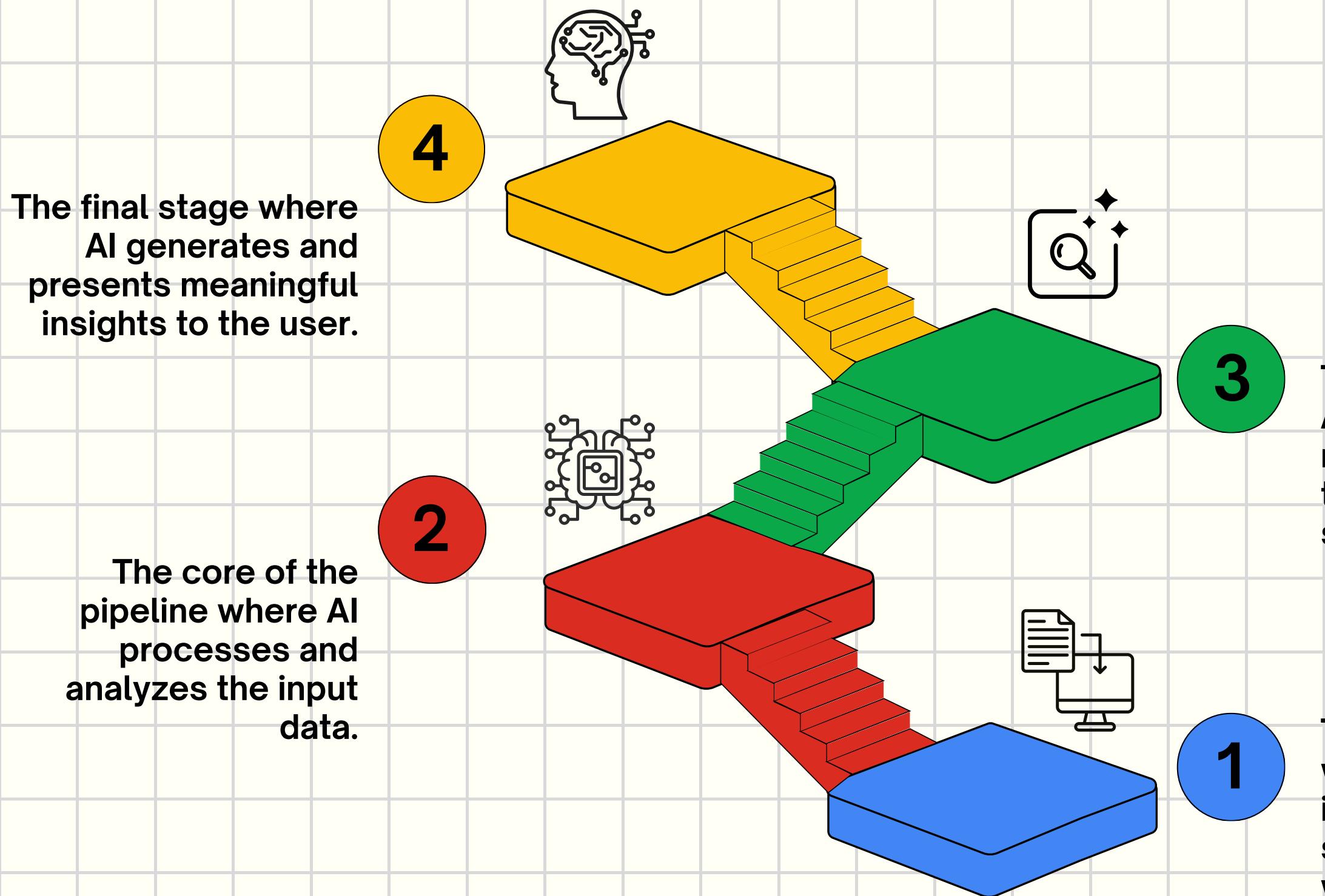
Backend AI Processing Flow (System View)

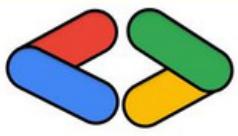


Cloud Deployment Diagram



Architecture diagram of the proposed solution





Google Developer Groups
On Campus • MITS-DU Gwalior



DATA SCIENCE AGENT

Launch Console

AUTONOMOUS AI FOR DATA SCIENCE

DATA SCIENCE AGENT

Autonomous AI for End-to-End ML

Upload your data. Describe your goal. Let AI handle profiling, modeling, visualization, and strategic insights autonomously.

Chat Now

DATA SCIENCE AGENT

Launch Console

How it Works

From raw data to actionable intelligence in 4 steps.

- 01 Ingest Data**
Upload your raw CSV, JSON, or Parquet files directly to the secure environment.
- 02 Define Objective**
Describe what you want to achieve in natural language. 'Predict churn' or 'Find outliers'.
- 03 Agent Execution**
The agent orchestrates tools to clean, transform, and model your data autonomously.
- 04 Receive Assets**
Get fully trained models, performance metrics, and interactive explainable reports.

Ready to automate your workflow?

Build smarter ML workflows with AI autonomy. Join the next generation of data scientists.

Data Profiling Report

Data Profiling Report

Overview Variables Interactions Correlations Missing values Sample

Overview

Brought to you by YData

| Overview | Alerts 34 | Reproduction |
|-------------------------------|-----------|--------------|
| Dataset statistics | | |
| Number of variables | 22 | |
| Number of observations | 175947 | |
| Missing cells | 451789 | |
| Missing cells (%) | 11.7% | |
| Duplicate rows | 0 | |
| Duplicate rows (%) | 0.0% | |
| Total size in memory | 118.0 MiB | |
| Average record size in memory | 703.3 B | |
| Variable types | | |
| DateTime | 2 | |
| Numeric | 12 | |
| Categorical | 4 | |
| Text | 4 | |

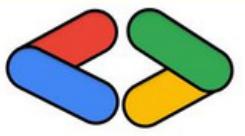
Snapshots
of the MVP

DATA SCIENCE AGENT

Powerful Orchestration

Not just a chatbot, but a true system of intelligence.

- Autonomous ML Pipelines**
End-to-end automation from profiling to deployment without manual coding.
- 82+ Specialized Tools**
An extensive arsenal for cleaning, statistical testing, and predictive modeling.
- Dual LLM Intelligence**
Orchestrated by Groq (for speed) and Gemini (for deep reasoning).
- Session Memory**
Maintains context across complex workflows, allowing for iterative refinement.
- Visual Insights**
Automatic generation of publication-quality charts and explainability reports.
- Cloud Run Ready**
Deploy your optimized models directly to production-grade cloud environments.



Future Development



Kaggle Integration

Real-time performance, data drift, and model health tracking



Model Registry

Versioned model storage with performance tracking



Vertex AI Integration

Scalable training, experiment tracking, and MLOps pipelines



Team Collaboration

Shared workspaces for teams to build and review models together



Automated Retraining

Periodic model retraining using new data and drift detection



Multi-User Authentication

Support secure login, user roles, and project isolation

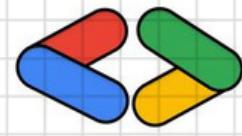
Links of Project:

GitHub Public Repository : <https://github.com/Pulastya-B/DevSprint-Data-Science-Agent>

Demo Video Link (3 Minutes):

https://drive.google.com/drive/u/0/folders/18bfEhsLfah9mZaSST6XVbvH0YI_t434Q

MVP Link: <https://data-science-agent-cezn.onrender.com>



Google Developer Groups
On Campus - MITS-DU Gwalior



Thank you!

