Datagrep

datagrep

Project Purposes and Goals

- Our Goal: Empower non-technical users to query a database, with an agent creating a data pipeline to answer.
- Datagrep Alignment: This aligns with Datagrep's mission to convert plain-English requests into data pipelines, drastically reducing time-to-insight and ensuring auditability and repeatability.
- Datagrep Modes:
  - Temporary Pipelines: For quick, one-off data extraction.
  - Permanent Pipelines: Can be built from scratch and exported to BI tools like Power BI and Tableau.
- Agent Capabilities: The agent manages intent mapping, joins, filters, transformations, schema inference, and data profiling.
- Agent Guardrails: Includes lineage, scheduling, and quality checks.

Important Names

Kaaustaaub Shankar (shankaks@mail.uc.edu)

Jay Kothari (kotharjy@mail.uc.edu)

Dhiren Mahajan (mahajadn@mail.uc.edu)

Advised by Bo Brunton, Head of Product Strategy @ Pantomath

Project Abstract

Abstract:

We propose Datagrep, an AI helper that turns **plain-English requests** from analysts and business users into ready data pipelines. It works in two modes:
(1) **temporary pipelines** to quickly pull data for one-off needs, and
(2) **permanent pipelines** built from scratch and exported to popular **BI tools** like Power BI and Tableau.

The agent maps intent to sources, joins, filters, and transforms, does schema inference and data profiling, and adds simple guardrails like lineage, scheduling, and quality checks. By combining N**L-to-query** and **NL-to-pipeline,** Datagrep cuts time-to-insight from weeks to minutes while keeping work auditable and repeatable.

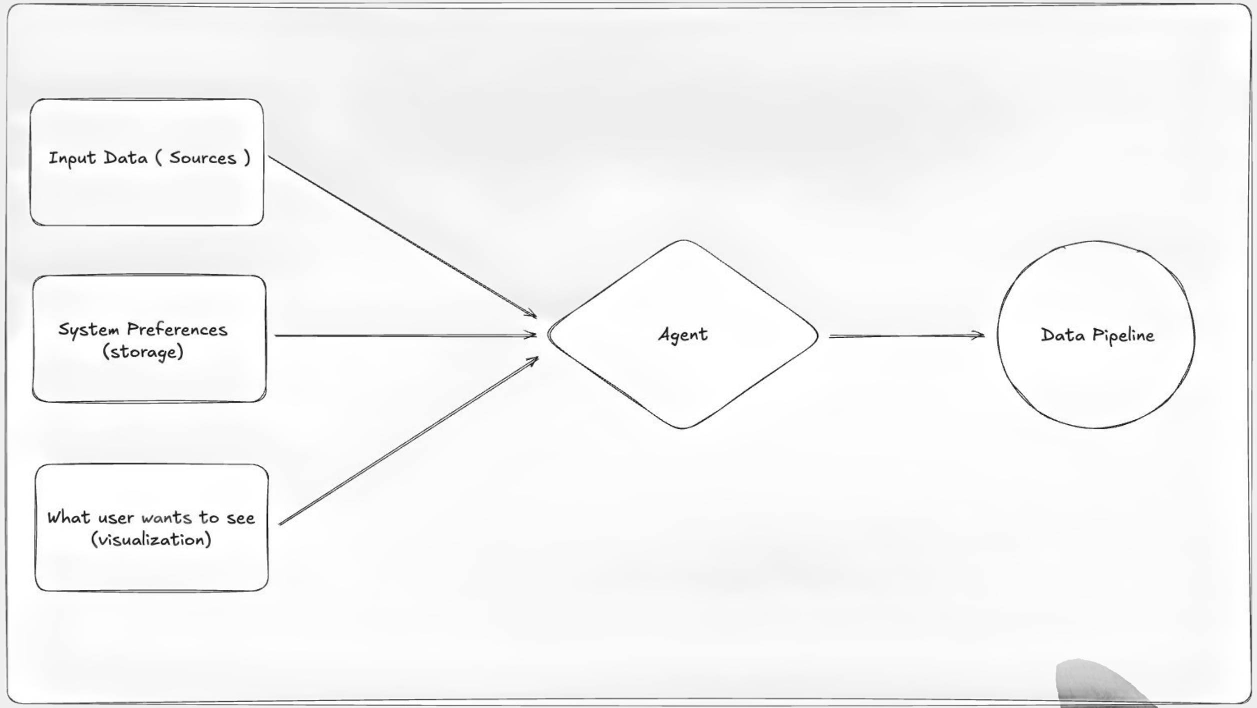
Keywords:
NL2Pipeline, plain-English to pipeline, analyst self-service, citizen developer, ad-hoc/temporary pipelines, BI export, semantic layer, lineage, governance, scheduling, observability.

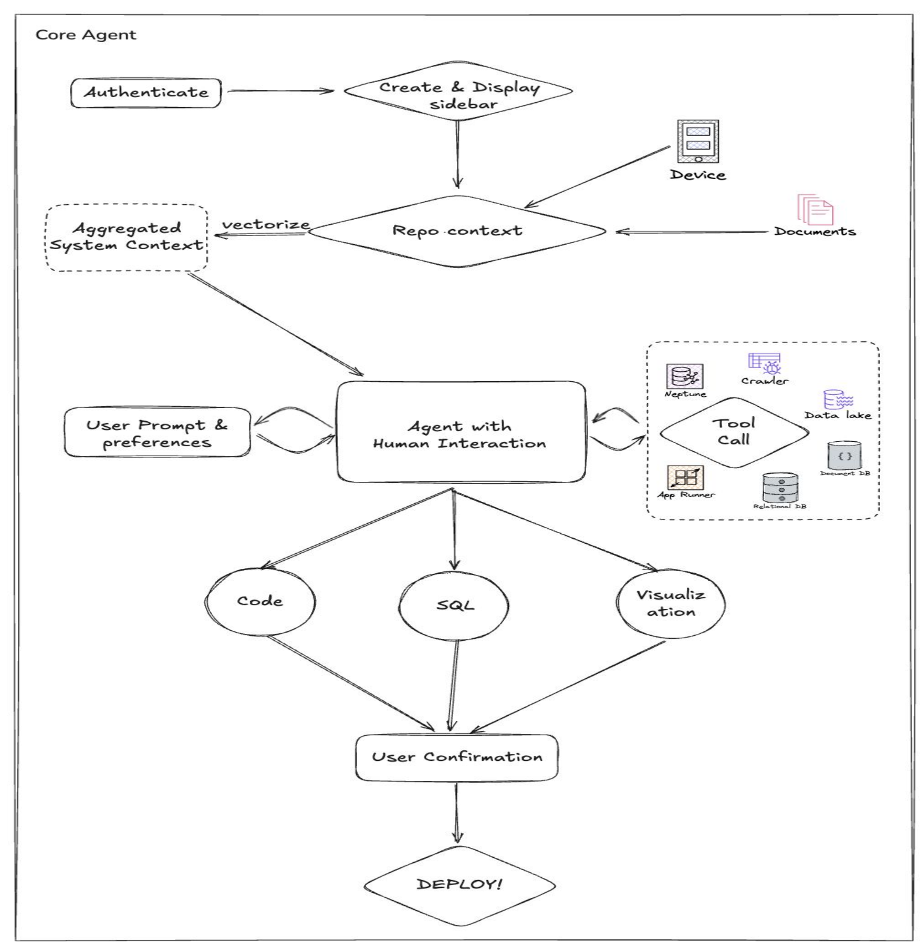
User Stories

The Head of Sales Perspective
As the Head of Sales, I want my analysts to be able to connect data from our various sales tools, so I can get a complete view of my team's performance without waiting on a long technical queue. This is critical for our GTM (go-to-market) strategy, as it allows us to perfect our PLG (product-led growth) model. By getting immediate data on our top users—seeing their activity and expansion potential—my team can focus its efforts on the accounts most likely to convert and generate significant revenue.

The Head of Data Perspective
As a head of data, I want to empower my non-technical stakeholders to fulfill their own data requests through a simple interface, so that I can clear our team's backlog of tasks and focus my expensive engineering talent on high-impact, strategic data architecture projects.

Design Diagram


Input Data (Sources)
System Preferences (storage)
What user wants to see (visualization)
Agent
Data Pipeline


Core Agent
Authenticate
Create & Display sidebar
Device
Aggregated System Context
vectorize
Repo context
Documents
Neptune
Crawler
Data lake
Tool Call
Document DB
App Runner
Relational DB
User Prompt & preferences
Agent with Human Interaction
Code
SQL
Visualiz ation
User Confirmation
DEPLOY!

Project Constraints

- **Economic:**
  a. **Model Costs:** Utilizing larger, more powerful AI models (like GPT-5-Thinking, Gemini-2.5-Pro) is cost-prohibitive.
  b. **Strategic Model Selection:** We'll employ a mix of AI models (GPT-4o, 4o-mini) based on the complexity of the data pipeline generation task.
  c. **Cost-Effective Tools:** We're leveraging free or student-provided cloud services (Azure for Students, AWS) for non-AI components.
- **Legal:**
  a. **Data Source Safety:** We exclusively use synthetically generated data or publicly licensed datasets to avoid IP rights, privacy, and ownership issues.
- **Ethical:**
  a. **Bias Mitigation:** We employ anonymized data and public datasets to minimize potential biases in the AI's output.
  b. **Transparency & Disclaimers:** Clear disclaimers will be provided to acknowledge the AI's limitations and ensure users understand it's a tool, not a replacement for human judgment.

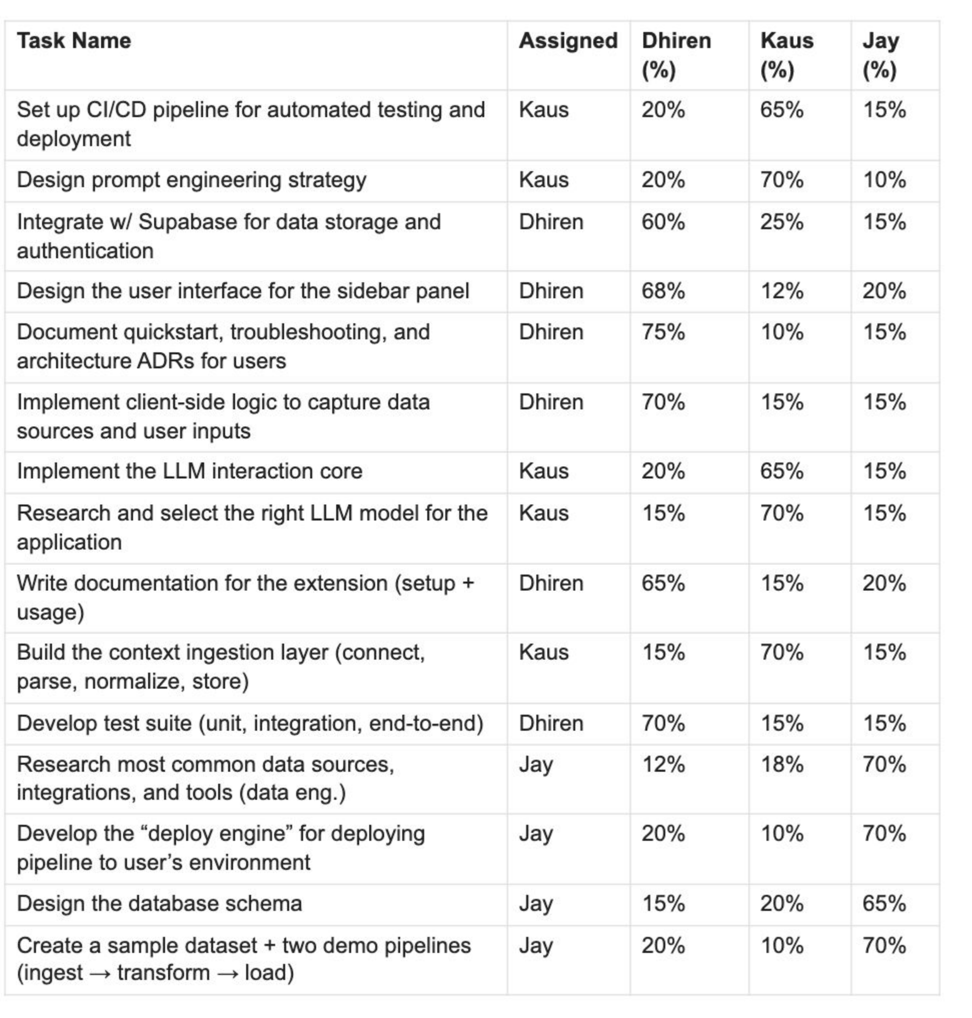
Project Progress

- Design Diagrams
- User Stories
- Preliminary Market Research and Product Market Fit
- Code-Storming

Expected Accomplishments

- Design Specifications
- Market Research
- Project Scaffold
- UI wireframes

Division of Work



Task Name	Assigned	Dhiren (%)	Kaus (%)	Jay (%)
Set up CI/CD pipeline for automated testing and deployment	Kaus	20%	65%	15%
Design prompt engineering strategy	Kaus	20%	70%	10%
Integrate w/ Supabase for data storage and authentication	Dhiren	60%	25%	15%
Design the user interface for the sidebar panel	Dhiren	68%	12%	20%
Document quickstart, troubleshooting, and architecture ADRs for users	Dhiren	75%	10%	15%
Implement client-side logic to capture data sources and user inputs	Dhiren	70%	15%	15%
Implement the LLM interaction core	Kaus	20%	65%	15%
Research and select the right LLM model for the application	Kaus	15%	70%	15%
Write documentation for the extension (setup + usage)	Dhiren	65%	15%	20%
Build the context ingestion layer (connect, parse, normalize, store)	Kaus	15%	70%	15%
Develop test suite (unit, integration, end-to-end)	Dhiren	70%	15%	15%
Research most common data sources, integrations, and tools (data eng.)	Jay	12%	18%	70%
Develop the “deploy engine” for deploying pipeline to user’s environment	Jay	20%	10%	70%
Design the database schema	Jay	15%	20%	65%
Create a sample dataset + two demo pipelines (ingest → transform → load)	Jay	20%	10%	70%

Vision for Demo at Expo

A visitor stops by our booth, curious about a dataset but doesn’t know how to code. With **Datagrep**, they simply type a question in a chat-like interface like *“Which products sold best this quarter?”* and instantly get insights, charts, or tables.

No code. No dashboards. Just a conversation that turns data into answers, making exploration simple and approachable for anyone.