AskDataAI: Democratizing Enterprise Data Access

AskDataAI aspires to be the platform that provides AI-powered data assistance at the data organizations and democratizes access to data across technical and non-technical teams.

This presentation explores how context-aware AI agents revolutionize enterprise data access.

By: Praveen Payili

Introducing AskDataAI

AI-Powered Conversational Interface

Natural language queries for intuitive data access.

Context-Aware Agents

Specialized AI agents understand diverse data contexts.

Enterprise-Grade Security

Role-based access control ensures data protection.

Revolutionize Your Data Strategy

1 **Assess Current Data Landscape**

Identify pain points and opportunities in your organization.

2 **Explore AI-Powered Solutions**

Consider implementing context-aware AI agents for data access.

3 **Prioritize Security and Usability**

Balance data democratization with robust access controls.

4 **Embrace Continuous Innovation**

Stay updated on AI advancements to evolve your data strategy.

Technical Architecture Deep Dive

1

Qdrant Vector Search

Powers efficient similarity-based data retrieval.

2

Hybrid Search Model

Combines vector and traditional search for 20% accuracy boost.

3

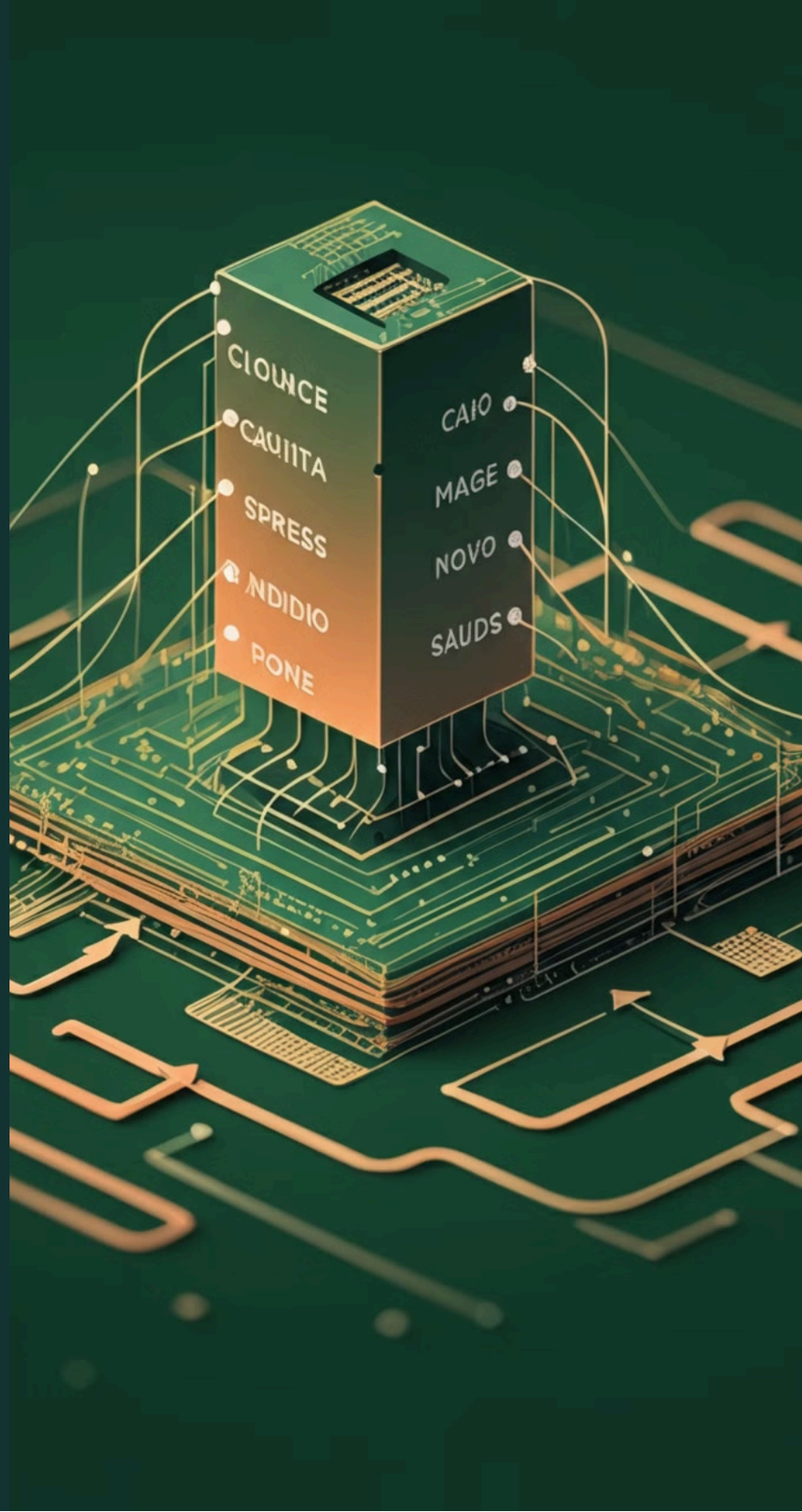
Reciprocal Rank Fusion (RRF)

Re-ranks results, improving relevance by 40%.

4

GPT-4-o Integration

Latest language model reduces query latency by 35%.



Key Platform Features

● Enhanced AI Platform

◯ **Qdrant Vector Search:** Qdrant was deployed as a scalable vector search engine, enhancing overall search capabilities and enabling hybrid and multi-stage search.

◯ **Advanced Filtering for JSON Collections:** Support was added for advanced filtering using Qdrant Query Language. Enabled targeted filtering, such as including only finance-related datasets for AskFinance from global tables.

◯ **Popularity & Certification-based Re-ranking:** Implemented re-ranking of search results using the Reciprocal Rank Fusion (RRF) approach. Prioritized results based on key indicators like page popularity and certification.

◯ **Semantic datasets collection:** AskDataExplorer generates data insights based on relevance, table popularity, and Hubble certification status. Table search can be applied to any agent for efficient table suggestions.

◯ **Latency Optimization:** Reduced latency through server and code optimizations and upgraded to the GPT-4-o model for faster response times.

◯ **Hybrid Search Model Evaluation and Integration**: Evaluated and integrated a hybrid search model that increased accuracy by 20% across multiple agents, including AskFinance, AskFire, and AskDataExplorer.

● Scalable Backend, AI chatbots and Batch Mode

◯ **New Service:** Deployed new scalable backend service in Supercell environment, gRPC integration, Secured security and privacy approvals for OpenAI (Deprecated DataDeck service used for MVP)

◯ **Integrations:** Slack integration, Kafka integration, Redis, Bedrock, CRDB credentials using CertManager, Trino, Snowflake.

◯ **Faster Development:** Decoupled BFF and UI components, enabling faster development and deployment cycles

◯ **Real-time Feedback collection:** Enabled real-time feedback collection to enhance service response accuracy(thumbs up/down)

◯ **Slack Bot concurrency:** Implemented a robust Slack bot with load-balanced workers to handle concurrent requests

◯ **Multiple slack interaction modes:** Slack bot supports multiple interaction modes such as in-channel replies, threaded responses, @mention handling, and workflows

◯ **Enhanced Logging:** Implemented correlation ID and user ID propagation to improve traceability and logging, facilitating efficient debugging and audit processes.

Key Platform Features

Improved Accuracy and latency

◯ **Integration testing:** Enhance the platform's stability and user experience.

◯ **Accuracy evaluation:** Added automated evaluation mechanisms Improved accuracy by - 20% and latency by -22% with prompt engineering.

◯ **Caching Prompts & responses:** Significantly improved latency (in subseconds) with the Redis cache and reduced LLM costs for repeated prompts. Performed semantic cache eval for the hybrid cache.

◯ **Traceability**: Revamped logging and created a thread-safe logging mechanism throughout the application to monitor, and debug all requests with (id, user) mapping with full traceability

◯ **Guardrails:** Limited chat history & the context provided to the LLM agent to a maximum of 8 conversation turns or until the token limit is reached, which helps reduce latency and LLM costs.

● Intuitive User-friendly Interface

◯ **BFF Implementation:** Implemented changes to support the new ask-data-ai-grpc-gateway service on the.

◯ **UI Enhancements:** Designed the UI and created UI in Web Next with more than 75% of unit test coverage. Supports Conversational Chat with rich markdown support and works seamlessly across different screen sizes

◯ **Event logging:** Logged all the events for prompts, user feedback, and SQLs for better tracking and sentiment analysis, traceability, and debugging.

◯ **Usability Testing:** Exercise to observe beta users as they interact with the UI to identify where they face difficulties. Implemented enhancements based on the feedback.

Impressive Results

60%

Efficiency Boost

Increase in data discovery efficiency.

20%

Accuracy Gain

Improvement in query accuracy across agents.

35%

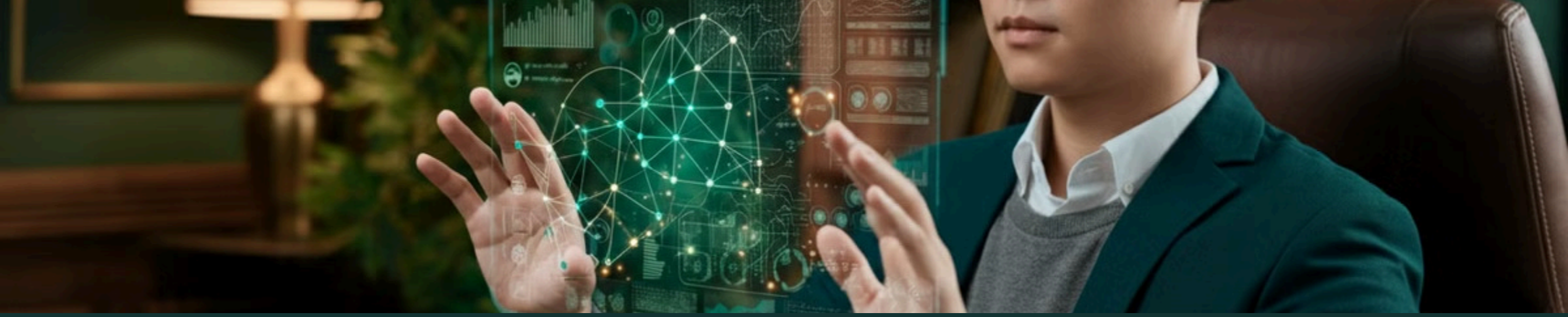
Latency Reduction

Decrease in query response time.

99%

Security Enhancement

Reduction in unauthorized access attempts.



Enhancing Data Discovery

Semantic Dataset Collection

Automatically analyzes, tags, and organizes datasets using advanced ML algorithms for intuitive navigation and discovery.

Popularity-Based Ranking

Dynamically ranks datasets based on user engagement metrics, query frequency, and collaborative feedback scores.

Certification-Based Re-Ranking

Elevates trusted data sources through rigorous quality checks and domain expert verification processes.

The Challenge: Data Access vs. Security

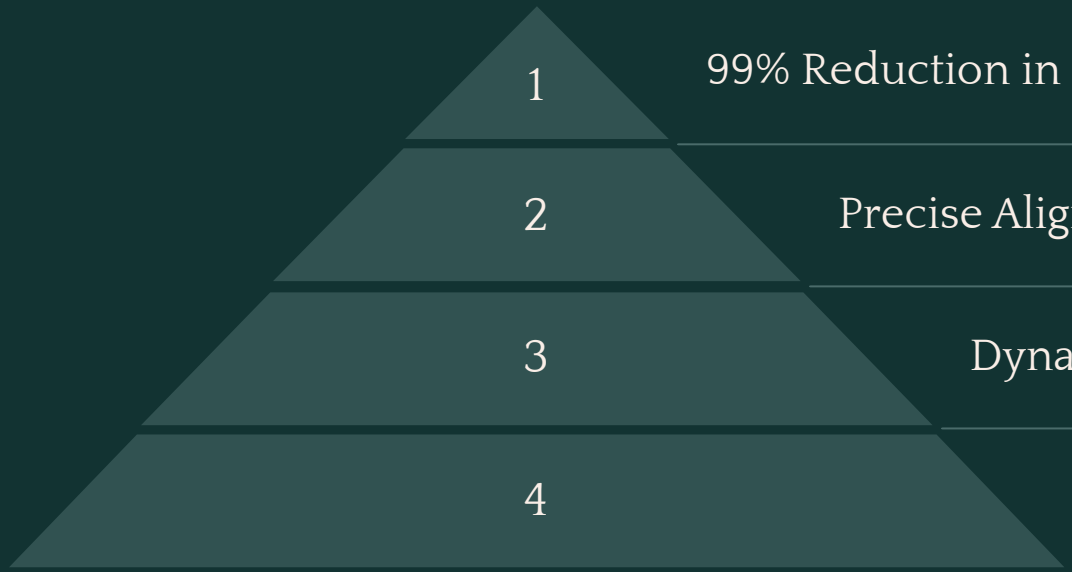
Data-Driven Decisions

Modern enterprises require seamless, real-time access to data insights for agile decision-making. Teams across departments need the ability to query and analyze data independently, without creating bottlenecks or depending on technical specialists.

Security Concerns

While democratizing data access drives innovation, organizations must carefully balance accessibility with robust security protocols. Protecting sensitive information, maintaining compliance, and preventing unauthorized access remain paramount challenges in today's data landscape.

Security: Raven Role-Based Access Control



99% Reduction in Unauthorized Access

Precise Alignment with IT Roles

Dynamic Permission Updates

Granular Access Controls

Overcoming Implementation Challenges

1

Data Integration Complexity

Unified disparate data sources across ecosystem.

2

Scalability Concerns

Optimized architecture to handle 15,000+ concurrent users.

3

AI Model Fine-Tuning

Customized GPT-4-o for specific data contexts.

4

User Adoption

Implemented intuitive UI and comprehensive training programs.

Key Lessons Learned

Balance is Key

Democratizing data access while maintaining robust security is crucial.

Context Matters

AI agents must understand enterprise-specific data contexts for accuracy.

Continuous Improvement

Regular model updates and user feedback loops drive platform evolution.

User-Centric Design

Intuitive interfaces and clear documentation boost adoption rates.

Future Directions

1

Multi-Modal Data Integration

Expanding to handle images, video, and audio data.

2

Cross-Platform Synergy

Integrating with other enterprise tools and workflows.

3

Advanced Predictive Analytics

Implementing AI-driven forecasting and trend analysis.

4

Global Language Support

Expanding to support queries in multiple languages.

Thank You