BI WORLDWIDE

Democratising Data:

How Generative AI Empowers Non-Technical Teams with Insights

Table of Contents

Executive Summary

Gen AI, powered by Azure OpenAI and BERT, is revolutionising the way modern businesses engage and interact with data, turning it into a strategic advantage.

By enabling natural language interaction, Gen AI empowers non-technical teams to query, analyse and visualise data seamlessly — eliminating technical bottlenecks that slow decision-making. With AI-led automation at the core, organisations can accelerate insight generation, reduce dependency on specialists and drive smarter business outcomes, at scale.

- Covert natural language queries into SQL for intuitive data exploration
- Connect to multiple databases and run complex queries without manual intervention
- Automate data retrieval, processing and visualisation for faster decision-making
- Deliver real-time actionable insights for business agility
- Bridge the gap between raw data and business intelligence for operational efficiency
- Scale seamlessly as data volumes grow, ensuring sustainability and cost-effectiveness

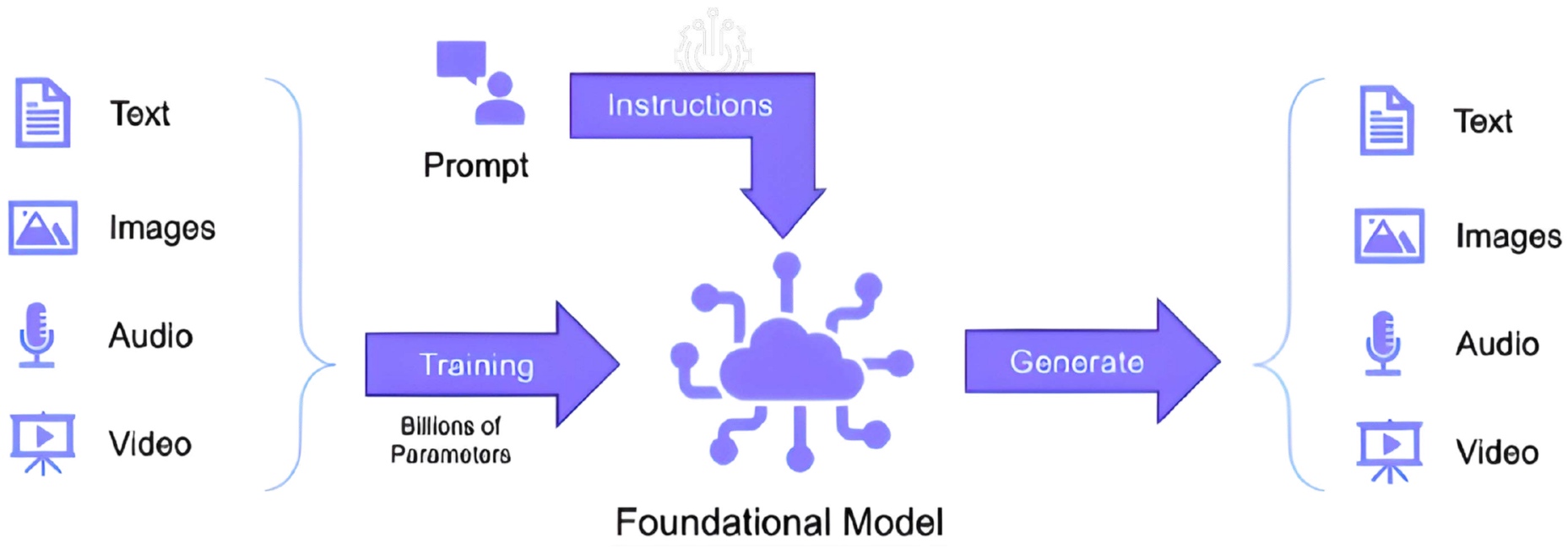
The AI-powered approach helps unlock the full potential of data, strengthening collaboration across teams, fuelling innovation and achieving cost-efficiency. This white paper outlines the strategy, architecture and technical considerations to build effective Gen AI solutions, providing a roadmap on how organisations can democratise data across teams. Simultaneously, it presents real-world scenarios for harnessing the power of Gen AI in BI WORLDWIDE's channel engagement solution.

Ready to deep dive? Let's start with what Gen AI truly is and how it's transforming the data landscape for business teams.

Introduction to Gen AI

So, what is Gen AI? Let's break down its core capabilities and the challenges organisations face when deploying such advanced technologies.

Gen AI refers to a set of advanced AI techniques, capable of creating new content derived from the training data they have been fed with, combined with additional user-provided context. This content can encompass text, codes, images, audio and video.



At its core, Gen AI relies on **Large Language Models (LLMs)**, trained using diverse, often publicly available datasets. Application users provide prompts or instructions to these models to generate output in various formats (text, images, audio or video), depending on the specific model being deployed.

Challenges of Using Gen AI

The potential of Gen AI is huge; however, it poses critical challenges:

➜ Quality & Reliability

LLMs tend to hallucinate, so ensuring the quality, accuracy and reliability of outputs is essential.

➜ Ethical & Societal Risks

Gen AI raises ethical considerations, such as the creation of deepfakes, leading to privacy concerns.

Computational Costs & Environmental Impact

The high computational costs and environmental footprint of Gen AI (energy consumption for image generation equivalent to charging a phone), must be considered.

Intellectual Property Concerns

Ownership of AI-generated content is an unresolved legal debate. Copyright of the content should be determined and training of the models using copyrighted material should be avoided.

Governance

Appropriate frameworks for the development and deployment of Gen AI technologies are essential. Concerns around accuracy (most recent information needs to be available for meaningful answers) and privacy (properly tagging data as internal, confidential, sensitive or subject to privacy regulations) must be addressed.

Providing Custom Context and Private Data

Apart from the above challenges, businesses also often struggle to access domain-specific insights, as foundational models are typically trained on publicly available content.

Now let's dive deeper into this – exploring customisation strategies, from simple prompts to fine-tuned models.

Diverse approaches exist to provide custom context to the models, enlisted below by increasing level of difficulty (development effort, AI skills, computing costs and hardware needs):

Prompt Engineering

The simplest method – it involves providing specific instructions to the model, that can be guided by prompt templates. The approach is highly flexible for adapting the LLM and prompt templates, and ideal for use cases requiring minimal domain context.

Retrieval Augmented Generation

RAG enhances output quality by providing the particular context for response generation, leveraging proprietary, company-owned data. It reduces hallucinations, maintains high flexibility (to change data sources, embeddings, LLM, vector database) and enables access control.

Fine-tuning

This incorporates more context into the foundational model by adjusting parameters for a specific industry or use case. It's useful for specialised fields but prone to hallucination, bias from flawed training data entry and lacks access control.

Training a Custom Foundational Model

The most advanced path – this enables a high customisation but requires trillions of curated tokenised datasets, sophisticated hardware infrastructure, expert ML teams and substantial budgets.

Ready to take a closer look? Let's dive now into Retrieval-Augmented Generation pipelines and their role in modern Gen AI solutions.

RAG Pipelines: A Detailed Overview

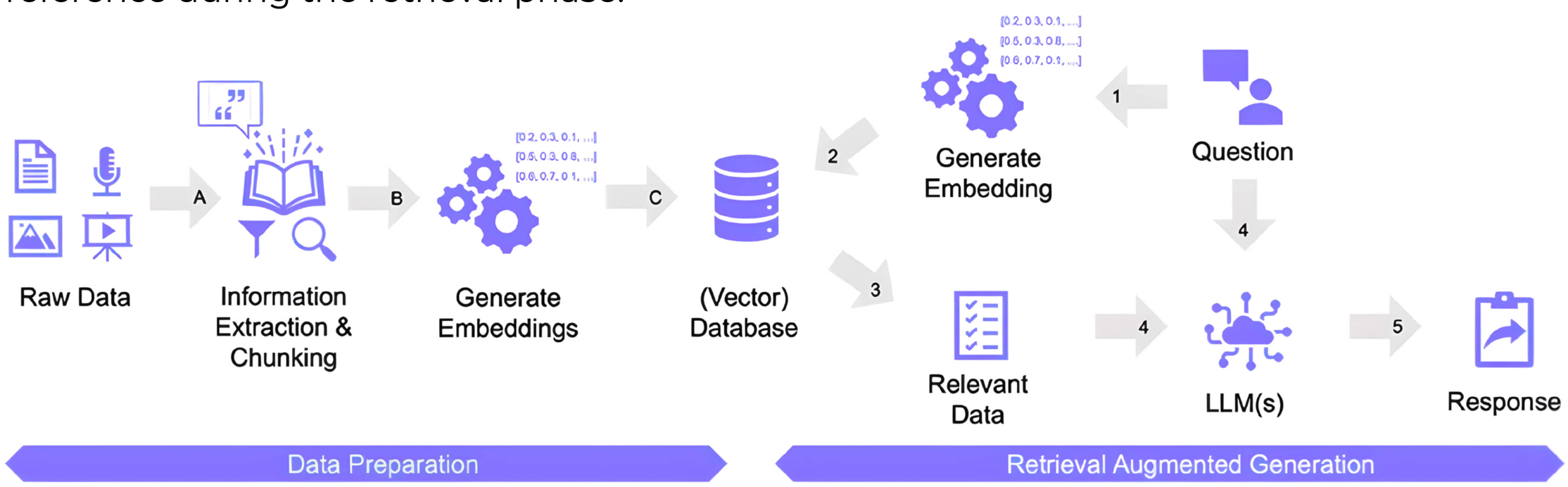
RAG pipelines are pivotal in modern Gen AI. Essentially, they involve a streamlined two-phase process: data preparation and data retrieval.

✳ Phase 1: Data Preparation

In this phase, raw data — text, audio or other formats — is extracted and segmented into smaller, manageable chunks. These chunks are converted into embeddings and stored in a vector database, alongside their metadata. Keeping metadata linked to the chunks ensures that each piece can always be traced back to its original source for accurate reference during the retrieval phase.

✳ Phase 2: Data Retrieval

This phase is initiated by a user prompt or question, that gets converted into an embedding and matched against the stored database to find the most relevant chunks. The chunks serve as context, along with the original query, for the Large Language Model (LLM) to generate a precise, well-informed response.

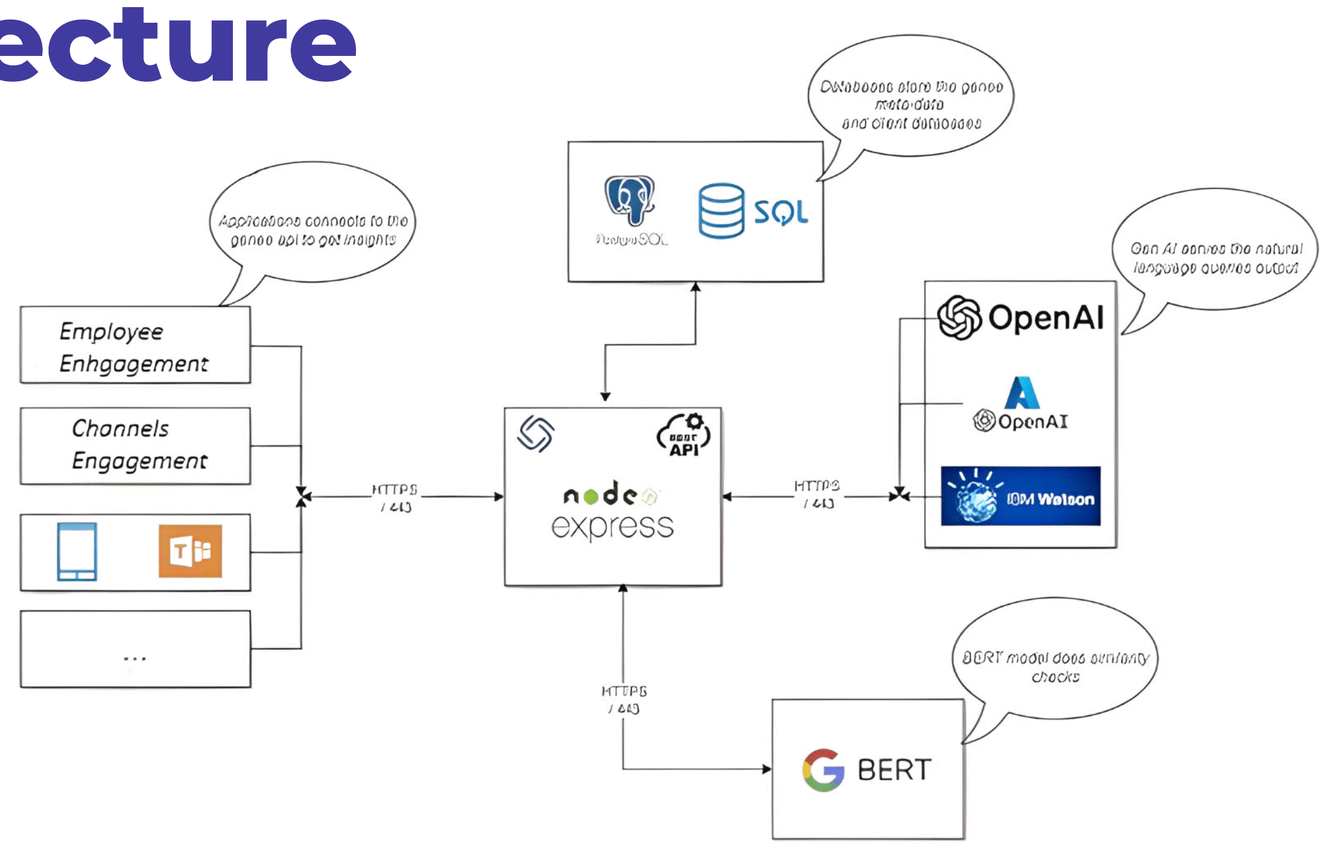


Structure of a RAG Pipeline

While this is a simplified overview, real-world implementation is far more nuanced. Key considerations include how to effectively chunk and extract information from sources like PDFs or documentation, as well as how to define and measure relevance when re-ranking results.

Wondering how this fits into the bigger picture? Here's an overview of the full solution architecture and each technology's contribution.

Solution Overview and Architecture



The solution integrates cutting-edge technologies to deliver a seamless natural language-to-SQL experience for business users. Here's how each component contributes to the platform's functionality:

✳ Node.js with Express (API Gateway and Orchestrator)

The Node.js Express server acts as the central orchestrator. It securely handles API requests from user interfaces and coordinates all downstream interactions between the database, AI models and response delivery.

Role: Translates front-end queries into backend actions.
Business Impact: Ensures fast, reliable and scalable request handling across multiple client apps.

✳ OpenAI / IBM Watson (Natural Language Processing Engine)

We use Gen AI models from OpenAI and IBM Watson to interpret user intent in natural language and convert it into executable SQL queries.

Role: Converts questions like "Show me last month's revenue" into syntactically accurate SQL.

Business Impact: Eliminates the need for SQL knowledge, enabling business users to query data on their own.

✳ PostgreSQL / SQL Database (Data Storage and Processing)

PostgreSQL databases serve as the central data repository, maintaining both metadata and client-specific datasets, while executing AI-generated SQL queries to retrieve or manipulate data.

Role: Executes the actual database queries generated by AI.
Business Impact: Leverages existing enterprise data securely and efficiently to support real-time decision-making.

✳ Google BERT (Semantic Embedding and Similarity Engine)

BERT powers semantic understanding by converting user queries into vector embeddings, capturing the context and meaning of queries, just beyond keywords.

Role: Enables semantic search, query duplication and personalised suggestions by comparing similarity between current and past queries.

Business Impact: Elevates user experience by delivering faster, more relevant results and supports intelligent caching for performance optimisation.

✳ Client Applications (BI WORLDWIDE's Employee & Channel Engagement Apps, Microsoft Teams and More)

Multiple client applications connect to the system via secure HTTPS endpoints to submit user queries and view results.

Role: Front-end interface for end users.
Business Impact: Integrates with existing tools used by teams, enabling analytics without disrupting workflows.

However, the big question arises – How do we keep the outputs reliable? Let's now examine how hallucinations are handled and prevented when generating SQL queries with AI.

Handling Hallucination in AI-Powered Query Generation

One of the biggest challenges in implementing Large Language Models (LLMs) like OpenAI's GPT for business data applications is **hallucination** — where the model generates plausible-sounding but flawed or fabricated outputs. In natural language-to-SQL translation, this can lead to invalid queries, unreliable insights or misleading interpretations of business performance.

However, BI WORLDWIDE's platform is purpose-built to minimise hallucinations through deliberate architectural design.

Use of Metadata-Only Inputs for LLMs

Instead of exposing raw or sensitive business data to external models, we restrict the input scope strictly to **metadata** — such as:

Table names	Column names and data types
Table relationships (foreign keys, joins)	Indexes and schema structures

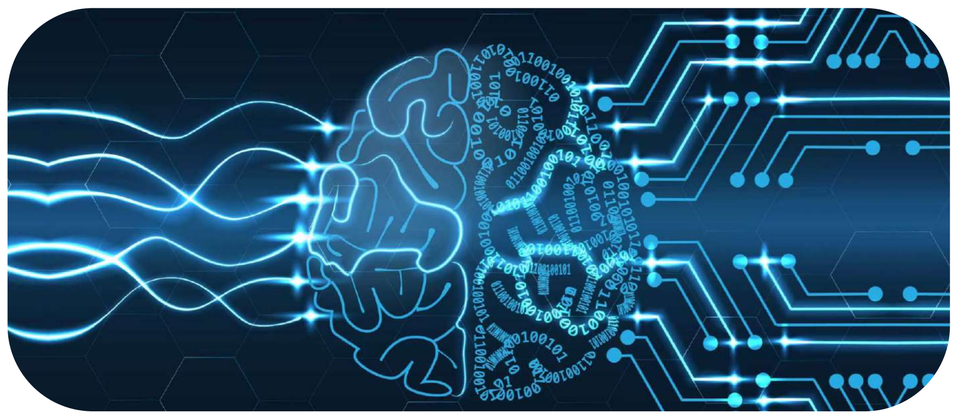
By constraining the LLM's context to well-defined metadata, the model generates queries grounded in the actual database structure, reducing the possibility of fabricating unknown tables or fields.

Delegation of Query Interpretation to Local BERT Model

The bulk of query interpretation is powered by a **locally deployed BERT-based model**, significantly reducing reliance on external generative systems. This model is responsible for:

- Interpreting user queries semantically
- Matching similar historical queries using vector embeddings
- Ranking query intent candidates before generation

As BERT operates deterministically within a controlled domain, it provides predictable, reproducible interpretations, minimising the creativity causing hallucinations in open-domain models.



Controlled Usage of LLMs for SQL Syntax Generation

Generative AI (OpenAI / IBM Watson) is leveraged **only to generate SQL syntax** after completing semantic interpretation. The prompt context provided is limited to:

- A clean description of what the user wants (parsed by BERT)
- The relevant table schema and metadata
- Strict prompt formatting templates

By constraining the input to a limited, well-structured space, we effectively reduce the scope for LLM hallucinations. The model is not tasked with inferring data logic or making assumptions—it simply produces valid SQL, based on a predefined structure.

Query Validation and Fallback Mechanisms

Every SQL query generated is rigorously passed through a **validation engine** before execution:

- Syntax validation
- Schema conformity checks
- Dry-run parsing against the actual database (without execution)

If the generated query fails validation or doesn't match schema expectations, the system automatically falls back to either:

- A refined re-prompt using metadata context
- A BERT-suggested previously successful similar query

The process ensures that users only receive executable and semantically valid SQL.

Continuous Learning from User Feedback

User interactions with generated queries (for instance, edits, approvals or rejections) are monitored and fed back into the system for reinforcement:

- Frequently accepted queries are prioritised for similar future inputs
- Hallucinated outputs are flagged and excluded from future generation paths
- Domain-specific prompt tuning is continuously improved

Outcome: Reliable, Accurate and Trustworthy SQL Generation

By design, BI WORLDWIDE's system delivers natural language-to-SQL conversion with exceptional accuracy, consistency and transparency. Anchoring query generation in structured metadata and validated history — while minimising reliance on probabilistic inference — enables us to eliminate hallucinations in real-world scenarios. The result is a platform that businesses can trust for reliable, high-stakes decision-making.

Want to understand how this works in action? Let's dive into technical implementation – from user apps to backend processing and semantic indexing.

Technical Implementation

The solution empowers business users to query enterprise data effortlessly using natural language. By integrating RESTful APIs, Gen AI, semantic embedding models and a relational database, it delivers accurate and actionable data insights.

User Interaction and API Invocation

- The user opens a connected app like Microsoft Teams or BI WORLDWIDE's Channel/Employee Engagement apps.
- They enter a natural language query, such as: "What were our top 5 products by revenue last quarter?"
- The application sends an **HTTPS POST request** to the Genee API (hosted on Node.js/Express), along with the query and relevant contextual meta-data (such as user ID, tenant ID and filters).

Request Processing in Node.js/Express

- The Express server orchestrates key request-handling tasks:
  - **Validation** of incoming requests
  - **Authentication** and **authorisation,** ensuring users are allowed to query the content.
  - **Logging** of every query for audit and performance analysis.
- Once validated, the API constructs a payload and transmits it to the Gen AI layer through a new HTTPS request.

Natural Language to SQL via Generative AI

- The payload is directed to OpenAI or IBM Watson, configured with a do-main-specific prompt template, such as:

```
"Convert the following user question into a PostgreSQL query. The
table schema is: [schema details]. Question: [user query]"
```

- The AI model interprets the natural language input and returns the SQL query. For instance:

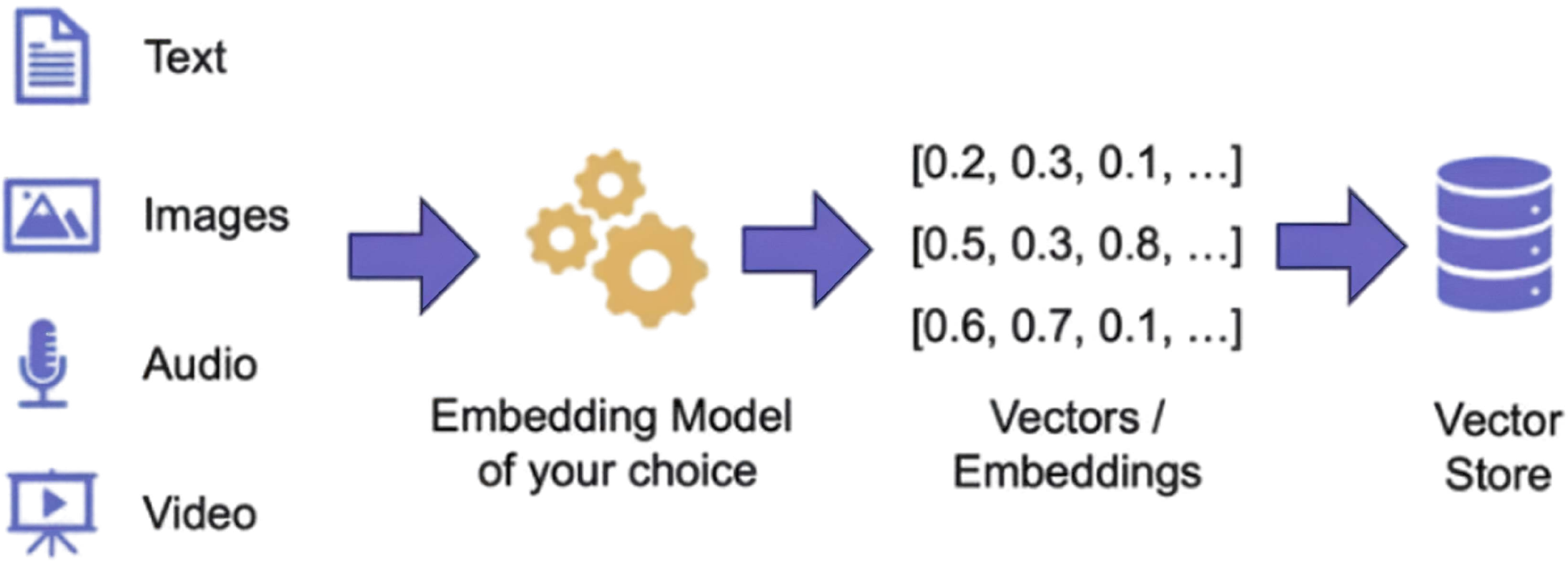
```sql
SELECT product_name, SUM(revenue) as total_revenue
FROM sales
WHERE sale_date BETWEEN '2024-01-01' AND '2024-03-31'
GROUP BY product_name
ORDER BY total_revenue DESC
LIMIT 5;
```

- Before execution, the API validates the generated SQL syntax to ensure accuracy.

Query Execution on PostgreSQL

- The validated SQL query is executed against the client's specific PostgreSQL database.
- Results are retrieved and if required, enriched with contextual metadata (such as, column names, formats and more).

Optional: Semantic Indexing with BERT



- In parallel, the original user query is directed to **Google BERT** model.
- The query is embedded into a vector representation and stored in a vector database (such as, Pinecone or FAISS), enabling:
  - Future **semantic similarity** checks
  - Query recommendations, informed by previous queries
  - Performance optimisation (for instance, caching)

```
CREATE TABLE word_embeddings (
    text STRING PRIMARY KEY,
    embedding FLOAT_VECTOR(4)
);
```

```
INSERT INTO word_embeddings (text, embedding)
VALUES
    ('Exploring the cosmos', [0.1, 0.5, -0.2, 0.8]),
    ('Discovering moon', [0.2, 0.4, 0.1, 0.7]),
    ('Discovering galaxies', [0.2, 0.4, 0.2, 0.9]),
    ('Sending the mission', [0.5, 0.9, -0.1, -0.7]);
```

```
| text                    | _score     |
|-------------------------|------------|
| Discovering galaxies| 0.917431   |
| Discovering moon    | 0.909090   |
| Exploring the cosmos| 0.909090   |
| Sending the mission | 0.270270   |
```

- Below is a demonstration of searching for content similar to 'Discovering Galaxy' in your table. The KNN_MATCH function needs to be leveraged, combined with a sub-select query that returns the embedding associated with 'Discovering Galaxies'.

```
SELECT *, 1 - (embedding <=> '[...]'::vector) AS similarity
FROM word_embeddings
ORDER BY similarity DESC
LIMIT 5;
```

```
| text                    | _score     |
|-------------------------|------------|
| Discovering galaxies| 1          |
| Discovering moon    | 0.952381   |
| Exploring the cosmos| 0.840336   |
| Sending the mission | 0.250626   |
```

Curious about the real-world impact of Gen AI? Let's check out some practical scenarios in BI WORLDWIDE's channel engagement solution.

Real-world Scenarios of Using Gen AI

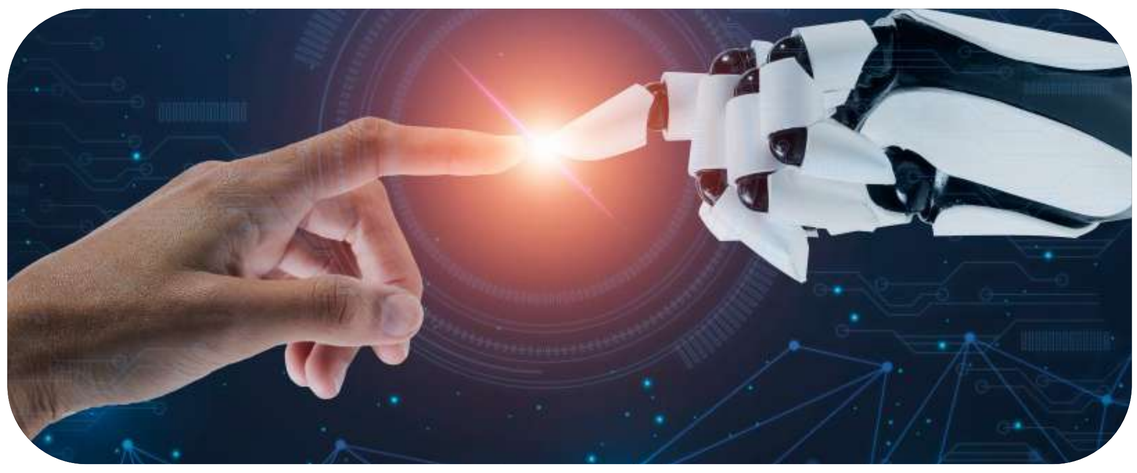
BI WORLDWIDE's Channel Engagement solution enables a versatile use of Gen AI, for tasks ranging from – reports creation to building custom dashboards, reports display and more – empowering users to transform complex data mines into strategic insights.

Here are some real-world scenarios:

- **Reports Creation** with **Reports Studio:** You can transform your reporting requirements effortlessly into SQL queries in the channel engagement solution's Reports Studio. Create natural language prompts that can be previewed, validated and stored as reusable Datahubs. You can also create preconfigured dashboards for common reporting needs or build custom dashboards with tailored filters, charts and tables for focused insights and clear data visualisation – all with seamless preview capabilities.

- **Reports Display** in **Partner Portal** and **Achievo App**: Showcase your reports seamlessly in the channel engagement solution's Partner Portal and Achievo App, that offer a streamlined and user-friendly experience. With Report Widgets, the generated reports can be embedded directly into relevant pages and menus with role-specific visibility. Moreover, managerial reports can be filtered by hierarchy for precise oversight, while standard reports ensure quick snapshots into transactional performance for smarter insights.

Now, let's talk numbers. Here is a glimpse into cost analysis and the tangible ROI organisations can expect from this approach.



Cost Analysis & ROI Evaluation

Our AI-powered solution integrates PostgreSQL with pgvector, streamlining vector embeddings storage and search, without the overhead of specialised databases. This approach ensures cost-effective, scalable and easily manageable architecture within the existing infrastructure.

Infrastructure & Hosting

PostgreSQL with pgvector on AWS RDS, Azure Database or self-managed VMs

OpenAI API Integration

Utilised for SQL generation and embedding, when not processed locally

Backend & Compute

- Built on Node.js/Express/Next.js stack
- Hosted on low-to-mid-tier cloud instances (for instance, AWS EC2, Azure App Service, etc.)

Return on Investment (ROI)

→ Empowering Business Users

Embedding AI-powered data search directly into business applications enables line managers and operational leaders to tap into data in natural language — without constantly relying on technical teams. This ensures:

- Quick assessment of team performance, customer trends and KPIs
- Faster, more confident decisions
- Minimised delays in request/response cycles with data teams

Business Impact ←

When insights are readily accessible:

- Partners and business units can take **proactive, data-driven actions**
- Organisations unlock noteworthy improvements in efficiency and responsiveness
- Key performance metrics, such as sales throughput, resolution time and engagement elevate owing to greater data visibility

➜ Improved Analyst Productivity

- Data analysts can focus on **higher-value KRAs** by offloading routine queries
- A **30–50% reduction in manual query workload** can be achieved
- This translates into saving **several dozens of hours per month** across midsized teams

➜ Optimised Query Reuse

- Vector-based search supports quick discovery of similar or past queries
- Repetitive efforts are cut down and user experience gets enhanced
- Dependence on external APIs usage is reduced, minimising costs significantly

➜ Cost-Efficient Vector Search

- Embedding vector queries within PostgreSQL eliminates the need for specialised databases
- Caching and embedding reuse minimise expensive external AI API calls
- This results in a leaner, more cost-effective infrastructure

Want to stay ahead of your competitors? Let’s now uncover the strategic business benefits and competitive edge that Gen AI solutions deliver.

Strategic Benefits & Competitive Advantages

Strategic Benefits

Seamless Integration: Built on PostgreSQL and open-source tools, our solution integrates easily into existing tech stacks and infrastructures, fast-tracking adoption and minimising disruption.

Cost-Effective AI Insights: Embedding vector search within PostgreSQL reduces costs by eliminating expensive third-party vector DBs and cutting down AI API overhead through embedding reuse.

Accelerated Innovation: AI-driven automated query generation minimises manual workload, enabling faster insights and decision-making across teams.

Scalable & Flexible: Cloud-agnostic deployment (AWS, Azure or on-premises) supports evolving business needs without vendor lock-in, driving growth.

Simplified Operations: A unified platform for both relational and vector data reduces operational overheads, simplifying security, strengthening compliance and streamlining maintenance.

Competitive Advantages

Rapid Time-to-Market: Leveraging familiar, existing tech and open-source tools enables faster rollout and quicker value delivery than competitors, relying on complex, multi-system architectures.

Advanced AI Capabilities: Automated SQL generation enables non-technical users to gain insights faster and autonomously, reducing reliance on specialised analysts and speeding workflows.

Cost Efficiency with Embedding Reuse: Intelligent caching reduces costly external AI API calls, enhancing performance and lowering ongoing operational overheads.

Unified Data Platform: Combining vector and relational queries in a single, integrated PostgreSQL engine simplifies architecture and management, against competitors relying on multi-layered stacks.

Flexible Deployment Options: Cloud-agnostic framework provides the freedom to deploy on your own terms, avoiding vendor lock-in, common with proprietary solutions.

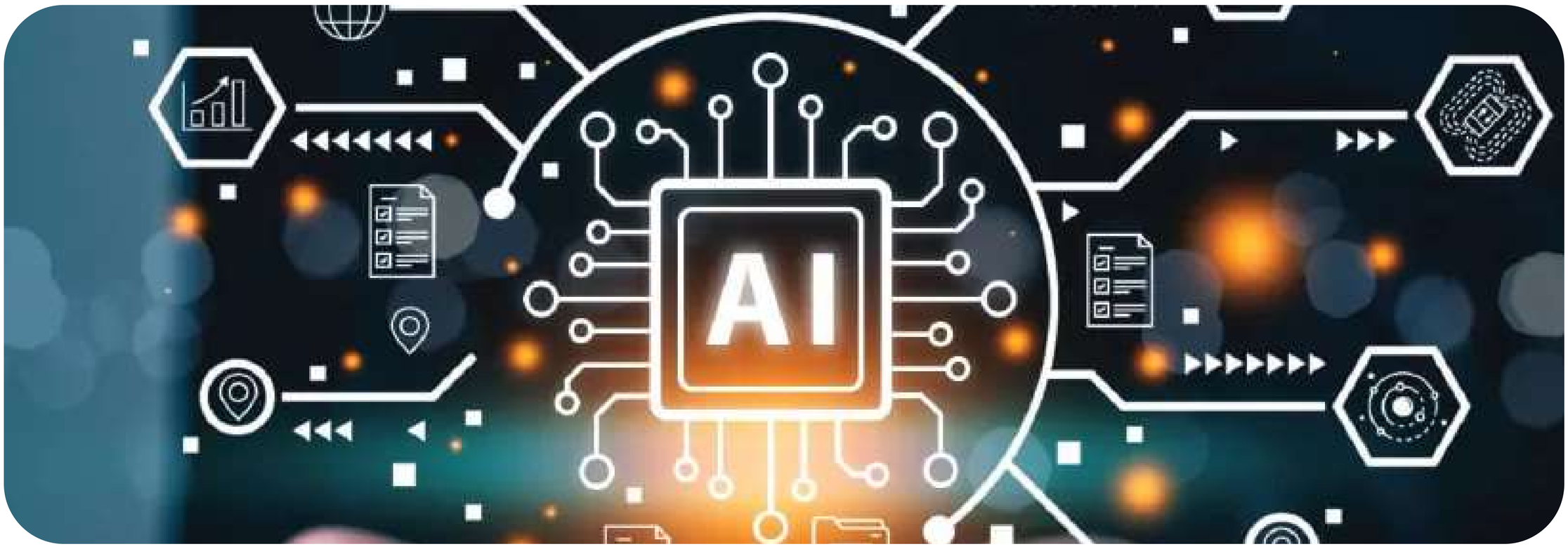
Enhanced User Productivity: Automating complex queries offloads analysts from manual query writing, driving focus on higher-value tasks. This results in accelerating innovation and data-driven decisions.

Beyond doubt, Gen AI empowers modern businesses to lead the data-driven future. Here's a glimpse of its transformative potential and opportunities:

Key Takeaways

Gen AI, powered by Azure OpenAI and BERT, is transforming how organisations tap into the potential of data mining — eliminating technical barriers and enabling natural language access to complex data insights. This whitepaper demonstrates how modern-day organisations can empower non-technical teams to query data, automate workflows and accelerate decision-making — all without writing a single line of SQL.

By bridging the gap between raw data and business intelligence, this AI-driven approach takes agility and collaboration to the next level. Simultaneously, it helps unlock cost-efficiency and enables scaling effortlessly with evolving data demands. Organisations that adopt this strategy position themselves to unlock deeper insights, foster innovation and stay ahead of the curve in today's data-first world. This is more than a shift in tools – it represents a strategic transformation in how data is accessed, interpreted and harnessed to drive enduring business value.



About BI WORLDWIDE

BI WORLDWIDE is a global leader in designing loyalty and engagement solutions that inspire your employees, channel partners and sales teams to achieve more. What sets us apart in the industry? We are not just program providers, we are your KPI-first consultative partner, creating the right end-to-end engagement journeys, powered by actionable behavioural science, AI, deep data analytics and human-centric design, that work for your people, brand and industry, driving results that matter. Over the past 75 years, we have been helping Fortune 500 and bluechip organisations across industries attract, engage & retain top industry talent, build strong, profitable channel partner ecosystems and run high-impact sales incentives, driving measurable business outcomes.