panoply
by SQream

PANOPLY WHITEPAPER

How to Leverage AI in Each Part of Your Data Stack

AI is more than just a buzzword; it's a versatile tool that's become integral to our personal and professional landscapes. Explore how AI tools can automate tasks, increase efficiency, and reduce costs across your data stack.

AI is Expanding

In 2023, it's nearly impossible to escape the influence of artificial intelligence (AI) in our daily lives. Whether scrolling through our social media feed, reading the latest news, or simply walking down the street, AI is a constant presence. It's not just about futuristic robots or self-driving cars; AI has become integral to our personal and professional landscapes.

For consumers, AI shapes the recommendations we see, the products we buy, and even the content we consume. For marketers, it's a powerful tool for reaching and engaging with audiences. And for tech companies, AI drives innovation, from code generation to data analysis. AI, in short, is everywhere.

But AI is more than just a buzzword; it's a versatile tool that can enhance every aspect of your data stack. Whether you're a small to mid-sized business (SMB) or a mid-market company, this guide will explore how AI can automate tasks, reduce manual workloads, increase efficiency, and save money across your data stack.

However, remember that AI is not a replacement for human intelligence; it's a complementary tool that requires ongoing monitoring to ensure accuracy.

Let's dive in and see how AI can seamlessly integrate with each component of your data stack, from unstructured data collection to data governance and security. Since the field is constantly changing, we'll focus on where you can look for AI solutions within your data stack rather than specific tool recommendations.

Data Ingestion

Data ingestion is the bedrock of your data stack, where raw data from diverse sources is collected and prepared for analysis. AI-driven enhancements in this phase can significantly impact your data operations' quality, speed, and efficiency.

Automating Data Extraction

Unstructured data, such as images, text, and video, presents a goldmine of information. However, manually extracting valuable insights from these sources can be time-consuming and costly. AI-powered solutions offer a more efficient and cost-effective way to tackle this challenge.

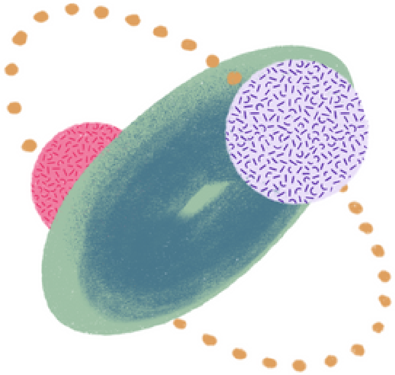
AI-powered tools can automate data extraction from unstructured sources, delivering a profound boost in efficiency. For instance, companies like Zebra Medical Vision have leveraged AI to enhance medical imaging diagnostics. While AI's exact role in Zebra's workflow may vary, the principle remains the same: AI assists in making sense of complex, unstructured medical images, reducing the burden on medical professionals and speeding up diagnoses.

AI has democratized access to processes that were once the domain of specialized solutions. Whether you're in healthcare, marketing, or any industry dealing with unstructured data, AI can automate the extraction of valuable information, making it accessible to everyone.

Detecting Data Sources

Data sources are like moving targets; they change frequently, introducing new elements and structures. AI-powered tools can act as vigilant sentinels, automatically scanning all data sources to detect new additions. AI can identify these changes and trigger the ingestion process without human intervention, whether a fresh application programming interface (API) service or additional fields in your production tables.

Mapping Data Sources



Understanding the structure and schema of your data sources is crucial for effective data management. AI can rapidly analyze the intricacies of your data sources and map them to their respective data models. But it doesn't stop there; AI can provide a comprehensive map of your data warehouse.

Imagine having a multi-connection map visually representing your data warehouse's tables and relationships. This holistic view not only simplifies data management but also helps you better understand your data assets. With AI-driven mapping, you can confidently navigate your data stack, knowing where your data resides and how it's connected.

Improving Data Quality

Data quality is paramount in any data stack, but ensuring the data ingested is clean and reliable is often a daunting task. AI can come to the rescue, offering powerful tools to enhance data quality.

AI-driven data quality processes include deduplication, anomaly detection, error identification, and handling inconsistencies or missing data. These automated checks and corrections not only save time but also reduce the risk of inaccurate data making its way into your analytics or decision-making processes.

Real-Time Complex Ingestion

Real-time data ingestion is becoming increasingly important in today's business environment where saving time = saving money.

Leveraging AI here can elevate real-time ingestion to a more sophisticated level. It reduces processing time and ensures that ingested data is always current, even when it goes through multiple preprocessing steps before ingestion. This real-time complex ingestion capability saves time and manual effort and ensures your data's accuracy and relevance.

Automated Error Detection

Every data ingestion process is susceptible to issues due to unmaintained processes or unexpected changes in incoming data structures. Automation capabilities in AI tools shine when detecting and resolving these issues.

AI can identify anomalies and deviations from expected data patterns, triggering automated actions to minimize or even prevent errors and delays in data deliveries. This proactive approach ensures data accuracy and reliability.

For instance, if your organization relies on data from external sources, AI can automatically detect and notify you of any disruptions in data feeds, allowing you to take corrective actions promptly.

Data Transformation

Data transformation involves converting, normalizing, and structuring raw data to make it suitable for analysis. Panoply alone can handle many transformations without AI, like data cleaning, de-duplicating, and handling missing values. But for more advanced data transformation tasks like feature engineering or dimensionality reduction, AI can significantly expedite this process by automating initial transformations.

Traditionally, data transformation has been carried out using ETL (Extract, Transform, Load) processes.

However, with the advent of AI, new methods such as ELT (Extract, Load, Transform) and Reverse ETL have gained prominence, each offering distinct advantages in different scenarios.

ETL: Extract, Transform, Load

ETL has been the go-to data transformation method for many years. It is favored when the processed data takes precedence over preserving the raw data.

Here's how ETL works:

- Extract: Data is extracted from various source systems, such as databases, applications, and APIs.
- Transform: The extracted data is cleaned, enriched, and transformed into a structured format. This phase often involves data normalization, aggregation, and quality checks.
- Load: The transformed data is loaded into a target data warehouse or database, making it available for analysis and reporting.

AI-driven ETL tools automate many aspects of data transformation, from data cleansing to pattern recognition. These tools can intelligently identify and correct data errors, reducing the need for manual intervention and speeding up the transformation process.

ELT: Extract, Load, Transform

ELT represents a shift in data transformation strategy. It is well-suited for organizations prioritizing time optimization and dealing with high volumes of data.

ELT flips the ETL process, emphasizing data loading before transformation.

The ELT process typically follows these steps:

- Extract: Data is extracted from source systems and loaded directly into a target data warehouse or storage system.
- Load: Data is ingested into the target storage without immediate transformation. This step enables organizations to handle large data volumes efficiently.
- Transform: Data transformation occurs within the data warehouse or storage system itself. Advanced data processing tools and AI algorithms transform the data on demand.

AI enhances ELT by enabling real-time or near-real-time data transformation within the warehouse. Machine learning models can be applied to transform data as it is ingested, allowing for more flexible and responsive data processing.

Panoply follows an ELT approach, enabling users to load data directly into their target location and perform transformations.

It supports the automatic transformation of NoSQL data from sources like Elasticsearch, DynamoDB, and MongoDB into structured tables.

Reverse ETL: A New Paradigm

Reverse ETL is a relatively recent addition to the data transformation landscape. Its unique focus on connecting data stores to customer-facing or action-based systems has garnered attention.

It enables organizations to derive shared real-time insights across applications, ultimately personalizing customer interactions at scale. Here's how:

- Extraction: Reverse ETL extracts data from various data stores, including data warehouses and CRM or ERP systems.
- Loading: Data is loaded into customer-facing or action-based systems, facilitating real-time access to critical insights.
- Transformation: Transformation in Reverse ETL is often geared toward making data more accessible and actionable for end-users.

AI's role in Reverse ETL is:

- Personalization: AI algorithms can analyze customer behavior and preferences, offering personalized product recommendations or content.
- Real-Time Decision Support: AI enables real-time decision-making by providing insights at the point of customer interaction.
- Integration: AI-powered integration tools help streamline data flow between data stores and customer-facing systems.

Data Storage

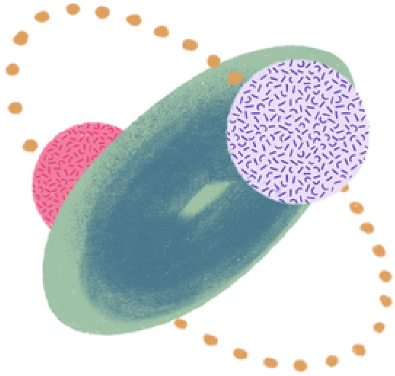
Data Warehousing

Data warehousing is the heart of an organization's infrastructure, serving as the repository for vast structured and semi-structured data. AI can enhance data warehousing in several ways, making it more efficient, secure, and insightful.

Optimizing Queries

One of the primary functions of a data warehouse is to facilitate querying and data retrieval. AI tools optimize these queries by analyzing historical query patterns and usage.

Here's how AI can enhance query performance:



- Indexing: AI-driven indexing identifies the most frequently queried data, allowing faster access to relevant information. It can automatically create and update indexes to improve query speed.
- Caching Strategies: AI can implement intelligent caching strategies, ensuring that frequently accessed data is readily available in memory, further enhancing query speed.
- Query Optimization: AI algorithms analyze query execution plans and suggest optimizations. For instance, they can recommend using more efficient join algorithms or query rewriting to improve performance.

Query Suggestions

One common challenge in data warehousing is the inconsistency of queries. Teams often pose similar questions differently, leading to data discrepancies and confusion. AI can help standardize queries by identifying common joins, filters, and groupings.

AI-driven processes can recognize patterns in query usage and recommend standardized query templates. This ensures the team queries data with the same logic, reducing discrepancies and improving consistency.

Predictive Analytics

AI's predictive analytics capabilities are a game-changer in data warehousing. AI tools can analyze historical data, identify patterns, and predict future trends when integrated into your data warehouse.

This empowers organizations to:

- Make informed decisions: Predictive analytics provides insights into future trends, enabling proactive decision-making.
- Optimize operations: Businesses can optimize their operations and resource allocation by forecasting demand or identifying potential issues in advance.
- Enhance customer experiences: Predictive analytics helps personalize customer experiences by anticipating their needs and preferences.
- Mitigate risks: AI can identify potential risks and vulnerabilities, allowing organizations to take preventive measures.

Anomaly Detection

Data warehouses store sensitive and critical data, making security paramount. AI algorithms can serve as vigilant sentinels, continuously monitoring activity within your data warehouse. They watch for red flags indicating security breaches, unauthorized access, or performance issues.

AI helps organizations adhere to data privacy regulations like the General Data Protection Regulation (GDPR) and avoid costly non-compliance penalties by identifying anomalies in user behavior, data access patterns, or system performance. It also ensures the integrity and security of your data warehouse.

Data Compression and Storage Optimization

Optimizing storage is critical to data warehousing, and AI can make a substantial difference in this area. AI can analyze data usage patterns and identify data compression and storage optimization opportunities.

AI-driven processes can determine which data is frequently accessed and which is rarely used. It can then automate the movement of less frequently accessed data to cost-effective cold storage services. This dynamic storage management ensures that data is readily available when needed, minimizing storage costs.

Capacity Planning

Many data warehousing services allocate capacity based on predefined configurations. AI services can change the game by automatically identifying when more capacity is needed and where. This dynamic allocation ensures that resources are used efficiently and cost-effectively, eliminating the need for static capacity schedules.

By continuously monitoring usage patterns and system performance, AI can recommend real-time capacity adjustments, ensuring your data warehouse operates smoothly during peak periods.

Monitoring and Alerts

While various alerting systems exist, they typically work reactively, notifying users only after an issue has occurred. AI can transform monitoring and alerting into a proactive process.

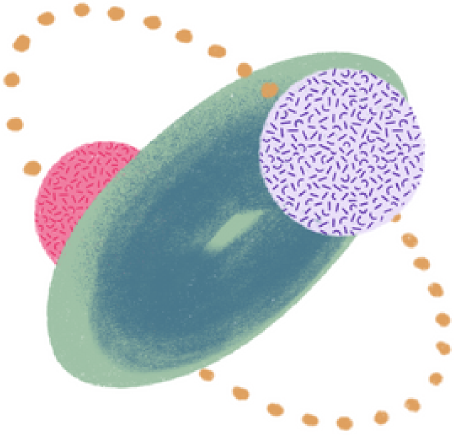
AI-driven monitoring continuously tracks data warehouse usage and performance. It can detect potential issues before they escalate into critical problems. By providing early warnings and insights, AI empowers data professionals to take preemptive actions, maintaining data warehouse health.

Panoply takes a proactive approach to monitoring and alerts by continuously assesses data pipeline performance and usage, alerting users to potential issues before they impact operations. This minimizes downtime and ensures data availability.

NLP (Natural Language Processing)

Natural language processing is a branch of AI that enables machines to understand and interact with human language. When integrated into data warehousing, NLP can:

- Facilitate data exploration: Users can ask questions in plain language. NLP can translate these queries into structured query language (SQL) or other query languages, making data exploration more accessible to non-technical users.
- Automate documentation: NLP can automatically generate documentation for data tables and schemas, reducing the burden of manual documentation tasks.
- Enhance search: NLP-powered search engines can retrieve relevant data and insights more effectively, improving data discovery within the warehouse.



Data Lineage an Auditing

Data lineage and auditing are crucial for maintaining data integrity and compliance. AI can automate the tracking and auditing of data as it flows through the data warehouse.

AI-driven data lineage tools can create comprehensive records of data movements, transformations, and access. These records serve as a transparent audit trail, enabling organizations to trace the origin and history of data and facilitating compliance with regulatory requirements.

Data Lakes

Data lakes have become the cornerstone of modern data architecture, allowing organizations to store vast amounts of structured and unstructured data flexibly and cost-effectively. AI is playing a pivotal role in optimizing the utilization of data lakes.

Automated Data Discovery

Data lakes are often vast reservoirs of information, making it challenging to discover and utilize relevant datasets efficiently. AI can be a guiding light in this data wilderness. AI-powered tools can scan and categorize data during the ingestion process into the data lake, making finding and accessing the datasets you need easier.

Schema Inference

Data within data lakes can be a mix of structured, semi-structured, and unstructured formats, making schema design and integration complex.

AI tools can decipher semi-structured and unstructured data structures, streamlining the schema design process and facilitating seamless data integration.

Data Cleansing and Enrichment

Data quality is paramount for data lakes to be valuable assets. AI algorithms contribute significantly to data cleansing and enrichment.

They identify and resolve errors, detect missing values, and link data with external sources. This automated data enhancement process ensures data integrity and reliability.

Data Cleansing and Enrichment

Metadata is crucial in understanding and managing data within a data lake. AI can simplify and expedite the process of metadata extraction from various sources.

AI algorithms can automatically extract metadata from various sources, including text-based documentation, images, videos, and data content. This automated metadata extraction enhances and shortens the data cataloging process, providing valuable insights into the nature and context of data assets within the data lake.

Data Collaboration and Sharing

Data collaboration and sharing are essential aspects of modern data lakes. AI can facilitate collaboration by suggesting relevant datasets, analyses, or insights to users based on their usage patterns and preferences. This proactive approach fosters data-driven collaboration within organizations.

Pattern Recognition

Pattern recognition is a fundamental AI capability that enhances data analysis within data lakes. AI algorithms can identify and extract meaningful patterns and trends from the vast data stored in data lakes.

By recognizing patterns within the data, AI aids data professionals in uncovering valuable insights, making data-driven predictions, and identifying anomalies or outliers. This pattern recognition capability is particularly valuable for organizations seeking to extract actionable insights from their data lakes.

Data Governance and Security

Effective data governance and security are paramount in data lake management. AI can aid in enforcing governance policies, ensuring data quality, and detecting security threats, helping organizations maintain control and compliance.

Lineage Tracking

Understanding data lineage is essential for data governance and compliance. AI automates data lineage tracking, providing a clear and auditable record of data's origins and transformations. This transparency is indispensable for maintaining data integrity and complying with regulations.

Unlocking the Power of AI in Your Data Stack

Artificial intelligence can revolutionize every aspect of your data stack, from collecting unstructured data to governance and compliance. AI is a dynamic field, and as it continues to evolve, so will its applications in your data stack. By harnessing the power of AI, SMBs to mid-market enterprises can streamline operations and stay ahead of the competition in this data-driven era.

Remember, before you can implement AI into your data stack, you need a solid starting point like a Panoply data warehouse to enable your favorite AI tools.

Inspired? Book Panoply demo to experience the power of a managed wata warehouse plus ELT and visualizations – your end-to-end solution – and start your data stack transformation.

Book a demo

panoply
by SQream