Data for the AI: the HARD choice

– Marco Dal Pino

DATA SATURDAYS

Saturdays Sofia 2024

05 | October | 2024

--------Gold sponsor

Microsoft awara IT INSPIRIT Baringa

Welcome to

DATA SATURDAYS

D Saturdays

A Sofia

T Bulgaria

A 2024

Big thanks

DATA SATURDAYS

D

A

T

A

Microsoft

INSPIRIT
THE DATA PLATFORM COMPANY

awara IT

Baringa

scalefocus

Kinkao_Sofia
THAIFOOD + HOMEMADE

SOFIA TECHPARK

Original Customer Environment

Brown Field

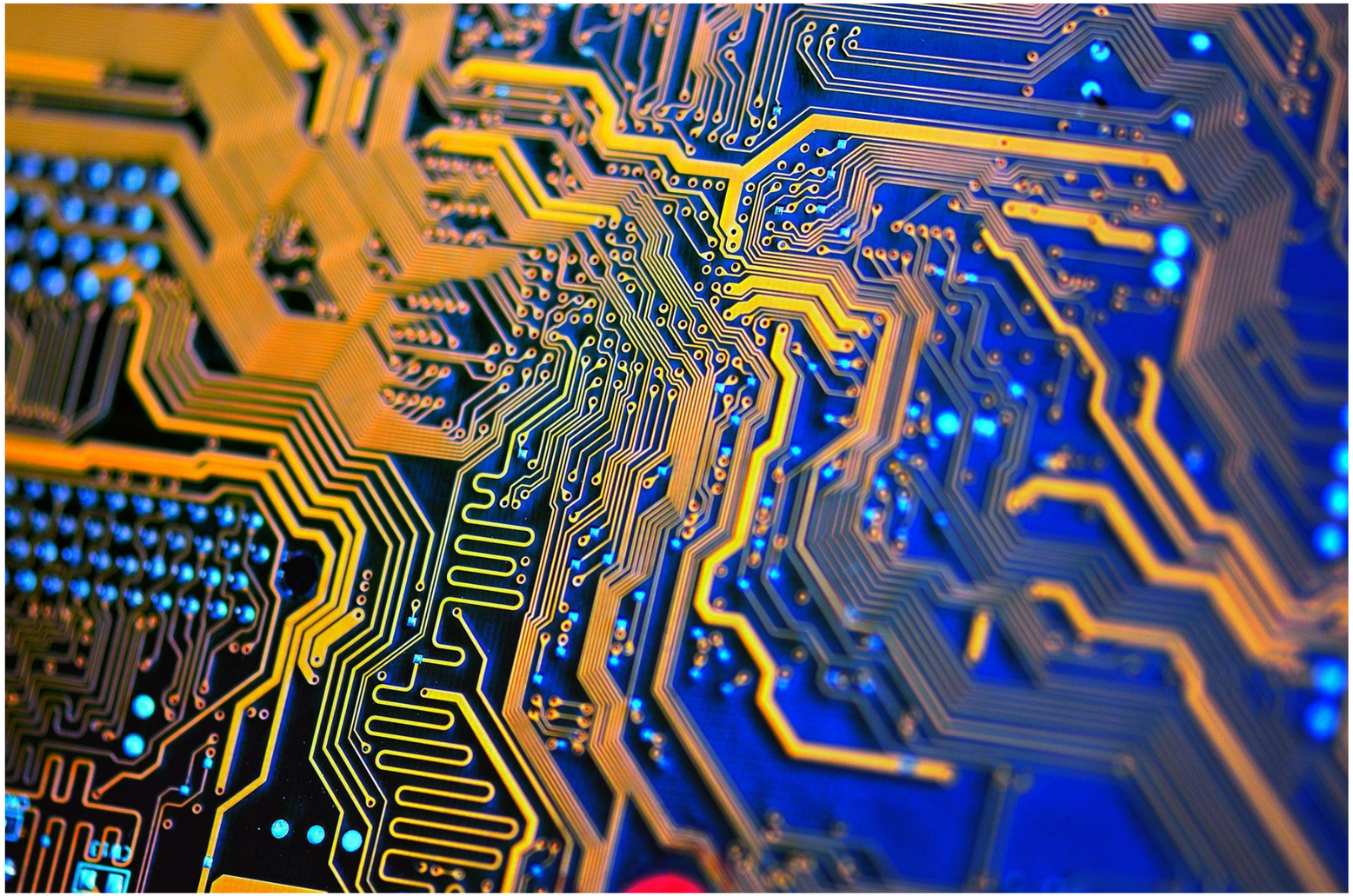
- Need to work with pre-existent Environment
- Modernize/Migrate
- Take the best from Customer Env.

Green Field

- *“Best Choice”*
- Cost
- Efficiency
- Compatibility
- Future Proof

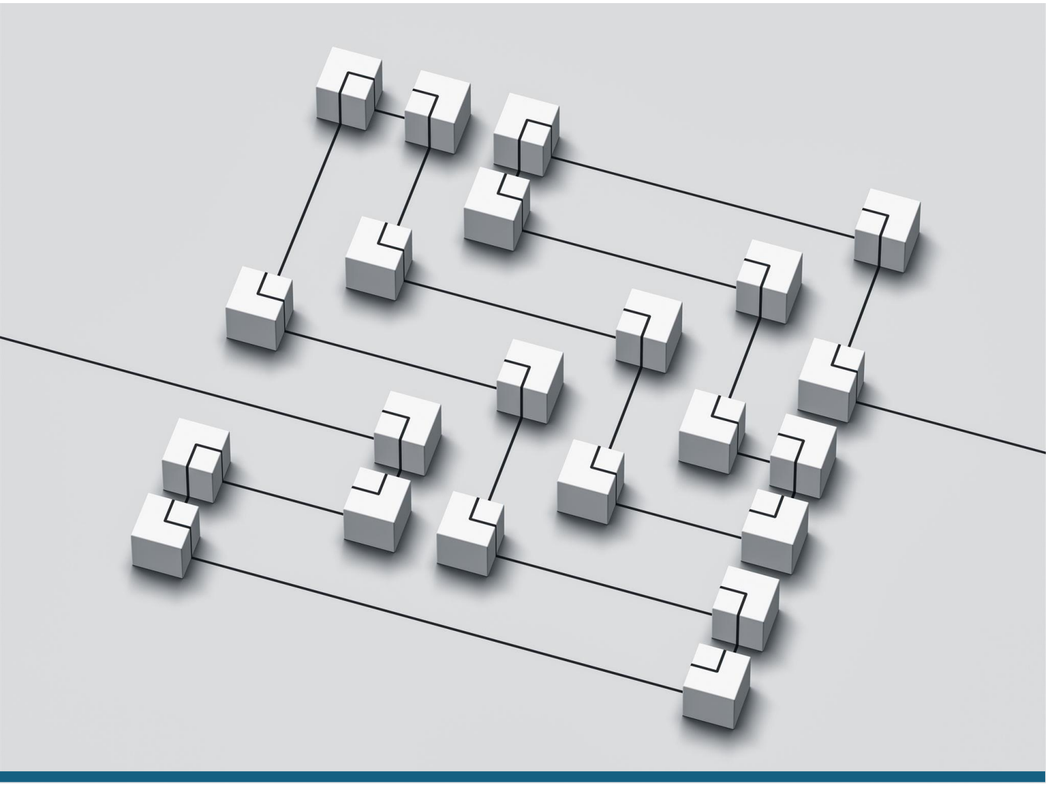
Choosing the Right Database for AI Projects

Key Factors to consider while selecting an AI database



Types of Databases for AI

Relational Databases



Suitable for Structured Data

Relational databases are suitable for structured data storage and are widely used in various applications, including AI applications.

Robust Data Integrity and Security

Relational databases offer robust data integrity and security features to ensure data consistency and protection against unauthorized access.

Inflexibility

Relational databases can be inflexible and have limited ability to store unstructured data or adapt to evolving data models.

Performance Issues

Relational databases can have performance issues with complex queries and large datasets, which can affect the overall performance of the application.

NoSQL Databases

Unstructured and Semi-Structured Data

NoSQL databases are designed to handle unstructured and semi-structured data, which makes them a suitable choice for AI applications that involve large data sets.

Horizontal Scalability

NoSQL databases can scale horizontally, which allows them to handle large amounts of data effectively.

Challenging Query

NoSQL databases can be challenging to query, which can make it more difficult to get the data you need quickly.

Data Integrity and Security

NoSQL databases can lack data integrity and security features, which can make them less suitable for applications that require high levels of data protection.

Graph Databases

Suitable for AI Applications

Graph databases are designed to handle complex relationships between data, making them ideal for AI applications that involve knowledge graphs, as they can be highly performant and offer flexibility in handling complex queries.

Challenges in Scaling

While graph databases can handle complex relationships between data with ease, they lack data integrity and security and can be challenging to scale. These challenges can significantly impact the performance of the database, leading to slower queries.

Understanding Database Requirements for AI Projects

Scalability

Horizontal scaling is a key requirement for AI databases to handle large data sets and meet the demands of AI applications. It allows databases to add more nodes to distribute the data and queries across the nodes.

Performance

The performance of an AI database can be affected by various factors, such as the size of the data set, the complexity of the queries, and the number of nodes in the cluster. It is important to optimize each of these factors to achieve optimal performance.

Flexibility

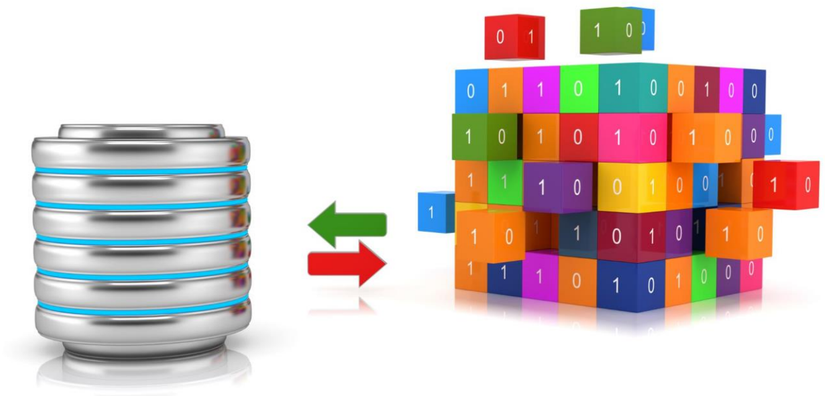
Data Types

AI databases must be able to handle various data types, including structured, semi-structured, and unstructured data, making them suitable for a wide range of applications.

Complex Relationships

AI databases must be able to handle complex relationships between data, making them suitable for knowledge graphs and other applications that require sophisticated data modeling.

Data Integration



Multiple Data Sources Integration

AI databases must be able to integrate data from multiple sources including data warehouses, data lakes, and real-time data streams. This process is essential to generate accurate insights and improve business decision making.

Data Quality Management

AI databases must be able to handle data quality issues, such as missing or inconsistent data. Proper data quality management ensures that the data is accurate and reliable, leading to more accurate insights and business decisions.

Data Quality

Data quality is crucial for AI applications as it can affect the accuracy and reliability of AI models. Ensuring data quality involves validating data, checking for inconsistencies, and identifying potential errors.

Data Structure

Flat Data Structure

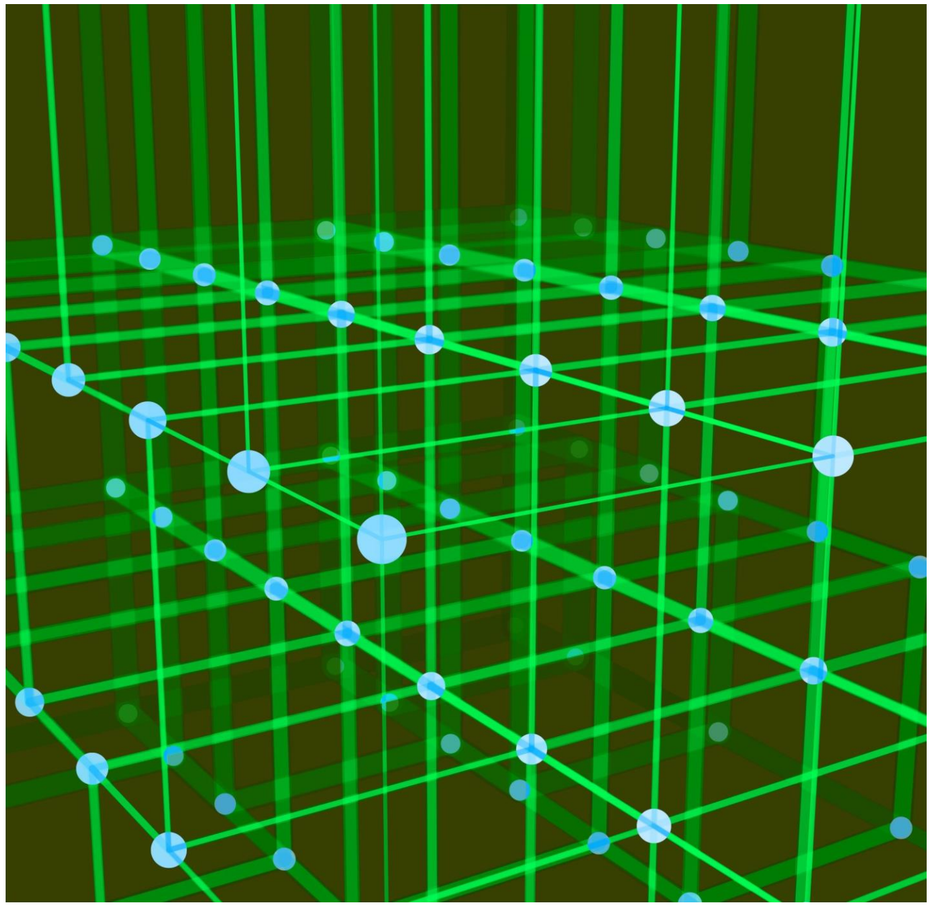
A flat data structure is a simple and straightforward way to organize data. It consists of a single table with rows and columns, similar to a spreadsheet.

Time-Series Data

Time-series data is a structure that is used for data that changes over time. It is commonly used in applications such as finance, weather forecasting, and IoT devices.

Graph Databases

Graph databases are used to store and manage data that has a complex, interconnected structure. They are commonly used in social networks, recommendation engines, and fraud detection systems.



Data Volume

The amount of data to be stored is a key factor to consider when selecting a database, as it impacts scalability and performance. Projects that require storing large amounts of data may require specialized solutions.

Query Requirements

Query Complexity

The complexity of queries that will be run on the database is an important consideration when selecting a database. Some projects may require complex queries, which can impact query performance and optimization.

Query Performance

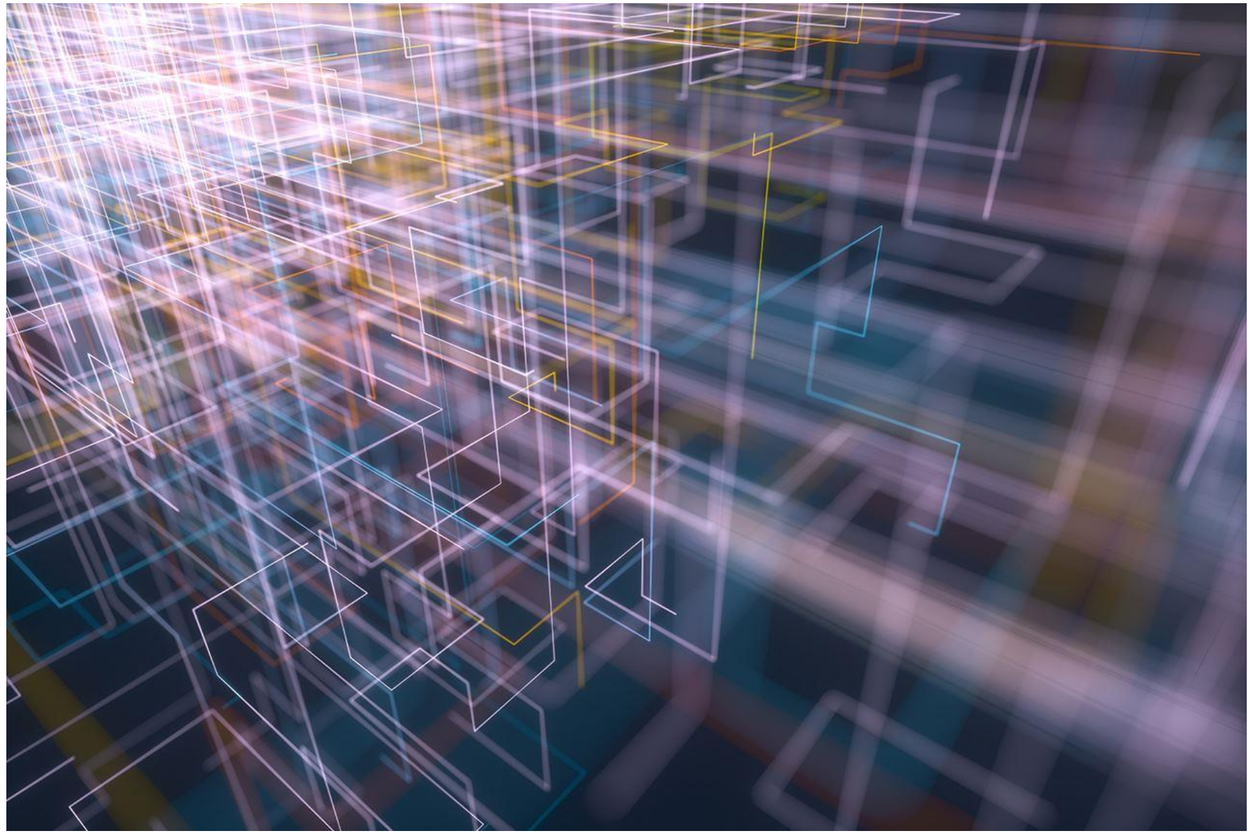
Query performance is important for any database application. When selecting a database, it's essential to consider its performance characteristics, including query optimization and indexing.

Data Visualization

Data visualization can help identify trends and patterns in large data sets. When selecting a database, consider its support for data visualization tools and techniques.

Choosing the Right Database for Large-Scale AI Projects

Scalability

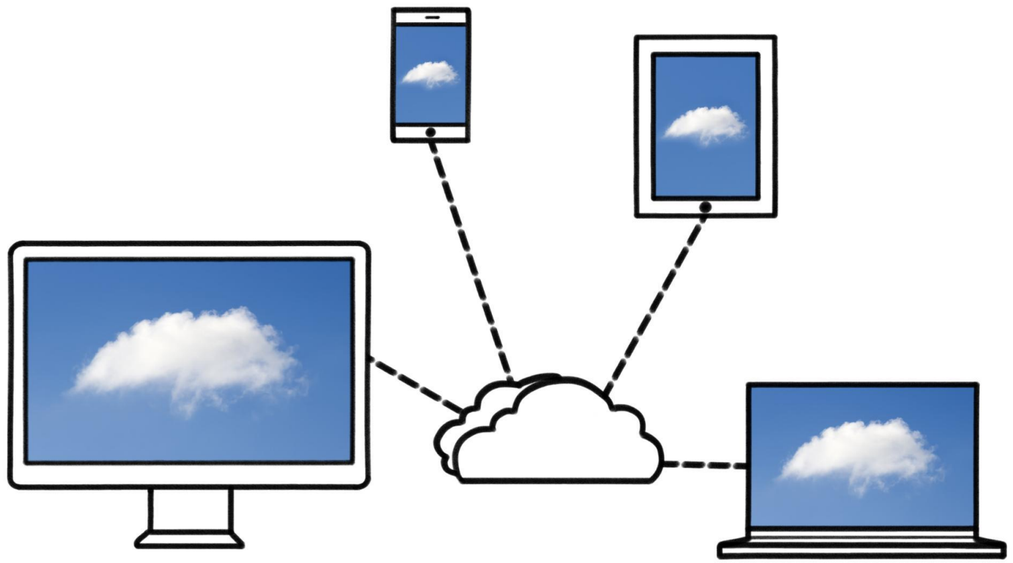


Scalability is critical for large-scale AI projects as it ensures databases are able to efficiently distribute data across multiple nodes, enabling fast and reliable access to data.

Latency

Latency is a critical factor for large-scale AI projects as it affects the speed at which databases can handle high volumes of requests. Low latency is essential to ensure that AI models can be trained and deployed quickly.

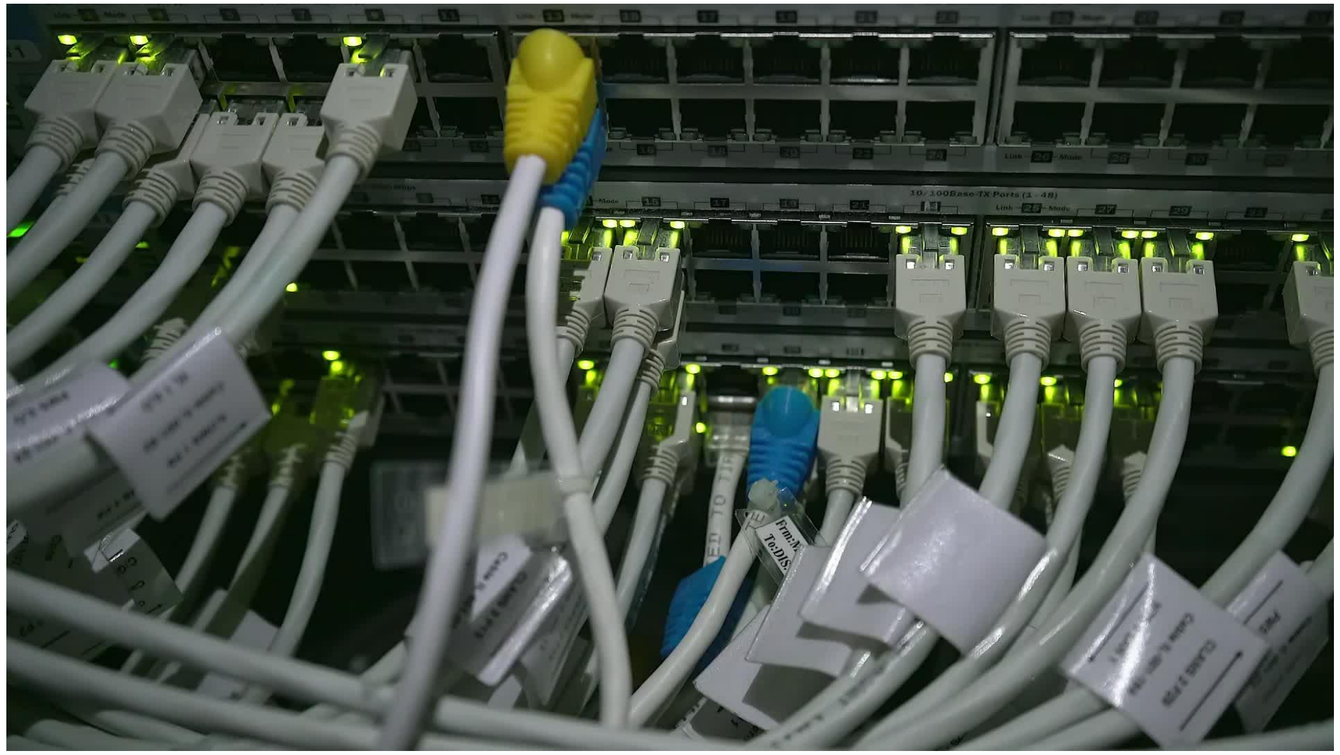
Availability



High availability is essential for large-scale AI projects. Databases must be designed to handle failures and ensure that data is always available for AI model training and inference.

Choosing the Right Database for Real-time AI Projects

Data Velocity



Data velocity is critical for real-time AI projects. Databases must be capable of handling high volumes of incoming data and providing low latency access to that data.

Data Granularity

Real-time AI projects require databases to store and retrieve highly granular data, such as individual sensor readings, to ensure accurate and effective AI model training.

Concurrency

Concurrency is essential for real-time AI projects as databases must be able to handle multiple requests concurrently and maintain data consistency across all requests.

A couple of suggestions

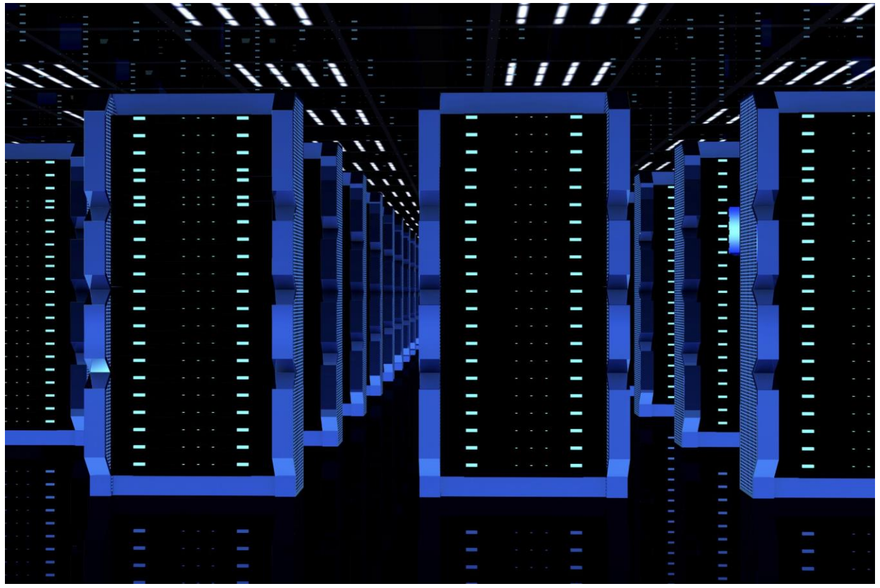
Microsoft Fabric

1. **Unified Platform**: Microsoft Fabric is an end-to-end analytics and data platform designed to unify data movement, processing, ingestion, transformation, real-time event routing, and report building[1].
2. **Components**: It integrates various components such as Data Engineering, Data Factory, Data Science, Real-Time Analytics, Data Warehouse, and Databases into a single platform[1].
3. **OneLake**: This is a unified data lake storage that centralizes data storage, allowing for seamless data access and management without the need for data duplication[2].
4. **AI Integration**: AI capabilities are embedded within the platform, facilitating the transition from raw data to actionable insights[1].
5. **SaaS Foundation**: Operating on a Software as a Service (SaaS) model, it simplifies analytics requirements by integrating services from Power BI, Azure Synapse Analytics, and Azure Data Factory[1].
6. **Real-Time Analytics**: It supports real-time data processing and analytics, enabling timely insights and decision-making[2].
7. **Governance and Security**: Fabric ensures centralized administration, governance, and security, with features like data sensitivity labels and row-level access controls[1].
8. **User Roles**: Tailored experiences for different user roles such as data engineers, scientists, and warehousing professionals[1].
9. **Familiar Tools**: Supports familiar data languages like T-SQL, PySpark, Scala, and SparkR, making it accessible for various data professionals[2].
10. **Collaboration**: Facilitates data sharing and collaboration across different data roles and organizational silos[2].

Microsoft SQL Server

Relational Data Store

Microsoft SQL Server is a relational database management system that is designed to store and manage structured data. It is suitable for storing large amounts of data and provides features such as indexing, querying, and transactions.



Parallel Query Execution

Parallel Query Execution

Microsoft SQL Server can execute queries in parallel, which can improve performance for large-scale data processing workloads by simultaneously processing multiple query fragments.

Distributed Computing

SQL Server can support distributed computing through features such as PolyBase.

Native Vector data type Support

Vector Datatype support

Starting from may 2024 SQL Server offer (in preview) native support for a new Vector datatype opening to easily build AI-enabled solution using an enterprise ready, secure, scalable platform

- 30+ years in IT (Developer, Architect, Consultant, PM, Trainer)
- Speaker, Community addicted
- IoT Influencer
- Microsoft Certified Trainer

Marco Dal Pino
Technical Consultant
Microsoft

https://www.linkedin.com/in/marcodalpino

https://about.me/marcodalpino

https://twitter.com/marcodalpino

info@contoso.blog

https://www.twitch.tv/dpcons
https://www.twitch.tv/techchat

Microsoft MVP Most Valuable Professional Reconnect

DotNet Toscana

Resources

- EAP for Vector Support Refresh - Introducing Vector type - Azure SQL Devs' Corner (microsoft.com)

Our next !

JS.TALKS ();

- Workshop Day | 22 | Nov | 2024
- Conference Day | 23 | Nov | 2024
- Innovation Forum "John Atanasoff"
- Sofia Tech Park