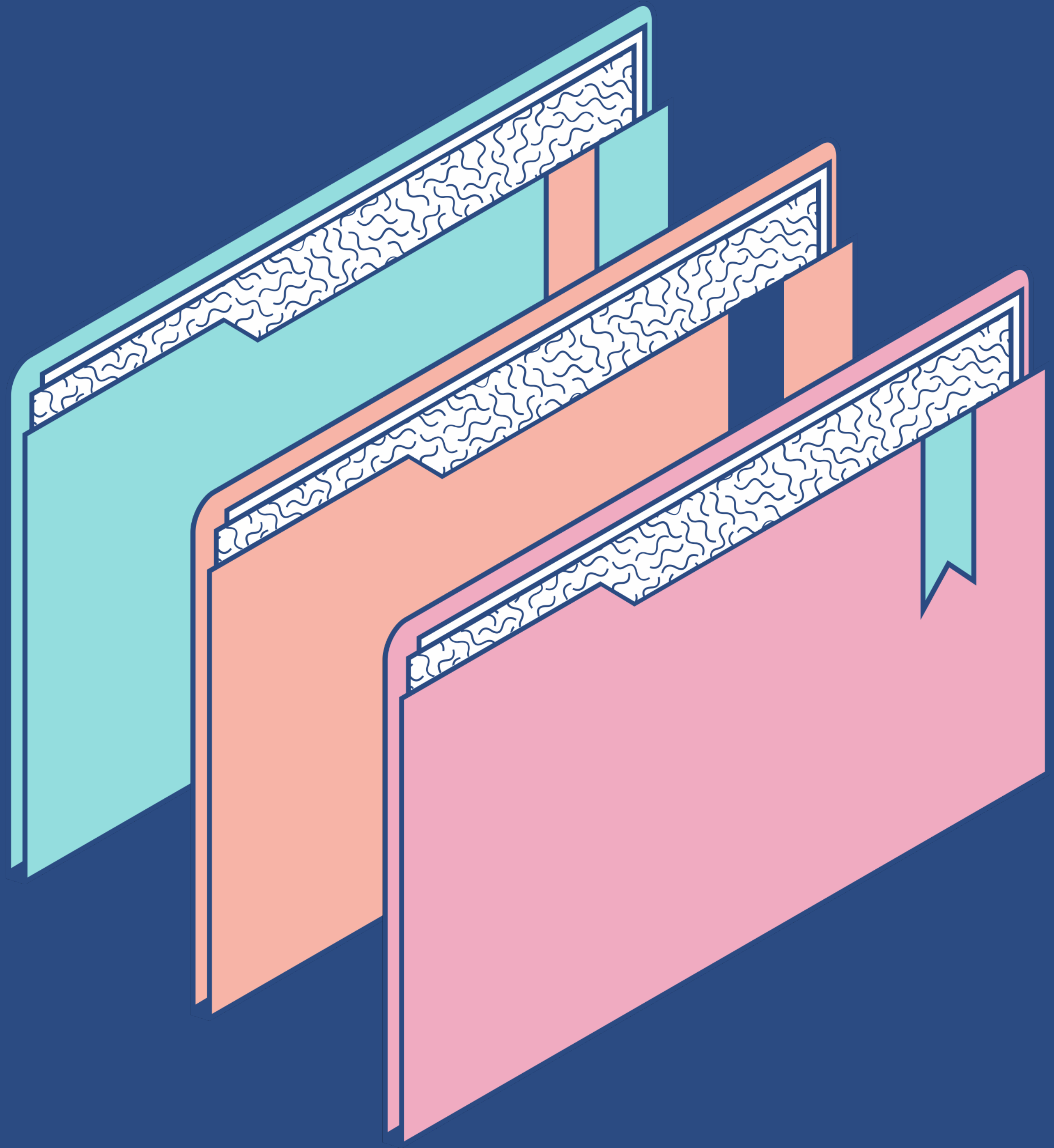
TECHNOLOGY PRESENTATION

Statistical and Machine Learning Technologies Applied to Databases

Minjun Kim, Nabin Kim, Seohee Yoon



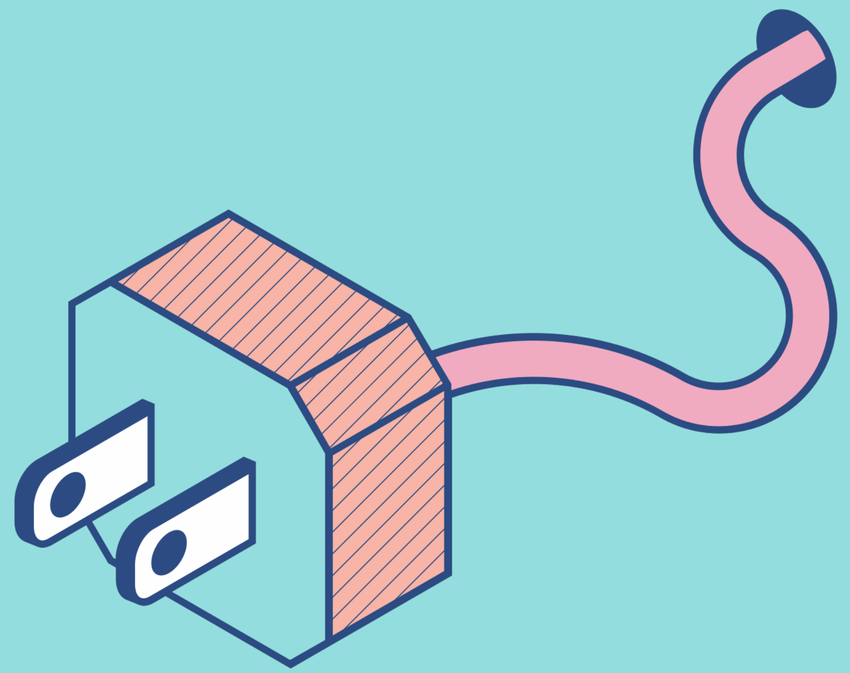
Agenda

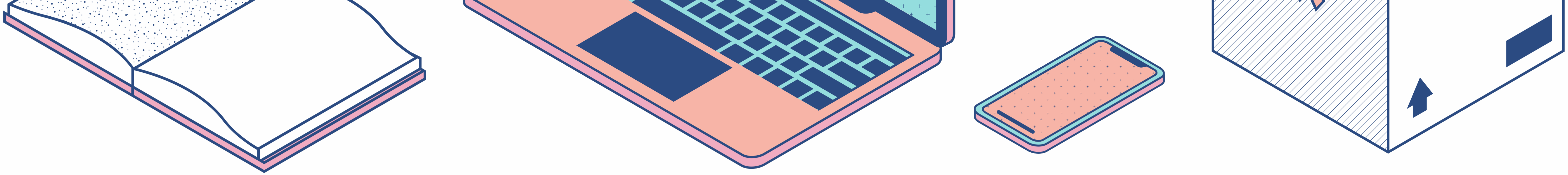
- Basic Definitions and Concepts
- Features and Functions
- Historical Development
- Products Comparison
- Details on Select Products
- Sample Applications
- Market Standing and Future Prognosis
- Relevant problems and Research

- Basic Definitions and Concepts
- Features and Functions
- Historical Development
- Products Comparison

Basic Definitions and Concepts

- Database Management Systems (DBMS): Software designed to store, retrieve, and manage data in databases. They ensure data integrity, security, and efficiency in data handling.
- Machine Learning: A subset of AI that enables systems to learn from data, identify patterns, and make decisions.
- Statistical Methods: Techniques rooted in statistics and probability theory, used to interpret data, identify trends, and make predictions.





Features and Functions

Automated Data Analysis

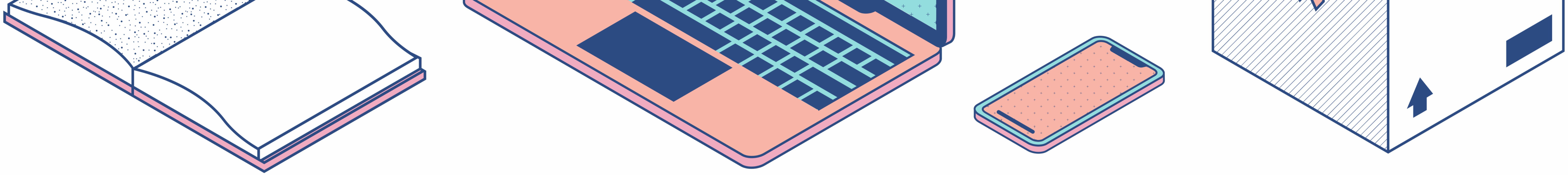
- **Pattern Recognition**: ML algorithms can easily identify patterns and trends in the data
- **Anomaly Detection**: Statistical methods are used to detect outliers or anomalies in the data.

Real-time Analytics

- **Stream Processing:** ML models can analyze data streams in real-time, enabling immediate insights and responses.

Predictive Analysis

- **Forecasting:** ML models, especially those based on time series analysis, can predict future values for a given dataset.
- **Customer Behavior Prediction:** By analyzing historical data, ML models can predict customer behaviors, such as purchasing patterns



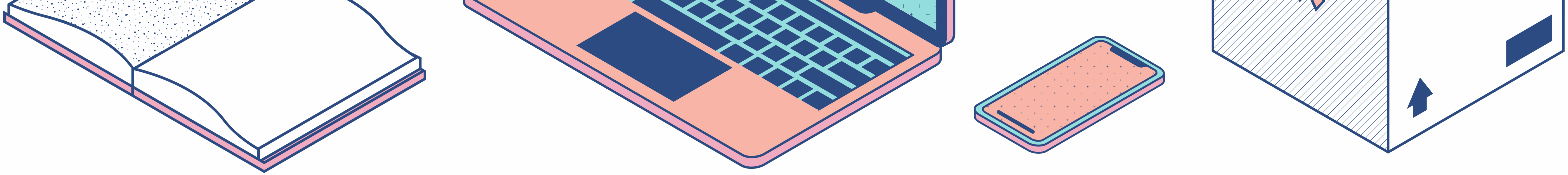
Features and Functions (Cont.)

Query Optimization

- **Dynamic Query Planning:** Predict the most efficient way to execute a query based on historical performance data, optimizing resource usage and reducing execution times.
- **Index Management:** Predict which indexes will be most beneficial for query performance, reducing the overhead of manual index tuning.

Data Management

- **Data Cleaning:** Automate the process of identifying and correcting errors in data, ensuring higher quality and reliability of the database.
- **Data Compression:** Intelligently compress data, balancing between compression rates and query performance, hence optimizing storage costs and speed.



Features and Functions (Cont.)

Natural Language Processing (NLP)

- **Semantic Query Processing:** Allows users to interact with the database using natural language queries, making data access more intuitive for non-technical users.
- **Sentiment Analysis:** Automate sentiment analysis, providing quick insights into user sentiments and trends.

Security and Compliance

- **Intrusion Detection:** Statistical anomaly detection techniques can identify potential security breaches by monitoring access patterns and flagging unusual behavior.
- **Data Privacy:** ML algorithms can help in implementing data masking and anonymization techniques, ensuring that sensitive information is protected in compliance with data privacy regulations.

Historical Development

1

1960S-1970S

Emergence of DBMS

Database Management Systems revolutionalize data storage and retrieval

2

1980S

Data Mining

Combined data analysis and statistics to discover patterns and relationships in large datasets.

3

1990S-2000S

Maturing of AI

Machine learning began to mature, with algorithms and models that could learn from data and improve over time.

4

2000S-2010S

Big Data and Cloud Computing

The explosion of big data introduced many challenges, hence more development. With Cloud Computing, database systems became more scalable and flexible.

5

2020S

AI-Driven Database Systems

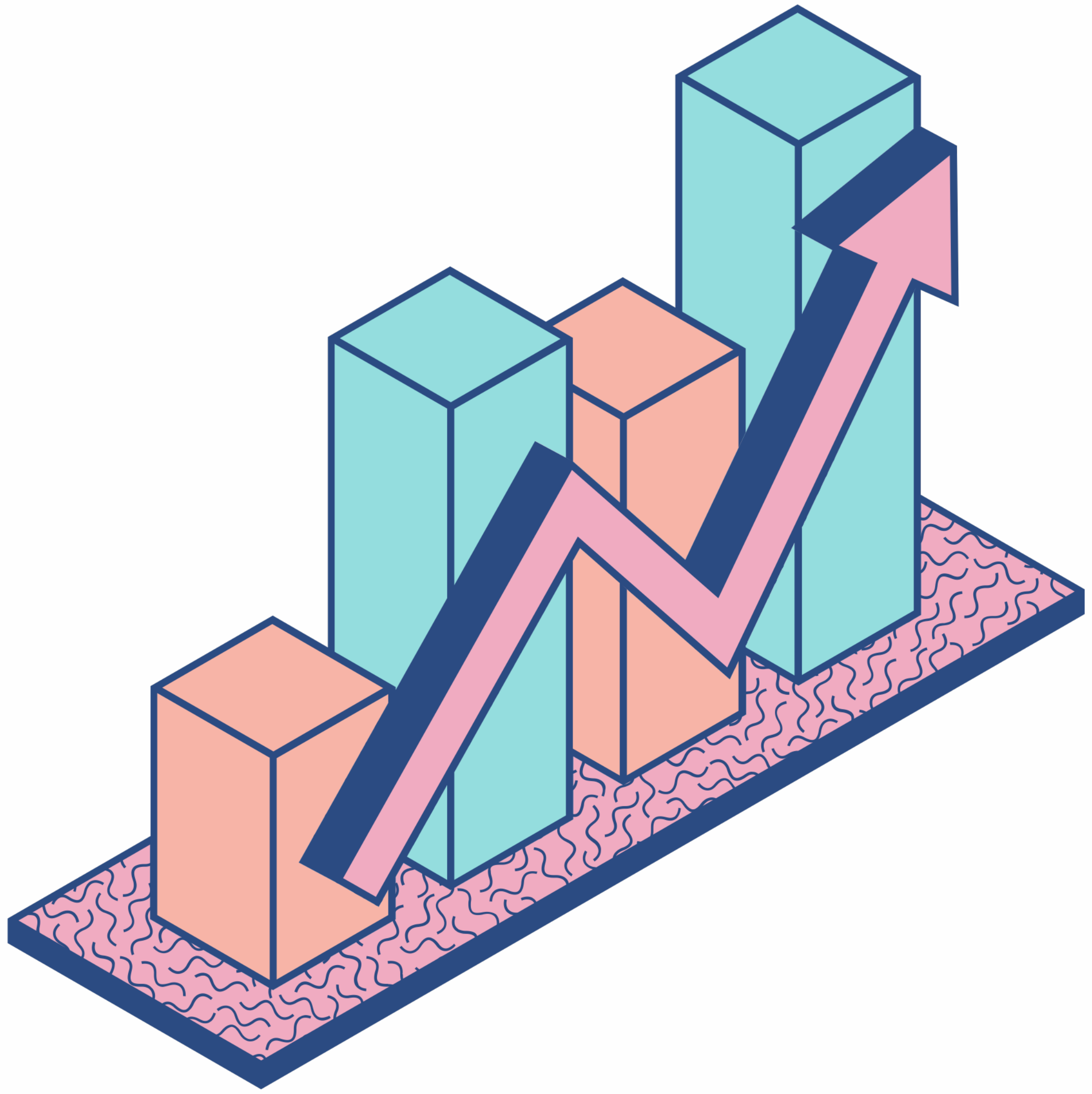
The current era is filled with AI-driven database systems that integrate ML models for real-time analytics, NLP, and automated decision-making.

Comparative Analysis of Systems: Technical Features and Functionality

Cloud Data Warehouse Solutions

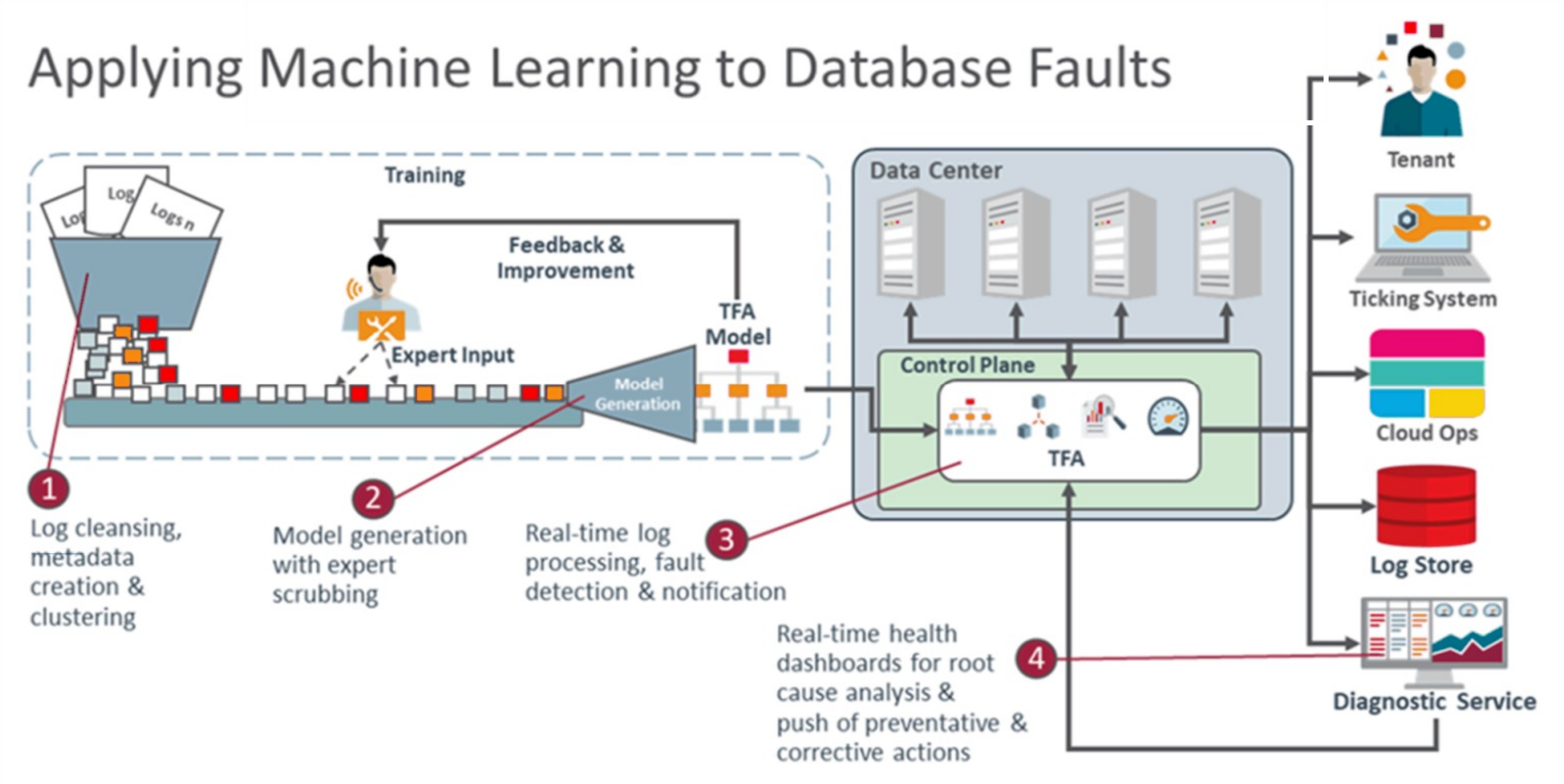
COMPARISON OF STATISTICAL & ML FEATURES

FEATURE	ORACLE ADW	IBM DB2 WAREHOUSE	AWS REDSHIFT	GOOGLE BIGQUERY	SAP DATA WAREHOUSE	SNOWFLAKE
Built-in ML integration	Y	Y	Y	Y	Y	Y
Scalability	Y	Y	Y	Y	Y	Y
ETL tool connectivity	Y	Y	Y	Y	Y	Y
In-memory analysis	Y	Y	N	Y	N	N
Real-time data sharing	Y	N	N	Y	Y	Y
Automated Data Transfer	N	N	N	Y	Y	N



Oracle Autonomous Database

- Cloud-based database solution
- Automates traditional management tasks
- Self-driving, Self-securing, Self-repairing
  - Self-driving: Automate provisioning, tuning, scaling.
  - Self-securing: Automate data protection, patching, access control.
  - Self-repairing: Automate failure detection, failover, repair.



Applying Machine Learning to Database Faults

1 Log cleansing, metadata creation & clustering

2 Model generation with expert scrubbing

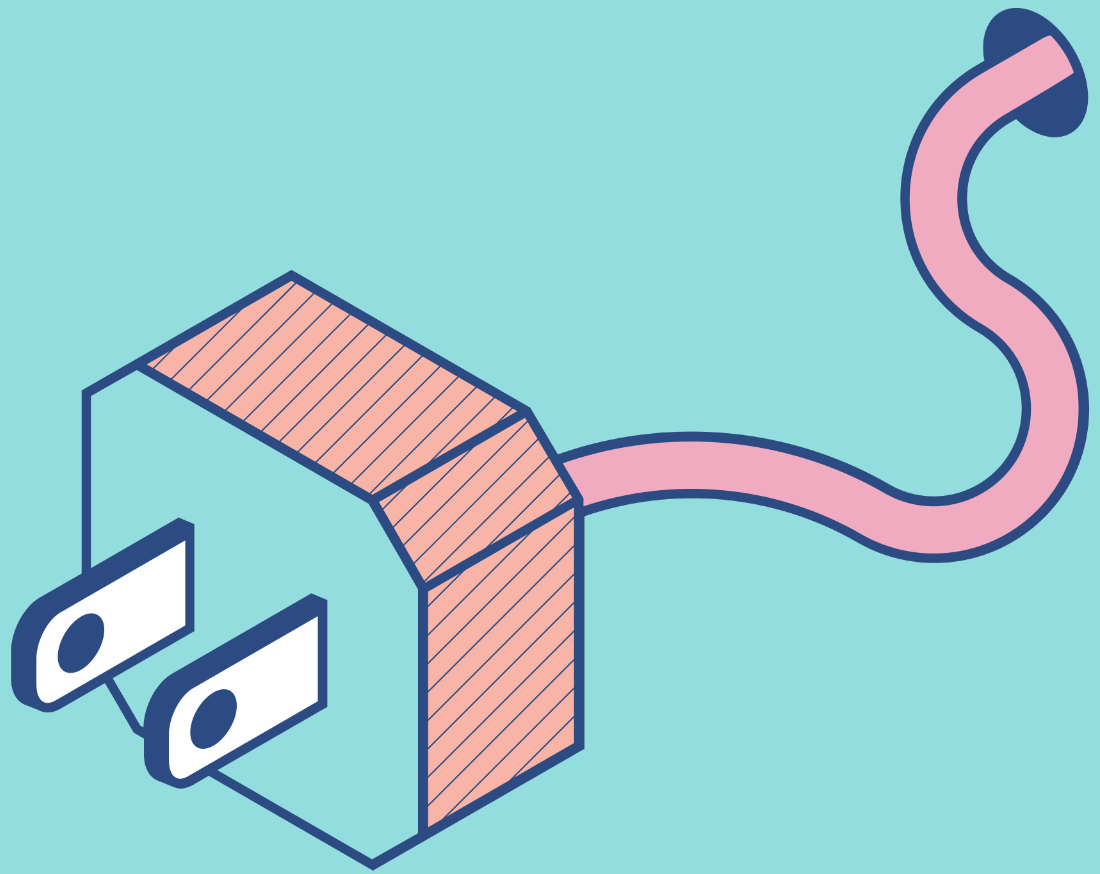
3 Real-time log processing, fault detection & notification

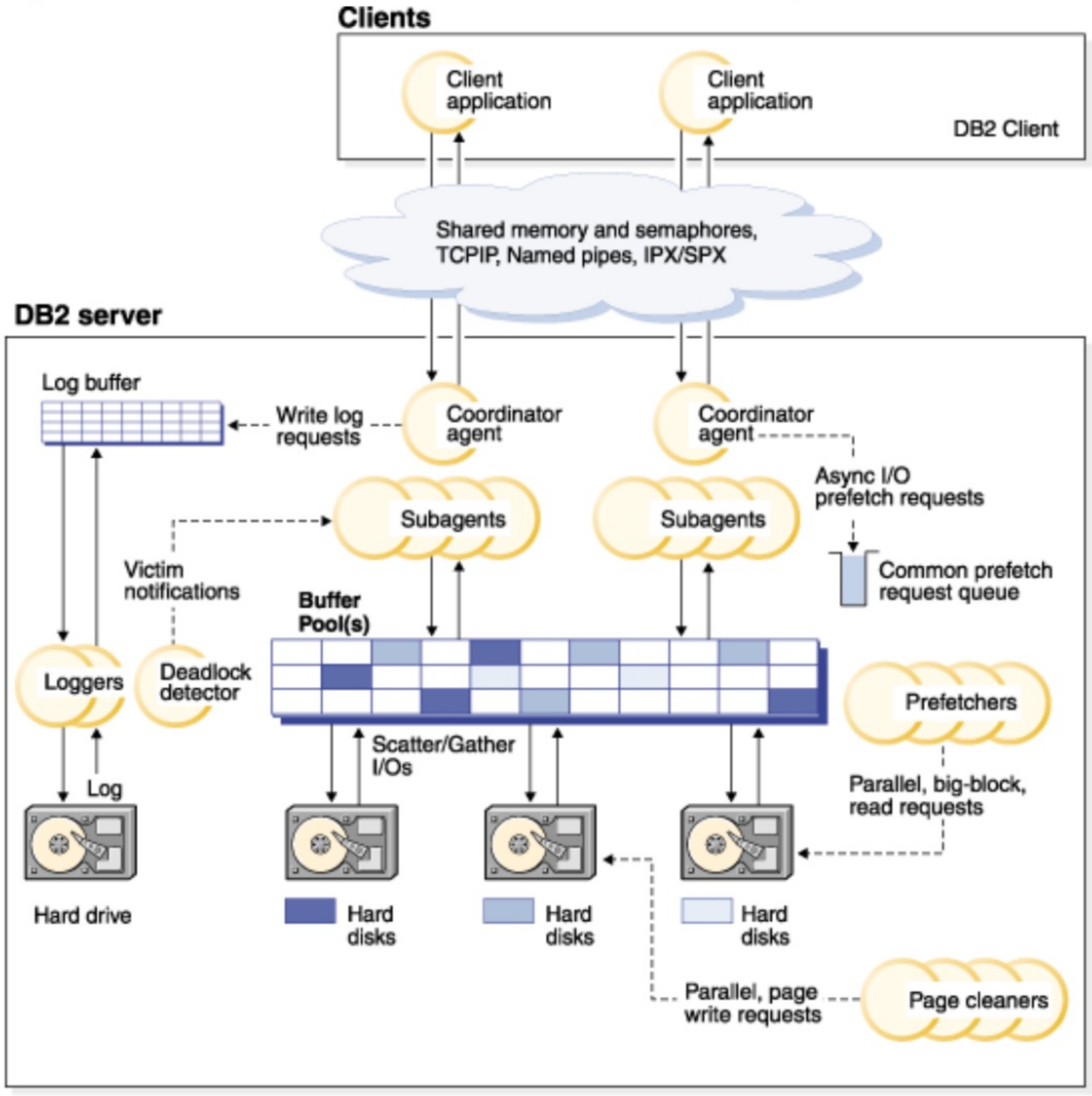
4 Real-time health dashboards for root cause analysis & push of preventative & corrective actions

Schubmehl, D. S., & Olofson, C. O. (2018, February). Applying Machine Learning to Database Faults. Oracle’s Autonomous Database: AI-Based Automation for Database Management and Operations. https://www.oracle.com/a/ocom/docs/database/idc-oracles-autonomous-database.pdf

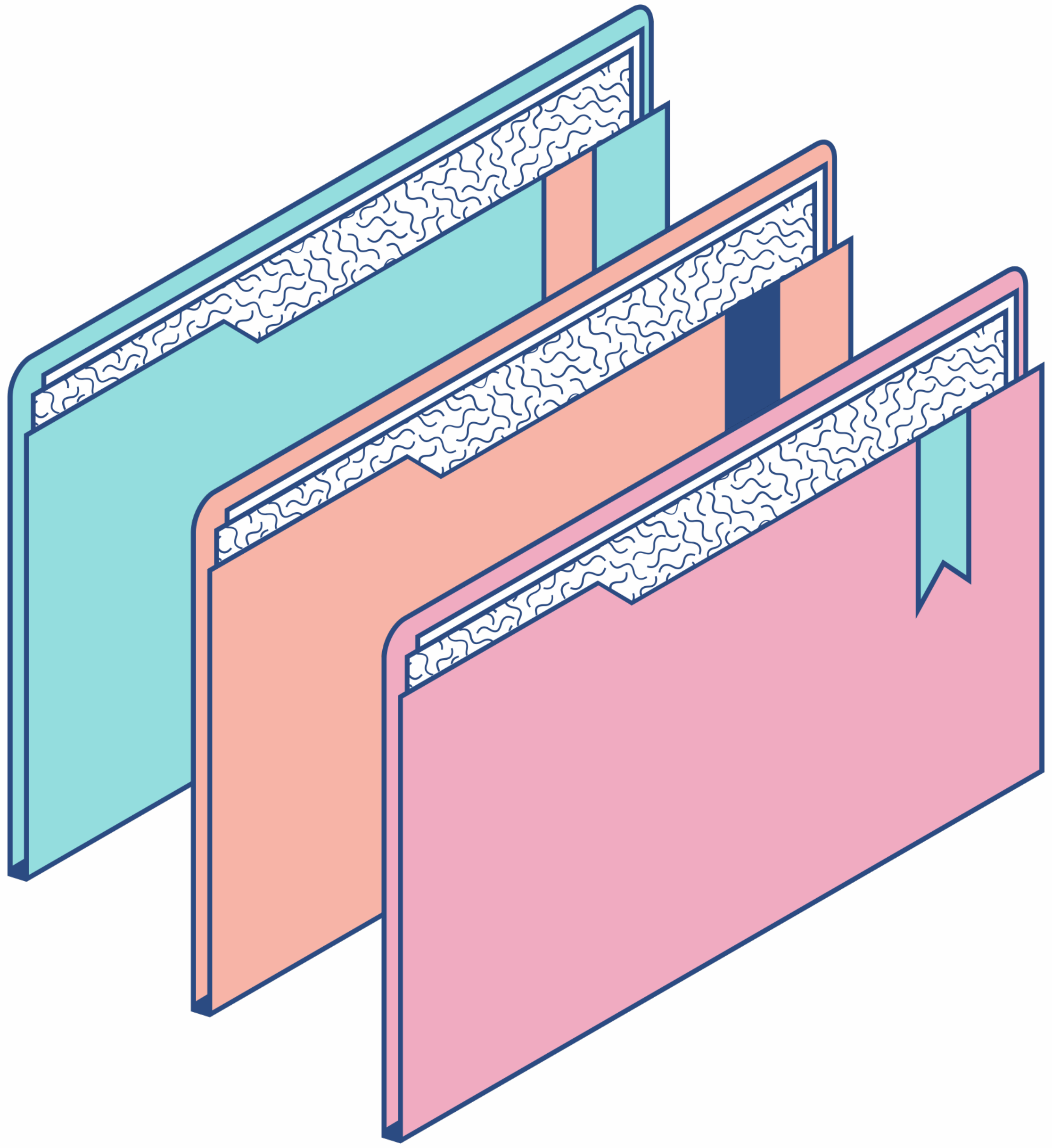
IBM Db2

- IBM Db2 is a cloud-native database designed for low latency transactions and real-time analytics at scale.
- Offers continuous availability, refined security, and effortless scalability for mission-critical applications.
- Utilizes AI-driven insights to enhance database performance and efficiency.





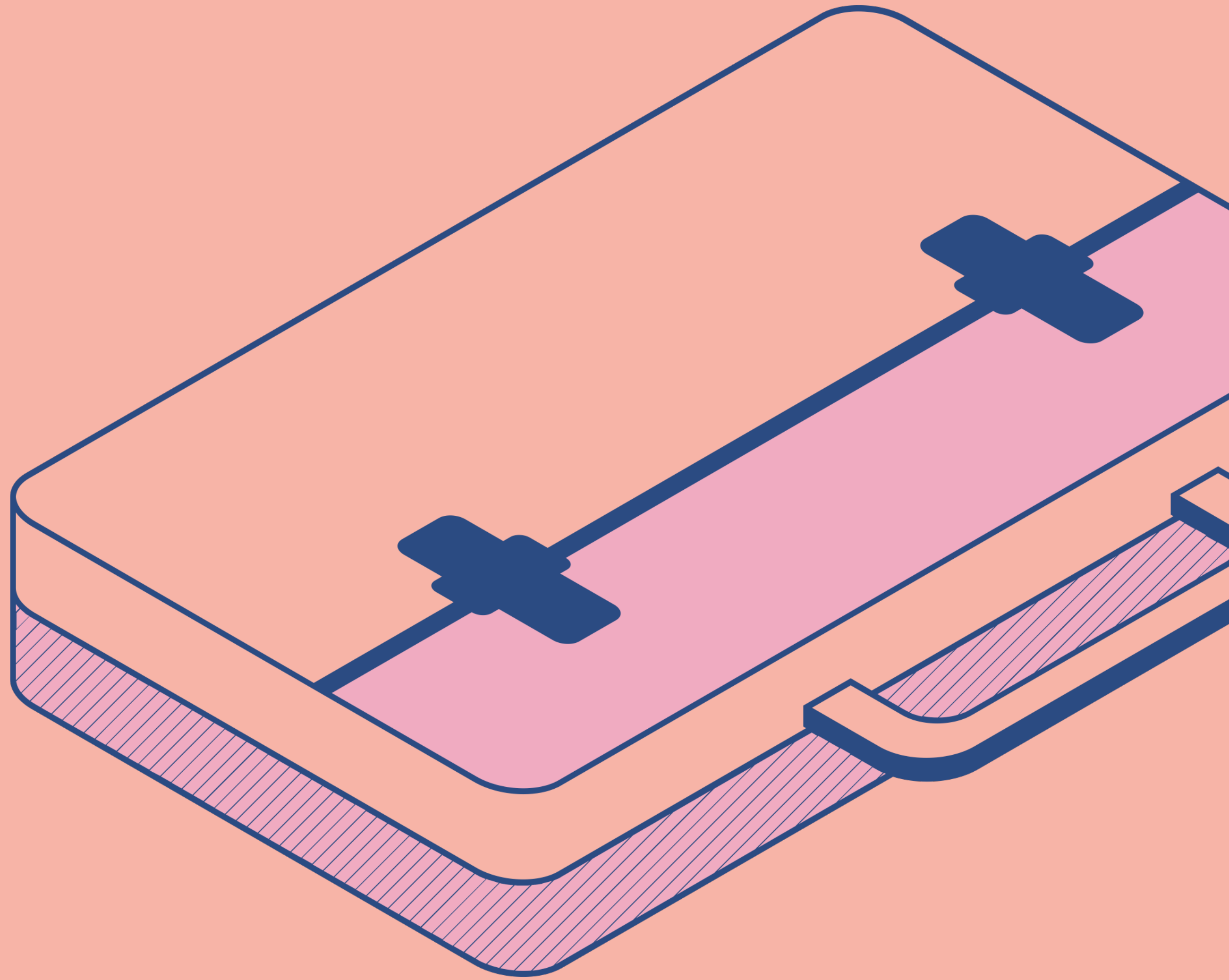
IBM documentation. (n.d.). https://www.ibm.com/docs/en/db2/11.5?topic=architecture-db2-process-overview



Automatic Statistics Collection

- Db2 optimizer relies on catalog statistics for efficient query access planning.
- Automatic statistics collection eliminates the need for manual intervention
- Collection can occur synchronously at statement compilation time
- Asynchronously in the background, ensuring real-time or interval-based updates.

Companies using Oracle Autonomous Database



Siemens

Optimize the Information Collection Process

[Decimal Point Analytics Private Limited]() (DPA)

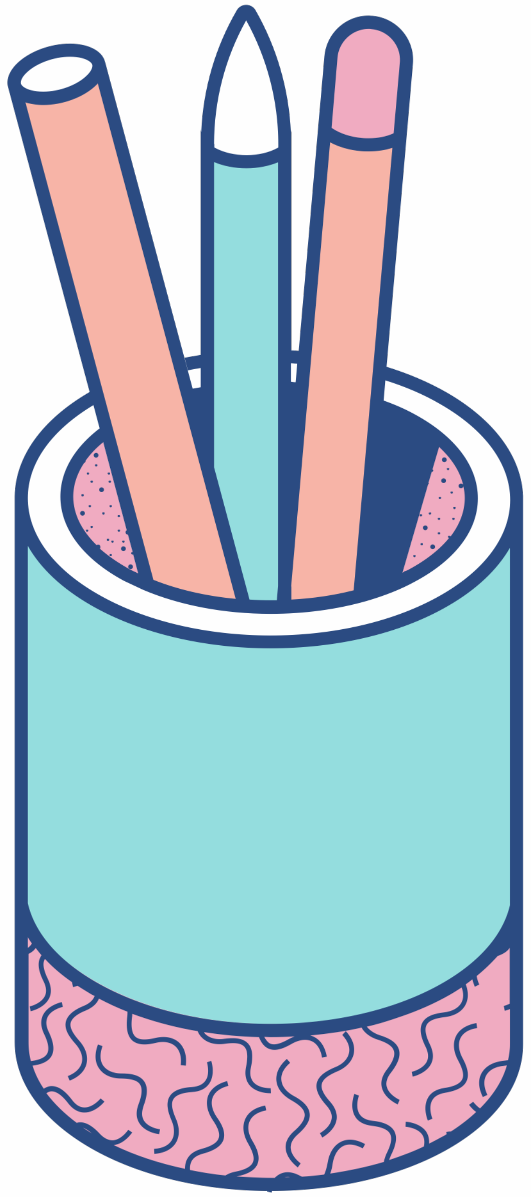
Reduce the data processing time and increase the data quailty and accuracy

Mestec

Reduce time consumed by IT processes

Outfront

Speed the process of loading third-party data



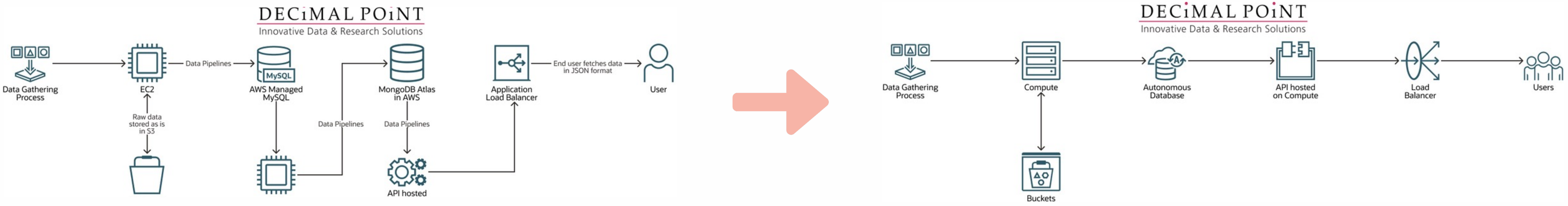
Case Study: Decimal Point Analytics Private Limited (DPA)

- **Goal**:
  - Find new architecture to reduce complexity of existing workflow and possibility of error

- **Solution**
  - Replace complex multi-database architecture on Amazon Web Service (AWS) using MySQL and Mongo Atlas with single converged Autonomous Database on Oracle

Case Study: Decimal Point Analytics Private Limited (DPA)

- **Result**:
  - Improved customer satisfaction through higher quality and 15% faster time-to-production of projects.
  - Increased application performance through simpler architecture and Oracle Autonomous Database capabilities.
  - Significant operational cost savings of approximately 10% through less infrastructure, application, and database management.



Baer, H. B., & Spiegel, J. S. (2023, June 2). DPA overcomes project complexity, improves performance with Autonomous Database. Oracle. https://www.oracle.com/customers/dpa-case-study/

Companies using IBM DB2

Owens-Illinois

Optimize storage footprint and improve transaction speed

[Marriott International]

Better performance in its analytics on 140 million+ Marriott Bonvoy members.

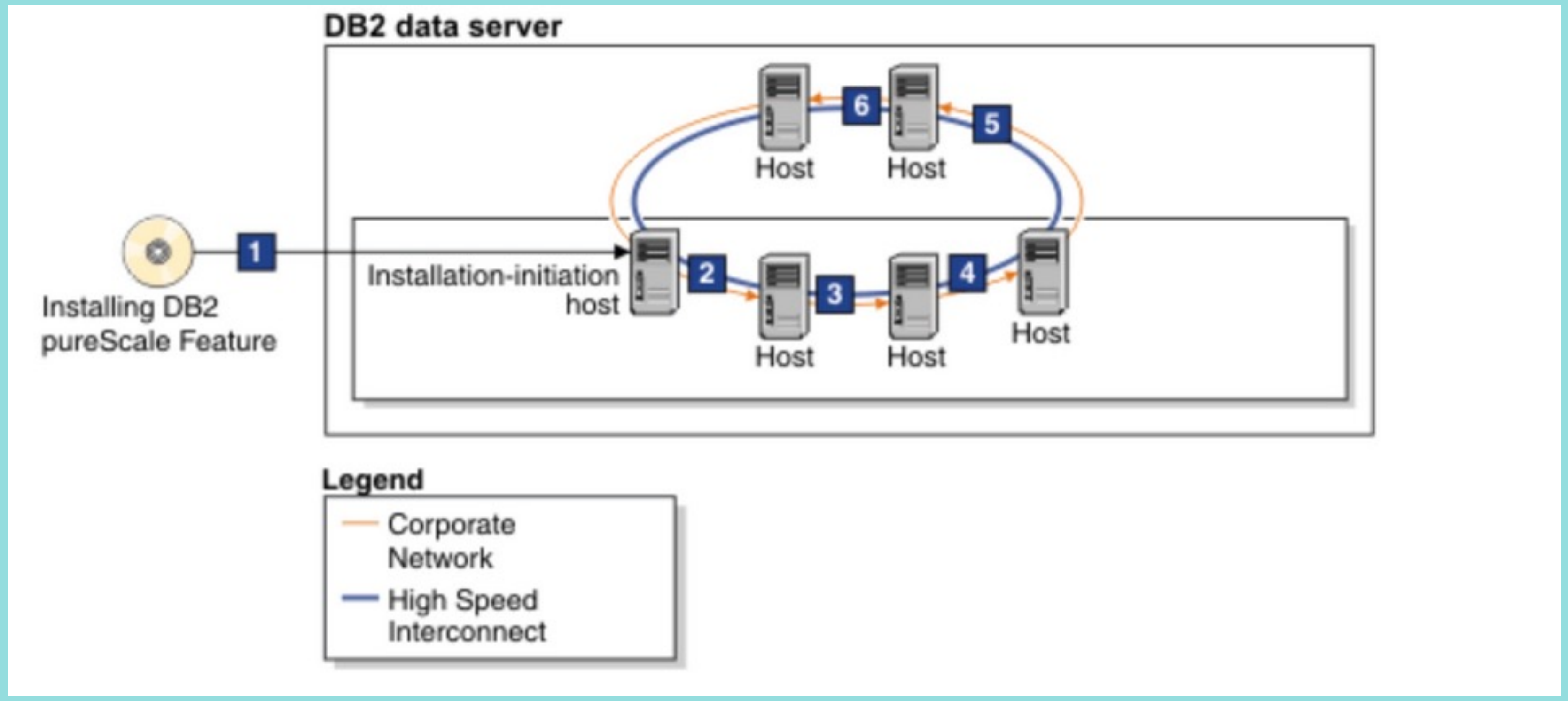
State Bank of India

Deliver a more streamlined customer experience and developed more targeted service offerings

Case Study: Puma

Goal: Find a technology for their database environment to process the rising transaction load

Solution: Used the IBM Db2 pureScale Feature to manage and simplify large database systems



IBM. (n.d.-b). IBM Db2 Warehouse. https://www.ibm.com/products/db2/warehouse

Result:

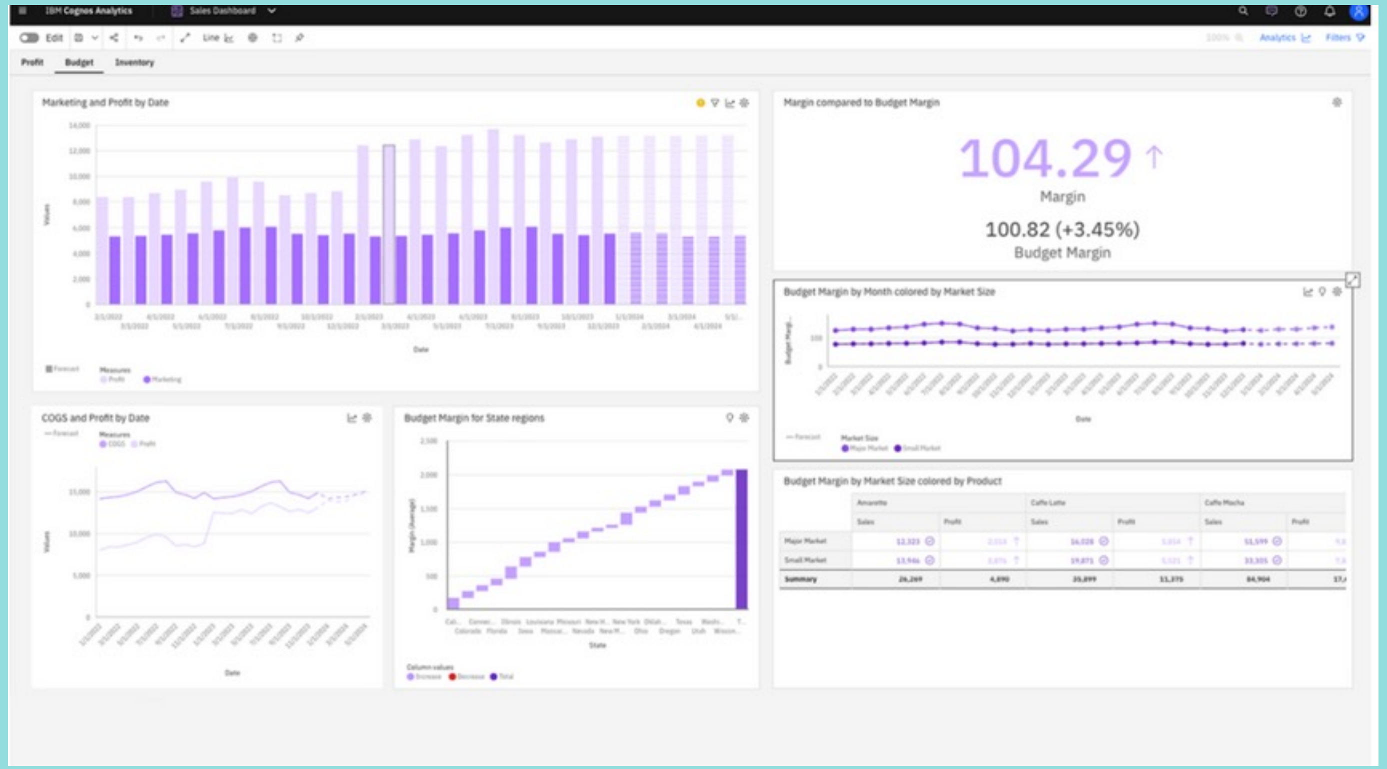
- Load tests found that the four-server pureScale cluster supported four to five times more users than before.
- Improved availability allows server upgrades without taking the system offline, and a failed server causes another member to take on its workload.

Case Study: Marriott International

Goal: Find data platform to get real-time insights by analyzing customers’ data and tools to consume, produce, and share data simultaneously for better serving

?

Solution: Utlized the IBM Db2 warehouse built in Db2

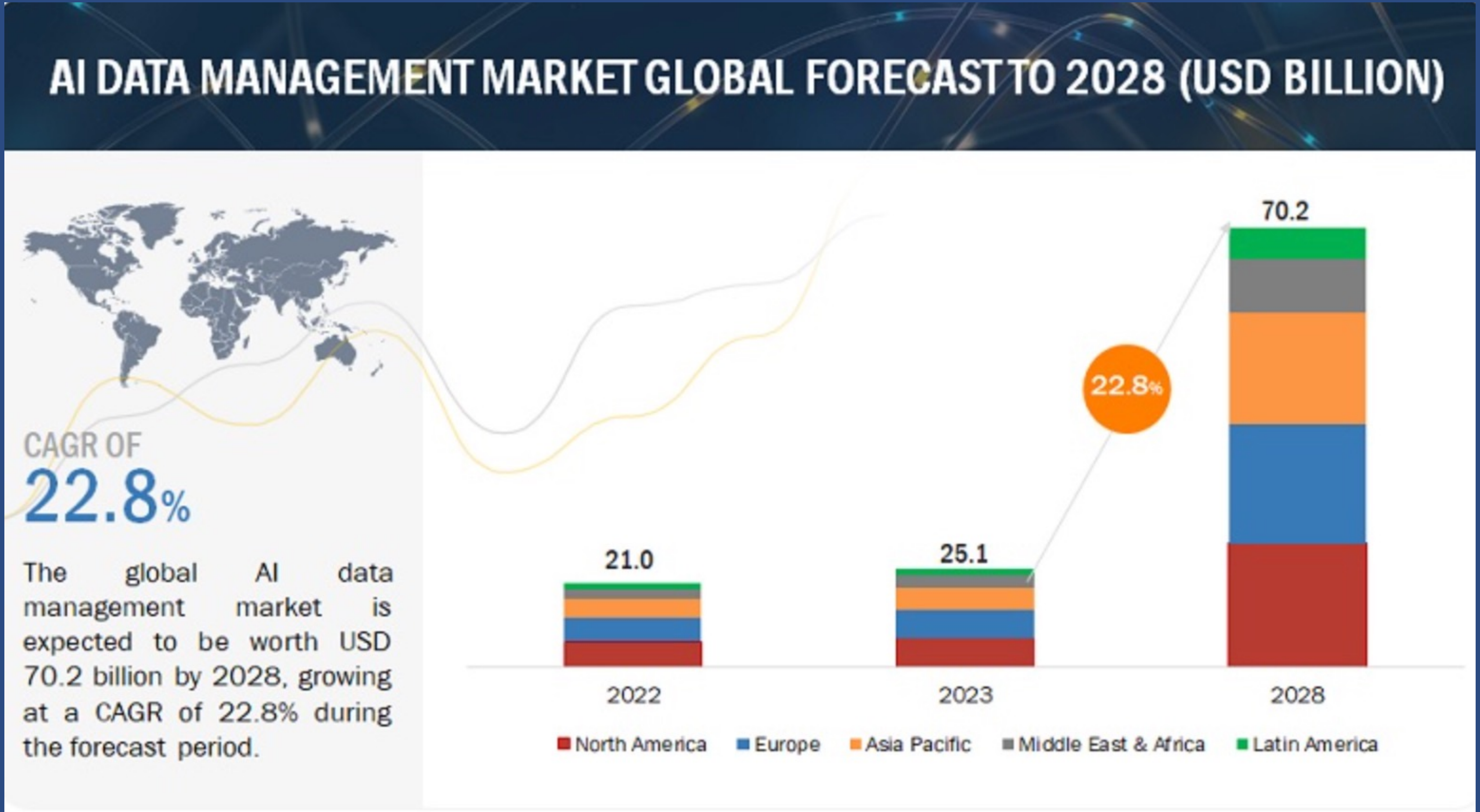


IBM. (n.d.-b). IBM Db2 Warehouse. https://www.ibm.com/products/db2/warehouse

Result:

- Marriott International achieves 90% faster performance in its analytics on 140 million+ Marriott Bonvoy members.
- Build real-time dashboards and reports with a combination of in-memory and column-store data retrieval to offer personalized service

Marketing Data



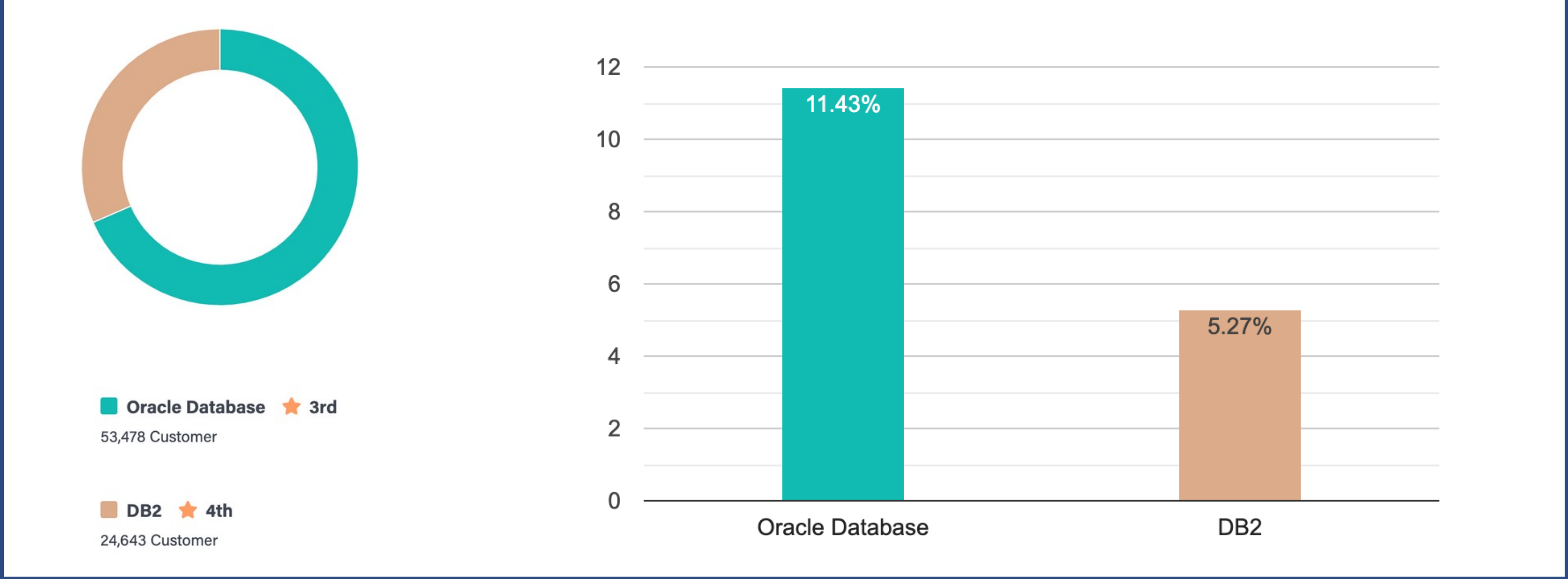
AI DATA MANAGEMENT MARKET GLOBAL FORECAST TO 2028 (USD BILLION)

CAGR OF 22.8%

The global AI data management market is expected to be worth USD 70.2 billion by 2028, growing at a CAGR of 22.8% during the forecast period.

MarketsandMarkets. (n.d.). AI Data Management Market size, share and global market Forecast to 2028 | MarketsandMarkets. https://www.marketsandmarkets.com/Market-Reports/ai-data-management-market-69639242.html

Comparing the Oracle Database and IBM DB2



6sense. (n.d.). Oracle Database vs DB2: Relational Databases Comparison. https://6sense.com/tech/relational-databases/oracledatabase-vs-db2

Future prognosis

USING GENERATIVE AI FOR DATA MANAGEMENTBUT IS IT THE FUTURE?

Data Quality Improvement
improving data quality by identifying and correcting errors, inconsistencies, and missing values in databases

Automated Code Generation
assist developers in generating code for database applications, including query optimization, schema design, and data transformation tasks

Relevant Problems

ANY RELEVANT PROBLEMS THAT RESEARCHERS ARE WORKING ON RELATED TO THIS AREA

Natural Language Processing for Database Queries

Problem:
Interacting with databases typically requires knowledge of specific query languages, which can be a barrier for non-technical users.

Research Direction:
Leveraging NLP to enable natural language querying of databases, making data access more intuitive and accessible to a broader audience.

Advanced Query Optimization

Problem:
As data volumes grow, finding the most efficient way to execute queries becomes increasingly complex and critical for performance.

Research Direction:
Applying ML algorithms to predict the best query execution plans based on historical data, considering various factors like data distribution, query complexity, and system load.

Researches

RESEARCHES RELATING TO THE PROBLEMS PRESENTED PREVIOUSLY

“Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning”

- Addresses the challenge of translating natural language queries into structured SQL queries, a critical aspect of enhancing databases with natural language processing (NLP) capabilities.
- Presents an approach using a sequence-to-sequence (seq2seq) model combined with reinforcement learning to generate SQL queries from natural language inputs. This method allows the model to learn from user feedback and improve its query generation capabilities over time.

“Learning to Optimize Join Queries With Deep Reinforcement Learning”

- Discusses the challenge of optimizing join queries in databases using deep reinforcement learning (DRL).
- Presents a DRL-based model that learns to select the most efficient join order for a given query. This approach is significant because it moves beyond traditional static rules or cost models, utilizing a learning model that can improve and adapt over time.

Zhong, V., Xiong, C., and Socher, R., “Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning”, *arXiv e-prints*, 2017. doi:10.48550/arXiv.1709.00103.
Krishnan, S., Yang, Z., Goldberg, K., Hellerstein, J., and Stoica, I., “Learning to Optimize Join Queries With Deep Reinforcement Learning”, *arXiv e-prints*, 2018. doi:10.48550/arXiv.1808.03196.

Questions