Group A1- Data Mining Tools and Products; OLAP Products

Isabela Fernandez, Jacob King, Sean Liu, Dora Sasson

What is Data Mining?

- Discovering insights and patterns from large datasets to make decisions [2]
- Transforming raw data into useful knowledge [2]
- Uses machine learning (e.g. classification, clustering, regression) to analyze and infer upon data [2]
- Integrates with data analysis and visualization tools [1]
- Also known as knowledge discovery in data (KDD) [2]

[1] Hillier, W. (2022, December 7). *The 7 best data mining tools in 2024*. CareerFoundry. https://careerfoundry.com/en/blog/data-analytics/best-data-mining-tools/

[2] *What is data mining?*. IBM. (n.d.). https://www.ibm.com/topics/data-mining

What is OLAP?

- OnLine Analytical Processing
- Analyzes multidimensional data from multiple datasets to optimize business decisions [1, 2]
- Makes quick, accurate decisions [2]
- Architecture: [1]
  - data warehouse- info from multiple sources
  - ETL (extract, transform, load) tools- retrieve and prepare data
  - OLAP server- backend machine
  - OLAP database- prevents overburdening data warehouse
  - OLAP cubes- rigid multidimensional array of information

[1] What is OLAP? - online analytical processing explained - AWS. (n.d.-b). https://aws.amazon.com/what-is/olap/

[2] *What is OLAP?*. IBM. (n.d.-b). https://www.ibm.com/topics/olap

Applications in the Business Context

- Data Mining [2]:
  - Sales and marketing- optimizing marketing, improving customer loyalty
  - Education- student performance evaluation
  - Fraud detection- identifying fake user accounts, especially in banks
  - Optimization- improving efficiency of reducing costs
- OLAP [1]:
  - Sales and financial reports
  - Resource management
  - Supply and demand forecasting
  - Workflow management

[1] GeeksforGeeks. (n.d.). *OLAP applications*. GeeksforGeeks. https://www.geeksforgeeks.org/olap-applications/

[2] *What is data mining?*. IBM. (n.d.). https://www.ibm.com/topics/data-mining

Brief History of Data Mining

- ML math foundations: Bayes' Theorem (1763), regression analysis (1805) [2]
- ML technology: neural networks (1943), development of databases (1970s) [2]
- 1960s: Originated as a subset of AI [1]
- 1970s: Apriori algorithm- unsupervised learning to identify associative relations between different variables in a dataset [1]
- 1990s: Powerful computing and storage popularized data mining [1]

[1] Mishra, L. (2023, July 18). *Data Mining: History, techniques, advantages, and example*. Success Guaranteed IIBA Training. https://www.adaptiveus.com/blog/business-analyst/technique/data-mining/

[2] What is data mining? A beginner's guide (2022). Rutgers Bootcamps. (2021, December 23). https://bootcamp.rutgers.edu/blog/what-is-data-mining/

Brief History of OLAP

- Before 20th century: Digital data stored in punch cards [1]
- 1962: OLAP first proposed as a programming language [1]
- 1975: first OLAP product released to support marketing [1]
- 1982: OLAP extended for financial applications [1]
- 1990s: Microsoft released OLAP services [1]

[1] *OLAP and business intelligence history*. OLAP.com. (n.d.). https://olap.com/learn-bi-olap/olap-business-intelligence-history/

Data Mining- Typical Features and Functions

- Ease of use- informative visualizations, intuitive user interfaces, minimal coding [1]
- Compatibility across data file formats, operating systems, and programming languages [1]
- Speed and efficiency [1]
- Scalability for large volumes of data [1]
- ML abilities- data cleaning, data exploration, prediction, classification, business intelligence, etc. [2]

[1] Bennett, T. (2023, October 6). *Top 9 data mining tools for Business Insights*. Integrate.io. https://www.integrate.io/blog/data-mining-tools/

[2] Sharma, R. (2022, August 30). *7 data mining functionalities every data scientists should know about*. upGrad . https://www.upgrad.com/blog/data-mining-functionalities/

Common Data Mining Tools

- Oracle Data Miner: [1]
  - ❖ Drag-and-drop features
  - ❖ Ability to concurrently run multiple processes
  - ❖ Automatically runs multiple ML algorithms for comparison
  - ❖ Designed for ease and accessibility

[1] Data miner. (n.d.-a). https://www.oracle.com/big-data/technologies/dataminer/

Common Data Mining Tools

- Orange:
  - Hands-on training [1]
  - No coding from users, and interactive data visualizations [1]
  - Extensions for working with external data, NLP, text mining, etc. [1, 2]
  - Support for most file formats [1]

[1] Orange Data Mining. (2024, January 10). https://orangedatamining.com/

[2] IONOS. (2023, March 1). *Data Mining Tools for better data analysis*. IONOS Digital Guide. https://www.ionos.com/digitalguide/online-marketing/web-analytics/a-comparison-of-data-mining-tools/

Common Data Mining Tools

- Dundas BI:
  - Easy-to-use, intuitive interface for data exploration and analysis [1]
  - Interactive data visualizations and displays in multiple formats [2]
  - Streamlined workflow through automated data preparation [1]
  - Compatibility with virtually any extension [1]

[1] *Dundas Bi*. insightsoftware. (n.d.). https://insightsoftware.com/dundas/

[2] Simplilearn. (2024, February 19). *Top 14 data mining tools you need to know in 2024 and why*. Simplilearn.com. https://www.simplilearn.com/data-mining-tools-article

Common Data Mining Tools

- Altair RapidMiner: [1]
  - Instant automated data cleaning
  - Rapid data visualization, without need for writing code
  - Integrates with R code
  - Uses traditional statistical methods and newest ML models

[1] *Altair® RapidMiner®*. altair.com. (n.d.). https://altair.com/altair-rapidminer

Common Data Mining Tools

- Apache Spark: [1]
  - ❖ Compatible with Python, Java, Scala, SQL, and R
  - ❖ Executes SQL faster than most data warehouses (built on distributed SQL engine)
  - ❖ Analyzes petabytes of data as given
  - ❖ Integrates with frameworks including TensorFlow, PyTorch, PowerBI, and Tableau

[1] *Apache sparkTM - unified engine for large-scale data analytics.* Apache SparkTM . (n.d.). https://spark.apache.org/

Common Data Mining Tools

- Knime:
  - Automation of routine tasks [1]
  - Analyzes data without need for code [1]
  - Access to all major ML libraries [2]
  - Great for predictive analytics and business intelligence [2]

[1] KNIME. (n.d.). https://www.knime.com/

[2] IONOS. (2023, March 1). *Data Mining Tools for better data analysis*. IONOS Digital Guide. https://www.ionos.com/digitalguide/online-marketing/web-analytics/a-comparison-of-data-mining-tools/

Common Data Mining Tools

- Amazon Elastic MapReduce (EMR): [1]
  - Optimized, flexible resource utilization
  - Secure, scalable data storage (e.g. Hadoop, Dynamo DB, Redshift)
  - EMR Studio- IDE that supports multiple languages, debugging, and visualizations

[1] Amazon EMR features - big data platform - amazon web services. (n.d.). https://aws.amazon.com/emr/features/

OLAP- Typical Features and Functions

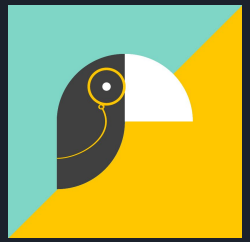
- Detailed, comprehensive data analysis [1]
- Ability to quickly access and manipulate enormous volumes of data [1]
- Informative data visualizations and intuitive user interfaces [2]
- Integration with databases and datasets, including cloud data [1]

[1] Gray, P. (2024, January 22). *24 best OLAP tools reviewed for 2024*. The RevOps Team. https://revopsteam.com/tools/best-olap-tools/

[2] *Characteristics of OLAP - javatpoint*. www.javatpoint.com. (n.d.). https://www.javatpoint.com/characteristics-of-olap

Common OLAP Tools

- Toucan: [1]
  - ❖ Uses storytelling and high-quality visuals to communicate data and trends
  - ❖ No need to code when preparing to data
  - ❖ Convenient database and statistics functionalities
  - ❖ Integration with cloud data warehouses (e.g. Snowflake, AWS Redshift, Google BigQuery)



[1] *Data storytelling for your customers*. Toucan. (n.d.). https://www.toucantoco.com/en/?r=rev-bolapt

Common OLAP Tools

- Tableau: [1]
  - ❖ Automates routine tasks
  - ❖ Intuitive interface for creating visualizations and applying ML models
  - ❖ Natural language processing to answer users' questions
  - ❖ Efficient features for data preparation (Tableau Prep), cloud-based analytics (Tableau Cloud), and visualizations (Tableau Desktop)

[1] *Business Intelligence and Analytics Software*. Tableau. (n.d.). https://www.tableau.com/

Common OLAP Tools

- Adverity: [1]
  - Processes and analyzes data without need for code
  - Built-in packages for business, sales, and marketing
  - Easy to monitor and schedule incoming data
  - Compatible with data lakes, cloud storage, and visualization software

[1] *The Integrated Data Platform for teams that run on Data*. Adverity. (n.d.). https://www.adverity.com/?r=rev-bolapt

Common OLAP Tools

- Domo: [1]
  - ❖ Creates business apps with minimal coding from users
  - ❖ Drag-and-drop ETL interface
  - ❖ Integration with cloud data warehouses
  - ❖ AI-powered, intuitive data visualizations

[1] *Discover the domo data experience platform*. Domo. (n.d.). https://www.domo.com/?r=rev-bolapt

Common OLAP Tools

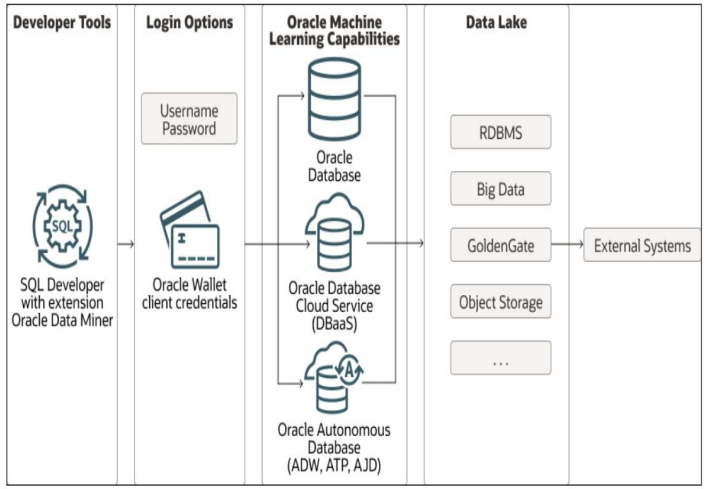
- Sisense: [1]
  - ❖ Efficient AI/ML and analytics engine for predictions
  - ❖ Compatible in any cloud infrastructure
  - ❖ Customizable, interactive data visualizations
  - ❖ Exports data and insights to common apps (e.g. Excel)

[1] *Build intelligent analytics into your products*. Sisense. (n.d.). https://www.sisense.com/

Analysis of Oracle Data Miner - Architecture

- Extension of Oracle SQL Developer
- ML to create, run, and manage workflows
- ODMRSYS schema as a dedicated system repository
- Uses Oracle Database features
  - Oracle Machine Learning for SQL
    - Provides the model building, testing, and scoring capabilities
  - Oracle XML DB
    - Manages the metadata
  - Oracle Machine Learning for R
    - User-defined R functions
  - Oracle Text
    - Text mining integrated with the modeling process.
  - Oracle Scheduler
    - Scheduling the Oracle Data Miner workflows.

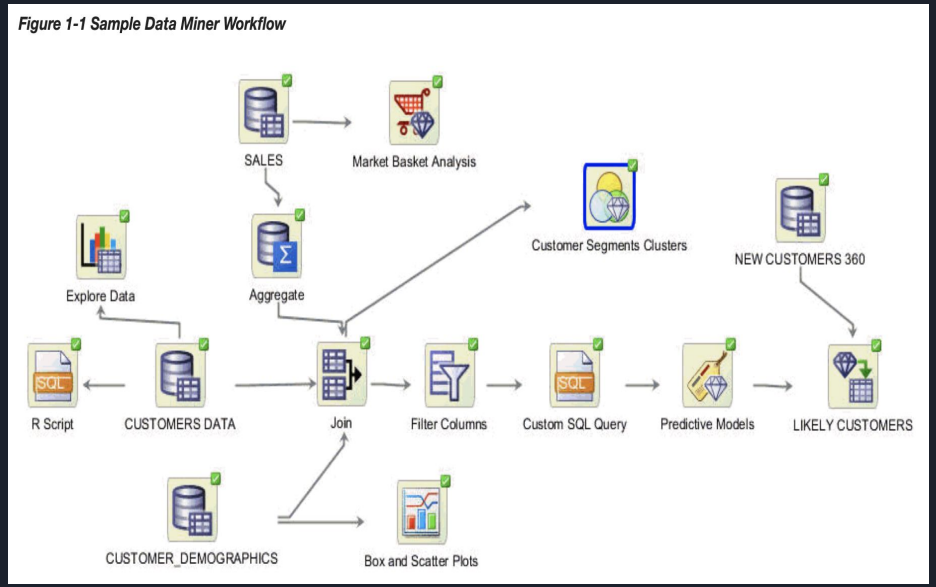
Figure 1-1 Oracle Data Miner Architecture



Oracle Data Miner System Overview. Oracle Help Center. (2023, March 31). https://docs.oracle.com/en/database/oracle/sql-developer/23.1/dmrig/oracle-data-miner-overview.html#GUID-E952F2CB-CEE5-47D4-934D-5A81957E9833

Analysis of Oracle Data Miner - Scenario

- Discover hidden patterns, relationships, and insights in their data.
- Data mining requires a problem definition, collection and cleansing of data, and model building
- Time spent in understanding data
- Enables data analysts to:
  - Work directly with data inside the database
  - Explore the data graphically
  - Build and evaluate multiple data mining models
  - Apply models to new data
  - Deploy predictions and insights throughout the enterprise



Data miner user's guide. Oracle Data Miner. (2015, June 16). https://docs.oracle.com/cd/E55747_01/doc.41/e58114/introodmr.html

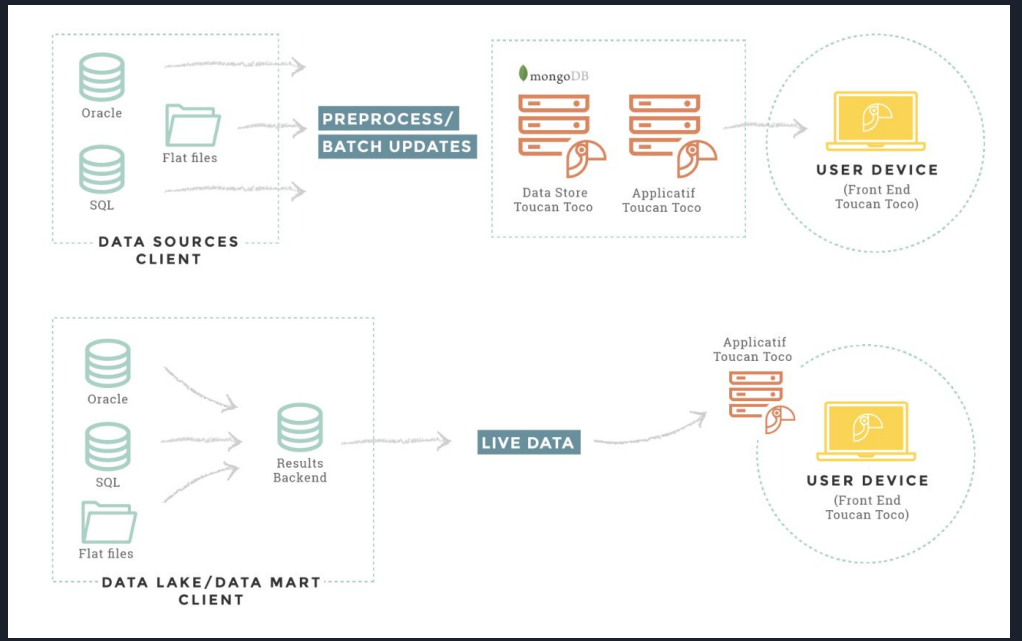
Analysis of Oracle Data Miner - Modes of Operation

- Mostly used to:
  - Predict customer behavior
  - Target best customers
  - Discover customer clusters, segments, and profiles
  - Identify customer retention risks
  - Identify promising selling opportunities
  - Detect anomalous behavior

Data miner user's guide. Oracle Data Miner. (2015, June 16). https://docs.oracle.com/cd/E55747_01/doc.41/e58114/introodmr.html

Analysis of Toucan OLAP Product - Architecture

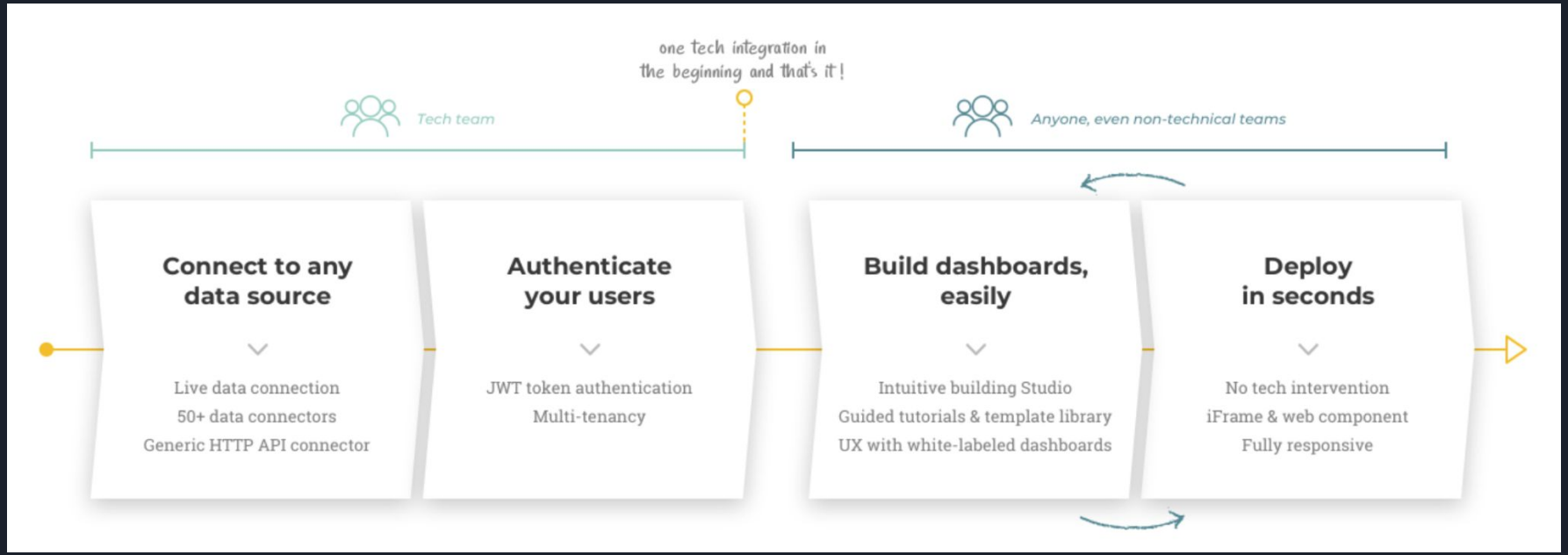
- A user centric approach to data management for analytics
- Prioritize user's requirements instead of technological aspects
- Motivating Question: "Can we guarantee a response time below 1 second from our data infrastructure?"
- Data is pre-aggregated and indexed with a user and business centric approach



Data infrastructure. Data infrastructure - Toucan Doco documentation. (n.d.).
https://docs.toucantoco.com/embed/02-data-infrastructure.html#the-toucan-s-way-a-user-centric-approach-to-data-management-for-analytics

Analysis of Toucan OLAP Product - Scenario

- Allows users to create a leading customer-facing analytics experience



Explore the customer-facing analytics platform. Toucan. (n.d.). https://www.toucantoco.com/en/product-overview

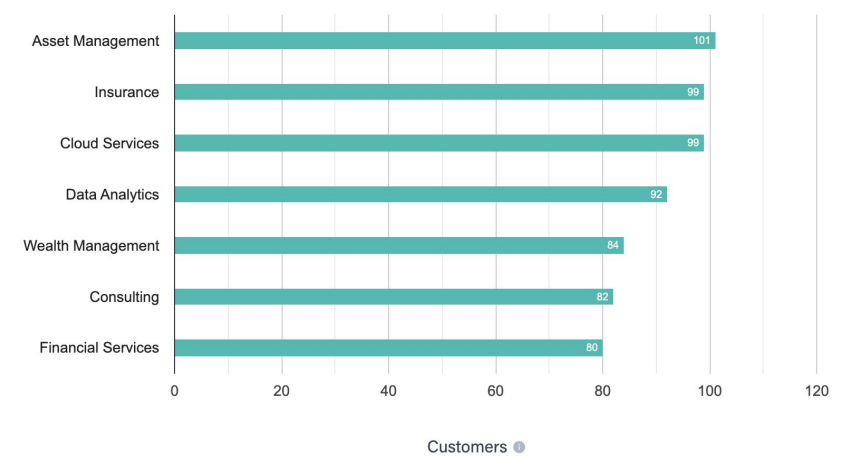
Analysis of Toucan OLAP Product - Modalities

- Storytelling Studio
  - Storytelling Framework and Guided Design to build pixel-perfect reports code- and design-free
- Broadcast and Embed
  - Embed Toucan in your SaaS product or workflow
- ActionCenter
  - Fit collaboration habits, integrates to Microsoft Teams and Slack
- YouPrep
  - No-code, on-the-fly transformation of data
- AnyConnect
  - Work seamlessly with live data coming from Snowflake, MongoDB, AWS
- Allows users to create a leading customer-facing analytics experience

Explore the customer-facing analytics platform. Toucan. (n.d.). https://www.toucantoco.com/en/product-overview

Major Companies Using Oracle Data Miner

- Amazon
- JP Morgan Chase & Co.
- Coca - Cola Company
- Deutsche Bank
- Netflix
- Lockheed Martin Corporation



Oracle Data Mining - market share, competitor insights in Data Mining. 6sense. (n.d.). https://6sense.com/tech/data-mining/oracle-data-mining-market-share

Financial Institutions and Oracle Data Mining

Use cases of Oracle Data Miner in Financial Institutions:

- Risk Management
- Customer Segmentation and Targeting
- Fraud Detection
- Marketing Analytics
- Portfolio Management
- Compliance and Regulatory Reporting

Deutsche Bank use of Oracle Data Miner

- Partner with Oracle
- Invested in modernizing its technology infrastructure
- The bank will run 10,000 Oracle Databases
- Leverage of Oracle Exadata Cloud@Customer, including Oracle Data Miner
- Used for customer segmentation, fraud detection, risk assessment and market analysis
- Predictive models and algorithms to extract insights of data.
- Integration with applications and workflows
- Continuous monitoring and optimization

Ricadela, A. (2021, Sep. 10). *Deutsche Bank's technology transformation includes benefits of Oracle's private database cloud.* Financial Services Blog. https://blogs.oracle.com/financialservices/post/deutsche-bank-technology-transformation-oracle-private-database-cloud

Benefits

- Automation of decision-making, so increasing efficiency
- Scale and complexity of data
- Enhanced Risk Management
- Gain of real-time insights
- Competitive advantage in the financial industry

Ricadela, A. (2023, May 2). *Amid a rising data tide, Deutsche Bank passes a tech milestone.* Oracle Connect. https://www.oracle.com/cloud/deutsche-bank-technology-transformation-oracle-private-database-cloud/

Marketing Data for Oracle Data Miner

Target: data scientists and business analysts, working with large datasets within Oracle databases.

Market Share: 17.56% in the Data Mining category

Customers: 2,734 in 10 countries

Competition: from established products like IBM SPSS Modeler, SAS Enterprise Miner, and open-source solutions like RapidMiner, KNIME, and scikit-learn.

Marketing Strategy: focus on promoting its broader data analytics solutions, which encompasses Data Miner, highlighting their deployment of predictive models, enabling data exploration and simplifying model through a GUI.

Oracle Data Mining Vs RapidMiner : In-Depth Comparison. 6sense. (n.d.).
https://6sense.com/tech/data-mining/oracledatamining-vs-rapidminer

Marketing Data for Toucan

Target: data storytelling and business intelligence software market

Annual Revenue: <50M USD

Competition: from established products like Tableau, Qlik, and Microsoft Power BI

Marketing Strategy: emphasizes the unique value proposition of transforming data into engaging stories, dedicated blog on storytelling, feature customer success stories

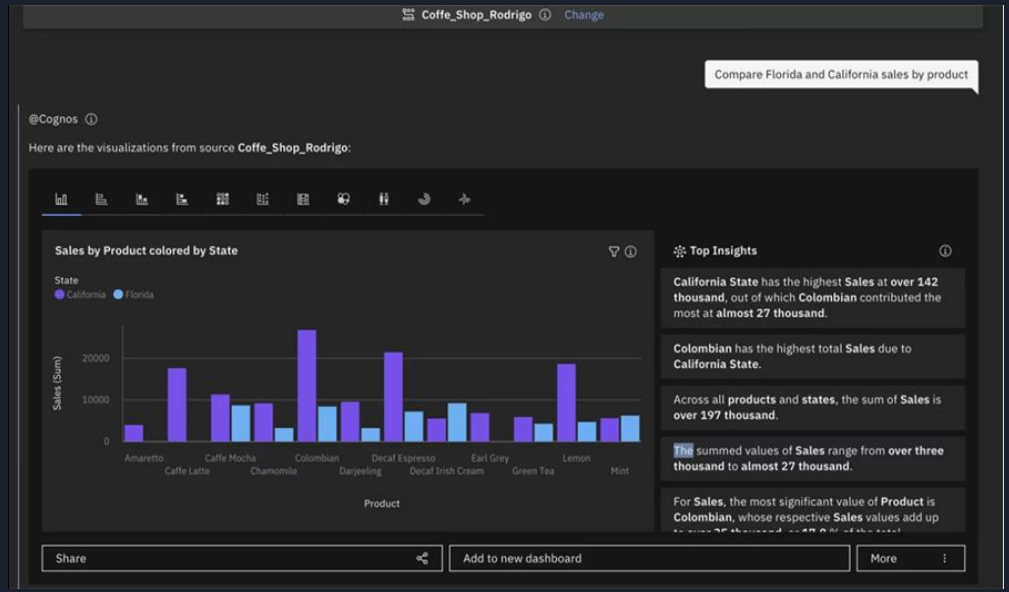
Toucan Reviews. Gartner (n.d.). www.gartner.com/reviews/market/analytics-business-intelligence-platforms/vendor/toucan/product/toucan

Changes in Data Mining/ OLAP Software

- Data Mining and OLAP Software appear to be expanding in three different ways
  - Accessibility: Making low/ no-code options more powerful
  - Range: Increasing the types of data that can be mined
  - Interoperability: Adding new interoperability with new types of databases

Accessibility

- The big trend for Data Mining software right now is adding AI assistants
  - ❖ Turning insights into natural language [1]
  - ❖ Turning prompts into SQL queries [2]
- Likely to become central feature for low/no-code data mining



Natural language in IBM Cognos Analytics [1]

[1] Leach, A. (2023, June 6). *Enabling AI-powered business intelligence across the Enterprise.* IBM. https://www.ibm.com/blog/enabling-ai-powered-business-intelligence-across-the-enterprise/
[2] *AI Chatbot for Apps*. MicroStrategy. (n.d.). https://www.microstrategy.com/enterprise-analytics/ai-chatbot-for-apps

Range

- Data mining has recently expanded to several new fields
  - Image Mining [1]
  - IOT
  - NLP
  - Audio Mining [2]
- While there are basic low-code computer vision software options, many current data mining software do not fully integrate it
- Future data mining software will likely incorporate these new fields

[1] A. Ennouni, Y. Filali, M. A. Sabri and A. Aarab, "A review on image mining," 2017 Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 2017, pp. 1-7, doi: 10.1109/ISACV.2017.8054968.

[2] P. Kumar, J. Kaur, R. Sandhu, M. Wamique and A. Yadav, "An Extensive Review on Different Strategies of Multimedia Data Mining," 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India, 2023, pp. 707-712, doi: 10.1109/IITCEE57236.2023.10091056.

Interoperability

- Current data mining software can access data from many different sources
  - ❖ Oracle Data Miner interops with Oracle Database, Spark, Hadoop and other data sources [1]
- However, current software will need to be able to interact with new types of databases, such as graph and vector databases, if they want efficient data retrieval and storage

[1] *Oracle Data Miner*. Oracle. (n.d.). https://www.oracle.com/big-data/technologies/dataminer/

Increase in Importance

- More and more data is being generated
  - Around 329 million TB of data is generated every day [1]
  - Enterprises only use around 1/3 of their available data [2]
- The barrier to entry for data mining software has decreased dramatically
  - Easier to generate and store data with more integrations
  - Low/ no-code software
- Being able to use data mining / OLAP software might become an essential skill

[1] Duarte, F. (2023, December 13). *Amount of Data Created Daily.* Exploding Topics. https://explodingtopics.com/blog/data-generated-per-day
[2] Leach, A. (2023, June 6). *Enabling AI-powered business intelligence across the Enterprise.* IBM. https://www.ibm.com/blog/enabling-ai-powered-business-intelligence-across-the-enterprise/

Current Problems

- The scale of big data
  - Data mining algorithms can become slow on very large data sets [1]
  - Not all data can be loaded into memory at one time
- Privacy concerns
  - More data means more information that can be leaked [2]
  - Customer data should be protected
  - For medical data, patients must opt in for their data to be used as training data [3]

[1] V. Kolici, F. Xhafa, L. Barolli and A. Lala, "Scalability, Memory Issues and Challenges in Mining Large Data Sets," 2014 International Conference on Intelligent Networking and Collaborative Systems, Salerno, Italy, 2014, pp. 268-273, doi: 10.1109/INCoS.2014.50.
[2] R. Josphineleela, S. Kaliappan, L. Natrayan and A. Garg, "Big Data Security through Privacy – Preserving Data Mining (PPDM): A Decentralization Approach," 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2023, pp. 718-721, doi: 10.1109/ICEARS56392.2023.10085646.
[3] Bennett, B., & Matta, N. M. (2023, November 9). *Beware privacy risks in training AI models with health.* Frost Brown Todd Attorneys. https://frostbrowntodd.com/beware-privacy-risks-in-training-ai-models-with-health-data-3/

Research Paper 1

- *Mining Conditional Functional Dependency Rules on Big Data* by Li et al.
- Conditional functional dependency discovery algorithms are used in data cleaning
- Three issues with current big data CFD discovery algorithms:
  - Current algorithms are meant for small datasets
  - Big data datasets can be dirty
  - Datasets can be larger than main memory
- This paper presents a new algorithm that uses a small representative set that is easily scalable and works linearly with data size
- Shows theoretical advances in data mining

M. Li, H. Wang and J. Li, "*Mining conditional functional dependency rules on big data*," in Big Data Mining and Analytics, vol. 3, no. 1, pp. 68-84, March 2020, doi: 10.26599/BDMA.2019.9020019.

Research Paper 2

- *A Hybrid Data Mining Method for Tunnel Engineering Based on Real-Time Monitoring Data From Tunnel Boring Machines* by Leng et al.
- Tunnel boring machines gather a lot of data such as stress, gas pressure, and current machine status. This data largely goes unused
- This paper discusses a framework for data mining and successfully implemented it
- Shows experimental application of data mining resources

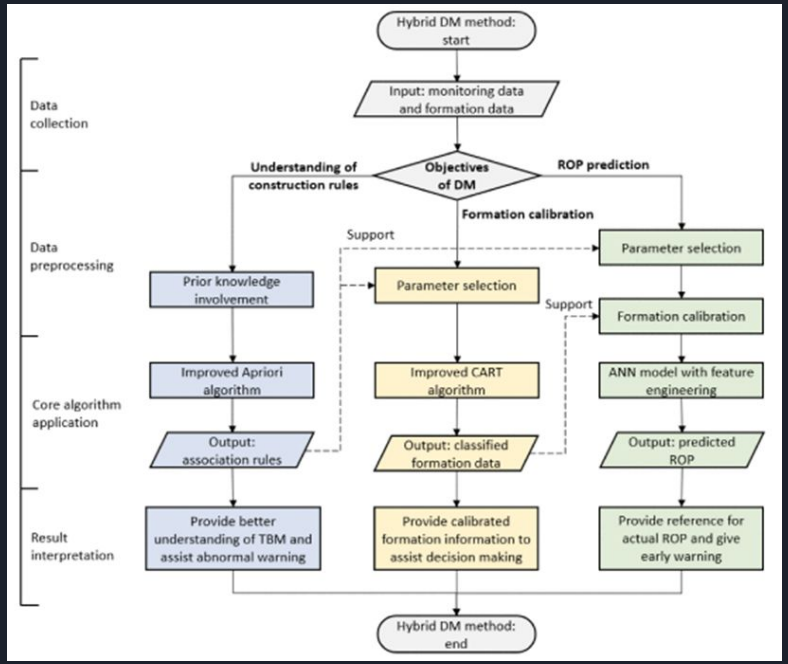


Figure 1. Framework of hybrid DM method

S. Leng, J. -R. Lin, Z. -Z. Hu and X. Shen, "A Hybrid Data Mining Method for Tunnel Engineering Based on Real-Time Monitoring Data From Tunnel Boring Machines," in IEEE Access, vol. 8, pp. 90430-90449, 2020, doi: 10.1109/ACCESS.2020.2994115

Questions and Answers