Data Mining Issues and Data Mining Tools

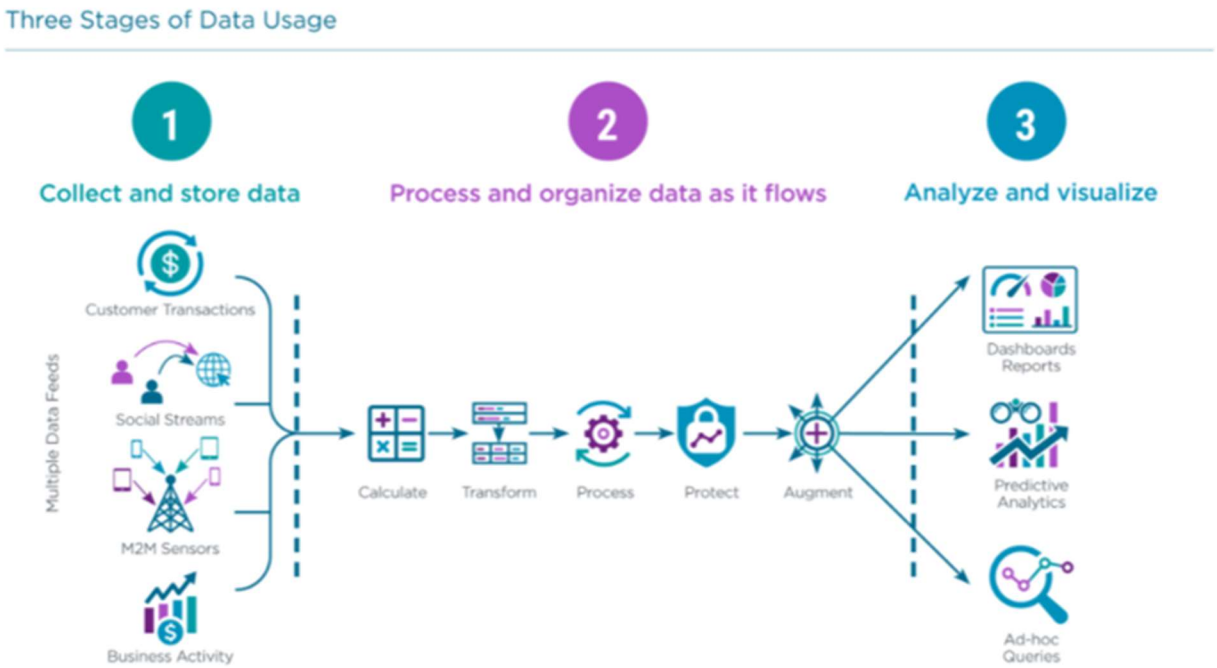
Omar Alomari, Osama Kfaween, Yousef Shaban

1932849 , 1937349 , 1932107

ABSTRACT

Data warehouse and data mining are two components that complement each other, in which a warehouse focuses on storing data in a structured database, while data mining focuses on extracting facts and useful information from that storage in order to apply these results to business needs.

From medium to huge, nearly all industries and organizations can benefit gracefully from this process!



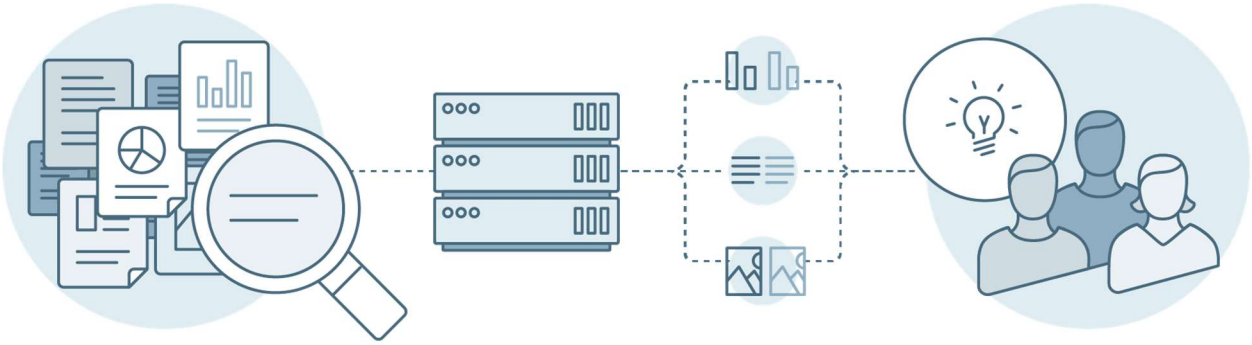
comptia.org/content/articles/how-is-data-mining-used-in-marketing

INTRODUCTION

Building upon the concept that **every successful leader must develop data-driven strategies** [1], it is essential to acquire reports and statistical analysis that is obtained through meaningful historical records.

Better yet, it is mentioned that 80% of organizations that apply business intelligence find the process of data mining important. [2]

So, what are a data warehouse and data mining? Why are they complementary? And how do they solve business needs?



springernature.com/gp/researchers/text-and-data-mining

Data Warehouse

medium.com/@geeksagar/what-is-data-warehouse-simple-definition-dde9fc98200e

What is a data warehouse?

According to Oracle, a data warehouse is a sort of management system that stores historical data, which is designed to support business intelligence (BI) and particularly analysis. That can be useful for data scientists and business analysts.[3]

Main characteristics of a data warehouse:

1. Non-volatile
2. Time-Variant
3. Subject Oriented
4. Integrated

Why is data warehousing so important?

A data warehouse is consistent, which means it offers reliable quality data, provides historical precise data, has a huge storage amount, and plays a key role in cost reduction.

Data Mining

ovhcloud.com/en/learn/what-is-data-mining/

What is data mining?

Data mining is a process, in which the techniques evolve around examining massive amounts of data, which the process could be conducted in parallel or distributed computing.

Main characteristics of data mining:

1. Manage Datasets
2. Predictions
3. Business Enhancement
4. Large Calculations

Why is data mining so important?

Data mining assists organizations in making better business decisions, it may minimize risks, and increase sales. It encourages not just smart decision-making but also detects trends and different types of fraud.

Data Mining Issues

- **Noisy and Incomplete Data**
  Data Mining is the way toward obtaining information from huge volumes of data. This present reality information is noisy, incomplete, and heterogeneous. Data in huge amounts regularly will be unreliable or inaccurate. These issues could be because of human mistakes, blunders or errors in the instruments that measure the data, so how will we deal with the data which is incomplete and noisy?

  By using binning sorted data is placed into bins or buckets. Bins can be created by equal-width (distance) or equal-depth (frequency) partitioning. On these bins, smoothing can be applied.
- **Data privacy and security**
- **Complex Data**
- **Improvement of mining algorithms**
- **Visual presentations**
- **Others such as Performance, data visualization, etc.** [4]

Data Mining Tools

- **RapidMiner**

  is a free open-source data science platform that features hundreds of algorithms for data preparation, machine learning, deep learning, text mining, and predictive.

- **Oracle data mining**

  is a component of oracle advanced analytics that enables data analysts to build and implement predictive models. it contains several data mining algorithms for tasks like classification, regression, anomaly detection, prediction, and more.

- **Weka**

Weka is an open-source machine learning software with a vast collection of algorithms for data mining. it was developed by the university of Waikato, in New Zealand, and it's written in JavaScript.[5]

Why use Weka for data mining and machine learning?

Because the Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.

Which algorithms are used in Weka? [6]

1- Logistic Regression
2- Naive Bayes
3- Decision Tree
4- k-Nearest Neighbors
5- Support Vector Machines

Bibliography

[1] https://www.herzing.edu/blog/what-data-warehousing-and-why-it-important

[2] https://www.iberdrola.com/innovation/data-mining-definition-examples-and-applications

[3] https://www.oracle.com/in/database/what-is-a-data-warehouse/

[4] https://www.jigsawacademy.com/blogs/data-science/data-mining-challenges

[5] https://www.tutorialspoint.com/weka/what_is_weka.html

[6] https://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/