

Literature Review

This section examines existing research on data mining, business intelligence, and analytics visualization, focusing on keywords:

- Data mining
- Business intelligence
- Power BI
- Tableau
- Key Performance Indicators (KPIs)
- Metrics
- Visualization tools
- Self-Service BI
- OLAP (Online Analytical Processing)
- ETL (Extract, Transform, Load) features

Scope:

The review is limited to thesis papers and journal articles from reliable websites and publications between 2020 and the present.

Objectives:

1. Analyse existing research on data warehousing, OLAP, data mining, and modern analytical tools (Tableau and Power BI).
2. Identify research gaps in well-documented findings.

The Impact of Data Warehousing in Business Intelligence

This book, authored by Martins (2020), examines the role of data warehousing in business intelligence applications. It outlines the essential steps for extracting, processing, and evaluating stored datasets to support querying and decision-making.

Key Concepts:

- Data warehousing: reservoirs for structured datasets from various sources (operational data, business information, health records, etc.)
- Characteristics: non-volatile, historic, and specific
- Data collection: extraction, transformation, and loading (ETL) into the data warehouse
- Data storage: meeting decision-makers' needs, representing information, and summarizing data
- Data analysis: querying and fetching information

Data Warehousing Process:

1. Data collection (ETL)
2. Data storage
3. Data analysis

Data Warehouse Schema and Dimensions:

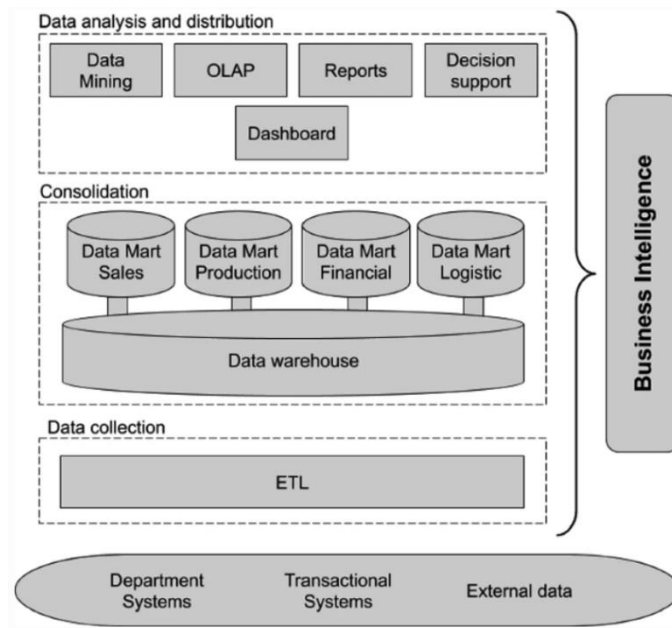


Figure 1: Data warehousing Processing

Methodology:

The study employs a thorough literature review (2011-2020) using four approaches and research questionnaires from seven global digital libraries.

Key Findings:

1. Real-time filtering processes are implemented at the ETL stage.
2. Structured databases: MongoDB, HBase, Cassandra, CouchDB, Neo4j, Hive, OrientDB, Infinispan, and Terrastore.
3. Unstructured databases: Hadoop, Apache Storm, GridGain, HPCC, and MapReduce.

Research Gaps:

The study recommends improving processing methods.

Relevance to Our Work:

This research relates to our project, as we encountered similar challenges during data processing using Microsoft Management Studios, a relational database provided by our institute.

Challenges Encountered:

- 1. Importing metadata
- 2. Cleaning data
- 3. slow query performance
- 4. Missing values
- 5. Integrating data from different sources

By acknowledging these challenges, our research aims to develop more efficient data processing methods.

Challenges in Data Processing

Data processing poses significant challenges, particularly when working with relational databases. These challenges include:

Relational Database Issues

Issue	Description
Duplication	Data duplication during ETL processes
Performance	Slow query performance
Reliability	Data inconsistency and errors

Data Integration Challenges

- Managing heterogeneous content and diverse quality problems using Microsoft SQL Server Integrated Services (SSIS)
- Integrating data from multiple sources
- Ensuring data consistency and quality

Data Preparation

Data manoeuvring prepares raw data for analysis by:

- Handling heterogeneous dataset sources
- Integrating and developing data
- Continuously monitoring data quality

Time-Consuming Process

Data integration and preparation are time-consuming due to:

- Merging information from different sources
- Requiring proper analysis and testing

As highlighted by Azeroual (2020), effective data processing requires careful attention to these challenges.

Algorithm/Approach	Methodology	Challenges Addressed	Ref.	Tool/Data Structure	Shape of Data	Evidence
Comparative review of DWH ETL tools	comparative review based on near real-time ETL approaches like: Change Data Capture (CDC), Trickle and Flip approach, real time data cache (RTDC) approach	market value and relevance of ETL tools in data science industry, and growing need of real time data analysis from structured and unstructured data sources	[30]	ETL tools (Informatica, Datastage, Ab Initio, Oracle Data Integrator, SSIS)	Relational, PDF, XML etc	Industry insights for the relevance of tools
Survey of design approaches for DWH from social media	literature review	exploitation of data from the web	[4]	Multidimensional model for DWH (Conceptual, Logical)	heterogeneous social media data	No evidence provided in this literature review
Identification of challenges of ETL implementation for near real-time environment	literature review to find challenges and solution approaches for ETL implementation	high availability, low latency and horizontal scalability features for functionality	[15], [25]	Change data capture (CDC), CDC log-based, real-time data cache (RTDC), Trickle and Flip	Relational	No evidence provided in this literature review
Optimized foreign-key join algorithm for OLAP workloads	foreign-key join algorithm instead of general-purpose hash joins	to enable surrogate key index to be efficient for foreign key joins in DWH workloads for both hardware accelerators and CPU	[59]	array-store oriented foreign key, Xeon Phi and NVIDIA K80 GPU platforms	Relational	proposed approach is evaluated using synthetic and real-life datasets
Performance Analysis of Not Only SQL Semi-Stream Join Using MongoDB	join module of stream with disk based data	efficient stream processing for NoSQL data for real-time DWH	[24]	MongoDB	NoSQL, Relational	Experimentally evaluated using synthetic and real-life data sets
Framework for big DWH dealing in both real-time and offline modes	first component: real-time data ingestion of both streaming and offline data generated by communication service providers (CSP) and coding, second component: big data ETL module	to effectively organize raw data, and implement complex and more intelligent use-cases that help in improving core networks and other areas of CSP	[29]	NiFi, Kafka, Spark Streaming, Hadoop	Relational	No experimental evidence provided
Architecture of Striim's streaming ETL engine, a distributed streaming ETL and intelligence platform	to handle the demands of modern data pipelines, transformation engine has been designed. Open Processor component of this engine enables users to develop join functionality or run machine learning models on the input data streams	to run low-latency transformation logic on input data streams using modern approaches of query optimization and execution. To enable declarative data filtering and updation on streaming real-time data	[35]	Striim transformation engine, SQL, CDC, Kafka, open source streaming ETL engine KSQL	Relational	Experimentally evaluated based on synthetic and real-life datasets
Overview of the existing data quality approaches in the ETL process	comparative review of some commercial ETL tools considering highlighted data quality characteristics (three quality dimensions considered for comparison: performance, reliability, deduplication). ETL tools: Talend Data Quality (TDQ), Talend Data Integration (TDI), Pentaho Data Integration (PDI), Informatica Data Integration (IDI) and Microsoft SQL Server Integration Services (SSIS)	management of data from internal and external sources: data with heterogeneous content and diverse quality problems	[39]	ETL tools: TDQ, TDI, PDI, IDI, SSIS	Relational	Comparative evaluation based on TDQ and TDI tools

Table 1: screenshot of Survey results on real-time processes and challenges faced source [Mehmood, 2020]

Relational vs Non-Relational Databases

When handling large datasets, choosing the right database management system is crucial. This section explores the differences between relational and non-relational databases.

Relational Databases

Characteristics:

- Extensively used database storage systems
- Include data storage and retrieval procedures
- Information stored in tables with predefined rows and columns
- Handle smaller amounts of datasets

- Four primary operations: INSERT, DELETE, UPDATE, and SELECT using SQL
- Handle complexities using schemas and structured languages

Example: Microsoft Management Studios (used in our work)

Non-Relational Databases (NoSQL)

Characteristics:

- Handle large-scale data for web 2.0 and social networking applications
- Store data generated from social media platforms and websites
- Schema values are not fixed
- Data stored in document form
- Use programming languages like C++ and JavaScript

Key Differences

Relational Databases	Non-Relational (NoSQL)	Databases
Data Structure	Tables with predefined rows and columns	Flexible schema, document-based
Scalability	Limited	High scalability
Data Size	Smaller datasets	Large-scale datasets
Query Language	SQL	Varied (e.g., C++, JavaScript)

Our Work

We utilize Microsoft Management Studios, a relational database, to manage metadata from multiple tables and sheets from different sources.

Next Steps

The following sections will discuss:

1. Data cleaning and extraction processes for business intelligence platforms
2. Visualization techniques using BI tools

Reference:

Samydurai (2022)

Relational vs Non-Relational Databases

When handling large datasets, choosing the right database management system is crucial. This section explores the differences between relational and non-relational databases.

Relational Databases

Characteristics:

- Extensively used database storage systems
- Include data storage and retrieval procedures
- Information stored in tables with predefined rows and columns
- Handle smaller amounts of datasets
- Four primary operations: INSERT, DELETE, UPDATE, and SELECT using SQL
- Handle complexities using schemas and structured languages

Example: Microsoft Management Studios (used in our work)

Non-Relational Databases (NoSQL)

Characteristics:

- Handle large-scale data for web 2.0 and social networking applications
- Store data generated from social media platforms and websites
- Schema values are not fixed
- Data stored in document form
- Use programming languages like C++ and JavaScript

Key Differences

Relational Databases	Non-Relational Databases (NoSQL)	
Data Structure	Tables with predefined rows and columns	Flexible schema, document-based
Scalability	Limited	High scalability
Data Size	Smaller datasets	Large-scale datasets
Query Language	SQL	Varied (e.g., C++, JavaScript)

Our Work

We utilize Microsoft Management Studios, a relational database, to manage metadata from multiple tables and sheets from different sources.

Next Steps

The following sections will discuss:

1. Data cleaning and extraction processes for business intelligence platforms
2. Visualization techniques using BI tools

Reference:

Samydurai (2022)

Data Mining Techniques and ETL Processing

Various techniques are employed in data mining processes. Our research utilizes some of these techniques, providing detailed explanations.

Research Methodology

1. Database Management Studios as data warehouse
2. SQL queries for data processing
3. OLAP queries for trend analysis
4. Data import into Tableau and Power BI for performance analysis

ETL Processing using OLAP Techniques

A study in the banking industry demonstrates data warehousing and ETL processing (Aziz, 2021).

Key Steps

1. Data gathering and transportation to database location
2. Data entry into facts table
3. Data transfer from facts table to OLAP cube

Tools Used

1. Microsoft Visual Studio 2019 (relational database management system)
2. Kaggle datasets
3. Power BI (OLAP analysis)

Findings

The study showcases the effectiveness of OLAP-based approaches in enhancing bank intelligence.

Business Intelligence

Business Intelligence encompasses various data processing and management techniques, including:

1. Data mining
2. Data warehousing
3. ETL processing
4. OLAP analysis
5. Data visualization (Tableau, Power BI)

Reference:

Aziz (2021)

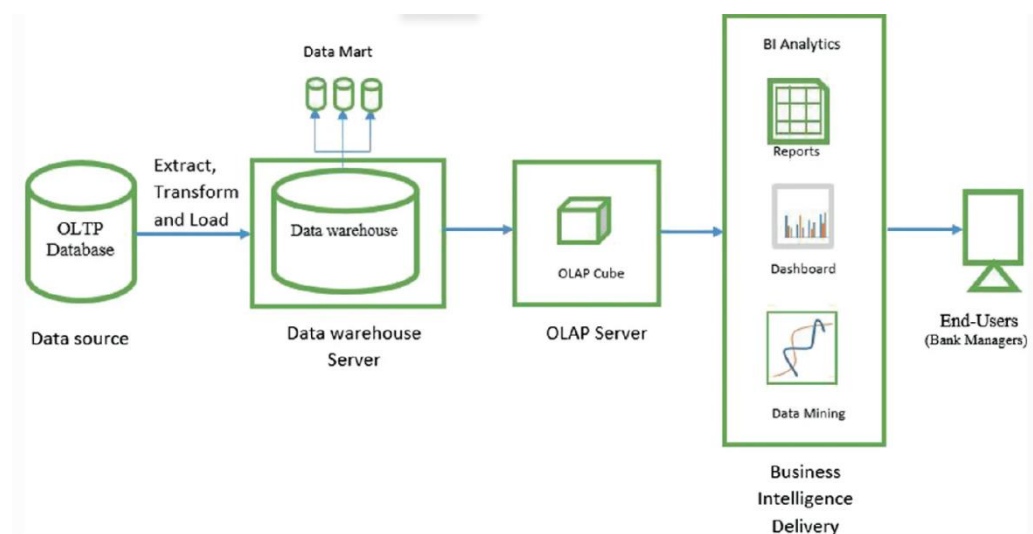


Figure 2: Analysing bank data using Data Mining Approach, Source [Aziz, 2021]

The Role of Self-Service Visualization Tools in Data Analysis

Real-time data processing has become increasingly challenging for companies, including manufacturing industries (Sousa et al., 2021). To address this, business intelligence techniques and artificial intelligence are being applied.

Research Objectives

1. Comparative analysis of existing applications for real-time data processing and visualization.
2. Evaluation of tools' ability to facilitate data processing and visualization.

Methodology

1. Comprehensive evaluation of technologies and tools.
2. Secondary research: books, journals, reports, papers, dissertations, and internet research.
3. Analysis of 200 publications.

Findings

1. Power BI emerged as a top tool with a customizable interface.
2. Power BI components: Power Query, Power Pivot, Power Map, Power View, Power BI Q&A, Power BI Desktop, Website, and Power BI Mobile Apps.

Importance of Business Intelligence Tools

1. Intuitive information analysis for real-time data.
2. Visualization principles for clear information presentation

References

- Sousa et al. (2021)
- Zhang et al. (2021)
- Bigliardi (2020)

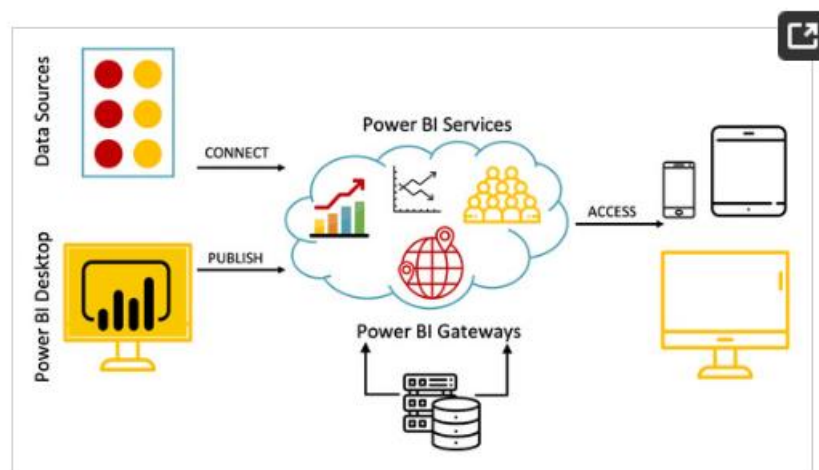


Figure 3: Screenshot showing Power BI basic architecture, Source [(Sousa, 2021)]

Tableau and Power BI: Features, Benefits, and Drawbacks

Tableau

Distinctions:

- Best Data Visualization (DM Review)
- Best Data Analysis (PC Magazine, 2005)
- Best Intelligence Solution (Software Information Industry, 2008)

Components:

- Tableau Desktop
- Tableau Server
- Tableau Online

Benefits:

- Simplicity of use
- Great performance
- Large number of connections
- Impressive visualization capabilities
- Simple access to various data sources
- Quick responsive dashboards
- No programming knowledge required (Batt, 2020)

Drawbacks:

- High cost
- Limited BI features
- Poor after-sales support
- Expensive employee training

Power BI

Drawbacks:

- Difficulty handling complex table relationships
- Poor configuration visuals
- Limited data handling capacity (10 GB free, 100 GB premium)

- Steep learning curve due to multiple components
- Limited flexibility in DAX language (Lamba, 2022)

Benefits:

- Data modification capabilities
- Formula-based data creation (DAX)

Data Visualization

Importance:

- Understanding data through visualization
- Making informed decisions

Steps:

1. Understanding data
2. Obtaining data
3. Data interpretation
4. Data cleaning
5. Data extraction
6. Refining data
7. Interaction
8. Presentation (Krishnan, 2022)

Using Power BI

Steps:

1. Select File
2. Choose file type (Excel, Text/CSV, XML, JSON, PDF)
3. Connect to data source
4. Specify file location

Tableau's Advantages

- Handles large, dynamic datasets
- Machine learning techniques
- Supports C, C++, Java, Python
- Ideal for public presentations (Majnarić, 2021)

References:

- Lamba (2022)
- Batt (2020)
- Mahalle (2021)
- Krishnan (2022)
- Majnarić (2021)

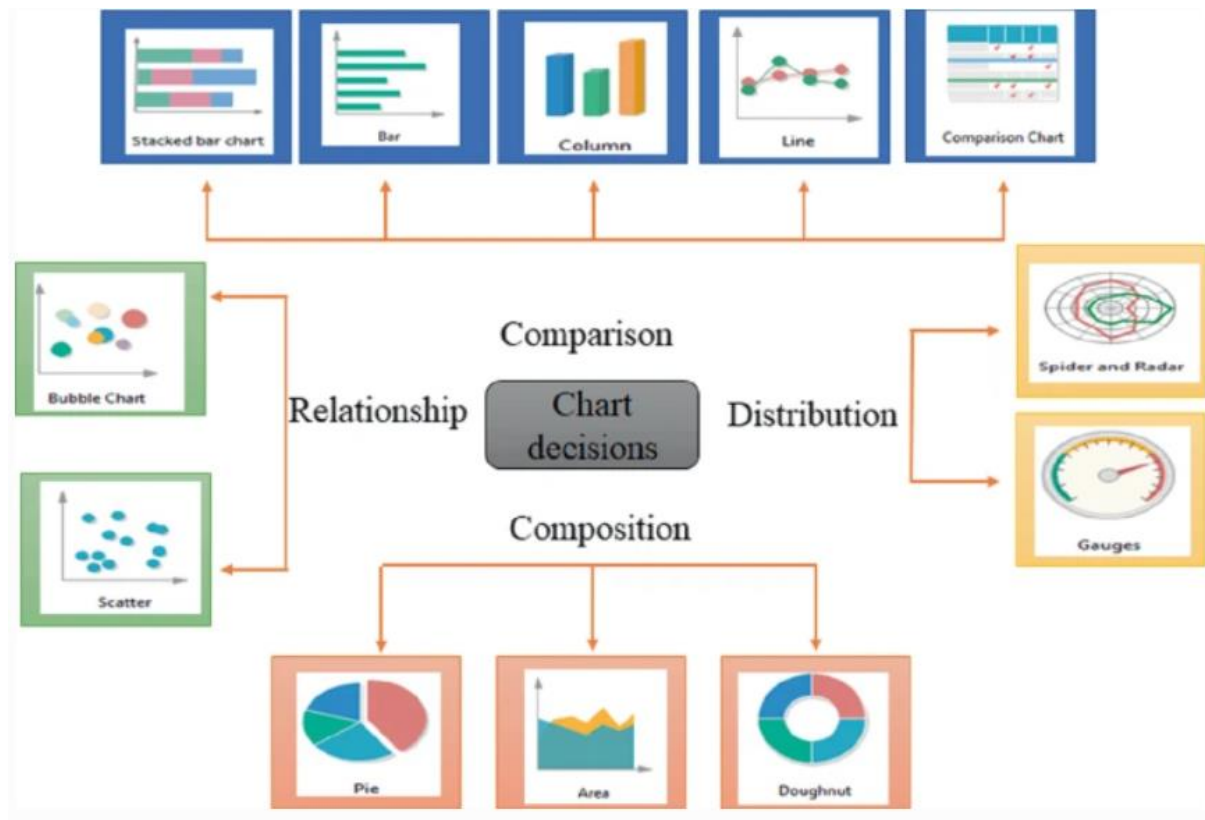


Figure 4:Chart classification and appropriate chart for each type of application, Source [Mahalle, 2022].

Data Classification and Collection: A Comparative Analysis of Business Intelligence Tools

This study evaluates 15 open-source business intelligence tools, including Tableau and Power BI, based on user reviews and ratings from 12 review websites (Srivastava, 2021).

Methodology

1. Comparison of 15 open-source business intelligence tools
2. Analysis of usability, efficiency, and ease of access
3. Standard deviation formula used to calculate error deviations

4. Color-coded representation of standard deviation values

Results

The study successfully tested the 15 tools and concluded that:

1. Kyubit BI is the top-rated business intelligence tool
2. Kyubit BI excels in efficiency and usability for business organizations

Key Findings

1. Tableau and Power BI ranked among the top tools
2. User reviews highlighted strengths and weaknesses of each tool
3. Standard deviation analysis revealed varying levels of error deviations

Implications

This study provides valuable insights for business owners to:

1. Choose the most suitable business intelligence tool
2. Evaluate tool efficiency and usability
3. Consider user reviews and ratings in decision-making

Reference:

Srivastava (2021)

BI Tool	S.D Value	Level
Zoho analytics	0.862	III
Sisense	0.581	II
Microsoft Power BI	0.822	III
Board	0.661	II
Tableau	0.4370	I
Looker	0.555	II
Datapine	1.167	IV
Microstrategy Analytics	0.996	III
Klipfolio BI	0.464	I
Birst	0.77	II
Clicdata BI	3.14	IV
kyubit BI	0.1	I
Dundas BI	1.058	IV
SAP Business Objects BI	0.996	III
Cluvio BI	0.489	I

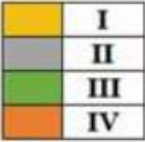


Figure 5: Analysis based on SD Source: [Srivastava, 2022]

Business Process Management (BPM) cycle. This research study focuses on dashboard design issues.

Research Objective

This study examines the importance of constructing dashboards tailored to specific business processes, considering:

1. Key Performance Indicators (KPIs)
2. Potential stakeholders as Business Intelligence dashboard users

Importance of Analytical Dashboards

To maintain control over organizational business processes, analytical dashboards are vital in the BPM life cycle.

Key Considerations

1. Dashboard design must align with business process characteristics
2. Effective monitoring and control over business process status
3. Stakeholder engagement through user-friendly dashboard interfaces

Reference

Sousa (2019b)

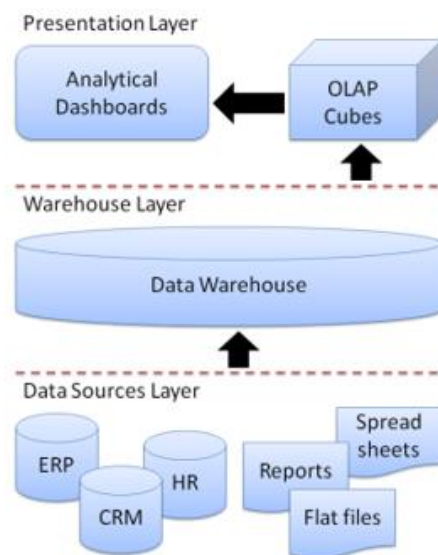


Figure 6: Simplified View of the dashboard's architecture Source [Orlovskyi, 2020]

Table 2: Overview of Existing Summary Papers

Article	Year	Research Problem	Methodology	Significant Results	Limitations
Data Visualisation using Tableau	2020	The goal of this thesis was to study different visualization methods, understand the various tools, and compare Tableau and Excel tools.	This research work mainly consists of two parts: theoretical and practical. The theoretical part is based on literature review that provides understanding about numerous data visualization methods, and provides understanding about widely used top applications.	Using the historical data the researcher found out that that calculations in excel which is an extension of power BI analysis show that applications such as Tableau provided a good result in terms of creation of a dashboard.	At the end of the study, main differences between Excel and Tableau are listed. Microsoft Excel has less distinct benefits as compared to Tableau.
Software Tools for Conducting Real-Time Information Processing and Visualization in Industry: An Up-to-Date Review	2021	The review presented in this document aims at providing an up-to-date review of the various tools available to perform visualisation tasks.	A critical review of the available tools and technologies for real time information processing and their application in manufacturing is carried out	The software tools Microsoft Power BI and the Tableau Suite help transform data into appealing and intuitive information displayed on dashboards.	As for future work, it is expected to identify and showcase bleeding-edge technology in Complex event processing, as well as it may impact manufacturing and daily lives.
Seaborn: statistical data visualization	2020	Functions in the seaborn library expose a declarative, dataset-oriented API that makes it easy to translate questions about data into graphics that can answer them.	The author uses complete graphics from a single function call with minimal arguments, seaborn facilitates rapid prototyping and exploratory data analysis. Hence offering extensive customization options.	As a result graphs are created through the seaborn interface using more advanced applications like defining composite figures with multiple arbitrary prototypes will require importing data and using functions	It does not implement the formal Graphics and cannot be used to produce quality visualizations
Business-intelligence framework for visualization and its associate text narration	2020	Chart visualization as delivering information point alone is not enough because reading charts only is not enough but fully understanding the messages and derive different conclusion is the problem	The author implements a framework called JavaScript to make linking connections between reading text areas and their associated charts and vice versa.	Visualization wrapper is successfully developed with different libraries like java script and python which can be used together under the wrapper without extracting form different libraries	Writing long text to guide users is never a good practice to the deliver information is also not a good dashboard presentation practice since it becomes the same as reading a book to match figures and captions match figures by their captions.
CODAS: Integrating Business Analytics and Report Authoring	2022	Elaborated information making it difficult to convey information on dashboard caused by data changes leading to a need to facilitate authoring of interactive reports in business analytics.	Creating a web-based prototype called CODAC for authoring data-driven reports for business analysis allowing business analysts to effectively communicate information using stories with a graphical user interface	The overall feedback about CODAS was positive. participant one and two (P1,P2) liked that it was easy to add story pieces (both chart and text) in the system and quickly arrange the sections to form a narrative	More intelligent features for authoring data-driven reports need to be developed. They include features like automatic content generation for recurring reports, recommendation of visualization charts and more.
Collaboration of Business Intelligence and cloud Computing and selecting the best Cloud Business Intelligence Solution	2021	The benefits and obstacles experienced by companies when using cloud computing and business intelligence arising from the collaboration of the two software's in the third part	Four service providers of both technologies were selected and the criteria were determined according to the needs of the company. Analytical hierarchy process, which is one of the multi-criteria decision-making method, was involved in the selection of cloud business intelligence solution	This study defends the concept of using cloud business intelligence for companies from small to big sized in order to become competitive. The study shows that Power BI provides the best software as it best fits the company needs and objectives	The result will be affected when the priority for these criteria changes in the training and support of customer service, when the criteria for measurement increases, the results will change to Tableau
Comparative Quantitative Analysis of Leading Business Intelligence Software Platforms	2022	This paper is a quantitative comparative analysis approach on 3 BI leading software platforms (Power BI, QlikView and Tableau) to find out what tool has more capabilities than the other.	Core methods and capabilities of BI are discussed along with a generalised stack of BI software architecture. Considered amongst the top BI platforms were Power BI, QlikView, and Tableau which were contrasted in terms of their B capabilities.	In conclusion the author says that QlikView which is one of the BI tools, has the best rating yet Power BI and Tableau are more preferred by users	The suggested methods should be generalised in the future so that it may be used to assess diverse software products from other industries. Other similar measures might be taken into consideration, the proposed approach could be improved, and the corresponding software component created
Visual Analytics Towards Prediction of Employee Erosion Through Data Science Tools	2020	This paper seeks to evaluate the effectiveness of Visualisation in predictive analytics using rich R visuals and strong visualisation tools	23,326 rows of corporate data, including demographic and employment characteristics of employee, were analysed for data visualisation effectiveness using a prediction algorithm. The solution overview for graphical report visualisation and employee turnover prediction.	Power BI creates reports that can be hosted online, present data graphically, and provide users with an interactive experience. Tableau can produce significant insights and carry out clustering to produce pieces of deducible facts	Visualization poses difficulties for a knovist using R and its capabilities, but non experts can effectively transform and analyse their data using programmes like Power BI and Tableau

Performance Management in Business Process Management Cycle

Monitoring and regulating corporate business processes is crucial in the Business Process Management (BPM) cycle. This research study focuses on dashboard design issues to optimize performance management.

Research Objective

This study examines the importance of constructing dashboards tailored to specific business processes, considering:

1. Key Performance Indicators (KPIs)
2. Potential stakeholders as Business Intelligence dashboard users
3. Real-time data analytics for informed decision-making

Importance of Analytical Dashboards

To maintain control over organizational business processes, analytical dashboards are vital in the BPM life cycle, enabling:

1. Data-driven decision-making
2. Process optimization
3. Performance measurement and evaluation

Key Considerations

1. Dashboard design must align with business process characteristics
2. Effective monitoring and control over business process status
3. Stakeholder engagement through user-friendly dashboard interfaces
4. Integration with existing business systems and tools

Benefits of Effective Performance Management

1. Improved operational efficiency
2. Enhanced business agility
3. Better decision-making
4. Increased transparency and accountability

Reference

Sousa (2019b)

