Design and Implementation of Big Data Visualization for Student Housing Analysis

Khairul Anwar Sedek*, Nurfarah Fatini Mohd Maulus, Mohd Nizam Osman, and Nor Arzami Othman

College of Computing, Informatics and Mathematics,
Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Malaysia

Received Date: 20 August 2023
Acceptance Date: 15 September 2023
Published Date: 1 November 2023

Abstract. The Student Affairs Division of UiTM Perlis is responsible for non-academic matters. However, the current e-Kolej system lacks comprehensive residential data, making it impossible for staff and students to analyse and use the data effectively. The purpose of this study is to address this issue by developing a residential dashboard that provides the platform for college statistics. By providing staff members with access to and analysis of relevant data, the dashboard will support informed decision-making and efficient management of student housing. The phases of the project include a requirements analysis, design, implementation, and usability assessment. This method employs Apache Hive for large data warehouses, HiveQL for data analytics, and Microsoft Power BI for data representation. Thirty participants provided valuable feedback during the test, validating the dashboard's usability and effectiveness in providing users with the required data. According to the evaluation data, the average overall score is 4.71. Future projects include adding more information from other perspectives to the dashboard and integrating the dashboard with the college database system at UiTM Perlis Branch remain.
Keywords: Big data, data visualization, data analytic, information system, knowledge management

1 Introduction

As technology advances and the use of digital technologies widen, the amount of digital data continuously increasing that generating tremendous amounts of data through the increased use of mobile devices. three characteristics of big data are known as 3Vs: volume, variety and velocity (Laney, 2001). Volume refers to the amount of data, variety refers to the number of types of data and velocity refers to the speed of data processing.

Data visualisation is the visual or graphical representation of data (Bikakis, 2018). It allows decision makers to view analytics presented visually in order to get insights. Big data and data visualisation are related because data visualisation is essential for storytelling by organising data into an easier-to-understand format and showing trends and outliers (Gandhi & Pruthi, 2020)

University dorms are of significant importance following education (Daliri Dizaj & Hatami Khanghahi, 2022). However, not all university students are promised college housing. Students must register for college online and await acceptance or denial of their application. College facilities are available to qualified individuals depending upon college availability.

The university must have an appropriate allocation of restricted college spaces for students utilising an effective college information system that provides information on college housing placement. The university's present system does not provide any visualisation of college information. This project's objective is to build and develop a dashboard for UiTM Perlis Branch students that displays residential information utilising big data visualisation. This dashboard will enable the detection of trends and give staff and students with crucial information about student housing. The data will contain the total population, gender distribution, the ratio of residents to non-residents, a list of colleges, and residential capacity use. In addition, residents can be categorised by course, semester, programme, hometown, extracurricular activities, and academic performance. This project has the ability to improve the university's college system by offering comprehensive insights and increasing overall efficiency.

2 Literature Review

Today there are millions of data sources that produce data very quickly. These data sources are accessible on an international market. Social media platforms and networks like Facebook produce more than 500 terabytes of data every day, making them among the largest data providers (Pence, 2014). In addition, data is available in various forms including structured, semi-structured and unstructured data including images, movies, text, and other types.

One of the benefits of data visualization in higher education student affairs analytics is that it can help institutions to organize, analyze and report data in a more effective and efficient way. Data visualization can enable users to explore and interact with data, discover patterns and trends, answer questions quickly, make data-informed decisions, and share their findings with others123. For example, data dashboards can provide a comprehensive and customizable view of key indicators and metrics that can inform student success initiatives1. Data visualization can also help to engage and empower students by providing them with feedback, guidance and recommendations based on their own data.

Big data is a term that describes a collection of enormous and complex data sets and volumes, as well as data management tools, social media analytics, and real-time data (Ishwarappa & Anuradha, 2015). Big data is about data volumes and large data sets, measured in terabytes or petabytes. Big data is a huge amount of data that cannot be stored, processed or analysed with conventional tools (Riahi & Riahi, 2018). Special advanced tools are required to manage all data. Big data technology is designed to process, analyse, and store these large amounts of data and is used in many industries such as agriculture, medicine, gaming, education, business, etc.

The properties of big data identified as 3Vs (Volume, Velocity, and Variety) as early as 2001 (Laney, 2001). IDC introduced the fourth V (Value) in 2011 (Gantz & Reinsel, 2011) and the fifth V (Veracity) in 2017 (Jain, 2017), making big data known as the 5Vs (Volume, Velocity, Variety, and Value, and Veracity).

Big data is a vast collection of complex data sets used in various industries, including agriculture, medicine, gaming, education, and business. It is classified into the 5Vs and consists of three layers: Data Access, Data Storage, and Data Processing. Big data solutions consist of four stages: data ingestion, storage, analytics, and application integration.

Big data distributed architecture for education consists of three layers: (1) Data Access Layer, Data Storage Layer, and Data Processing Layer (Alblawi, 2018). The Data access layer collects educational data sources such as student logs, student records, and historical data. The second layer is big data storage using HBASE and Hadoop Distributed File System (HDFS). The third layer provides the sentiment and predictive analytic. Common storage options include HDFS and Hive, while the processing layer uses Hadoop and Apache Spark. Data analytic and visualization stages are used to clean, sort, analyse, and visualize data for smart applications, such as management decisions and academic dashboards.

The solution and approaches for big data consists of four stages: data ingestion, data storage, data analytics, and application or operational integration (Bhattarai et al., 2019). There are many choices of solutions and approaches that can be used in each stage. To choose the approaches, it depends on data type, speed, reliability, delivery, file size, scalability, and durability(Amare & Šimonová, 2021). For example, the RESTful API technique is reliable for data acquisition from integration of web data (Cui et al., 2020). Streaming tools such as Kafka and Flume are used for data that has high velocity from web server data sources (Sharma et al., 2021). Sqoop can be used for data synchronization between Relational Database Management System (RDBMS) and HDFS (Li & Zhang, 2020).

In the storage stage, the common data storage is HDFS and Hive (Li & Zhang, 2020). Hadoop is efficient and swift storage and processing of big data that uses a master-slave architecture to store and access data (Zhai et al., 2021). Apache Hive is an open-source relational database system for analytic big-data applications that was developed on top of Hadoop for large-scale, reliable processing, and it already supports SQL compatibility using HiveQL (Camacho-Rodríguez et al., 2019). HiveQL manages unstructured data in HDFS using strong SQL skills for data analysts (Małysiak-Mrozek et al., 2022).

The data processing layer uses Hadoop framework for big data analytic techniques based on two main components: Map Reduce and HDFS (Sharma et al., 2021). Apache Spark is engine for large-scale data analysis over a diversity of workloads with lightning fast processing (Deshai et al., 2020). It is a flexible procedure and reduces the drawbacks of Map Reduce. Other data visualization technology has been developed by computer graphic community such as D3, VegaLite, VizQL, Tableau, and Microsoft Power BI (Qin et al., 2020).

Additionally, there is a data analytic and visualization stage for cleaning, sorting, analysing, and visualize data in order to develop a smart application (Mjhool et al., 2019). Apache Zeppelin provides integrated environment for complete data ingestion, data discovery, data analytics and data visualization quickly and efficiently (Cheng et al., 2018). Two examples of applications based on big data visualizations are management decisions dashboard (Lucio et al., 2018) and academic dashboard (Destiandi & Hermawan, 2018).

Aprillia et al. (2021) propose a project on developing a dashboard as a monitoring system for the distribution of government aid at Balai Besar Perikanan Budidaya Air Payau (BBPBAP) Jepara, Indonesia. The project has produced two dashboards, The Ujitek Dashboard, and The Dukungan Teknis Dashboard. Targeted user of this dashboard is e-Coordinator and Sub-Coordinator of the Technical Support Group.

Lucio et al. (2018) has developed a project to construct a dashboard as a tool for visualising management decisions in order to substantiate valuable university management recommendations. The dashboards display applicant demand for educational specialties, programme rank, pricing policy, employer demand, and the number of graduates enrolled for Independent External Evaluation using charts, including bar graphs, line graphs, and area graphs. All the exhibited information will make it easier for users to rationalize the university's useful ideas. This project's drawback is the incomprehensibility of the exhibited dashboard charts and data, particularly by the public users.

Destiandi and Hermawan (2018) propose an intelligent business approach for academic dashboards. The constructed academic dashboard aid in improving the quality of education through better decision making. The dashboard provides academic-related data, including the total number of study programmes, the total number of students enrolled in tertiary institutions, the total number of graduates, and the total number of permanent lecturers, as well as a gauge graph displaying the Key Performance Indicator (KPI) of student graduates and the overall lecturer ratio. The dashboard interface is designed with all the information are provided on a single page, making it easy for the user to analyse the information quickly. The dashboard also featured easy-to-understand charts and graphs and provided information that does not require specialised knowledge.

Nizam et al. (2023) propose residential map visualization to provide house rental for student. The project uses integration of Google Map visualization and Whatsapp API using PHP MySQL development platform. The system improves rental management between admin, landlord, and tenant, making the process more effective and smoother.

Aprillia et al. (2021), Lucio et al. (2018), Destiandi and Hermawan (2018), and Nizam et al. (2023) use visualisation to improve decision making for stakeholders in their organization. However, they do not use visualisation for the same purposes. *Table 1* depics the summary of related work. Aprillia et al. (2021) develop visualisation dashboards for monitoring the Indonesian government's aid distribution using Microsoft Power BI. Lucio et al. (2018) develop dashboard for visualising management decisions to justify valuable proposals in university management. Destiandi and Hermawan (2018) also develop academic dashboards based on an intelligent business approach using PHP for ETL processes and application development and a database management system for data warehouses. Nizam et al. (2023) develop web-based academic dashboards using the PHP MySQL platform to provide house rental search for non-residential university students. All four works developed web-based academic dashboards. None of these projects integrate big data visualisation platforms to develop their academic dashboards. This work develops an academic dashboard to provide analytical information on the residential college of UiTM Branch using a big data visualisation platform based on the Apache Hadoop big data platform. Hence, this work proposes a different approach to developing the dashboard using a big data platform. It can prepare the university for future demands of big data analytics when the data grows to support all data requirements from all 35 campuses of UiTM and 187,994 students in 2023 enrollment (Mohd Janor, 2023). Therefore, a big data platform will provide universities with a scalable and reliable data analytic solution.

Table 1: Related Work

Title	Issues	Method	Result
Dashboard Design as Monitoring System Distribution of Government Assistance at Balai Besar Perikanan Budidaya Air Payau (BBPBAP) Jepara, Central Java (Aprillia et al., 2021)	Ineffective presentation of data causes difficulties for interested parties to understand and draw conclusions about problem solutions	Using Microsoft Power BI to develop both dashboard	Dashboard, Monitoring, Balai Besar Perikanan Budidaya Air Payau (BBPBAP) Jepara
Using MS Power BI Tools in the University Management System to Deepen the Value Proposition (Shaulska et al., 2021)	The traditional model of forming value propositions at universities is mainly focused on internal capabilities; market demand trends are more often derived factors, leading to inefficient decisions	Using Microsoft Power BI to develop both dashboards.	university management, university, value proposition, performance factors, behavioral factors
Business Intelligent Method For Academic Dashboard (Destiandi & Hermawan, 2018)	Management of tertiary institutions in Indonesia requires appropriate and fast academic reports that can inform strategic decisions	Use PHP to perform ETL process.	Business Intelligent Academic Dasboard KPI Business Intelligent Life Cycle
Web-Based Student Residential Searching Integrated with Google Maps and WhatsApp API (Osman et al., 2023)	Residential map visualisation using PHP Web applications, Google Maps, and WhatsApp API integration	Provide students with the tools they need to overcome the problem of finding a rental room that suits their needs.	The solution makes the rental management process between admin, landlord, and renter more efficient and streamlined through residential map visualisation

3 METHODOLOGY

Figure 1 depicts the methods used for this study. There are four phases: requirement analysis, design, development, and evaluation.

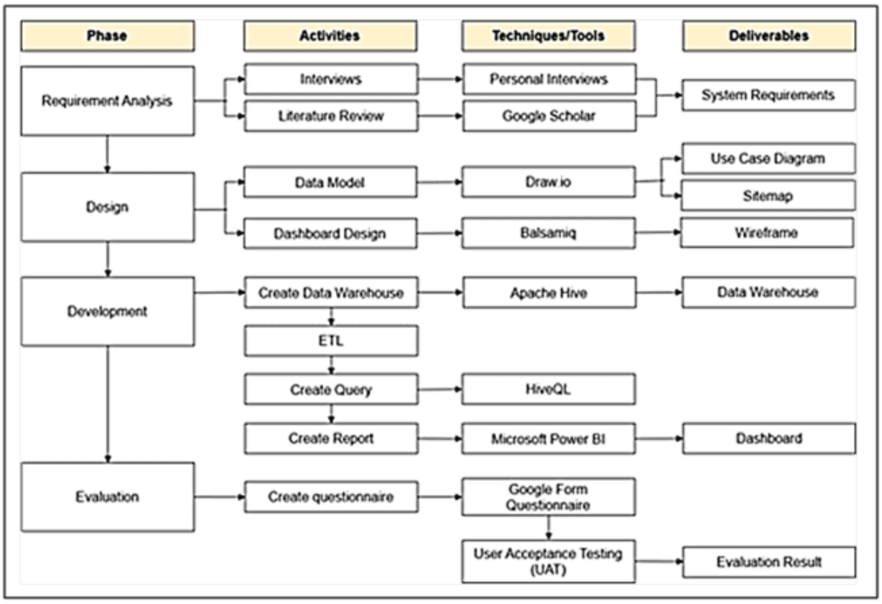

Figure 1: Research Methodology

3.1 Requirement Analysis

The functional requirements (FR) specify the functions, features, and capabilities of dashboard implementation to meet the project's objectives. This article identifies seven FR, as shown in Table 2.

Table 2: Functional Requirements

FR No.	Function	Function Description
1.	Display Basic Residential Statistic	The dashboard can display basic statistics of the residential like total population in UiTM Arau, Perlis, total population by resident type, resident or non-resident, and total population by gender.
2.	Display Colleges List	The dashboard able to show all colleges list available in UiTM Arau, Perlis and number of residents resides in each college.
3.	Display Residential Capacity	The dashboard able to display the overall residential capacity and residential capacity for each college in university.

4.	Visualize Residents by Few Categories	The dashboard able to visualize number of residents categorized in gender, college, program, course, semester, hometown, household income, pointer, and curricular merits.
5.	Filtering Residential Statistics	The dashboard able to filter the residential statistics by gender, male or female, and by colleges. The statistics will be displayed depending on the filter chosen for more detailed information for each college.
6.	Forecasting Population	The dashboard able to forecast population of residents and non-residents students in UiTM, Arau Perlis.
7.	Display Basic Residential Statistic	The dashboard can display basic statistics of the residential like total population in UiTM Arau, Perlis, total population by resident type, resident or non-resident, and total population by gender.

Figure 2 illustrates the use case diagram for the requirement modelling for the student residential dashboard. There are three actors: staff, students, and administrators. Both staff and students will only be able to view the dashboard, while the administrator will be able to import any residential data, export statistics and dashboards, generate reports, and gather required data.

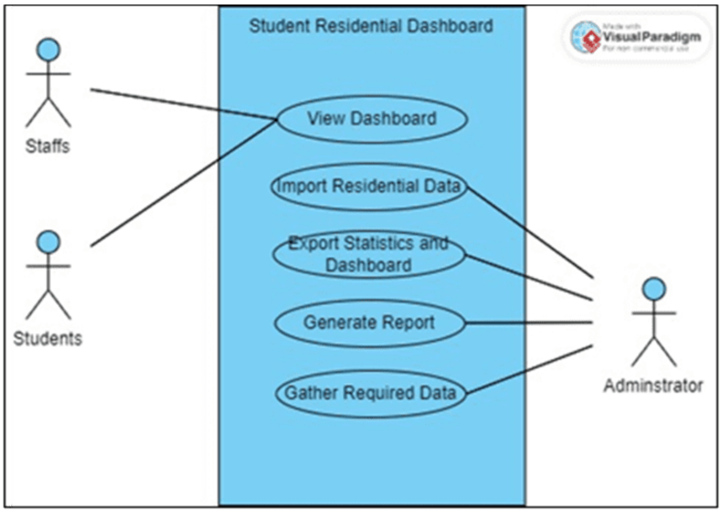

Figure 2: Use Case for Student Residential Dashboard

Non-functional requirements (NFR) are concerned with the quality characteristics, and limitations that define the system's overall behaviour and performance. Table 2 outlines four NFR for the dashboard.

Table 3: Non-Functional Requirements

No.	NFR	NFR Description
1.	Performance	The dashboard should have quick response times to ensure that data and information are shown quickly, even when many concurrent users are accessing the system.
2.	Reliability	The dashboard should be reliable, limiting downtime and guaranteeing that it is available and functional for users in accordance with the desired service level.
3.	Usability	The dashboard should be user-friendly, providing an intuitive interface, clear navigation, and effective organisation of information to enhance the user experience and facilitate ease of use.
4.	Compatibility	The dashboard should be compatible with different web browsers, operating systems, and devices, ensuring consistent functionality and appearance across various platforms.

3.2 Design

3.2.1 Dashboard Sitemap

Figure 3 depicts the dashboard sitemap that represents the hierarchical representation of the pages and structure of a system. It assists in organizing and categorizing the system's pages which can help to enhance and improve the system’s usability. The dashboard consists of four primary pages including the homepage, overview, residence, and forecasting. Each page is designed with specific elements and content that are carefully selected to fulfil the project purpose.

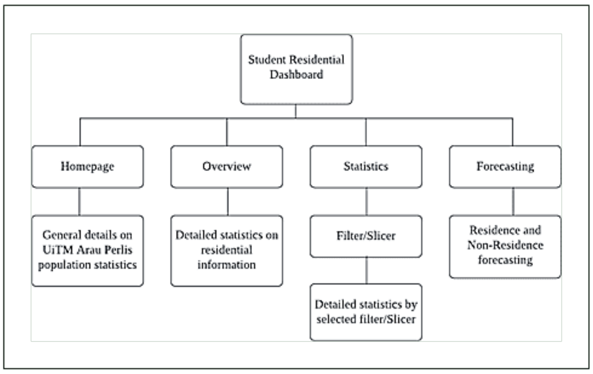

Figure 3: Dashboard Sitemap

3.2.2 Dashboard Wireframe Design

Figure 4 illustrates user interface dashboard design using wireframes. It has four pages, including the homepage (a), overview (b), residence (c), and forecasting (d). The first wireframe is the homepage, which consists of a dashboard banner and college residents' statistics such as male population, female population, student population, resident student population, and non-resident student population. The overview page contains general information about the overall residence, such as total residents, total male and female residents, a list of all colleges available on campus, and the overall residence capacity. The residence page provides in-depth details of college information like the number of residents by program, semester, and course. It also provides filters for gender and colleges at the top of the dashboard page. The forecasting page contains two-line charts that aim to predict the future numbers of resident and non-resident students.

Figure 4: Dashboard Wireframe

3.2.3 Big Data Pipeline

Figure 5 shows the big data pipeline from raw data source to big data platfrom and data visualization.

3.3 Implementation

The pipeline starts with data acquisition from raw data sources. The data source is saved in a CSV file. Then, the raw dataset is extracted from the Windows file system for the Extraction, Transformation, and Loading (ETL) process. The dataset file is transferred to Apache Hadoop big data platform on a local CentOS Linux system that runs under a VMware Workstation virtual machine. The virtual machine runs Cloudera

Distributed Hadoop (CDH) 5.13. The CDH contains Apache Hadoop and other related projects where all the components are 100% open-source under the Apache License. The dataset is loaded from the local Linux file system into the Apache Hadoop distributed data system using Apache Hive Warehouse for scalability and fault-tolerance big data storage. In data analysis uses HiveQL queries. The result of the data query is transferred back to the Linux file system in CSV files. Finally, the CSV files is transferred to Microsoft Power BI to create data visualisations and dashboards.

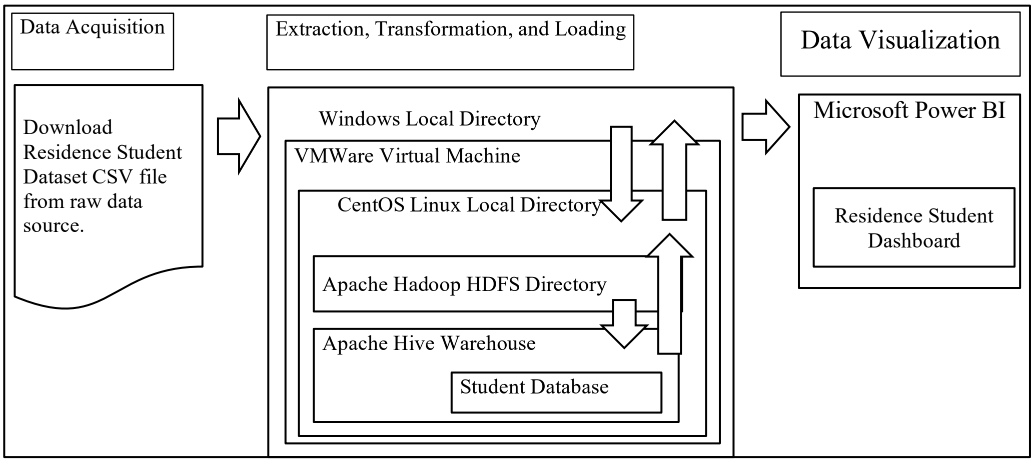

Figure 5: Big Data Visualization Pipeline

3.3.1 ETL

The implementation starts with the ETL process. The ETL process consists of three steps (Nwokeji et al., 2018). In step 1, data was extracted from various sources (local disc, DBMS, streaming data, and many more) in various formats such as txt, csv, and xls. In the data extraction process, the dataset used in this project was retrieved from the Student Affairs Division (SAD). The dataset is received by email from the SAD officer. The dataset containing the college capacity dataset and residence is in comma-separated values (CSV) format.

The second step is data transformation, which ensures that the data is in a uniform and usable format, enabling precise and meaningful analysis to be conducted. During this phase, the data is prepared and modified to ensure its consistency, quality, and suitability for further analysis. Data transformation involves activities like eliminating duplicate entries, handling missing values, and standardising formats. Additionally, unnecessary columns are removed, and new columns may be introduced to streamline the data analysis process. The names of the columns and the values are replaced to ensure the dataset is suitable for further processing. Figure. 6 (a) shows the raw data of the student resident before the cleaning process. Figure 6 (b) shows the dataset that has been cleaned. The header row is removed before it is loaded into the data warehouse. The gender value in the second column is changed from “L” to “Male” and from “P” to “Female”. The value of the race column changed from “melayu” to “Malay”.

	A	B	C	D	E	F	G	H	I	J	K
1	Bil	Jantina	Kaum	Nama Program	Nama Kolej	Nama Fakulti	Kursus	Semester	Status	No bilik	No Katil
2	1	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-001	B-001-1
3	2	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-001	B-001-2
4	3	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-001	B-001-3
5	4	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-001	B-001-4
6	5	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-002	B-002-1
7	6	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-002	B-002-2
8	7	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-002	B-002-3
9	8	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-002	B-002-4
10	9	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-005	B-005-3
11	10	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-005	B-005-4
12	11	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-007	B-007-1
13	12	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-007	B-007-2
14	13	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-007	B-007-3
15	14	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-007	B-007-4
16	15	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-011	B-011-2
17	16	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-011	B-011-3
18	17	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-004	B-004-1
19	18	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-004	B-004-2
20	19	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-004	B-004-3
21	20	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-004	B-004-4
22	21	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-021	B-021-4
23	22	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-021	B-021-1
24	23	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-023	B-023-3
25	24	L	melayu	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	r-residen	B-023	B-023-2

(a)

1	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
2	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
3	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
4	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
5	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
6	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
7	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
8	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
9	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
10	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
11	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
12	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
13	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
14	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
15	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
16	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
17	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
18	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
19	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
20	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
21	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
22	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
23	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
24	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident
25	Male	Malay	Diploma	BERINGIN	Sport Science and Recreation	SR111	1	Resident

(b)

Figure 6: Dataset (a) before cleaning process and (b) after cleaning process

The last step is loading the cleaned data into the data directory on the local disc. To keep the data organized, a folder called "data" is created in the data warehouse to store the dataset in the "studentdata.csv" file in the "/user/cloudera/data" directory before it is loaded into the Apache Hive database.

The student database and table are created in the Apache Hive database. Figure 7 shows the creation of the "Student" table using HiveQL's create table statement. The table is created with nine columns based on the structure of the student dataset in the CSV file.

```
hive> CREATE TABLE student (
    > studentID int,
    > gender String,
    > races String,
    > programme String,
    > college String,
    > faculty String,
    > course String,
    > semester String,
    > resident String
    > ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LOCATION '/user/cloudera/data/student';
OK
Time taken: 0.143 seconds
```

Figure 7: Create Student Table Statement in Hive Terminal

The "studentdata.csv" file is then loaded into the table "Student" that has been created in the Hive in the extraction process as shown in *Figure 8*.

```
hive> LOAD DATA LOCAL INPATH '/home/cloudera/data/studentdata.csv' INTO TABLE student;
Loading data to table student.student
Table student.student stats: [numFiles=1, totalSize=258650]
OK
Time taken: 0.432 seconds
```

Figure 8: Hive Command to load data from studentdata.csv file into the student table

3.3.2 Data Analytic

HiveQL is used to query data in the student database. HiveQL is a query language for Hive to process and analyse structured data, which makes MapReduce tasks more understandable and easier for developers using familiar Structured Query Language (SQL) (Sahal, 2022). The results of the queries are stored in the report table using the CREATE TABLE [report_table] AS [SELECT statement] statement. Each SELECT statement is to query that defined in the functional requirements above. Then, the result table is transferred to the local directory before it is loaded into Microsoft Power BI to create the data visualisation using the INSERT OVERWRITE LOCAL DIRECTORY statement. *Figure 9* shows the query statement that creates the report table totalPopulationGender from an aggregate query of the number of students by gender.

```
hive> CREATE TABLE totalPopulationGender AS SELECT gender, count(studentID) as NumStd
    > FROM student
    > GROUP BY gender;
Query ID = cloudera_20230710231111_9656ab0a-718b-4f06-a4a4-af331358a65c
Total jobs = 1
```

```
hive> SELECT * FROM totalPopulationGender;
OK
Female  3501
Male    2017
Time taken: 0.061 seconds, Fetched: 2 row(s)
```

Figure 9: Query the number of students population by gender and save the query result in the totalPopulationGender table

Figure 10 shows the HiveQL query to find the number of residents by course and semester.

```
Time taken: 0.069 seconds, Fetched: 4 row(s)
hive> CREATE TABLE totalPopulationCourse  AS SELECT course, count(studentID) as NumStd
    > FROM student
    > GROUP BY course;
Query ID = cloudera_20230710232424_4fd415d8-e16e-490c-901e-88a8675666bd
```

Figure 10: Number of residents by course and semester query

All report tables created from the above step are transferred to the local directory using the INSERT OVERWRITE LOCAL DIRECTORY statement as shown in *Figure 11*. This allows the table to be converted to csv files prior to data transfer to Power BI for data visualisation creation.

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/data/extract2'
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > SELECT * FROM totalPopulationGenderResident;
Query ID = cloudera_20230710233535_a30c457e-8f63-437c-92db-d349fbf8a567
```

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/data/extract'
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > SELECT * FROM totalPopulationGender;
Query ID = cloudera_20230710233232_33a24ed6-d205-469a-99b2-28bbe4f38c
```

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/cloudera/data/extract3'
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > SELECT * FROM totalPopulationResidentStatus;
```

Figure 11: Transferring report tables from Hive to local directory in csv files

3.3.3 Basic Students Population Scorecard Visualization

The visualisation and dashboard development processes utilise Microsoft Power BI software. After the data is cleaned, transformed, and analysed using Apache Hive, it is then loaded into Microsoft Power BI software to be visualised using suitable techniques and to develop the dashboard.

The dashboard includes scorecard charts as illustrated in *Figure 12*. The scorecards provide a concise overview of basic student population information on campus. The scorecards show key metrics such as (a) the total student population, (b) the male and female population, and (c) the total residence in a residence college. The scorecard is used to allow users to quickly grasp essential information about the student population at the university.

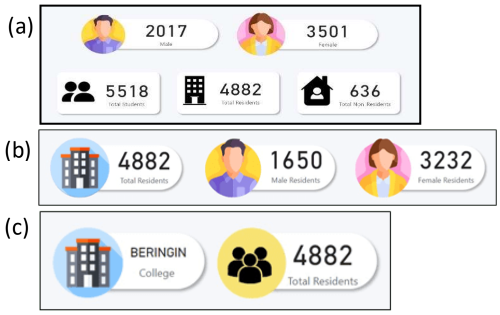

Figure 12: Student Population Scorecard Visualization

3.3.4 Residence Student Capacity Data Visualization

The college lists are visualised in a bar chart, with the number of residents residing in each college, allowing for a comparison of the number of students by college and programme. The bar chart focuses on 14 colleges that are available on campus. *Figure 13*(a) presents the visual representation of this data in the form of a horizontal bar chart. *Figure 13* (b) presents the number of students by programme level in a vertical bar chart.

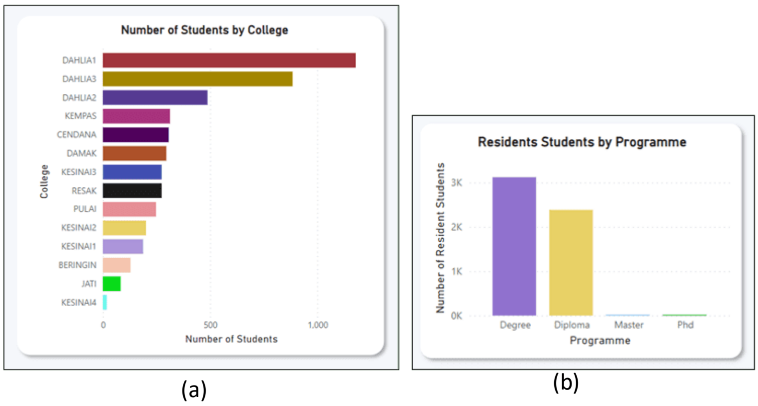

Figure 13: Bar Chart of (a) Number of Students by College (b) Number of Students by Programme

3.3.5 Tree Map Visualization of the Number of Residents by Semester and Program

The number of residents by semester is visualised using a tree map chart, as shown in *Figure 14*. A tree map is a visual representation of hierarchical data using nested rectangles. The main purpose of a tree map is to display the relative sizes or proportions of different categories within a dataset. In this chart, the main categories of the tree map are the semester, and the subset is the course in each semester. There are seven semesters in total, and the sizes of the rectangles depend on the value of the residents.

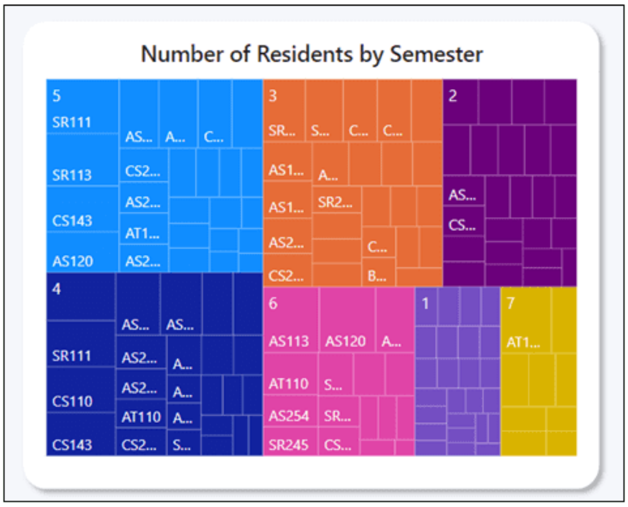

Figure 14: Tree Map Visualization of Number of Residents by Semester

3.3.6 Comparison of Number of Residents and Non-Residents

Figure 15 shows residents and non-resident students' visualisation in a line chart from the year 2020 until 2022. There are two lines in the chart: the blue one indicates non-resident students, and the orange one indicates resident students. These two lines are combined in one chart to allow for the identification of any patterns or relationships between the number of residents and non-resident students.

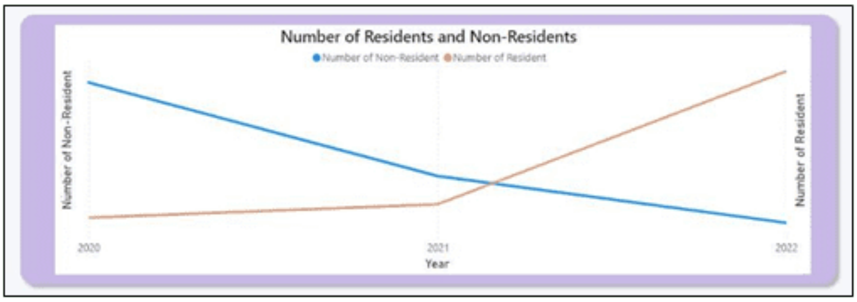

Figure 15: Line Chart of Number of Residents and Non-Residents

3.3.7 Forecasting of Residents

The following line graph displays the number of residents over the years, along with a forecasting projection for the following three years as shown in Figure 15. The orange line represents the current total number of residents, while the black line illustrates the projected values for future years. Additionally, a dotted blue line is included to indicate the current residence capacity available at UiTM Arau, Perlis. This combination allows staff members to assess and ensure sufficient college allocation for students and better decision making based on the forecasted number of residents.

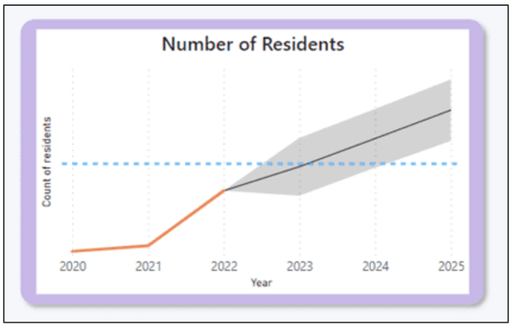

Figure 16: Line Chart of Number of Residents Forecasting

3.3.8 Dashboard

Figure 17 depicts the Student Residential Dashboard Homepage, which consists of four pages: (a) the homepage, which consists of the student residential dashboard banner and a few basic campus statistics; (b) the overview of the residential page, which displays general residential information to users; (c) the residence page, which provides in-depth detail on the residential information; and (d) the forecasting page, which provides the forecasting of the future number of residential students.

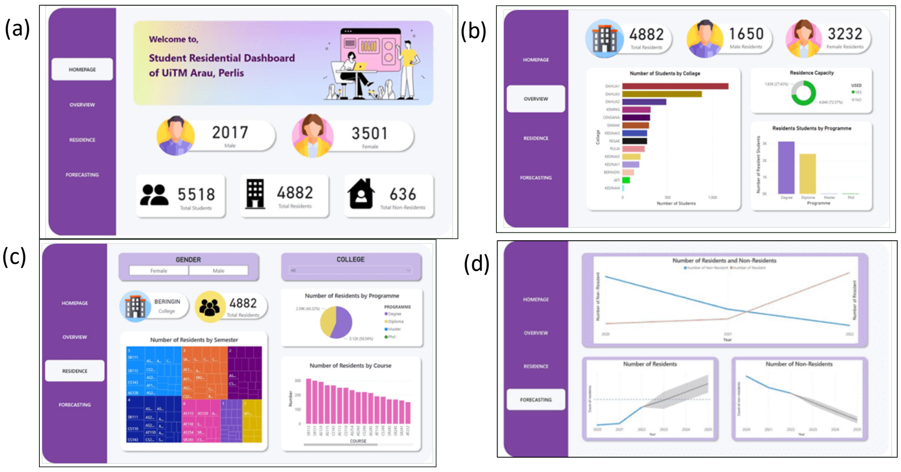

Figure 17: Student Residential Dashboard Homepage

4 Results

4.1 Evaluation Design

User Acceptance Test (UAT) is employed for the evaluation. The Google Form platform is used to generate the questionnaire, and a link to the dashboard is added to the first portion of the questionnaire for responders to access prior to submitting their responses. Only students and personnel of the UiTM Perlis Branch are used to choose the sample of users.

The questionnaire is divided into three sections. The first section is demographic background, which consists of age, status, and gender. The second section is the usability questions on the dashboard, which comprise a set of ten questions. The third section consists of the comments and suggestions section, allowing them to give their own feedback on the dashboard.

In the evaluation process, a total of 30 participants were involved, and their responses were collected using a Google Form. The collected data was analysed and processed to evaluate and determine the usability of the developed dashboard. This analysis and processing of the data allowed for insights into how effectively the dashboard met the needs and expectations of the users, as well as identifying any areas for improvement.

Figure 16 shows the bar graph of mean score for all items. The overall average score of 4.71 indicates that respondents are satisfied with the usefulness of the student residence dashboard. The average score is between 4.57 and 4.97, indicating that respondents are satisfied with the dashboard. 4.57 is the lowest score on the second question (I can easily understand the content of the dashboard). It can be construed to suggest that it may take some time for the respondent to comprehend the dashboard's content. The highest average score on the first question is 4.97. (I find the dashboard interface pleasant and well-organized). In light of this, it can be determined that the dashboard interface is visually appealing in terms of the colour scheme, iconography, and organisation of the charts and graphs. Overall, the built dashboard is user-friendly and beneficial for both students and staff at the UiTM Perlis Branch.

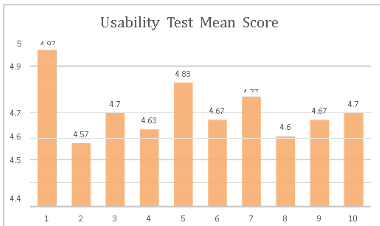

Figure 18: Usability Test Mean Score

5 Conclusion

The Student Residential Dashboard in the UiTM Perlis Branch was designed with three primary research goals in mind. The initial objective is to assess the need for a residential dashboard for students at the UiTM Perlis Branch. In building the system, the system requirements are gathered through a comprehensive research study based on prior projects and interviews with the staff. The research and interviews gathered provided as references and suggestions for the dashboard development project.

This project's second purpose is to build a Student Residential Dashboard System for UiTM Perlis Branch. This purpose is accomplished by implementing the methods outlined in the introduction. A wireframe drawing, a use case diagram, and a dashboard sitemap are created during the dashboard development process. Data is loaded into the big data platform based on the Apache Hadoop distributed data platform and the Apache Hive data warehouse. After the data analysis using HiveQL, the data is loaded into Microsoft Power BI for visualisation. Several relevant charts are used to display the data on the dashboard. The dashboard is then made public, and users can view it using the link that has been supplied. The dashboard has efficiently displayed information on general college-related details. However, the dashboard is unable to provide specifics regarding the college's future admissions capacity because it simply displays the college's overall capacity. This work has been successful in the implementation of student residential dashboards using a big data visualisation platform.

The final objective is to assess the residential dashboard's usability for UiTM Arau, Perlis, students. Google Form's usability was evaluated using User Acceptance Test (UAT) to accomplish this objective. The results of the usability testing demonstrated that the created dashboard is advantageous for users to acquire college-related information and assists college officials in making better allocation decisions.

The dashboard can be enhanced by incorporating data from various domains, such as demographic factors, academic performance, and extracurricular activities. For instance, it is important to analyse trends about the student's current hometown, household income, most recent current grade point average (CGPA), and extracurricular involvement. Personnel from the Student Affairs Division can use this data to gain a deeper understanding of applicants and make informed judgments regarding the prioritising of college room allocations for suitable students.

In addition, the system can be improved by integrating the dashboard to the college database system at UiTM Perlis Branch so that the dashboard can be updated in real-time, hence making the dashboard more efficient because it allows for live updates. This would enable greater flexibility and ensure that the displayed information is always updated and reflective of the most recent changes.

During the usability test, participants requested adding informative tooltips to the Power BI dashboard. Tooltips can greatly enhance the user experience by providing more context and direction. The tooltips could provide relevant information, explanations, or definitions for data items, visualisations, or dashboard elements. By reducing uncertainty and enhancing usefulness, tooltips can significantly enhance the user experience. They can aid users, especially those who are inexperienced witbvh the data or Power BI's capabilities, in navigating the dashboard correctly.

Acknowledgments

This work was supported by the College of Computing, Informatics, and Mathematics and Student Affair Division, Universiti Teknologi MARA, Perlis Branch, who provided valuable comments and the raw dataset for this work.

References

Alblawi, A. S. (2018). Big Data and Learning Analytics in Higher Education: Demystifying Variety, Acquisition, Storage, NLP and Analytics (arXiv:1801.06052). arXiv. https://doi.org/10.48550/arXiv.1801.06052

Amare, M. Y., & Šimonová, S. (2021). Learning analytics for higher education: Proposal of big data ingestion architecture. https://doi.org/10.1051/shsconf/20219202002

Aprillia, D. A., Noranita, B., & Nugraheni, D. M. K. (2021). Dashboard Design as Monitoring System Distribution of Government Assistance at Balai Besar Perikanan Budidaya Air Payau (BBPBAP) Jepara, Central Java. 2021 5th International Conference on Informatics and Computational Sciences (ICICoS), 187–192. https://doi.org/10.1109/ICICoS53627.2021.9651776

Bhattarai, B. P., Paudyal, S., Luo, Y., Mohanpurkar, M., Cheung, K., Tonkoski, R., Hovsapian, R., Myers, K. S., Zhang, R., Zhao, P., Manic, M., Zhang, S., & Zhang, X. (2019). Big data analytics in smart grids: State-of-the-art, challenges, opportunities, and future directions. IET Smart Grid, 2(2), 141–154. https://doi.org/10.1049/iet-stg.2018.0261

Bikakis, N. (2018). Big data visualization tools. arXiv Preprint arXiv:1801.08336.

Camacho-Rodríguez, J., Chauhan, A., Gates, A., Koifman, E., O'Malley, O., Garg, V., Haindrich, Z., Shelukhin, S., Jayachandran, P., Seth, S., Jaiswal, D., Bouguerra, S., Bangarwa, N., Hariappan, S., Agarwal, A., Dere, J., Dai, D., Nair, T., Dembla, N., … Hagleitner, G. (2019). Apache Hive: From MapReduce to Enterprise-grade Big Data Warehousing. Proceedings of the 2019 International Conference on Management of Data, 1773–1786. https://doi.org/10.1145/3299869.3314045

Cheng, Y., Liu, F. C., Jing, S., Xu, W., & Chau, D. H. (2018). Building Big Data Processing and Visualization Pipeline through Apache Zeppelin. Proceedings of the Practice and Experience on Advanced Research Computing, 1–7. https://doi.org/10.1145/3219104.3229288

Cui, Y., Kara, S., & Chan, K. C. (2020). Manufacturing big data ecosystem: A systematic literature review. Robotics and Computer-Integrated Manufacturing, 62, 101861. https://doi.org/10.1016/j.rcim.2019.101861

Daliri Dizaj, M., & Hatami Khanghahi, T. (2022). Students' residential preferences: A case study is dormitories of University of Mohaghegh Ardabili. Journal of Asian Architecture and Building Engineering, 21(4), 1348–1363. https://doi.org/10.1080/13467581.2021.1941987

Deshai, N., Venkataramana, S., Sekhar, B. V. D. S., Srinivas, K., & Saradhi Varma, G. P. (2020). A Study on Big Data Processing Frameworks: Spark and Storm. In S. C. Satapathy, V. Bhateja, J. R. Mohanty, & S. K. Udgata (Eds.), Smart Intelligent Computing and Applications (pp. 415–424). Springer. https://doi.org/10.1007/978-981-32-9690-9_43

Destiandi, N., & Hermawan, A. (2018). Business Intelligent Method For Academic Dashboard. Bit-Tech, 1(2), Article 2. https://doi.org/10.32877/bt.v1i2.42

Gandhi, P., & Pruthi, J. (2020). Data Visualization Techniques: Traditional Data to Big Data. In S. M. Anouncia, H. A. Gohel, & S. Vairamuthu (Eds.), Data Visualization: Trends and Challenges Toward Multidisciplinary Perception (pp. 53–74). Springer. https://doi.org/10.1007/978-981-15-2282-6_4

Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. IDC Iview, 1142(2011), 1–12.

Ishwarappa, & Anuradha, J. (2015). A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. Procedia Computer Science, 48, 319–324. https://doi.org/10.1016/j.procs.2015.04.188

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6(70), 1.

Li, Y., & Zhang, D. (2020). Hadoop-Based University Ideological and Political Big Data Platform Design and Behavior Pattern Mining. 2020 International Conference on Advance in Ambient Computing and Intelligence (ICAACI), 47–51. https://doi.org/10.1109/ICAACI50733.2020.00014

Lucio, R., Campbell, M., Detres, M., & Johnson, H. (2018). Using Dashboards to Engage Faculty in Improving Academic Programs and Courses. INTED2018 Proceedings, 1844–1852. https://doi.org/10.21125/inted.2018.0327

Małysiak-Mrozek, B., Wieszok, J., Pedrycz, W., Ding, W., & Mrozek, D. (2022). High-Efficient Fuzzy Querying With HiveQL for Big Data Warehousing. IEEE Transactions on Fuzzy Systems, 30(6), 1823–1837. https://doi.org/10.1109/TFUZZ.2021.3069332

Mjhool, A. Y., Alhilali, A. H., & Al-Augby, S. (2019). A proposed architecture of big educational data using hadoop at the University of Kufa. International Journal of Electrical and Computer Engineering, 9(6), 4970.

Mohd Janor, R. (2023). Kebolehpasaran global: Hala tuju UiTM 2023/Prof. Ts. Dr. Roziah Mohd Janor.

Nwokeji, J. C., Aqlan, F., Anugu, A., & Olagunju, A. (2018). Big Data ETL Implementation Approaches: A Systematic Literature Review (P). SEKE, 714–713.

Osman, M. N., Othman, N. A., & Sedek, K. A. (2023). A Web-Based Student Residential Searching Integrated with Google Maps and WhatsApp API: A Case Study of UiTM Arau Perlis. Journal of Computing Research and Innovation, 8(2), Article 2. https://doi.org/10.24191/jcrinn.v8i2.372

Pence, H. E. (2014). What is Big Data and Why is it Important? Journal of Educational Technology Systems, 43(2), 159–171. https://doi.org/10.2190/ET.43.2.d

Qin, X., Luo, Y., Tang, N., & Li, G. (2020). Making data visualization more efficient and effective: A survey. The VLDB Journal, 29(1), 93–117. https://doi.org/10.1007/s00778-019-00588-3

Riahi, Y., & Riahi, S. (2018). Big data and big data analytics: Concepts, types and technologies. International Journal of Research and Engineering, 5(9), 524–528.

Sahal, R. (2022). How Apache MapReduce Handles Big Data Query?

Sharma, G., Tripathi, V., & Srivastava, A. (2021). Recent Trends in Big Data Ingestion Tools: A Study. In R. Kumar, N. H. Quang, V. Kumar Solanki, M. Cardona, & P. K. Pattnaik (Eds.), Research in Intelligent and Computing in Engineering (pp. 873–881). Springer.

Shaulska, L., Yurchyshena, L., & Popovskyi, Y. (2021). Using MS Power BI Tools in the University Management System to Deepen the Value Proposition. 2021 11th International Conference on Advanced Computer Information Technologies (ACIT), 294–298. https://doi.org/10.1109/ACIT52158.2021.9548447

Zhai, Y., Tchaye-Kondi, J., Lin, K.-J., Zhu, L., Tao, W., Du, X., & Guizani, M. (2021). Hadoop Perfect File: A fast and memory-efficient metadata access archive file to face small files problem in HDFS. Journal of Parallel and Distributed Computing, 156, 119–130. https://doi.org/10.1016/j.jpdc.2021.05.011