Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute InformingScience.org

IJIKM.org

Volume 18, 2023

A NEW MODEL FOR COLLECTING, STORING, AND ANALYZING BIG DATA ON CUSTOMER FEEDBACK IN THE TOURISM INDUSTRY

Thanh Ho	University of Economics and Law, Ho Chi Minh City, Vietnam; Vietnam National University, Ho Chi Minh City, Vietnam	thanhht@uel.edu.vn
Van-Ho Nguyen*	University of Economics and Law, Ho Chi Minh City, Vietnam; Vietnam National University, Ho Chi Minh City, Vietnam	honv@uel.edu.vn
Thien Le	University of Economics and Law, Ho Chi Minh City, Vietnam; Vietnam National University, Ho Chi Minh City, Vietnam	thienlb.ktl@uel.edu.vn
Hoanh-Su Le	University of Economics and Law, Ho Chi Minh City, Vietnam; Vietnam National University, Ho Chi Minh City, Vietnam	sulh@uel.edu.vn
Thon-Da Nguyen	University of Economics and Law, Ho Chi Minh City, Vietnam; Vietnam National University, Ho Chi Minh City, Vietnam	dant@uel.edu.vn
Thi Cam-Tu Mai	University of Economics and Law, Ho Chi Minh City, Vietnam; Vietnam National University, Ho Chi Minh City, Vietnam	tumtc@uel.edu.vn
Thi-Anh Tran	University of Economics and Law, Ho Chi Minh City, Vietnam; Vietnam National University, Ho Chi Minh City, Vietnam	anhtt@uel.edu.vn
Hoai-Phan Truong	University of Economics and Law, Ho Chi Minh City, Vietnam; Vietnam National University, Ho Chi Minh City, Vietnam	hoaiphan@uel.edu.vn

Accepting Editor Natasha Boskic | Received: February 1, 2023 | Revised: April 12, April 17, 2023 | Accepted: April 19, 2023.

Cite as: Ho, T., Nguyen, V.-H., Le, T., Le, H.-S., Nguyen, T.-D., Mai, T. T.-C., Tran, T.-A., & Truong, H. P. (2023). A new model for collecting, storing, and analyzing big data on customer feedback in the tourism industry. *Interdisciplinary Journal of Information, Knowledge, and Management*, *18,* 225-249. https://doi.org/10.28945/5107

ABSTRACT

Aim/Purpose	In this study, the research proposes and experiments with a new model of collecting, storing, and analyzing big data on customer feedback in the tourism industry. The research focused on the Vietnam market.
Background	Big Data describes large databases that have been "silently" built by businesses, which include product information, customer information, customer feedback, etc. This information is valuable, and the volume increases rapidly over time, but businesses often pay little attention or store it discretely, not centrally, thereby wasting an extremely large resource and partly causing limitations for business analysis as well as data.
Methodology	The study conducted an experiment by collecting customer feedback data in the field of tourism, especially tourism in Vietnam, from 2007 to 2022. After that, the research proceeded to store and mine latent topics based on the data collected using the Topic Model. The study applied cloud computing technology to build a collection and storage model to solve difficulties, including scalability, system stability, and system cost optimization, as well as ease of access to technology.
Contribution	The research has four main contributions: (1) Building a model for Big Data collection, storage, and analysis; (2) Experimenting with the solution by collecting customer feedback data from huge platforms such as Booking.com, Agoda.com, and Phuot.vn based on cloud computing, focusing mainly on tourism Vietnam; (3) A Data Lake that stores customer feedback and discussion in the field of tourism was built, supporting researchers in the field of natural language processing; (4) Experimental research on the latent topic mining model from the collected Big Data based on the topic model.
Findings	Experimental results show that the Data Lake has helped users easily extract information, thereby supporting administrators in making quick and timely decisions. Next, PySpark big data processing technology and cloud computing help speed up processing, save costs, and make model building easier when moving to SaaS. Finally, the topic model helps identify customer discussion trends and identify latent topics that customers are interested in so business owners have a better picture of their potential customers and business.
Recommendations for Practitioners	Empirical results show that facilities are the factor that customers in the Vietnamese market complain about the most in the tourism/hospitality sector. This information also recommends that practitioners reduce their expectations about facilities because the overall level of physical facilities in the Vietnamese market is still weak and cannot be compared with other countries in the world. However, this is also information to support administrators in planning to upgrade facilities in the long term.
Recommendations for Researchers	The value of Data Lake has been proven by research. The study also formed a model for big data collection, storage, and analysis. Researchers can use the same model for other fields or use the model and algorithm proposed by this study to collect and store big data in other platforms and areas.
Impact on Society	Collecting, storing, and analyzing big data in the tourism sector helps government strategists to identify tourism trends and communication crises. Based on that information, government managers will be able to make decisions and

	strategies to develop regional tourism, propose price levels, and support innovative programs. That is the great social value that this research brings.
Future Research	With each different platform or website, the study had to build a query scenario and choose a different technology approach, which limits the ability of the solution's scalability to multiple platforms. Research will continue to build and standardize query scenarios and processing technologies to make scalability to other platforms easier.
Keywords	data lake for big data, data collection, data storage, customer feedback, topic modeling, tourism

INTRODUCTION

According to a survey and compilation by Luca Clissa (2022), as of 2021, Amazon S3 has stored 100 trillion objects. Assuming 5MB per object, the volume of data that Amazon S3 is storing is up to 500EB (Exabyte). Meanwhile, according to Luca Clissa's summary, in 2021, every minute, 240 thousand photos are uploaded to Facebook, 65,000 photos are uploaded to Instagram. Meanwhile, every day, the amount of video uploaded to the YouTube system is equivalent to 720 thousand hours. The above figures have painted a picture of the huge data that exists around us, promising to bring huge benefits and value in the field of Big Data analysis.

The tourism industry is one of the industries that plays an important role in the economic growth of countries. Accordingly, 10.4% of global Gross Domestic Product (GDP) and 330 million jobs were an excellent contribution to the tourism industry in 2019 (Nhandan, 2022). However, the COVID-19 pandemic has caused great damage to the tourism industry. In Vietnam, in 2020, the number of international visitors decreased by 78.7% compared to 2019, domestic tourists decreased by about 50%, 530 trillion is the loss estimated by the General Statistics Office (T. T. Nguyen, 2022). Also, according to the research, thousands of companies operating in the tourism sector went bankrupt when they could not survive due to the government's prolonged blockade order. However, COVID-19 is also considered as an opportunity to purge weak companies; therefore, entering the post-COVID period, businesses that are surviving in the industry will be strong ones. This industry will not stop competing to recover from the recent difficult period.

Based on the objective conditions of the tourism industry, combined with the rapid development of technology in the fourth industrial revolution, the study sees how big data technology can help businesses, especially businesses in the tourism sector in recovering and developing tourism after the pandemic. Therefore, this study focuses on the tourism sector. However, Big Data in general and Big Data in tourism in particular are being stored in a distributed and heterogeneous manner. In each enterprise data is also stored in different departments, especially within SMEs when there is no centralized data warehouse to aggregate, process, and analyze data for businesses to support, help businesses make decisions (Azeroual & Theel, 2018). Meanwhile, in the current digital era, customer exchanges, comments, or reviews on online platforms are an extremely valuable resource for businesses, but businesses have not focused on exploiting these (H. Le et al., 2017; Rajendran et al., 2023). Based on this motivation, the research focuses on building a Big Data collection and storage model to help businesses synthesize and analyze data more efficiently. The study conducted an experiment by collecting customer feedback data on several booking platforms and exchanges on travel forums in Vietnam.

This study is divided into five parts. In the next section, related works will be presented as a basis for the solution proposed by this study. In part three, the study will present in detail the proposed model and the experimental process. In part four, the study will discuss the results and evaluate the results with the criteria of Big Data. Finally, the conclusion will be presented in part five, re-evaluating the research points that have been made and limitations, as well as the development direction of the research.

RELATED WORKS

Big Data is a term that describes data that has a large volume, fast data generation rate, and a complex structure that makes it difficult to collect, store and process data. The concept of Big Data was first described by Douglas Laney (2001) through the 3V model: (1) Volume; (2) Velocity; (3) Variety. Over time, with the change of technology, when it comes to Big Data, the 5V model has gradually been replaced by the 3V model with the original 3V and further developed 2V, including (4) Veracity; (5) Value (Demchenko et al., 2014). The 5V model will also not stop when new technologies arrive or other perspectives are considered when looking at Big Data, specifically the authors in the study by Khan et al. (2019) mentioned the 51V model of Big Data, which includes some Vs such as Visualization and Availability.

Nowadays, the term Big Data has become popular, and the value that Big Data brings to businesses has also been verified in many fields (Ciampi et al., 2021; Monino, 2021; Wang et al., 2022), so enterprises are constantly applying technological advances to collect and build large data warehouses to support businesses in decision making.

Research by Wah et al. (2007) introduced a framework for building the library data warehouse. The Extract-Transform-Load (ETL) process has been centralized in processing data from various sources. The goal of this paper is to explore steps in developing a data warehouse for a library system and, therefore, provide a framework that simplifies the process of developing a library data warehouse. This study also proved that building a data warehouse with only internal data is not enough; the value from external data is much larger and businesses need to pay attention and focus on mining this amount of data. Research by Hamoud and Obaid (2013) has also proposed building a data warehouse for the medical field. The results from the study indicate that Online Analytical Processing (OLAP) can be used to analyze data and find relationships between many factors and provide a good view of the data from multiple perspectives. The use of Data Warehouses and OLAP techniques is a good choice to help professionals and analysts meet goals. However, along with the change in analysis needs, combined with the development of science and technology, Data Warehouse faces many challenges. Research by Armbrust et al. (2021) has shown several points, including incorrect data, staleness, and high costs. Figure 1 presents an illustration of the various aspects and components for building a data warehouse and data lake.

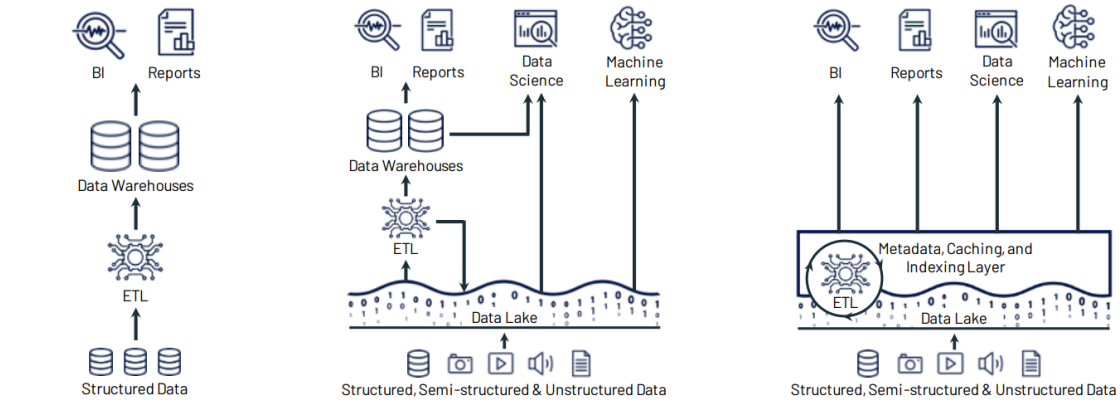

Figure 1: Building Data Warehouse and Data Lake (Armbrust et al., 2021)

Today, Data Lake and Data Warehouse are all widely used to store Big Data, but they are not interchangeable terms. The Data Lake is a vast collection of raw data; its purpose still undetermined. A Data Warehouse is a repository of structured, filtered, and processed data for a specific purpose. There is even an emerging data management architecture trend of the data lake house, which combines the flexibility of the Data Lake with the data management capabilities of the Data Warehouse.

Table 1: Compare Data Warehouse and Data Lake (Amazon Web Services [AWS], n.d.)

CRITERIA	DATA WAREHOUSE	DATA LAKE
Data	Relational data from transactional systems, operational databases, and line of business applications	All data, including structured, semi-structured, and unstructured
Schema	Often designed prior to the data warehouse implementation but also can be written at the time of analysis. (schema-on-write or schema-on-read)	Written at the time of analysis (schema-on-read)
Performance	Fastest query results using local storage	Query results getting faster using low-cost storage and decoupling of compute and storage
Data quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (i.e., raw data)
Users	Business analysts, data scientists, and data developers	Business analysts (using curated data), data scientists, data developers, data engineers, and data architects
Analytics	Batch reporting, BI, and visualizations	Machine learning, exploratory analytics, data discovery, streaming, operational analytics, big data, and profiling

Based on the comparison in Table 1, the study selected Data Lake as the storage technology in the current study. The development of Big Data is based on unstructured data; in this study, customer feedback data is also unstructured data.

In the tourism sector, studies (Alcántara-Pilar et al., 2017; Jackson, 2016) have looked at Big Data as a means to lead businesses through growth challenges. These studies focus on qualitative and descriptive statistical methods. However, in the process of exchanging and responding to information on online platforms, customers generate a huge amount of data (T. Le et al., 2022; V. H. Nguyen & Ho, 2023). All these data will create great value when businesses conduct analysis to understand customer experience, but currently, there are not many studies combining qualitative and experimental methods. There are many reasons for this, one of which is that these data are often stored in a distributed way, so there needs to be a method for a Data Lake to be able to aggregate large amounts of semi-structured and unstructured data. With a centralized Data Lake, researchers and businesses can easily analyze the data. This was a void that motivated this study.

Many researchers and studies (Doreswamy et al., 2017; Lin et al., 2021) have developed the Big Data analysis, processing, and storage model. They use Hadoop Distributed File System (HDFS) technology to store Big Data and apply technologies in the Hadoop ecosystem to collect and process data. However, this faces many obstacles including difficulty in accessing HDFS technology and related technologies in the Hadoop ecosystem. To solve this problem, this study uses Amazon S3 storage service because of its ease of use, high security, and availability that make Amazon S3 easily accessible (Han, 2015). Services like Amazon Web Services (AWS) Batch are also easy to configure, due to their service-oriented (Software as a Services – SaaS) design (Bracci et al., 2012). In addition, cloud-based services only really charge when users use the service, which is the reason why researchers choose the AWS platform to test solutions to optimize costs (Mukherjee, 2019).

Olmedilla et al. (2016) conducted data collection using the XPath language, which is designed to support and transform Extensible Markup Language (XML) documents and analyze Hypertext Markup Language (HTML) code. Using XPath language alone is not sufficient in the current fast-paced technology era. Therefore, this study combines XPath method (which analyzes HTML code), APIs, and JavaScript-related techniques to collect the necessary data from online platforms. Using API and Python code in research also ensures easy re-use and vendor change (Thien et al., 2021).

Today there are many systems that use Hadoop for Big Data analysis and processing. The biggest advantage of Hadoop is that it is based on a parallel programming model for big data processing, MapReduce, which allows for scalable computing, flexibility, fault tolerance, and low cost. This reduces the processing time for large data and maintains speed, avoiding delays when the data volume increases. Although there are many strengths in parallel computing and high fault tolerance, Apache Hadoop has a disadvantage that all operations must be performed on the hard drive, which reduces the computation speed many times. To overcome this drawback, Apache Spark was born. Apache Spark can run 10 times faster than Hadoop on hard disk and 100 times faster when running on RAM (Drabas & Lee, 2017).

Apache Spark is an open-source cluster computing framework originally developed in 2009 by AMPLab. Later, Spark was given to Apache Software Foundation in 2013 and developed to date.

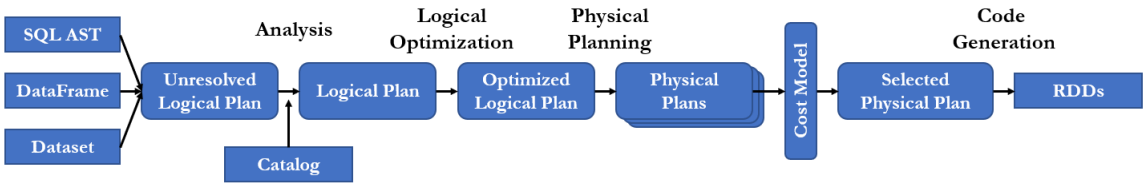

Figure 2: Spark Execution Flow (Drabas & Lee, 2017)

Figure 2 depicts the Spark Execution Flow process, including several main stages: (1) The code written is first noted as an Unresolved Logical Plan, if it is valid Spark converts this into a Logical Plan; (2) The Logical Plan is passed through the Catalyst Optimizer to apply the optimized rules; (3) The Optimized Logical Plan is then converted into a Physical Plan; (4) The Physical Plan is done by the Spark executors (Rimmalapudi, 2023).

Apache Spark gets the support of high-level libraries such as streaming data, Structured Query Language (SQL) queries, machine learning, and graph processing. Not only do these standard libraries increase developer productivity, but it also ensures seamless connectivity for complex workflows. Therefore, this study chose PySpark as the Big Data processing technology.

PROPOSED MODEL AND METHODOLOGY

The new model shown in Figure 3 is proposed with four layers: Data Sources, Collection, Storage, and Analysis. At the Data Sources layer, the study presents several booking and information exchange platforms in the field of tourism that the research conducted experimentally. Next, at the Collection layer, the research introduces methods to collect data based on Python code, combining cloud-based technologies. Then, the data will be integrated into the Data Lake for Big data, which is also the Storage layer of the proposed model. Finally, the experimental study analyses the latent topic in a small set of data that the study has extracted based on the topic model using PySpark technology (Apache Spark, n.d.).

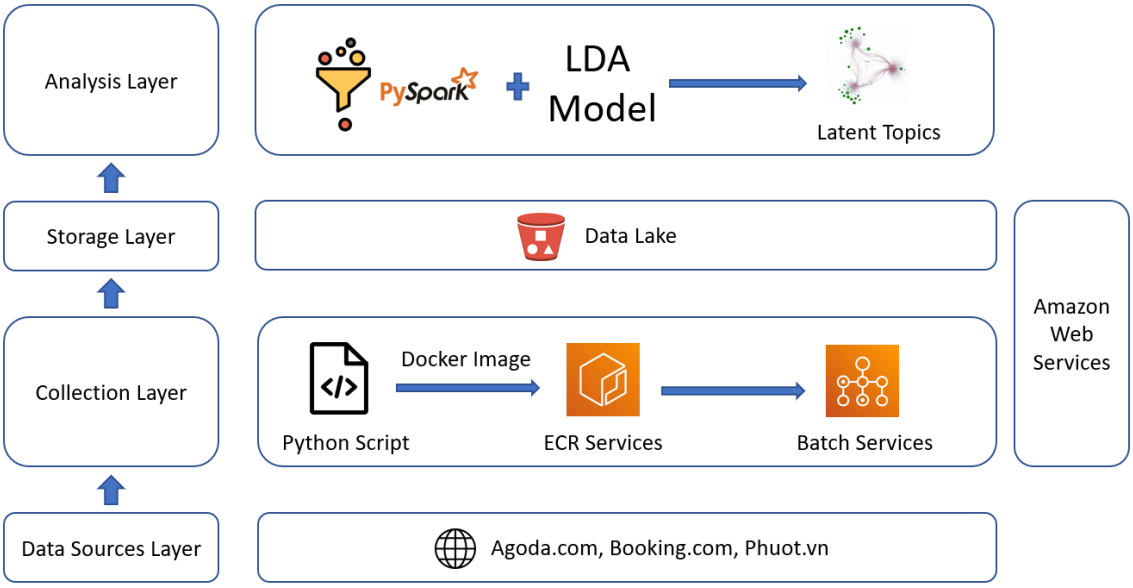

Figure 3: A new model and methodology for collecting, storing and analyzing Big Data

Data Sources Layer

Agoda.com is an online hotel booking platform headquartered in Singapore, founded in 2003. Hotels on this platform are concentrated in the Asia-Pacific region. In 2008, Agoda was voted the best accommodation provider in Asia by TravelMole website. Booking.com is one of the world's largest online booking platforms with presence in 228 countries and territories, headquartered in the United States and established in 1996. Meanwhile, Phuot.vn is a tourism information exchange forum for the community of people who love travel in Vietnam. Here, users can post to share travel moments and travel experiences. Users can also interact with each other by commenting.

According to statistics from Statista in 2020 (M. N. Nguyen, 2022), Booking and Agoda are the two online booking platforms with the largest market share in Vietnam. This study focuses on Vietnam tourism; therefore, research was conducted using data collected from Booking and Agoda. On the other hand, Phuot.vn website was selected because it is one of the diverse tourism news exchange platforms with a large number of active members, constantly updating information in Vietnam. In addition, due to different purposes of use, the information structure of platforms such as Agoda and Booking compared to Phuot.vn will be very different. This is also one of the reasons to research and choose these websites for the purpose of collecting and storing information, to diversify data sources, data structures and data types.

This research focuses on collecting and storing customer feedback, but does not separate other relevant information, including (1) Which hotel is that feedback from, which hotel is in? Which city in Vietnam? (2) What are customers comments? (3) Feedback belongs to which article, etc.

Figure 4 is a sample of user feedback on the Booking platform, with easy to see feedback, including (1) Respondent; (2) Feedback language; (3) Evaluation date; (4) Overall feedback (5) Positive feedback; (6) Negative feedback; (7) Evaluation score; (8) Room type; (9) Date of hire; (10) Number of rental days; (11) Guest type. In addition, this feedback is evaluated for a specific hotel, so the research is conducted to collect more hotel information, hotel address information, type of room booked by customers, etc. With the Agoda platform, the study also found that there are similarities in information structure compared to the Booking platform, so the study also collects similar fields of information.

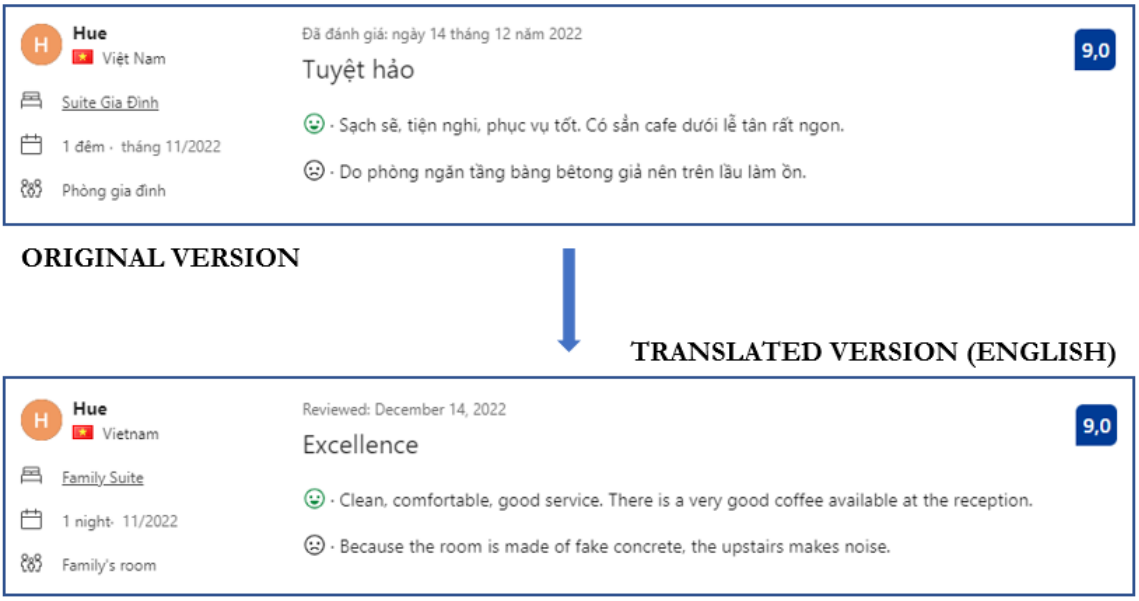

Figure 4: Comment form from Booking.com website
(Source: Captured from Booking.com by Authors)

Unlike Booking and Agoda, Phuot.vn is an information exchange forum, so the data is stored in the form of articles and a set of comments according to articles, not containing information such as hotels, room types, etc. Figure 5 is a sample post on the forum when users ask about travel experiences in the Northwest. Analyzing these articles and feedback, the study found that it is possible to identify tourism trends in the Vietnamese community. Combined with comments from booking platforms, the research will have a clearer and more comprehensive view of tourism trends and customer experiences in the field of tourism, with a focus on Vietnam tourism.

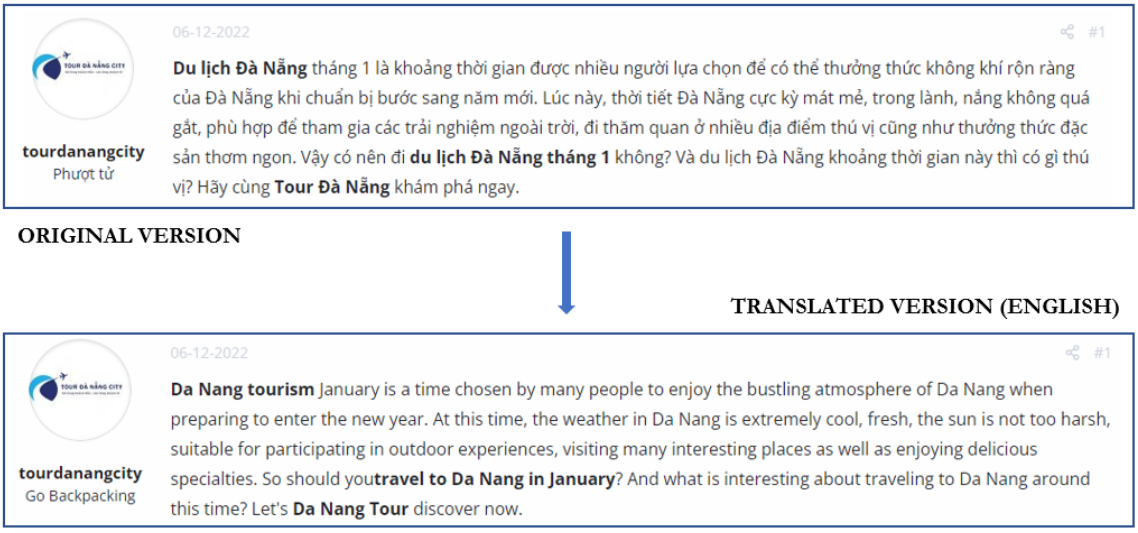

Figure 5: Comment form on Phuot.vn website
(Source: Captured from Phuot.vn by Authors)

COLLECTION LAYER

With each platform, the data and processing technology will be different. In this research scope, the study introduces only some general concepts and algorithms that the research has approached to collect data. Algorithm 1 describes how the study collects data from the Agoda platform.

Algorithm 1: Data Collection Algorithm

```
M = [ ]
Initialize a list of 63 provinces and cities in Vietnam
P = ['an giang, 'ba ria – vung tau', 'bac lieu' … 'ha noi', 'ho chi minh' …]
for p in P do
    (1) Information about the province/city you are looking for
    H = Gather a list of hotels in the province/city p
    for h in H do
        (2) Hotel details
        - Address
        - Images
        - Utilities
        - …
        R = Gather a list of hotel customer feedback
        for r in R do
            (3) Customer feedback r
            - Room type
            - Travel type
            - Response date
            - Feedback (General, Positive, Negative)
            - Rating
            - ….
            M += {(1), (2), (3)}
        end for
    end for
end for
```

As shown in Figure 6, the first step in the data collection process is to identify the target website and the data within that website to be collected. Once the objective is defined, the website is searched for, and its structure is identified. The next step is to determine the crawling method, and environment for web scraping is set up, which includes selecting a browser, creating Hypertext Transfer Protocol (HTTP)/Hypertext Transfer Protocol Secure (HTTPs) requests, and defining input parameters for the web scraping tool or library. After that, the website is accessed using HTTP/HTTPs requests, and the data is parsed using the web scraping tool or library. The parsed data is then stored in a database or file for further use. Finally, the code is optimized to crawl more websites, and any errors that occur during the web scraping process are fixed.

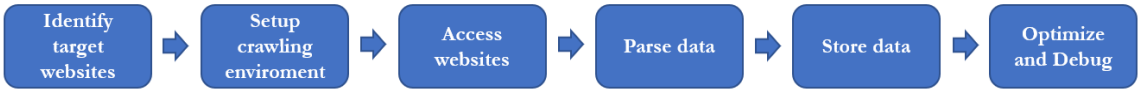

Figure 6: The flow of the data source collection process

The territory of Vietnam is divided into 63 provinces/cities. Therefore, a set of 63 provinces/cities in Vietnam is created and stored in the database. Research uses Application Programming Interface (API) to collect hotel information of each province/city. Next, the research extracts customer feedback comments from each hotel. Finally, a set of information related to customer feedback, combining hotel information is created.

Currently, Agoda and Booking platforms are designing APIs in two approaches, Rest API and GraphQL. Research uses these two types of APIs for data collection. Meanwhile, with Phuot.vn website, the research combines API and HTML code analysis methods because this website design technology is different from Agoda and Booking platforms, so research cannot just use API method to collect data. In general, in each specific case, the research is conducted using API method or HTML code analysis method, combining JavaScript-related techniques to extract data.

For data extraction scripts, the study uses Python code to execute. Next, the study uses Docker services to package Python code and related libraries into Docker Image and push Docker Image to Amazon Elastic Container Registry (ECR) service. Then the study uses the AWS Batch service to execute Docker Image from ECR.

Figure 7 depicts the process of scheduling and executing the AWS Batch service. Research uses the Amazon EventBridge service to schedule a day, a week, or a certain amount of time to execute a task, in this case launching the AWS Batch service. From there, a Worker/Job is created, which executes all the code in the Docker Image, which are the dimensions to collect the study's data. Finally, the specified data is stored in the Amazon S3 Bucket.

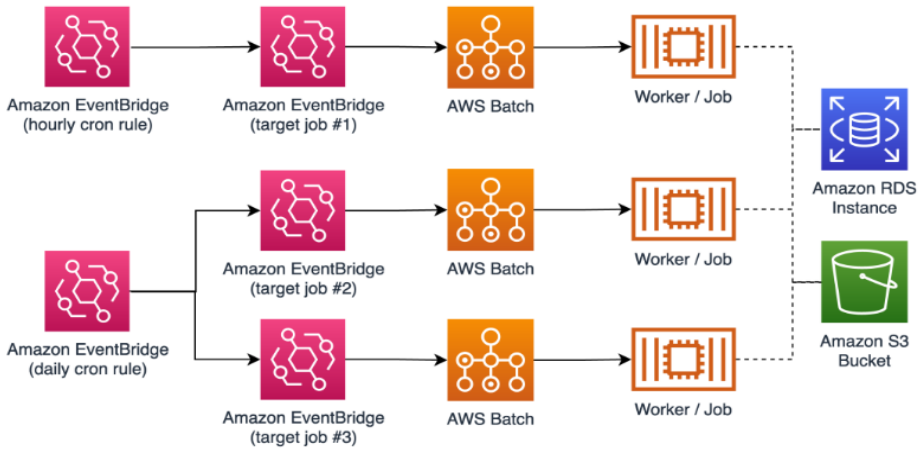

Figure 7: AWS Services for Data Collection (Denot et al., 2021)

This study applies asynchronous processing technology to speed up data collection. To put it simply, if implementing sequential techniques, it is necessary to collect data from An Giang province, then it will be possible to collect comment data from customers in Ba Ria - Vung Tau province. However, when using asynchronous technology, the system will create two workflows to independently query the information of An Giang province and Ba Ria - Vung Tau province. This is a new technology supported by Python since version 3.5 with the Async/Await keyword.

Storage Layer

Research uses Amazon S3 service to store data after collecting from booking platforms and websites in the field of travel. Data is stored in two main forms: (1) Semi-structured data in the form of JavaScript Object Notation (JSON) files to store hotel information, customer feedback on booking platforms, and information exchange, comment on tourism on forums; (2) Pictures of the hotels, about the comments. Data after being stored is considered a Data Lake.

Analysis Layer

In this part, the study will conduct a test of topic model Latent Dirichlet Allocation (LDA) combined with PySpark technology. The study extracts a small amount of data from the Data Lake to conduct experiments.

Data from the Booking and Agoda platforms has been pre-labelled as positive comments or negative comments. In the experimental range, the study extracted 45,793 negative Vietnamese comments to examine the latent topics that customers are complaining about regarding the services and products of hotels. The LDA topic model was introduced in 2003 by Blei et al. Currently, the LDA model is still considered as state-of-the-art in the field of identifying latent topics. There have been many studies using LDA model in many fields such as E-Commerce (Li et al., 2022; Santosh et al., 2016), Tourism (Huang et al., 2018), Education (Ho & Do, 2018). However, previous studies only used single-threaded processing technology, not focusing on multi-threaded processing technology such as gensim (Řehůřek & Sojka, 2010) and scikit-learn (Pedregosa et al., 2011) libraries. Therefore, the PySpark technology approach to support Big Data analysis is also a new point in this study when applied to the LDA model.

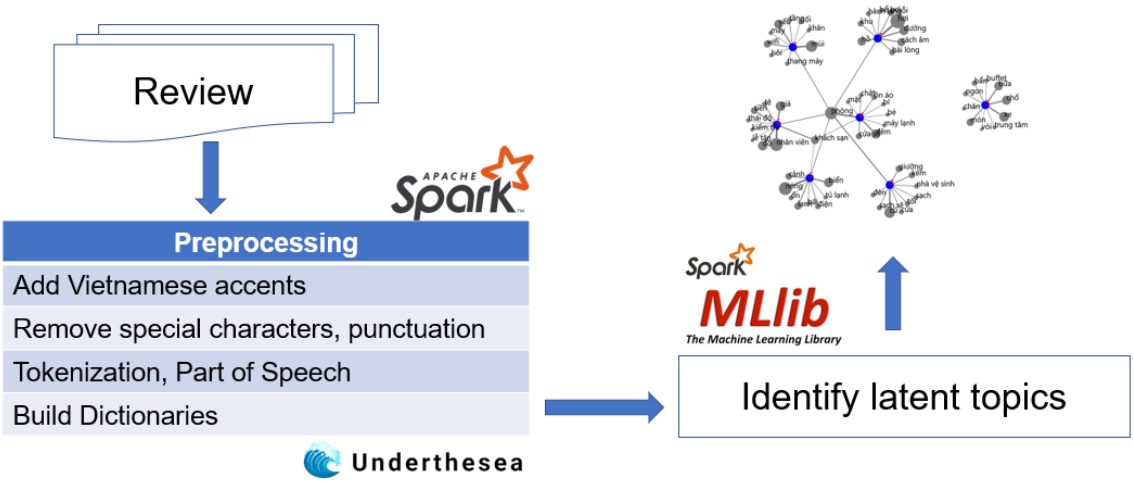

Figure 8: The process of implementing the Topic Modelling

In general, the process of implementing the topic model will be the same (Figure 8), regardless of the technology. The steps taken are still data preprocessing, determining the optimal number of threads. Research uses GitHub (n.d.), a library that combines PySpark techniques for data preprocessing. Then, the study proceeds to identify latent topics using the LDA model. The algorithm of the LDA model is available in Spark MLlib.

EXPERIMENTAL RESULTS AND DISCUSSION

DATA LAKE FOR BIG DATA

In this section, the study presents structured information about the data that has been collected. The study selected a data sample from the Booking platform for illustration. The information structure consists of two parts: information about the hotel and the customer feedback for each hotel. This study is focusing on Vietnamese data, which may be confusing for some readers. Therefore, the study presents translations from Vietnamese into English of some essential information.

Figure 9: Sample of hotel information collected from the Booking.com platform

The Data Lake contains hotel data with information such as hotel name, hotel address, total current reviews, and photos. With the photos, the research has extracted the information storage structure, the Booking platform uses the domain "https://cf.bstatic.com" as the root Uniform Resource Locator (URL). Then study stores the relative path of each image. From there, by combining the root URL and the corresponding relative path, the researcher can easily view the image and download it for storage (Figure 9). The study archived in two forms: (1) URL access to the image; (2) Images are downloaded and stored in Amazon S3 Bucket.


```
{
    "id": 2595165,
    "reviewer_name": "Hope",
    "reviewer_country": "Việt Nam",
    "stay_length": "1 đêm",
    "room_type": "Căn Hộ Superior",
    "rental_date": "Tháng 6-2022",
    "review_date": "ngày 24 Tháng 6 năm 2022",
    "Travel_type": "Cặp đôi",
    "review_overall": "Great Staff",
    "review_score": " 10 ",
    "like": "Phòng sạch sẽ, nhân viên lễ tân nhiệt tình",
    "dislike": null,
    "language": "vi",
    "images_urls": null
},
```


Figure 10: Sample of customer feedback information collected from Booking.com platform

Figure 10 is a sample of customer feedback for a hotel that the study tested, with information collected such as reviewer, date of rental, date of feedback, room type, positive comments, average rating, and negative comments. Data is stored by hotel, each response is an object belonging to the hotel's response collection, stored in JSON structure.

The data from Figure 10 is a Vietnamese response. Therefore, the prospective study is presented in tabular form with an English translated version for the reader's convenience (Table 2)

Table 2: Description of the data that have been collected and stored from the experimental process, including original version and English version

PROPERTY	ORIGINAL VERSION	TRANSLATED VERSION (ENGLISH)
id	2595165	2595165
reviewer_name	Hope	Hope
reviewer_country	Việt Nam	Vietnam
stay_length	1 đêm	1 night
room_type	Căn Hộ Superior	Superior Apartment
rental_date	Tháng 6-2022	June 2022
review_date	ngày 24 Tháng 6 năm 2022	June 24, 2022
travel_type	Cặp đôi	Couple
review_overall	Great Staff	Great Staff
review_score	10	10
like	Phòng sạch sẽ, nhân viên lễ tân nhiệt tình	Clean room, enthusiastic front desk staff
dislike	null	null
langugage	vi	vi
images_urls	null	null

Data collected in the form of HTML code analysis is often referred to as text data of the string data type. However, in essence, the review_score attribute has a numeric data type of decimal (double/float). Meanwhile, rental_date and review_date properties have data type of date. During the analysis, the research will consider the data rules to convert the appropriate data type. The study only pays attention to this point so that researchers avoid errors in the processing, technically, the data type conversion of these cases is not too difficult.

BIG DATA COLLECTION AND 5V MODEL

Volume

Table 3 shows that, in the Vietnam market alone, the research has collected more than 1 million customer feedbacks from two platforms Agoda (763,911 comments) and Booking (409,092 comments). In addition, the study also collected 190,872 exchanges from Phuot.vn website. The study considers this data set to be quite large.

Velocity

With many customers actively interacting, the three sites above generate a large amount of data and increase rapidly over time. Considering the data of the Booking platform from Table 3, within only about 40 months (6/2019 – 8/2022), in the Vietnamese market alone, the Booking platform has generated more than 400 thousand responses, an average of about 10 thousand responses per month. The study considers this to be a fast data generation rate.

Table 3: Description of the data that have been collected and stored from the experimental process

CRITERIA	AGODA.COM	BOOKING.COM	PHUOT.VN
Time	7/2007 – 8/2022	6/2019 – 8/2022	5/2007 – 8/2022
Number of hotels (Number of articles)	6,362	12,477	5,216
Number of comments	763,911	409,092	190,872
Number of languages	36	45	1
Number of comments in Vietnamese/English/Uncategorized*	93,942 / 433,600 / 4,039	89,925 / 55,640 / 191,115	190,872 / 0 / 0

** The Agoda and Booking platforms allow customers to specify the language to use to respond. The study considers this attribute to group languages. For uncategorized comments, the study will use the spacy-langdetect library to determine the language type based on customer comments.*

Variety

The study collected, processed, and stored data in two forms: (1) Semi-structured data (Hotel information, comment information) and (2) Unstructured Data (Image). The study considers that the currently collected data set is diverse.

Veracity

Agoda and Booking are two of the most popular booking platforms in the world. Meanwhile, Phuot.vn is a tourism information exchange forum used by a large number of young Vietnamese people. Therefore, the study that considers data from these sources is reliable.

Value

The value of customer feedback has been covered by many studies (Huang et al., 2018; H. Le et al., 2017; Rajendran et al., 2023). The present study also derives a lot of value from the large amount of customer comment data collected, especially comments and exchanges in Vietnamese. This amount of data will support the Vietnamese research community in the field of natural language processing. As a result, the study considers this data set to be of great value.

Topic Modelling

Figure 11 depicts the process of implementing the LDA model using PySpark. The first step is to initialize SparkContext, and, then, load the data from the Data Lake hosted in AWS S3. Next, the study proceeds to pre-process the data with steps such as adding Vietnamese accents, removing special characters, separating words, and building a dictionary. After that, the study carried out vectorization using IF-IDF technique. Next, the research was conducted using the library "pyspark.ml.clustering" with the LDA algorithm, the input parameter is the number of K topics, the experimental study was from K = 2 to K = 15. For each test, research extracted the "Perplexity" score. When running completed 14 models with K of different topics, the study reselected the good model as the model with the lowest "Perplexity" score. The study reruns the algorithm with the best K-index and displays the results as latent topics.

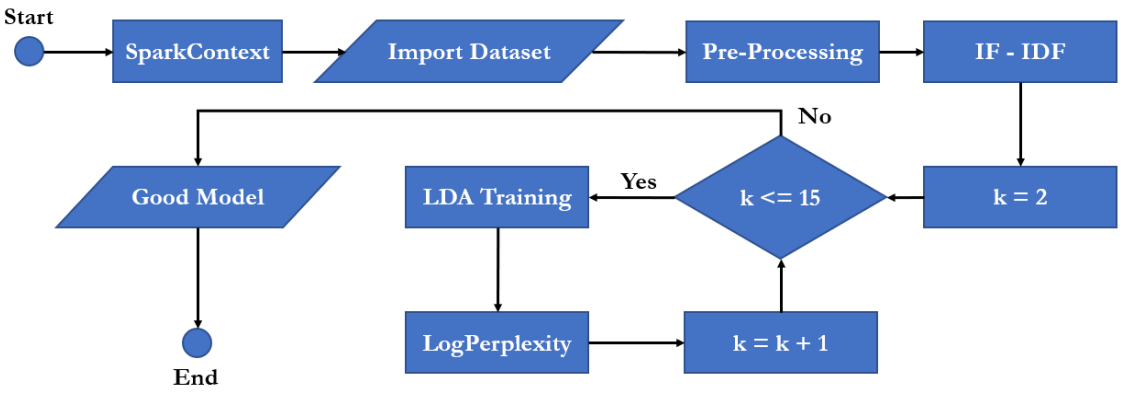

Figure 11: The process of implementing the LDA model using PySpark

In Table 4, the study presents the parameters in the LDA model using PySpark that the study used. In general, the study reuses the baseline model with the default value, the study only adjusts the number of topics K and records the Perplexity value to determine the good model.

Table 4: The main parameters used for the LDA model in PySpark (Apache Spark, n.d.)

ATTRIBUTES	VALUE	NOTE
checkpointInterval	10	Set checkpoint interval (>= 1) or disable checkpoint (-1). Example: 10 means that the cache will get checkpointed every 10 iterations. Note: this setting will be ignored if the checkpoint directory is not set in the SparkContext
docConcentration	None	Concentration parameter (commonly named "alpha") for the prior placed on documents distributions over topics ("theta").
featuresCol		LDA is given a collection of documents as input data, via the featuresCol parameter. Each document is specified as a Vector of length vocabSize, where each entry is the count for the corresponding term (word) in the document. Feature transformers such as pyspark.ml.feature.Tokenizer and pyspark.ml.feature.CountVectorizer can be useful for converting text to word count vectors.
k	[2, 15]	The number of topics (clusters) to infer. Must be > 1
keepLastCheckpoint	True	(For EM optimizer) If using check pointing, this indicates whether to keep the last checkpoint. If false, then the checkpoint will be deleted. Deleting the checkpoint can cause failures if a data partition is lost, so set this bit with care

ATTRIBUTES	VALUE	NOTE
learningDecay	0.51	Learning rate, set as an exponential decay rate. This should be between (0.5, 1.0] to guarantee asymptotic convergence
learningOffset	1024.0	A (positive) learning parameter that down-weights early iterations. Larger values make early iterations count less
maxIter	10	Max number of iterations (>= 0)
optimizeDocConcentration	True	Indicates whether the docConcentration (Dirichlet parameter for document-topic distribution) will be optimized during training
optimizer	Online	Optimizer or inference algorithm used to estimate the LDA model. Supported: online, em
seed	None	Random seed
subsamplingRate	0.05	Fraction of the corpus to be sampled and used in each iteration of mini-batch gradient descent, in range (0, 1]
topicConcentration	None	Concentration parameter (commonly named "beta" or "eta") for the prior placed on topic distributions over terms
topicDistributionCol	topicDistribution	Output column with estimates of the topic mixture distribution for each document (often called "theta" in the literature). Returns a vector of zeros for an empty document

In the next section, the study presents the results of the LDA model using PySpark.

The LDA model uses the Perplexity score to evaluate the optimal number of topics of the experimental data set (Blei et al., 2003). Figure 12 depicts the Perplexity score according to the number of topics from 2 to 15. The results show that from a small dataset of negative responses, the LDA model identifies 7 optimal latent topics with a Perplexity index of 5.1456. The detailed results of the LDA model are presented in Tables 5, 6 and 7.

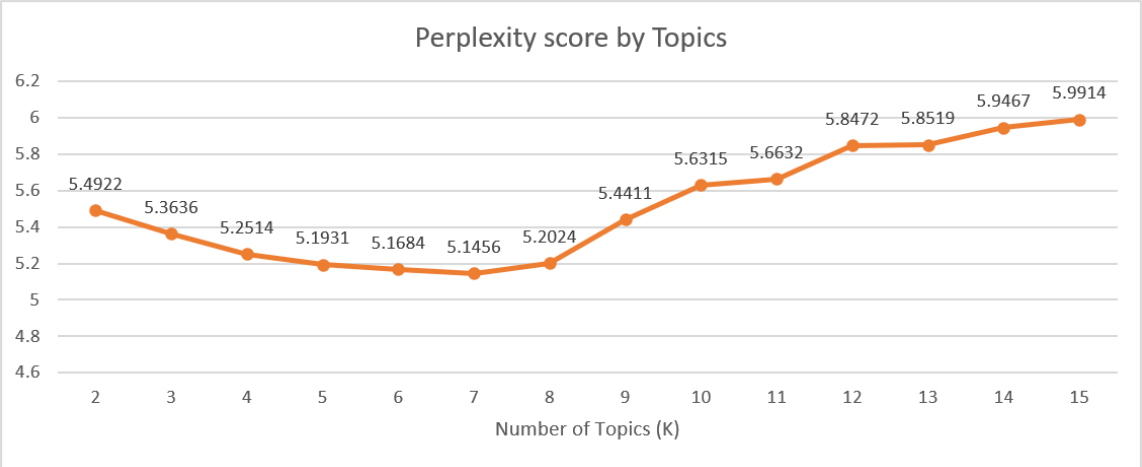

Figure 12: Perplexity score by Topics of a small negative dataset (Source: Authors)

Table 5: Top 10 keywords with the biggest weight on topics 1-3

TOPIC 1 (FACILITIES)		**TOPIC 2 (FOOD)**		**TOPIC 3 (FACILITIES)**	
Keyword	***Weight***	***Keyword***	***Weight***	***Keyword***	***Weight***
mùi (smell)	0.0461	xe (car)	0.0399	cũ (old)	0.0433
yếu (weak)	0.0362	chỗ (place)	0.0373	phòng (room)	0.0338
wifi	0.0290	bữa (meal)	0.0357	giường (bed)	0.0270
phòng (room)	0.0248	món (dish)	0.0302	kém (bad)	0.0253
tầng (floor)	0.0220	bẩn (dirty)	0.0221	tối (dark)	0.0223
thang máy (elevator)	0.0186	ngon (tasty)	0.0207	sạch sẽ (clean)	0.0220
hôi (foul)	0.0186	trung tâm (center)	0.0174	nhà vệ sinh (toilet)	0.0206
máy (machine)	0.0177	buffet	0.0168	đèn (light)	0.0198
gối (pillow)	0.0176	chăn (blanket)	0.0157	cửa (window)	0.0196
khăn (towel)	0.0173	vòi (water tap)	0.0140	sạch (clean)	0.0194

Table 6: Top 10 keywords with the biggest weight on topics 4-5

TOPIC 4 (FACILITIES)		TOPIC 5 (LOCATION)	
Keyword	*Weight*	*Keyword*	*Weight*
phòng (room)	0.0504	nóng (hot)	0.0527
đêm (night)	0.0380	biển (sea)	0.0383
cửa sổ (window)	0.0249	lạnh (cold)	0.0291
ồn ào (noisy)	0.0212	cảnh (scene)	0.0274
chật (tight)	0.0182	điện (electricity)	0.0223
khách sạn (hotel)	0.0177	phòng (room)	0.0223
bí (cramped)	0.0176	bãi (space)	0.0192
máy lạnh (air conditioner)	0.0173	ổn (fine)	0.0186
mặt (face)	0.0171	tủ lạnh (refrigerator)	0.0177
bé (small)	0.0166	khách sạn (hotel)	0.0155

Table 7: Top 10 keywords with the biggest weight on topics 6-7

TOPIC 6 (PRICE – EMPLOYEE)		TOPIC 7 (FACILITIES)	
Keyword	*Weight*	*Keyword*	*Weight*
nhân viên (employee)	0.0523	hơi (pretty)	0.0595
giá (price)	0.0395	đường (street)	0.0374
đồ (item)	0.0384	hồ (lake)	0.0346
khách sạn (hotel)	0.0323	cách âm (soundproof)	0.0324
tiền (money)	0.0281	muỗi (mosquito)	0.0265
phòng (room)	0.0244	phòng (room)	0.0253
tệ (bad)	0.0223	hài lòng (satisfied)	0.0246
thái độ (attitude)	0.0181	khu (area)	0.0238
kiểm tra (check)	0.0155	hành lang (lobby)	0.0208
lễ tân (receptionist)	0.0153	bể bơi (pool)	0.0190

Based on experimental results, topic 1 has keywords like "smell", "wifi", "room", "floor", "elevator", "machine", "pillow", and "towel", so study marked topic1's label is "Facilities". Topic 6 has keywords like "price" and "money", so research marked the topic related to "Price". In addition, keywords like "employee", "bad", "attitude", and "receptionist" are marked related to the topic "Employee". Summarizing the results, the study found that the majority of users are complaining about the "Facilities". In addition, customers also have negative discussions about the services such as "Food", "Location", "Price", and "Employee". Based on this result, the administrator will see what topics and keywords customers are complaining about. From there, managers have plans and strategies to improve service quality to attract customers.

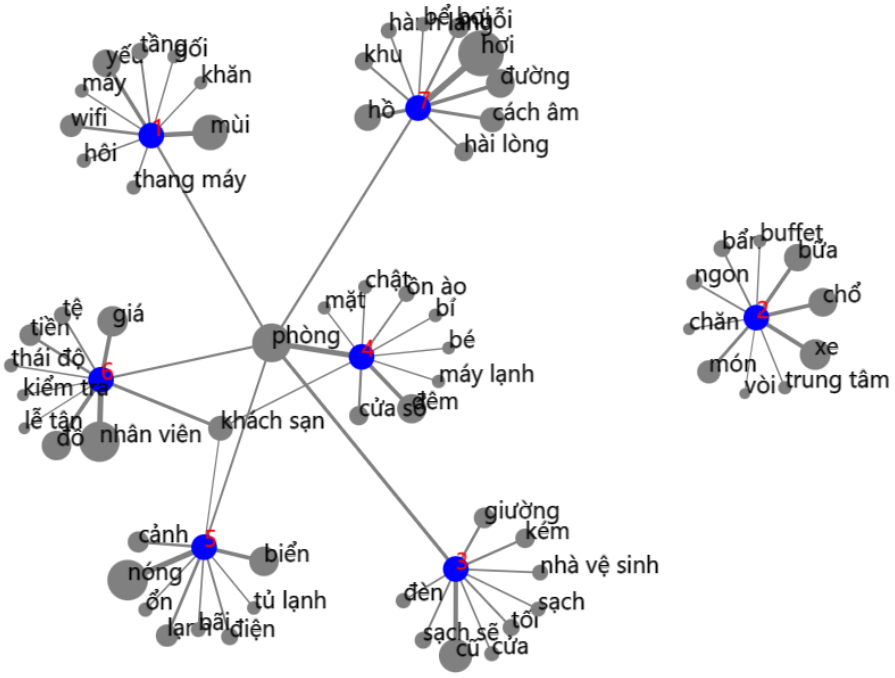

Figure 13: The graph network model considers the relationship between topics and keywords

The study also built a graph network to visualize the relationship between topics and keywords to identify overlapping keywords. This is also an approach for administrators to capture information more easily.

Based on the results from Figure 13, the study found that only 2 keywords "khách sạn - hotel" and "phòng – room" were common keywords. It is easy to see that these 2 keywords are the two main keywords in the field of tourism – hotel. Therefore, the study considers, this clustering result is good.

The study also visualized a graph network model (Figure 14) version based on keywords translated from Vietnamese to English for readers' convenience.

In addition to common keywords like "hotel" and "room". In the English version, there are two more nodes in common, "window" and "bad". This illustrates the difference when translating meanings between languages. In Vietnamese, "tệ" and "kém" are considered two different shades of meaning. However, when interpreted in English, they all mean "bad", so between topic 3 and topic 6 there is a common keyword "bad". Likewise, topic 3 and topic 4 share the same keyword as "window". Readers can see the detailed translation from Tables 5 to 7 to see more clearly Vietnamese keywords and English translation.

This is also a new way of looking at things. When translating the meaning of a word into another language, the original view will change, from there, the analysis aspects will be more diverse.

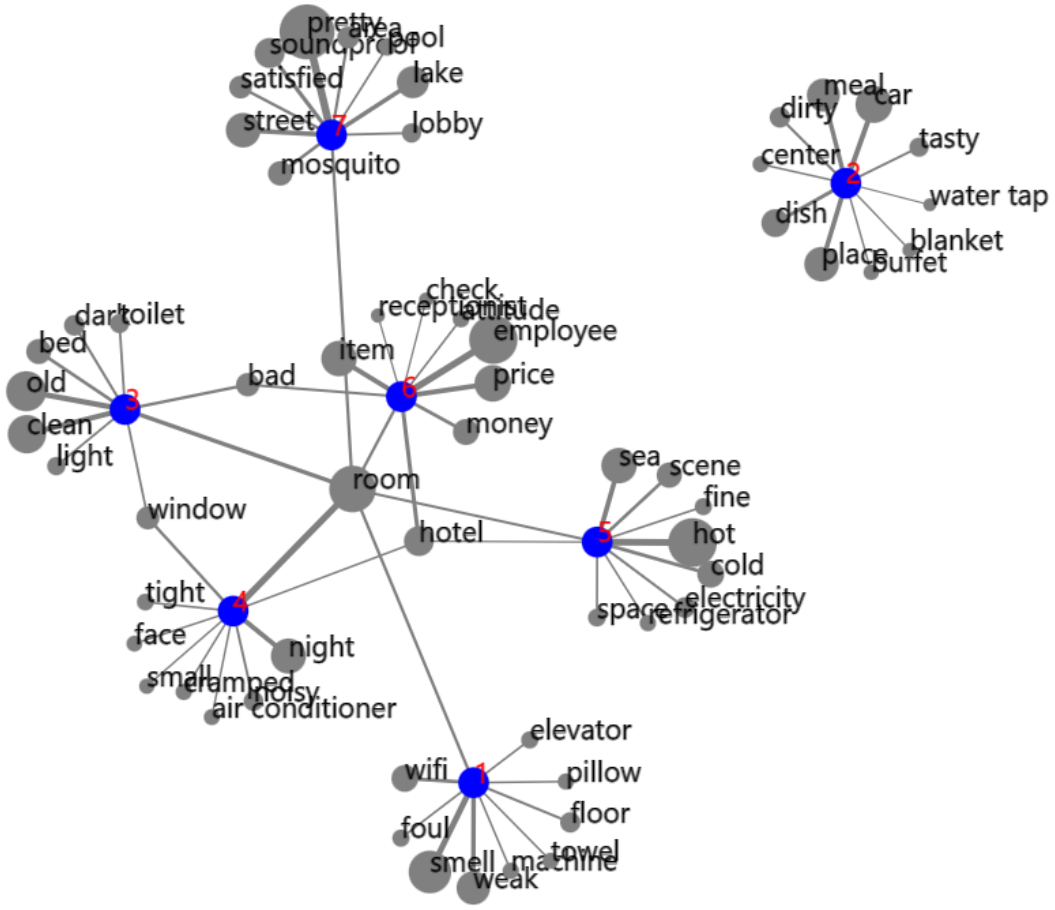

Figure 14: The graph network model considers the relationship between topics and keywords
(translated English version)

Discussion

The Data Lake is capable of storing structured, unstructured, and semi-structured data of any size. In particular, it can also store data in its original format without being too strict, not limiting the amount of space, records, or files. This allows users to use a variety of data formats, while increasing analytical capabilities across platforms. Building a Data Lake gives administrators more flexibility in data analysis. This is also a step forward compared to traditional research when building a Data Warehouse in which only structured data storage is allowed. This study has proven what the collection and experimental model that the proposed study can do, and this model can be applied to different domains.

The study also tested the Topic Model analysis based on PySpark Big Data processing technology. As a result, from the model, administrators can use the LDA model integrated with Big Data technology to analyze topics that customers are interested in. From the understanding of customer complaints, businesses/administrators will continue to develop strategies to improve product/service quality in order to retain customers and build loyal customers.

Conclusion and Future Works

This study focuses on building a Big Data collection and storage model. A set of data has been collected and stored by the study. In general, the current data set only partially meets the standards of Big Data. However, the study discussed and analyzed it to evaluate it as a large data set, which is the

result of the model proposed by the study, thereby proving that the current model and solutions are effective and can be applied in practice. In summary, the study has met the four initial objectives: (1) building a model to collect and store Big Data; (2) experimenting with solutions on a number of booking platforms and travel forums using cloud computing technology; (3) building a large data warehouse storing customer feedback and exchanges in the field of tourism, supporting the community of natural language researchers; (4) experimental LDA model based on Big Data technology. However, the study also encountered some limitations, with each different platform or website the study had to build a query scenario and choose a different technology approach, which limits the ability to the solution's scalability to multiple platforms. Research will continue to build and standardize query scenarios and processing technologies to make scalability to other platforms easier.

ACKNOWLEDGMENTS

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number DS2022-34-01.

REFERENCES

Alcántara-Pilar, J. M., del Barrio-García, S., Crespo-Almendros, E., & Porcu, L. (2017). Toward an understanding of online information processing in e-tourism: Does national culture matter? *Journal of Travel & Tourism Marketing, 34*(8), 1128–1142. https://doi.org/10.1080/10548408.2017.1326363

Amazon Web Services. (n.d.). *Data warehouse vs. Data lake vs. Data mart – Comparing cloud storage solutions – AWS*. https://aws.amazon.com/compare/the-difference-between-a-data-warehouse-data-lake-and-data-mart/

Apache Spark. (n.d.). *LDA – PySpark 3.3.2 documentation*. https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.clustering.LDA.html#pyspark.ml.clustering.LDA.checkpointInterval

Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics. *Proceedings of the 11th Annual Conference on Innovative Data Systems Research (CIDR '21)*, *volume 8* (Article 29). Chaminade, USA. CIDR. https://15721.courses.cs.cmu.edu/spring2023/papers/02-modern/armbrust-cidr21.pdf

Azeroual, O., & Theel, H. (2018, March). The effects of using business intelligence systems on an excellence management and decision-making process by start-up companies: A case study. *International Journal of Management Science and Business Administration*, *4*(3), 30–40. https://doi.org/10.18775/ijmsba.1849-5664-5419.2014.43.1004

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bracci, F., Corradi, A., & Foschini, L. (2012, July). Database security management for healthcare SaaS in the Amazon AWS Cloud. *Proceedings of the 2012 IEEE Symposium on Computers and Communications (ISCC)* (pp. 812-819). Cappadocia, Turkey. IEEE. https://doi.org/10.1109/ISCC.2012.6249401

Ciampi, F., Demi, S., Magrini, A., Marzi, G., & Papa, A. (2021, February). Exploring the impact of big data analytics capabilities on business model innovation: The mediating role of entrepreneurial orientation. *Journal of Business Research*, *123*, 1–13. https://doi.org/10.1016/J.JBUSRES.2020.09.023

Clissa, L. (2022, February 20). Survey of big data sizes in 2021. *ArXiv:2202.07659v2*. https://doi.org/10.48550/arXiv.2202.07659

Demchenko, Y., de Laat, C., & Membrey, P. (2014, May). Defining architecture components of the big data ecosystem. *Proceedings of the 2014 International Conference on Collaboration Technologies and Systems (CTS 2014)* (pp. 104–112). Minneapolis, MN, USA. IEEE. https://doi.org/10.1109/CTS.2014.6867550

Denot, A., Attanayake, S., & Dray, J. (2021, August 18). Scaling Laravel jobs with AWS Batch and Amazon EventBridge. *AWS Partner Network (APN) Blog*. https://aws.amazon.com/blogs/apn/scaling-laravel-jobs-with-aws-batch-and-amazon-eventbridge/

Doreswamy, Gad, I., & Manjunatha, B. R. (2017, April). Hybrid data warehouse model for climate big data analysis. *Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies (ICCPCT 2017)*. Kollam, India. IEEE. https://doi.org/10.1109/ICCPCT.2017.8074229

Drabas, T., & Lee, D. (2017). *Learning PySpark*. Packt Publishing Ltd.

GitHub. (n.d.). *apache/spark: Apache Spark – A unified analytics engine for large-scale data processing*. https://github.com/apache/spark

Hamoud, A., & Obaid, T. (2013). Building data warehouse for diseases registry: First step for clinical data warehouse. *International Journal of Scientific & Engineering Research*, *4*(7), 636-640. https://doi.org/10.2139/ssrn.3061599

Han, Y. (2015). Cloud storage for digital preservation: Optimal uses of Amazon S3 and Glacier. *Library Hi Tech*, *33*(2), 261–271. https://doi.org/10.1108/LHT-12-2014-0118

Ho, T., & Do, P. (2018). Social network analysis based on topic model with temporal factor. *International Journal of Knowledge and Systems Science*, *9*(1), 82–97. https://doi.org/10.4018/IJKSS.2018010105

Huang, C., Wang, Q., Yang, D., & Xu, F. (2018, March 14). Topic mining of tourist attractions based on a seasonal context aware LDA model. *Intelligent Data Analysis*, *22*(2), 383–405. https://doi.org/10.3233/IDA-173364

Jackson, S. (2016). Prediction, explanation and big(ger) data: A middle way to measuring and modelling the perceived success of a volunteer tourism sustainability campaign based on 'nudging.' *Current Issues in Tourism*, *19*(7), 643–658. https://doi.org/10.1080/13683500.2014.898616

Khan, N., Naim, A., Hussain, M. R., Naveed, Q. N., Ahmad, N., & Qamar, S. (2019, May). The 51 V's of big data: Survey, technologies, characteristics, opportunities, issues and challenges. *Proceedings of the International Conference on Omni-Layer Intelligent Systems* (pp. 19–24). Crete, Greece. ACM. https://doi.org/10.1145/3312614.3312623

Laney, D. (2001). *3D data management: Controlling data volume, velocity, and variety*. META Group. https://studylib.net/doc/8647594/3d-data-management--controlling-data-volume--velocity--an...

Le, H. S., Lee, J. H., & Lee, H. K. (2017, April). Analyzing visitors' preferences on tourism accommodation services by opinion mining. *The Journal of Internet Electronic Commerce Research*, *17*(2), 111–127. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07160326

Le, T., Nguyen, V. H., & Ho, T. (2022, October). A model of discovering customer insights in tourism sector approach to Vietnamese reviews analytics. *Proceedings of the 2022 9th NAFOSTED Conference on Information and Computer Science (NICS 2022)* (pp 205–210). Ho Chi Minh City, Vietnam: IEEE. https://doi.org/10.1109/NICS56915.2022.10013410

Li, L., Fu, L., & Zhang, W. (2022). Impact of text diversity on review helpfulness: A topic modeling approach. *Interdisciplinary Journal of Information, Knowledge, and Management*, *17*, 87–100. https://doi.org/10.28945/4922

Lin, X., Liu, K., & Li, Y. (2021). BI warehousing system based on big data. *E3S Web of Conferences*, *257*, 02015. https://doi.org/10.1051/E3SCONF/202125702015

Monino, J. L. (2021). Data value, big data analytics, and decision-making. *Journal of the Knowledge Economy*, *12*, 256–267. https://doi.org/10.1007/s13132-016-0396-2

Mukherjee, S. (2019, March 07). Benefits of AWS in modern cloud. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3415956

Nguyen, M. N. (2022, February 15). *Leading online travel agencies in Vietnam as of November 2020*. https://www.statista.com/statistics/1201342/vietnam-most-used-online-travel-agencies/

Nguyen, T. T. (2022, May 13). Solutions to recover the human resources for Vietnam's tourism industry after the COVID-19 pandemic. *Industry and Trade Magazine*. https://tapchicongthuong.vn/bai-viet/giai-phap-phuc-hoi-nguon-nhan-luc-nganh-du-lich-viet-nam-sau-dai-dich-covid-19-88642.htm

Nguyen, V. H., & Ho, T. (2023). Analyzing online customer experience in hotel sector using dynamic topic modelling and net promoter score. *Journal of Hospitality and Tourism Technology*, *14*(2), 258–277. https://doi.org/10.1108/JHTT-04-2021-0116

Nhandan. (2022, February 10). WTTC: *Du lịch và Lữ hành có thể đóng góp 8,6 nghìn tỷ USD [WTTC: Travel and tourism could contribute $8.6 trillion]*. Ministry of Culture, Sports, and Tourism. https://bvhttdl.gov.vn/wttc-du-lich-va-lu-hanh-co-the-dong-gop-86-nghin-ty-usd-20220210090158147.htm

Olmedilla, M., Martínez-Torres, M. R., & Toral, S. L. (2016). Harvesting big data in social science: A methodological approach for collecting online user-generated content. *Computer Standards & Interfaces*, *46*, 79–87. https://doi.org/10.1016/J.CSI.2016.02.003

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Rajendran, S., Srinivas, S., & Pagel, E. (2023). Mining voice of customers and employees in insurance companies from online reviews: A text analytics approach. *Benchmarking: An International Journal*, *30*(1), 1–22. https://doi.org/10.1108/BIJ-12-2020-0650

Řehůřek, R., Sojka P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (pp. 45-50). Valletta, Malta. https://doi.org/10.13140/2.1.2393.1847

Rimmalapudi. (2023, February 14). *Spark internal execution plan*. https://sparkbyexamples.com/spark/spark-execution-plan/

Santosh, D. T., Babu, K. S., Prasad, S. D. V., & Vivekananda, A. (2016). Opinion mining of online product reviews from traditional LDA topic clusters using Feature Ontology Tree and Sentiword-net. *International Journal of Education and Management Engineering*, *6*(6), 34–44. https://doi.org/10.5815/ijeme.2016.06.04

Thien, L., Phe, B., & Thanh, H. (2021). Cloud-based business intelligence solution: A case study on financial dataset. *Journal of Asian Business and Economic Studies*, *32*, 83–101. http://jabes.ueh.edu.vn/Home/SearchArticle?article_Id=e389985f-36a8-4757-8809-48fb1a8bd3e0

Wah, T. Y., Peng, N. H., & Hok, C. S. (2007, November). Building data warehouse. *Proceedings of the 24th South East Asia Regional Computer Confederation Conference*, *volume 15* (pp. 51-56). Bangkok, Thailand: South East Asian Regional Computer Confederation (SEARCC). https://static.aminer.org/pdf/PDF/000/307/364/acquiring_and_integrating_external_data_into_data_warehouses_are_you.pdf

Wang, J., Xu, C., Zhang, J., & Zhong, R. (2022). Big data analytics for intelligent manufacturing systems: A review. *Journal of Manufacturing Systems*, *62*, 738–752. https://doi.org/10.1016/J.JMSY.2021.03.005

AUTHORS

Thanh Ho (Ph.D.) received M.S. degree in Information Technology from University of Information Technology, VNU–HCM, Vietnam in 2009 and Ph.D. degree in Information Technology from University of Information Technology, VNU-HCM, Vietnam in 2018. He is currently an Associate Professor, Senior Lecturer in Faculty of Information Systems, University of Economics and Law, VNU–HCM, Vietnam. His research interests are data mining, data analytics, business intelligence, social network analysis, machine learning and big data. The author can be contacted at: thanhht@uel.edu.vn

Van-Ho Nguyen (M.A.) received the B.S. degree in Management Information System (MIS) from the Faculty of Information Systems, University of Economics and Law (VNU–HCM), Vietnam, in 2015, and the Master's degree in MIS from the University of Economics Ho Chi Minh City, Vietnam, in 2020. He is currently a Lecturer at the Faculty of Information Systems, VNU-HCM. His current research interests include business analytics, business intelligence, data analytics, and machine learning. The author can be contacted at: honv@uel.edu.vn

Thien Le (M.Sc.) received B.S degree in Management Information System from Faculty of Information Systems, University of Economics and Law, VNU-HCM, Vietnam, in 2019, and the Master's degree Information Systems from the University of Information Technology Ho Chi Minh City, VNU-HCM, Vietnam, in 2023. His current research interests include business intelligence, data analytics, and digital transformation. The author can be contacted at: thienlb.ktl@uel.edu.vn

Hoanh-Su Le (Ph.D.) received the B.E in Electronics and Telecommunication, MSc in MIS and MBA degrees from Vietnam National University HCM City in 2009 and 2011. He received Ph.D. degree in E-business from Pukyong National University, Republic of Korea in 2016. Since 2011 he has been a faculty member and currently Dean (Tenure) of Faculty of Information Systems at University of Economics and Law, Vietnam National University Ho Chi Minh City (VNU HCM). His research interests are data analytics, artificial intelligence and fintech. The author can be contacted at: sulh@uel.edu.vn

Thon-Da Nguyen (Ph.D.) is a Lecturer and Researcher in Information Systems at the University of Economics and Law, VNUHCM, Vietnam. He received a PhD in Information Systems at the Posts and Telecommunications Institute of Technology (2021). Before that, in 2013, he earned a master's degree in Computer Science from the University of Information Technology, VNUHCM, Vietnam. His research interests include data mining, pattern mining, sequence analysis and prediction, text mining, big data analytics, and recommender systems. The author can be contacted at: dant@uel.edu.vn

Thi Cam-Tu Mai (Ph.D.) received a Ph.D. degree in Economics in University of Economics and Law, Vietnam National University Ho Chi Minh, Vietnam in 2016. She is currently a Lecturer in the Faculty of International Economics relations, in Economics in University of Economics and Law, Vietnam National University Ho Chi Minh City, VNU-HCM. Her research interests are corporate social responsibility, international business, customer analytics, and foreign direct investment. The author can be contacted at tumtc@uel.edu.vn

Thi-Anh Tran (Ph.D. student) is a Lecturer and Researcher in Faculty of Information Systems at the University of Economics and Law, VNUHCM, VietNam in 2015. She received a Master's degree in Management information systems from the Ho Chi Minh City University of Technology, VNUHCM, Vietnam. Now, she is a Ph.D student at University of Economics and Law, VNUHCM, Vietnam. Her research interests include text mining, big data analytics, customer experience in e-commerce, and data-driven CRM. The author can be contacted at: anhtt@uel.edu.vn

Hoai-Phan Truong (M.Sc.) works as a Full Lecturer and Researcher in Faculty of Information Systems at the University of Economics and Law, VNUHCM, Vietnam. He received a Master's degree in Computer Science from the University of Science, VNUHCM, Vietnam in 2002. His research interests are computer science, AI, knowledge base systems, and machine learning. Besides that he is interested in computer networking. The author can be contacted at hoaiphan@uel.edu.vn