INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT047-3-M-BDAT

BIG DATA ANALYTICS AND TECHNOLOGIES

APDMF2112DSBA(DE)(PR)

APRIL 2022

TITLE: BIG DATA TOOLS PROPOSAL IN TOURISM

LEE KEAN LIM

TP065778

LECTURER: DR. V. SIVAKUMAR

ABSTRACT

Current technologies have provided people with the ease of travelling which spurred the development of the tourism sector. The tourism sector is a rapidly growing industry which can provide high revenue generation for a nation. However, advances of information technologies have brought the tourism sector into an information-intensive business. Data in high volume and speed is generated and consumed by the tourism sector each day. Legacy systems in dealing with such data is insufficient and inefficient. Therefore, the adoption of Big Data platform is highly valued by companies. This study proposes a Big Data ecosystem for a tourism company named Five Star Tourism with the need to handle and process Big Data efficiently and effectively. In addition, the ecosystem would require analytics module to perform data analytics to assist the company in predicting search trends quickly and cost-effectively. The proposed Big Data ecosystem would comprise of 12 Big Data tools that would work together to achieve the needs of Five Star Tourism in Big Data analytics. The tools proposal would be based on existing use cases and be adapted to the current company. The implementation of the proposed Big Data ecosystem would allow Five Star Tourism to better manage Big Data and derive value from the data to gain a competitive advantage in the market.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

HDFS ..... Hadoop Distributed File System

IoT ..... Internet of Things

SQL ..... Structured Query Language

UGC ..... User-generated content

SECTION 1

INTRODUCTION

1.1 INTRODUCTION

The tourism sector is a big business making up about 10% of the world gross domestic product and it is continuously growing and pose a huge future potential in the global wealth and employment. The availability of current information technologies has propelled the tourism sector to generate and consume vast amount of data. Therefore, data generated in the tourism sector is considered as Big Data which exhibit the characteristic of 3Vs, known as Volume, Velocity, and Variety. Volume refers to the size of data generated and consumed, velocity refers to the speed of data transmission, and variety refers to the different types of data available. In addition, Big Data solutions have provided more opportunities from deriving insights from the data. The tourism sector heavily relies on feedback and client interactions to evaluate their business performance and to identify the demands of the customers. In which, Big Data techniques have been widely applied in the sector to achieve various business objectives. Example, Claude (2020) utilized web search data from Amsterdam to understand the behaviors and patterns of tourists based on web searching activities to predict the number of tourist likely to visit to Amsterdam in the future months.

However, before being able to utilize the Big Data solutions, companies are required to set up a Big Data ecosystem to support such capabilities. A popular Big Data ecosystem is the Hadoop ecosystem. The Hadoop ecosystem is a framework used for processing high volume of data very quickly. Example, a multinational hotel group implemented Big Data ecosystem to integrate high volume of corporate data and competitor data to build a real-time dashboard to compare and analyze different hotel prices (biSmart, 2019). In which, these tools have allowed the hotel to quickly reflect on their hotel prices to capture more customers and ultimately increase company profits.

1.2 BACKGROUND

The use of Big Data analytics has provided more opportunities to the tourism sector. As huge volume of data is generated daily and waiting to be explored to derive valuable insights. There are a few Big Data analytics use cases typically adopted in the tourism sector (ActiveWizards, 2019). The following include but not limited to examples of Big Data analytics used in the tourism sector:

- Customer segmentation, which divides customers into different groups according to their behaviors and preferences to provide a tailored service to the specific group.
- Customer sentiment analysis, which analyze text data to capture emotional elements of customers as feedbacks and reviews heavily influence decision making of customers in choosing and planning for the next destination or hospitality.
- Recommendation engines, where travel and booking web platforms provide the most optimal travel packages and offers to customers based on their previous search and preferences.
- Predictive analytics, which mainly provides dynamic pricing and fare forecasting based on seasonal patterns to optimize profit returns for companies.
- Route optimization, which the complexity of trip planning can be partially automated and optimized to minimize travel cost and distance while providing optimum time management.

Big data in the tourism sector is typically divided into three categories (Becha *et al.*, 2020). First category of Big Data is the user-generated content (UGC) data. UGC data is typically generated through sharing of opinions and feedbacks by people in social media networks and forums. Second category of Big Data is device data. This is where devices like Internet of Things (IoT) and sensors continuously generate data from the tracking of movements of tourists. Final category of Big Data is transaction data where this type of data mainly comprises of daily operational and transactional data specifically in the tourism sector. Table 1.1 shows some examples of data in each of the categories mentioned.

Table 1.1: Examples of data sources

Data categories	Examples of data
User-generated data	Reviews, ratings, feedback, complaints, recommendations, photos, videos, audio
Device data	Global positioning system, mobile roaming, Wireless Fidelity, meteorological data
Transaction data	Web search data, food purchase, airline tickets purchase, accommodation bookings, attraction ticket purchase

As the data generated in the current age comes in many varieties, traditional systems such as relational databases and data warehouses are not able to cater for most data due to majority of the current data generated is in unstructured or semi-structured format. In addition, traditional systems use vertical scaling which requires higher processing power, memory, and storage for high volume of data which is inefficient and ineffective. Furthermore, data in traditional systems are stored in silos which can be a challenge to combine them for further analysis. Therefore, the introduction of Big Data technology is here to solve the issue of processing large volume of data in a quick manner. The implementation of Big Data ecosystem is provided to aid the processing of Big Data efficiently and effectively to derive insights from the data to achieve specific goals and ultimately increase profit for companies. One such popular Big Data ecosystem widely adopted is the Hadoop ecosystem. Hadoop was the pioneer in the Big Data scene and provided many real-world benefits to companies enabling them to store and process large volume of data. The framework mainly comprises of four modules that work together collectively to provide ingestion, storage, processing, and management of the data.

The data ingestion framework is the process of sourcing data from external sources to a centralized storage which can be accessed and used for further analysis. Two types of data ingestion are available. First type is batch ingestion where a group of data that is large in volume are gathered and delivered together as a batch. This process can be triggered by conditions or by scheduled basis. Second type is streaming ingestion where a continuous flow of data is going into the storage which is typically used for real-time analytics. This process is constantly locating and pulling data from various sources thus consumes more resources. Table 1.2 shows some of the data ingestion tools available in the market.

Table 1.2: Examples of data ingestion tools

Data ingestion tools			
Apache Flume	Apache Sqoop	Facebook Scribe	Apache Chukwa
Netflix Suro	Apache Samza	Apache NiFi	HIHO
Apache ManifoldCF	Cloudera Morphline	Apache Kafka	Fluentd

There are challenges typically faced in the data ingestion process. First challenge concerns the security and compliance of data where data from external sources may pose security vulnerabilities to the internal system. In addition, validating the ethical and legality of data can be difficult as data comes in at very large volume and at high speed. Second challenge is regards to the speed of data transfer as different data sources utilize different infrastructures. In which, a data source with particularly slow transmission speed as compared to other sources would hinder the entire data transmission process and potentially introduce errors. Third challenge revolves around the quality of data received. Although the data are available in large volume, but not all data is relevant and useful. In which, irrelevant and incorrect data causes analytics to be less reliable and slowing down the analysis processes.

Next is the data storage framework. The typical data storage framework uses the Hadoop Distributed File System (HDFS) which is a distributed file system efficient for storing Big Data and runs on commodity hardware with built in fault-tolerance. In addition, on top of HDFS there is NoSQL databases and NewSQL database to support real-time data processing. Table 1.3 shows some of the data storage tools available which covers distributed file system, NoSQL databases, and NewSQL databases.

Table 1.3: Examples of data storage tools

Distributed File System	**NoSql Databases**		**NewSql Databases**
Apache HDFS	Apache Hbase	RocksDB	VoltDB
Lustre File System	Apache Cassandra	Redis	SenseiDB
XtreemFS	MongoDB	Giraph	SAP HANA
Ceph File System	DynamoDB	TitanDB	BayesDB

There are challenges typically faced in HDFS. First challenge concerns the latency of data access. HDFS was not designed for low latency data transmission. Instead, HDFS was designed to handle large volume of data and batch processing thus having high latency. Second challenge is regards to size of data stored in HDFS. HDFS is not efficient for large amount of small sized data files storing due to the high-capacity design. Third challenge involves the modification of files stored in HDFS which is not supported, as HDFS adopts the write-once-read-many model.

Third framework is the data processing framework. The typical processing component in the Hadoop ecosystem is MapReduce which process large volume of data using distributed and parallel algorithms for structured and unstructured data. In addition, on top of MapReduce there are analytical components that provides further processing on the data to derive insights. Examples of the components include distributed programming, machine learning engines, graph processing engines, etc. Implementation of these components are dependent on the requirements of the companies in specific of what they want to achieve. Therefore, a wide selection of components is available to cater for specific demands. Table 1.4 shows some examples of the data processing tools available.

Table 1.4: Examples of data processing tools

Distributed Programming	**SQL on Hadoop**	**Machine Learning**
Apache Ignite	Apache Mahout	Apache Hive
Apache MapReduce	WEKA	Apache HCatalog
Apache Pig	H2O	Apache Drill

Lastly is the management framework which functions as a coordination, delegation, and synchronization among the resources and components within the ecosystem. This provides a centralized service to easily manage and oversee the clusters at scale. In addition, the security component can be placed under this framework which controls authorization of data usage and user accessibility. Table 1.5 shows some of the management tools available.

Table 1.5: Examples of management tools

Service Programming	Scheduling	Security	System Deployment
Apache Zookeeper	Apache Oozie	Apache Sentry	Apache Ambari
Apache Avro	Apache Falcon	Apache Knox Gateway	Apache Mesos
Apache Karaf	Shedoscope	Apache Ranger	Horthonworks HOYA

1.3 RESEARCH QUESTION

For the purpose of this study, the following questions will be addressed:

1. How does Big Data support decision making in tourism management?
2. What are the challenges of implementing Big Data solution in the tourism sector?
3. What are the tools that will allow tourism sector to derive strategic value from Big Data?

1.4 AIM & OBJECTIVES

1.4.1 Aim

The aim of this study is to propose a Big Data ecosystem to be implemented in the tourism sector to predict hotel and flight search trends in a faster and cheaper manner.

1.4.2 Objectives

The objectives of the study are as followed:

1. To identify the impact of Big Data in decision making specifically in tourism management.
2. To identify the challenges of implementing Big Data solution in the tourism sector.
3. To propose a Big Data ecosystem that enables strategic value to be derived from Big Data of the tourism sector.

SECTION 2

CASE STUDY

This section shows some of the case studies of implementing Big Data ecosystem in companies specifically in the tourism sector. Table 2.1 shows the case studies of companies relating to the tourism sector applying Big Data platform to achieve specific objectives. The table would describe the aim of why the companies are adopting Big Data platforms, the Big Data tools adopted, and the challenges faced by each company when trying to adopt a Big Data ecosystem.

Table 2.1: Case studies of Big Data platform implementation

References	Title	Aim	Big Data Tools		Challenges
Caesar (2022)	Big data ecosystem case study [Air Pacific Travel]	- Reduce time in flight and data search - Maximize revenue generation	- HDFS - MapReduce - YARN - Hive - Pig - HBase - Mahout - Zookeeper	- Oozie - Sqoop - Flume - Ambari - Apache Drill - Apache Spark - Solr & Lucene	- Difficulty in processing raw log data generated by flight searches and hotel searches - Existing infrastructure does not have fault tolerance and bound to system failure
Strickland (2020)	How Wyndham Hotels & Resorts improved their data capabilities to get to know their guests better	- Manage business more effectively - Build better customer relation	- AWS SFTP - Amazon Kinesis - AWS RedShift Spectrum - Amazon CloudWatch - Amazon Elasticsearch - Amazon QuickSight		- Still adopting the legacy platform to ingest data which is slow and prone to errors - Varying formats of data records in different hotels based in different locations

	[Wyndham Hotels & Resorts]				
Olson (2018)	Big data at United Airlines [United Airlines]	- Improving customer experience - Improve employee experience - Revenue generation - Improve operational reliability	- Teradata - Hortonworks platform - Ambari - Hive - NiFi - Apache Apex	- Apache Flink - Hbase - Spark - Apache Ranger - Apache Atlas	- Bookings & flight schedule data in constant motion and require a tool that can quickly refresh the data every 24 hours to identify the changes happened
Spanos (2015)	Making the case for Hadoop in a large enterprise [British Airways]	- Utilizing publicly available competitor data to facilitate decision making - Maximize revenue generation for each flight	- HDFS - YARN - Kerberos - Ranger - Ambari - Falcon - Zookeeper - Oozie - Sqoop	- Flume - Kafka - Pig - Storm - HCatelog - Hive - Hbase - Hue	- Moving data from storage to analytic phase is a challenge as lack of technologies that can be easily adopted by the analyst in their company
Jonathan Seidman (2010)	Using Hadoop and hive to optimize travel search [Orbitz Worldwide]	- Establish a long-term infrastructure to store large amount of data - Provide data accessibility to developers and analysts - Ad-hoc querying of data and rapid deployment for reports and analytics	- HDFS - MapReduce - Hive - R		- Massive volume data generation each day and not being able to capture such data as system is unable to cope with the supply - Existing data infrastructure insufficient to store and process the incoming data

Based on Table 2.1, all the companies are producing vast amount of data and realized the potential of Big Data that can bring more opportunities to their company. The companies are facing a common issue before the adoption of Big Data platform which is the storage and processing of huge volume of data. Legacy systems are unable to handle such volume of data efficiently and effectively thus the companies move on to adopt the Big Data platform. In addition, failure of legacy system in providing fault tolerance has provided negative impact to companies. The common aim by the different companies of adopting Big Data platform is revenue management. In which, the tourism sector is heavily influence by seasonality thus being able to predict the demand would allow resources to be more efficiently allocated to maximize profits. It is observed that different companies used a different set of tools dependent on multitude of factors to cater for specific needs.

SECTION 3

BIG DATA SOLUTION PROPOSAL

3.1 INTRODUCTION

In this section, a Big Data ecosystem will be proposed for Five Star Tourism to handle the storage and processing of Big Data to predict the hotel and flight search trends in a faster and cost-effective manner. Therefore, factors to be considered when proposing such ecosystem includes the type of data involved, the type of analysis to be expected, the types of output to be achieved, and data governance.

3.2 BIG DATA TOOLS PROPOSAL

The following section proposes Big Data tools for each of the components in the Big Data ecosystem.

Distributed File system

Hadoop Distributed File System (HDFS) – According to Emms (2020), a comparison between nine file systems used for Big Data and the survey found that HDFS was the highest rated distributed file system and adopted by many. HDFS is a scalable, distributed, and portable file system that provides high-throughput access to application data. It is highly fault-tolerant and deployable on commodity hardware. In addition, reliable in storing very large files across a large cluster. Data in the tourism sector consist mostly of unstructured format especially UGC data which comprises of videos and photos which may take up a large chunk of the storage memory. Therefore, HDFS is suitable to be implemented in Five Star Tourism to handle the high-volume log data. Example of HDFS applied the tourism sector can be observed in Orbitz Worldwide which implemented HDFS due to legacy system of insufficiency in storing large volume of data (Jonathan Seidman, 2010).

NoSQL Databases

MongoDB – MongoDB is a document-based NoSQL database which can model complex objects. It is useful for the variety of data coming in from different sources. In addition, querying objects

is performed via simple selections and logical expressions which can be easily adapted by developers and analysts. MongoDB implementation in Five Star Tourism would help the company to read and write data from the database in a quick manner and allow users to easily learn and write queries on demand. An example of travel site who uses MongoDB is Expedia. Where MongoDB has allowed Expedia to change its database schema during operation with zero downtime and maintained the customer experience (MongoDB, 2016).

Data ingestion tools

A comparison between Sqoop, Flume, and Kafka is performed by Vohra (2020). In which, all three of the data ingestion tools are suitable to be applied in the current domain as each of them serve a different purpose.

Sqoop – Sqoop mainly handles relational databases which consist of structured data. Data from devices such as IoT and sensors can be transformed into structured format and stored in a relational database thus Sqoop can be utilized to handle such data import and export from HDFS. Sqoop in Five Star Tourism can simplify and increase the efficiency of data import process from a relational database to HDFS. Example, Pasarakonda (2020) imported data stored in MySQL database to HDFS using Sqoop to identify top 10 customers with highest purchasing amount.

Flume – Flume mainly handles log data which consist of unstructured data. As the tourism sector generates a lot of unstructured data, Flume is required to transfer large quantities of data from external source into a centralized data store. Complement to Sqoop, Flume can assist Five Star Tourism to import log data from other sources into the HDFS. Use case of Flume can be seen in the works of Rodrigues and Chiplunkar (2019), where real time twitter feeds are extracted using Flume and stored in HDFS before being processing for sentiment analysis.

Kafka – Kafa is mainly used for building real-time data streaming pipelines. This is particularly useful in the tourism sector as booking or ticketing data and activities are in a constant motion and would require frequent refresh to obtain the latest information such as notifying a delay or cancellation of booking, notifying the availability of slots, and microservices. Implementing Kafka in Five Star Tourism allows the setup of a dashboard to stream incoming data displaying the current

status of work which can assist the company to optimize workflow. As mentioned by Baer (2018), Kafka has been widely used for recommendation engines. However, as newer technologies are coming out, better alternatives to Kafka are preferred such as MapR Streams and Amazon Kinesis Firehose.

SQL on Hadoop

Hive – It is a data warehouse that supports the use of SQL to query data and perform analysis on large volume of data in HDFS. By using SQL queries, a wider group of developers and analysts can work with the ecosystem on commodity computing clusters as SQL is easy to learn and adopt. By implementing Hive in Five Star Tourism, time consumption for querying and processing log data can be reduced which lead to a faster hotel and flight search. Aman (2015) performed flight data analysis using Hive and mentioned that the SQL like language for querying is easily adaptable and able to reduce time spent on writing programs.

Distributed Programming

Spark – Spark is a framework for writing fast and distributed programs. It provides in-memory computing for the ecosystem, which is faster than MapReduce that uses disk execution. Typically running between 10 to 100 times faster than MapReduce. Implementing Spark in Five Star Tourism would allow analysis computation to be much faster and achieve real-time analytics to provide insights to the company as quickly as possible. Example of use case in TripAdvisor a travel company, has implemented Spark to process large amount of data from user interactions and destination details to provide trip planning recommendations based on the historical choices and preferences of the user (Kumar, 2019).

Pig – It is an execution engine for parallel data flows which includes a language called Pig Latin for data operations that can handle both structured and unstructured data. In addition, capable of developing own functions for read, write, and process data. Implementing Pig in Five Star Tourism would complement the data querying from Hive and transform data into useful insights. Flight data analysis and visualization using PIG has made companies improved their sales and passenger flow due to the efficiency of Pig Latin which is higher than normal SQL queries (Ishan Meena, 2019).

Service Programming

ZooKeeper – It is a coordination service for distributed applications ensuring the distributed system functions collectively by providing synchronization, serialization, and coordination over large clusters. Implementing ZooKeeper would help Five Star Tourism to optimize the maintenance schedule of the ecosystem to ensure minimum downtime. A review between users of AVRO and ZooKeeper found that Avro has met their business objectives better than ZooKeeper with majority of Avro users coming from the real estate sector (G2, 2022). However, when comparing in terms of product supports such as feature updates and roadmaps, ZooKeeper is more preferred. In overall, enterprise favor ZooKeeper over Avro while mid-sized marketer favor Avro.

Scheduling

Oozie – It is the workflow scheduler that manages the execution and work path of workflows. It can integrate with Pig, Hive, and Sqoop by providing workflows to each of the components. Oozie is a server-based web application. In addition, Oozie is scalable and provides the function to interrupt, start, pause, and rerun failed tasks. Oozie in Five Star Tourism can act as an automatic job schedular which allows raw data to be processed every day to receive latest updates automatically. Example use case in an airline company, flight data is required to be processed through three layers of transformation (Martínez-Prieto *et al.*, 2017). Whereby Oozie is utilized to coordinate each layer to ensure the transformed data is stored in specific repository and be accessible for further data analysis.

Machine Learning

Mahout – It is a machine learning and data mining library that supports scalability. It is useful for building recommender systems which are beneficial in the tourism sector in ensuring effective management and improving customer experience. Mahout in Five Star Tourism would assist the prediction of flights and hotels search trends that would allow the company to implement targeted marketing strategies to better capture the attention of the customers. Air Pacific Travel has utilized Mahout for clustering, collaborative filtering, classifications, and frequent pattern mining to predict the flight and hotel data web search trends to achieve improvement in company ranking in the market (Caesar, 2022).

Deployment

Ambari – It is a web-based tool providing a dashboard for provisioning, managing, and monitoring the health and status of the Big Data ecosystem clusters. Ambari in Five Star Tourism provides easy management and configuration to the clusters in the ecosystem. In the case of Air Pacific Travel, Ambari has provided a reliable interface for administration control thus making Big Data cluster management easier (Caesar, 2022). In addition, Ambari is highly extensible and customizable which provides custom services to Air Pacific Travel. This in result has improved work efficiency of Air Pacific Travel.

Figure 3.1 illustrates the Big Data ecosystem based on the proposed Big Data tools as mentioned, which provides an overview of the tools involved.

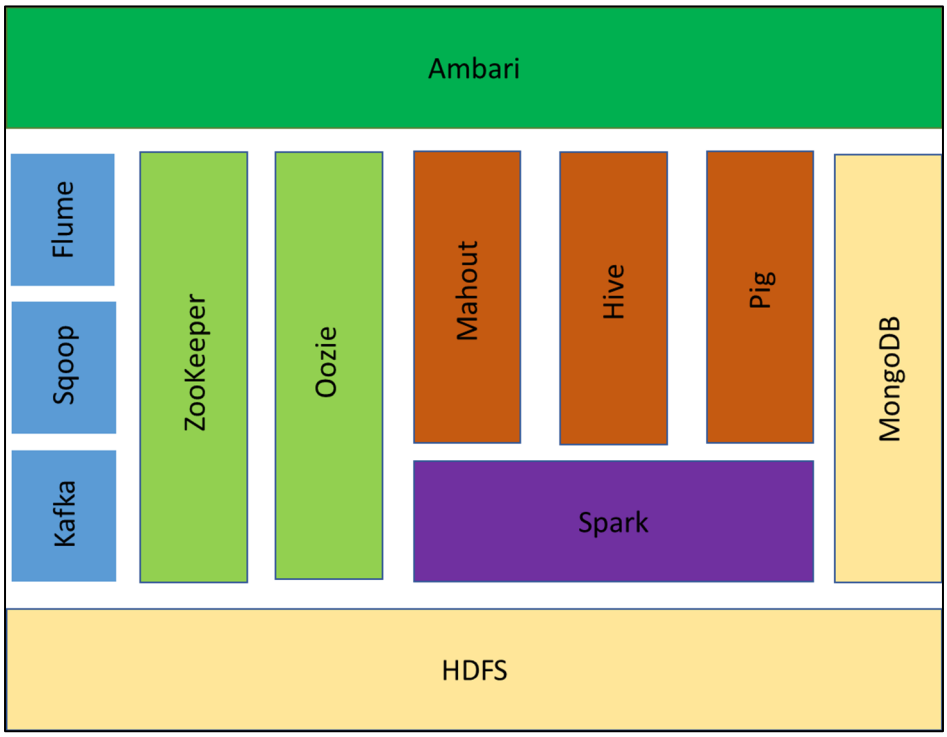

Figure 3.1: Proposed Big Data ecosystem

3.3 SECURITY, SOCIAL AND ETHICAL ISSUES

The previous section has proposed a Big Data ecosystem for Five Star Tourism to handle the storage and processing of Big Data to achieve hotel and flight search trends in a faster and cost-effective manner and ultimately helps Five Star Tourism to achieve competitive advantage in the sector. However, as voluminous data is handled by multiple users in the organization, it is possible for the data to be misused. This is due to the ecosystem lacking the implementation of a security layer. This is a serious problem as the data consisting of customer information should be remained confidential and used responsibly. Therefore, data authorization and accessibility are important which tracks the user who is using the data to ensure data is used responsibly and for specific legitimate tasks. It is highly suggested to implement a security layer in the ecosystem to ensure security and confidentiality of the data. In addition, to guard against cyber threats which are a prominent problem in the current age. Several tools can be implemented in the Big Data ecosystem to ensure security of the data is maintained (Caesar, 2022). Firstly, Kerberos can be implemented which is a network security protocol that is used to encrypt customer sessions when browsing the website of the company to ensure data security of the customer is maintained. Secondly, implementing traffic encryption which encrypts external sourced data that is ingested into the HDFS. Lastly, implementing HDFS file and directory accessibility permission to ensure only very specific users and clients are allowed to access specific data.

SECTION 4

CONCLUSION

A Big Data ecosystem was proposed to Five Star Tourism to cater for the problem of storage and processing of Big Data. In addition, providing the capability to predict trends in flights and hotels web search to provide tailored services to targeted audience. 12 Big Data tools have been proposed in the ecosystem that work collectively to achieve the needs of Five Star Tourism. The proposed ecosystem would allow processing of high volume and velocity of structured to unstructured data. In addition, introducing fault tolerance to ensure minimal downtime experience in the system thus reducing the overall maintenance costs. Furthermore, the use of SQL like languages in Hive and Pig would allow more users to easily adapt and perform simple analytics. Analytical engines are implemented for the company to automatically derive valuable insights from the data which helps achieve business goals and improve customer experiences. Data at the current age is an important asset for every company. Therefore, implementation of data security and privacy is highly encouraged to ensure data utilized in Five Star Tourism is ethical and within the bound of law. In turn, this would increase trust and confidence of customers using Five Star Tourism.

REFERENCES

ActiveWizards. (2019). *Top 7 Data Science Use Cases in Travel*. Retrieved from https://www.kdnuggets.com/2019/02/top-7-data-science-use-cases-travel.html

Aman. (2015). *US flight data analysis using hive*. Retrieved from http://www.lifeisafile.com/flight-analysis/

Baer, T. (2018). *Kafka is establishing its toehold*. Retrieved from https://www.zdnet.com/article/kafka-is-establishing-its-toehold/

Becha, M., Riabi, O., Benmessaoud, Y., & Masri, H. (2020). *Applications of Big Data in Tourism: A Survey.* Paper presented at the Advanced Data Mining and Applications, Cham.

biSmart. (2019). How big data technologies can improve tourism. Retrieved from https://blog.bismart.com/en/big-data-technologies-tourism

Caesar, M. (2022). *Big Data Ecosystem Case Study*. Retrieved from https://caesarmario.medium.com/big-data-ecosystem-case-study-a85b15ed46a8

Claude, U. (2020). Predicting tourism demands by google trends: a hidden markov models based study. *Journal of System Management Sciences, 10*(1), 106-120.

Emms, S. (2020). *9 Best File Systems for Big Data*. Retrieved from https://www.linuxlinks.com/filesystems/

G2. (2022). *Compare AVRO and ZooKeeper*. Retrieved from https://www.g2.com/compare/avro-vs-zookeeper

Ishan Meena, R. A., Vijayditya Sarker, Nadhini. (2019). Flight Data Analysis Using PIG. *International Journal of Engineering and Advanced Technology, 8*(4).

Jonathan Seidman, R. V. (2010). *Using Hadoop and Hive to Optimize Travel Search*. WindyCityDB 2010. Retrieved from https://fdocuments.net/reader/full/using-hadoop-and-hive-to-optimize-travel-search-windycitydb-2010

Kumar, N. (2019). *Apache Spark Use Cases & Applications*. Retrieved from https://www.knowledgehut.com/blog/big-data/spark-use-cases-applications

Martínez-Prieto, M. A., Bregon, A., García-Miranda, I., Álvarez-Esteban, P. C., Díaz, F., & Scarlatti, D. (2017, 17-21 Sept. 2017). *Integrating flight-related information into a (Big) data lake.* Paper presented at the 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC).

MongoDB. (2016). *With Expedia Online Travel Gets Personal*. Retrieved from https://www.mongodb.com/customers/expedia

Olson, J. (2018). Big data at United Airlines [Video]. Youtube. https://www.youtube.com/watch?v=WaSbJ6Xqyjo.

Pasarakonda, R. (2020). Use-case on Sqoop, HDFS, Hive, and Spark.

Rodrigues, A., & Chiplunkar, N. (2019). A new big data approach for topic classification and sentiment analysis of Twitter data. *Evolutionary Intelligence*. doi:10.1007/s12065-019-00236-3

Spanos, A. (2015). Making the case for Hadoop in a Large Enterprise [Video]. Youtube. https://www.youtube.com/watch?v=xLe_UA9R_Kk.

Strickland, S. (2020). *How Wyndham Hotels & Resorts improved their data capabilities to get to know their guests better*. Retrieved from https://www.pwc.com/us/en/library/case-studies/wyndham-data-architecture.html

Vohra, D. (2020). *Comparing Apache Sqoop, Flume, and Kafka*. Retrieved from https://www.techwell.com/techwell-insights/2020/05/comparing-apache-sqoop-flume-and-kafka