**Chapter 1**

**THE PROBLEM AND ITS SETTINGS**

**Introduction**

Enterprises today, both nationally and globally, are continuously seeking competitive advantages. It has become an undeniable saying that information is the key to determining how to gain such a competitive advantage. The problem today is how to deal with the mountains of raw data which existing information systems are collecting, processing, developing, and spreading. We are in the midst of an outburst of data. Somewhere hidden in this explosion of data are the clues management needs to define their strategies to make the best use of their competitive position. In this picture, technology has inserted the concept of Data Warehousing and Data Analytics as an alternative to deal with the well-known information overload described above.

Data Warehouses exist to help the management and decision makers transform raw data into information and to help management identify key trends. It helps the enterprise foresee predictable events and act in anticipation of those events and helps management understand the entire picture of what has already happened. It allows them to develop a good systemic understanding of events, thus allowing focused reactions to those events, such as redefining and re-engineering business processes to take advantage of that understanding. A clear need to allowing management to accomplish all of the above is the fact that the data to be analyzed has to be accessible and flexible, and it has to be available in a format that is usable. To date, too much weight has been employed on the raw technology which is the concept of Big Data, and still not enough.

This paper discusses how real-time and analytics could be a solution to this need. As the concept of faster transaction evolves, the synchronism between transactional data and data warehouses, statically implemented and has been redefined. Traditional data warehouse systems have static structures of their schemas and relationships between data and therefore are not able to support any dynamics in their structure and content. Their data is only periodically updated because they are not prepared for continuous data integration. For real-time enterprises with needs in decision support purposes, real-time data warehouses seem to be very promising. In this paper, we present a methodology on how to adapt data warehouse schemas and user-end OLAP queries for efficiently supporting churches or mission groups toward their goal of reaching more people.

**Background of the Study**

This study is undertaken to improve the process of providing accurate information in a timely manner and building the foundation for an integrated reporting environment. The goal of the research is to provide a central repository for related information, such as key contacts, volunteers, small groups, outreach events, and materials.

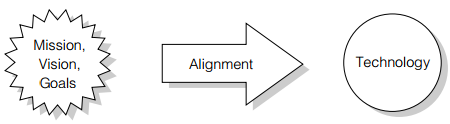
Currently, the data is hosted on various sources resulting in silos of information where data retrieval and compilation can be time-consuming. Also, some of the data come in the forms of Excel spreadsheets. The ability to do any custom analysis or drill down capabilities is not available.

Some of the challenges facing the users include the timely access to accurate data. With the use of the system, there will be a central place to get the data needed for contacts, volunteers, small groups, materials, outreach events, and attendance, and can provide automated reports and dashboards.

**Theoretical Framework**

**Business Model Framework**

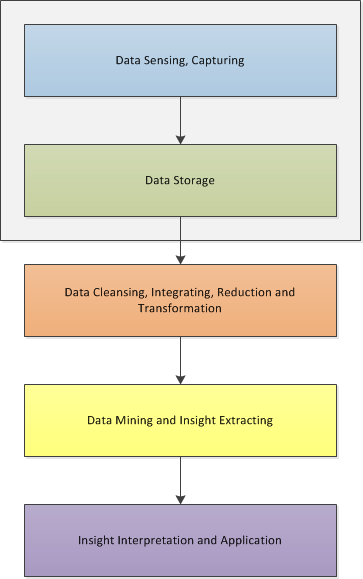
Figure 1: Critical Success Factors



No technology will ever be successful if it is not aligned with the organization's mission, vision, and goals. Therefore, it is crucially important to identify such elements of the study. This model greatly helps the system in its business flow process.

**Big data driven safety decision-making model**

Figure 2: Big-data-driven safety decision-making model

****

The figure above shows the theoretical framework of this research based on the study of Huang, Wu, Wang and Ouyang (2018) in Big-data-driven safety decision-making the stages in developing a big-data driven decision-making system.

Stage 1 involves Data sensing, capturing and storage. First, the researcher considers the inputs for the system and how its gathered. The researcher gathers data through an existing system, devices, web pages or encoded by the user. Second, the researcher considered these data storage. It may be structured, unstructured or semi-structured. Stage 2 involves Data Cleansing, integration by using middleware technology, reduction, by dimensionality reduction and compression and transformation by normalization, aggregation and discretization. Next stage is Data mining which involves choosing the best algorithm applicable to the data gathered and Insight extracting. Last Stage is Insight interpretation and application, it could be a descriptive insight, showing what has happened and what is happening to help in making strategic decisions.

**Conceptual Framework**

Figure 3: Conceptual Framework of the study.

1.Related literature and studies about Big Data

2. Survey Questionnaires filled out by the church or mission group

3. Use of theories and models to support the study.

1. Apply statistical treatment and analyze data gathered from related literature and studies.

2. Analyze the gathered information through the use of the questionnaires.

3. Development of the YAVATAN project

4. Comparison of algorithms

1. Finished YAVATAN

2. Evaluation of the churches or mission groups.

3. Conclusion and Recommendation of the study.

**INPUT**

**PROCESS**

**OUTPUT**

The figure above shows the source of data for this study which are the survey questionnaires filled out by the church or mission group. Those data would be used as a basis for the development of the YAVATAN project, then it will also be evaluated**.**

**Statement of the Problem**

This study aims to explore and develop a system using real-time database and analytics for church missions in achieving the Great Commission.

Specifically, the study sought to answer the following:

1. What are the issues encountered by the respondents in the existing system for recording, consolidating and summarizing the data for contacts, volunteers, small groups, materials, outreach events, and attendance?
2. What data mining algorithm could be utilized for information extraction, summarizing and clustering and how was it used in the development of the system?
3. What is the level of acceptance of the respondents concerning the proposed project for church or mission-related transactions, in terms of:

3.1. Efficiency;

3.2. Usability;

3.3. Reliability;

3.4. Functionality

4. What is the level of agreement of the respondents in using Predictive Analytics for Yavatan’s version 2?

5. What would be the ideal Yavatan system’s platforms?

**Scope and Limitations**

This research discusses the data mining procedures and other existing big data technologies that could help in improving the process of the churches or mission groups in terms of data storage and data processing.

This study takes the bottom-up approach, starting right in highly focused and targeted data marts directed at specific critical areas. The philosophy here has to do with deciding between addressing short-term tactical requirements to help individual modules and long term strategic planning issues regarding data architectures.

The Bottom-Up approach favors the use of lesser, more focused applications that can avoid the pitfalls of the Top-Down approach by basically limiting the range of the implementation. This method also shows simpler data archaeology problems, and these are usually limited data sets, limited user views, and a good understanding of the data requirement and how they relate to the business problem. In this approach, each church or mission group is responsible for extracting whatever data they need and using their own database for decision support at their level.

**Significance of the Study**

This study is beneficial to the churches and mission groups that can pull the analytical insights from data sets in the community and partner them up with the current trends in their congregation, the larger dataset can help in creating new ministries and developing a strategic plan. Eighty percent of the world's data is unstructured, and most churches or mission group do not even try to use this data to their advantage. The trend of Big Data is making way for new opportunities and new challenges for businesses across industries. The churches also have a lot of unstructured data. The data should be stored and analyzed by Big Data techniques and method. This approach is useful for better analysis. By using computerized or electronic processes, the system provides speed and reliability since computational errors as well as errors in posting will be reduced because of less human intervention. And importantly, by knowing the current condition in a specific area, it can help in creating strategies in reaching out the inhabiting people there. With these, the system can determine and provide their real needs and delivering the solutions in the most effective way possible.

## **Significance of the Study**

This study determines the respondents’ assessment and preferences in using real-time database and analytics for church missions in achieving the Great Commission, therefore, will be deemed important to the following:

**To the church or mission group leaders,** as this study helps those responsible in the outreach planning and handles small groups in the actual event.

**To the mission volunteers,** as this study helps the volunteer looking for the outreach events, they would like to participate in.

**To the Future Researchers**, as this study helps other researchers in the same topic of conducting a study about real-time processing of data and analytics, as it serves as a resource and reference for the future researchers.

**Definition of Terms**

**Data warehouse**. A large store of data accumulated from a wide range of sources within a company and used to guide management decisions.

**Data Stream**. Is a sequence of digitally encoded coherent signals (packets of data or data packets) used to transmit or receive information that is in the process of being transmitted.

**Database**. Is a structured set of data held in a computer, especially one that is accessible in various ways.

**Business Intelligence**. Is an umbrella term that refers to a variety of software applications used to analyze an organization's raw data. BI as a discipline is made up of several related activities, including data mining, online analytical processing, querying and reporting.

**OLAP (Online Analytical Processing)**. Is a powerful technology for data discovery, including capabilities for limitless report viewing, complex analytical calculations, and predictive “what if” scenario (budget, forecast) planning.

**OLTP (Online transaction processing).** Is a class of information systems that facilitate and manage transaction-oriented applications, typically for data entry and retrieval transaction processing.

**Drill-down**. Is a data warehouse term which refers to the ability to start at a summary piece of information and through a mouse or keyboard entries work your way down to the lowest detail available.

**Chapter 2**

**REVIEW OF RELATED LITERATURE AND STUDIES**

This chapter presents the existing theories and knowledge related to the study, which serve as a guide to its development. Given the lack of local studies pertaining to the use and application of real-time database and data mining in the Philippines, the foreign studies have sufficed the necessary data for the ideas needed by the proponent to conceptualize the study.

# Big Data

Big data means large datasets which are large with diverse data types that it becomes difficult to process using traditional platforms therefore require new technologies as said by (Russon, 2013) and Big data technologies are used for several tasks.

According to Chen and Zhang (2013), the four V’s of Big Data are: (a) Volume, represents the actual amount of data available; (b) Velocity, the speed at which data is being created and how fast it must be processed to meet business needs; (c) Variety, refers to the multiple data sources and formats in which data can be presented; and (d) Veracity, denotes the uncertainty surrounding data caused by inconsistency and incompleteness.

**Data Storage**

# Data warehousing

The concept of data warehousing evolves when we require a huge space for data storage and along with that the data that is converted into useful information can be used in the future to perform an analysis. As stated by Mahboubi, Ralaivao, Loudcher, Boussaid, Bentayeb, and Darmont (2017) Data warehousing and OLAP applications must handle complex data that are not only numerical or symbolic these days. Dewan, Aggarwal and Tanwar (2014) state that “A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process” where the term subject-oriented means that a data warehouse should focus on an area for e.g. sale, the term integrated means that data is gathered from multiple sources due to which data can be in different terminologies and formats and time-variant points towards the historic collection of data in the data warehouse.

According to Kimball "A data warehouse is a copy of transaction data specifically structured for query and analysis," it defines that data warehouse consists of same data as in the source system but it is structured or transformed to perform analysis on it. A data warehouse is developed through a process called ETL (Extract transform and load) according to which data is extracted from heterogeneous data sources which are working for the organization then the data is sent to the staging area which is also called transformation in which the data quality is improved and data is cleansed and is brought to the same format as it is present in the data warehouse, then the loading process which is done periodically as per the traditional approach. During the uploading of new data, the data warehouse is not available for query session.

Sathish and Vyedhi (2013) stated that data warehousing is specifically intended to provide vital strategic information. In the 1990s, organizations began to achieve competitive advantage by building data warehouse systems. Traditionally data warehouse stores the historical data, for instance, relevant information needs to be delivered as fast as possible to knowledge workers or decision systems that rely on it to react in a near real-time manner, according to the new and most recent data captured by an organization's information system

According to Jameel, Memon, and Shaikh (2017), a sufficient amount of information has become a requirement and need of a human being in order to succeed in today's environment. The information needs to be integrated from multiple sources and adjusted to a certain level.

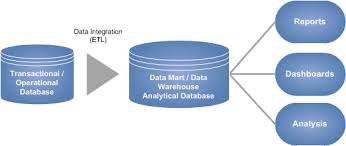
Now, data warehouses have been improved to support more than just strategic reporting, analytics, and daily forecasting. Organizations are investing significant resources to integrate valuable information contained in their data warehouse into their day-to-day operations. Incorporating business intelligence into decision making helps the organizations to optimize their performance throughout the day.

# Real-time Data Warehousing

However, to achieve these efficiencies, data must be supplied in a real-time environment. There are many data integration technologies that serve the data acquisition needs of a data warehouse in organizations, and the demand for low-latency data is causing IT organizations to evaluate a wide range of approaches: intraday batch Extract, Transform, and Load (ETL) processes as well as real-time Change Data Capture (CDC) techniques. In one Oracle White Paper (2012), it is said that Business time is increasingly moving toward real time. As organizations look to grow their competitive advantage, they are trying to uncover opportunities to capture and respond to business events faster and more rigorously than ever. Today, the majority of competitive advantage comes from the effective use of IT. Therefore, from that standpoint, the key to achieving faster and accurate Business Intelligence (BI) is a robust enterprise data warehouse combined with an enterprise analytics framework. Across the enterprise, each fact of the business gathers data through a collection of activities. The quality of these decisions depends not only on the complexity level of the analytics applications that run on the data warehouse but also on the underlying data. Data has to be complete, accurate, and trusted. The longer it takes to capture and process data, the lower the value of information. Replication is one-way from the active to the standby system. The standby systems are not running mission-critical online applications; they may be used for query and other non-update type services.

# Data Warehouse Architecture and Extraction, Transformation and Loading (ETL)

Figure 4: The figure above shows the Data Warehousing Architecture



To enable timely and consistent analysis, there should be combining of highly available systems with active decision engines to allow near real-time information dissemination for data warehouses. Cumulatively, this is the basis for zero-latency analytical environments. The real-time data warehouse provides access to an accurate, integrated, consolidated view of the organization's information and helps to deliver near real-time information to its users. This requires efficient ETL techniques enabling continuous data integration, the focus of this paper. By adopting real-time data warehousing, it becomes necessary to cope with at least two radical data state changes. First, it is necessary to perform continuous data update actions, due to the continuous data integration, which should mostly concern row insertions. Second, these update actions must be performed in parallel with the execution of OLAP, which due to its new real-time nature, will probably be solicited more often. Therefore, the main contributions of this research are: (1) Maximizing the freshness of data by efficiently and Rapidly integrating most recent OLTP data into the data warehouse; (2) Minimizing OLAP response time while simultaneously performing continuous data integration; and (3) Maximizing the data warehouse’s availability by reducing its update time window, in which users and OLAP applications are off-line.

Before a data warehouse could be implemented, an integrated architecture and a companion implementation methodology needed to be adopted. ETL processing occurs between data sources and the data warehouse, between the data warehouse and data marts and may also be used within the data warehouse and data marts.

## **Data Sources**

Integrated data repositories (IDRs) are poised to become a foundational element of successful and efficient data warehouse projects by providing the coordinated data sources.

This component of our data warehouse architecture (DWA) is used to supply quality data to the many different data marts in a flexible, consistent and cohesive manner. It is a ‘landing zone' for inbound data sources and an organizational and re-structuring area for implementing data, information and statistical modeling. This is where business rules which measure and enforce data quality standards for data collection in the source systems are tested and evaluated against appropriate data quality business rules/standards which are required to perform the data, information and statistical modeling described previously. A powerful and flexible extraction, transformation and loading (ETL) process is to use Structured Query Language (SQL) views on host database management systems (DBMS) in conjunction with a good ETL tool.

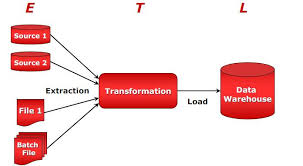
## **Data Marts**

This component of the data warehouse architecture may manifest as visible relational tables, OLAP cubes, Pre-determined parameterized and non-parameterized reports, Ad-hoc reports, Spreadsheet applications with pre-populated worksheets and pivot tables, Data visualization graphics, Dashboard/scorecards for performance indicator applications. Typically a business intelligence (BI) project may be scoped to deliver an agreed upon set of data marts in a project. Over time as more and more data marts and logical data models (LDM's) are built the conceptual data model (CDM) becomes more complete.

## **Extract, Transform, Load (ETL) Process**

The ETL process involves extracting the data from source systems, transforming it according to business needs, and loads the results into the target data warehouse. The quality of data is dependent on the efficiency of the ETL process.

Figure 5: The figure above shows the ETL Process



In the study of North, Thomas, Richardson, and Akpess (2017) their primary goal is to provide managers and executives the tools and information needed to make informed decisions concerning data warehousing, to understand the processes and technology involved, and to identify individuals' responsibilities. The information is presented in clear, understandable terms and is designed for decision-makers with little or no Information Technology (IT) background.

As said by Runtuwene and Tangkawarow (2017) Such solution typically allows managers to have a clear picture of their organization’s performance within the industry. It allows managers to clearly see key performance indicators (KPIs) on different dimensions (time, region, product, etc.).

In the study of Chhabra, Kumar, and Pahwa (2016) they used Data warehouses as the repository of data from a wide range of sources that provide analytical results for making significant business decisions. As a data warehouse has to support different complex queries, therefore, the design technique should be different from the traditional ones. The best-suited technique is an object-oriented data warehouse design technique which is a collection of objects that interact with each other. That further enhances the decision making capability of Data Warehouse. In their paper, they illustrated a technique for designing a relational schema from an object model, represented in UML form and then transformed it into a data warehouse.

Sam Madden (2016) once had a study where the application developed involves producing reports over Google’s ads infrastructure: Google executives want to see how many ads each Google property is serving, and how profitable they are, and Google’s customers want to see how many users are clicking on their ads, how much they are paying, and so on. At a small scale, solving this problem is straightforward, new ad click and sales data are appended to a database file as they are sent from the processing system, and computing the answer to a particular query over the data involves reading the contents of the data file to compute a running total of the records in the groups the user is interested in. Making this performance at the scale of Google Ads, where billions of clicks happen per day, is the challenge addressed by the Mesa system. Some of the highlights of this work include: (1) The use of batch updates and append-only storage. New data is not added one record at a time but is aggregated into batches that are sent into Mesa. Instead of merging these batches into the existing storage, these batches are simply written out as additional files that need to be scanned when processing a query. This has the advantage that existing files are never modified, so queries can continue to be executed while new data is added without worrying about new data being partially read by these existing queries; (2) Massive scale parallel query processing. Each query is answered by one query processing node, but there can be hundreds or thousands of compute nodes. They can each answer queries independently because of the use of append-only storage: query processors never need to worry that the files they are scanning will change while they are running, and query processors never wait for each other to perform operations; (3) Atomic Updates, because some care is needed to atomically install update batches, such that they are either not seen at all or are seen in their entirety. Mesa labels each update with a unique, monotonically incrementing version number, which is periodically communicated to each query processor. Once a query processor learns of a new version number, it will answer queries up to and including that batch and is guaranteed that the files containing the batch have been completely written and will not change. This means it can take some time (a few seconds to a minute) for a query processor to see a new update; and (4) Unusual Data Updates, because database systems are optimized for high throughput but lack several features that are a requirement of the Mesa solution. First, Mesa fits neatly into the elegant modular (layered) software architecture stack Google has built: It runs on top of Colossus (their distributed file system) and provides a substrate on which advanced query processing techniques (like their F1 system) can be built. Layering software this way allows different engineering teams to maintain code and allows different layers to service multiple clients. Many existing data processing systems are much more monolithic and would be difficult to integrate into the Google software ecosystem. Second, conventional databases were not built to replicate data across multiple data centers. Traditional systems (typically) use a single-master approach for fault tolerance, replicating to a (read-only) standby that can take over on a master failure. Such a design will not work well if datacenter failures or network partitions are frequent.

Sultan (2016) discussed the core concept of the data warehouse and what are some requirements to build a data warehouse and the industrial perspective and need for a data warehouse and where can it be used and why it is so important for the growth of any industry and how the concept of real-time data warehousing and real-time business intelligence be used in the industry and how much it will be beneficial for the one who adopts it.

Moving with the requirement of organization and the growth of technology the concept of near real-time data warehousing which can be considered as an alternative of a real-time data warehouse where minor delays will not create many problems for the business. In this paper, a solution was also proposed by using CDC (Change data capture).

According to Sultan (2016), organizations now demand live data into their data warehouse so that analysis can be performed on the current statistics of the business to make better business decisions. A concept of real-time data warehousing is evolving with different approaches so that live data or at least most recent data can be provided to the data warehouse so that efficient business strategies can be designed to enhance business performance.

Since the establishment of the data warehouse concept a number of ideas that have responded to the problems of the last decade, were created. The new solution is the Stream Data Warehouse (StrDW), which was presented by Gorawski and Gorawska (2014) along with many other models and solutions. The Stream Data Warehouse is a solution that combines the idea of the data warehouse with stream data processing model benefits. Therefore, it is more complex and sophisticated tool. Data Stream Processing Model. The data stream processing model is one of the solutions to the problem of handling huge volumes of data. The foundation of systems based on this model is an assumption of data streams existence, i.e. infinite collections of data that supply system in an unpredictable manner. Moreover, the order and content of data volumes are independent of the system. Streams are used only for transmission of data in a form of tuples. The data stream can also be defined as interrelated thematic elements (records) belonging to an unlimited set of tuples and timestamps. A tuple is the smallest possible portion of data containing information about a particular phenomenon or state of a tested situation at a specific point in time. However, a feature that most distinguishes this particular model from others is the fact that tuples are processed on the fly and then deleted. Stream Data Warehouse's Architecture. The Stream Data Warehouse creates a platform for processing a variety of stream data sets stored in a Stream Materialized Aggregate List in a form of an analytic information. The StrDW concept assumes the existence of many collaborating engines.

In one study by Jain, Gyanchandani, Khare, Singh, and Rajesh (2017) the researchers introduced the meaning of enormous information and Big Data challenges. Referring privacy and security aspects in big data and presented existing techniques of anonymization using MapReduce framework and the study of Wang and Wang (2016) showed that with the technology adoption, the managerial function could be improved and information system could be upgraded. Since the technology is reshaping the world of, it is significant for people to exploit technologies.

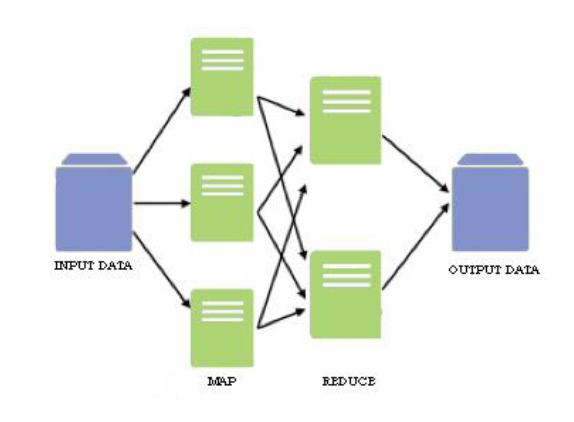
Analyses of the gathered data in the study of Ram (2013) indicated that most organizations are able to use the information from their data warehouses for marketing, decision making, financial planning, discovering trends, and product development. Some of the organizations could use the information for the purposes of fraud detection, loss prevention and knowing who their customers are. Benefits that the organizations would like to derive by using data warehousing and data mining tools include: a single version of the truth, greater insights into customer benefits, keep alongside new developments, empower frontline staff by providing them with self-service tools which would enable them to analyze data and make decisions.

# Hadoop Ecosystem

Many organizations are taking into consideration the better way to collect, process and analyze big data and have turned to a newer class of technologies which includes Hadoop and other related tools such as YARN, MapReduce, Spark, Hive, and Pig as well as NoSQL databases. These technologies form the core of Hadoop Ecosystem that supports the processing of bulky and various datasets across clustered systems.

According to Das and Kumar (2013), Hadoop is a software-framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, massive processing power and the ability to handle virtually limitless concurrent tasks or jobs. Hadoop is one of the most interesting and promising new technologies related to big data as said by Agrawal, Joshi, and Velez (2016).

Figure 6. The Apache Hadoop with HDFS and MapReduce.



In some cases, Hadoop clusters and NoSQL systems are used as landing pads and staging areas for data before it gets loaded into a data warehouse for analysis, often in a summarized form that is more beneficial to relational structures. Increasingly though, big data vendors are pursuing the concept of a Hadoop data lake that serves as the central storehouse for an organization's incoming streams of raw data. In that way, subsets of the data can be filtered for analysis in data warehouses and analytical databases, or it can be analyzed directly in Hadoop using batch query tools, stream processing software and SQL on Hadoop technologies that run interactive, ad hoc queries written in SQL.

Potential pitfalls that can trip up organizations on big data analytics initiatives include a lack of internal analytics skills and the high cost of hiring experienced analytics professionals. The amount of information that's typically involved, and its variety, can also cause data management headaches, including data quality and consistency issues. In addition, Roy and Agarwal (2015) said that integrating Hadoop systems and data warehouses can be a challenge, although various vendors now offer software connectors between Hadoop and relational databases, as well as other data integration tools with big data capabilities.

According to Russon (2013), the Hadoop Distributed File System (HDFS), MapReduce, and various Hadoop tools will be the software products most aggressively adopted for BDM in the next three years. Others include complex event processing, NoSQL databases, in-memory databases, private clouds, in-database analytics, and grid computing.

## **Components of Hadoop Ecosystem**

The following are the components of Hadoop Ecosystem according to various authors Sagiroglu and Sinanc (2013) and Monteith et. al (2013):

### Storage

HDFS is the primary distributed file system used by Hadoop applications that run on large clusters of commodity machines. HDFS clusters are consist of a NameNode that manages the file system metadata and DataNodes that store the actual data. These are used as storage of large imported files from applications outside of the Hadoop ecosystem and for the staging of imported files to be processed by Hadoop applications.

HBase is a distributed, column-oriented database. HBase uses HDFS as its primary storage and supports both batch-style computations using MapReduce and random reads. It serves as a storage of large data volumes even up to billions of rows on clusters of commodity hardware, bulk storage of logs, documents, real-time activity feeds and raw imported data. It handles the consistent performance of reads and writes to data used by Hadoop applications. It stores data that can be aggregated or processed using MapReduce functionality.

HCatalog is a table and storage management layer for Hadoop that enables Hadoop applications like Pig, MapReduce, and Hive to read and write data into a tabular form as opposed to files. It also provides REST APIs so that external systems can access these tables’ metadata. It serves as a centralized location of storage for data used by Hadoop applications. It is a reusable data store for sequenced and iterated Hadoop processes like ETL. HCatalog is also used as storage of data in a relational abstraction

### Processing

MapReduce is a distributed data processing model and execution environment that runs on large clusters of commodity machines. It uses the MapReduce algorithm which breaks down all operations into Map or Reduce functions. It is used for Aggregation which includes counting, sorting, filtering, and stitching on large data sets, the scalable parallelism of Map or Reduce tasks, distributes task execution and machine learning.

Pig is a scripting SQL based language and execution environment for creating complex MapReduce transformations. Functions are written in Pig Latin language and translated into executable MapReduce jobs. Pig also allows the user to create extended functions using Java. Pig is used as a scripting environment to execute ETL tasks on raw data in HDFS. It is a SQL based language for creating and running complex MapReduce functions. It is used in data processing, stitching, schematizing on large datasets.

### Querying

HIVE is a distributed data warehouse built on top of HDFS to manage and organize large amounts of data. Hive provides a query language based on SQL semantics called HiveQL which is translated by the runtime engine to MapReduce jobs for querying the data. It is used as a schematized data store for housing large amounts of raw data. It has SQL-like Environment that executes analysis and querying tasks on raw data in HDFS and integrates with outside RDBMS applications.

### External integration

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data into HDFS. Flume transports large quantit0lies of event data using a streaming data flow architecture that is fault tolerant and failover recovery ready. It handles the transportation of large amounts of event data such as network traffic, logs, and email messages. It streams the data from multiple sources into HDFS. It ensures guaranteed and reliable real-time data streaming to Hadoop applications.

## **HADOOP Projects**

According to Schacter (2014), one of the companies that implemented Hadoop in their corporation is Pontis, it is a business involved in developing online marketing automation software for the telecommunications industry. Their solutions replace traditional, mass segmentation telco tactics with multi-dimensional context marketing that enables communications service providers to create continuous, on-going dialogs with their customers and respond to its personal needs and preferences. Some of their current clients include leading global telecommunication companies such as Vodafone, Telefónica, and VimpelCom. Pontis chose MapR as their goal architecture to integrate Hadoop into their organization and built a real-time operational analytics environment on Hadoop. By gradually transitioning from their legacy architecture to MapR, Pontis accomplished higher scalability and lower Total Cost of ownership in gradual steps. While the earlier system hit the wall at 100 million transactions a day, Pontis is now able to deal with 400 million customers daily.

# HBase

Apache HBase is one of the most popular non-relational databases built on top of Hadoop and HDFS. It is also known as the Hadoop database. It is a Wide-column store based on Apache Hadoop and on concepts of BigTable

# Polybase

As written in Polybase Guide (2017) PolyBase is a technology that accesses and combines both non-relational and relational data, all from within SQL Server. In SQL Server 2016, it allows you to run queries on external data in Hadoop or Azure Blob Storage. Queries are optimized to push computation to Hadoop. In Azure SQL Data Warehouse, you can import data from Azure Blob Storage and Azure Data Lake Store.

Transact-SQL (T-SQL) statements are used to import and export data back and forth between relational tables in SQL Server and non-relational data stored in Hadoop or Azure Blob Storage.

Storing and handling high quantities of data is challenging the current Relational Database Management Systems. Big Data and its related products came to help in this matter and NoSQL databases arise with the purpose to offer better solutions and feature to handle massive amounts of data with higher performance sometimes almost near real-time. Caouchbase, MongoDB, and RethinkDB are three different NoSQL databases extensively used in the market nowadays and comparing in the single thread and multiple thread scenarios, MongoDB scored the highest (Pereira, Morals and de Freitas, 2017).

In the study of Pandey and Ushaker (2016), they reviewed the work carried out in the area of commit processing in both distributed environment and distributed real-time environment. The traditional commit protocol and its variants are for distributed database systems. A representative set of real-time commit protocols have also been proposed which are specifically used in the real-time environment. Most of the real-time commit protocols improve the performance by allowing access to the dirty data for executing transactions. However, real-time commit protocols can still have improvements in the performance such as optimized memory use, creating less temporal objects needs to be investigated for getting better results.

# Firebase

In the recent years, Google had their own NoSQL database, As used in the project of Kumar, Akhi, Gunti, and Reddy (2016) in Implementing Smart home using firebase, it is a technology that allows us to make web applications with no server-side programming so that development turns out to be easier and quicker. Firebase is able to take every essential steps such as verifying users, storing data, and implementing access rules with little configuration. It also supports the iOS, web, OS X, and Android clients. Applications using Firebase can store, control, synchronize and use data in real-time in a variety of examples. Writing server-side code is not necessary when using firebase, or to deploy a complex server framework to get an app started with Firebase.

According to Ernesti, Fuchs, Kronenberg, and Lexhagen, (2017), Big data information sources can significantly increase the prediction performance. As stated by Schimd, Masson, and Hirsbrunner (2013), over the years, many different toolkits have been created to support the development of applications which started from APIs to address specific problems that were being explored at the time. Then used in the synchronization and conflict handling of events by cloud-based real-time backend solutions like Firebase.

In the Hikester project by Khatipov, Mazzara, Negimatzhanov, Rivera, Zakirov, and Zamaleev (2018) they used Firebase to develop a real-time event management system which uses data mining and data collecting API for recommending events based on the user social profile. When users create some event, the neural networks are notified about it and starts to analyze that data and then recommendation system send notifications to all users which may be interested in the newly created event.

It is a flexible backend with a ton of good uses. It cuts down development time and avoids upsetting servers and information storage because it gives cloud benefit. Data is stored as native JSON. Data is protected because that the Firebase requires 2048-piece SSL encryption for all information exchanges. Data is reflected and moved down to different secure areas, so there are negligible chances of information loss. It integrated nicely with popular systems such as Angular JS.

# Data Processing

# Data Science

Data science is a revived term for discovering knowledge from data (Dhar, 2013). It is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data. Possibly, the most closely related concept to data science is data mining—the actual extraction of knowledge from data via technologies that incorporate these principles. Data science is the connective tissue between data-processing technologies including big data and data-driven decision making. (Provost and Fawcett, 2013)

Data science involves much more than just data mining algorithms. It also involves methods and methodology for visualizing data and a large portion of what has traditionally been studied within the field of statistics. In order for data science to flourish as a field, we must think beyond the algorithms, techniques, and tools in used. Instead, focus on the core principles and concepts beyond the techniques, and also the systematic thinking that fosters success in data-driven decision making. These data science concepts are general and very broadly applicable.

# Machine Learning

Machine learning serves as the technical basis of data mining. It is used to extract information from the raw data in databases—information that is expressed in a comprehensible form and can be used for a variety of purposes. The process is one of simplification, taking the all the data and conclude whatever structure lie behind it. It is interpreted as the gaining of structural descriptions from examples.

Insights gained by the user are of most interest in the majority of practical data mining applications; indeed, this is one of machine learning’s major advantages over classical statistical modeling. (Witten, Frank and Hall, 2010)

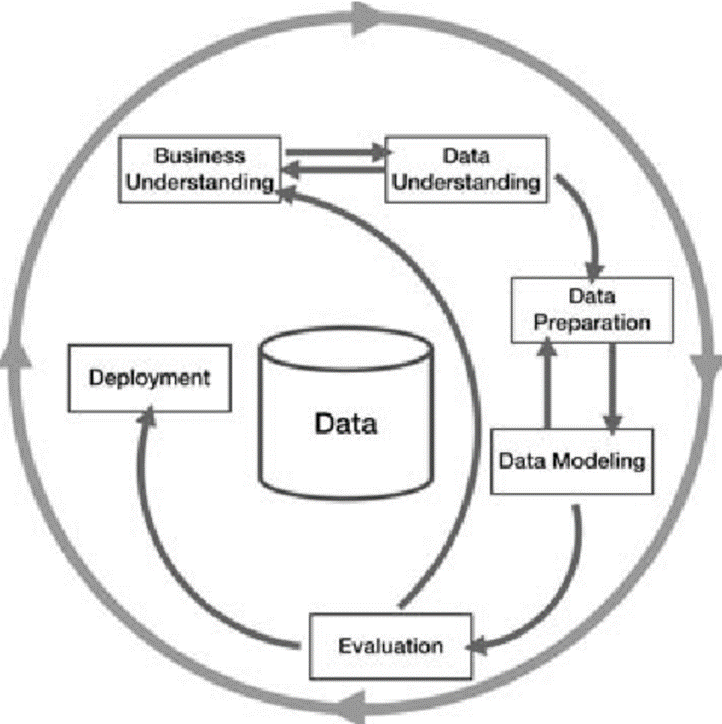
According to Chen & Zhang (2013) and Jordan & Mitchell (2015), the most evident characteristic of machine learning is to discover knowledge and make intelligent decisions automatically. It addresses the question of how to build computers that improve automatically over time under uncertainty also known as online learning. It is one of today’s most rapidly growing technical fields, which is at the intersection of computer science and statistics, and at the core of artificial intelligence and data science. In relation to Big data, there is a need to improve machine learning algorithms, both supervised learning and unsupervised learning, to cope with it. For example, Support Vector Machine (SVM), which is a very fundamental algorithm used in classification and regression problems, suffers from serious scalability problem in both memory use and computation time. Parallel SVM (PSVM) is introduced recently to reduce memory and time consumption. Several frameworks also exist like Map Reduce, DryadLINQ, and IBM parallel machine learning tools that have capabilities to enhance machine learning.

# Knowledge Discovery

According to Osei-Bryson and Barclay (2015), Knowledge discovery concerns the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine.

Nowadays researchers with strong industrial engagement realized the need from DM to KDD to deliver useful knowledge for the business decision making. The process of getting useful knowledge from data to solve business problems can be done systematically, traditionally, one standard, named CRISP-DM (Cross-Industry Standard Process for Data Mining)

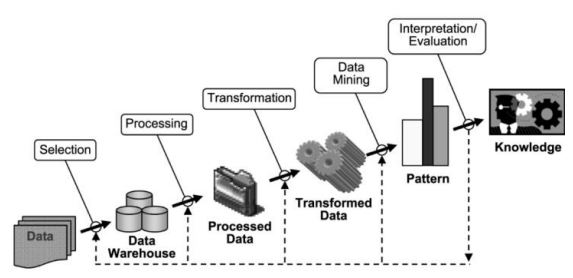
Figure 8: Cross-Industry Standard Process for Data Methodology (CRISP-DM)



According to Ramman, Shamsuddin, Hasan, and Haris, (2016), the figure above the first phase is business understanding where to understand what is really to be accomplished. This task involves more detailed fact-finding about all the resources, assumptions and other factors that should consider in determining the data analysis goal. The second phase is data understanding that investigates a variety of descriptive data characteristics. The third phase is data preparation which is the most difficult and time-consuming element in the KDD process. The goal is to choose relevant data from available data and to represent it in a form which is suitable for the analytical methods that are applied. Data preparation includes activities like data selection, filtering, transformation, creation, integration, and formatting. The fourth phase is modeling which is the use of analytical methods or algorithms. There are many different methods and the most suitable one must be chosen. This phase is also verifying the quality of the model such as testing in the independent data matrix, and others. The fifth phase is evaluation where the interpretation and evaluation of the discovered knowledge. In a decade, CRISP representation of DM process seems to become more dominant. However, using this traditional framework represented some issues when the deployment stages are taken. The framework life cycle is sequential and linear. In the real world scenarios, challenges always come from specific domain problems which back to the goal of DM towards business concerns, hence the objectives and goals of applying KDD are basically problem-solving to satisfy real user needs.

Data mining also popularly known as Knowledge Discovery in Databases (KDD) according to authors Sathiyamoorthi (2017), Wu and Olson (2016), Redgert (2017) Thirumagal, Suganthy, Mashima and Kesavaraj (2014) refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. It includes extracting useful information from the huge amount of data present in the database. The information from the data can be used to discover beneficial trends for the organization’s process, which is helpful to discover trends and perform other analytical activities. While data mining and knowledge discovery in databases or KDD are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

Figure 9: Data Mining is the core of the Knowledge Discovery Process



The following is a brief description of the nine-step KDD process according to Data Mining and Knowledge Discovery Handbook, starting with a managerial step: (1) Developing an understanding of the application domain; (2) Selecting and creating a data set on which discovery will be performed; (3) Pre-processing and cleansing; (4) Data transformation; (5) Choosing the appropriate Data Mining task; (6) Choosing the Data Mining algorithm; (7) Employing the Data Mining algorithm; (8) Evaluation; and (9) Using the Discovered Knowledge.

The data can be analyzed from the different point of view and different types of useful information can be obtained from these analyses. For Data Mining there are different classes of statistical methods as stated by Das (2016) like the Classical statistics (regression, curve fitting etc.), Induction of symbolic rules and neural networks.

# Data Mining and Analytic Tasks

The data mining and analytic task are categorized into following: (1) Similarity Learning; (2) Rule / Pattern Matching; (3) Outlier / Anomaly Detection; (4) Clustering, defined as the task of organizing data into sensible groups. Many different clustering algorithms exist, which may all produce very different partitions of the same data set. (Craenendonck and Blockeel, 2015); (5) Graph / Network Analysis; (6) Pattern matching; (7) Descriptive Statistics; and lastly, (8) Comparative Analysis, which is a common technique in statistics. It refers to the task of quantitatively and qualitatively comparing two or more objects with each other. These objects can be for example the unemployment rates for two countries or the budgets of two municipalities. Descriptive statistics and visualizations are the basis for comparative analysis. Possible axis for comparative analysis are e.g. temporal, geographical, and by sector. Therefore, comparative analysis is related but not limited to time series analysis. Regression and correlation matrices are terms in comparative analysis.

**Web Mining**

This section divides web mining into three categories depending on the type of data i.e. Web Content Mining, Web Structure Mining and Web Usage Mining.

Web content mining

Is the mining, scanning, and extraction of text, videos, graphs, and pictures from web documents. It is also known as text mining. Two types of approaches are used in web content mining. The two approaches are the database approach and the agent-based approach. The database approach helps in retrieving the semi-structured data from web documents. The agent-based approach searches relevant information and helps in organizing the collected information. Web content mining analyzes the content of web resources. Content data correspond to the collection of facts a web page was designed to convey to the users. Most of the data available on the web are unstructured data. Two different points of view of web content mining are the information retrieval view and the database view. The main goal of Web content mining from information retrieval view is to improve the filtering and finding of the information to the users. The main goal of a database view is to manage the web data (Srividya, M., D. Anandhi and M. I. Ahmed, 2013).

Web Content mining has the following approaches to mine data: unstructured mining, structured mining, semi-structured mining and multimedia mining.

### Web Structure Mining

The goal of web structure mining is to generate a structural summary about web pages and websites. It shows the relationship between the user and the web. It discovers the link structure of hyperlinks at the inter-document level. It also helps in discovering the structure of the document which is used in revealing the structure the structure of web pages and it's possible to compare the web page schemes. This is further divided into two types that are based on the kind of structural information used.

a) Hyperlinks: Hyperlinks help in connecting web pages to a different location either in the same web page or on a different web page. A hyperlink is divided into two categories i.e. intra-document hyperlink and inter-document hyperlink. Intra-document hyperlink connects different part of the same page whereas inter-document hyperlink connects two different pages.

b) Document Structure: The content within the web page can be organized in a tree structure that is based on various HTML and XML tags.

### Web Usage Mining

Is the process of finding out what users are looking for on the internet. It tries to discover useful information secondary data derived from the interaction of users while surfing web. There are three phases of web usage mining. The three phases are: (a) Preprocessing, it helps in retrieving the raw data from web resources and then processes the data; (b) Pattern Discovery: After preprocessing the data, the data is used for discovering patterns; and (c) Pattern Analysis, after discovering the pattern the pattern is analyzed and then the pattern is checked. If the pattern is correct then it is implemented on the web to extract the information from the web. (Saini and Pandey, 2015)

**Time series analysis**

As defined in European Union’s Horizon 2020 Research and Innovation Programme (2017), IBM SPSS Forecasting (2017), and author Gamboa (2017) includes methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the input data. The main assumption in analyzing time series is that the successive values of a variable represent consecutive measurements of equally spaced time intervals. There are two main goals in time series analysis: a.) identify an internal structure that we have to consider, such as autocorrelation, trend and seasonal variation and b.) forecast future values of the desired variable/measure based on previously observed values.

One approach to address the big data problem is through data analytics. While traditional statistical analysis and data analytics are often used interchangeably—and share many similarities—data analytics extends the focus to the analysis of larger datasets gathered from a wide variety of data sources (Davenport, 2013). In expanding the scope of traditional statistics, many now consider machine learning and data mining technologies - which are grounded in statistics - as important aspects of data analytics. It follows from the definition of data analytics that 'business analytics' is data analytics, but applied in a business environment to address business issues. Closely related to applying data analytics to address the big data problem has been the rise of the data science field.

# Data Mining Algorithms under Clustering

There are two basic types of clustering algorithms (Kaufman & Rousseeuw 1990): partitioning and hierarchical algorithms.

# K-means Clustering

A simple, unsupervised learning algorithm that is often used with big data sets, often as a way of pre-clustering or classifying into larger categories that other algorithms can further refine. It has some other inherent problems that make it best suited to large-scale, high-level clustering.

The k-means algorithm is a simple iterative method to partition a given dataset into a user-specified number of clusters, k.

# Mean Shift Clustering

Mean shift is a general non-parametric mode of finding/clustering procedure for regions containing a high density of data (Carreira-Perpin´an, 2013). Unlike the classic K-means clustering approach, there are no embedded assumptions on the shape of the distribution nor the number of modes/clusters.

# Density Based Clustering of Application with Noise (DBSCAN)

Density Based, Eliminate noise points, performs clustering in remaining points when it works well, it is resistant to noise and can handle clusters of different shape and sizes. (Tan, Karpatne, Steinbach & Kumar, 2018)

# Expectation –Maximation (EM) using Gaussian Mixture Models (GMM)

A clustering algorithm used for knowledge discovery. It uses clustering to predict data models that can be used in other statistical analysis methods.

# Related Systems

LBS- location-based service (LBS) is a software-level service that uses location data to control features. As such LBS is an information service and has a number of uses in social networking today as information, in entertainment or security, which is accessible with mobile devices through the mobile network and which uses information on the geographical position of the mobile device. LBS can be used in a variety of contexts, such as health, indoor object search, entertainment, work, personal life, etc. LBS is critical to many businesses as well as government organizations to drive real insight from data tied to a specific location where activities take place. The spatial pattern that location-related data and services can provide is one of its most powerful and useful aspect where the location is a common denominator in all of these activities and can be leveraged to better understand patterns and relationships. LBS include services to identify a location of a person or object, such as discovering the nearest banking cash machine (ATM) or the whereabouts of a friend or employee. LBS include parcel tracking and vehicle tracking services. LBS can include mobile commerce when taking the form of coupons or advertising directed at customers based on their current location. They include personalized weather services and even location-based games. They are an example of telecommunication convergence. This concept of location-based systems is not compliant with the standardized concept of real-time locating systems (RTLS) and related local services, as noted in ISO/IEC 19762-5 and ISO/IEC 24730-1. While networked computing devices generally do very well to inform consumers of days old data, the computing devices themselves can also be tracked, even in real-time.

The main idea of the project is to improve the connectivity of the citizens to the emergency services with the help of the latest technology. With the rising demand for technology, people started to increasingly indulge in online stuff. The project will help people to be in touch with the services 24/7 and get rid of their problems and they even don't have to go the emergency services for doing tasks like booking an appointment in a hospital or lodging. They can do all the above-mentioned tasks and much more just by some clicks on their handsets and can easily communicate with the services from anywhere around. It is a smart city project and it aims at simplifying the lives of the citizens through technology. (Samantha, Khan and Sushmita, 2018)

The push notification is used in many web applications to send updates and alerts to the subscribed users. The notification is sent to the user’s device using cloud messaging server. Travel Helper uses push notification to alert the user to start the travel based on saved preference to reach a destination on time. Figure 5 shows the overview of cloud messaging components for push notification. (Murgolo, Hoffman, and Penzentadler, 2017)

# Data Security

The ability to link information from multiple sources, especially when that information is collected by different organizations, or for different purposes, raises significant legal and ethical questions (Crawford, 2014).

In the research of Asadi Someh, Breidbach, Davern, & Shanks (2016), they conceptualized big data analytics as social process and identified ethical issues it can cause. This includes data ownership, data control, awareness, privacy, data quality, data sourcing, data sharing or disclosure, algorithmic decision making and presentation.

**Synthesis of the Reviewed Literature and Studies**

A real-time data is a pre-requisite for the concept of real-time business intelligence in which reporting is done on live data and dashboards and graphs move with current data.

Several approaches were proposed by different researchers which include some new architectures and upgraded hardware equipment to achieve the concept of real-time data warehousing, out of which some approaches were successful but the most common reason behind the failure of some approaches was the data quality provided by the source system, as we know that data warehouse is based upon a process ETL in which the transformation phase is the most time consuming and important process, this process can be optimized if the data quality from the source systems can be improved.

So to achieve real-time data warehousing an approach that can be followed is that the quality of data should be improved so that the data from the source system can directly be sent to the data warehouse without a time and resource consuming process called transformation so that data can be present in the data warehouse for analysis and reporting and along with that hardware should also be improved so that the uploading time can also be reduced.

Previous researches have shown that in projects where the technology has been perceived as a failure, the problem does not usually lie with the technology itself, but rather with the way in which the technology was applied. It is often applied to the wrong business problem, at the wrong scale, with insufficient training, and planning, and with little or no thought to how users need to access the data. Some analysts’ statistics show that over 50% of warehouse projects fail to meet their stated objectives. In order to lessen the risks associated in those projects, the project must work closely with the business community that will benefit from the warehouse and also need to have a solid grounding from a financial return perspective. Data warehousing technology can benefit organizations at different levels and scales of implementation.

When it comes to processing of the data gathered, mean-shift is appeared to be helpful in this kind of research. The advantages of mean-shift clustering from other clustering algorithms are It makes no model assumptions (other than using a specific kernel), unlike Gaussian mixture models or K-means, It is able to model complex clusters having non-convex shape. The user need only set one parameter, the bandwidth, which has an intuitive physical meaning of local scale, and this determines automatically the number of clusters. This is often more convenient than having to select the number of clusters explicitly. It has no local minima, thus the clustering it defines is uniquely determined by the bandwidth, without the need to run the algorithm with different initializations. Knowing such concern, proper handling and an awareness of the issues surrounding the ethical use of collected data is vital for researchers entering this industry.

**Chapter 3**

**RESEARCH METHODOLOGY**

This chapter presents the methods and procedures used in pursuit of the development of the study. This deals with the research design, method of research, sources of data, research instrument, data gathering procedure, ethical considerations and proposed system architecture.

**Research Design**

The researcher chose a descriptive research design by the use of survey questionnaire because it best served to answer the questions and the purpose of the study. The survey research is one in which a group of people or items is studied by collecting and analyzing data from only a few people or items considered to be representative of the entire group. This means only a part of the population is studied and findings from this are expected to be generalized to the entire population by the use of a questionnaire.

The researcher also used a Data mining tool to simulate different clustering algorithms for identifying the best algorithm to be utilized in mining based on different criteria.

Lastly, Developmental method was used to develop a system that will serve as a solution to the existing problems discovered considering the ideal platform used.

**Method of Research**

The research is carried out by the following logical steps, these are:

1. Advice and guidance - First of all the researcher will seek advice and guidance from the church or mission group leader. By doing this the researcher will be informed about the processes, principles, and best practices.
2. Permission - The researcher will ask permission from the church or mission group. This step was essential because their data are confidential
3. Personnel - The researcher will inform their IT Personnel which would also serve as the system administrator of the proposed software.

**Sources of Data**

The researcher gathered the related data from the selection of tribes or people of the Philippines, religion, church locations, dialect and weather data from the web thru mining. The key contacts, volunteers, small groups and outreach events data are gathered from the specific church or mission group. The website of Philippine Statistics Authority and Open Data Philippines is also used as a source of data.

**Research Instrument**

A questionnaire and face-to-face interview were used to collect data from the involved church or mission group. The purpose of the questionnaire was to gather data that would help with the arrangement of the interview. Data such as the name and phone number of the contact person are necessary for arranging interviews. Accordingly, questions that are asked for such data were included in the questionnaire.

Clustering is a division of data into similar groups; each similar group is called a cluster. Object in a cluster are similar or close to each other. Clustering algorithms can be implemented via number of different approaches. The researcher conducted the simulation and comparison between different clustering algorithms on Rapidminer that offers Education Program License. This paper shows that study and comparison between mean shift, k-means method, and density-based method.

ISO 9126 is an international standard for the evaluation of software. The standard is divided into four parts which addresses, respectively, the following subjects: quality model; external metrics; internal metrics; and quality in use metrics. These characteristics are broken down into sub characteristics. Functionality is the essential purpose of any product or service. Once a software system is functioning, as specified, and delivered the reliability characteristic defines the capability of the system to maintain its service provision under defined conditions for defined periods of time. One aspect of this characteristic is fault tolerance that is the ability of a system to withstand component failure.

Usability only exists with regard to functionality and refers to the ease of use for a given function. This characteristic is concerned with the system resources used when providing the required functionality. The amount of disk space, memory, network etc. provides a good indication of this characteristic.

The researcher did not encounter any issues on the process and system (if any) that is required for the research, as it not affects the interview questions that are formulated accordingly.

**Data Gathering Procedure**

The aim and objectives of the research were achieved by accumulating a sample of church or mission group. At the outset, the researcher did not know which group would be willing to participate in the research. In order to get such a sample and gather data from them, these are the procedures used by the researcher divided into three steps:

1. Send letters to the Head of mission group, requesting them to participate in the research. A large population was chosen in order to increase the possibility of getting a representative sample. However, the chances of getting a representative sample would have been better if letters were sent nationwide.

2. Send a questionnaire, in order to gather data that would help with the arrangement of an interview. When replies were received from contact persons who were willing to take part in the research, the researcher will give them the Questionnaire and request them to complete and return the documents then suggest a date and time that would be convenient for them to participate in the interview.

3. Interview was conducted to gather the required data using the contact details obtained from questionnaires.

**Ethical Consideration**

The following ethical guidelines was put into place for the research period:

1. The protection of the privacy of research participants was ensured.
2. An adequate level of confidentiality of the research data was ensured.

**Data Case Analysis**

For the Statistical Algorithm, mean-shift clustering was used, and this is how mean-shift clustering works:

1. Selects a data point of interest.

2. Induced a circle of a specified radius around the point of interest.

3. Collected all data points within the circle and compute their mean.

4. Moved the center of the circle to the mean.

5. Repeated 3 & 4 until convergence. Each iteration will move "uphill" on the density gradient of the data distribution until it reaches the top of the hill (a local maximum).

6. Repeat 1-5 for all data points. Points that converge to the same local maximum are members of the same cluster. The number of clusters is the number of local maxima.

For higher dimensions, replace "circle" with sphere (3-D) or hypersphere with four (4) and higher dimensions. This algorithm depends on a choice of radius, which determines the granularity of the search for local maxima.

**Software Development**

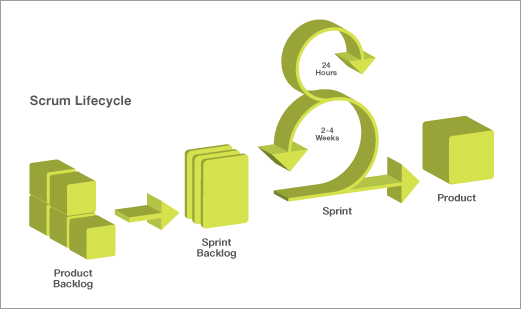
The following software was used for System Development:

* Database

MySQL – Uses text mining to gather useful information from the web between the source datasets to the real-time database.

* Interface for Transactions, Query, and Reports
  + - PHP
    - Javascript
* R and RapidMiner – data mining tools used to simulate different clustering algorithms.

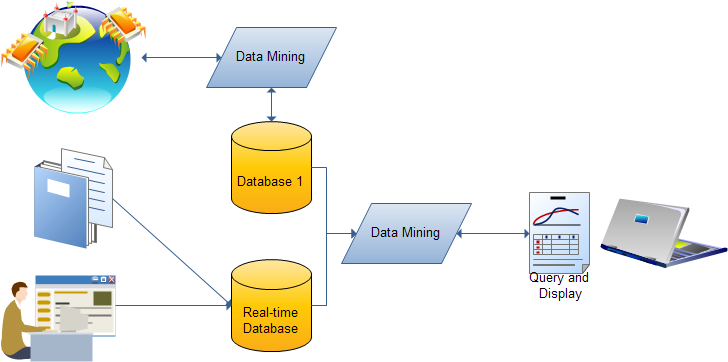
Figure 10. Agile Scrum Development Life Cycle



The figure above shows the agile scrum development life cycle which was used in the development of the proposed system. Agile scrum development is an iterative and incremental framework for managing software/product development. It is a full strategy where development team work as a unit to reach a common goal in a short period of time.

**Deployment Diagram**

Figure 10: Deployment Architecture of the YAVATAN



The figure above shows the source of database and data mining project which are the list of people of the Philippines from the web and then the key contacts, volunteers, small groups and outreach events of the mission group. Those data would be transformed and stored in the database, then it will be processed through to come up with the analytics using Mean Shift algorithm**.**

**System Architecture**

Figure 11: System Architecture of Yavatan

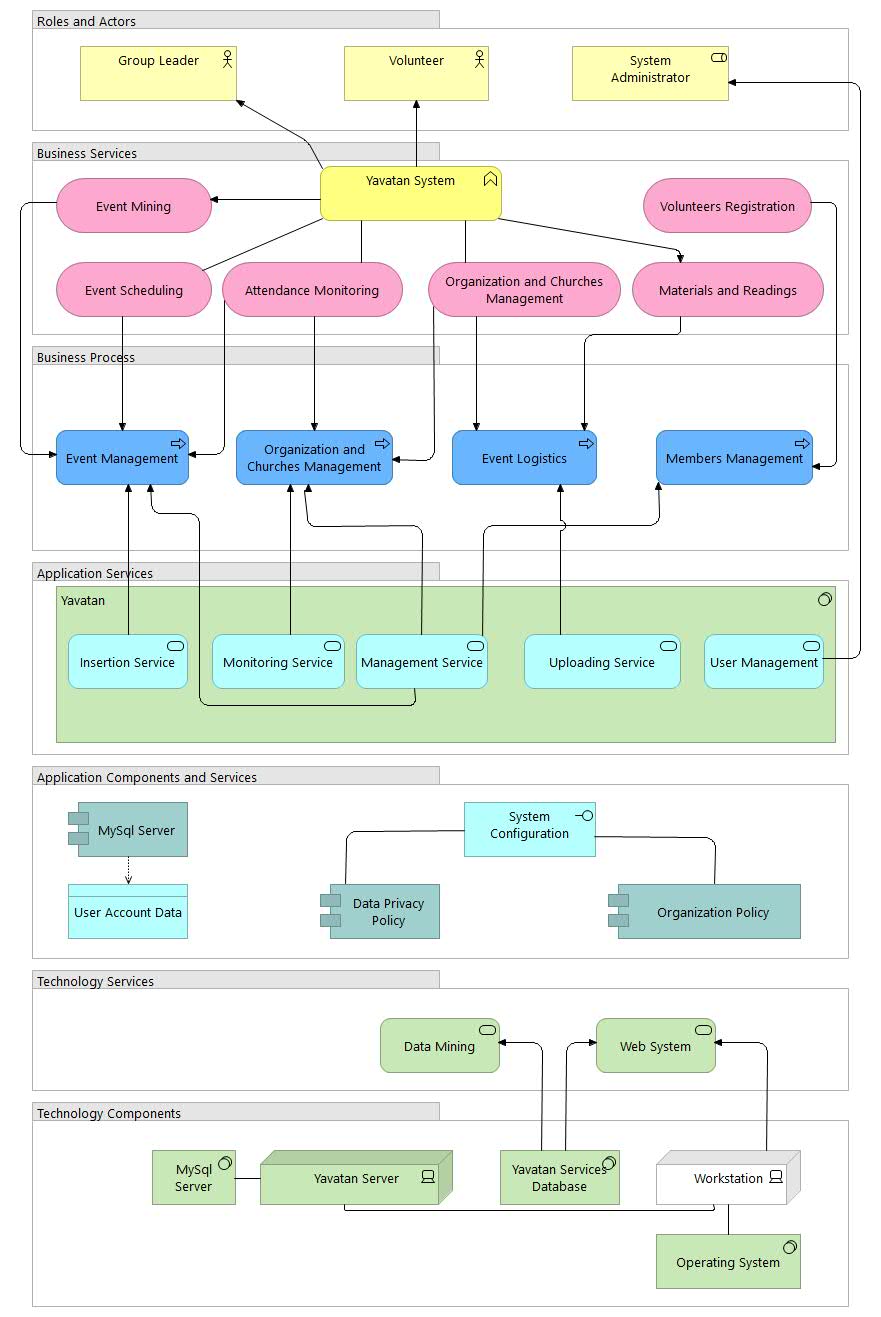


Figure 11 shows the components of the Yavatan system – It has 7 classifications consist of the roles and actors, business services, business process, application services, application components and services, technology services and technology components. There are 3 roles and actors in the system, these are the group leaders, volunteers (members), and the system administrator.

In the business services, Yavatan includes events mining, events scheduling, attendance monitoring, organization and churches management, volunteers’ registration, and the uploading of materials and readings.

The business process of Yavatan System Architecture contains 4 activities such as event management, organization and churches management, event logistics, and members management.

For the Application Services, it has 5 modules named as Insertion Service, Monitoring Services, Management Services, Uploading Services, and the User Management. The system used MySQL Server for storing the data of the users and creates a system configuration to update the Data Privacy Act and the organization policies of the governing bodies the will use the Yavatan system.

The system used technology services such as Clustering Algorithm and Web system. There are 4 technology components involved in the system, these are MySQL Server, Yavatan Server, Yavatan Services Database and the Workstation for the operating system.

**REFERENCES**

Russom, P. (2013). Managing big data. TDWI Best Practices Report, TDWI Research, 1-40.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314-347.

Mahboubi, H., Ralaivao, J. C., Loudcher, S., Boussaïd, O., Bentayeb, F., & Darmont, J. (2017). X-wacoda: an xml-based approach for warehousing and analyzing complex data. *arXiv preprint arXiv:1701.08033*.

Dewan, S., Aggarwal, Y., & Tanwar, S. (2014). Review on Data Warehouse, Data Mining and OLAP Technology: As Prerequisite aspect of business decision-making activity.

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling*. John Wiley & Sons.

Sathish, M. and Vyedhi, S. (2013). *Data warehousing Issues and Problems in the Real World* Entities, International Journal of Engineering Science Invention Volume 2, Issue 7

Jameel, F., Memon, F., Shaikh, P., & Memon, N. A. (2017). *Implementing a Business Intelligence Environment in Local Market of Pakistan*, Mehran University Research Journal of Engineering & Technology, Volume 36, No. 2

September 2012, Oracle White paper, Data Integration for Operational Data Warehousing

North, M., Thomas, L., Richardson, R., & Akpess, P. (2017). Data warehousing: A practical managerial approach. *Computer Science and Information Technology*, *5*, 18-26.

Runtuwene, J. P., & Tangkawarow, I. R. (2017). Vertical Information System: A Case Study of Civil Servant Teachers’ Data in Manado City. Indonesian Journal of Electrical Engineering and Computer Science, 6(1), 42-49.

Chhabra, R., Kumar, P., & Pahwa, P. (2016). An approach to Design Object Oriented Data Warehouse. International Journal of Research and Engineering, 3(3), 54-56.

Madden, S. (2016). Technical Perspective Mesa Takes Data Warehousing to New Heights.

Sultan, U. (2016). *REAL TIME DATA WAREHOUSING* (Doctoral dissertation, Institute of Business Administration).

Gorawski, M., & Gorawska, A. (2014, May). Research on the stream ETL process. In *International Conference: Beyond Databases, Architectures and Structures* (pp. 61-71). Springer International Publishing.

Jain, P., Gyanchandani, M., Khare, N., Singh, D. P., & Rajesh, L. (2017). A Survey on Big Data privacy using Hadoop Architecture. *International Journal of Computer Science and Network Security (IJCSNS)*, *17*(2), 148.

Wang, Y., & Wang, Z. (2016). Integrating Data Mining Into Managerial Accounting System: Challenges and Opportunities. *Chinese Business Review*, *15*(1), 33-41.

Ram, S. (2013). Benefits of using data warehousing and data mining tools.

Das, T. K., and Kumar, P. M. (2013). *Big Data Analytics: A framework for unstructured data analysis*. International Journal of Engineering Science & Technology, 5, 1, 153.

Agrawal Sr, M., Joshi Sr, A. S., & Velez, A. F. (2017). Best Practices in Data Management for Analytics Projects.

Roy, P. and Agarwal, R. (2015). *A Review on Technologies in Business Analytics*. International Journal of Science, Technology & Management, 4, 1.

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In Collaboration Technologies and Systems (CTS), 2013 International Conference on (pp. 42-47). IEEE.

Monteith, J. Y., Mcgregor, J. D., and Ingram, J. E. (2013). *Hadoop and its Evolving Ecosystem*. In IWSECO@ ICSOB.

Schacter, L. (2014). *Pontis Transitions From a Legacy Architecture to a Big Data Platform*. MapR Technologies Inc.

*PolyBase Guide*. (n.d.). Retrieved May 28, 2017, from Microsoft, <https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-guide>

Pereira, D. A., Ourique de Morais, W., & Pignaton de Freitas, E. (2018). NoSQL real-time database performance comparison. International Journal of Parallel, Emergent and Distributed Systems, 33(2), 144-156.

Pandey, S., & Shanker, U. (2016). Transaction execution in distributed real-time database systems. In Proc. Int. Conf. Innov. Inf. Embed. Commun. Syst (pp. 96-100).

Kumar, K. N., Akhi, K., Gunti, S. K., & Reddy, M. S. P. (2016). Implementing smart home using firebase. International Journal of Research in Engineering and Applied Sciences, 6(10), 193-198.

Höpken, W., Ernesti, D., Fuchs, M., Kronenberg, K., & Lexhagen, M. (2017). Big data as input for predicting tourist arrivals. In Information and communication technologies in tourism 2017 (pp. 187-199). Springer, Cham.

Schmid, O., Lisowska Masson, A., & Hirsbrunner, B. (2014). Real-time collaboration through web applications: an introduction to the Toolkit for Web-based Interactive Collaborative Environments (TWICE). Personal and Ubiquitous Computing, 18(5), 1201-1211.

Khatipov, R., Mazzara, M., Negimatzhanov, A., Rivera, V., Zakirov, A., & Zamaleev, I. (2018). Hikester-the event management application. arXiv preprint arXiv:1801.06400.

Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64-73.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51-59.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Osei-Bryson, K. M., & Barclay, C. (Eds.). (2015). Knowledge Discovery Process and Methods to Enhance Organizational Performance. CRC Press.

Abdul Rahman, F., Shamsuddin, S. M., Hassan, S., & Abu Haris, N. (2016). A Review of KDD-Data Mining Framework and Its Application in Logistics and Transportation. *International Journal of Supply Chain Management*, *5*(2), 77-84.

Sathiyamoorthi, V. (2017). Data Mining and Data Warehousing: Introduction to Data Mining and Data Warehousing. In *Web Data Mining and the Development of Knowledge-Based Decision Support Systems* (pp. 312-337). IGI Global.

Olson, D. L., & Wu, D. (2017). Predictive Models and Big Data. In *Predictive Data Mining Models* (pp. 95-97). Springer Singapore.

Redgert, R. (2017). Measuring Data Quality in a Data Warehouse Environment.

Thirumagal, R., Suganthy, R., Mahima, S., & Kesavaraj, G. (2014). ETL TOOLS IN DATA MINING A REVIEW.

Das, D. (2016). Data Mining of Indian Stock Market from April, 2015 to March, 2016 Using Curve Fitting Technique. *Power*, *653*(618), 034.

Van Craenendonck, T., & Blockeel, H. (2015, June). Using internal validity measures to compare clustering algorithms. In Benelearn 2015 Poster presentations (online) (pp. 1-8).

Irfan, M. S. D. A. M. (2013). WEB MINING AND ITS CATEGORIES–A SURVEY. International Journal Of Engineering And Computer Science, 2(04).

Saini, S., & Pandey, H. M. (2015). Review on web content mining techniques. International Journal of Computer Applications, 118(18).

European Union’s Horizon 2020 Research and Innovation Programme. (2016, June 1). *2016 OpenBudgets.eu: Fighting Corruption with Fiscal Transparency.* Requirements for Statistical Analytics and Data Mining*.* Retrieved from

http://openbudgets.eu/assets/deliverables/D2.3.pdf

*IBM SPSS Forecasting.* Retrieved May 21, 2017, from Smart Vision Europe website, http://www.sv-europe.com/wp-content/uploads/2011/08/IBM-SPSS-Forecasting.pdf

Gamboa, J. C. B. (2017). Deep Learning for Time-Series Analysis. *arXiv preprint arXiv:1701.01887*.

Davenport, T. H. (2013). Analytics 3.0. Harvard Business Review, 91(12), 64-72.

Mihai, D., & Mocanu, M. (2015). Statistical considerations on the k-means algorithm. *Annals of the University of Craiova-Mathematics and Computer Science Series*, *42*(2), 365-373.

Carreira-Perpinán, M. A. (2015). A review of mean-shift algorithms for clustering. arXiv preprint arXiv:1503.00687.

Tan, P., Steinbach, M., Karpatne, & A., Kumar, V. (2018). Introduction to Data Mining 2nd Edition. Pearson

Samanta, A., Khan , F., Sushmita, S. R. (2018). Public Information and Welfare System. *International Journal of Emerging Technologies in Engineering Research (IJETER) Volume 6, Issue 4. April.* EverScience Publications

Murgolo, F., Hoffman, M., & Penzenstadler, B. (2017). Efficient Travel Decision Making Using Web Application Based on MVC Architechture. *University of Mumbai, India.*

*Crawford, K. (2014). Big Data Stalking. Scientific American, 310(4), 14-14.*

Asadi Someh, I., Breidbach, C. F., Davern, M., & Shanks, G. (2016). Ethical implications of big data analytics. Research-in-Progress Papers, 24.

**Appendix 1**

**QUESTIONNAIRE**

PART I. RESPONDENT’S PROFILE

Name (Optional): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Church \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

How long have you been in the Mission Ministry?

🞏 < 1 year

🞏 1-3 years

🞏 4-6 years

🞏 7-10 years

🞏 more than years

Do you find it difficult consolidating reports?

🞏 Yes

🞏 No

Do you want to have an automated system specifically for Missions?

🞏 Yes

🞏 No

Would you like to use the system anywhere and anytime?

🞏 Yes

🞏 No

PART II. LEVEL OF AGREEMENT OF THE RESPONDENTS ON THE ISSUES ENCOUNTERED IN THE EXISTING SYSTEM.

Directions: Please respond to the checklist on your level of agreement on the issues encountered by the respondents in the existing system for recording, consolidating and summarizing the data for contacts, volunteers, small groups, materials, outreach events, and attendance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 5 – Strongly Agree | 1. – Agree | 3 – Slightly Agree | 2 – Disagree | 1– Strongly Disagree |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| RECORDING, CONSOLIDATING AND SUMMARIZING REPORTS | Strongly Agree | Agree | Slightly Agree | Disagree | Strongly Disagree |
| There is no single source of truth or no single repository of data of contacts, volunteers, small groups. |  |  |  |  |  |
| There is no mechanism to check and identify redundant data |  |  |  |  |  |
| Reports are consolidated after the submission of reports from other areas – not real time |  |  |  |  |  |
| Summarized data can not be drilled down as these are manually generated |  |  |  |  |  |
| Analytics are not available for relevant decision making by missionaries |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PREPARATION OF MATERIALS, EVENTS, AND KEEPING TRACK OF ATTENDANCE | Strongly Agree | Agree | Slightly Agree | Disagree | Strongly Disagree |
| Materials are not readily available because of the absence of a foreknowledge on the details of the people for mission |  |  |  |  |  |
| Limited inputs analysis is done due to dependencies on some matters which references are not available most of the time |  |  |  |  |  |
| Collated and consolidated attendance reports are difficult to filter and classified and grouped |  |  |  |  |  |
| Individual status of witnessed individuals during mission is not properly monitored for follow up |  |  |  |  |  |

PART III. LEVEL OF AGREEMENT ON THE ACCEPTATBILITY OF THE DEVELOPED SOLUTION.

Directions: Please respond to the checklist on your level of agreement on the acceptability of the developed solution in terms of functionality; reliability; usability; efficiency; maintainability; and portability.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Efficiency | Strongly Agree | Agree | Slightly Agree | Disagree | Strongly Disagree |
| Yavatan provides appropriate response time, processing time and throughput rates when performing the various functions (e.g uploading/ downloading files, online testing) under stated conditions |  |  |  |  |  |
| The system has an acceptable level of performance when large numbers of users are accessing the system at any one time |  |  |  |  |  |
| The Yavatan system can handle large documents |  |  |  |  |  |
| It can locate operations & information quickly |  |  |  |  |  |
| The system perform a sequence of operations (data input) with economy of motion |  |  |  |  |  |
| Reliability |  |  |  |  |  |
| Maturity – Yavatan system is capable of accepting work arounds on functionalities which are not present |  |  |  |  |  |
| Fault Tolerance – Yavatan has the ability to maintain a specified level of performance or continue  functioning in the event of software fault(s) |  |  |  |  |  |
| Recoverability /Error Handling – Yavatan allows users to take corrective action once an error has been recognized |  |  |  |  |  |
|  |  |  |  |  |  |
| Usability |  |  |  |  |  |
| Understandability – Yavatan uses standards in its UI and documentations |  |  |  |  |  |
| Learnability – Yavatan displays straightforwardness of performing tasks and maintains a level of pleasant User experience |  |  |  |  |  |
| Operability – The system exhibits clarity and consistent instructions |  |  |  |  |  |
| Attractiveness – Yavatan offers richness in the pleasant presentation of interface |  |  |  |  |  |
| Functionality |  |  |  |  |  |
| General Features |  |  |  |  |  |
| Yavatan capable of providing functions which meet the stated and implied basic needs of professionals under specified conditions of usage (what the software does to meet needs) |
| The system ensures a secure set of user privileges (role-based access control), which determine  permission levels (creation and updating data) that  users need to control, manage, and update content. |  |  |  |  |  |
| The system successfully implements identified algorithms for better performance |  |  |  |  |  |
| Maintenance of the following which are important for missions are well taken by Yavatan |  |  |  |  |  |
| Mission Contacts |  |  |  |  |  |
| Volunteers for mission activities |  |  |  |  |  |
| Small groups |  |  |  |  |  |
| Materials Management |  |  |  |  |  |
| Outreach events Monitoring |  |  |  |  |  |
| Planning/Scheduling (Category: Student-Led, Young Professionals, Community) |  |  |  |  |  |
| People/Tribes |  |  |  |  |  |
| Language |  |  |  |  |  |
| Logistics |  |  |  |  |  |
| Weather API |  |  |  |  |  |
| SMS notification to key contacts within the area |  |  |  |  |  |
| Attendance (encoding of new contacts) |  |  |  |  |  |
| Follow up SMS (By Group) |  |  |  |  |  |
| Automatic Suggestions by Yavatan |  |  |  |  |  |
| Suggestion of Events to Volunteers within the area (location -based recommendation) |  |  |  |  |  |
| Suggestion of Contacts to Group Leaders |  |  |  |  |  |
| Suggestion of Reading Materials to Contacts |  |  |  |  |  |
| Generation of Missions’ Reports and Analytics |  |  |  |  |  |
| Number of Key Contacts with Mapping |  |  |  |  |  |
| Number of Key Volunteers with Mapping |  |  |  |  |  |
| Number of events conducted with Mapping |  |  |  |  |  |
| After Outreach details and counts of people on Gospel conversation, Exposed, Prayed to receive, and follow up |  |  |  |  |  |

PART IV. LEVEL OF AGREEMENT ON THE USE OF PREDICTIVE ANALYTICS FOR YAVATAN.

Directions: Please respond to the checklist on your level of agreement on the acceptability of having the ff. predictive analytics in YAVATAN’s future version.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PREDICTIVE ANALYTICS | Strongly Agree | Agree | Slightly Agree | Disagree | Strongly Disagree |
| A report that predicts the number of attendees per area drilled down to the lowest barangay or tribe |  |  |  |  |  |
| A report that predicts the number of people who will be ready to accept the Gospel depending on the religious affiliations |  |  |  |  |  |
| A report that predicts the kinds of material for production for the yearly mission activities |  |  |  |  |  |
| A report that predicts the number and types of volunteers needed for the next year’s activities |  |  |  |  |  |
| A report that predicts which area to avoid or be cautious about within the year |  |  |  |  |  |

---- Thank you---