EXPLORING THE BIG DATA ANALYTICS TOOLS FOR STRUCTURED AND UNSTRUCTURED DATA

Kyaw Kyaw Min[1], Pwint Phyu Khine[2], Soe Mya Mya Aye[3]

Abstract

Today, as technologies advance, so does the data usage. This makes many organizations need to handle a lot of data creating huge data collection which affects all sectors. A larger amount of data assumingly could give a better output however manipulating such amount of data with different structures becomes a challenge. A specific generation of analytical tools is increasing in use to handle such an increase in the data, structured and unstructured data. With appropriate big data analytical tools, an organization or individual can manipulate large amounts of diverse data to solve specific problems. The major objectives of the research are to analyze some popular big data analytics tools for structured and unstructured data using case studies to support better analytics.

Keywords: Big Data, Big Data Analytics

Introduction

Big data can be defined as datasets in which a large amount of volume requires more advanced and sophisticated technologies to obtain, store, transform, manage, and analyze for extracting useful valuable information. According to Waller & Fawcett (2013) define big data as datasets that are too large for traditional data processing systems and therefore require new technologies to process them. Sabharwal and Shah Jahan Miah (2021), describe it as the traditional enterprise machine-generated data and social data. The extraction of valuable information from a large amount of data is called big data analytics. Big data analytics is the process used to examine a large amount of different structured datasets to extract most of the useful information to uncover unknown correlations, hidden patterns, market trends, customer preferences and which help organizations to make better business decisions.

The first three characteristics of big data are volume, velocity, and variety. Additional characteristics of big data are variability, veracity, visualization, and value. Understanding the characteristics of Big Data is the key to learning its usage and application properly. Among Big Data characteristics, Variety is one of the study aspects which can further be classified into structured data, and unstructured data. Big data is defined as a huge volume of structured and unstructured data. Semi-structured data is considered to be included in unstructured data although another opinion is to consider semi-structured data as standalone data.

The need of storing and analyzing these kinds of diverse data is a big challenge. Big data challenges encompass data collection, data storage, data analysis, search, sharing, transfer, and querying, updating, information, and data source as data are differ in structures, types, sources and tools and techniques to handle them. According to Katal (2013), its main functionalities are to provide different corporate to enhance quicker access and smart decision in the importance of database management. As data variety differ, so are the tools used for big data analytics. Big data analytics tools help organizations and businesses in saving time and cost providing faster and better insights to make decisions. With appropriate big data analytical tools, an organization or

[1] Department of Computer Studies, University of Yangon
[2] Department of Computer Studies, University of Yangon
[3] Department of Computer Studies, University of Yangon

individual can obtain, store, transform and analyze large amounts of data to solve specific problems. The major objectives of the research are to understand some of the most popular Big Data Analytics Tools used for Structured and Unstructured Data, and to analyze some case studies using these tools to support better analytics. This research describes the characteristics of three software tools for big data analytics based on data variety characteristics-SPSS for structured data, Tableau for semi-structured and Python for unstructured data.

Types of Big Data

Regardless of effort to store and analyze traditional data, most big data is unstructured (or) not structured in nature, which requires different techniques and tools to store, process, and analyze. The structured data and unstructured data which differ based on data variety specific characteristics are described detail in Table 1.

Table 1: Example of Structured and Unstructured Data

Structured Data	Unstructured data
Highly organized information Define how they will be stored with rigid data types and structures	Data with different specific structure that could not be stored in a traditional row and column format.
Mostly relational structure with Entity relationship (Tables as schema, rows as records, columns as fields) e.g. transaction data, traditional RDBMS (Relational Database Management System)	Semi-structured data e.g. data files such as spreadsheets in CSV (Column Separated Values), location (Longitude, Latitude), log files, XML files. Unstructured data e.g. (Images, audio, videos, log files, etc.)
Databases, data warehouses	Data lakes, non-relational databases
Predetermined Formats with fixed structures for creating a model defined data type and some restrictions on the data.	An array of different formats with undefined data structures

To analyze the unstructured data advanced analytical knowledge is needed. Sometimes, it is very tough to manage and analyze completely due to different forms of inherent structures. Semi-structured data is a form of structured data that does not conform to the formal structure of data models associated with relational models or other forms of data tables. This semi structured data has been added to unstructured data in the research.

Big Data Analytics

Big data analytics is a process of inspecting, differentiating, and transforming data with different data structure characteristics to extract useful information. Big data analytics has emerged as an important tool for supporting managerial decision-making. Jeble et al (2018) suggests that big data discovery efforts can reveal previously unknown findings which can result in insights that are helpful for managerial decision-making. Khan and et al (2014) have presented that Big Data Analytics is a strategy used to analyze huge information sets containing various qualities of information.

Three Levels of Big Data Analytics

Data analytics addresses information obtained through observation, measurement, or experiments about an occurrence of interest. Data analytics aims to extract as much information as possible that is relevant to the subject under consideration. Further, classified the entire field of big data analytics can be classified into the three-level based on the depth of analysis -

(a) descriptive analytics- to use data to understand past and present
(b) predictive analytics- to analyze current and historical facts to make predicting future events
(c) prescriptive analytics- to use optimization techniques and suggests what should be done to optimize results, it addresses decision making and efficiency

Big Data Analytics Tools

There are different types of Big Data Analytics tools available under data analytics that help to improve the process such as data analysis, data cleansing, data mining, data visualization, data integration, data storage, and management. It is essential to know about which data analytical tools match the best for given data. There were nearly more than a hundred tools available to perform various steps of analytics such as SPSS, Spark, Talend, Khime, Tableau, PowerBI, Python, R, Julia, Hadoop, Hive, Sqoop, Mahout, Storm, Teradata, HBase, SAS, RapidMiner, and so on. Selecting a tool from the vast number of existing tools is a challenging task. When choosing analytical tools to use, it is useful to grasp the dimension of the software's market share and popularity.

To demonstrate the different dataset usage, two datasets are used. Employee Dataset Real World Dataset for Civil Service Academy (Lower Myanmar) Trainees For Structured Data and Semi-Structured Data), and Olivetti faces dataset from the scikit-learn library For Unstructured Data. Location such as city from real world employee dataset is used to demonstrate the Semi-structured data usage.

There are various analytics tools for analytics such as low-cost light weight spreadsheets such as excel. Data Analytics is the process of analyzing datasets to draw results, based on information they get. It is popular in commercial industries, scientists, and researchers to make a more informed business decision. There are many Big Data Analytics Tools for Structured Data such as SPSS, SAS, Khime, RapidMiner and so on. Big data analytical tools for unstructured data are more diverse and complex. Tools such as Tableau, PowerBI for esthetic analytical visualization, Programming tools such as Python, R, Julia for complex and advanced data analytics, Hadoop for file processing, Pig, Sqoop for data transfer, Hive for data warehouse and management, and so on. With big data, more commercialized or open-source tools are needed to use for analytics. Commercialized tools such as SPSS, Tableau, etc. are developed specially for the analysis of big amount of data. Open source tools could offer the advanced data analytics due to their resourceful libraries. The process of Big Data analytics helps organizations to better understand the information which is present within the sets of data. This research is conducted particularly based on three big data analytical tools, namely SPSS for structured data, Tableau for semi-structured data, and Python as unstructured data analytics.

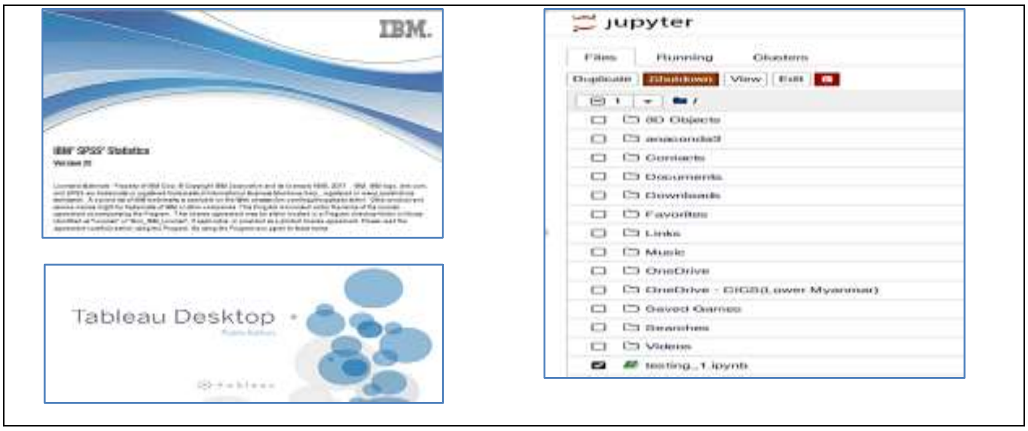


Figure 1: Big Data Analytics Tools “SPSS, Tableau and Python”

Big Data Analytics Tools for Structured Data

SPSS

SPSS (Statistical Package for the Social Sciences) from IBM Statistics, is a software package used for the analysis of statistical data. The origin of SPSS is in the field of social sciences; however, its use has further expanded into a wide variety of science for the analysis of data. SPSS become a popular data analytics platform especially for structured data. The most widely used powerful algorithms such as Decision Tree, K-Nearest Neighbor, even some neural network algorithms have been included in in-built solutions.

Data Processing for SPSS

Scaling including normalizing and transforming data such as binning could easily performed in SPSS. It is also possible to perform dimensionality reduction although data must be conformed with the rigid relational structure. As a use case for to perform data Processing in SPSS, real-world dataset of employee information who attended the Central Institute of Civil Service - Lower Myanmar for a period of time is used. E.g. variable Gender (Sex) into Male or Female (Binary). Table 2 describes some of the variables (fields) that are preprocessed for further analysis. Scaling including normalizing and transforming data such as binning could easily be performed in SPSS. Possible to perform dimensionality reduction although data must be confirmed with the rigid relational structure. The relevant data fields required for analyzing are first categorized such as variable Gender (Sex) into Male or Female (Binary). Table 2 describe some of the variables (fields) that are preprocessed for further analysis.

Table 2: Type the Field on Data View in SPSS

No	Variable Name	Type	Note
1	Sex	Categorial (Binary)	Male=0, Female=1
2	Age Group	Categorical (Numeric)	21-25 = “1”, 26-30 = “2”, 31-35 = “3”, 36-40 = “4”, 41-45 = “5”, 45+ = “6”

3	Qualification	Categorical	Bachelor = “1”,
			Honest = “2”,
			Master = “3”,
			M.Res = “4”,
			Ph.D = “5”,
			Diploma = “6”,
			Other = “7”

4	Rank	Ordinal	Officer = “1” Head of Branch = “2” Assistant Manager = “3” Assistant Engineer = “4” Head of Workshop = “5” Coaching Officer = “6”

SPSS provide Data View for data values, and variable view for data definitions. Qualification should be ordinal data. However, there is an assumption a degree holder might have related or unrelated diploma degree. Therefore, data is categorized as nominal (categorical data). Rank is ordinal data although there may have some other values with different name such as officer(administration) and assistant lecturer (university) has the same ordinal value. Figure 2 explore some features (variables) in above dataset as case study. Average age of the trainees is around the age of 30 i.e. regardless of rank, trainees in this age are more willing to get the diploma Figure 2(a). Scatter plot for fields Age, Qualification, Department are used to visualize each employee as a dot on three-dimensional predictor space to explore the dataset. Male and female employees are represented with blue and red dot color respectively. It indicates that females are more interested in getting the diploma for career advancement Figure 2(b).

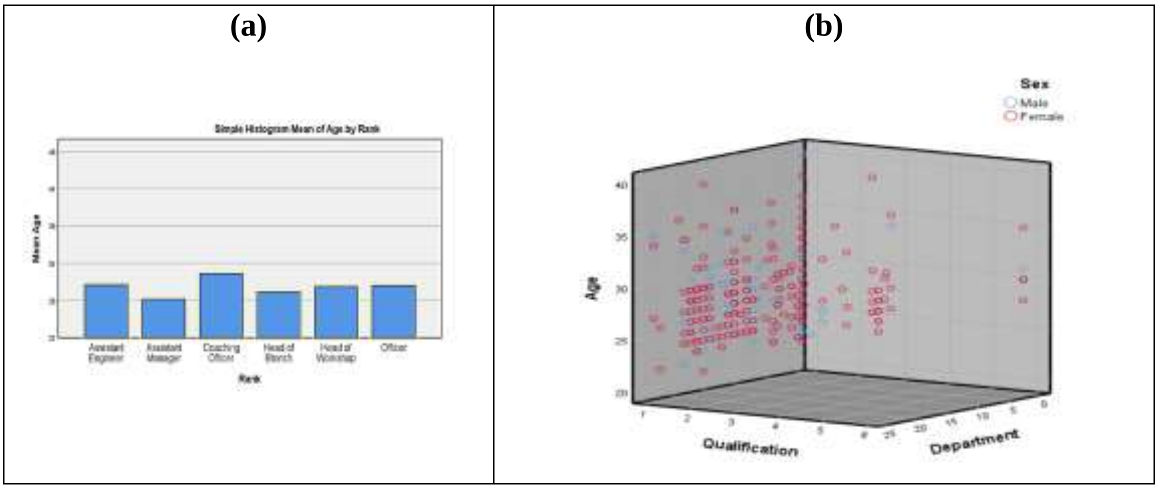


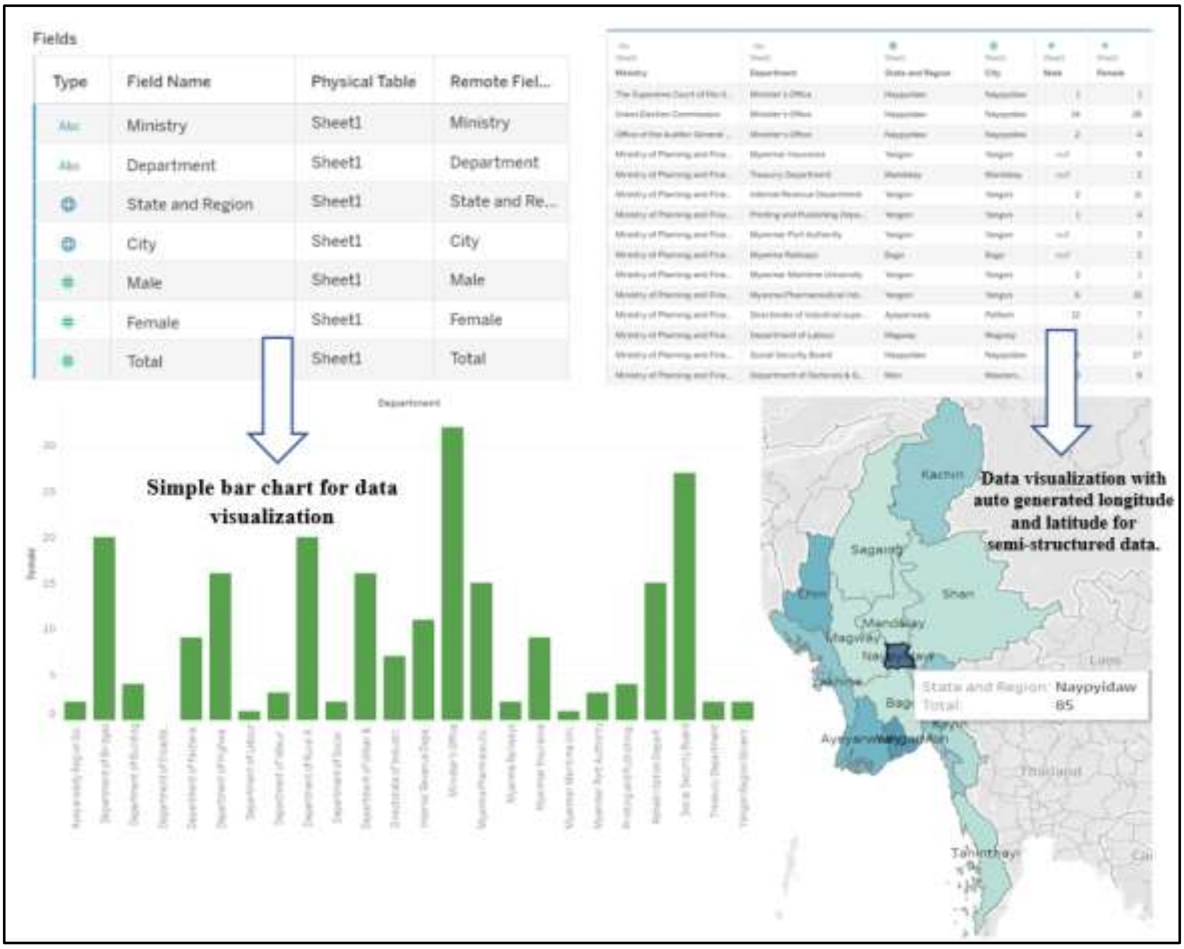
Figure 2: Exploratory analysis for structured data in SPSS

Tableau

Tableau is a visual-oriented data analytical tools for business intelligence and analytics with a variety of integrated products for visualizing and understanding data. Unlike standalone SPSS which only support batch processing, Tableau could support both batch and online processing (data could be streamed). Another prominent characteristic is Tableau could support the inherent semi-structured data.

Data Processing for Tableau

Tableau can handle different types of data structures – structured e.g. Microsoft Excel, Access, many renowned SQL databases formats, and semi-structured such as JSON, XML, etc. It also handles unstructured data, however, both SPSS and Tableau are not very good at flexibility as they could only allow limited user defined routines. Collected data is loaded or connected through Tableau for further analysis. Tableau can analyze not only structured but also semi-structured data as it can automatically translate some of the data feature. For example, Location data which has little or no sense except identity could be interpreted in Tableau as semi-structured data (with longitude and latitude). The results obtained after such data analysis can be generated in the form of a more sophisticated visualization report generated by Tableau. Figure 3 shows the relation of State and Region and number of Civil Services.



Fields

Type	Field Name	Physical Table	Remote Fiel...
Abc	Ministry	Sheet1	Ministry
Abc	Department	Sheet1	Department
🌐	State and Region	Sheet1	State and Re...
🌐	City	Sheet1	City
#	Male	Sheet1	Male
#	Female	Sheet1	Female
#	Total	Sheet1	Total

Simple bar chart for data visualization

Data visualization with auto generated longitude and latitude for semi-structured data.

State and Region: Naypyidaw
Total: 85

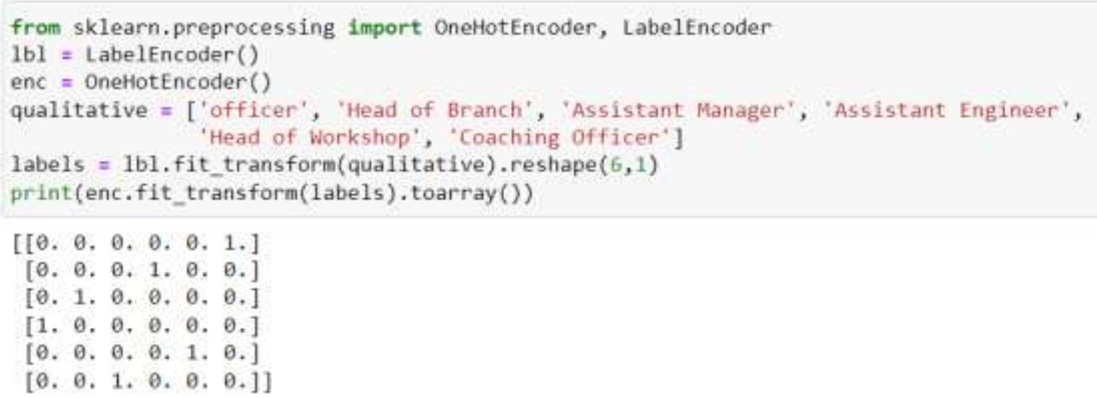
Figure 3: The relation of State and Divisions and number of Civil Services

Big Data Analytics Tools for Unstructured Data

Data Analytics is the process of analyzing datasets to draw results, based on information they get. It is popular in commercial industries, scientists, and researchers to make a more informed business decision. There are many Big Data Analytics Tools for Unstructured Data such as Python, R, Julia, Power BI, including tools from Hadoop ecosystems such as Hadoop, Hive, Pig, Spooq, other big data analytic tools such as Spark, Mahout, and so on. Among them, Python, a very widespread use for big data analytic and management is explored.

Python

Python is a computer programming language often used to build websites and software, automate tasks, and conduct data analysis. Python is a general-purpose language, meaning it can be used to create a variety of different programs and isn't specialized for any specific problems. Python can be considered a tool, analyze structured data, semi structured and unstructured data. Programming knowledge is required to use python. In addition, in-depth understanding is also required. However, the most used algorithms can be included in the famous python machine learning and data visualization libraries and can be used with a few lines of code providing required flexibility. For example: Data preprocessing for analysis in python is described in following figure 4. Unstructured text data could easily be encoded for further analysis.

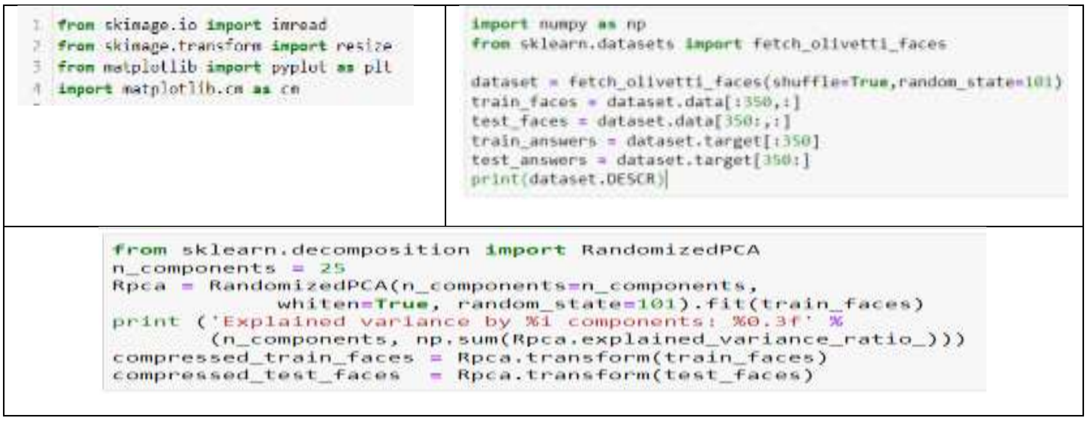


```python
from sklearn.preprocessing import OneHotEncoder, LabelEncoder
lbl = LabelEncoder()
enc = OneHotEncoder()
qualitative = ['officer', 'Head of Branch', 'Assistant Manager', 'Assistant Engineer',
               'Head of Workshop', 'Coaching Officer']
labels = lbl.fit_transform(qualitative).reshape(6,1)
print(enc.fit_transform(labels).toarray())
```

```
[[0. 0. 0. 0. 0. 1.]
 [0. 0. 0. 1. 0. 0.]
 [0. 1. 0. 0. 0. 0.]
 [1. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 1. 0.]
 [0. 0. 1. 0. 0. 0.]]
```

Figure 4: data preprocessing in python

The weakness of both SPSS and Tableau is their inability to handle with unstructured data. To demonstrate the use case of unstructured data, Olivetti face dataset from scikit-learn library is used. Data Visualization could be done easily by importing necessary libraries to classify an image. First, Python handle unstructured data image dataset by interpreting the image data into data set with numerical values. Unstructured image data is transformed, processing, and refined to be able to use for training. Dataset containing the image is simply split into training set and test set to analyze the face of an image. Then, randomized Principal Component Analysis algorithm is used in training of the data.



```
from skimage.io import imread
from skimage.transform import resize
from matplotlib import pyplot as plt
import matplotlib.cm as cm
```

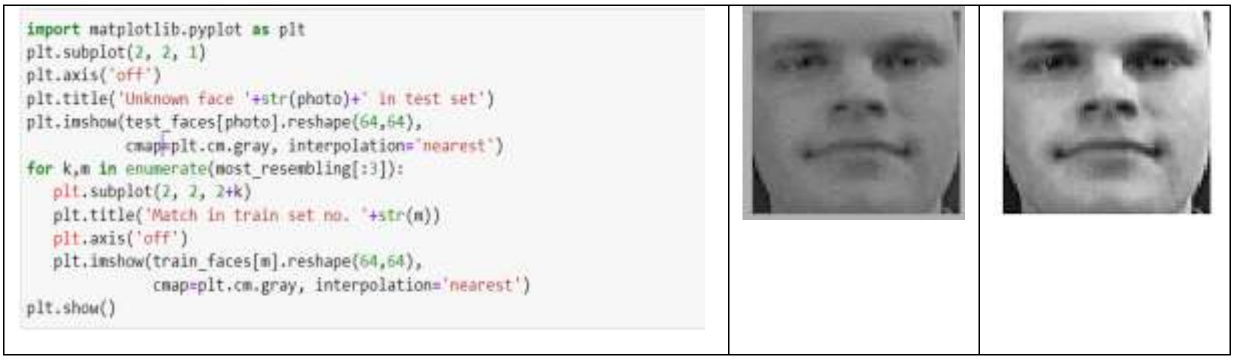
```
import numpy as np
from sklearn.datasets import fetch_olivetti_faces

dataset = fetch_olivetti_faces(shuffle=True,random_state=101)
train_faces = dataset.data[:350,:]
test_faces = dataset.data[350:,:]
train_answers = dataset.target[:350]
test_answers = dataset.target[350:]
print(dataset.DESCR)
```

```
from sklearn.decomposition import RandomizedPCA
n_components = 25
Rpca = RandomizedPCA(n_components=n_components,
            whiten=True, random_state=101).fit(train_faces)
print ('Explained variance by %i components: %0.3f' %
        (n_components, np.sum(Rpca.explained_variance_ratio_)))
compressed_train_faces = Rpca.transform(train_faces)
compressed_test_faces  = Rpca.transform(test_faces)
```

Figure 5: Image face dataset import program in python

After data is trained, unstructured image data can be interpreted and visualized easily. However, these visualizations could be impossible or difficult in SPSS and tableau. The demonstration is done as follow. First, a script is written to find a face from test data set that is most similar in the trained data. The image is searched in the training dataset set. Necessary image processing steps are performed in both the training set and test set such as changing into the greyscale image to reduce the inference of color in determining the similarity. he most similar image from the train set is retrieved as a result.



```
import matplotlib.pyplot as plt
plt.subplot(2, 2, 1)
plt.axis('off')
plt.title('Unknown face '+str(photo)+' in test set')
plt.imshow(test_faces[photo].reshape(64,64),
           cmap=plt.cm.gray, interpolation='nearest')
for k,m in enumerate(most_resembling[:3]):
   plt.subplot(2, 2, 2+k)
   plt.title('Match in train set no. '+str(m))
   plt.axis('off')
   plt.imshow(train_faces[m].reshape(64,64),
              cmap=plt.cm.gray, interpolation='nearest')
plt.show()
```

Figure 6: Image classification program in python

Results and Discussion

This paper first describes the literature review of big data, data variety based on structured and unstructured data, the nature of data analytics, and introduce different big data tools. Big Data Analytics Tools is one of the most important tools to support decision-making in every organization, business, and Research field. The success of an organization is critically linked to effective analytics. These Big Data Analytics Tools provide effectiveness and are essential for all organizations and businesses.

Table 3: Comparison of Big Data Analytics Tools

Tools	Data Variety	Processing	Skill Requirements
SPSS	Structured Data	Batch processing, Standalone	Statistics, Data Visualization, Domain knowledge
Tableau	Structured, Semi-structured	Batch, and Online Analytical, Both Standalone and Crowd sourced	Data Analytical Skill, Statistics, Understanding of SQL and NoSQL data stores
Python	Structured, Semi-structured, Unstructured	Could support all degrees of processing-batch, complex event processing, near real time, real-time, Stream processing. Required advanced knowledge	Programming Skill, Data Science, Database knowledge-both SQL and NoSQL, Statistical Knowledge, Understanding of Advanced algorithms and data structures

Analytics Tools for Structured data explored. In SPSS, Structured Data Field can be analyzed to get the required results. SPSS provides data analysis for descriptive and exploratory statistics providing easy data transformation, analysis, and visualization. However, it is rigid and could only support batch processing. Tableau can support both structured data and unstructured data (semi-structured data). Tableau support online processing by allowing distributed computing. Python, a programming language, can also be used as a tool for big data analytics although it requires the advanced understanding of data structures and algorithms. There are other data analytical oriented programming languages such as R and Julier which also provide famous mathematical, statistical, data analytical and data visualization. Table 3 describes the comparison of above explored three data analytical tools -SPSS, Tableau, and Python.

Conclusion

In this research, both real-world and pre-built datasets are used to explore the potential of big data analytics tools SPSS for structured data, Tableau for semi-structured data, and Python for unstructured data in areas such as data identification, data storage, data filtering and extraction, data analysis, and data visualization. Big Data tools can provide insights and benefits to decision-makers for better decisions in various areas.

SPSS provide structured data fields can easily analyze, however, it is rigid and could only support batch processing. Tableau can support both structured data and unstructured data (semi-structured data), and supports online processing by allowing distributed computing, however, it is limited as a preparatory analytical tool. Python, a programming language, can also be used as a tool for big data analytics and able to handle all structured, semi-structured, and unstructured data, however it requires an advanced understanding of data structures and algorithms, mathematical, statistical, data analysis, and data visualization methods.

Big data analytics is of great significance in this age of data overflow and can provide insights and benefits to decision-makers in various areas different sectors and industries, such as civil service management, healthcare, retail, telecommunication, manufacturing, etc. Future research will be conducted an in-depth look, the pros and cons of these analytics tools, and experimental analysis about the currently explored big data analytical tools. Some rapidly renowned tool such as Power BI and other big data tools will also be explored. By using Big Data Analytics Tools, faster and better decision-making for organizations and businesses can be achieved with reduced cost and time.

Future research will be conducted with an in-depth look, at the pros and cons of these analytics tools, and experimental analysis of the currently explored big data analytical tools. Some rapidly renowned tools such as Power BI and other big data tools will also be explored. By using Big Data Analytics Tools, faster and better decision-making for organizations and businesses can be achieved with reduced cost and time.

Acknowledgments

I would like to thank Dr. Nwe Nwe Win, Professor, and Head (Retired), Department of Computer Studies, University of Yangon, for her guidance to perform this research.

I extend my wholehearted thanks to Dr. Soe Mya Mya Aye, Professor and Head, Department of Computer Studies, University of Yangon for her kind and valuable advice and help. I am also thankful to Dr. Khin Sandar Myint, Associate Professor and Daw Pwint Phyu Khine, Demonstrator, Department of Computer Studies, University of Yangon for their interest and advice given on the research.

References

Jeble, S., Sneha Kumari, Yogesh Patil, (2018), “Role of Big Data in Decision Making”, Operation and Supply Chain Manaagement, Vol. 11, No. 1, 2018, pp (36-44), ISSN 1979-3561/EISSN 2759-9363.

Katal, A., M. Wazid and R.H. Goudar, (2013), “Big Data: Issues challenges tools and Good Practices”, Sixth International Conference on Contemporary Computing (IC3), pp(404-409), http://dx.doi.org/ 10.1109/ ic3.2013.6612229.

Khan, N., et. al., (2014), “Big Data: survey, technologies, opportunities, and challenges”, Scientific World Journal. doi: 10.1155/2014/712826.

Sabharwal, R. and Shah Jahan Miah, (2021), “A new theorical understanding of Big Data Analytics Capabilities in organizations: A thematic analysis”, Journal of Big Data 8, Article number: 159(2021).

Waller, M.A., Stanley E. Fawcett, (2013), “Data Science, Predictive Analytics, and Big Data”, Journal of Business Logistics, Vol. 34, Issue 2, pp (77-84).