Data Analytics Guideline

Prepared for
INTOSAI Working Group on IT Audit

INTOSAI
Working Group on IT Audit

2019

INTOSAI
Goal Chairs Collaboration
PSC – CBC – KSC

Quality Assurance Certificate of the

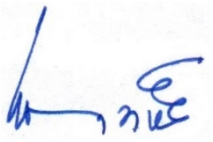
Chair of INTOSAI Working Group on Information Technology Auditing (WGITA)

This is to certify that ***Data Analytics Guideline,*** which is placed at **level 2 (two)** of Quality Assurance as defined in the paper on "Quality Assurance on Public Goods developed outside Due Process" approved by INTOSAI Governing Board in November 2017, has been developed by following the Quality Assurance processes as detailed below:

i. The project proposal was developed by the team in consultation with INTOSAI WGITA Members;
ii. The project was discussed during the 26th WGITA Meeting in Seoul in 2017, the 27th Meeting in Sydney in 2018 and the 28th Meeting in Fiji in 2019;
iii. The draft project output was circulated among team members and WGITA members; and was exposed for 45 days (from 10 May 2019 to 24 June 2019) for comments.

The product is consistent with relevant INTOSAI Principles and Standards. The structure of the product is in line with the drafting convention of non-IFPP documents.

The product is valid until **September 2025** and if not reviewed and updated by **September 2025**, it will cease to be a public good of INTOSAI developed outside the Due Process.



Rajiv Mehrishi
Chair of INTOSAI Working Group on Information Technology Auditing

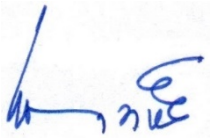
INTOSAI
Goal Chairs Collaboration
PSC – CBC – KSC

Quality Assurance Certificate of the

Chair of INTOSAI Knowledge Sharing and Knowledge Services

Based on the assurance provided by the Chair of INTOSAI **Working Group on Information Technology Auditing** (WGITA) and the assessment by the Goal Chair, it is certified that ***Data Analytics Guideline*** which is placed at **level 2 (two)** of Quality Assurance as defined in the paper on "Quality Assurance on Public Goods developed outside Due Process" approved by INTOSAI Governing Board in November 2017 has been developed by following the Quality Assurance processes as detailed in the Quality Assurance Certificate given by the Working Group Chair.

The product is valid until **September 2025** and if not reviewed and updated by **September 2025**, it will cease to be a public good of INTOSAI developed outside the Due Process.



Rajiv Mehrishi
Chair of INTOSAI Knowledge Sharing and Knowledge Services Committee

Table of Contents

1. Introduction

1.1. Purpose

This document is intended to provide general guideline on data analytics implementation in auditing. It is organized to provide data analytics concept and an outline of generic processes of implementing the data analytics practices for users who already familiar with traditional audit methodology. Several good practice standards for data analytics have been seamlessly adopted into the guidelines, without focusing on particular technology and methodology.

1.2. Data Analytics

There is no consensus on the definition of Data Analytics (DA) since it is frequently interchangeable with Data Analysis. The relation between Data Analytics and Data Analysis is somehow similar to the relation between Informatics and Information. Data Analysis is the broader term of Data Analytics. Data Analytics is the part of Data Analysis that requires the intensive use of computation power and scientific approaches.

AICPA has defined data analytics in audit as "the science and art of discovering and analyzing patterns, identifying anomalies, and extracting other useful information in data underlying or related to the subject matter of an audit through analysis, modeling, and visualization for the purpose of planning or performing the audit."[1]

The main goal is to enhance audit quality, in particular, to respond to a business environment characterized by pervasive use of IT, increased availability of large amounts of data, and increased use of IT-based data analytic tools and techniques by audited entities of all types and sizes.

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains.

Data analysis often falls into two phases: exploratory and confirmatory. The exploratory phase isolates patterns and features of the data and reveals them forcefully to the analyst. Through an exploratory analysis, the analyst may uncover patterns that represent irregularities in the dataset. These patterns may lead the analyst to revise the model if the model does not meet the reasonable assurance. As a consequence, the process will be repeated until the model satisfied the requirement.

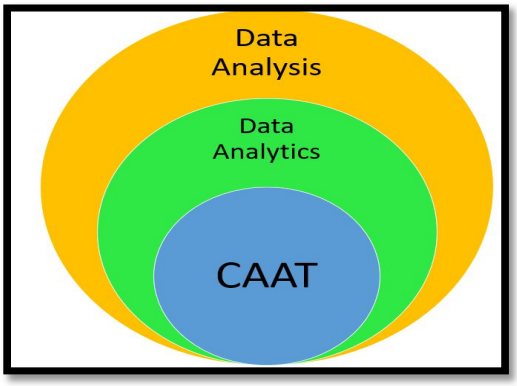
In contrast, confirmatory data analysis quantifies the extent to which deviations from a model could be expected to occur by chance. Confirmatory analysis uses the traditional statistical tools of inference, significance, and confidence. Confirmatory analysis tests the statistical hypotheses.

[1] https://www.aicpa.org/InterestAreas/FRC/AssuranceAdvisoryServices/DownloadableDocuments/AuditAnalytics_LookingTowardFuture.pdf

Exploratory data analysis is sometimes compared to detective work: it is the process of gathering evidence. Confirmatory data analysis is comparable to a court trial: it is the process of evaluating evidence. Exploratory analysis and confirmatory analysis can, and should, be proceeded side by side.

In this document, Data Analytics is regarded as a computation process of Data Analysis. The computation process involves several phases such as collecting data, cleansing data, analyzing data, and deploying data.

Data Analytics are not specifically referred to a Generally Accepted Audit Standard in term of implementation of Computer Assisted Audit Techniques (CAATs). Data Analytics can be regarded as the evolutionary form of CAATs. Using Data Analytics, auditors are able to explore the data deeper and visualize the data in order to get broader range of audit objectives. The following picture draws the relationship between Data Analysis, Data Analytics, and CAAT,



1. Relationship between Data Analysis, Data Analytics, and CAAT

The purpose of Data Analytics in many organizations is to add a competitive advantage by enabling information-based decision making. To ensure the successful use of Data Analytics practices, it is necessary to consider using a goal-based approach rather than problem-based approach. While the problem-based approach starts the process with the question of how to solve the problem, goal-based approach starts the Data Analytics Process with the question of what we want to achieve. It can help the DA Process to use Objectives and Key Results (OKR) method to transparently align and prioritize resources towards a common goal.

In all, Data Analytics enhances the quality of information-based decision-making process. Data Analytics enables Supreme Audit Institution (SAI) to apply various techniques to obtain relevant insights such as pattern, relationship, and cluster in a set of data. Also, Data Analytics may enrich the SAI's management dashboard or Business Intelligence through an interactive data visualization. In addition to that, the Data Analytics process in this document adapted from several frameworks such as CRISP-DM[2], 3Cs[3] (Collect, Create, Consume), and Microsoft Data Science Life Cycle[4].

[2] ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf
[3] https://www.dqlab.id/data-science/
[4] https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

2. Initial Stage of Data Analytics Process

Data Analytics Process is a collection of processes starting with the identification of a business need. The goal of this initial stage is to define key variables whose metric is relevant to determine the success of this whole process. The output of this process is a relevant data and the source of the data.

Two main tasks of this initial stage are as follow.

- Identifying the target

  An ultimate objective of this task is to identify the key business variables in which the analysis needs to figure out. These variables then become the target of the proposed analytical model. Some examples of such goals are budget forecast and probability of an expenditure being fraudulent.

  Defining the target needs sharp questions that are relevant, specific, and unambiguous. The question will determine the appropriate algorithm that will be implemented in further process. Typical question and its appropriate algorithm are as follows.

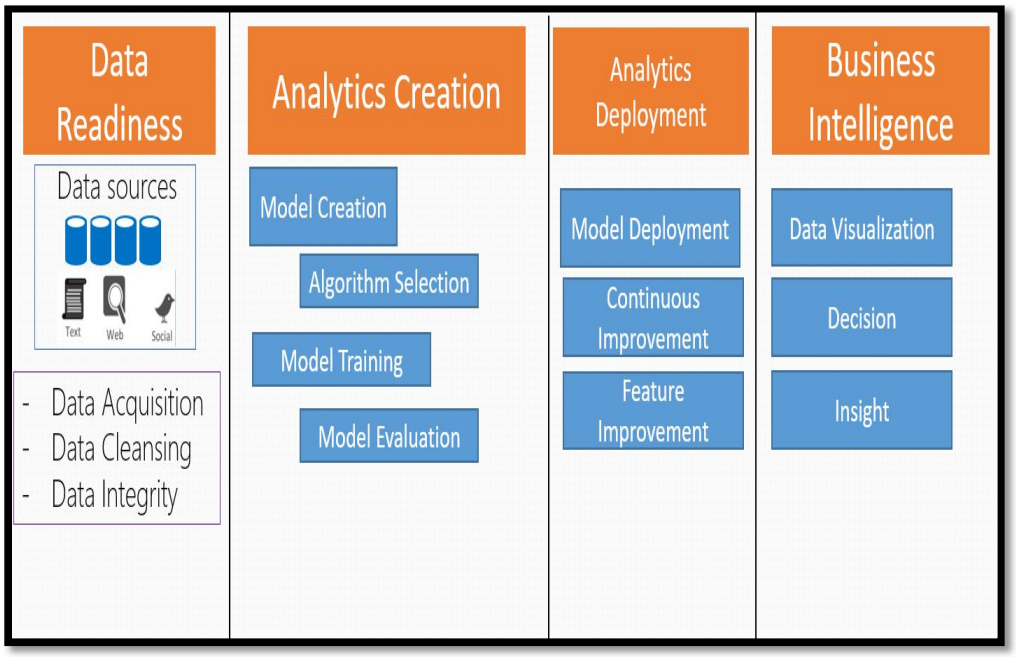
  - How much or how many? → Regression
  - Which category? → Classification
  - Which group? → Clustering
  - Is this weird? → Anomaly Detection
  - Which option should be taken? → Association

- Identifying the data source

  The output of this task is the list of data that are available and required for the analysis. The output of this process will be the input for the next process. A typical document of this task is the data dictionary.

The following picture shows the further processes after completing this initial stage. The processes are sequential in nature. However, like in the concept of Software Development Life Cycle, the process may generate a cycle of Data Analytics Process, i.e., the output of the last process could be proceeded as an input for the first process. For instance, the output of Business Intelligence may rise a new target (an input) for either the new or the improved analytic process, thus, initiate a new iteration of Data Analytics Process.

.



2. Data Analytics Process

The following chapters explain each stage of the data analytics process as shown in the picture above in greater detail..

3. Data Readiness

In the arena of Electronic Data Processing, there is a common phrase **Garbage In, Garbage Out**. It means if there a small part of unclean data into analytics, there will only be a nonsensical result, making the analyses useless. The Data Readiness is the condition in which the data is available and ready for analytics, with no garbage in it.

3.1. Data source identification

In the first stage of achieving Data Readiness, SAI should start with identifying the source of data required for analytics. There are two sources of data, i.e., the data that resides on SAI's premise (Internal) and the data that resides on other places (External) such as Auditee's premise, on the websites, or in the cloud storage.

3.1.1. Internal

Some examples of Internal Data Source are:
- Data generated through Audit Process
- Audit Entity Profile
- Any other audit-related data available in SAI's Data Center.

3.1.2. External

Some examples of External Data Source are:
- Audit Entity's Data which includes financial and non-financial data
- Other data available in public domain.

After all information regarding the data have been identified, auditors could start the ETL Process. ETL process consist of all processes starting from how the data is collected until the data is ready for analysis. ETL is the abbreviation of Extract, Transform, and Load. In this guideline, these three processes are labeled as Data Acquisition, Data Cleansing, and Data Porting.

3.2. Data Acquisition

This process identifies the type of data being collected and the method of collecting the data. The process assumes that collecting data from Internal SAI is not an issue. Therefore, the focus of this process is about collecting the data from external sources, i.e., auditee's premise and public domain.

3.2.1. Data type

Data type is the attribute of the data that tells the user on how to interact with such data. The common data types are as follow.

- String
  This type of data contains alphanumeric character. This type of data is not designed for mathematical calculation. Some examples of this data are employee name, employee identity number, address, and invoice number.

- Numeric

This type of data contains only numeric data that represents quantitative/ measurable information. Some example of this data type are an area of the city, the Invoice amount, and the sample size.

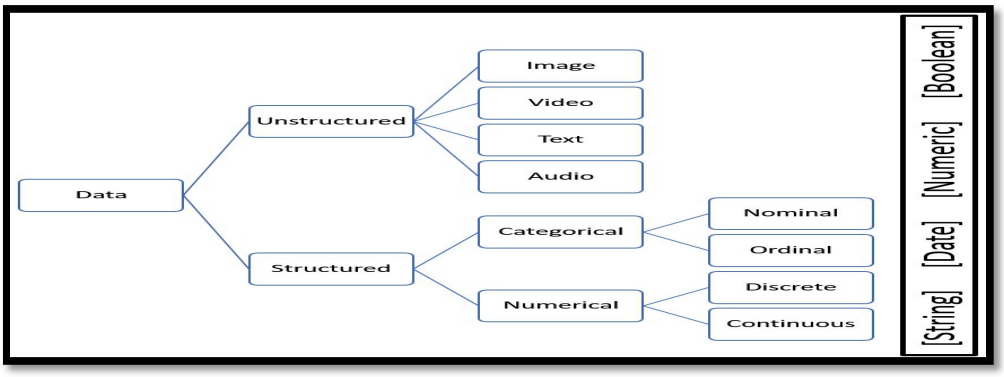
- Date
  This type of data represents a date value such as birthdate, invoice date, and report date.

- Boolean
  This type of data contains only a condition of True or False. Other variations of its content are Yes or No, and 1 or 0.

On top of these four common data types, there are two kinds of data based on their format, i.e., Structured Data and Unstructured Data. Structured Data is the data that comprises of two elements; row and columns. The structured data is often referred to a tabular form. A structured data is the form of data that is ready for an analysis process. Structured Data may contain a numerical or categorical value. Numerical value could be either a discrete value or continuous value. A discrete value contains only a certain value such as number of auditors, number of employees, and number of digits. A continuous value contains any value such as company's profit, width of a bridge, and cash balance. Categorical value may contain nominal and ordinal value. Nominal value is not intended for ordering purposes, instead, it may be useful for grouping the data. Some examples of nominal value are employee's name, gender, audit opinion, and assertion. Ordinal value, on the other hand, is intended for ordering. Some examples are Likert Scale, Academic Grading, and Profitability Ratio.

Another type is Unstructured Data. Unstructured Data comprises any kind of data which are far from tabular form such as Text, Video, Audio, Image, and Spatial. A large proportion of publicly available third party data, which are external in nature, are of this type. Unlike structured data, the unstructured data is not ready for analysis process. Certain preliminary processes are required for making it "ready".

Following diagram depict the tree of data.



3. Data Tree

3.2.2. Access Method

In many cases, auditors get the data from auditee's premise through a provisioned access to the specific system. Typical methods of obtaining data from the auditee are the read-only

access to the database, backup-restore mechanism, and delivering the requested data through LAN-WAN or by creating a data dump file from the data base by an appropriate query covering the requested data elements.

- Read-only access to the database

  Using this method, auditors are able to query the data per their need. This method offers auditors a high degree of freedom in selecting the data and arranging the data to fit the need. However, this method requires an extensive knowledge on creating query and the data structure itself. Without appropriate knowledge on the query and data structure, auditors may be lost in the forest of data. Also, without appropriate knowledge on the query, this method may contribute the degradation of system performance.

- Backup-Restore mechanism

  This method is a kind of cloning auditee's database. This method is relatively safer than the previous one. Auditors conduct data analytics in an isolated database, therefore, it does not impact the operational information system of auditee. However, auditors should have the same database management system. If the auditee uses Oracle, then the auditors should also have Oracle Database Management System.

- Delivering the requested data through LAN-WAN or VPN

  This method limits the auditor's interaction with auditee's database management system. Auditee put the file needed by auditors on the location in which auditor has right to access the file through organization network using Wi-Fi or Cable. In the same intention, it is possible for auditee to send the requested data to auditors through organization network or through Internet.

3.2.3. Data Extraction

Once the auditors know what kind of the data that they need and how to access such data, they can start to extract the data. Data extraction is important because the data that auditors need are stored in several locations such as from a database management system, a website, and a file. Also, auditors need to extract data in order to avoid the risk of altering the original source. It aslo helps narrowing down the dataset to cover only the necessary data elements for the purpose of analytics.

3.3. Data Cleansing

After receiving the data, the next process is the data cleansing. Data cleansing is the hardest part of data analytics process. This process is established on top of the assumption that the data come from extraction process are still dirty. Consequently, the data from extraction process cannot be loaded straightforward to the new storage or the new database.

Dirty data is the information that is either incorrect, corrupt, or missing. These three qualifiers cause the imbalance of the data. Auditors may deal with this situation in the analytics process. The imbalance of the data may defect the data quality since it may violate the five principles of

data quality, i.e., Validity, Accuracy, Consistency, Completeness, and Uniformity. Followings are what auditors should do for cleansing the data

3.3.1. Incorrect Data

In this qualifier, information has been incorrectly added to the database. Sometimes, this symptom is identified using our general knowledge or common sense. Some examples of incorrect data are as follows.

- The date of '04/12/2018' can be regarded as either December 4th, 2018 or April 12th, 2018.
- A transaction dated '05/09/2017' was included in a data set of transaction for year 2018.
- Reversed Longitude and Latitude value.
- Negative value of a wefare payment to a beneficiary.Age which is in 1000 years

Incorrect Data affects the Validity, Accuracy, and Consistency, thus, lowering the quality of the data.

3.3.2. Corrupt Data

This qualifier was caused by system either during transmission or during extraction. The data originally have been correct in the source dataset, however, there are several events that made it corrupt. The followings are typical events that lead to a corrupt data.

- The source dataset has been physically damaged
- The source dataset has been altered by another software
- The source dataset has been extracted in an unadvisable manner.

Some examples of corrupt data are as follows.

- The long numeric value that is converted into a string with exponential sign, e.g., a value of 1,000,000,000.00 was converted into string "1E+09"
- Incompatible Carriage Return character for Line Spacing.
- Inappropriate use of column separator when generating a quasi-csv file.
- Unicode problem

Corrupt data affects the Validity, Accuracy, Completeness, Consistency, and Uniformity.

The procedures that could be conducted to fix the corrupt data are:

- Re-extract the data form its original source to identify some procedures that may corrupt the data during the extraction process;

- Confirm to the person-in-charge of the data extraction to see if they can explain what the actual data should be;

- Exclude the rows that contain corrupt data from further process; being to be analyzed or being loaded into the database.

If these three procedures do not satisfy in resolving the problems, such corrupt data may then be labelled as the missing data.

3.3.3. Missing Data

This qualifier occurs when certain information does not exist in the dataset. This qualifier is a common topic in the data analytics. Human error is the primary factor of this problem.

Missing data affects the Validity, Accuracy, Completeness, Consistency, and Uniformity.

The available methods that could be conducted to fix the missing data are:

- Predict the missing data.
- Leave it as it is
- Remove the record or column, which contains missing data, entirely.
- Replace the missing data with mean/median value if the missing data is a numerical value.
- Type the value of missing data by exploring correlation and similarities.
- Introduce a dummy variable for the missing data.

The activity of dealing with missing data is not a mandatory activity especially if it requires modification of the content. Altering the data source may reduce the originality of the data thus violating the concept of chain of custody. Altering the data due to the existence of missing data depends on the degree of auditor's reasonable assurance. Other methods such as re-inquiry, to some extent, are still relevant. However, all process of altering the data should be well-documented in order to maintain the chain of custody.

3.4. Data Porting

Once the data are considered free from error, auditors can load the data into the target database or file. However, loading data into auditor's workplace can sometimes cause problems such as missing of cleaning up some dirty data. Consequently, after completing this process, auditors should take time to manually look through the data for the last time before running the analytic algorithm.

As in Computer-Aided Audit Technique, auditors should make sure that they work with auditable data. The following are typical techniques to make sure the data are ready for further analysis.

- Control Total

  This technique requires comparison of number of records between the original dataset and the target dataset. In addition to number of records, it is also necessary to sum up the value of certain or all numerical column and compare it to the initial dataset.

- Checking the columns for skewness

Using this technique, auditors check the top *n* and the bottom *n* rows. This information may be useful for further analysis stage.

- Checking the columns that are susceptible to corruption

  This technique is to ensure that all corrupted data are solved. This procedure includes check all columns that are most prone to error such as date and numeric.

- Checking the text value

  If the original dataset contains a free-form text, sometime the target dataset has a default length which is lesser than the length of text from original dataset. This technique is to ensure the length of the text is not trimmed.

At this stage, auditors have questions and relevant datasets. The next part is the creation of analytics to answer such questions based on the clean and reliable data.

4. Analytics Creation

Data Analytics are defined on chapter 1. Also, its overlap with CAAT was explained on the same chapter. This chapter will discuss some basic algorithms that are commonly used in Data Analytics after the data is ready for further analysis.

Analytics creation involves the inclusion, aggregation, and transformation of available data to generate the features that form the proposed business cases.

4.1. Model Creation

There are three approaches in Model Creation for generating the insight. These three approaches are Descriptive Analytics, Diagnostic Analytics, Predictive Analytics.

4.1.1. Descriptive Analytics

Descriptive Analytics is the process of Data Analytics that creates an overview of the data. It is the basic type of analytics which is frequently used by not only auditors but also engineers. It's taking historical data and summarizing it into something that is understandable by public. Summarizing, Cross tabulation, and Grouping are the common techniques to conduct Descriptive Analytics.

Using Descriptive Analytics, auditors are able to indicate a statistical synopsis of the data. For instance, as follows:

- In Year 2019, Government Revenue from Taxes is 75% of total Government Revenue
- The demographic structure of government employee this year is dominated by employee in range of 30 – 40 years old. Among these group, 75% is women.
- 37% of National Budget Realization is a Capital Expenditure in all ministry offices.

4.1.2. Diagnostic Analytics

Diagnostic Analytics is the process of Data Analytics that offers an integrated information to the auditor. Diagnostic Analytics enable auditors to find out the degree of integration among information and identify the reason of why something happened.

The benefit of Diagnostic Analytics can be derived from these three categories.

- Identification of Outlier
  Using the result of Descriptive Analytics, Diagnostic Analytics can further evaluate some information more detail to find out some outliers. These outliers may help auditors to answer the question raised in a business case. Outlier has one or more attributes that may contribute to a risk of misstatement. Also, outlier may provide auditors with useful information in designing or modifying audit procedures.

  The following are some examples of outlier.

- There was some amount of state subsidies from Ministry of Finance to seven villages where there is no poverty in those villages. Most of subsidies were transferred to villages with poverty.

- There are cars whose odometer indicate the increment 74.000KMs within one month. Most of cars can reach on average 19000KMs per month.

- Information Discovery
  Information Discovery in Diagnostic Analytics enable auditors to trace all data that relate to an anomaly in data. Often, Information Discovery requires auditors to look for patterns outside the existing data sets. Also, it might require additional data from other sources. A common process such as filtering and grouping could be used for identifying the common attributes of outliers, emphasizing on the nature, cause, and what is the impact at the level of assertion level.

- Uncovering the Causal Relationship
  This benefit can be obtained using regression analysis, filtering, and time-series data analytics. Supported by existing theory, Diagnostic Analytics is able to identify correlations and determine if any of them are causal in nature. Regression analysis is a statistical technique that could be applied in a time-series or a cross-sectional data. A regression analysis of historical data may help in identifying the existence of casual relationship between two variables, for example, causal relationship between amount fo debt and capital expenditure within the last ten years.

4.1.3. Predictive Analytics

Predictive Analytics is the process of Data Analytics that creates the estimation about the likelihood of an upcoming output or outcome. Among the three approaches in Data Analytics, Predictive Analytics is the most complex process as it is considered the fundamental of machine learning.

Three issues should be taken into account when auditors want to develop predictive analytics. These issues are as follow.

- A target

  Target in Predictive Analytics is the information that we would like to guess what will happen. In statistic terms, it can be referred to a Dependent Variable. There are two types of measurement in the target, i.e., continuous along predefined interval and categorical. A typical example for continuous target is predicting the amount of sales. And, a typical example of categorical target is predicting whether a credit card transaction is “fraud” or “no fraud”. The categorical target can be two or more than two classes.

- Indicators

  Combination of information that all together have impact to the target. In statistic term, it can be referred to a collection of Independent Variables. The process of identifying indicators requires solid academic references such as a theory or best practices as a basis. Without strong basis, the result might be spurious. In case of predicting whether a credit card transaction is fraud of no fraud, the indicators are several information attached to the credit card that, simultaneously, estimate the probability of fraud of the transaction. Some

typical indicators are location of transaction, average amount of transaction after the last transaction, amount of current transaction, expired date, and card issuer.

- Sufficiency of historical data

  In order to predict something, sufficient historical data is important. The sufficiency of the data is relative to the algorithm selection. In general, the more data you have, the more reliable the prediction is. All indicators and the target should be available in the historical data. In order to be more accurate in predicting the target, large amount of historical data is needed. Ceteris peribus, predicting target based on 10000 previous credit card transaction is statistically more accurate than predicting target based on 100 previous transaction.

- Proper Algorithm

  There are several algorithms available for conducting predictive analytics. These algorithms form two approaches in conducting predictive analytics. These approaches and the related algorithm are as follows:

  - Supervised

    A concept of predicting the value of variables based on historical data to a predefined class, for instance, predicting whether the credit card transaction is fraud or no-fraud, predicting whether the audit opinion would be Unqualified, Qualified, Disclaimer, or Adverse, and predicting whether the email is spam or no-spam. In the following is the list of some algorithms for conducting Predictive Analytics using Supervised Approach.

    i. Decision Tree[5]

    ii. Support Vector Machine[6]

    iii. Naïve Bayes Classifier[7]

    iv. Random Forest[8]

  - Unsupervised

    A concept of predicting the value of variables based on historical data to a group of similar attributes, for instance, predicting the group of accounts whose similar risk level, identifying the similar vendor's name and address, and recommending books to a customer. In the following is the list of some algorithms for conducting Predictive Analytics using Unsupervised Approach.

    i. K-Means[9]

[5] https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/
[6] https://dataaspirant.com/2017/01/13/support-vector-machine-algorithm/
[7] https://dataaspirant.com/2017/02/06/naive-bayes-classifier-machine-learning/
[8] https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/
[9] https://www.datascience.com/blog/k-means-clustering

ii. Hierarchical Clustering[10]

iii. DBSCAN[11]

iv. Apriori[12]

- Accuracy

  The accuracy depends on the algorithm used in the Predictive Analytics. Accuracy is the key factor of determining the proper algorithm. It is a common practice to use several algorithms simultaneously. The algorithm that provide the best accuracy among others should be chosen as the most suitable algorithm.

4.2. Model Training

This process applies only if auditors conduct a Predictive Analytics. Model Training is the process that involves several processes such as splitting the data into two part, i.e., for training and for testing, selecting algorithm, and tuning the statistical feature.

- Splitting the data

  The available historical data is split into two parts, i.e., part for training and part for testing. There is no consensus on the size of training data and testing data. The common practice is the proportion of 80% for training data and 20% for testing data.

- Selecting algorithm

  In this step, auditors choose the algorithm for conducting Predictive Analytics. There are three types of algorithm in Predictive Analytics. They can be distinguished depending on the measurement level of the target. These types are:

  - Classification
  - Clustering
  - Regression

- Tuning statistical feature

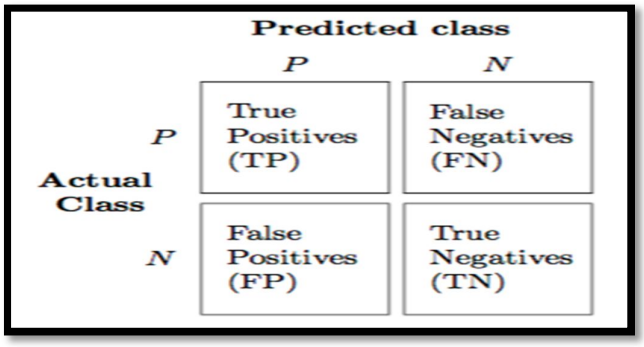
  Each algorithm has its parameters than can be used to optimize the result in term of accuracy, processing time, and process efficiency.

4.3. Model Evaluation

This process applies only for classification in a Predictive Analytics. Three tools are available to measure the performance of the model. These tools are Confusion Matrix, Receiver Operating Characteristic (ROC), and Area Under the Curve (AUC).

[10] https://dataaspirant.com/2018/01/08/hierarchical-clustering-r/
[11] https://www.datanovia.com/en/lessons/dbscan-density-based-clustering-essentials/
[12] https://en.wikipedia.org/wiki/Apriori_algorithm

Confusion Matrix is a table for explaining the accuracy of a classification model on a set of test data for which the true values are known. This table shows a level of accuracy of predicting the values and the actual values. The following picture depicts the Confusion Matrix.



4. Confusion Matrix

To get the understanding of confusion matrix, recall to a sample in Predictive Analytics (on page 15) about predicting the credit card transaction whether the transancation is fraud or no fraud. Following is the sequential process for predicting the credit card transaction.

#1. Obtain the historical data of Credit Card Transaction, include the variable that contains the status of transaction, i.e., fraud or no-fraud. Let's say, there are 4600 Credit Card Transaction.

#2. Split historical data into two part and name it as Trained Data and Test Data with composition is 80% [3680 records] for train data and 20% [920 records] for test data. The trained data is a group of data that will be used by a predictive analytics algorithm.

#3. Apply the selected predictve analytics algorithm to the Trained Data. This will result a Predictive Model. Predictive Model is the Mathematical Model that represents the pattern of historical data in deciding the fraud or no-fraud transaction.

#4. Apply the Predictive Model to the Test Data. This will result an information about the accuracy of the prediction.

For example, among these 920 transactions in the Test Data, the result after applying it to the predictive model indicate the following information;

- There are 200 transactions predicted as fraud [Positive] and the original test data shows it is fraud [True].
- There are 700 transactions predicted as no-fraud [Negative] and the original test data shows it is no-fraud [True].
- There are 13 transactions predicted as no-fraud [Negative] and the original test data shows it is fraud [False].

- There are 7 transaction predicted as fraud [Positive] and the original test data shows it is no-fraud [False].

The Evaluation of the model will add the new information of predicted fraud or no-fraud transactions. In some cases, the predicted value may be different from the existing value. To find out how significant is the differences, auditors can use Confussion Matrix.

There are four conditions that are mapped to the confusion matrix. These conditions are:

i. True Positive (TP)

A condition where the predicted is **positive** and it is **true** according to the fact value.

ii. True Negative (TN)

A condition where the predicted is **negative** and it is **true** according to the fact value.

iii. False Positive (FP)

A condition where the predicted value is **positive** and it is **false** according to the fact value.

iv. False Negative (FN)

A condition where the predicted value is **negative** and it is **false** according to the fact value

Back to the case of credit card fraud, each TP, TN, FP, and FN is 200, 700, 13, and 7, respectively. As a result, the Confussion Matrix of the predictive model can be drawn as follow.

TP 200	FP 13
FN 7	TN 700

There are three indicators for measuring the reliability of the predictive model. These indicators are as follows;

i. Recall

Recall indicates the level of how much the model predict accurately (correct prediction on fraud transactions) among all positive cases (all transactions labeled fraud). The value should be higher.

The formula is: TP / (TP + FN).

Put the value to the formula, the Recall = 96.6% [200 / (200 + 7)]

ii. Precision

Precision indicates the level of acuracy of actual positive (correct prediction on fraud transactions) among all predicted positive (all transactions predicted fraud).

The formula is: TP / (TP + FP).

Put the value to the formula, the Precision = 93,9% [200 / (200 + 13)]

iii. Accuracy

Accuracy indicate the level of all correct prediction among all transactions.

The formula is: (TP + TN) / (TP + TN + FN + FP).

Put the value to the formula, the Accuracy = 97.8% [(200 + 700) / (200 + 700 + 13 + 7)]

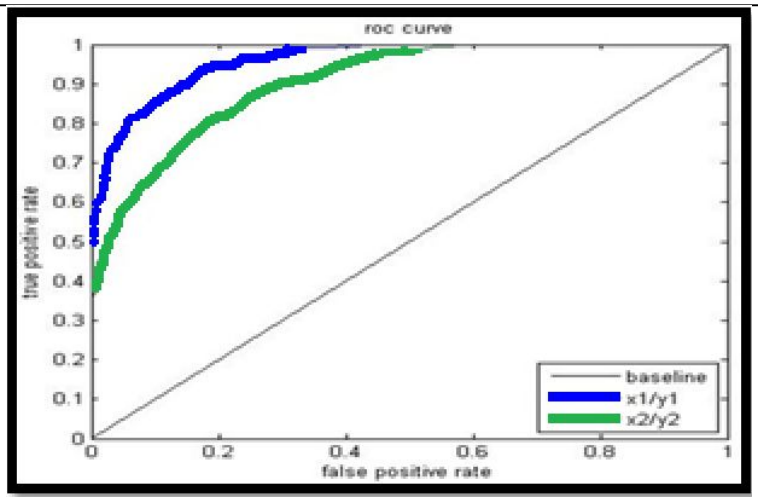
Receiver Operating Characteristic (ROC) is a graph that represents the performance of a classification model at all classification thresholds. This graph plots two parameters; True Positive Rate and False Positive Rate.

- False Positive Rate (FPR) = False Positive / (False Positive + True Negative)
- True Positive Rate (TPR) = True Positive / (True Positive + False Negative)

On the ROC Chart, at least, there are two plots; the base line and the result of classification algorithm. The closer the plot is to the baseline, the less accurate the test is.

If there are two algorithms simultaneously tested with the same data set, the algorithm whose plot is the farthest from baseline is the best algorithm among them. The farthest plot represents the model that is able to distinguish the classification with no-significant overlap.

To illustrate, in Figure 5, the ROC Chart shows that the algorithm with the blue plot is better than the algorithm with the green plot.



5. ROC Chart

Another tool is Area Under the ROC Curve (AUC). AUC has been proposed as the alternative metric as a complimentary of ROC Curve. Many existing learning algorithms have been modified in order to seek the classifier with maximum AUC.

In the process of Data Analytics, it is relevant to put a quote from Peter Norvig, Director of Research of Google.

"We don't have better algorithms; we just have more data. More data beats clever algorithm, but better data beats more data."

5. Analytics Deployment

The output of Analytic Creation is the most accurate model to answer the question or to meet the business case. A model as a result of Data Analytics process, should be stored in an appropriate format, depend on the software used for both generating and consuming it. This process adopts the concept of reusability in which the model can be used repetitively without a necessity to run and to test the accuracy of the model.

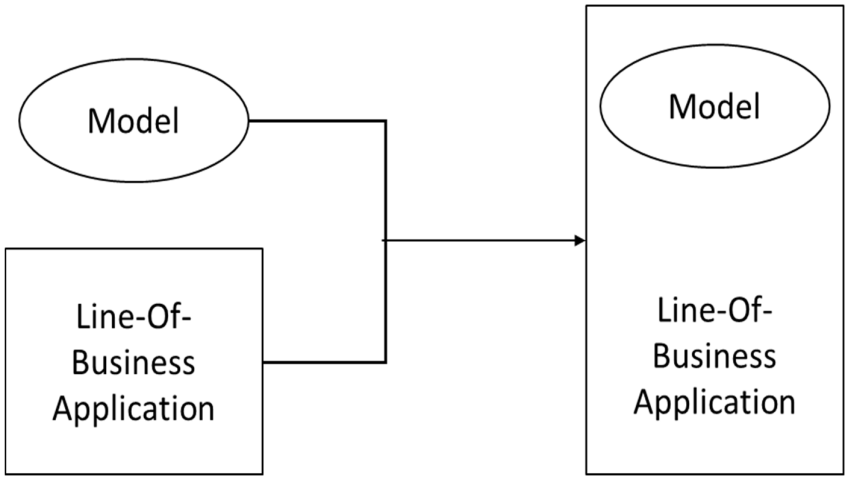
For instance, a model for predicting the likeliness of credit card transaction is fraud can be stored somewhere so that the Line-Of-Business Application can use the model to immediately raise a warning if the credit card transaction is suspected to be a fraud. By doing so, it is not necessary for the application to find pattern from large amount of historical data every time transaction occurs.

5.1. Model Deployment

Model Deployment is the process of integrating the model into an existing information system. The model is stored in a production environment. As a result, other applications such as Line-Of-Business (LOB) and office applications are able to consume the model. There are two common scenarios of deploying the model.

- Deploying the model as an extension of existing LOB Applications.

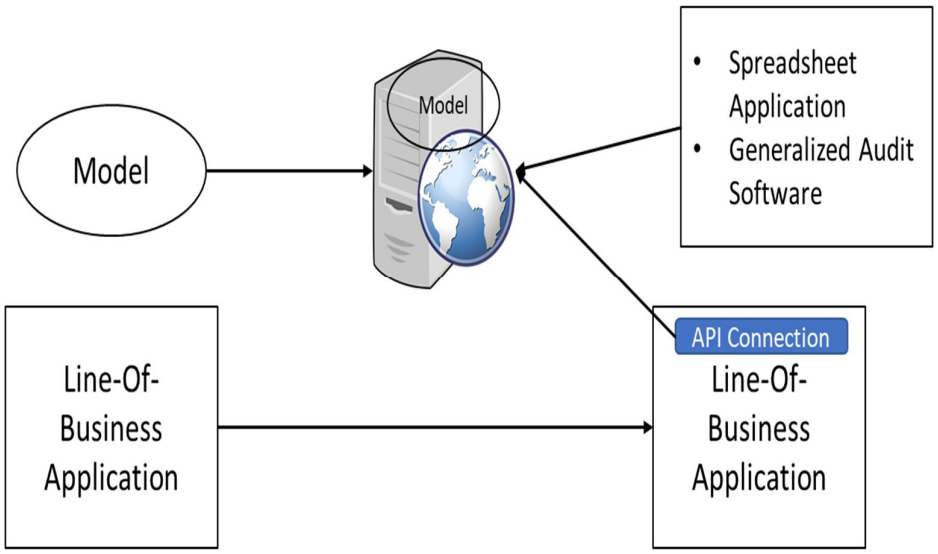
  This scenario requires some modifications of the LOB Application. Consequently, any changes of the model lead to a version upgrade of the LOB Application. The following picture briefly describes this scenario.



6. Deploying Model as an extension of LOB Application

- Deploying the model as an API Web Service.

  This scenario is more convenient than the previous. In this scenario, the model is deployed on the server computer, usually a web server. By deploying the model as an API Web Service, modification of the existing application is only necessary once. In general, the change of the model require redeployment to the server but does not require a modification of the LOB Application. Furthermore, deploying the model as an API Web Service enables the spreadsheet or other software to consume the model. The following picture briefly describes this scenario



7.Deploying Model as an API Web Service

As shown on the picture, there is an opportunity to consume the model using traditional software; Spreadsheet, Generalized Audit Software, and any other softwares that have features of accessing the API Web Service.

5.2. Continuous Improvement

When the model has been deployed on the shareable environment, more user has access to the model. The more user consumes the model, the more data confirms the accuracy of the model. In the beginning period of deployment, the accuracy of the model may be steady. As more data are applied to the model, there can be a situation where the accuracy of the model starts weakening. If the accuracy is weakening, the model should be redeveloped using the new data. Consequently, this would lead the process back to the Model Training (Section 4.2).

For instance, in case of credit card fraud prediction, if the model gradually generates false prediction after some period of its impementation, the model should be retrain using new credit card transactions. This process is performed either regularly or based on certain measures.

5.3. Feature Improvement

In contrast to Continuous Improvement, the Feature Improvement does not require the redevelopment of the model using new data. Instead, it requires the redevelopment of the model using new configuration such as, new variable and new statistical parameter value.

For instance, in case of credit card fraud prediction, if the model generates more false prediction since the beginning of its impementation, the model should be retrained using new indicators or new prediction method.

In some circumstances, the combination of Continuous and Feature Improvement is necessary to get the optimal accuracy level.

6. Business Intelligence

Business intelligence (BI) is a collection of techniques and tools used to transform raw data into meaningful information through visualization for business analysis[13]. In other words, BI integrates the results of Data Analytics and the power of Data Visualization.

6.1. Data Visualization

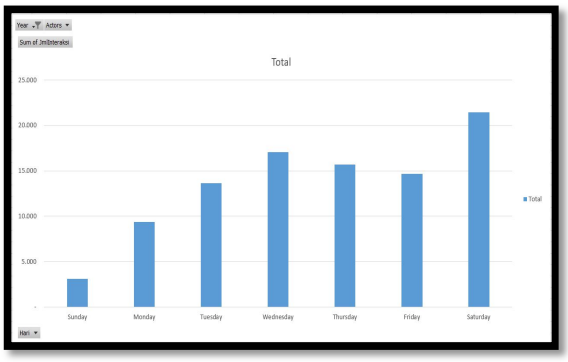
Data Visualization is the process of presenting the result of Data Analytics. Data Visualization hide the complexity of the Data Analytics process from the end-user. There are two types of Data Visualization.

- Static Visualization

  This type of visualization is referred to the traditional way of displaying the data either in tabular or graphical mode. Creating this type of visualization can be easily done by a traditional spreadsheet such as Microsoft Excel and LibreOffice Calc, and GAS such as ACL and IDEA.

  The followings are typical examples of Static Visualization.

Year	2017
Employee	(All)
Row Labels	**Sum of JmlInteraksi**
Sunday	3.115
Monday	9.377
Tuesday	13.663
Wednesday	17.083
Thursday	15.713
Friday	14.652
Saturday	21.426
Grand Total	**95.029**



[13] https://competency.aicpa.org/media_resources/211947-utilizing-business-intelligence-to-your-benefit

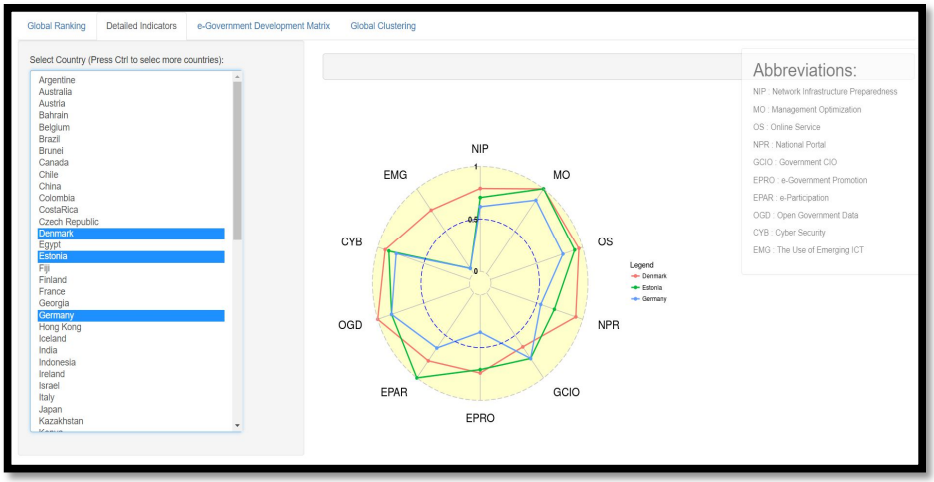
Country	Salesperson	2011 Sum of Units	2011 Sum of Order Amount	2012 Sum of Units	2012 Sum of Order Amount	2013 Sum of Units	2013 Sum of Order Amount
UK	Bromley	232	24,756.89	228	40,396.64	73	9,894.51
	Coghill	81	4,029.25	39	4,657.11		
	Farnham	170	14,055.87	44	5,892.65	17	2,560.40
	Gillingham	397	40,826.37	276	17,181.58	202	14,519.68
	Gloucester	209	31,433.16	143	19,691.89	135	17,667.20
	Rayleigh	422	59,827.19	268	41,903.64	131	15,232.16
UK Total		**1,511**	**174,928.73**	**998**	**129,723.51**	**558**	**59,873.95**
USA	Bromley	58	7,553.95	27	3,654.00	7	1,101.20
	Callahan	623	49,400.07	337	43,263.95	200	18,059.50
	Coghill	885	120,626.31	520	46,505.90	405	49,945.11
	Farnham	699	89,663.20	506	73,360.59	217	15,663.56
	Finchley	699	95,850.36	487	55,787.97	302	30,861.76
	Fuller	539	71,168.14	473	73,524.18	170	17,811.46
USA Total		**3,503**	**434,262.03**	**2,350**	**296,096.59**	**1,301**	**133,442.59**
Grand Total		**5,014**	**609,190.76**	**3,348**	**425,820.10**	**1,859**	**193,316.54**

- Dynamic Visualization

  Dynamic Visualization, in a simple term, can be formulated as a Static Visualization plus a feature of Interactivity. Not only interactivity but also animation can be included in a visualization.

  A common feature of dynamic visualization is the clickable on most area of visualization. For example, in a tabular based visualization, the cell or the value can be either clicked or right-clicked to go through a more detail information linked to it.

  The following picture illustrate a dynamic visualization. The visualization provides the user with the ability to compare one object to others. In this example, comparing Denmark, Estonia, and Germany.



8. Dynamic Visualization

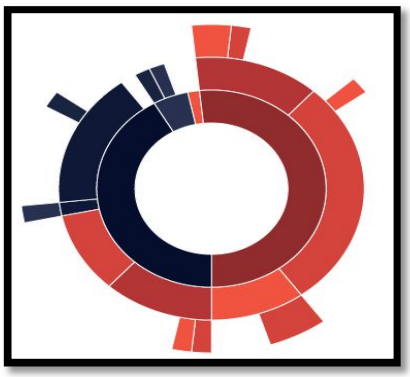
6.2. Insight

In information science, there is a concept of the level of human mind understanding and connectedness. The level is arranged as data, information, knowledge, and wisdom consecutively. Insight is located between information and knowledge. Data visualization is essential to uncover the insight of datasets.

There are various Data Visualization types[14] for exposing some interesting information and gaining the insight. The following are common types of visualization that relate to gaining the insight in auditing. This could be helpful for auditor when identifying some irregularities.

- Sunburst Diagram

  A Sunburst Diagram[15] is used to visualize hierarchical data, depicted by concentric circles. The circle in the center represents the root node, with the hierarchy moving outward from the center. A segment of the inner circle bears a hierarchical relationship to those segments of the outer circle which lie within the angular sweep of the parent segment.

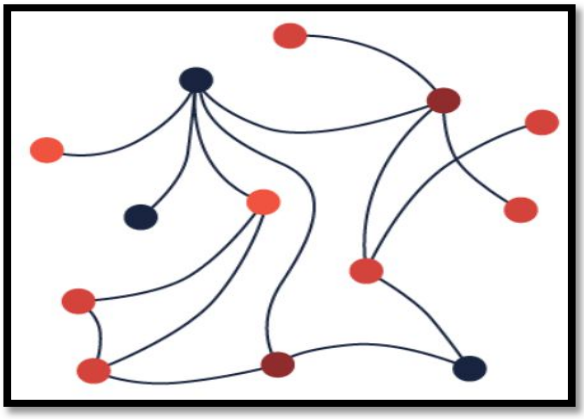


9. Sunburst Diagram

- Network Diagram

  Network Visualization[16] (also called Network Graph) is often used to visualize complex relationships between a huge number of elements. A network visualization displays undirected and directed graph structures. This type of visualization illuminates relationships between entities. Entities are displayed as round nodes and lines show the relationships between them. The vivid display of network nodes can highlight non-trivial data discrepancies that may otherwise be overlooked.

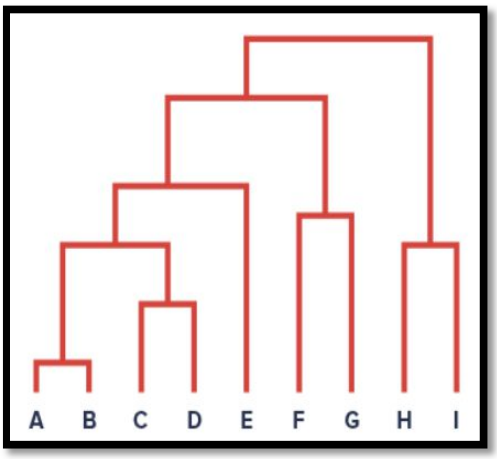
[14] https://datavizproject.com/data-type/
[15] https://datavizproject.com/data-type/sunburst-diagram/
[16] https://datavizproject.com/data-type/network-visualisation/



10. Network Diagram

- Dendrogram

  A dendrogram[17] is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.



11. Dendrogram

6.3. Decision Support

The Implementation of Data Analytics helps SAI and its auditors to use data as a basis for decisions and conclusions.

Decision Supports is the ultimate goal of Data Analytics and Data Visualization.

[17] https://datavizproject.com/data-type/dendrogram/

7. Data Analytics in Audit

7.1. Potential use of DA in audit

DA can contribute to every phase of the audit

- Audit planning, whether strategic, macro, micro (entity level) or engagement planning;
- Understanding the entity and its environment and assessing the risks of material misstatement;
- Evaluating the design and implementation, and testing the operating effectiveness of internal controls;
- Substantive testing, both analytical procedures and tests of details; and
- Concluding and reporting.

DA is relevant to and has the potential to significantly improve audit procedures throughout the audit. Examples include procedures for the following:

- Identifying and assessing fraud risk
- Performing external confirmation procedures, especially the identification of high risk items for confirmation
- Auditing accounting estimates
- Obtaining an understanding of related party relationships and transactions
- Obtaining evidence about the valuation of investments, the existence and condition of inventory, as well as the completeness of litigation, claims, and assessments
- Identifying material subsequent events
- Evaluating whether there is substantial doubt about the entity's ability to continue as a going concern

DA integration with traditional audit methodology enable auditors to reap the potential benefits, regardless of type of audit performed.

7.2. Considerations in Determining Which DA to Use to Meet the Objective of the Audit Procedure

The data analytics literature distinguishes between two different modes of analysis, exploratory and confirmatory. Exploratory DA is bottom-up and inductive. It starts with the data and the auditor asking questions such as, "What does the data suggest is happening? Does the data

suggest something might have gone wrong? Where do the risks appear to be? Are there potential fraud indicators? On what assertions should we focus? What models and approaches appear to be optimal for analytical procedures?" Exploratory DA is most useful in audit planning—understanding the entity and its environment, identifying and assessing the risks of material misstatement, and designing further audit procedures.

Confirmatory DA, on the other hand, is top-down and deductive. It starts with audit objectives and assertions. It tends to be model-driven with the auditor asking questions such as, "Is the subject matter consistent with my model (that is, with expectations)? Are there deviations that are individually significant or that form a pattern, such that they indicate the potential presence of material misstatement?" Confirmatory DA is used to provide the auditor with substantive or controls assurance about whether management's assertions are materially correct—ultimately, whether the financial statements are free from material misstatement.

The use of visual exploratory techniques can help auditors see patterns, trends, and outliers that are otherwise hidden, and reveal relationships between variables that could be the foundation for a confirmatory model. Confirmatory techniques are more formal and tend to be more mathematical and analytical (Behrens 1997); for example, they might utilize multiple regression analysis or the extraction and summarization of transactions meeting certain risk criteria. However, there is no bright line distinction between exploratory and confirmatory DA, and they tend to be used iteratively. For example, initial exploratory techniques may suggest a fruitful confirmatory model to be used for substantive analytical procedures, but the residuals from that model (actual minus expected) may lead to the discovery of additional factors that can be used to improve the model. Some of the same techniques can be used for exploratory and confirmatory analytics.

Examples of matters an auditor may consider in determining which DA to use, and the methods and tools to use in applying it, include the following:

- Whether the DA is to be used in risk assessment, test of controls, substantive procedures, or in helping to form an overall audit conclusion
- The nature and extent of the account balances, classes of transactions, and related assertions for which the DA is being used
- The persuasiveness of the audit evidence, including, where applicable, the level of precision the DA is intended to provide
- The types of risk of material misstatement it is expected to respond to when used in a substantive procedure
- Whether the DA is intended to be focused on any combination, or all, of the following:
  - ✓ Organizing data into some form of hierarchy to enable further analysis (for example, sorting or classification)
  - ✓ Determining the key attributes of specified types of accounts or classes of transactions
  - ✓ Searching for data with specified characteristics
  - ✓ Developing an estimate of a value or another attribute

- ✓ Identifying data that has attributes that are outside of specified ranges (for example, values or frequencies of occurrence that are significantly higher or lower than would normally be expected in the circumstances)

- ✓ Identifying data having similar attributes when that would not normally be expected in the circumstances

- ✓ Determining whether there are relationships (for example, correlations or causal relationships) among variables

7.3. Relation to Applicable Auditing Standards

There is a risk associated with the use of new and innovative techniques for which there is not a strong framework within the standards.

GAAS[18] does not specifically mention the use of data analytics techniques. However, the lack of reference to data analytics beyond mention of traditional CAATs in GAAS should not necessarily be viewed as a barrier to their adoption more broadly.

This lack of reference to data analytics in GAAS does not also result in a view that gathering information from the use of data analytics does not necessarily reduce the procedures required by GAAS today. Some procedures may or may not be reduced as a result of the information gained from the use of data analytics in an audit assignment.

Many similarities can be drawn between DA and CAATs. DAs could be applied manually to discover and analyze patterns, identify anomalies, and extract other useful information in data. However, in practice, they would seldom be performed without using a computer. In that regard, DAs might be viewed as an evolutionary form of CAATS that have, for example, enabled the auditor to make more effective use of data visualization techniques and help achieve a broader range of audit objectives. On the otherhand, DA has a greater applicability in audit planning, whereas CAATs have limited application in planning. Also, DA has more applicability in terms of testing conformity.

7.4. Relevance and Reliability of Data

Auditor must design and perform audit procedures that are appropriate in the circumstances for the purpose of obtaining sufficient appropriate audit evidence.

The sufficiency and appropriateness of audit evidence are interrelated. Sufficiency is the measure of the quantity of audit evidence. The quantity of audit evidence needed is affected by the auditor's assessment of the risks of misstatement (the higher the assessed risks, the more audit evidence is likely to be required) and also by the quality of such audit evidence (the higher the quality, the less may be required). Obtaining more audit evidence, however, may not compensate for its poor quality.

Appropriateness is the measure of the quality of audit evidence; that is, its relevance and its reliability in providing support for the conclusions on which the auditor's opinion is based. The

[18] Generally Accepted Audtng Standards (GAAS)

reliability of evidence is influenced by its source and by its nature, and is dependent on the individual circumstances under which it is obtained.

7.4.1. Relevance

Relevance deals with the logical connection with, or bearing upon, the purpose of the audit procedure and, where appropriate, the assertion under consideration. For financial audit, the relevance of information to be used as audit evidence may be affected by the direction of testing. For example, if the purpose of an audit procedure is to test for overstatement in the existence or valuation of accounts payable, testing the recorded accounts payable may be a relevant audit procedure. On the other hand, when testing for understatement in the existence or valuation of accounts payable, testing the recorded accounts payable would not be relevant, but testing such information as subsequent disbursements, unpaid invoices, suppliers' statements, and unmatched receiving reports may be relevant.

7.4.2. Reliability

The reliability of information to be used as audit evidence, and therefore of the audit evidence itself, is influenced by its source and its nature, and the circumstances under which it is obtained, including the controls over its preparation and maintenance where relevant. Therefore, generalizations about the reliability of various kinds of audit evidence are subject to important exceptions. Even when information to be used as audit evidence is obtained from sources external to the entity, circumstances may exist that could affect its reliability. For example, information obtained from an independent external source may not be reliable if the source is not knowledgeable, or a management's expert may lack objectivity. GAAS has some generalization about reliability of audit evidence, two of which are discussed bellow:

- The reliability of audit evidence is increased when it is obtained from independent sources outside the entity. However, when using data analytics, auditor cannot assume that data from third-party sources is complete and accurate. External data obtained from third-party data providers may only be an aggregation of data obtained from multiple sources and may not have been subject to procedures to validate completeness, accuracy and reliability of data that is needed in an external audit context.

- The reliability of audit evidence that is generated internally is increased when the related controls, including those over its preparation and maintenance, imposed by the entity are effective. When using data analytics, this means auditor have to consider and document some aspects of general IT controls and application controls, particularly:

  - ✓ The level of general IT controls testing, and the impact of the results of that testing; and

  - ✓ The impact of any deficiencies in general IT controls and application controls upon which the auditor intends to rely in order to conclude that the data from the IT system is sufficiently reliable for the auditor's purpose.

When performing data analytics, especially in data cleansing phase for dealing with missing data, auditor must consider reliability requirements when choosing what actions or techniques to take. For example, when used in audit planning stage, it may be acceptable to use prediction

to fill-in the missing value, but such action may not be acceptable when performing substantive analytical procedures.

7.5. Addressing Circumstances in Which DA Identifies a Large Number of Items for Further Consideration

When DA involves 100 percent of items in sizeable populations, the auditor may initially identify a large number of items requiring some form of auditor consideration to ensure that risk is sufficiently low. In some cases, items initially identified using a DA may, in fact, represent a previously unidentified risk or a higher level of risk of material misstatement than initially assessed, control deficiencies, or misstatements. In other cases, some or all the items identified using the DA may not, in fact, represent those types of matters (that is, those items may be what are sometimes called "false positives").

In determining whether the items identified warrant an audit response, further attention may not necessarily involve the performance of an investigation of each individual item identified. For example, the auditor's response might include one or more of the following:

- More clearly defining the characteristics of the data that are likely to be indicative of matters that require an audit response and then re-applying the DA using these more clearly defined characteristics.

- Identifying subgroups within the population of items that initially appear to warrant further attention and designing and performing additional procedures that may effectively and efficiently be applied to each subgroup. That further analysis might, for example, provide evidence that a subgroup does not represent a risk of material misstatement, control deficiencies, or misstatements. On the other hand, the follow-up analysis might indicate that the items in a subgroup require further response from the auditor. The nature, timing, and extent of additional procedures required would take into account the relevant characteristics of the items in the subgroup.

- Applying a different DA, or another procedure, that might more clearly identify those items that represent a risk of material misstatement, control deficiencies, or misstatements.

7.6. Documentation

GAAS do not currently require the auditor to retain all of the information used in selecting items to test, but require the auditor to document the identifying characteristics of the specific items or matter tested. The documentation requirements need not be any different when making use of data analytics. Auditor may record the scope of the procedure and identify the population analyzed or tested. GAAS do not require (nor, in many cases, is it practicable) to include in the audit file, or incorporate by reference, all the data analyzed or tested using an audit procedure.

The documentation may include the following:

- Objectives of the procedure

- Risks of material misstatement that the procedure intended to address at the financial statement level or at the assertion level

- The sources of the underlying data and how it was determined to be sufficient and appropriate (as necessary in the context of the nature and objectives of the DA being performed)
- The DA and related tools and techniques used
- The tables or graphics used, including how they were generated
- The steps taken to access data, including the system accessed and, when applicable, how the data was extracted and transformed for audit use
- The evaluation of matters identified as a result of applying the DA and actions taken regarding those matters
- The identifying characteristics of the specific items or matters tested
- The individual who performed the audit work and the date such work was completed
- The individual who reviewed the audit work performed and the date and extent of such review

8. Data Analytics Project Management

The need for project management structure around data analytics (DA) is especially important because of the iterative nature of DA. In general, project management in DA is the same as project management in other activities. There are five processes, i.e. initiating, planning, executing, monitoring and controlling, closing.

8.1. Initiating

In this phase, auditor should define and identify some things.

- audit objectives
- audit approach to meet objectives
- audit tests to be performed

Auditor should also consider some issues.

- Can data analytics be used to perform the testing?
- Does the audit team have the resources (people, time, and technology) to perform the analytics?
- Is the data available?

Deliverable of initiating phase:

- Kick off meeting documents
- Preliminary document

8.2. Planning

There are some important things that should be done by auditor at this phase.

- Define requirements of analytics
- Identify data sources and criteria
- Create time estimates (budget) for each analytic objective
- Prioritize analytics

Deliverable of planning phase:

- Data Analytics Audit Work Plan (stating the audit objectives of the analytic, audit procedures, man-days, assigned personnel)
- IT Infrastructure understanding documents

- Data Dictionary from the audited entity

8.3. Executing

In the execution phase of DA, auditor performs some steps in sequence.

- Retrieve data
- Validate data
- Code analytic routines – use scripts to capture logic and to allow for re-runs
- Confirm results
- Re--code as necessary

Deliverable of executing phase:

- Clean extracted data
- Documented scripts and queries
- Documented audit evidence
- Presentation to Management

8.4. Monitoring & Controlling

Auditor should monitor and control at least two things.

- Completed objectives
- Time and budget

Before deciding to proceed with the DA project, auditors should ensure being aware of some issues.

- Were additional areas to examine identified?
- Does it make sense to continue?

Deliverable of monitoring and controlling phase:

- Supervising Document from team leader or supervisor

8.5. Closing

There are some questions that should be answered before DA project is closed.

- Have we met the defined objectives?
- Were additional areas to exam identified?
- What are our lessons learned?

- How did the analytic effort enhance the audit?

Deliverable of closing phase:

- Lesson learned document
- Final report on results

9. Glossary

Data Analytics in Audit – the science and art of discovering and analyzing patterns, identifying anomalies, and extracting other useful information in data underlying or related to the subject matter of an audit through analysis, modeling, and visualization for the purpose of planning or performing the audit

CAAT – Computer assisted audit techniques (CAATs) refer to the use of technology to help evaluate controls by extracting and examining relevant data.

Model – simply a mathematical equation that describes relationships among variables in a historical data set

Line of Business (LoB) Application – a software for running and supporting a core business of an organization such as Human Resource MIS, Audit MIS, and Financial MIS.

Application Programming Interface (API) – a specification of possible interactions with a software component, usually built as a software for providing another software to communicate and collaborate with each other.

10. References

1. PnP Team, *"MSF Agile"*, Microsoft, 2005.
2. ISSAI 1500 – Audit Evidence.
3. Audit Analytics and Continuous Audit: Looking Toward the Future, AICPA, 2015.
4. Exploring The Growing Use Of Technology In The Audit - With A Focus On Data Analytics, IAASB, 2016.

11. Appendices

11.1. Example of Data Analytics Used in Identifying Potential Shell Company

Background Information

In auditing a ministry, there is a task to investigate a potential fraud in the procurement unit. Specifically, a complaint in the Whistle Blowing System claimed that some staff in the procurement unit were living extravagant lifestyles bankrolled by the vendor company. To see whether the claim was true or false, auditors analyzed the vendor data for shell companies. Shell companies are fictitious vendors created by an employee to commit fraud and embezzle money from ministry. If the ministry does not have independent vendor authorization, procurement staff may be able to add those shell vendors. The employee, or fraudster, often use his/her home address as the shel company's address.

At first, based on previous training on CAAT, the auditor tried to compare the address of vendors with those of the ministry's employees using simple join operation. However, the auditor was not satisfied with the result since the method only identifies the *exact match*.

Identifying Target (Question)

In order to perform Data Analytics, auditors decided to find the similar address between vendor's address and employee's address. Auditors must use alternative method in comparing these addresses.

Data Readiness

Auditor collect the vendor and employee table and reshape them so that the data is realiable for further process. The reshaping process covers all process include identification, acquisition, and cleansing. The following picture briefly depict the process.

Vendor Table

VENDOR_ID	Address
KD-00001	Jl. Pulo Bambu No. 15, Kota Tenggara Lama
KD-00002	Taman Vivo Indah, Blok AA No. 7
KD-00003	Meta Residences, No. 32C
KD-00005	Jln. Tegal Sari Indah, No. D87 -- Kota H
KD-00006	Perum Pluto, Blok C No. 1
KD-00008	Kali Mars Cluster, No. 24C
KD-00010	Perum Venus, Gg. Harimau No. 1A
KD-00012	Pulo Bambu No. 15, Kota Tenggara Lama
KD-00015	Jl. Puri Arteri Raya, No. 88 - Kota T
KD-00020	Jln. Manggis II, Gang Buntu No. 1
KD-00021	Puspa Loka, No. 98B, Kota Y
KD-00024	Perum Maju Permai Persada Indah, Gang Kenari No. 3

Employee Table

EMPLOYEE_ID	Address
EMPL_001	Gang Bulan Desember III, No. 9
EMPL_002	Apartemen Kecapi Indah, Lt. 16 No. 1610
EMPL_003	Jalan. Kebon Jahe, No. F16 - Kota E
EMPL_004	Cluster Ikan Mas, Taman Intan No. 2
EMPL_005	Jalan Hang Tuah, No. 11, Kota DM
EMPL_006	Boulevard Raya Residences, Blok AA2 No. 88
EMPL_007	JL. Pahlawan, No. 69CCD
EMPL_008	Asrama Pelajar No. 22 A - Pondok Bima Sakti
EMPL_009	Jl. Bintang Supernova, No. 78
EMPL_010	Jl. Wisma Tenteram Saja, No. A22
EMPL_011	Asrama Perawat IV, No. 1 - Kota D

Analytic Creation

There are various method, algorithm, and tools to compare the similarity between two text, in this case, between vendor's address and employee's address. The common method is to measure the distance between these two texts. In some modern Audit Softwares, there is a function of use of fuzzy logic to identify duplicates.

If auditors are satisfied enough with the result, the auditors may proceed further using this method and tool. This analytics process will result a clustering among the record with percentage probability of similarity, which is called similarity index. The similar texts were grouped into one cluster.

Following is the result of the process.

grouping	SUB_ID	CODE	Address	RECCOUNT
26	KD-00001	V	Jl. Pulo Bambu No. 15, Kota Tenggara Lama	3
26	KD-00012	V	Pulo Bambu No. 15, Kota Tenggara Lama	3
26	KD-00778	E	Jalan. Pulau Bambu No. 15 - Kota Tenggara Lama	3
13	EMPL_073	E	Taman Bunga Langit, Jl. Utara No. 3	2
13	KD-00076	V	Taman Bunga Langit, Jl. Utara No. 3	2
16	EMPL_072	E	Taman Bunga Langit, Jl. Barat Laut No. 6	2
16	KD-00099	V	Taman Bunga Langit, Jl. Barat Laut No. 6	2
60	KD-00044	V	Kompleks Pelaut Tangguh, No. 5A	2
60	KD-00492	V	Kompleks Pelaut Tangguh, No. 5A	2
79	EMPL_019	E	Jalan. Pulau Sentosa No. 133	2
79	KD-00066	V	Jl. Pulau Sentosa No. 133	2
81	EMPL_065	E	Lucky Beruntung Apartment , Lt.5 No. 4	2
81	KD-00116	V	Apartemen Lucky Beruntung, Lt. 5 No. 4	2
96	EMPL_054	E	Apartemen Bukit Merah Annex Plaza, Lt 3 No. A1	2
96	KD-00128	V	Apartemen Bukit Merah, Annex Plaza, Lt 3 No. A1	2

11.2. Example of Data Analytics Used in Comparing Government Price from a Procurement Agency

Background Information

In auditing a ministry's procurement, there is a task to investigate a potential of misused discount in the ministry's procurement. In a procurement system, government institution entitled a discount on any product published in vendor's e-commerce website, thus, the price for government is always lower than that for ordinary buyer. To see whether the discount was applied or not, auditors analyzed the data from national procurement agency whom the government should buy the product from. National procurement agency is a marketplace for those who want to sell product to government.

At first, based on previous training on CAAT, the auditor tried to collect the product id and product price from national procurement agency. In the database, there is a product table that consists of product id, product name, and product price for government, product price for non-government, and the product's URL address from vendor's e-commerce website. Since the product is very large, it is not possible for auditor to inspect URL address for each product to compare the government price with published price.

Identifying Target (Question)

In order to perform Data Analytics, auditors decided to collect product price from vendor's e-commerce website and compare the price with the government price. Auditors must use alternative method in collecting data from *public domain*.

Data Readiness

Auditor collected the list of products bought by the ministry and created a simple join with the product table from national procurement agency. The following picture shows the table used by auditor for further analysis. The structure of the table is as follows:

- PROD_ID : Product Identification Number
- PROD_NAME : Product Name
- RETAIL : Product Price for non government buyer
- GOV_PRICE : Product Price for Government buyer
- PROD_URL : URL for the product from vendor's e-commerce website

PROD_ID	PROD_NAME	RetailPrice	Government	PROD_URL
3319126116	Compact Pro 3 [CMP3]	18.100.000,00	16.471.000,00	https://vendors.co.id/compact-pro-3-cmp3-sku3319126116
3317578384	SE 1102 C11	37.323.000,00	35.600.000,00	https://vendors.co.id/se-1102-c11-sku3317578384
3317932822	Liebert [GXT3000L-MTPlus]	20.100.000,00	18.369.500,00	https://vendors.co.id/liebert-gxt3000lmtplus-sku3317932822
3317929815	Liebert [GXT3000-MTPlus]	20.350.000,00	18.598.000,00	https://vendors.co.id/liebert-gxt3000mtplus--sku3317929815
3316442320	LP5-31T [16732]	100.100.000,00	98.280.000,00	https://vendors.co.id/lp531t-16732-sku3316442320
3316442417	LP6-31T [16731]	105.600.000,00	103.680.000,00	https://vendors.co.id/lp631t-16731-sku3316442417
3316395081	Luxriot VMS Enterprise (unlimited channels) [TV-VMS999]	33.157.000,00	31.167.500,00	https://vendors.co.id/luxriot-vms-enterprise-unlimited-channels-tvvms999-sku3316395081
3317569751	DiskStation [DS1517+] - 5 Years Warranty	16.900.000,00	15.990.000,00	https://vendors.co.id/diskstation-ds1517-5-years-warranty-sku3317569751
3319298582	Medium Format Mirrorless Digital Camera Body Only [X1D-50c]	136.000.000,00	133.880.000,00	https://vendors.co.id/medium-format-mirrorless-digital-camera-body-only-x1d50c-sku3319298582
3317290488	Catalyst 2960L 8-Port Gigabit PoE SFP Managed Switch [WS-C2960L-8PS-LL]	15.633.200,00	13.507.000,00	https://vendors.co.id/catalyst-2960l-8port-gigabit-poe-sfp-managed-switch-wsc2960l8psll-sku3317290488
3316442514	LP8-31T [16730]	127.050.000,00	124.740.000,00	https://vendors.co.id/lp831t-16730-sku3316442514
3319176944	9PX 11000VA [9PX11KiRT31] - 3 Years Warranty	90.634.375,00	88.270.000,00	https://vendors.co.id/9px-11000va-9px11kirt31-3-years-warranty-sku3319176944
3318687482	SE 1102C31	42.000.000,00	39.600.000,00	https://vendors.co.id/se-1102c31-sku3318687482
3316442611	LP10-31T [16729]	134.200.000,00	131.760.000,00	https://vendors.co.id/lp1031t-16729-sku3316442611
3317291070	Catalyst 2960L 24-Port Gigabit SFP Managed Switch [WS-C2960L-24TS-AP]	19.373.200,00	16.893.000,00	https://vendors.co.id/catalyst-2960l-24port-gigabit-sfp-managed-switch-wsc2960l24tsap-sku3317291070
3318462442	Speak 810 UC [7810-209]	8.220.000,00	8.203.500,00	https://vendors.co.id/speak-810-uc-7810209-sku3318462442
3316446879	LP10-33 [19263]	161.700.000,00	158.760.000,00	https://vendors.co.id/lp1033-19263-sku3316446879
3316446685	LP15-31 [16716]	161.700.000,00	158.760.000,00	https://vendors.co.id/lp1531-16716-sku3316446685
3316921985	40GBASE-LR4 Single-mode [DEM-QX10Q-LR4]	114.526.070,00	111.500.500,00	https://vendors.co.id/40gbaselr4-singlemode-demqx10qlr4-sku3316921985
3316446782	LP20-31 [16717]	170.940.000,00	167.832.000,00	https://vendors.co.id/lp2031-16717-sku3316446782
3318177456	All-in-One Aspire AC22-860 (Core i5-7200U)	9.299.000,00	9.239.000,00	https://vendors.co.id/allinone-aspire-ac22860-core-i57200u-sku3318177456
3316446976	LP20-33 [19275]	194.810.000,00	191.268.000,00	https://vendors.co.id/lp2033-19275-sku3316446976
3319126213	Compact Pro 6 [CMP6]	39.490.000,00	35.935.500,00	https://vendors.co.id/compact-pro-6-cmp6-sku3319126213
3316928193	DXS-3600 8-Port Gigabit Managed Switch [DXS-3600-16S/ESI]	141.440.217,00	137.704.000,00	https://vendors.co.id/dxs3600-8port-gigabit-managed-switch-dxs360016esi-sku3316928193
3317688382	Taurus 10kVA	153.494.080,00	149.489.500,00	https://vendors.co.id/taurus-10kva-sku3317688382
3316447073	LP30-33 [19283]	239.800.000,00	235.440.000,00	https://vendors.co.id/lp3033-19283-sku3316447073
3318177359	All-in-One Aspire AC22-860 Non Windows (Core i5-7200U)	8.399.000,00	8.339.000,00	https://vendors.co.id/allinone-aspire-ac22860-non-windows-core-i57200u-sku3318177359

Analytic Creation

The common terminology to extract some content from the website is web scraping. Using web scraping, computer accessed all URL address one by one. Computer through an algorithm found the data pattern on the website to extract a component. The following was the result of web scraping combined with the pre-existing data.

New columns are added:

- PublishPrice : Product Price extracted from website
- Diff : The difference between Government Price and Publish Price.

PROD_ID	PROD_NAME	RetailPrice	Government	PublishRate	Diff	PROD_URL
3319126116	Compact Pro 3 [CMP3]	18.100.000,00	16.471.000,00	18.100.000,00	- 1.629.000	https://vendors.co.id/compact-pro-3-cmp3-sku3319126116
3317578384	SE 1102 C11	37.323.000,00	35.600.000,00	37.323.000,00	- 1.723.000	https://vendors.co.id/se-1102-c11-sku3317578384
3317932822	Liebert [GXT3000L-MTPlus]	20.100.000,00	18.369.500,00	20.100.000,00	- 1.730.500	https://vendors.co.id/liebert-gxt3000lmtplus-sku3317932822
3317929815	Liebert [GXT3000-MTPlus]	20.350.000,00	18.598.000,00	20.350.000,00	- 1.752.000	https://vendors.co.id/liebert-gxt3000mtplus--sku3317929815
3316442320	LP5-31T [16732]	100.100.000,00	98.280.000,00	100.100.000,00	- 1.820.000	https://vendors.co.id/lp531t-16732-sku3316442320
3316442417	LP6-31T [16731]	105.600.000,00	103.680.000,00	103.680.000,00	-	https://vendors.co.id/lp631t-16731-sku3316442417
3316395081	Luxriot VMS Enterprise (unlimited channels) [TV-VMS999]	33.157.000,00	31.167.500,00	31.167.500,00	-	https://vendors.co.id/luxriot-vms-enterprise-unlimited-channels-tvvms999-sku3316395081
3317569751	DiskStation [DS1517+] - 5 Years Warranty	16.900.000,00	15.990.000,00	15.990.000,00	-	https://vendors.co.id/diskstation-ds1517-5-years-warranty-sku3317569751
3319298582	Medium Format Mirrorless Digital Camera Body Only [X1D-50c]	136.000.000,00	133.880.000,00	133.880.000,00	-	https://vendors.co.id/medium-format-mirrorless-digital-camera-body-only-x1d50c-sku3319298582
3317290488	Catalyst 2960L 8-Port Gigabit PoE SFP Managed Switch [WS-C2960L-8PS-LL]	15.633.200,00	13.507.000,00	13.507.000,00	-	https://vendors.co.id/catalyst-2960l-8port-gigabit-poe-sfp-managed-switch-wsc2960l8psll-sku3317290488
3316442514	LP8-31T [16730]	127.050.000,00	124.740.000,00	127.050.000,00	- 2.310.000	https://vendors.co.id/lp831t-16730-sku3316442514
3319176944	9PX 11000VA [9PX11KIRT31] - 3 Years Warranty	90.634.375,00	88.270.000,00	90.634.375,00	- 2.364.375	https://vendors.co.id/9px-11000va-9px11kirt31-3-years-warranty-sku3319176944
3318687482	SE 1102C31	42.000.000,00	39.600.000,00	42.000.000,00	- 2.400.000	https://vendors.co.id/se-1102c31-sku3318687482
3316442611	LP10-31T [16729]	134.200.000,00	131.760.000,00	134.200.000,00	- 2.440.000	https://vendors.co.id/lp1031t-16729-sku3316442611
3317291070	Catalyst 2960L 24-Port Gigabit SFP Managed Switch [WS-C2960L-24TS-AP]	19.373.200,00	16.893.000,00	19.373.200,00	- 2.480.200	https://vendors.co.id/catalyst-2960l-24port-gigabit-sfp-managed-switch-wsc2960l24tsap-sku3317291070
3318462442	Speak 810 UC [7810-209]	8.220.000,00	8.203.500,00	8.203.500,00	-	https://vendors.co.id/speak-810-uc-7810209-sku3318462442
3316446879	LP10-33 [19263]	161.700.000,00	158.760.000,00	161.700.000,00	- 2.940.000	https://vendors.co.id/lp1033-19263-sku3316446879
3316446685	LP15-31 [16716]	161.700.000,00	158.760.000,00	161.700.000,00	- 2.940.000	https://vendors.co.id/lp1531-16716-sku3316446685
3316921985	40GBASE-LR4 Single-mode [DEM-QX10Q-LR4]	114.526.070,00	111.500.500,00	114.526.070,00	- 3.025.570	https://vendors.co.id/40gbaselr4-singlemode-demqx10qlr4-sku3316921985
3316446782	LP20-31 [16717]	170.940.000,00	167.832.000,00	170.940.000,00	- 3.108.000	https://vendors.co.id/lp2031-16717-sku3316446782
3318177456	All-in-One Aspire AC22-860 (Core i5-7200U)	9.299.000,00	9.239.000,00	12.699.000,00	- 3.460.000	https://vendors.co.id/allinone-aspire-ac22860-core-i57200u-sku3318177456
3316446976	LP20-33 [19275]	194.810.000,00	191.268.000,00	194.810.000,00	- 3.542.000	https://vendors.co.id/lp2033-19275-sku3316446976
3319126213	Compact Pro 6 [CMP6]	39.490.000,00	35.935.500,00	39.490.000,00	- 3.554.500	https://vendors.co.id/compact-pro-6-cmp6-sku3319126213
3316928193	DXS-3600 8-Port Gigabit Managed Switch [DXS-3600-16S/ESI]	141.440.217,00	137.704.000,00	141.440.210,00	- 3.736.210	https://vendors.co.id/dxs3600-8port-gigabit-managed-switch-dxs360016sesi-sku3316928193
3317688382	Taurus 10kVA	153.494.080,00	149.489.500,00	149.489.500,00	-	https://vendors.co.id/taurus-10kva-sku3317688382
3316447073	LP30-33 [19283]	239.800.000,00	235.440.000,00	239.800.000,00	- 4.360.000	https://vendors.co.id/lp3033-19283-sku3316447073
3318177359	All-in-One Aspire AC22-860 Non Windows (Core i5-7200U)	8.399.000,00	8.339.000,00	12.699.000,00	- 4.360.000	https://vendors.co.id/allinone-aspire-ac22860-non-windows-core-i57200u-sku3318177359

There are two new columns, i.e. Publish Price and Diff. Publish Price is the price of the product extracted from vendor's e-commerce website. Diff is the difference between Government Price and Publish Price. If the value of Diff is Zero or Positive, then the Government Price is greater than or equal to Publish Rate. The condition of Zero or Positive Diff value indicates a violation of government price.

11.3. Data Science/Advanced Analytics Quick Start

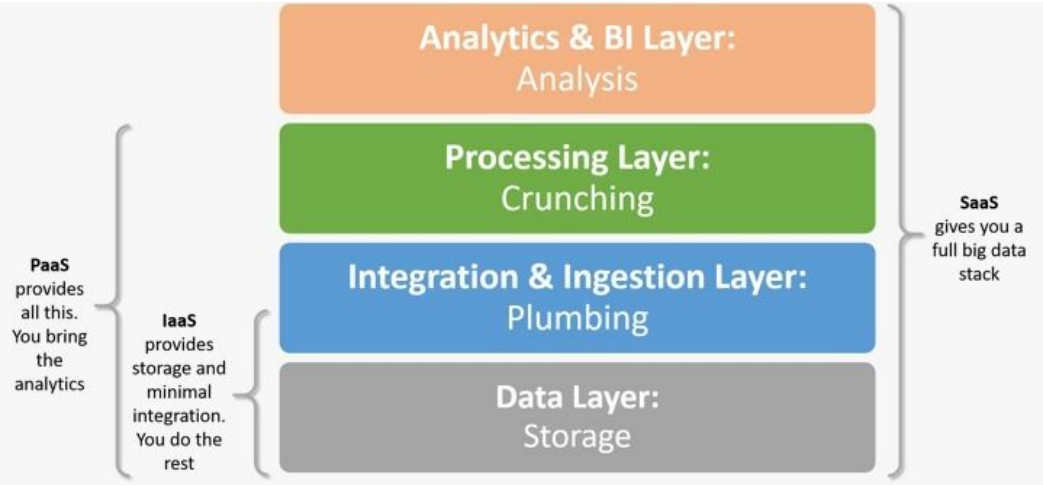
Analytics is a game changer. It is revolutionizing how individuals, businesses, and society can use technology. Analytics is also flexible; it can be used merely to analyze limited data for a single task, or it can change the entire business landscape. The full value of analytics can be realized only when applied to integrated data from multiple sources and when insights are immediate and actionable. As with any organizational capability, analytics should be explored gradually to understand what value could be gained from it, but this exploration should be done in a way that enables it to grow so more value can be obtained. Viewing big data analytics as an ecosystem provides the understanding of how to chart the way to start small while enabling growth to achieve advanced levels of maturity and value. By observing the success or failure in building a big data analytics capability in small and large organizations, several recommendations can be adopted.

Start small but tackle real problems
Common sense dictates developing big data analytics capability in small steps. But this action does not mean applying analytics to an insignificant problem. A real problem can demonstrate the value of analytics and pave the way to address other, perhaps bigger, issues. Starting small could mean leveraging cloud computing platforms to avoid having to invest and set up a huge infrastructure. It could also mean using consultants instead of going through a hiring process that can take a significant amount of time and effort, and may not provide all the skills needed.

Start from where the organization is now, and decide where you want it to be
Although big data analytics is a relatively young discipline, it inherits components from a number of more established areas such as business intelligence (BI), information management, and enterprise data warehouses (EDWs). Hence, many organizations are not necessarily starting big data analytics from absolute zero. If they have capabilities in these areas, they can harvest some of their data, skills, and tools to start their journey. SAIs should consider who their data analytics and BI users are, the people that are going to both use and benefit from analytical tools, and conduct in-depth interviews with them. Find out their use

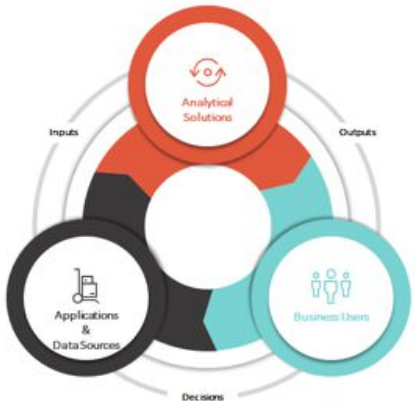
cases, what tool they use, what the tools can and cannot do, and if they have the knowledge required to make the fullest use of their tools.

Each SAI needs to decide where it wants to be and chart its course accordingly. For example, most SAIs may not need their own big data analytics infrastructure and may be able to leverage cloud-based big data analytics such as Big Data Infrastructure as a Service (IaaS), Big Data Platform as a Service (PaaS), Big Data Software as a Service.



Enable connected data, insights, and actions

There are some key obstacles that prevent organizations from maximizing the value that can be derived from big data analytics. These obstacles are disconnected—fragmented, incomplete, and not integrated—data; disconnected insights (hunches and pet theories with no support by data); and disconnected actions (insights that are not actionable). From the outset, make sure the data-to-insights-to-actions loop is closed.

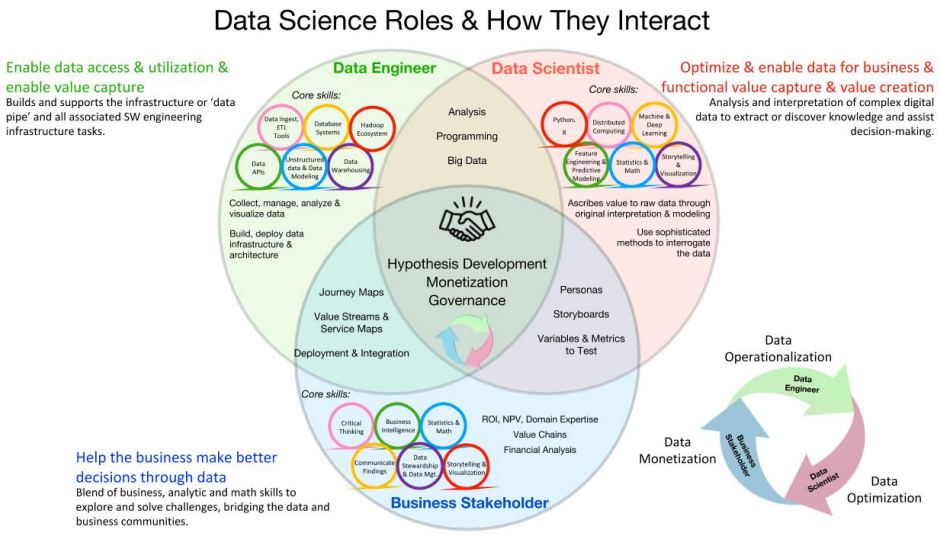


Engage all the key roles of the analytics ecosystem

One key challenge organizations face is recruiting people with the skills to navigate all the data they have access to. Skills shortage in big data analytics is significant and is predicted to escalate. In particular, many organizations are unable to fill the data scientist role that they have deemed so critical for big data analytics. Data scientists have contributed immensely to the development of the big data analytics phenomenon. However, as organizations increasingly embark on this journey, a more strategic, cost-effective, and sustainable approach than current attitudes is needed. A successful initiative depends on more than a single talent. It requires many roles—business and technical, internal, and external—all working collaboratively within a common vision, culture, and architecture. In short, it requires an analytics ecosystem.

The fact that data scientists are very hard to find and expensive are not the only problems. The unbalanced, almost exclusive focus on the role has diverted attention from some key aspects required to establish successful and sustainable big data analytics capabilities. Some organizations have confused the skills with the individual. While the combination of mathematical, statistical, and coding skills are vital for big data analytics, these skills can be acquired and developed across a team of existing staff, not just within a single individual, and augmented by external consultants and modern analytics tools.

One of the reference for a data science team and how they interact is depicted in the following figures[19]



[19] https://www.kdnuggets.com/2018/09/winning-game-plan-building-data-science-team.html

Role, responsibilities, and tools used by specific role is as follows

DATA SCIENCE RESPONSIBILITIES

Role	Data Engineer	Data Scientist	Business Stakeholder
Responsibilities	▪ Collect, manage, analyze, and visualize data ▪ Manage data infrastructure and architecture ▪ Develop dataset processes for modeling, mining, and production ▪ Improve data reliability, efficiency, and quality ▪ Develop data pipeline infrastructure ▪ Develop scale out and scale up solutions	▪ Translate business issues into data science ▪ Ascribe value to raw data through original interpretation and modeling ▪ Interact with data using sophisticated machines ▪ Prepare data for use in predictive and prescriptive modeling ▪ Perform feature engineering and identify hidden patterns in data ▪ Automate using prescriptive and predictive analytics in scale ▪ Engage stakeholders through stories	▪ Help business make better decisions through data ▪ Satisfy business queries using data ▪ Deploy critical thinking when reviewing data ▪ Deploy math skills when reviewing data ▪ Undertake business intelligence activities ▪ Communicate findings using understandable language and content ▪ Data stewardship and metadata management e.g. Definition, business requirement, business rule anda data elements ▪ Subject matter expert
Languages, SW, and Tools	▪ ETL / ELT ▪ Data virtualization ▪ Cleaning ▪ Architecture ▪ Presentation ▪ Presto ▪ MapReduce / MPP ▪ Hadoop Stack ▪ SQL ▪ Python / R ▪ Java ▪ PHP ▪ Visualization ▪ Pentaho ▪ Informatica ▪ Ab Initio ▪ Talend ▪ Unix, linux, shell, perl ▪ Spark, Pig, Hive, Flume ▪ AngularJS, D3 ▪ Kafka ▪ Natural Language Processing ▪ Git ▪ Google Cloud Platform, AWS	▪ TensorFlow ▪ Scala ▪ C / C++ ▪ Python ▪ Theano ▪ Keras ▪ PyTorch ▪ Visualization ▪ Tableau ▪ Caffe2 ▪ Google Cloud Platform ▪ ML libraries, e.g.: Beam, NumPy, SciPy,Pandas sci-kit-learn,dplyr,ggplot2 ▪ Big query ▪ AWS ▪ Linux, Windows, Unix ▪ Apache Hadoop ▪ Spark ▪ Deep Learning ▪ Ensemble ▪ SPSS, SAS, STATA, PHP, Matlab	▪ Math ▪ Statistics ▪ Dashboards ▪ Reports ▪ MS Excel ▪ BI Tools ▪ SQL ▪ Data stewardship ▪ Metadata management ▪ Vbscript ▪ Visual Modeling ▪ Facilitation skills ▪ Lean and agile methodologies

The data scientist role is crucial for a big data analytics program. However, if an organization neglects business stakeholder/the data steward, analysis can be performed on the wrong data, security and privacy considerations can be compromised, or there may be many other undesired business risks and consequences. Without business stakeholder, organizations risk the disconnected data-insights-actions syndrome mentioned previously. In other words, organizations may end up with successful experiments that cannot be put into production or achieve the desired business outcomes.

12. Contributors

1. **Project Team:**

   Team Leader:

   SAI of Indonesia

   Team Members:

   SAI of Bangladesh, SAI of Brazil, SAI of Ecuador, SAI of Georgia, SAI of India, SAI of Iran, SAI of Iraq, SAI of Japan, SAI of Malaysia, SAI of Pakistan, SAI of South Africa, and SAI of USA

2. **Others:**

   SAI of Philippines, SAI of Peru, SAI of Mexico, SAI of Finland, SAI of Afghanistan, SAI of Lithuania, SAI of Germany, SAI of France, and SAI of India who have given valuable inputs during the exposure period.