# Data Analytics Guideline

Prepared for
INTOSAI Working Group on IT Audit

INTOSAI
Working Group on IT Audit

2019

# Table of Contents

# 1. Document Version and Authorization

## 1.1. Purposes

This section of document is to provide the chain of custody of this document and the authorization.

## 1.2. Document Version

| Version | Date | Author | Description |
|---|---|---|---|
| 0.5 | 29 March 2019 | SAI Indonesia | Initial Draft |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

## 1.3. Document Authorization

PIC SAI Indonesia,

***Name***

Date:

***Name***

Reviewed by,

***Name***

*Organization*

Date:

Authorized by,

***Name***

*Organization*

Date:

# 2. Introduction

## 2.1. Purpose

This document provides the audiens with the concept of data analytics and an outline of generic processes of implementing the data analytics practices. Also, this document outlines some considerations for using data analytics in audit processes.

## 2.2. Data Analytics

**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 these forcefully to the analyst". If a model is fit to the data, exploratory analysis finds patterns that represent deviations from the model. These patterns lead the analyst to revise the model, and the process is repeated.

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.

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, proceed side by side".

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.

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 reffered to a Generally Accepted Audit Standar in term of implementation of 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 purpose of Data Analytics in many organizations is to add a competitive advantage by enabling information-based decision making. To ensure the successful of Data Analytics practices, it is important to use a goal-based approach rather than problem-based approach.

In all, Data Analytics enhances the quality of information-based decision-making process. Data Analytics enables 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.

# 3. 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 fraudulen.

  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 follow.

  - How much or How many? → Regression
  - Which Category? → Classification
  - Which Group? → Clustering
  - Is this weird? → Anomaly Detection
  - Which option should be taken? → Recommendation

- 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.

# 4. 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.

## 4.1. Data source identification

The first stage of achieveing 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.

### 4.1.1. 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.

### 4.1.2. 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 Management.

## 4.2. 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, i.e., auditee's premise and public domain.

### 4.2.1. Data type

Data type is the atribut 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 and relevant mathematical sign such as minus, and decimal sign. Some example of this data type are an area of the city, the Invoice amount, and the sample size.

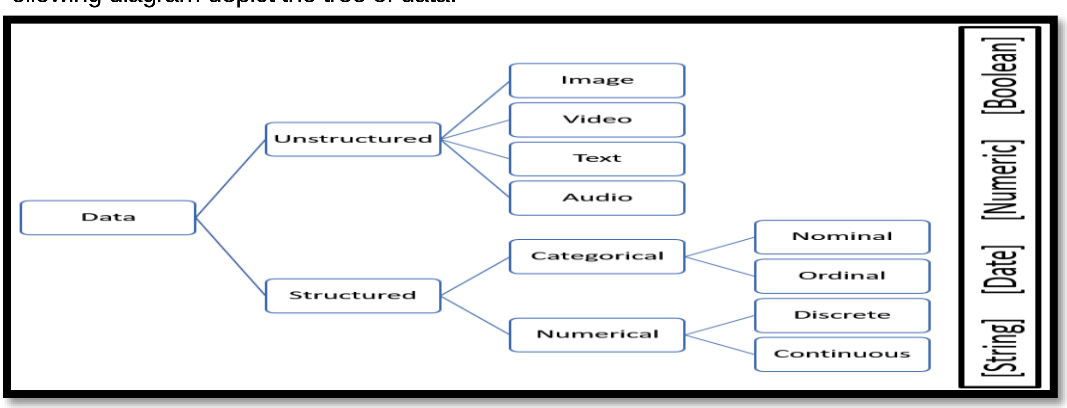
- Date
  This type of data represents a date value such as birthdate, invoice date, and report date.

- 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 structure 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. 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.




1. Data Tree

#### 4.2.2. 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 a read-only acces to the database, backup-restore mechanism, and delivering the requested data through LAN-WAN or VPN.

- 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 of 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 shuld 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 Wifi or Cable. In the same intention, it is possible for auditee to send the requested data to auditors through organization network or through Internet.

#### 4.2.3. 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.

### 4.3. 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

### 4.3.1. Incorrect Data

In this qualifier, information has been incorrectly added to the database. Sometimes, this symptom is identified using our general knowledges 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 12nd, 2018.
- A transaction dated '05/09/2017' was included in a data set of transaction for year 2018.
- Reversed Longitude and Latitude value.

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

### 4.3.2. 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 damage
- The source dataset has been altered by another software
- The source dataset has been extracted in an unadvisable mean.

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.
- Unappropriate 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 then labelled as the missing data.

### 4.3.3. 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.

## 4.4. Data Management

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.

# 5. 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 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 will the proposed business cases.

## 5.1. Model Creation

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

### 5.1.1. Descriptive Analytics

Descriptive Analytics is the process of Data Analytics that creates an overview of the data. Summarizing, Crosstabulation, and Grouping are the common technique to conduct Descriptive Analytics.

Example:

- In Year 2019, Government Revenue from Taxes is 75% of total Government Revenue

### 5.1.2. Diagnostic Analytics

Dignostic 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.

- Information Discovery
  Information Discovery in Diagnostic Analytics enable auditors to trace all data that relate to an anomaly data. Often, Information Discovery requires auditors to look for patterns outside the existing data sets. Also, it might require additional data from other sources.

- 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.

### 5.1.3. Predictive Analytics

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

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.

- 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.

- Proper Algorithm

  There are several algorithms available for conducting predictive analytics. Some of them are:

  - Support Vector Machine
  - Decision Forest
  - Neural Network
  - Linear Regression

- Random Forest

- 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 choosen as the most suitable algorithm.

## 5.2. 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 concensus 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

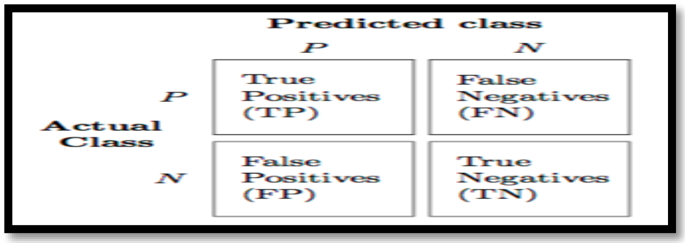
- Tuning statistical feature

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

## 5.3. 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).

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.




2. Confusion Matrix

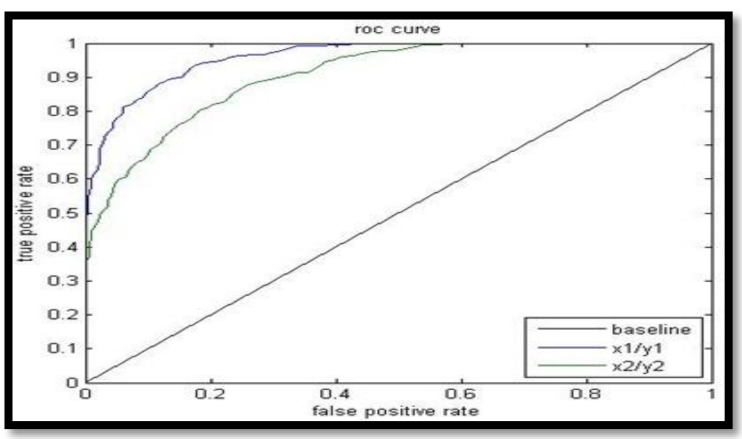
Receiver Operating Characteristic (ROC) is a graph that represents the performance of a classification model at all classification thresholds. This graph plots two parameters; True Positive Rate and False Positive Rate.

- 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 looser the plot to the baseline, the better the plot 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 3, the ROC Chart shows that the algorithm with the blue plot is better than the algorithm with the green plot.




3. 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.

# 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[1]. 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 two types of Data Visualization.

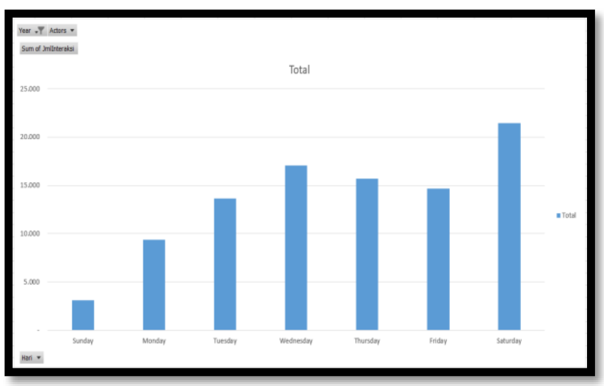
- Static Visualization

  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** |




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

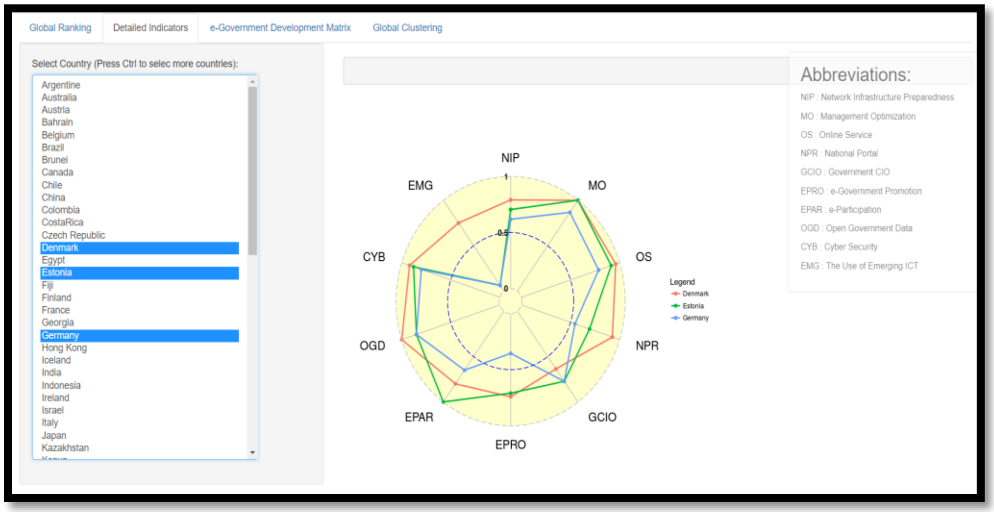
| Country | Salesperson | 2011 Sum of Units | 2011 Sum of Order Amount | 2012 Sum of Units | 2012 Sum of Order Amount | 2013 Sum of Units | 2013 Sum of Order Amount |
|---|---|---|---|---|---|---|---|
| UK | Bromley | 232 | 24,756.89 | 228 | 40,396.64 | 73 | 9,894.51 |
| | Coghill | 81 | 4,029.25 | 39 | 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 interactiviy 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 provide user with the ability to compare one object to others. In this example, comparing Denmark, Estonia, and Germany.




4. Dynamic Visualization

## 6.2. Insight

In information science, there is a concept of the level of humand 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.

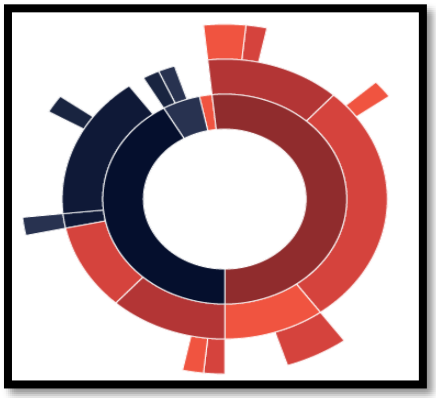
There are various Data Visualization types[2] for exposing some interest information and gaining the insight. The followings are commont types of visualization that related to gaining the insight in auditing. This could helpful for auditor when identifying some irregularities.

- Sunburst Diagram

  A Sunburst Diagram[3] is used to visualize hierarchical data, depicted by concentric circles. The circle in the centre 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.

[2] https://datavizproject.com/data-type/

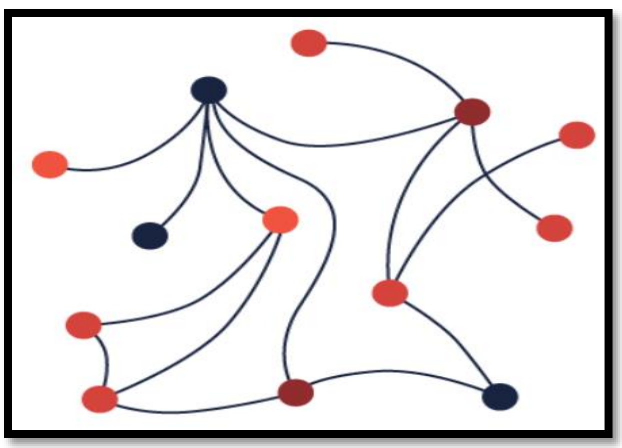
[3] https://datavizproject.com/data-type/sunburst-diagram/



5. Sunburst Diagram

- Network Diagram

  Network Visualisation[4] (also called Network Graph) is often used to visualise complex relationships between a huge number of elements. A network visualisation 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 be otherwise be overlooked.

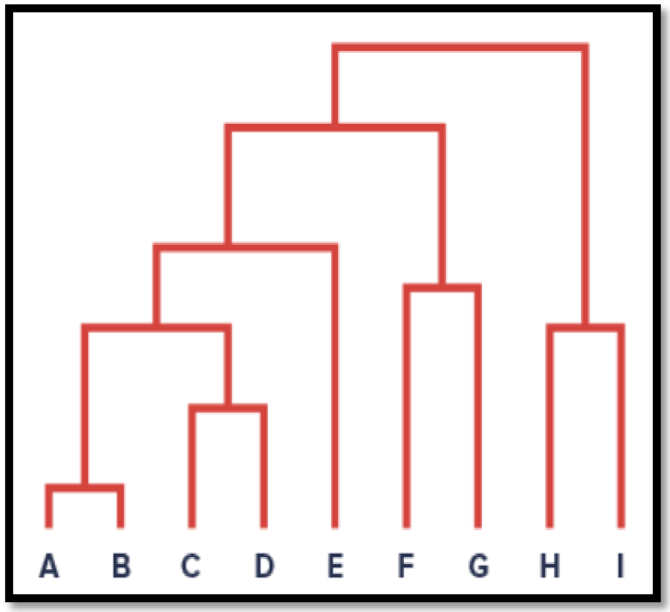


6. Network Diagram

- Dendogram

[4] https://datavizproject.com/data-type/network-visualisation/

A dendrogram[5] is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.




7. Dendogram

## 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.

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

# 7. Analytics Deployment

After analytics appears to be performing satisfactorily, it can be deployed into production for other applications to consume, such as: online websites, spreadsheets, dashboards, line-of-business applications, and back-end applications. This is usually done in one of two ways. Traditionally, the model is turned over to IT Department to translate into a production stack language to prepare for deployment into the production environment. Alternatively, setting up infrastructure that empowers data scientists to deploy models on their own as APIs is an option that's gaining popularity because it eliminates lags between data science and IT Department teams and gets results in front of decision makers faster.

# 8. Data Analytics in Audit

## 8.1. Definition

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."[6]

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.

## 8.2. 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

6 https://www.aicpa.org/InterestAreas/FRC/AssuranceAdvisoryServices/DownloadableDocuments/AuditAnalytics_LookingTowardFuture.pdf

- 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

## 8.3. 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

## 8.4. 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 do not prohibit the use of data analytics techniques. However, the lack of reference to data analytics beyond mention of traditional CAATs in GAAS may be viewed as a barrier to their adoption more broadly.

This lack of reference to data analytics in GAAS also results in some being of the view that gathering information from the use of data analytics does not necessarily reduce the procedures required by GAAS today, even if those required procedures now appear redundant as a result of the information gained from the use of data analytics.

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.

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

### 8.5.1. 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.

### 8.5.2. 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.

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

When DA involve 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.

## 8.7. 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

# 9. Data Analytics Project Management

In general, project management in data analytics (DA) is the same as project management in other activities. There are five processes, i.e. initiating, planning, executing, monitoring and controlling, closing.

## 9.1. 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?

## 9.2. 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

## 9.3. 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

## 9.4. 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 be some issues.

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

## 9.5. 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?

# 10. 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

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

# 11. 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.

## 12. Contributors

1.