INTOSAI
Working Group on Big Data

GUIDANCE ON CONDUCTING AUDIT ACTIVITIES WITH DATA ANALYTICS

September
2022

INTOSAI
Goal Chairs
Collaboration
PSC - CBC - KSC

Quality Assurance Certificate of the Chair of the INTOSAI Working Group on Big Data

This is to certify that ***Guidance on Conducting Audit Activities with Data Analytics*** which is placed at level ***3*** of Quality Assurance as defined in the paper on "Quality Assurance on Public goods developed outside Due Process" approved by the 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 WGBD Members;

(ii) The project was discussed during the 3th WGBD Meeting in Copenhagen in 2019,

the 4th Meeting(Virtual) in 2020 and the 5th Meeting(Virtual) in 2021;

(iii) The draft project output was circulated among team members and WGBD members for several rounds between December 2020 to August 2021; and was exposed for more than 50 days (from 23 August 2021 to 13 October 2021) for final comments.

The product developed 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 till **7 November 2025** and if it is not reviewed and updated by **7 November 2025**, it will cease to be a public good of INTOSAI developed outside the Due Process.

侯. 凯

Hou Kai
Auditor General of the National Audit Office of China
Chair of INTOSAI Working Group on Big Data

INTOSAI

Goal Chairs
Collaboration
PSC - CBC - KSC

Quality Assurance Certificate of the Chair of the Knowledge Sharing and Knowledge Services Committee

Based on the assurance provided by the **INTOSAI Working Group on Big Data (WGBD)** and the assessment by the Goal Chair, it is certified that the **Guidance on Conducting Audit Activities with Data Analytics** which is placed at level **3 (three)** of Quality Assurance as defined in the paper on "Quality Assurance on public goods developed outside Due Process" approved by the 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 till **7 November 2025** and if it is not reviewed and updated by **7 November 2025**, it will cease to be a public good of INTOSAI developed outside the Due Process.

Girish Chandra Murmu
Chair of Knowledge Sharing and
Knowledge Services Committee

FOREWORD

Emerging technologies including artificial intelligence, big data, and cloud computing make it possible for supreme audit institutions to collect data with wider coverage, transmit data in a faster speed, process data in a smarter way, and produce analysis results of higher quality. While shifting from the traditional way of validation to exploratory analysis in managing data, auditors are now able to analyze data from multiple levels and perspectives, and dig out more valuable information behind data. Remote data analysis also offers an effective means for supreme audit institutions to remain resilient in the face of public calamities.

We are pleased to present this Guidance on Conducting Audit Activities with Data Analytics, developed by the INTOSAI Working Group on Big Data, in a hope to help auditors verify, clean and analyze data promptly and effectively. The guidance identifies the concept, target, process, staffing abilities, and working mechanism for conducting audit activities with data analytics, and introduces the information system investigation with examples. The guidance explains the principle, mode, method, and source of data collection, and touches upon data standardization during data collection. The guidance further demonstrates how to develop analysis plans, and to analyze data using parsing tables, data analysis techniques and tools. Finally, the guidance lists the emerging risks in using data analytics for auditing, and discusses the issues of quality control and data security.

We would like to express our thanks to the Secretariat of the INTOSAI Working Group on Big Data who took the lead in planning and drafting the guidance. We also appreciate the SAIs of Austria, Ecuador, India, Peru, the Philippines, Thailand, and Türkiye for their efforts in preparing respective chapters. Our thanks also go to the SAIs of Estonia and Mexico for their valuable comments on the guidance.

We hope you find this guidance useful.

侯凱

Hou Kai

Auditor General of the National Audit Office of China
Chair of INTOSAI Working Group on Big Data

About INTOSAI Working Group on Big Data

INTOSAI Working Group on Big Data (WGBD) is a specialized working group approved by INTOSAI under Strategic Goal Three: Knowledge Sharing and Services. Its objectives are to identify the challenges and opportunities faced by SAIs in the era of big data, to summarize the knowledge and experience in the field of big data audit, and to strengthen bilateral and multilateral technical cooperation on big data.

Chair

China

Vice Chair

United States

Members

- Argentina
- Austria
- Bangladesh
- Belgium
- Bhutan
- Brazil
- Canada
- Chile
- Denmark
- Ecuador
- Estonia
- Finland
- Fiji
- France
- Georgia
- India
- Indonesia
- Kuwait
- Malaysia
- Mexico
- Netherlands
- New Zealand
- Norway
- Pakistan
- Peru
- Philippines
- Portugal
- Republic of Korea
- Russian Federation
- Samoa
- Senegal
- Slovakia
- Thailand
- Türkiye
- Ukraine
- United Kingdom
- Zambia

Observers

- AFROSAI–E
- Bulgaria
- European Court of Auditors
- Ireland
- Vietnam

CONTENTS

01 Overview of Conducting Audit Activities with Data Analytics

02 Data Collection and Preparation

03 Data Analytics and Utilization

04 Quality Control and Data Security

05 Accessory

01 Overview of Conducting Audit Activities with Data Analytics

1.1 Concept

The "Analysis and Utilization of Electronic Data" in the Practice Guideline means that audit institutions comprehensively integrate multi-industry electronic data (hereinafter referred to as data) around audit objectives and scope, investigation items, research, etc. for analysis and utilization.

1.2 Target

Data is correlated, analyzed and utilized to improve the auditors' capabilities for investigation, audit evidence collection, evaluation and judgment, and macro analysis with big-data technology, methods and tools as well as economic activity-related data in huge volume, multiple sources and diverse formats such as financial revenue and expenditure.

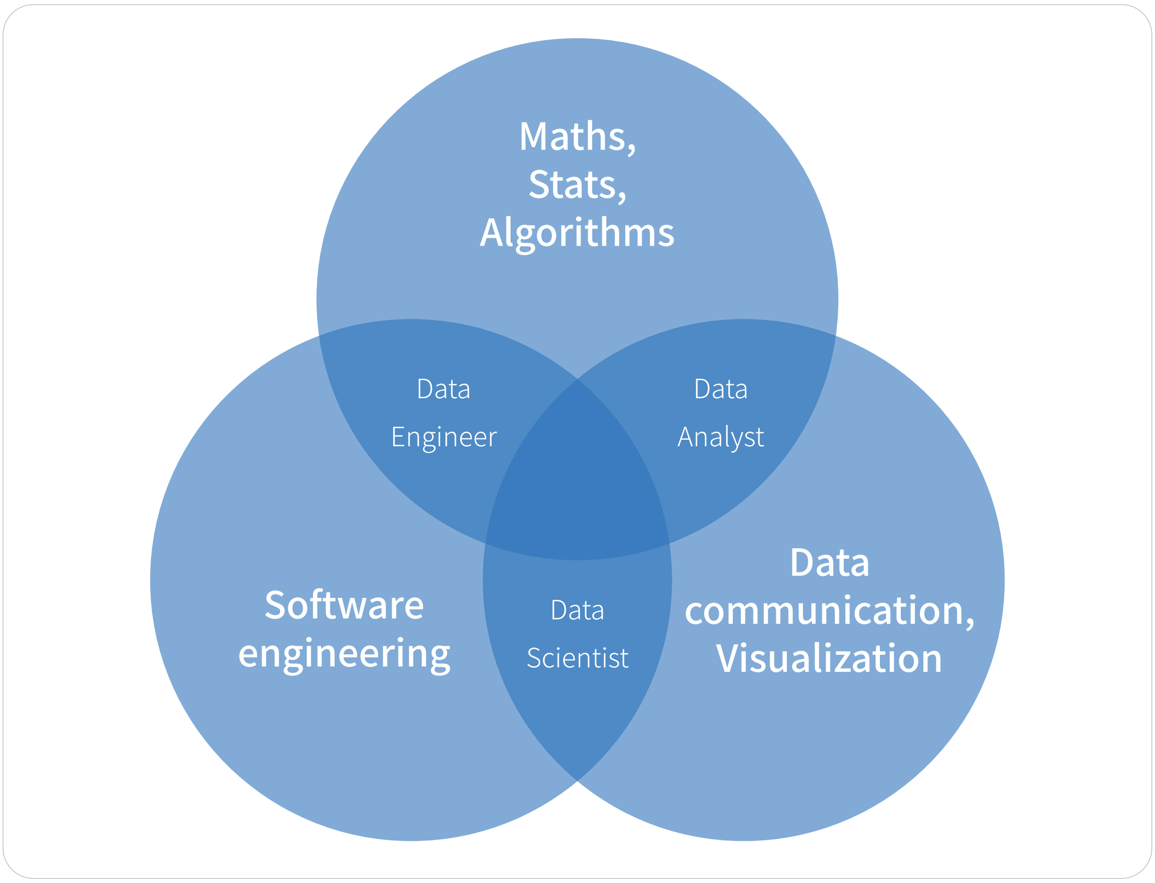
1.3 Process

Data analysis and utilization shall be carried out in accordance with the digital audit mode of "overall analysis, doubt identification, diversified verification, and systematic research".

Auditors shall effectively and timely handle data recovery, verification, cleaning, sorting, and processing.

1.4 Staffing Abilities and Team Requirements

1.4.1 Staffing Abilities

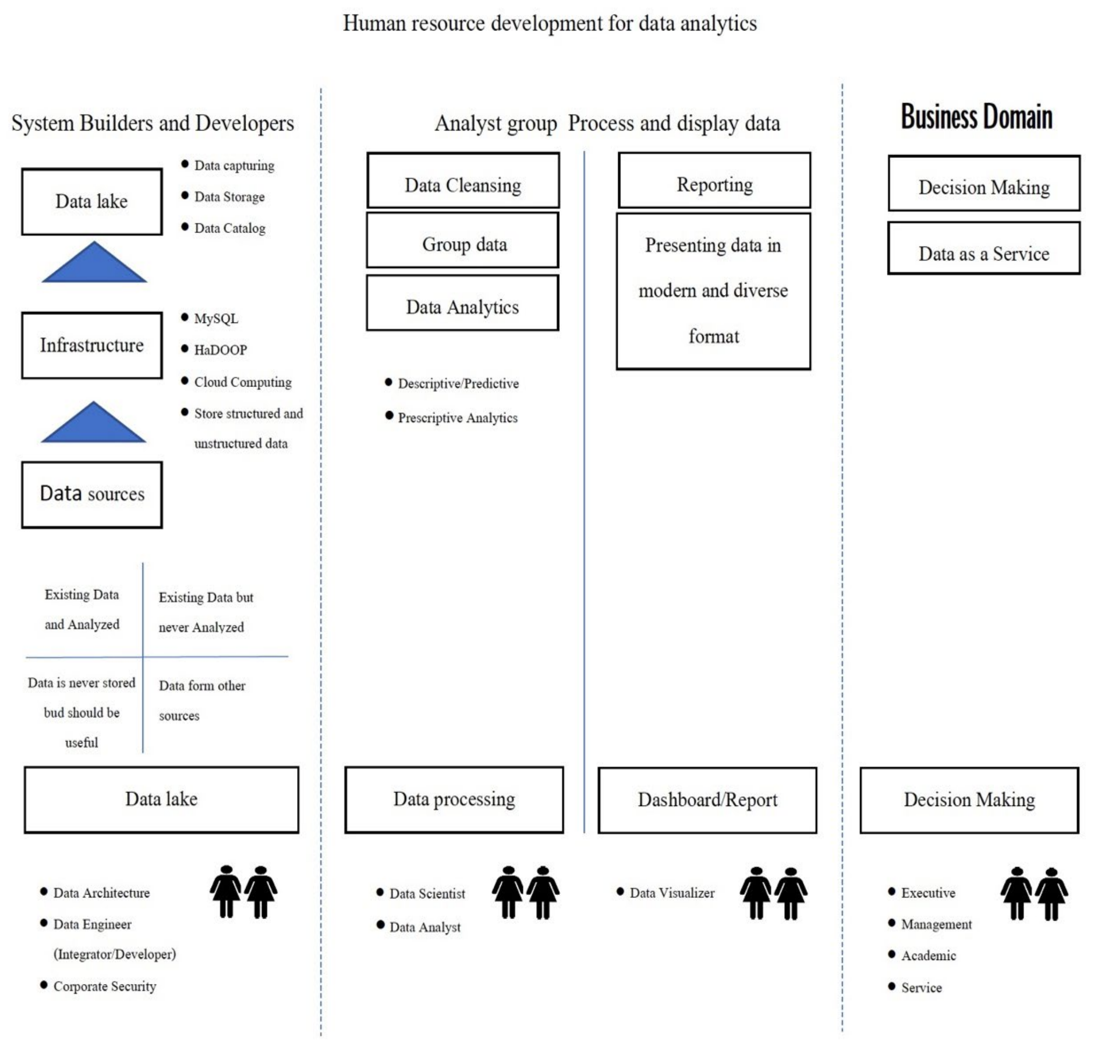


For successful data analysis, it is important to jointly work with an expert of the knowledge domain (the audit topic), a data analysis expert and a visualization expert.

The data analysis expert needs to know one or more programming languages (e.g., R, Python, C++, …) and some statistics. A visualization expert needs to master the relevant techniques and main tools of data visualization. All other related personnel need to master professional competence related with big data and computer.

audit institutions may carry out pre-audit training for auditors, so as to inform relevant personnel of information obtained from investigations and ensure a more targeted audit.

1.4.2 Team Requirements



Throughout data analysis, at least three teams shall be involved: data lake building team, data analyst team and decision-making team.

The data lake building team needs to collect, backup, restore, and store data with appropriate technologies and methods according to certain principles. This may require the participation of data architectures, data engineers, corporate securities, and so on.

The data analyst team needs to obtain the relevant data authorization according to the audit purpose for data cleaning, grouping, processing and analysis to present the analysis results and report to the decision-making team. This may require the participation of data scientists, data analysts, data visualizers, etc.

The decision-making team needs to evaluate the data analysis report and make the work plan. This may require the participation of all the auditors in the project. And these auditors should embrace professional competence, such as audit theory, relevant practices, and data analysis skills.

1.5 Working Mechanism

Reproducible audit is a working mechanism that auditors prefer to adopt. It is an audit conducted and documented in a manner that every auditor of a SAI can carry out the same analysis with the similar data and obtain the same result (derived from reproducible research).

◉ **There are four basic steps in each data analysis project:**

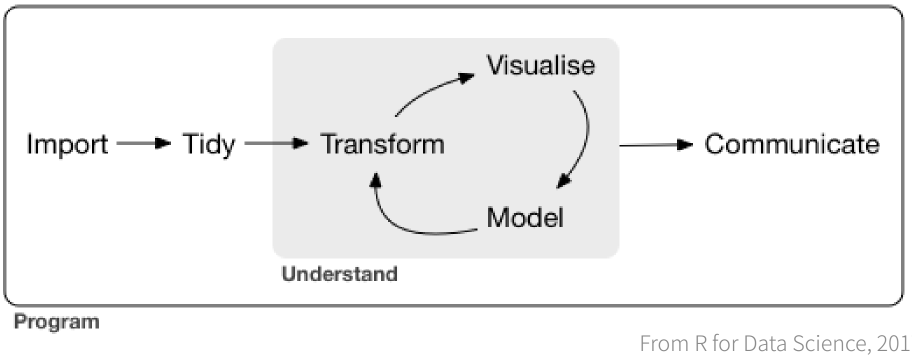
1. Collection of data for analysis
2. Data cleaning
3. Implementation of data analysis
4. Presentation of data analysis results

It should be part of a reproducible workflow, so literate programming tools (with the source code and documentation together saved in one file) are recommended to make sure that all steps of data analysis are reproducible. Literate programming tools including Python and R are quite helpful for quality control because traceability is essential for the work of the quality assurance representative.

◉ **Here are three basic recommendations for documentation of an audit project:**

1. Everything needs to be documented
2. All files for documentation should be human readable
3. A clear and easy-to-understand structure for documentation is required

1.5.1 Model of the Tools Needed in a Typical Data Science Project



From R for Data Science, 2017, p. IX

◉ Importing data

Importing data typically means that you take data stored in a file, database, or web application programming interface (API), and load it into a data frame in some data analysis software such as R. It is often a time-consuming step with many technical problems.

◉ Tidying data

> *"Tidying your data means storing it in a consistent form that matches the semantics of the dataset with the way it is stored."* ①

> *"Apart from tidying, there are many other tasks involved in cleaning data: parsing dates and numbers, identifying missing values, correcting character encodings (for international data), matching similar but not identical values (created by typos), verifying experimental design, and filling in structural missing values, not to mention model-based data cleaning that identifies suspicious values."* ②

◉ Transforming data

After tidying the data, you can transform the data:

- focusing on particular observations
- creating new variables from existing variables (like calculating debt to asset ratio from total assets and total liabilities)
- calculating summary statistics (like means)

Visualization of data

"A good visualization will show you things that you did not expect, or raise new questions about the data. A good visualization might also hint that you are asking the wrong question, or you need to collect different data." [3]

"Excellence in statistical graphics consists of complex ideas communicated with clarity, precision and efficiency. Graphical displays should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- be closely integrated with the statistical and verbal descriptions of a data set [4]

Modelling of data

A model is a summary of a data-set, for example, to assess if the collected data is related. Communicating the results of data analysis

Results of data analysis have to be clearly communicated, so different stakeholders can understand them. Expert of the knowledge domain (= the audit topic) and data analysis have to tell the story of data they collected and analyzed. To tell the story of data in an exciting manner, they have to

- understand the context,
- choose a meaningful visualization,
- eliminate unimportant data,
- focus attention to the essential messages.

[1] from R for Data Science, Hadley Wickham, p. X

[2] from Journal of statistical software by Hadley Wickham, August 2014, Volume 59, Issue 10

[3] from R for Data Science, Hadley Wickham, p. X

[4] from The Visual Display of Quantitative Information by Edward R. Tufte, p.13

A data analysis and utilization plan should be in place to: preliminary investigation, interpretation of relevant policies and regulations, objectives and ideas of data analysis and utilization, measures of data analysis and utilization, and data security.

1.5.2 Possible Challenges in Audits with Data Analysis

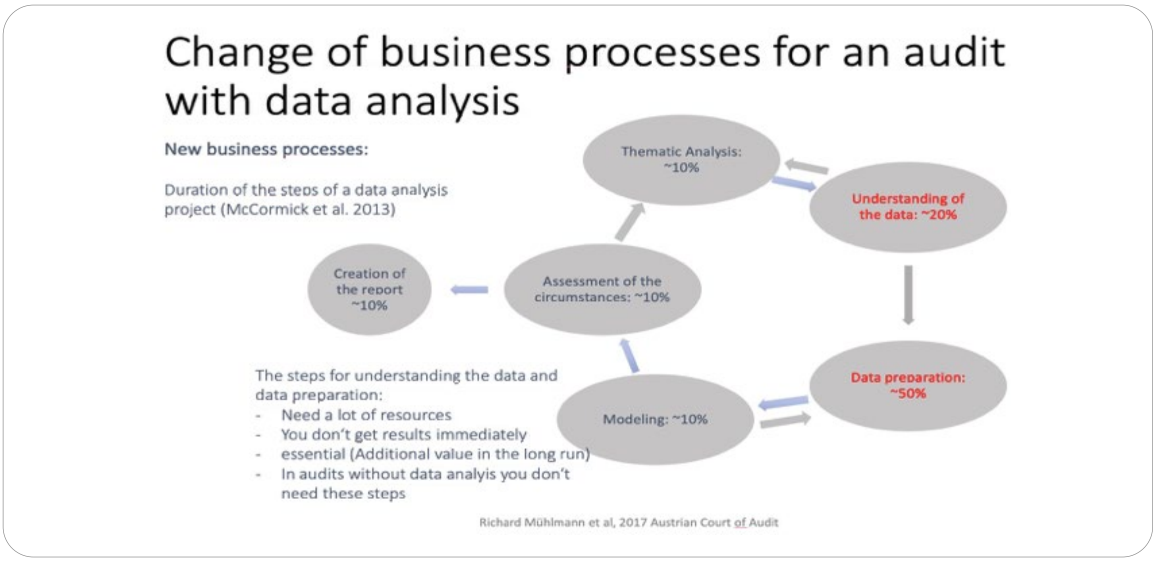
Change of the organizational structure of a SAI. For example, it may be necessary to establish an additional data analysis department and include one or more data analysis experts in each audit team.

Challenges for employees. For example, it may be necessary to carry out training on data analysis and hiring new employees with expertise in Data Analysis.

Change of Business processes. For example, more time will be needed for the audit preparation, so does the duration of an audit because data analysis is often time-consuming.

Data privacy and confidentiality. Performing big data analysis must conform with the national data privacy policies, particularly on sensitive personal data or personally identifiable information.

Compatibility issues. For example, the use of legacy systems and modern systems may result to ineffective data analysis if data is not available in expected formats and outputs.



Change of business processes for an audit with data analysis

New business processes:

Duration of the steps of a data analysis project (McCormick et al. 2013)

Thematic Analysis: ~10%
Understanding of the data: ~20%
Data preparation: ~50%
Modeling: ~10%
Assessment of the circumstances: ~10%
Creation of the report ~10%

The steps for understanding the data and data preparation:
- Need a lot of resources
- You don't get results immediately
- essential (Additional value in the long run)
- In audits without data analyis you don't need these steps

Richard Mühlmann et al, 2017 Austrian Court of Audit

1.5.3 Recommendations

A data analysis and utilization plan should be in place to include: preliminary investigation, interpretation of relevant policies and regulations, objectives and ideas of data analysis and utilization, data analysis and utilization measures, data security, etc.

1.5.4 Steps for Trainings

1.5.4.1 Step 1 - Defining the Problems/Policy Issues

◉ **Working Mechanism**

Providing knowledge on how to use big data.

The development schedule is divided into 3 groups as below:

1. Information decisions shall be implemented by the group of data users consisting of chief executives, directors, policy and academic workers, and inspectors;
2. Data shall be analyzed and presented by the group consisting of data scientists, data analysts, and data visualizers responsible for data grouping, analyzing and processing as well as presentation format development including a dashboard for data presentation;
3. System construction and development shall be carried out by the group consisting of data engineers, data architects, business analysts, project managers, corporate security, and IT operators responsible for designing and developing data infrastructure including maintenance and data management for continuous and safe use.

◉ **Related agencies**

State Audit Development Institute

1. Developing capabilities of data users, including chief executives, directors, policy and academic personnel, and service workers, in being able to define the problem or desired issue, analyze the data, and appropriately apply the results, especially, information technology workers of relevant agencies shall be at least able to supervise project management and realize safe data management and governance.
2. Establishing a knowledge management system and collecting operational data to facilitate continuity in operations.

1.5.4.2 Step 2 – Recruitment – Engaging Personnel in the Project

◉ **Based on a good understanding of the data analysis skill level, personnel can be divided into 3 main groups as follows:**

- The group equipped with personnel able to manage information systems and capable of data analysis, processing and visualization, but still need technical support or implementation in some aspects.
- The group under development with personnel understanding and able to provide information of needs at certain levels clearly and in need of technical support for system building and development as well as data analysis, processing and display.
- The group without information personnel but in need of using the information for policy decision making or management.

Based on understanding, personnel with specific skills and expertise from various departments can be pooled to deal with the same project in a working style of "Agile Team". Meanwhile, external consultants or private agencies may join as needed.

1.5.4.3 Step 3 – Human Resource Development in Project-based Learning

◉ **Personnel skills are mainly developed through learning in real work are as follows:**

- Setting possible targets for big data in accordance with the organization's key mission and relevant policies for an alignment between the project and the government's goal.
- Defining a curriculum framework for the data science skills development project which contains big data analysis and management with process steps including materials, pedagogy, learning outcome and assessment.
- Requiring the trainees to develop skills in practice and use results obtained with big data to analyze the benefits for the job and present to executives.

1.5.4.4 Step 4 – Using Information

◉ **Development results obtained according to Step 3 can be used by different groups:**

- The user group consisting of executives and the big data working group, which jointly determine the appropriate project inception and explore data according to the specified pilot problem in order to assess the possibility and transform the problematic needs into data analysis and problem analysis.
- The analysis group studying data distribution and exploring the relationship between variables with samples, which needs real data samples to specify additional data through exploratory data analysis and prepare data to create models or mathematical models for prediction using various techniques and algorithms as well as test the accuracy of the predictive analytical model and design the display method with appropriate data dimensions for the team to try and communicate with the management team and , in order to convert plans into further development on the data visualization dashboard.
- The information system construction and development group, which develops programs according to the mathematical model, and sets enables the program to automatically process the model according to the planned frequency and implementation and deployment.

02 Data Collection and Preparation

2.1 Information System Investigation

2.1.1 Definition of the "Information System Investigation"

The "Information System Investigation" refers to investigations and analysis carried out by audit institutions and auditors as required to get acquainted with the information system and relevant business processes of the auditee.

2.1.2 Content of the "Information System Investigation"

Below are some examples of information systems:

◉ **From audited entities:**

1. Management system;
2. Logistics system (acquisitions);
3. Finance system, payment system, etc.;
4. Budget System
5. Information system of the auditee and its security level;
6. Data of the auditee;
7. Main business processes of the auditee and its dependence on the information system;
8. Management agency and mode of the auditee related with the information system;
9. Operation environment of data analysis and utilization by the auditee;

10. Changes in the information system of the auditee over the previous year;

11. Feasibility of correlation analysis between the data of the auditee and that owned by the audit institutions.

- **From public entities:**

1. Audits system (from SAI);
2. Tax system administration: Single taxpayer registration;
3. State entity purchases;
4. National budget and execution;
5. Central risks;
6. Public records;
7. National civil registry: Citizen identification information;
8. Sworn declaration of income, assets, rents, interests, etc.

2.1.3 Identification of the Auditee’ s IT Systems

It is recommended to identify the auditee’ s IT systems (sources of potential data sets for the SAI) and collect background information (high level architecture, data structure and formats, business process of the auditee and its dependence on the information system, the operation environment of data analysis and utilization, etc.)

2.1.4 Assurance about the Quality and Reliability of Datasets

It is possible to obtain assurance from IS audits of controls (by the SAI, the internal auditor or any other source). In the absence of such assurance, necessary caveats have to be appended to the analytics insights drawn from such data.

2.1.5 Applicable Situations of the "Information System Investigation"

◉ **The Guideline is applicable to information system investigations as follows:**

1. Use of information systems such as accounting software and preservation of the accounting information in the form of data by the auditee;
2. Use of information systems such as business management software and preservation of the information on economic activities such as main business, other business and auxiliary system in the form of data by the auditee;
3. Investigations particularly listed in the audit plan;
4. Information system investigation on e-government and other projects.

2.1.6 Personnel Requirements of the "Information System Investigation"

Personnel involved in an information system investigation shall be qualified with professional competence related with big data and computer.

Computer professionals from audit institutions or outsourced companies shall be involved in the investigation if necessary.

2.1.7 Achieved Effect of the "Information System Investigation"

◉ **Investigators shall collect relevant documents focusing on the following items, and put forward useful opinions and suggestions:**

1. Audit objectives;
2. Audit items and priorities;
3. Audit items that have a great impact on audit objectives;
4. Importance and audit risks;
5. Organization mode and operating methods for data analysis and utilization;
6. Software, hardware, and data techniques and solutions necessary for audit;

7. Quantity and composition of auditors necessary for data analysis and utilization;

8. Scheduled starting and ending time, and budget for the audit work.

2.2 Data Collection

2.2.1 Principle of Data Collection

◉ **Useful, integral, accurate and timely.**

(1) Useful. Based on a comprehensive investigation of information systems, the audit team shall properly determine the range of collection according to the data analysis and utilization plan, and collect data around audit objectives;

(2) Integral. Collected data shall be satisfactory for audit, span continuous time slice, and cover an integral region (sector), so as to facilitate data analysis;

(3) Accurate. The authenticity of data shall be ensured by management means such as data transfer commitment and the accuracy of data shall be ensured by technical means such as data verification;

(4) Timely. Data collection and verification shall be completed at the earlier stage of the audit project. The deadline for submitting collected data shall be clarified to the auditee as required.

Data collection shall not interfere with the normal activities of the auditee as much as possible. In addition, the data shall not be directly collected from the production database.

2.2.2 Content of Data Collection

◉ **The audit team shall comprehensively collect data and necessary technical documents related to duty performance, mainly including financial data, management data, and business data.**

(1) Financial data. Auditors shall collect the complete financial data of relevant years and departments as required, and recover them in the audit database by computer professionals using data migration tools to generate a standard table for financial accounting;

(2) Management data. Management data is mainly composed of basic information and human resource information, which can be fully backed up by human resource management software, office system, etc., and the audit team may prepare an intermediate table for the auditee to accurately fill in;

(3) Business data. Before collecting business data, the audit team shall familiarize with the business process and information system functions of the auditee, require it to provide relevant business data and data dictionary, get acquainted with the name of data sheet, the meaning of key fields for key business, and the relationship between the tables, verify whether the background data structure by sampling the foreground data can be accurately comprehend, and generate a standard in the audit database after selecting the collection range upon data export, backup and recovery for data analysis.

2.2.3 Preparations for Data Collection

Before collecting data, the audit team shall clarify data confidentiality, and prepare different data collection plans based on the security level.

Ownership of data

MoUs, if necessary with the auditee

The auditee shall provide a data transfer commitment with a list of data.

2.2.4 Mode of Data Collection

The mode in which data will be collected – e.g., through APIs, SFTP transfer using XML/JSON (normally digitally signed), ODBC or equivalent connectivity, database restoration (in RDBMS/ERP formats), or conversion from other formats (PDF, Excel, CSV, etc.)

Existence (or otherwise) of an audit module developed as part of the auditee allows for exception-based querying and analytics insights (perhaps through external modules) (India)

Periodicity of data collection – Continuous (API-based) or otherwise

2.2.5 Method for Data Collection

It is recommended to export structured data (e.g., data in relational databases such as Oracle) through original database backup. For complex systems, an intermediate table shall be prepared, with data collected by foreground or background export. In contrast, unstructured data (Word, Excel, PDF, etc.) shall be copied completely, processed, and stored by different departments (regions) and years to facilitate future management.

◉ **The data can be collected by original database backup, foreground data export, or an intermediate table generated by the auditee.**

(1) Original database backup. Financial data and business data of a simple system shall be collected by original database backup. The audit team may require the auditee to provide technical professionals for data export and recovery in the audit database;

(2) Foreground data export. Information systems featuring small volume of data and complex backup and export (e.g., audit institutions can hardly recover data) can be collected by foreground export. Attention should be paid to the integrity of data while exporting, after which all data shall be integrated into the database;

(3) An intermediate table generated by the auditee. For complex information systems or those with large volume of data, an intermediate table shall be prepared as required together with relevant business data by the auditee, the latter of which in the information system shall be fully considered while preparing the intermediate table. While collecting data for the intermediate table, integrity shall be checked. The auditee shall also provide a script to generate the intermediate table, and the auditor shall logically check the table generated.

2.2.6 Source of Data Collection

A primary source of data collection is web services developed by the auditee or a public entity. Information consulting can be done in the entity's web system.

Besides, web scraping can be made to the auditee's web.

In many cases, it is ideal to equip with a legal framework that allows a SAI to access (preferably unrestricted access) information, or make agreements with public entities to get the fastest possible access to their respective databases.

2.2.7 Data Storage / Repository

(1) Data warehouse/repository development and maintenance;

(2) Data virtualization platform (for integrating multiple sources of data, creating a data abstraction layer, enabling information delivery/self-services and metadata/ schema management, practicing data governance, etc.);

(3) Architecture that supports a data analytics platform on top of the warehouse/ virtualization;

(4) Licensing arrangements for multiple database formats (if data is stored in native RDBMS formats);

(5) Archival arrangements;

(6) Infrastructure and scalability.

Conceptually, one of the questions that auditors need to address is "how do we prevent the data repository from becoming a data swamp ".

2.2.8 Classification and Security of the Auditee' Data

(1) Considering how the SAI maintains appropriate security over the auditee' data (broadly not less than the security level for the auditee' IT systems).

(2) Identifying personal and sensitive personal data/personally identifiable information, which may need to be handled depending on the data privacy regulations in force.

(3) Considering whether the auditee' data need to be anonymized before being handed over to the auditor.

(4) Arranging for encryption/HSM (Hardware Security Module) as needed.

(5) Getting access to governance – profiles, roles and mapping, logging etc.

(6) Ensure the traceability of the chain of custody of data, so that they are capable of being maintained as evidence.

2.3 Data Preparation and Standardization

2.3.1 Data Preparation

(1) After obtaining the data, it is essential to understand the data first by inquiring owners (entities to which the data corresponds). It is recommended to ask public entities to train auditors the usage of their respective systems and data structures they send, and provide auditors with all technical information that allows them to consult, analyze and process the data.

(2) Before data cleaning, sorting, processing and analysis, it is necessary to get a comprehensive understanding of the structure and content of data sheets and the relationship between them by the ways as follows:

- Consultation. Consulting the design of information systems, database design, design documents of data dictionary, and metadata or dictionary table in the database;
- Inquiry. Inquiring the information system administrator or software developer;
- Analysis. Familiarizing with the business processes and their logic relationships, analyzing the data handling process, and finding out the correspondence of the data sheets based on business flow charts, user manual for information system, and other documents;

- Comparison between foreground and background data. Finding out the correspondence of data elements in the background data sheets based on the foreground interface and functions of information systems;
- Review. Retrieving the original paper archives and vouchers, and reviewing the data elements in the corresponding sheets.

(3) ETL (Extract, Transform and Load) – High level strategy and approach, tools, implementation

(4) Identifying relevant tables/fields of interest

2.3.2 Data Standardization

(1) Promoting data standardization. If the plan for audit data of a certain sector is unavailable, auditors shall prepare a standard table by processing the original data and extracting data.

(2) Recovering and inputting structured, semi-structured and unstructured data.

(3) Cleaning and standardizing data – Removing outliers; logging cleaning.

(4) Integrating multiple data sets; merging and splitting data sets.

(5) Identifying the different platforms and applications used to understand the data, its entry points, and the normalization process.

2.4 Data Verification

2.4.1 Objective of Data Validation

To ensure the authenticity, integrity, relevance, usability and security of data sets while collecting data [①]. The approaches to data verification will obviously vary, if auditors use API-based extraction of data, as opposed to restoration of data from dumps / files, etc.

2.4.2 Methods of Data Verification

Data verification methods include comparing the quantity and volume of data with those of data in the data list provided by the auditee, verifying whether there is any omission, and checking the integrity constraints of relational models, etc.

[①] Authenticity - Data is created through the process it claims. Integrity - Data is complete, accurate and trustworthy. Relevance - Data is appropriate and relevant for an identified purpose. Usability - Data is readily accessible in a convenient manner. Security - Data is secure and accessible only to authorized parties.

For example, to check the identity documents of persons in the Civil Registry. There are identity documents not corresponding to the persons' names and certain inconsistencies of one or two figures (possible typing errors in the origin system. If such data is not validated, the analysis and subsequent results may not be valid for involving irrelevant persons when typing (by mistake) their identity documents in the source system. Obviously, it is much time-consuming to validate and correct the identity documents of thousands of persons.

◉ **Data can be verified using the following methods:**

(1) Comparing the quantity and volume of semi-structured and unstructured data with those of data in the data list provided by the auditee to verify whether there is any omission, and opening the file to verify whether the data is available and complete;

(2) Checking the integrity constraints of relational models, including physical constraints, primary key constraints, referential constraints, user-defined integrity constraints, etc.;

(3) Checking whether the data is consistent with the original information (such as accounting vouchers and business documents);

(4) Checking the total quantity of data and statistical indicators of main variables, including total quantity of records; checking whether the quantity of data sheets and that of records in the sheet are consistent with those in the data list provided by the auditee, whether the fields of data sheets are complete, whether there is any digital gibberish in data sheets and whether there is any omission of key field values; verifying the authenticity of values in the data sheets through calculation and aggregate. For example, checking whether the range of main variables is abnormal and consistent with those in the statement through the calculation of the maximum and minimum values of main variables, aggregate, and comparison with the business statement;

(5) Verifying the business rules, including the balance between debit and credits, discontinuous and repeated figures, date range, articulation, constraints of laws and regulations, etc.;

(6) Other applicable methods.

03 Data Analytics and Utilization

3.1 Analysis Plan Development

3.1.1 Preliminary Research

At this stage, it is necessary to gain a general knowledge of the target industry or sector together with its features, internal operations and external environment, in order to determine more appropriate programs and audit procedures for the entity, number of auditors to employ, estimated runtime, and the cost of reformulating the audit budget.

Auditors should consider the results of previous audit work related directly to the objectives and scope of the present audit to determine its feasibility for the target set and identify critical areas within the purpose of the exam.

The results of audit work are related to internal control principles, management principles and auditing standards issued by SAIs, the existing legal framework, directives and regulations issued by the systems of budget, treasury, public debt and government accounting, as well as other provisions for internal control.

These guidelines should be simple and clear in presentation and around a specific issue; they must be flexible, adaptive and regularly renewable with the progress in the modernization of government administration, orientation to achieve social or political goals, and use of public funds efficiently and effectively.

3.1.2 Interpretation of Law, Regulations, Rules and Policies

The objectives, set by the team, are clearly defined and must be socialized by all the work-related responsible personnel.

They are always the foundation of analysis and determine its scope, scheduling procedures, implementation and responsibilities. They can range from diagnosis to data evaluation.

- **The following aspects must be verified:**

(1) Reliability of information;

(2) Asset protection;

(3) Fulfillment of plans;

(4) Error detection;

(5) Effective system design;

(6) Efficient use of resource

3.1.3 Measures of Data Analysis and Utilization

- **These skills can be divided into two areas:**

(1) Techniques: These are related to the knowledge of technology use and its processes, including:

Testing and validation;

SQL query;

Data modeling;

Data analytics;

Report generation.

(2) Business: These are related to the knowledge of the business environment, including:

Understanding of the usage/application of technology to solve business problems;

Understanding of the business strategy of the company and industry trends;

Identification of key business identifiers;

Soft skills and the ability to transmit the results.

The cost of analysis is roughly proportional to the amount of audit data collected. The cost of analysis includes the time needed for grouping and reviewing the audit logs.

◉ **Steps suggested for data analysis are as below:**

(1) Ensuring the data to be correct, consistent, complete and free of duplication, with inaccurate or irrelevant parts removed, and the final data to be more easily combined with different sets of data to obtain accurate information and a thorough analysis;

(2) Analyzing special values instead of ignoring them because they are presented;

(3) Removing the corrupted data, which allows the remaining data to be correctly used by machines;

(4) Communicating the results with graphs and charts accurately and clearly for viewing.

3.1.4 Security of Data Analysis and Utilization

Evaluation results should be presented in different ways, depending on the case.

To begin, it is important have a report that fully presents the analysis process and results.

Once the report is finished, parts of it can be extracted to prepare summaries to be shared.

◉ **Structure of the report:**

- Title Page: names of authors and date
- Executive Summary
- Table of Contents
- Lists of Tables and Images
- Goal / Scope
- Design
- Purposes
- Methods and Tools
- Results
- Conclusions and Recommendations

3.1.5 Form of the Results

It includes thematic data analysis report, clues to doubts, audit method, audit information, etc.

3.2 Data Analysis According to Analysis Plan

3.2.1 Principle of Data Analysis

The data shall be analyzed and utilized by relevant parsing table, data analysis techniques, methods and tools, and based on different audit stages and objectives.

3.2.2 Formulating the Parsing Table

◉ **Before data analysis and utilization, auditors shall formulate the parsing table (to be analyzed) after comprehending the following contents and based on the audit objectives, tasks, priorities and audit items.**

(1) Business processing logic;

(2) Relationship between data, e.g., between accounting data and business data, internal data and external data, etc.

3.2.3 Establishing a Data Warehouse for Audit

During data analysis and utilization, auditors shall make full use of data warehouse technology, integrate data by theme, comprehensively and uniformly describe the data around analysis objects and relationship between data, and establish a data warehouse for audit.

3.2.4 Different Stages for Data Analysis

◉ **The audit planning stage, audit implementation stage and audit report stage are as follows:**

(1) While planning the audit, auditors may analyze the existing internal and external data owned by audit institutions, so as to properly determine audit projects.

(2) During audit, auditors may analyze the general conditions of the auditee by various data analysis techniques, methods and tools to determine the priorities, analyze specific matters, find out problems and doubts, verify the clues, and come to conclusions.

(3) While preparing the audit report, the audit results are created based on data analysis, so auditors shall analyze and evaluate the data utilization, summarize the practice in data preparation and analysis, and improve the audit efficiency.

3.2.5 Technique and Methods for Data Analysis

◉ **Data analysis techniques include query analysis, statistical analysis, multi-dimensional analysis, and data mining analysis.**

(1) Data analysis is classified into general analysis and thematic analysis.

General analysis: *the audit team comprehensively and systematically analyzes the general conditions of the auditee to get acquainted with its main characteristics, operation rules and development trends, and facilitate the determination of audit priorities.*

Thematic analysis: *the audit team further analyzes based on audit priorities and verify the clues to facilitate the collection of audit evidence.*

(2) Data analysis methods include structural analysis, trend analysis, ratio analysis, comparative analysis, logical analysis, mathematical statistics, feature discovery etc.

Structural analysis: *Revealing the overall structural relationship by calculating the proportion of each component of the analysis object in the total;*

Trend analysis: *Revealing the rules or abnormal changes through comparison and analysis of relevant data in several periods;*

Ratio analysis: *Selecting, calculating, and comparing the ratios;*

Comparative analysis: *Recalculating, comparing and verifying based on the articulation between various data;*

Logical analysis: *Calculating, checking and verifying whether the data is consistent with the business logic;*

Mathematical statistics: *Analyzing the state and law of data by various statistical methods;*

Feature discovery: *Summarizing data features based on audit practice; discovering unknown data features by data mining and other techniques; discovering data with the specified features by query analysis, multidimensional analysis and other techniques; discovering audit clues accordingly;*

Other methods

3.3 Application of New Technologies and Tools

3.3.1 Data Storage

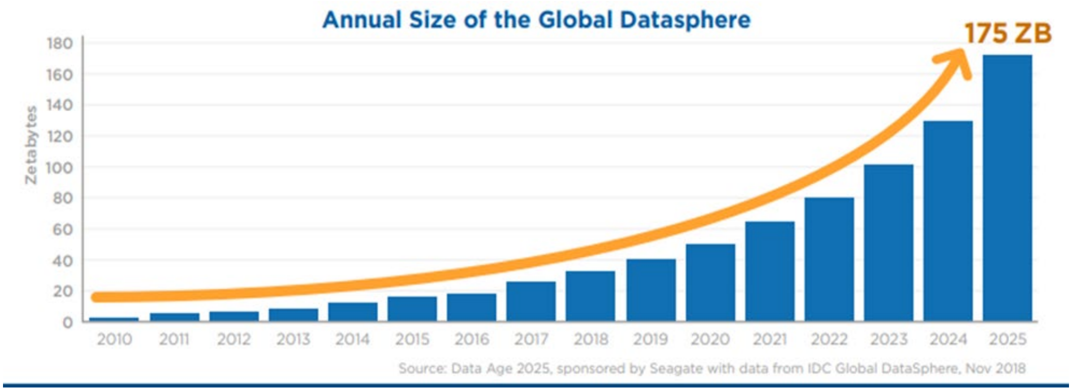
Data storage, despite its diverse definitions, briefly means collecting and maintaining data in a physical or digital area.

The first phase of data analysis is data collection. Without smoothly stored data, data analysis cannot be fulfilled. To derive meaningful conclusions through data analysis, the present and previous data should be stored and associated with each other in a certain area.

Data storage is also important in new auditing approaches. Data storage technologies and tools can be used for data analysis to facilitate the audit. Business intelligence programs are used in external and internal audits to detect audit risks, outliers, and frauds through the analysis of stored data. To maintain data, many tools and different technologies are used currently.

(1) Data storage technologies:

Currently, most of the transactions are moving to the digital area and this transition affects all aspects of business including audit. The data production rate which can be seen from the graph below is increasing exponentially.



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

To keep up with this gigantic amount of data, new storage sources are intrinsically needed. Besides traditional and emerging methods, there are a variety of surveys conducted in such areas such as a DNA storage system that is predicted to be more efficient and cheaper. As new methods of data storage emerge, audit management needs to be reviewed as well. Cloud storage and block chain are new trends to improve efficiency in data storage.

◉ Cloud storage

Cloud storage is a service of data storage accessible anywhere with any device. Additionally, the data stored in the cloud can be simultaneously accessed and downloaded by many users if they want.

The stored data is maintained in a certain data center owned a service provider, in which case users do not need to purchase hardware to store data. When needed available capacity can be decreased or increased.

In this system, the data stored using cloud technology is maintained in a place not owned by any data owner, making data safety first and foremost and requiring a reliable company or entrepreneur who provides data storage service. Otherwise, the data maintained in the cloud can be stolen and used sinisterly.

Cloud technology can be used to deliver an easier, faster and more efficient audit process: e.g., governments can constitute a platform based on cloud storage, to which the supreme audit institutions, auditee institutions and other stakeholders can have access

◉ Open data

◉ Block chain technology

Block chain technology is a method of data documentation on the internet. It relies on a technology called distributed ledger technology (DLT). Transactions between parties saved with the DLT in chronological order generate series of blocks, which form an interconnected chain. In this technology, each block refers to the other block before itself.

The technology works to maintain the data in a participant device which links to constitute a chain with other participant devices storing similar data. This is why the system is called block chain. The system maintains data in several devices trustworthily.

Also, it can be used to develop block chain applications, such as social networks, messengers, storage platforms, money transactions, etc.

Any type of data can be recorded on block chain in any form: identity information, transfer of money, an agreement between two parties, or even how much fuel a car has consumed in one hour. However, this requires confirmation from several devices such as computers and smartphones. Once a consensus is reached, the storing process will start. No record will be changed or removed without the permission of the recorder.

Also, where the information mentioned above is stored in more than one place or device, a variety of data servers will participate in the network. This is why the system is called a ‘decentralized’ system. There is no certain place where the data is kept.

Participant identities are secret and secured using cryptography but transactions that occur in the system can be seen by anyone at any time.

IoT

(2) Data storage tools:

Data storage tools help users manage their workloads, especially in big data areas. Some of the basic requirements sought by users are the company’ s reliability, storage scalability, and data security.

The preference for storage tools may change with users’ expectations and requirements. There are lots of platforms in the data storage area, aiming to provide data storage service with additional features. A few of the leading systems are given below.

- **Cloudera**

Cloudera uses cloud-based technology and offers the industry’s first enterprise data cloud. It includes multi-function analytics on a unified platform, shared data experience featuring consistent security and true hybrid capability with support for public cloud and multi-cloud, and on-premises deployments.

- **Google cloud storage**

Google cloud storage is an online file storage web service, aiming to store and access data on Google’s cloud platform infrastructure. It provides worldwide and highly durable object storage that scales to exabytes of data. The cloud platform offers different storage classes depending on the availability, which may change instantly after users optimize settings and performance.

- **Amazon web services**

Amazon web services, referred to as AWS, provides cloud-based service and offers a variety of storage products with some required features and different storage capacities that can be built to scale on demand according to users’ needs. AWS also offers a complete range of cloud storage services to support both application and archival compliance requirements. Clients may select from object storage, file storage, and block storage services as well as cloud data migration options and may design on the foundation of cloud IT environment.

- **IBM data storage**

IBM data storage provides cloud-based storage. There are three types of systems: object storage, block storage, and file storage. According to their needs, clients may choose one of them. The platform claims that its storage systems are scalable, configurable and available on-demand.

Additional:

Belladati, HortonWorks, Rackspace, OVH, Latisys, NetApp and Codera NoSQL etc.

3.3.2 Data Analytics

Data analytics is a method for examining data in order to gather information. Generally, it also applies to data analysis. Data analysis and data analytics can be used interchangeable. However, data analytics include the computation process of data analysis. To gather useful information from data analytics, different procedures are examined on raw data. Useful conclusions from raw data are generally drawn with the help of mechanical processes and algorithms.

◉ **Organizations, depending on needs, may use different data analytics procedures, which can be classified into several steps:**

1. Data collection process. Organizations should determine which data is collected to derive conclusions and then collect and store data in the database.
2. Cleaning process. Data cleaning guarantees no duplication, type error, or incompleteness. Then data are grouped by useful separators like age, demographic, income, and consumer.
3. Preparedness for data analysis. By using different data analysis tools, useful information is generated on data.
4. Application of generated information on data by related business entities in the decision-making process.

Data analytics can help organizations to yield optimal performance in the decision-making process. With data analytics, big data can be examined to produce useful information. SAIs can also use data analytics to improve information systems, business intelligence systems, and data analysis systems. Data analytics facilitates auditors to obtain information on data including patterns, relationships and models. In this way, auditors can easily visualize data and explore broad information around audit objectives, contributing to a more structured audit procedure and risk assessment.

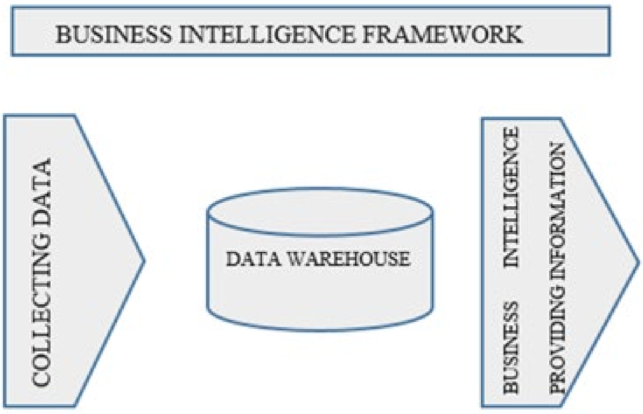
(1) Data analytics technology:

The volume of data used in the audit procedure has increased substantially recently. To obtain a clear understanding, SAIs must focus on the auditee and its environment and technology usage. Data analytics technology has become a part of auditors’ work in order to understand big data. There are different methods and software for data analysis. However, data analytics methods can be grouped into business intelligence and AI + machine learning.

◉ Business intelligence

Business intelligence includes advanced technologies and methods used to analyze data and provide information for the decision-making process. Firstly, organizations collect data in its warehouse and then draw conclusions with the help of BI systems. To draw conclusions on data and forecast future data sets, BI systems use different predictive analysis methods. Finally, BI systems gather the results of data analytics and make visual presentation. BI framework is composed of information collection, analysis, and presentation.

Figure 1-BI Framework



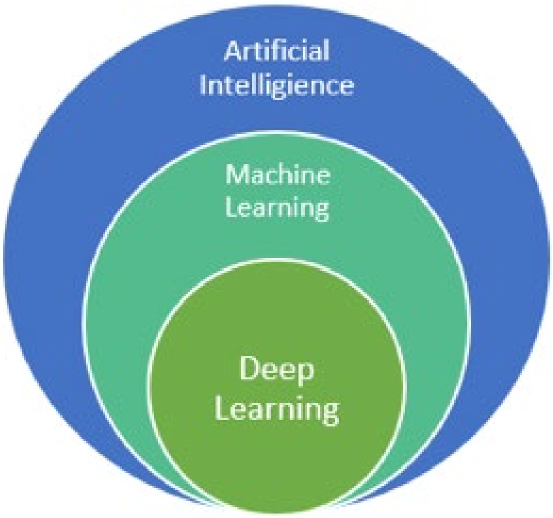
(Watson H. J., Wixom B. H., 2007, pg. 97)

Major benefits of BI systems include cost and time efficiency driven by the effective data analysis process. Useful information for the decision-making process is generated by the help of BI systems with less time and human resources. Moreover, SAIs can effectively leverage BI systems, which allows auditors to allocate their time more efficiently. Instead of analyzing data manually, with the results of BI systems, auditors can concentrate on their decision-making process and audit approach more efficiently. Results of BI systems help auditors give correct audit opinions.

◉ Artificial intelligence, machine learning and deep learning

Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) are tools used especially for big data analysis. Large amounts of data can be analyzed using AI, ML and DL tools to make predictions and conclude information on the data sets. AI is a general concept of computers and machines able to imitate human brain. AI is a broad term used in different areas where human logic is needed. Sometimes AI and ML are used interchangeable, but actually ML is a subclass of AI. ML can be defined as a program capable of automatically analyzing data and providing learning information. On one hand, ML can learn from big data by accessing and analyzing it; on the other hand, AI is capable of carrying out tasks like human brain to derive conclusions and make predictions about future values. ML begins with access to big data, followed by learning from data by experience, instructions and patterns. The ML system aims at computer learning from data without any human intervention. Basically, ML is about developing models, which are used by AI to infer the certain conditions. One step ahead from ML is DL. DL uses many aspects of ML software to produce extensive amount of information from raw data. DL includes artificial neural network, in each node of which new information is produced from data sets and summarized eventually by the help of AI.

Figure 2-AI, ML&DL Relations↵



ML provides analysis on huge amount of data. Analysis results include accurate and fast information for predictions, which allows for a more effective decision-making process. Big data with AI, ML and DL properties may be essential for organizations.

SAIs use AI, ML and DL properties extensively. Data about the auditee is analyzed deeply using these tools to produce proper results for auditors, who sometimes can analyze full data sets instead of samples, outliers and anomalies on data as well as each transaction. Moreover, with AI, ML and DL results and information from data, auditors can derive meaningful conclusions and infer audit objectives.

Supervised learning

Supervised learning uses past data, including past patterns and trends, to predict the future values. The machine starts with past trends in data sets, calculates the prediction function, and forecasts the future value. To create a more efficiently system, real values are compared with predicted ones, so errors are calculated and the model is adjusted accordingly.

Main algorithms used in supervised learning can be classified as:

a. Linear regression;

b. Logistic regression;

c. K-nearest neighbors;

d. Decision tree;

e. Random forest;

f. Gradient boosting machine;

g. Support vector machine (SVM);

h. Neural network.

Unsupervised learning

Unsupervised learning varies from supervised learning in the data used. If the data is not classified and labeled, the system can only infer a function, but not make any prediction. For future predictions, the system explores data and draws conclusions using datasets. Sometimes, supervised or unsupervised learning depends on the data type. When a data set includes both labeled and unlabeled data, semi-supervised learning is employed.

Main algorithms used in unsupervised learning can be classified as:

a. k means clustering;

b. Hierarchical clustering;

c. Neural network.

Reinforcement learning

Reinforcement learning, part of AI, is emerging relative to other machine learning methods. It produces actions and discovers errors in the environment mainly using the trial-and-error method.

(2) Data analytics methods:

Different methods are used in data analytics, generally classified into descriptive, diagnostic, predictive and prescriptive analytics, each of which has different insight on data. The process from descriptive to prescriptive analytics becomes increasingly complex to bring more useful information.

◉ Descriptive analytics

In descriptive analytics, data is summarized to provide a general overview and trend of data.

Raw data is summarized to be understood by the user. Descriptive analytics tries to answer the question "what has happened" and provides an understanding of past transactions in an organization. It involves the aggregation of individual transactions to offer a broader meaning and context and summarization of data via numerical or visual descriptions. Main statistical values like average, standard deviation, median and variance are tools for this method. In the audit, descriptive analytics is used to give a general overview of data and to identify potential risky areas.

Diagnostic analytics provides integrated information on data as an advanced form of descriptive analytics and tries to answer the question "why did it happen" or "how did it happen". It gives an understanding of the relationship between related data sets and identification of transactions/transaction sets with their behavior and underlying reasons. In this process, statistical techniques like correlation assist in the understanding of the causes for various events. Diagnostic analytics facilitates the identification of outlier data and causality inherent in data as well. It allows auditors to make a broader decision on the data set and come up with more details for reporting.

Predictive analytics is especially useful for predicting upcoming values. It assesses potential future scenarios using advanced statistical methods and tries to predict what will happen in the near future, making the sufficiency of historical data quite important. For the sake of generalization and prediction, data sets should include enough elements in terms of statistical sufficiency. It employs both traditional and AI and ML methods.

Prescriptive analytics assesses potential future scenarios using advanced statistical methods. As the name implies, it tries to predict "what will happen", "when will it happen" and "where will it happen" based on past data. Various forecasting and estimation techniques can be used to predict the future outcome of an activity.

◉ Traditional methods

In predictive analytics, mathematical methods that need basic computational skills are traditional. Forecast methods like trend line, regression and moving are basic traditional methods.

Nowadays, because of an increase in data size, more efficient ways of analysis on data is needed for predictions. Therefore, AI and machine learning can be used for predictions in advance.

Predictive analytics is especially useful for the audit work because in the presence of uncertainties, making predictions based on available data is important.

Prescriptive analytics takes over from predictive analytics and allows auditors to "prescribe" a range of possible actions as inputs such that outputs in future can be altered to the desired solution. In prescriptive analytics, multiple future scenarios can be identified based on different input interventions.

Prescriptive analytics not only predicts the future but also helps decide which action will be taken in the near future after such prediction.

Prescriptive Analytics helps auditors not only predict near future elements but also take the best action depending on prediction. (ISSA World Social Security Forum, 2019)

(3) Data analytics tools:

Basic data analysis tools:

- **Microsoft Excel**

As a basic, popular and widely used analytical tool almost in all industries, Excel becomes important when there is a requirement of analytics on the client’ s internal data. It analyzes the data summarized in a preview of pivot tables to help filter the data as per the client’ s requirement.

- **Microsoft Access**

Microsoft Access is a database management system (DBMS) from Microsoft that combines the relational Microsoft Jet Database Engine with a graphical user interface and software-development tools. Access databases provide many tools to maintain data quality. Lookup lists and validation rules for individual fields and records can be easily implemented in Access at the table level. Forms can add additional rules during data entry to respond to user selection and events. Access also offers referential integrity between tables to ensure consistent data across tables.

General audit software:

- **ACL**

Audit Command Language (ACL) is data extraction and analysis software for fraud detection and prevention, and risk management. It samples large data sets to spot irregularities or patterns in transactions that could indicate control weaknesses or fraud. ACL is a combination of scripts and workspaces. Most data manipulation tasks like creation of new formatted fields from raw data are performed in workspaces, where the syntax for scripting is fairly simple.

- **IDEA**

IDEA is a comprehensive, powerful and easy-to-use data analysis tool. It seamlessly connects IDEA data to third-party applications that support ODBC including Tableau, Power BI, MS Excel, among others. The Discover task is powered by IDEA’ s Analytic Intelligence to identify trends and outliers and automatically populate dashboards to be further refined. IDEA creates charts and field statistics on reusable dashboards, and profile data to easily identify patterns, trends, outliers, and correlations.

BI software tools:

◉ Tableau

Tableau is a powerful and the fastest growing data visualization tool used in the business intelligence industry. It helps simplify raw data in an easily understandable format. Data analysis is very fast with Tableau and the visualizations created are in a dashboard or worksheet. Data created using Tableau can be understood by any professional at any level in an organization. It even allows a non-technical user to create a customized dashboard.

◉ Qlik

QlikView is Qlik’ s classic analytics solution for rapidly developing highly-interactive guided analytics applications and dashboards, delivering insight to solve business challenges. The modern analytics era truly began with the launch of QlikView and the game-changing Associative Engine it is built on. QlikView is a BI data discovery product for creating guided analytics applications and dashboards tailor-made for business challenges.

Qlik Sense is a data analytics platform that sets the benchmark for a new generation of analytics. With its one-of-a-kind associative analytics engine, sophisticated AI, and high-performance cloud platform, people will be able to make better decisions daily, creating a truly data-driven enterprise.

Qlik makes it easy to connect and combine data from hundreds of data sources – from apps and databases to cloud services, files and more. Explore common data sources and connection options for Qlik Sense. Qlik has AI technology built in at a foundational level and has an Advisor suggests analyses and insights, automates analytics creation and data preparation processes, and supports natural language interaction for search-based and conversational analytics. It also includes machine learning for more relevant insights over time, business logic for customization, and leverages the Qlik Associative Engine for unmatched context awareness and peripheral vision.

◉ Sisense

Sisense is a business intelligence platform that lets users join, analyze, and picture out information they require to make better and more intelligent business decisions and craft out workable plans and strategies. With Sisense, all the data can be unified in visually appealing dashboards via a drag and drop interface. Sisense basically allows the user to turn data into highly valuable insights and then share them with colleagues, business partners, and clients via interactive dashboards.

- **Oracle BI**

Oracle Business Intelligence (BI) Enterprise Edition is a business intelligence server. It includes advanced business intelligence tools built upon a unified architecture. The server provides centralized data access to all business-related information in a corporate entity. It integrates data via sophisticated capabilities from multiple sources. Oracle BI Server is a query, reporting and analysis server and provides services to the other components of the Business Intelligence suite such as Data mining, Reporting, and Analytic Applications.

- **SAP Business Objects**

SAP Business Objects BI is a reporting and analytics business intelligence (BI) platform aimed at business users. It consists of a number of reporting applications that allow users to discover data, perform analysis to derive insights, and create reports that visualize the insights. SAP BO is intended to make reporting and analysis simple for business users so they can create reports and perform processes like predictive analytics without needing the input of data analysts.

- **Zoho, IBM Cognos Analytics, Looker**

Additional: MS Power BI, Datawrapper, Dundas BI, BOARD, Pentaho BI Suite, Suite, SAP Business Intelligence, SAS Business Intelligence, MicroStrategy, Domo, Yellowfin BI, TIBCO Spotfire, Hevo Data, Clear Analytics

Other Analysis Software Tools:

- **KNIME**

KNIME Analytics platform is one of the most popular open-source platforms used in data science to automate the data science process. KNIME has thousands of nodes in the node repository which allows the user to drag and drop the nodes into the KNIME workbench. A collection of interrelated nodes creates a workflow which can be executed locally as well as can be executed in the KNIME web portal after deploying the workflow into the KNIME server. KNIME helps to create the Guided Analytics process as a workflow in KNIME platform by helping to automate the process.

◉ R

R is a language and environment for statistical computing and graphics. It is a GNU (GNU is Not Unix) project which is similar to the S language and environment. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

◉ RapidMiner

RapidMiner is a data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, model validation and optimization.

◉ OpenRefine

OpenRefine is a standalone open-source desktop application for data cleanup and transformation to other formats, the activity known as data wrangling. It is similar to spreadsheet applications (and can work with spreadsheet file formats); however, it behaves more like a database.

◉ Orange

Orange is an open-source data visualization, machine learning and data mining toolkit. It features a visual programming front-end for explorative data analysis and interactive data visualization. Orange is a component-based visual programming software package for data visualization, machine learning, data mining, and data analysis. Orange components are called widgets and they range from simple data visualization, subset selection, and preprocessing, to empirical evaluation of learning algorithms and predictive modeling.

◉ Python

Python is an interpreted, high-level, general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python offers tools and techniques for data analysis that can be used at any step of the analysis process. It can be used for data extraction, data reading and preparing data for to be used by machine learning. Moreover, it is a good tool for data mining and data visualization.

Additional: Weka, Google Fusion Tables, NodeXL, Cloudera, Talend, and tailor made tools that can be developed.

3.3.3 Data Mining

Definition

Data mining is the process of finding anomalies, patterns and relationships within large volumes of integrated data sets to predict outcomes and future trends. It comprises of algorithms which enable gaining insights and knowledge. It is the knowledge discovery step during analysis which comes after data extraction, data cleaning and other data processing functions.

Data mining is a multidisciplinary process associated with areas such as data science, statistics, mathematics, geometry, machine learning, and artificial intelligence. Both structured data and unstructured data such as graphics and audio can be mined using right techniques and tools.

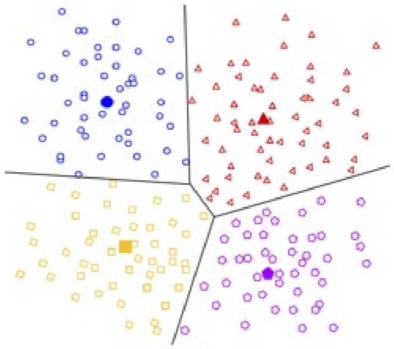
Data mining starts with data understanding, in order to understand and gain insights from the data set. Data will be prepared through cleaning, scrubbing, integration, transformation and reduction in order to secure a suitable mining technique used for clean and proper data. A model of the data set will be constructed to determine the mining technique. Lastly, evaluating and interpreting the results reached after applying the chosen data mining technique on the modelled data set.

Grouped as below, techniques used in data mining are grounded on a predictive, descriptive, pattern mining or anomaly detection model. Only some common types of them will be explained in the next section.

(1) Data Mining Techniques

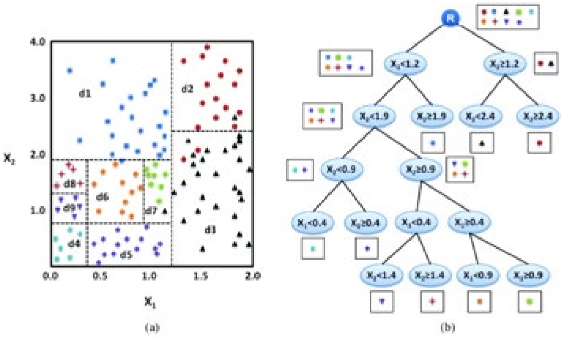
Classification is a method of assigning a label to unclassified data. In classification, firstly auditors built a model called the training set which includes elements correctly labeled to make predictions. There are many types of classification, some of which are named as decision trees, linear classifiers, kernel estimation, logistic regression, and random forest.

Figure 4-Clustering Model↵



www.developers.google.com↵

Figure 3-Decision tree model↵



www.databricks.com↵

Clustering is the process of making the groups of similar objects together. Data points with similar characteristics are distributed into groups. Clustering is divided into hierarchal clustering, density-based clustering, model-based clustering, among others.

Visualization is the process of converting textual or numerical data into meaningful images.

Decision Tree is used for nominal and numeric data values, while decision analysis is performed with the help of the tree-shaped structure.

Association Rules aim to find rules associated with frequently co-occurring items.

Association rules help find rules, patterns or casual structures associated with the frequently co-occurring items. The probability of the relationships between items within a large data set is calculated with this technique.

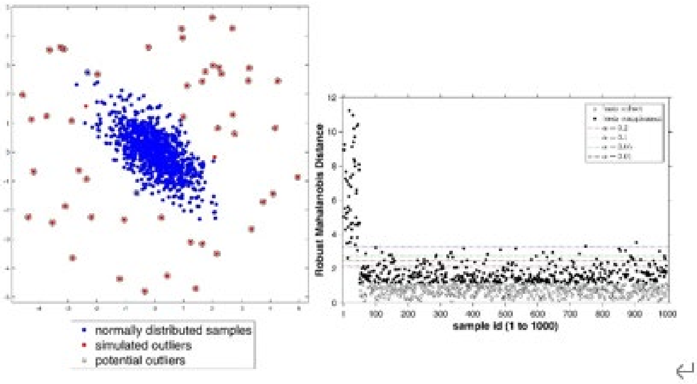
Artificial Neural Network (ANN) is a mathematical model inspired by biological neural networks.

A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation.

It is an information processing technique including nonlinear statistical data modeling tools that help discover the relationships or patterns among inputs and outputs.

Sequential Patterns is a data mining task specialized for analyzing sequential data to discover sequential patterns. It is the mining of frequently occurring items or subsequences as patterns when values are being delivered sequentially. Thus, sequential patterns are being discovered in a sequence database. Time series analysis is an example for sequential patterns.

Figure 6-Outlier Analysis↵



www.mathworks.com↵

Outlier detection is the process of finding data objects with behaviors that are very different from expectation. In outlier analysis, data objects different from the rest of the data are identified. The reason for outliers may be either error or fraud. Outlier Analysis identifies the deviations and discovers objects that have a high risk of being fraudulent.

(2) Data Mining Tools:

The most commonly used software in audit profession is generalized audit software (GAS), a standardized program with basic features. GAS is a shelf type of software package developed by professional auditing entities. Used mainly in the execution phase, it is applicable to any phase of audit. With features on a huge scale from data extraction to generate a report, it is used for data manipulation, mathematical computation, sample selection, risk assessment, result evaluation, fraud detection and summarization, among others.

However, in order to process vast amounts of data from various sources and discover valuable patterns and relationships amongst such data, auditors need to use data mining techniques in more advanced programs. There are some examples.

Most of these programs as mentioned above are used for data analytics, so they will be explained from their usage for data mining.

◉ RapidMiner

RapidMiner is a software platform developed for data mining, text mining, machine learning, and predictive analysis and business analysis techniques in an integrated environment. It is produced by the company of the same name in an open core model.

◉ Knime

KNIME is an open-source software. It is a modular platform integrating many components to build workflows and data analysis pipelines upon predefined components that are called nodes.

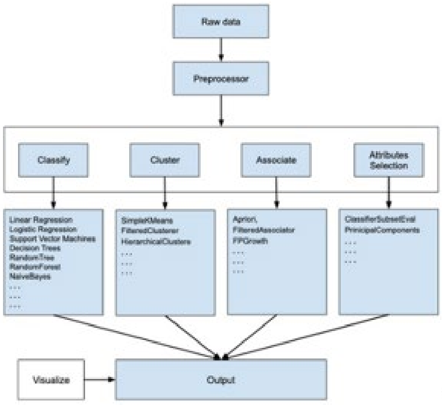
◉ Orange

Orange is an open-source tool that can be used for data mining. In Orange, data mining is being done through Python or visual programming. The tool can also be used for text mining, machine learning, exploratory data analysis and visualization.

◉ Weka

Weka is an open-source machine learning software tool. It is a collection of machine learning algorithms used for running data mining techniques and contains necessary tools for techniques such as classification, regression, clustering, association rules, and visualization.

Figure 7-Process of WEKA↵



↵

https://www.tutorialspoint.com/weka/what_is_weka.htm↵

◉ SAS Data Mining

SAS is a data mining package standing for Statistical Analysis System. It comprises of different programs which were originally developed for conducting analyses on statistical data. It supports SQL queries and SQL server, and Access can be used as a data source in SAS. Programs within SAS work together to store and modify data and run statistical analysis. It suits for advanced analytics, business intelligence, data management, and predictive analytics.

◉ R

R is a free software language and environment that supports statistical computing and graphics. It is highly extensible and provides a wide range of statistical and graphical techniques.

Additional: Oracle Data Mining, Apache Mahout, Sisense, Teradata, DataMelt, Board, Dundas BI

Potential Use of Data Mining in Audit

Since auditing involves the evaluation of large volumes of data, using a broad range of techniques will lessen the resources used, help to conduct a more focused audit, reduce the risks involved, and improve efficiency, effectiveness and transparency. As a result of organizing and analyzing data in a more efficient and effective way, using data mining techniques can ease the audit process.

As a risk assessment tool, data mining can be used to detect misstatements caused by either error or fraud. Outlier detection and Benford’ s Law are two analyses that can be used in fraud detection. Anomalies in the data which are carefully hidden in fraud can easily be discovered using either analysis.

The logical and associative relationships between data are discovered and divided into segments according to predetermined parameters using data mining techniques. By this way, data mining allows for data query to identify the susceptible activities.

Here are some examples:

- multiple payments to the same vendor for the same invoice;
- multiple payments to different vendors for a particular invoice;
- miscalculations;
- payments for non-rendered services;

3.3.4 Data Visualization

Definition

Data visualization is to present information and data using statistics, probabilities, charts, graphs, pivot tables, heat maps, geographical mapping, and other artifacts in order to discover trends, outliers and patterns in data.

In the digital era, data visualization techniques and tools are essential to analyze massive amount of data and make data-driven decisions. A good data visualization should combine communication, data science and design in order to make data more understandable and highlight the useful information for users. Thus, in order to attain a good visualization, not only well sourced, complete and clean data is needed but right types of charts, graphs, maps etc. should be chosen and designed.

Data visualization is crucial since it makes data easier to understand and remember as well as helps users ask better questions, make better decisions, discover unknown facts, outliers and trends, and visualize relationships and patterns quickly.

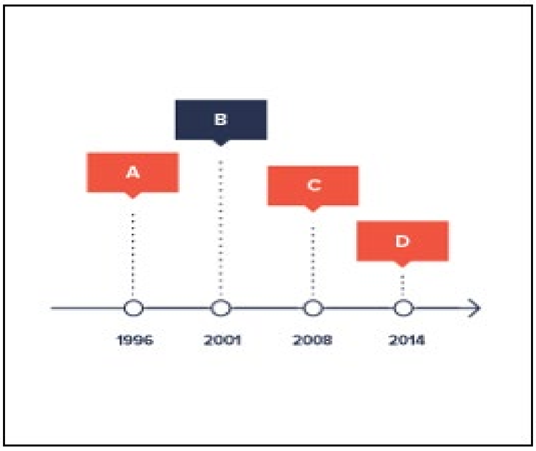
(1) Data Visualization Techniques

◉ Temporal

Temporal visualizations should be both linear and one-dimensional. They share a common start and finish data point and include items that either stand alone or overlap with each other. Temporal visualizations generally show all events before, after, or during some time period or moment.

Here are some examples for temporal visualizations: timelines, line graphs, Gantt charts, scatter plots, pie charts, stream graphs, arc diagrams/thread arcs, tree rings/concentric circle graphs, time series charts/graphs, stacked area charts, bar charts, heat maps, polar area diagrams, and alluvial diagrams.

Figure 8-Timeline↵



↵

www.datavizproject.com/data-type/timeline↵

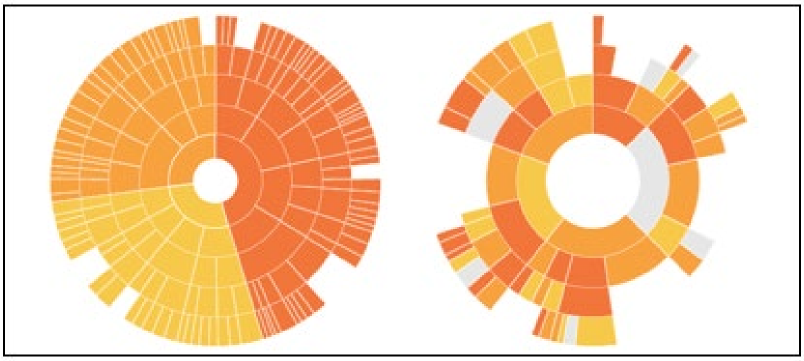
Linear and one-dimensional visualizations have a start and finish time and include items that may overlap each other.

Hierarchal

Hierarchical visualizations begin with a root entity, which has at least one “child node”, and every further child node has zero or more children, with each item having a link to one parent item (except the root). Items and the links between parent and child can have multiple attributes. These can be applied to items and links.

Here are some examples for hierarchical visualizations: dendrogram, phylogenetic tree, radial tree, hyperbolic tree, tree diagram, tree map, cone tree, radial hierarchy, ring charts/sunburst diagram, circle packing, and decision tree/flow chart.

Figure 9-Sunburst Diagram↵



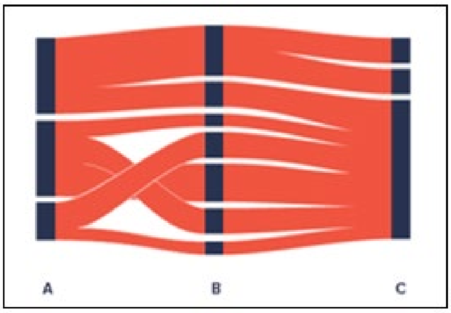
Larger groups encompass sets of smaller groups

Network

Network visualizations present the relationships between datasets within a network, in which datasets connect deeply with other datasets, and users can understand where items connect or identify the shortest or least costly paths connecting two items or traversing the entire network.

Here are some examples for network visualizations: matrix charts, adjacency list, node-link diagrams, word clouds, alluvial diagrams, hive plot, dependency graph, arc diagram, force-directed graph, hierarchical edge bundling, sankey diagram, and subway map.

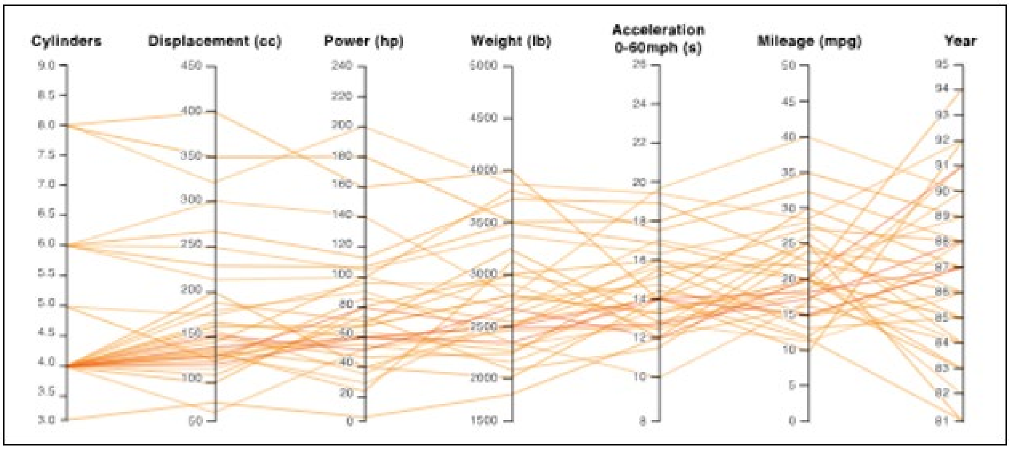
Figure 10-Alluvial Diagram↵



Multidimensional

Multidimensional visualizations include data elements with two or more dimensions in contrast to temporal visualizations, so there are always more than two variables in the mix to create a 3D visualization. They are generally the most eye-catching visuals depending on the many concurrent layers and datasets. These visuals can present key takeaways by breaking a lot of non-useful data.

Figure 11-Parallel Coordinate Plot↵



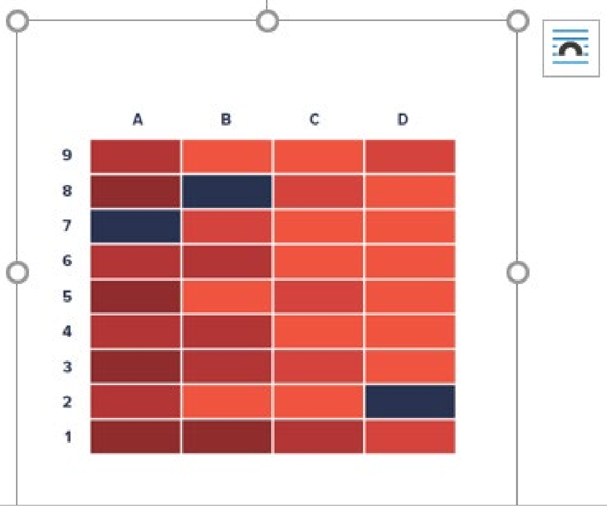
Here are some examples for multidimensional visualizations: scatter plots, pie charts, venn diagrams, stacked bar graphs, histograms, parallel coordinate plot, rank plot, tree map, line chart, step chart, area chart, heat map, matrices, parallel coordinates/sets, radar/spider chart, box and whisper plots, mosaic display, waterfall chart, pixel bar chart, and tabular comparison of charts.

- **Geospatial**

Geospatial visualizations describe present real-life physical locations and overlaying familiar maps with different data points. They create a holistic view of performance.

Here are some examples for geospatial visualizations: flow map, choropleth map, cartogram, heat map, and density map.

Figure 12-Heat Map↵



They can be configured as charts, plots, maps, diagrams and matrices.

(2) Data Visualization Tools

- **Tableau**

Tableau is a business intelligence tool for visually analyzing data packed with graphs, charts, maps and more. Users can create and distribute an interactive and shareable dashboard, which depicts trends, variations, and density of the data in the form of graphs and charts. Tableau can connect to files, relational and big data sources to acquire and process data.

- **Sisense**

Sisense is a business intelligence platform that has a very friendly user interface via a drag and drop user interface which allows charts and more complex graphics, as well as interactive visualizations, to be created with a minimum inconvenience. Sisense provides users with instant insights and allows for sharing with colleagues etc. via interactive dashboards.

- **HighCharts**

HighCharts is a chart library written in pure JavaScript that allows users to view and run interactive visualizations easily and conveniently. It enables fast and flexible solutions with the smallest need for specialist data visualization training.

- **Plotly**

Plotly is a library that enables more complex and sophisticated visualizations via its integration with analytics-oriented programming languages such as Python, R and Matlab. Plotly allows users to modify their dashboards, interactive graphs and the script to create their own programs and features to get the highest comprehension.

- **Qlik**

Qlik consists of QlikView and Qlik Sense which are used to access, present and explore data. Qlik Sense provides self-service data visualization and analytics capabilities. It does not have a build-and-publish approach, so users can drag and drop to build or extend visual analytics . QlikView provides an analytical application development platform that aims to enable analysts with minimal development expertise or experience to build and publish applications that include information governance and management capabilities.

- **Fusioncharts**

Fusioncharts is a widely-used, JavaScript-based platform. It can produce a considerable amount of different chart types with only one set of data entry and at the same time integrate with a wide range of devices, frameworks and platforms thus giving users a great deal of flexibility. Users also can choose one visualization from a wide range of “live” example templates available by plugging in their data sources as needed.

Additional: Zoho, Domo, Microsoft Power BI, Google Charts, Datawrapper, Infogram

◉ Potential Use of Data Visualization in Audit

Auditors need more advance techniques than traditional techniques since they are faced with major challenges in processing and analyzing data with the advent of such “Big Data”. Data visualization helps auditors gain better insights, draw better conclusions and ultimately improve the audit process.

Data visualization is an important tool for auditors to perform audit effectively and efficiently and to generate a meaningful audit report. Data visualization is effective from the beginning to the end of the audit. More specifically, when an auditor is in the phase of audit planning and scoping, data visualization can guide the auditor in adjusting the audit scope. Afterwards, the auditor may quickly identify the areas of interest, anomalies and trends as they become clear via the use of various types of graphics, tables or dashboards. On the other hand, an effective data visualization may conclude that there is no significant matter, which needs particular attention. As a result, auditors can ask better questions, make better decisions, understand data easier through data visualization, and generate a notable audit report.

3.4 Analysis Result Verification

3.4.1 Technique of Analysis Result Verification

The audit team shall verify the analysis results using the multi-dimensional analysis techniques, such as multi-dimensional, quick and flexible query and analysis by data slicing, dicing, drilling and rotation; visual display of query easily understood.

3.4.2 Storage of Data Analysis Results

Data analysis results shall be stored in the terminal which is a device specially used for data analysis by auditors. The terminal shall be kept in a special room for data analysis. While engaging in data analysis in the room, auditors shall not disclose the data and analysis results in any way.

3.4.3 Principles of Analysis Result Verification

Auditors shall ensure the data analysis results to be scientific, rational and accurate, and record the personnel and time of data analysts, data source, data analysis logic, legal basis, operation method, program script, analysis results, etc. Auditors shall pay attention to the analysis process and results.

3.5 Analysis Report

3.5.1 Preparations the Analysis Report

In this phase, the results will be shown along with the previous useful analyses in an orderly, meaningful, simple way. Findings should be summarized finally. The report usually starts with an introduction that states the purpose of writing the report from the client’ s perspective and give a taste of what is possible and include a necessary background.

3.5.2 Principles for the Analysis Report

(1) Less technical details and simple writing

Words should always exist, not just pictures and graphs.

Avoid flowery prose and extra details, to the point (bullets as needed).

Do not get bogged down in the weeds or the process, to miss reporting the outcome.

(2) Clear description of what you did and why

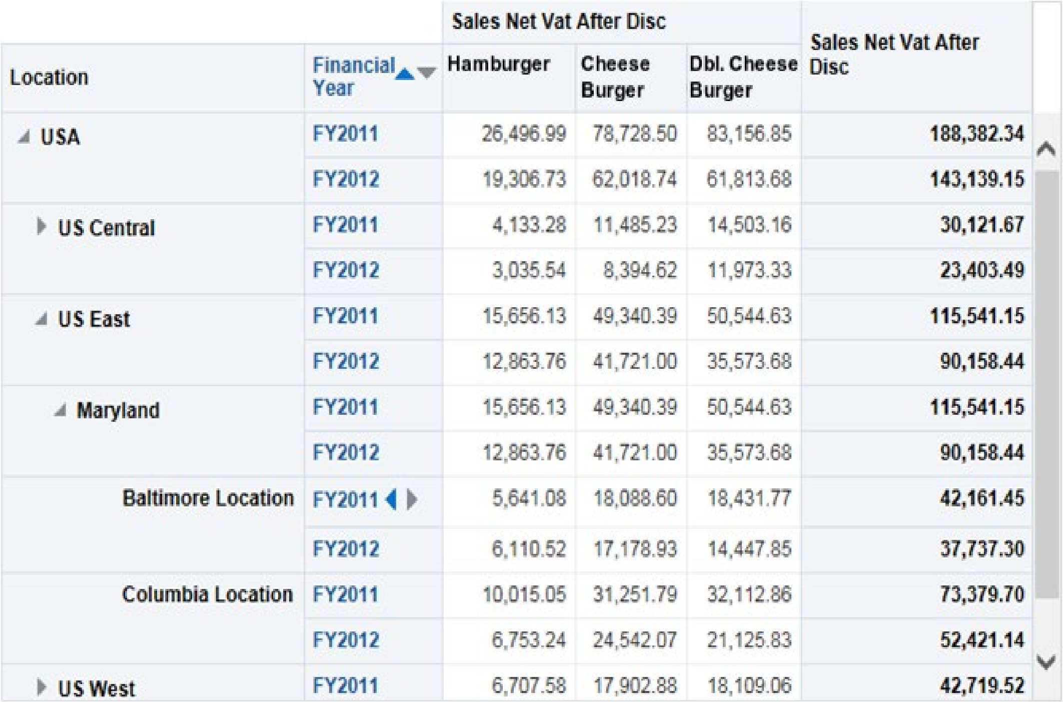
The report can be shown by tables, graphics, and words. It is recommended not to have more than 3 metrics in a table and data should always be segmented, not aggregated. This makes it easier to read.

Findings must be listed one by one; they can be grouped by a topic, for example: by person, number, public contracts, or another topic.

3.5.3 Display Form of the Analysis Report

The analysis report may be started with a brief introduction and then show all the information obtained, through tables and graphics.

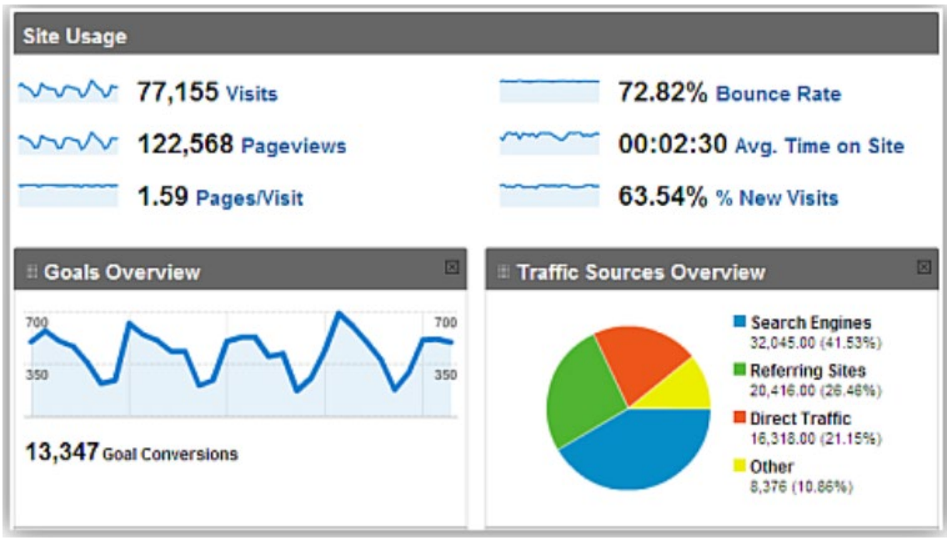
(1) Tables can be presented with some metrics,



		Sales Net Vat After Disc			Sales Net Vat After Disc
Location	Financial Year	Hamburger	Cheese Burger	Dbl. Cheese Burger	
USA	FY2011	26,496.99	78,728.50	83,156.85	**188,382.34**
	FY2012	19,306.73	62,018.74	61,813.68	**143,139.15**
US Central	FY2011	4,133.28	11,485.23	14,503.16	**30,121.67**
	FY2012	3,035.54	8,394.62	11,973.33	**23,403.49**
US East	FY2011	15,656.13	49,340.39	50,544.63	**115,541.15**
	FY2012	12,863.76	41,721.00	35,573.68	**90,158.44**
Maryland	FY2011	15,656.13	49,340.39	50,544.63	**115,541.15**
	FY2012	12,863.76	41,721.00	35,573.68	**90,158.44**
Baltimore Location	FY2011	5,641.08	18,088.60	18,431.77	**42,161.45**
	FY2012	6,110.52	17,178.93	14,447.85	**37,737.30**
Columbia Location	FY2011	10,015.05	31,251.79	32,112.86	**73,379.70**
	FY2012	6,753.24	24,542.07	21,125.83	**52,421.14**
US West	FY2011	6,707.58	17,902.88	18,109.06	**42,719.52**

In this example, it is clear to see the main figures by state and in the state under analysis, see it in more details.

(2) Graphics can be showed in the report



In this example, 4 graphics are displayed in one. They are clear and concise and there are many types of graphics available to use.

(3) Individual reports and consolidated ones are both available.

3.5.4 Basic Elements of the Analysis Report

The audit team shall properly record and prepare a data analysis report composed of the following basic elements:

(1) Title;

(2) Name of the auditee;

(3) Name of the audit project;

(4) Name of the audit institution;

(5) Contents;

(6) Time for completing the data analysis report;

(7) Name of data source (system and date)

3.5.5 Main Content of the Analysis Report

The main content of the analysis report includes basic information of data analysis, general analysis and results, and thematic analysis and results.

(1) Basic information of data analysis. Generally, it includes the scope, content and source of data, the main technical methods, the main basis for conclusion, etc.;

(2) General analysis and results. Generally, it includes the data required, methods and steps of analysis, and basis for judgment and conclusion;

(3) Thematic analysis and results. Generally, it includes the data required, methods and steps of analysis, basis for judgment and conclusion, and recommendations for further audit.

04 Quality Control and Data Security

4.1 New Risk of Conducting Audit Activities with Data Analytics

4.1.1 Risk of Data Authenticity and Integrity

There are always some risks in the authenticity and integrity of big data. When auditors conduct audit activities, they should have a clear understanding and observe due professional care. For example, authenticity is the basis of data value, but it is also a congenital defect of big data. If preventive measures are not sufficiently comprehensive, the wrong audit conclusions and audit risks may occur.

Auditees may deliberately modify the data or conceal the information, and these are just a few of the new risk of associated with conducting audit activities with data analytics.

4.1.2 Risk of the Causal Relationship and Correlation of Data

Big data analysis conducted by machines, which can only give the correlation between data, while detailed causal relationship needs further analysis taken by auditors. Focusing only on the big data analysis of correlation rather than causal explanation is likely to lead to wrong and even dangerous conclusions.

4.1.3 Risk of Data Collection, Storage, Analysis and Security Management

In the process of advancing big data auditing, auditors should attach great importance to new problems and new risks associated with in data collection, storage, analysis and security management.

The way to control the risk of big data audits and improve the audit quality is a critical challenge that will even affect the future development of big data audits.

4.2 Quality Control of Conducting Audit Activities with Data Analytics

4.2.1 Quality Control of Staff Requirement

During data analysis and utilization, auditors shall get acquainted with the basic information of the data required, carefully study the relevant business processes, and master the data analysis techniques, in order to avoid deviation from the actual conditions.

4.2.2 Quality Control of Data Processing

Auditors shall ensure data collection to be compliant and processing and analysis to be scientific, rational and accurate, and record the personnel and time of data analysis, data source, data analysis logic, legal basis, operation method, program script, analysis results, etc. Auditors shall pay attention to the analysis process and results.

4.2.3 Quality Control of Using Data Results

If auditors adopt data analysis results as audit evidence for violations of disciplines and laws, spot check and verification are required. If the relevant review or quality control departments have any question on the analysis process or audit conclusion, auditors shall make explanations. Auditors shall timely summarize the practice of data analysis and utilization, build a data analysis model based on the analysis process to truly and reliably show the business process and results of audit, and submit to the relevant departments upon check and approval.

Auditors shall make full use of available data using data analysis models and results of audit institutions to improve the efficiency and quality of data analysis.

4.2.4 Quality Control of Data Profiling

Data analysis and utilization shall be filed in accordance with the relevant archives management requirements, including relevant archives of data collection, authorized use, sorting, processing, data analysis, as well as the review opinions of analysis results, collection of audit evidence, and data analysis model.

Data management personnel shall check the state and operation log of data storage on a regular basis to ensure the prevention of data from damage due to sorting, processing, analysis and utilization.

4.3 Security Management in Data Collection, Storage and Analysis

4.3.1 Policies for Security Management

Each audit institution should define a set of policies to clarify their direction of and support for information security. At the top level, an overall "information security policy" needs to be implemented in the conduct of audit activities, particularly in the data collection, storage and use. Policies for security management should be reviewed at planned intervals or when significant changes occur to ensure their continuing suitability, adequacy, and effectiveness within each institution jurisdiction.

4.3.2 Principles and Methods of Security, Integrity and Confidentiality

All entities shall formulate and improve the system of data analysis, utilization and confidentiality, including data collection, receiving, recovery, authorized use, security, integrity, confidentiality, and accountability.

4.3.3 Security Management of Data Analysis Environment

Data analysis environment refers to the device and environment for data analysis, such as network, special place and analysis terminal. All entities shall create a data analysis environment as per relevant requirements on security and confidentiality. The audit team shall also create an on-site data analysis environment in accordance with relevant requirements on security and confidentiality.

(1) Both the network established for data analysis and the data analyzers should be non-secret-related. Therefore, technology reinforcement, protection and management shall be provided for them based on security and protection requirements on secret-related network and terminal.

(2) Data analysis terminal is a device especially set aside for data analysis by auditors. It shall be kept in a special room for data analysis. While engaging in data analysis in the room, auditors shall not disclose the data and analysis results in any way, Therefore, they could sign confidentiality agreements. Both the desktop and laptop that used for data analysis terminal should be marked "only for data analysis". They shall not be equipped with wireless devices such as wireless network card, Bluetooth module, wireless mouse, wireless keyboard. The operating system, database system, office system, analysis tools and other software installed in the data analysis terminal shall be copyrighted. Work-irrelevant software and tools shall be prohibited from being installed in the data analysis terminal.

4.3.4 Security Management of Data Analysis Process

(1) A strict system of approval and registration shall be built for accessing the places where data are received, recovered and stored. It is strictly prohibited to bring communication devices such as mobile phones that can take photos, videos, and recordings while accessing.

(2) In case of data collection outside a location, safe and controllable means shall be adopted to transfer data based on data confidentiality and relevant laws and regulations on confidentiality.

(3) The data analysis and utilization system formulated by each entity shall include the authorization management mechanism, application, approval and authorization procedures required for data use.

4.3.5 Security Management of Data Receipt, Storage and Output

(1) The security level of data shall be determined by data providers. Once received, the data shall be marked with the security level and appropriate protection shall be provided.

(2) If data at different security levels are stored in the same environment, the highest level shall prevail. The data identified as state secrets must be stored in a secret-related device and operate on a secret-related network. The secret-related data shall be managed in strict accordance with the relevant legal requirements. The non-secret-related data also shall be classified based on sensitivity and importance, and used upon approval. The secret-related data shall be managed separately in the secret-related environment, and the approval procedure shall be strictly managed in accordance with the relevant provisions.

(3) Data shall be output upon approval. The output port of data shall be controlled strictly, and data output shall be managed by a specially assigned person. Moreover, data output shall be registered with relevant documents kept well. Generally, the personnel responsible for data output shall not be engaged in data analysis or audit projects related to the output data. These functions shall be declared incompatible duties for internal control purposes.

4.3.6 Security Management According to Different Block Business

(1) All responsibilities pertaining to the planned security management should be defined and allocated to a specific actor within the audit institution. Each audit institution should identify areas of responsibility for the purpose of segregation to reduce unauthorized or unintentional modification or misuse of information collected, stored, or used by the audit institution.

(2) The implementing audit institution should have a defined policy on screening employees for auditing activities involving the collected and stored data. During employment, the audit institution should ensure that employees and contractors are made aware of and motivated to comply with their security obligations. A formal disciplinary process is necessary to handle information security incidents allegedly caused by auditors. Security aspects of any employee’ s departure from the organization or significant changes of roles within it should be managed as well.

(3) Mobile devices and teleworking

A policy and supporting security measures should be adopted to manage the risks introduced using mobile devices or when the audit institution allows the bring-your-own-device scheme in the workplace.

(4) IT asset management

All assets used for the collection, storage, and access of data should be inventoried and the owners should be identified to be held accountable for their security compliance. The policy of ‘acceptable use’ should be defined, and assets should be returned when personnel leave the organization. All information should be classified and labelled by its owners according to the security protection needed, and handled appropriately. Information storage media should be managed, controlled, moved and disposed of in such a way that the information content is not compromised.

(5) Access control

The audit institution’s requirements for access control over information assets should be clearly documented in an access control policy and procedures. The allocation of access rights to users should be controlled from initial user registration through to removal of access rights plus regular reviews and updates of access rights.

(6) Use of cryptography

There should be a policy on the use of encryption, plus cryptographic authentication and integrity controls such as digital signatures and message authentication codes, and cryptographic key management.

(7) Physical and environmental security

Each audit institution should define physical perimeters and barriers, with physical entry controls and working procedures to protect the premises where data are stored. Equipment and information should not be taken off-site without authorization, and must be adequately protected both on and off-site.

(8) Operation security

Audit institutions shall document the IT operating responsibilities and procedures, control the *changes to IT facilities and systems, manage capacity and performance, and adopt malware controls, including user awareness. Appropriate backups should be taken and retained in accordance with the backup policy. Technical vulnerabilities should be patched, and rules should be issued in place governing software installation by users. All of these should be considered in conducting audit activities with data analytics.*

(9) Information security incident management

Audit institutions should be equipped with defined responsibilities and procedures to consistently and effectively manage (report, assess, respond to and learn from) information security events, incidents and weaknesses, and to collect forensic evidence.

(10) Security aspect of business continuity management

The continuity of information security should be planned, implemented and reviewed as an integral part of the audit institution’s business continuity management systems. IT facilities should have sufficient redundancy to satisfy availability requirements for data.

(11) Compliance

The audit institution must identify and document its obligations against external authorities (e.g., privacy commission) and other stakeholders (e.g., the audited entity) in relation to information security, including privacy/personally identifiable information and cryptography. A separated and independent designated actor should routinely review employees’ and systems’ compliance with security policies, procedures etc., and initiate corrective actions where necessary.

05 Accessory

An example of extracts from SAI India’ s data analytics guidelines for data collection.

◉ Data classification

Classification refers to the process of arranging data into homogenous groups or classes according to common characteristics. Classification of information resources according to criticality and sensitivity is a critical element of an access control mechanism.

◉ Data security and sensitivity issues

Data security issues should be adequately addressed to ensure complete security and prevent any unauthorized access to data sets. Electronic records are easier and faster in making multiple copies, modifying data, deleting etc. than manual records. Data security protocols applicable to the audited entity may be followed by auditors for handling the acquired data sets. SAI should ensure the audited entity’ s dataset or any other sensitive dataset not to be shared with unauthorized persons/entities, and Personally Identifiable Information (PII) information may be protected to ensure data security and privacy.

◉ Archival Policies

The roles of all personnel dealing with data should be well defined, including which tasks each IT staff should perform as part of their work. Policies should apply to the employment of permanent staff, temporary staff, contractors and consultants. Access controls, logs and monitoring facilities should be enabled to limit the access at various levels, depending on the user’ s requirements and confidentiality.

◉ Ownership of Data

The ownership of data sets remains to the audited entity/third party data sources and auditors shall hold such data only in a fiduciary capacity. Once data sets are obtained from the data sources, the Head of Audit should enjoy the ownership right and exercise such controls on data security and confidentiality as envisaged for the data owner in the audited entity. All concerns and instructions of the data owner, if any, should be ascertained and kept in mind. The data provided by data sources must be kept in safe custody for reference and all analysis must be undertaken only in copies of source data. It is necessary to ensure compliance to rules, procedures and agreements on data security, confidentiality and use of the audited entity/third party within the overall framework of data protection and security prescribed by the SAI from time to time.

◉ Data Sharing

A data sharing protocol should be established between the auditee and the auditor. While collecting data, the authenticity, integrity, relevance, usability and security of data sets should be ensured. In order to ensure the integrity of data (i.e. – from missing), it is necessary to check the total number of records or numeric columns (hash totals). In order to ensure the completeness of data, it is necessary to obtain a certificate stating that the data is complete and the same as in the IT system of the audited entity at the time of receiving data.

It is necessary to establish a Memorandum of Understanding (MoU) or data sharing protocol containing specific details on the mode (API, data dump, ftp, etc.), frequency of production of digital data, methodology, etc. A MoU should be routed through authorized personnel in SAI, who is entrusted with data handling matters (for example: Chief Data Officer) for ensuring that all the required data is captured and clauses on safeguarding of the interest of SAI is protected. This will also ensure that the MoU has a common framework and multiple MoUs for similar requirements are not executed by different entities.

◉ Data Collection

Data collection primarily takes place from internal sources (mostly structured data – state accounting data) and external sources (both structured and unstructured data to be collected from the auditee and third-party sources to give us insights into the risks associated with the auditee).

Data collection is a systematic approach to gather and measure information from a variety of sources to get a complete and accurate picture of an area of interest. The IT system should be studied and understood while collecting data, which would facilitate the identification and requisition of relevant data. They can be complete databases, selected tables out of databases, selected data fields of tables in the databases or data pertaining to specific criteria/conditions for a particular period, location, class etc. Depending on the data size, this may be obtained in flat file or dump file formats. Where it is impossible to obtain the relevant data/tables for analysis, the entire data may be collected.

◉ Sources of Data

Data is available to audit today, in different forms and from different sources. But it’s important to understand the data types and sources before initiating the process of acquisition, preparation and analysis. Data sources can be broadly classified as below:

Internal Data: Data available with SAI
External Data: Data pertaining to audited entities
Third party Data: Data available in the public domain

They may be collected with automated methods (Application Program Interface (API) based, or through File Transfer Protocol (FTP)) or through flat files and database dumps (from RDBMS/

Native format in which the application creates the data/tables extracted from an audit module developed in the system or may be extracted in formats like pdf, csv, excel, txt, etc.

◉ Data Restoration

Data from data sources should be copied and restored in the auditor’s computer for further analysis. While using data in a dump/backup format, it is necessary to bring the data table to its original format through data restoration.

◉ Scalability

Data size, source and velocity are continuously evolving. The SAI’s infrastructure needs be adapted to support an increasing volume of data or users, so it should allow computer equipment and software programs for growth over time, rather than replacement.

◉ Licensing Arrangement

Since SAI need to deal with multiple databases, it is increasingly urgent to decide on licensing requirements of various database software. Auditors need to answer whether they may accept the data in a native form or in a particular form decided by audit institutions.

◉ Data Virtualization

Data virtualization integrates data from disparate sources without copy or moving, which gives users a single virtual layer spanning multiple applications, formats, and physical locations, contributing to faster and easier access to data. With the need of linking multiple databases and querying on the same as if they all are coming from single source, it is urgent to explore the tools like Denodo.

◉ Log Trails

In audit, the most important thing is documenting the audit trails and for the same reason all the steps taken while cleaning/preparing/standardizing the data must be captured and recorded to justify data assurance as needed.

◉ Cataloguing and Metadata

A data catalogue is a completely organized service enabling users to explore their required data sources and know the location of a data source for data connection. Metadata management allows for an understanding of data flow, not only as an index like data catalogue. It shows data movements and help build, improve and monitor the systems.

An example of extracts from SAI-India’ s data analytics guidelines for data preparation and standardization.

◉ Data Preparation and Standardization

Identified datasets may not always be in a desired form, size or quality for analysis, so data have to be prepared from an available format to a desired format. Data understanding is a prerequisite for the auditor to decide on the ‘desired format’ for subsequent analysis.

Data preparation is the process of organizing data for analysis. It involves activities such as restoration, importing of data, selection of databases/tables/records/fields, joining or appending of datasets, as well as cleansing, aggregation and treatment of missing values, invalid values, outliers and transformation. They may either be interconnected or be a series of independent steps. Data preparation is a project’ s[①] specific phase. Though broad steps may not vary significantly, the order of sub-processes or tasks involved may vary from project to project. Further, there may be a need to back track or repeat certain steps/tasks.

◉ Extract, Transform, Load

Extract, transform, load (ETL) is the general procedure of copying data from one or more sources into a destination system which represents the data differently from the source or in a different context other than the source.

[①] A project here refers to a data analytic project while conducting an audit or an analysis of data obtained from data sources not necessarily connected to an audit.

◉ Identification of Tables/Fields of Interest

To optimize computational speed and capacity, it is essential to keep relevant data variables for analysis only. Relevant fields/tables/variables of interest may have to be identified with utmost care, as all the procedural steps may have to be repeated again if, at a later stage, any additional field/table/variable is found to be relevant.

◉ Importing into the Analytical Tool

Most analytical tools provide options to read flat files in the software or connect to a database and read tables, while some provide the option of importing relevant columns/tables and changing the data type before reading the file into the platform only. They offer options to clean and enhance the data. Depending on data quality and quantity, auditors may choose to deal with data cleaning/enhancement within or outside the analytic platform, in a spreadsheet or RDBMS. The steps of importing and data cleansing may precede or follow each other depending on the datasets and the availability of suitable tools.

◉ Merging and Splitting Data Files

Data received from data sources may pertain to different periods, or locations, or may simply be split into different parts. To make the data amenable to analysis, it is essential to merge the data sets into one by appending the data files. Similarly, different data sets pertaining to an entity contain details on different functions/ parameters. In such case, all the data files may be merged together to get all the parameters in one file for analytics.

Data files can also be split to make the data sets leaner for greater efficiency. Files may be split based on the number of records or parameters. Files can be merged or split through the RDBMS or data analytic tools.

Data Cleaning

High-quality data which is clean, complete and devoid of errors is essential for good analysis. Data cleansing, data cleaning, or data scrubbing is the process of detecting and correcting or removing corrupt or inaccurate records from a record set, table, or database. It refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or filtering out the inaccurate or corrupt data. The process of data cleaning may involve removing typographical errors or validating and correcting values against a known list of entities or by cross checking with a validated data set. Data cleaning may involve rejecting or correcting records and verifying the existence of any invalid values. Further with new sources of data like IoT based sensors being used by auditee entity to collect data (e.g., for utility-based services such as electricity and water), noise reduction techniques will be becoming increasingly important in coming days.

Standardization

Data enhancement is also a data cleansing process where data is made more complete by adding related information. It involves activities such as harmonization of data and standardization of data. For example, appending the name of a bank with a bank code enhances data quality. Similarly, the harmonization of short codes (st, rd, etc.) to actual words (street, road, etc.) can be done. Data standardization is a means of changing a reference data set to a new standard, e.g., use of standard codes.

Missing Values and Other Data Preparation Steps

Missing values occur when there is no data value available for the variable in a field in the dataset. It is common but reduces the representativeness of the dataset and may distort inferences and conclusions drawn from data. Missing values can occur in random or with some patterns. Understanding the reasons for and the nature of missing values is important to appropriately handle the remaining data. Based on the nature of missing values, the data set should be appropriately treated by either deleting the missing values or assigning them with certain other values such as the mean, median or mode of the available values.

Other data preparation steps include deleting unwanted columns, formatting and renaming various columns, and inserting additional columns.

◉ Data Integration: Linking Multiple Databases

Data integration is the process in which the data collected from various data sources or different tables within the same data source are combined to get a final dataset for analysis. Data from different sources can be integrated based on any common field such as unique customer id, bill number, village name, etc. For example, to understand whether the coverage of beneficiaries under a certain social security scheme is correlated to population distribution, the beneficiary data may be linked (joined) to the census data at district level, taluk level or even at further granular levels. Understanding the meta data of different sources will help data integration.

While linking multiple data sets, it is not necessary to have a common field in the data set as data can be aggregated at a higher level for comparisons. For example, while it may be impossible to link an individual beneficiary in pension beneficiary database and BPL database, the data can be aggregated at village/block/district level to identify whether there is a mismatch between these numbers. The reasons for such mismatch can then be explored during a substantive check by audit.

An example of extracts from SAI-India’ s data analytics guidelines for data verification.

◉ Data Integrity

In order to ensure data integrity (i.e. – data missing), it is necessary to check the total number of records or numeric columns (hash totals). In order to ensure data completeness, it is necessary to take control measures, e.g., taxes collected by individual taxpayers should add up to the total tax collected in the tax office. The auditor should obtain a certificate stating that the data is complete and the same as in the IT system of the audited entity at the time of receiving data.

Auditors should undertake statistical testing or some sort of hypothesis testing to verify the veracity of data. A completeness certificate obtained from the auditee on the data collected should ensure:

- The data dump is full, complete and whole of actual data.
- There is no erasure, tampering or overwriting of original data.
- There is no data inconsistency and there was no loss of data during data migration from one system to another or backup or due to theft/hacking etc.
- There is no damage of data i.e., by destruction, alteration, modification, deletion or re-arrangement of any computer resources by any means.

◉ Data Reliability

Data reliability is a function of data authenticity, integrity, relevance and usability. It can be affected because of the methods for generation/capture of data.

Generally, auditors have limited means to ensure the reliability of the data when received from the auditee entity as its reliability can be assessed only in the audit process, during which analysis can reveal internal inconsistencies or incompleteness. However, auditors need to be vigilant about data reliability and exercise due precaution while obtaining data from the auditee entity. Generally speaking, when the manual and IT systems operate in parallel, the chances of errors in data become higher. Similarly, an MIS system involving manual data entry is likely to be less reliable than systems where MIS data is directly generated through an IT system.

INTOSAI

Working Group on Big Data

Secretariat of INTOSAI Working Group on Big Data

Add： National Audit Office of the People's Republic of China
No.17 Jin Zhong Du Nan Jie, Fengtai District, Beijing 100073, China

Tel: + 86 (10) 50992782

E-mail: intosai-wgbd@audit.gov.cn