Guidelines on Data Analytics

Office of the Comptroller and Auditor General of India

2017

Table of contents

Preface

Technology plays a significant role in modern day governance for enhancing delivery of public goods and services. The diverse technology systems are continuously producing volumes of data in disparate forms, throwing up immense opportunities for data analytics.

As a responsive Supreme Audit Institution, we have to be institutionally agile to keep pace with such developments and embrace the evolving opportunities in data analytics. The Big Data Management Policy, formulated in 2015, envisioned the broad contours of the data analytic framework for the Department. Creation of the Centre for Data Management and Analytics was the first step in establishing this framework.

The Guidelines for Data Analytics is a major initiative in institutionalising the practice and use of data analytics in the Department. These guidelines explain the concept of data analytics, outline the data analytic process and envisage development of data analytic models. Data analytics is an evolving discipline and therefore these guidelines would have to be periodically reviewed and updated.

I am sure that officers and staff of the Department would find these guidelines useful and would apply them purposefully towards enhancing the quality of public accounting and auditing.

Shashi Kant Sharma
Comptroller and Auditor General of India
September 2017

1. Data Analytics

Introduction

1.1 **Data analytics is the application of data science[1] approaches to gain insights from data. It involves a sequence of steps starting from collection of data, preparing the data and then applying various data analytic techniques to obtain relevant insights. The insights include, but are not limited to, trends, patterns, deviations, inconsistencies, and relationships among data elements identified through analysis, modelling or visualization, which can be used while planning and conducting audits.**

Data analytics adds a competitive advantage to enable information based decision making. As it is an evolving discipline, the possible utilities of data analytics are still under experimentation and exploration in both public and private sector.

1.2 These guidelines prescribe the methodology of employing data analytics in the auditing function of Indian Audit and Accounts Department (IA&AD). The data analytic principles and methods will, however, be applicable to the domains of accounting and administration.

1.3 These guidelines have been developed as a follow up of the Big Data Management Policy issued in September 2015 and subsequent initiatives in use of data analytics in IA&AD, particularly in audit. The guidelines draw on the existing guidelines on Performance Auditing, Compliance Auditing, Financial Auditing,

[1] Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information. Although the name Data Science seems to connect most strongly with areas such as databases and computer science, many different kinds of skills - including non-mathematical skills - are needed - An Introduction to Data Science - Jeffrey Stanton, Syracuse University.

Auditing Standards and other relevant instructions and manuals in IA&AD.

Scope for individual initiative and professional judgment

1.4 While these guidelines are prescriptive in nature, these are not intended to supersede the professional judgment of the Accountant General[2]. The Accountant General is expected to make situation or subject specific adjustments to the provisions set out in these guidelines. However, Accountants General will be expected to document the rationale of all significant departures from the guidelines and obtain authorization from the competent authority.

Data analytics and IA&AD

1.5 The IA&AD has a very broad audit mandate, which includes audit of Union Government and State Governments and extends to bodies or authorities such as statutory corporations, government companies, autonomous bodies constituted as societies, trust or not for profit companies, urban and local bodies and to any other body or authority whose audit may be entrusted to the Comptroller and Auditor General of India. Audits are conducted with reference to such accounts, vouchers and records as may be received in the audit office and/ or in the accounts office and may include online data, information and documents of the auditable entity. The Auditing Standards envisage obtaining sufficient and appropriate evidence to support the auditor's judgment as well as conclusions regarding the organisation, programme, activity or function under audit. This would involve study and analysis of data collected before and during the audits. With limited available resources, Audit undertakes a risk based audit approach and applies analytical

[2] The term Accountant General includes all heads of Departments (HoD) of the rank of Senior Administrative Grade and above, within the IA&AD.

procedures, test of controls and substantive checks on available and selected data during planning and execution of the audits.

With rapid computerisation, most of the activities of auditable entities are being recorded electronically, in various IT systems. These electronic records or 'data', if interpreted properly, can provide insights into past events, guide corrective action in the present and forecast future events thereby enhancing the efficiency of the auditor.

1.6 Data is available to audit today, in different forms and from different sources. Data analytics provides the potential to analyse these data sets and obtain insights to assist in the audit processes by identifying patterns, trends, descriptions, exceptions, inconsistencies and relationships in data sets and their variables. The insights so drawn would assist in setting the direction of the audits, by primarily identifying areas of interest or risk and in identifying exceptions.

Data analytics in Audit

1.7 Data analytics begins with identification and collection of various data sources for a particular audit. The analysis of data through various data analytic techniques will yield insights on the working of the audited entity. The risk areas or areas of interest identified through such an exercise will assist in identifying audit objectives and developing an Audit Design Matrix. Data Analytics will also assist in identifying the sample of audit units where substantive checks will be conducted.

1.8 The various analyses can then be built into a re-executable Data Analytic Model. This will ensure that results of data analysis can be used repetitively with periodic updating of data. Establishing a mechanism for receiving data periodically will be crucial for such an approach. The scope of the model once built can be expanded by

incorporating the feedback from substantive checks and bringing in additional data sources. Thus, data analytics in IA&AD is not envisaged to be a one-off process for a specific audit, but is expected to evolve over time.

1.9 The schematic diagram of the process is provided below:

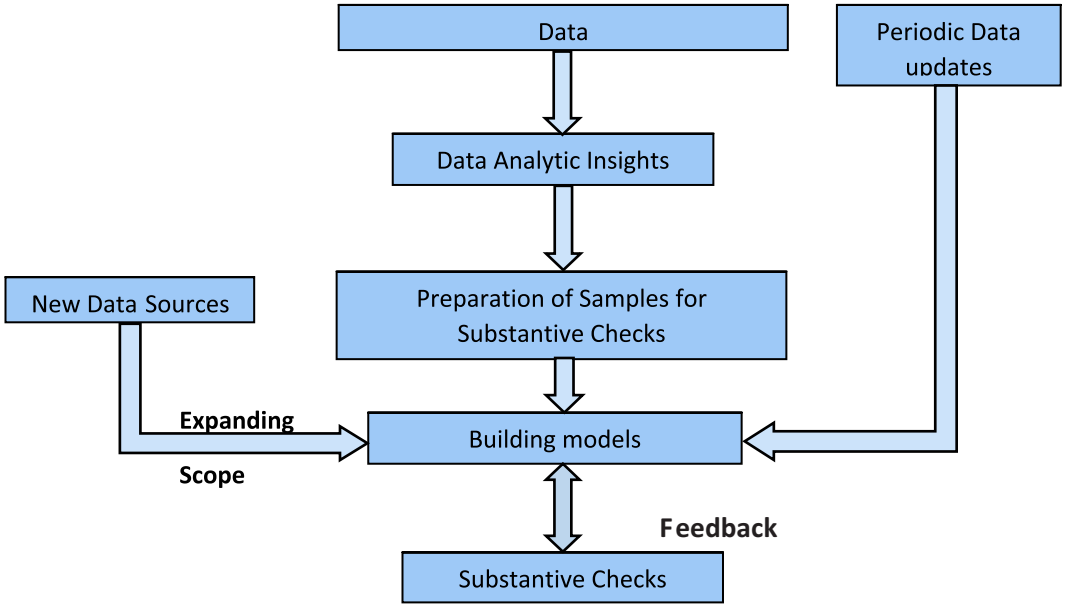

Figure 1- Data analytic process

The data analytic process has been explained in detail in subsequent chapters.

1.10 The Centre for Data Management and Analytics (CDMA) will be the nodal body for steering data analytic activities in IA&AD. CDMA will provide guidance to the field offices on data analytics and pioneer research and development in the future direction of data analytics.

In the structure envisaged for data analytics in IA&AD, data analytics is to be conducted by each field office as per its annual plan. The data analytic activities in a field office will therefore be the responsibility of the Head of Department (HoD), who will constitute a Data Analytic Group. The Data Analytic Groups constituted in the field offices under the charge of a Group Officer will be responsible for steering data analytics in the field offices. To obtain meaningful insights for audit from data analytics, knowledge in the area of audit will be essential. The exercise of data analytics is therefore envisaged as a collaborative effort with technical knowledge of the Data Analytic Groups and domain expertise from functional groups in the field office, complementing each other. An indicative role assignment for data analytic activities is provided at **Annexure 1.**

Hiring of external experts

1.11 In specialised areas, field offices could consider engagement of external experts, if such need is justified. Engagement of external experts should, however, be as per the guidelines issued by IA&AD from time to time. Some of the specialized areas for such hiring could be related to data handling, applying advanced data analytic techniques or management of data repository.

2. Data Acquisition and Preparation

2.1 Data analytic process encompasses data acquisition, data preparation, data analysis, results and analytic models. This chapter addresses identification and collection of data as well as handling of collected data and preparing it for analysis. It is however, important to understand the data types and their sources before initiating the process of acquisition, preparation and analysis.

Understanding data types

2.2 The core of data analytics is 'data'. Data can be measured, collected, analysed and visualized to give a meaningful interpretation of facts and reasons. Data can be understood and categorised as follows:

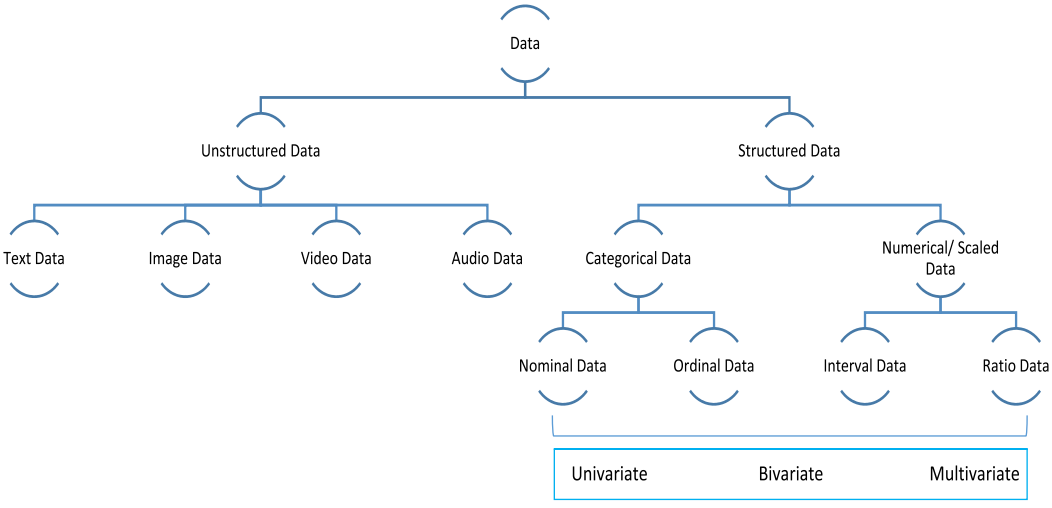

Figure 2- Types of data

- **Unstructured or structured data**: Unstructured data comprises data such as text, image, audio or video data, which cannot be readily 'tabulated' for statistical or mathematical analysis. Structured data on the other hand refers to data in tabular form. Structured data could be categorical or numerical.

- **Categorical or numerical data**: Categorical data could be nominal (data not amenable to ordering e.g., name, gender of a person) or ordinal (data amenable to ordering e.g., ranking based on quality of service: highly satisfied; satisfied; not satisfied). Examples of numerical data could be interval data (e.g. temperature which is amenable to identifying differences in values) or ratio data (e.g. expenditure of a company which can be compared as multiples of one another).

Operation	Nominal	Ordinal	Interval	Ratio
Count	✓	✓	✓	✓
Ordering of values		✓	✓	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Addition/Subtraction			✓	✓
Multiplication/Division				✓
Whether true zero exists				✓

Figure 3- Possible operations with types of data

- **Number of variables – Univariate, bivariate or multivariate data**: Based on the number of variables in a data set, it may be called univariate, bivariate or multi variate data. Univariate data has only one variable. It is essentially descriptive in nature. Analysis of univariate data involves summarization and identification of patterns in the data. Bivariate data has two variables and statistical analysis can be applied to understand the relationship between two variables. They can be represented on X-Y axis and visual representation through plots like scatter plot will be useful in understanding relationship patterns in this type of data. Multivariate data involves multiple variables. Statistical analysis would be required to analyse the data and to discover relationships and dependencies between the variables. Visual representation is a useful tool in

understanding the relationship patterns among different variables, and the plots can be drawn on three dimensions, X, Y and Z. Plots can thus include more than three variables using appropriate visualization approaches.

Sources of data

2.3 Identification of various sources of data available to the IA&AD is the corner stone of the data management framework. The Big Data Management Policy categorises various data sources as:

Internal data sources: This comprises

- Combined Finance and Revenue Accounts
- VLC data base
- GPF and Pension data in A&E offices
- Data generated through Audit process
- Any other data available in the department

External data sources: This comprises

a) **Audited entities' data** available with the department in its professional capacity which includes
   - Financial and non-financial data of audited entities
   - Programme specific data including beneficiary databases
   - Other data pertaining to audited entities

b) **Third party data** which comprises data available in the public domain and includes:
   - Data published by Government and statutory authorities like
     - Census data
     - NSSO data
     - Data published by the various Ministries/Departments
     - Data available in data.gov.in

- Reports of various commissions
- Other Reports and data pertaining to Union Government /States

- Other data available in public domain
  - Surveys and information published by NGOs
  - Industry specific information published by CII, FICCI/NASSCOM etc.
  - Sector specific information published by various organizations
  - Social media etc.

2.4 Field offices may encounter situations where the required data is available in manual form. The field offices should then decide whether the manual data can be converted into electronic form by creating electronic data sets. For instance the details contained in sanction orders received in audit offices may be converted into an electronic data file, which can be utilised for data analytics.

Data Identification

2.5 As a part of collecting and maintaining a comprehensive data base on auditable entities, field offices should formulate a mechanism for identifying availability of electronic data with audited entities/third party data within their jurisdiction and updating them periodically.

Data acquisition

2.6 Data acquisition involves obtaining access to and collecting data keeping in view the ownership, security and reliability of data collected.

Data access

2.7 Since IA&AD is not the owner of several data sources required for data analytics, data availability would remain a challenge in the

medium term. Exacerbating this problem is the reluctance by many of the audited entities to part with their data. Continuous persuasion and monitoring with the audited entities taking support from relevant provisions of the CAG's Duties, Powers and Conditions of Service, Act 1971 and Regulations on Audit and Accounts 2007 will be the way to address this issue.

2.8 Data may be provided to the auditors on the entity's sites through access to the system. This can be a read-only access without any transaction rights so that the system's performance is not affected. The data may be provided through backup files created in the entity's environment and shared on a removable media with the auditors. The data may also be shared electronically using electronic transfers though networks - LAN or WAN or internet or a VPN, as the case may be.

2.9 Indicated below is a progression in the way auditors can access data from their audited entities, starting from manual records to online, real time data sharing. However, it is not essential that the progression be sequential and auditors accessing only manual records may start accessing real time data electronically without going through the intermediate steps. The access to data solely depends on the capability of the auditors, the auditing environment and the level of access established between the two.

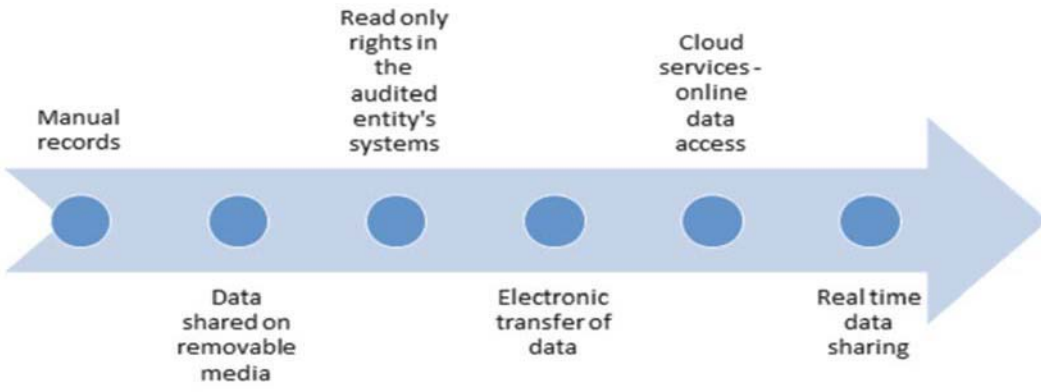

Figure 4- Access to data

2.10 One of the ways to deal with data access is through involvement of audit from the design stage of the IT systems when it may be possible to incorporate the data requirements of audit into the system design. This would facilitate acquisition of data in the requisite format. To ensure this, field offices would need to convey the data requirements for audit to the concerned entities at the stage of important system developments thereby facilitating access to requisite data when the system is operational. These data requirements could cover information sets to be acquired, format of data, mode of transfer and periodicity of data to be made available to audit. At the same time, access to the complete system or complete data, if required for any specific audit, such as performance audits, systems audits, IT audits, special audits etc., should not be precluded by involvement of auditors at the system development stage.

Data handling at different levels of data access modes

2.11 When the data is shared in removable media, the auditors need to have hardware compatible to run the media - CD, DVD, tape drive or an USB drive etc. Along with the capacity to run the media, the auditors need to have appropriate operating system and database

application (like the RDBMS) where the data can be read from the media. Thus, an environment similar to the source from which data is received is to be created to be able to read the data. Read only rights are typically the view rights granted to the auditors at the entity's systems which should facilitate viewing/copying of the requisite data. In electronic transfer of data, the data in file form is transferred using networks such as through mail, file transfer protocols etc. In online access, data is made available through cloud from a remote server. Real time systems provide access to live systems and the information contained therein in a real time mode. Real time data access provides the possibility of real time processing, thereby enabling the development of continuous auditing approaches through embedded audit modules[3].

All field offices should endeavour to evolve an appropriate data access mechanism with the data source organisations so as to access data on a periodic/real time basis into their data repository/ data analytic models.

Collection of data

2.12 Data collection is the systematic approach of gathering and measuring 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 identification and requisition of relevant data. These can be complete databases, selected tables out of the databases, selected data fields of tables in the databases or data pertaining to specific criteria/ condition for a particular period, location, class etc. Depending on the data size, this may be obtained in flat file or

[3] Embedded Audit Module- Audit module embedded/ integrated with the IT systems, thus receiving online data including real time data.

dump file formats. Where it is not possible to obtain the relevant data/tables for analysis the entire data may be collected.

2.13 While collecting data, the authenticity, integrity, relevance, usability and security of the data sets should be ensured[4]. For ensuring the integrity of data (i.e. – that some data is not lost), checks such as counting the total number of records or sum of numeric columns adding up to total (hash totals) may be undertaken. For ensuring that data is complete, completeness control measures should be undertaken, 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. An indicative template of such certificate is provided at **Annexure 2.** It should be ensured that only authorized personnel handle data transfers from the data sources to the auditors. The access to such data should be through appropriate access controls to prevent any unauthorized access to data.

Data from an entity not within audit jurisdiction

2.14 Field offices may require data sets whose ownership is not with auditable entities under their audit jurisdiction. The field office may then seek the assistance of the concerned field office which has audit jurisdiction over such auditable entities and the concerned field office should provide all assistance in obtaining the required data sets.

[4] Big Data Management Policy, Section IV-2. Data Management protocols have to ensure that data satisfies the following characteristics:
Authenticity - Data is created through the process it claims. **Integrity** - Data is complete, accurate and trustworthy. **Relevance** - Data is appropriate and relevant for the identified purpose. **Usability** - Data is readily accessible in a convenient manner. **Security** - Data is secure and accessible only to authorised parties.

Ownership of data

2.15 The ownership of the data sets remains that of the audited entity/ third party data sources and IA&AD holds this data only in a fiduciary capacity. Once the data sets are obtained from the data sources, the HoDs should assume the ownership of the data sets and should exercise such controls on security and confidentiality of the data as envisaged for the data owner in the audited entity. The concerns and instructions of the owners of data, 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 the source data. Compliance to all rules, procedures and agreements regarding data security, confidentiality and use of data of the audited entity/ third party must be ensured by audit within the overall framework of data protection and security prescribed by IA&AD from time to time.

Data security

2.16 In case of electronic records, making multiple copies, modifying data, deleting etc. are easier and faster when compared to manual records. Data security protocols applicable to the audited entity may be followed by the auditors for handling acquired data sets. The data analytics results, however, may be dealt with in the manner prescribed by IA&AD.

2.17 While handling data, the basic approach should be to limit, to the bare necessity, the number of personnel with access to the raw data and to establish a trail of personnel who have accessed data. Complete and chronological record of all data shared between data source owner and the auditor should be stored in an unaltered and secure manner. It should be ensured that computers which are used for data analytics are not connected to internet.

2.18 Given the sensitivity of the data obtained from the audited entity, it should be handled with due diligence to avoid any kind of unauthorised disclosure from auditors. Information Security measures of government[5], those specified in Information Systems Security Handbook of IA&AD, along with any specific agreement between the auditor and the data source owner should be followed to ensure confidentiality and security of data.

Data reliability

2.19 Data is said to be reliable when the data accurately captures the parameter it is representing. Data reliability is a function of authenticity, integrity, relevance and usability of data. Data reliability can be affected because of the methods of generation /capture of data. As IA&AD has to rely on data generated from other sources, it is important that reliability of each data source is understood *a priori* so that adequate caution can be exercised in its utilisation.

2.20 Generally auditors would have limited means to ensure reliability of data while receiving data from the auditable entity as reliability can be assessed only after using the data in audit process, when analysis could reveal internal inconsistencies or incompleteness. However, auditors need to be vigilant about data reliability and exercise due precaution while obtaining data from auditable entities. Generally, if the manual and IT system are operating in parallel, the chances of errors in data are 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. Information System audit of the IT system, if any conducted earlier, can provide insights on data reliability.

[5] Guidelines for use of IT Devices on Government Network dated 14 October 2014, *http://meity.gov.in/writereaddata/files/Guidelines%20for%20Use%20of%20IT%20Devices%20on%20Government%20Network%20_0.pdf*

2.21 Auditors need to clearly differentiate between the purposes for which the data set would be put to use while considering data reliability. Consideration of data reliability would be significantly higher for data sets planned for usage as audit evidence to support audit conclusions as compared to data sets planned for drawing broad insights while planning. The Big Data Management Policy mentions various third party data sources which can be used for audit in IA&AD. While third party data can strengthen the audit planning process, the auditor should use professional judgment while using such data sources as audit evidence and should ensure that it meets the criteria laid down as per auditing standards of CAG of India. For example, Survey Data of an academic institution related to sanitation can be used to identify issues in the sector and may feed into the sampling process (identifying high risk /low risk administrative units), along with other parameters in the audit planning stage. However, whether the analytic results of the survey data can be used as audit evidence depends on whether it satisfies the conditions, criteria and standards of audit evidence laid down for IA&AD.

Data preparation

2.22 The identified datasets, as available, may not always be in the desired form, size or quality for analysis. Hence the data would have to be prepared from the available format to the desired format. Understanding the data is a prerequisite for the auditor to decide on the 'desired format' of data for subsequent analysis.

2.23 Data preparation is the process of organizing data for analytic purposes. It involves various activities such as restoration, importing of data, selection of database/ table/ record /field, joining datasets, appending datasets, cleansing, aggregation and treatment of missing values, invalid values, outliers and

transformation. These activities may either be interconnected or be a series of independent steps. Data preparation is a project[6] specific phase. Though the broad steps may not vary significantly, the order of the sub-processes or tasks involved may vary according to the project. Further, there may be a need to back track or repeat certain steps/tasks.

Data restoration

2.24 The data from the data source should be copied and restored in the auditor's computer for further analysis. While using data in dump/ backup format, it will be necessary to bring the data tables to its original format through a data restoration process.

Before restoring a database backup/dump file, some basic information such as database software version, operating system, database size is required. Based on this information, an environment should be created to restore the backup/dump file, if not already present. Database restoration requires adequate technical knowledge of the database, as steps that need to be followed while restoring a database may vary according to the database software. While it may be possible to restore a lower version backup /dump file in a higher version of database software, it could involve compatibility issues, which should be confirmed from the Database Administrator.

Identification of tables/fields of interest

2.25 In order to optimize computational speed and capacity, it is essential that only the relevant data variables are kept for analytical purposes. Identification of the relevant field/table/variable of interest would have to be carried out with utmost care as all the

[6] A project, here, is a data analytic project either while conducting an audit, or while carrying out analytics of data obtained from data sources, which is not necessarily connected to an audit.

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

2.26 Most analytical tools have options to read flat files into the software or connect to a database and read tables. Some analytic software provide the option of importing only the relevant columns/tables and changing the data type before reading the file into the platform. The analytical tool itself offers various options to clean and enhance the data. Depending on the quality and quantity of data, the auditor may choose to do the data cleaning/enhancement within the analytic platform or outside, in a spreadsheet or RDBMS. The steps of importing and data cleansing may precede or follow each other depending on the datasets and availability of suitable tools.

Merging and splitting data files

2.27 Data received from the 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 will be essential to merge the data sets into one. This can be done by appending the data files. Similarly, different data sets pertaining to an entity contain details on different functions/ parameters. In such cases, all the data files may be joined together to get all the parameters into one file for analytics.

2.28 Data files, can also be split to make the data sets leaner thereby assisting efficient analytics. The files may be split either based on number of records or on number of parameters. Merging and splitting of files can be carried out through the RDBMS or the data analytic tools.

Data cleaning

2.29 Good 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.

2.30 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 any Bank Code enhances the quality of data. Similarly, harmonization of short codes (st, rd, etc.) to actual words (street, road, etc.) could be done. Standardization of data 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

2.31 Missing values occur when no data value is available for the variable in a field in the dataset. This is a common occurrence, which reduces the representativeness of the dataset and can distort inferences and conclusions drawn from the data. Missing values can occur in random or with some pattern. 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.

2.32 Other data preparation steps include deleting unwanted columns, formatting and renaming various columns and inserting additional columns (say inserting an additional 'YEAR' column for trend analysis).

Data integration: linking multiple databases

2.33 Data integration is the process whereby the data collected from various data sources or different tables within the same data source are combined to obtain the final dataset for analysis. Data from different sources can be integrated based on any common field such as Unique customer id, Bill number or Village name etc. For example, to understand whether the coverage of beneficiaries under a certain social security scheme is correlated to the 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[7] of different data sources will aid the process of data integration.

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

[7] Meta data is the data of other data sets. It contains information on the data sets in a manner to make it easier to identify the data sets.

3. Data Analysis and Modelling

Data analytic approaches

3.1 The data after due preparation is analysed to derive insights using various analytic approaches. The following approaches can be used in data analytics:

- **Descriptive analytics** tries to answer "what has happened". In descriptive analytics, raw data is summarized so that it can be understood by the user. Descriptive Analytics provides an understanding of the past transactions that occurred in the organisation. Descriptive analytics involves aggregation of individual transactions and thus provides meaning and context to the individual transactions in a larger perspective. It involves summarization of data through numerical or visual descriptions.
- **Diagnostic analytics** is an advanced form of descriptive analytics and tries to answer the question "why did it happen" or "how did it happen". Diagnostic analytics involves an understanding of the relationship between relatable data sets and identification of specific transactions/ transaction sets along with their behaviour and underlying reasons. Drill down and statistical techniques like correlation assist in this endeavour to understand the causes of various events.
- **Predictive analytics,** as the name implies, tries to predict, "What will happen", "when will it happen", "where will it happen", based on past data. Various forecasting and estimation techniques[8] can be used to predict, to a certain extent, the future outcome of an activity.

[8] Forecasting and Estimating techniques, involve use of past data, available knowledge / documents, assumptions and identified risks and form part of the operations research and quantitative techniques disciplines.

- **Prescriptive analytics** takes over from predictive analytics and allows the auditor 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.

Data analytic techniques

3.2 Data analytic techniques are employed to leverage the above approaches. The analytical techniques that use descriptive and diagnostic approaches would help the auditor to understand the auditable entity and to identify issues therein. A predictive technique like regression will help understand the behaviour of one (or more) variables based on the changes in the other set of variables. These analytic techniques can be broadly classified as Statistical and Visual[9].

- **Statistical techniques** are the use of statistical measures to derive insights about the dataset.
- **Visualisation techniques** are the use of visuals, graphs and charts to derive an understanding and insight into the dataset.

A combination of various statistical and visual techniques is usually employed for data analytics.

3.3 There are inbuilt algorithms for the above-mentioned approaches in data analytic software. However, there are no laid down sequential steps for application of data analytic techniques, which can broadly be described as the zoom out – zoom in – filter approach. The data is first understood at a bird's eye view, followed by a drill down, to understand the data at a deeper level. Subsequently, a filter is done or a query is run to extract results or

[9] Though most of the visualisation techniques like graphs and charts are essentially statistical, they are different in the sense that the understanding is derived not from mere statistical measures but from comparing, analysing and deriving insights visually.

exceptions, if necessary. For example, in a data set containing property tax demand and collection, one can understand the average range of tax demands/collections and range of taxpayers and their distribution across zones with a zoom out. With a zoom in, one can understand the correlation between variables and discern patterns of tax payments across zones. Subsequently, the tax arrears can be filtered at the highest risk zone identified. Further a regression analysis can also be done to see which zones are most likely to have maximum arrears in future.

Statistical techniques

3.4 Once the data is prepared, as a first step, descriptive statistics of the dataset can be produced to summarize the data in some way with each statistical measure describing the data set. This can be complemented by simple graphical representations such as line graphs, histograms or scatter diagrams. For example, the measures of central tendency describe the expected normal behaviour of the entity and its elements, with respect to a particular parameter or variable. The measures of spread indicates the distribution of the data points. Relationship between two or more variables can be explored or established using techniques of correlation[10] and regression[11]. Identification or segregation of important parameters can be done using regression, component analysis[12] or factor analysis[13]. Clustering[14] and classification[15]can be used to

[10] Correlation is used to measure the strength of association between two variables and ranges between -1 to +1.

[11] Regression analysis gives a numerical explanation of how variables relate, enables prediction of the dependent variable(y) given the independent variable.

[12] Principal Component Analysis aims to reduce the number of inter-correlated variables to a smaller set which explains the overall variability.

[13] Factor Analysis aims to group together and summarise variables which are correlated thereby enabling data reduction.

[14] Cluster analysis is a multivariate technique used to group individuals/variables based on common characteristics.
(Ref: www.statstutor.ac.uk)

[15] The process of arranging data into homogenous group or classes according to some common characteristics present in the data is called classification.
(Ref:http://www.emathzone.com/tutorials/basic-statistics/classification-of-data.html#ixzz4r2Rlugdu)

identify group(s) in the data sets based on one or more similarity. The results from different statistical tests need to be read together to get a final understanding of the dataset.

Data visualisation

3.5 Data Visualization serves the following two distinct purposes:

- **Exploratory Data Analysis (EDA)***:* It is an approach to analysing data sets to summarize their main characteristics, often with visual methods. Primarily, EDA is undertaken for seeing what the data can tell us beyond the statistical analysis and modelling.
- **Communication of findings / reporting**: Insights derived from data can be communicated to the users such as higher management or the readers of audit reports. Data visualisation is a powerful technique to communicate data analytic insights.

3.6 Data visualisation aims at achieving one or more of the following objectives:

- Comprehensibility: makes information and relationships easily understandable.
- Comprehensiveness: presents features/information for the entire selected data set/sample size as against selective reporting.
- Focussed communication: facilitates concise and 'to the point' communication.
- Reducing complexity: simplifying the presentation of large amounts of data.
- Establishing patterns and relationships: enables identification of patterns and relationships in the data.

- Analysis: promotes thinking on 'substance' rather than on 'methodology'. It focusses on the essence of the finding being communicated rather than on the procedure for communication.

The IA&AD Practitioner's Guide for use of Data Visualisation and Infographics should be referred to for principles of data visualisation.

3.7 It should be noted that a single technique will not give a comprehensive understanding of the data set(s). An auditor should apply a combination of Statistical and Visual techniques to derive insights. The suitability of techniques depends upon the dataset and the purpose of the auditor.

Population statistic instead of Sample statistic

3.8 With the help of modern data analytical tools, it is possible to analyse the whole of the data set. Thus, inferences on the population (all transactions included in the data set) can be made by analysing all transactions in the data set instead of inferring through samples. However, substantive checks will be required if the data set is not fully representative of the complete business process captured by the IT system.

Data analytic tools

3.9 Data analytics is a multistage process involving preparation, analysis and building of models, with different requirements at each stage. There are many powerful open source[16] and proprietary software[17] available for the purpose. No single tool can be termed comprehensive or suitable for all analytic or data

[16] Knime(www.knime.org), R(www.r-project.org), Python (www.python.org) , Weka, Rapidminer, SPAGO are some of the open source tools

[17] SAS, Tableau, MS Power BI17, Tidco Spotfire, Informatica, IBM Analytics, SPSS, D3J,Qlik etc. are some of the proprietary tools

extraction requirements. Some tools are useful in data preparation, while they may be found lacking in data visualisation. Similarly, there are tools with powerful visualisation features which lack in the ability to carry out advanced statistical analysis.

3.10 While auditors may explore and adopt any of the open source or proprietary software, care should be taken with respect to the sustainability of the tool and data security. When adopting a new analytic tool, the HoD should consider the issues of sustainability of the tool in terms of financial and human resources. The scalability (vis-a-vis size and variety of data sets) of the tool also needs to be kept in mind apart from the availability of the tool in future. The HoD should also ensure that the audited entities' dataset or any other sensitive dataset does not get shared in the server/cloud environment of the data analytic software with unauthorized persons /entities. By way of abundant caution, whenever usage of a new tool is being formalised in an office, approval for the same may be obtained from CDMA.

Data analytic results

3.11 The results of data analytics can be in the form of:

- Audit Insights
- Audit evidence

Audit insights

3.12 The auditor applies various statistical and visualisation methods to derive insights from the data. There are no laid down series of steps for deriving insights and it may involve backtracking and repeating of steps by way of iterations. The auditor should catalogue all insights that are thus derived. It must be kept in mind that while all statistical findings will describe some pattern(s), not all of them lead to new insights. The insights obtained from data analysis may confirm the previous understanding of the data/entity. The insights

thrown up by data analysis should, therefore, be appreciated collectively.

3.13 Domain knowledge is essential for appreciating the results derived from the data analytic process. The findings[18] generated using various analytic techniques should be catalogued and checked with the domain experts to understand their value and significance. These insights can then be used to identify the risk areas/ areas of interest for audit. A template for cataloguing and documenting the statistical findings and insights is provided at **Annexure 3.**

Audit evidence

3.14 The auditor applies professional judgement in evaluating the data analytic results for using them as audit evidence to support audit findings and conclusions. The data analytical results may have to be validated by other forms of evidence gathered through substantive checks. The data analytical results would qualify as audit evidence when they meet the requirements prescribed in the Auditing Standards.

[18] The findings from analytics here are different from audit findings. The analytic findings are as discovered through analysis, which lead to insights. The insights are pursued in audit through substantive checks to confirm an audit finding.

Data analytic models

3.15 Data Analytic model refers to the set of analytic tests leading to analytic results, which can be used repetitively by updating/ changing data. Building a model will ensure that risk analysis once done on specific dataset/s can be used repeatedly by using the same data set for subsequent years/ periods, once a mechanism for obtaining the data periodically is established. The process creating a data analytic model is explained with the following flow diagram:

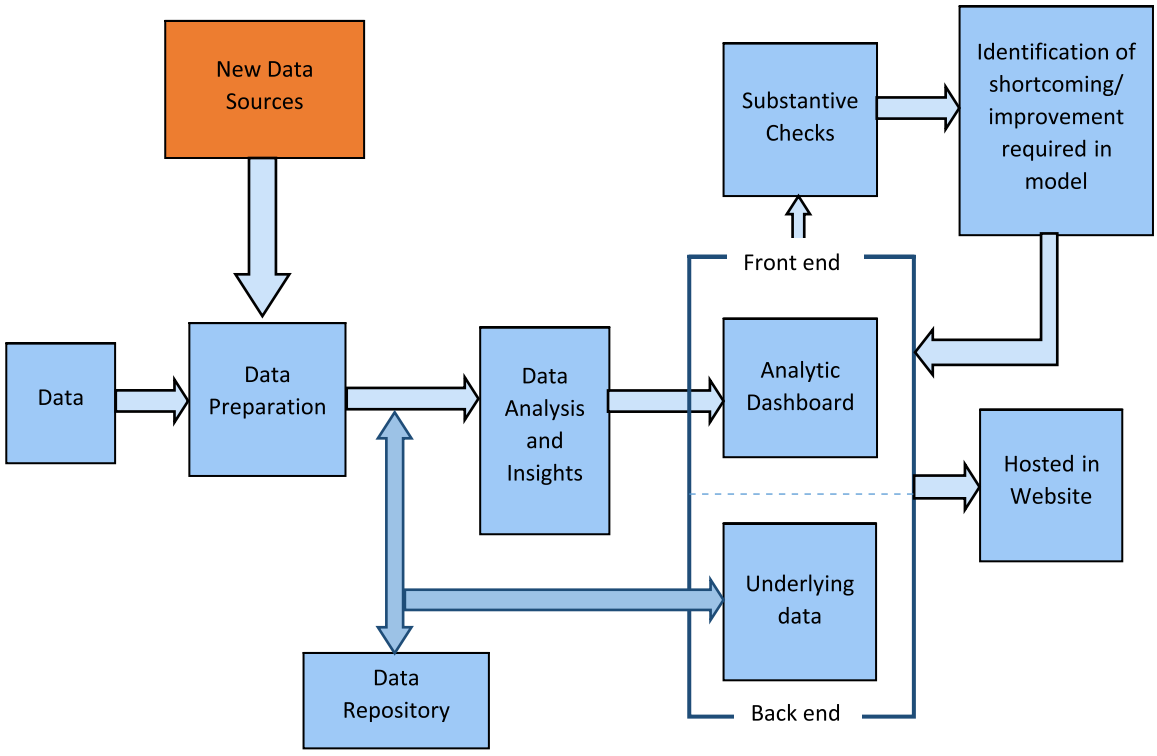

Figure 5 – Process flow of a data analytic model

3.16 To build a data analytic model, the following steps are followed:

- Data available from different sources are prepared for data analytics through restoration and cleaning of the data set.
- After the preparation stage, the data is stored in the data repository.
- Insights are drawn by applying different analytic techniques on the data sets fetched from the data repository and/or data available after data preparation.
- The relevant insights are converted into data analytic models. The models may comprise of equations, queries, workflows or dashboards[19].
- Once a model has been prepared, it should be submitted to CDMA for review and approval.

3.17 Data models could be developed on centralised or de-centralised data sources:

- **Centralised data sources**: If the data of the auditable entity/ sources is centralized, i.e., is available through a central database, a model can be built directly on the restored database. Alternately, relevant tables/ fields from the audited entity's database can be extracted to develop the model. When new data for subsequent periods are available, it may be incorporated by passing this through the restoration and cleaning (data preparation) stages before appending/loading into the model.
- **Decentralised data sources**: If the data of audited entity/ sources is decentralised, (i.e. – data from each audited

[19] Dashboard can be understood as an interface (usually interactive) which is used to showcase information/data in a more presentable manner. In data analytics, multiple insights (in the form of graphs/charts) can be brought together in a dashboard for understanding of the viewer.

entity sub unit is at different locations which are not connected seamlessly), then the model may be used at a sub unit by replacing the back end data of earlier sub unit with its data into the model.

- If data is being received on a real time basis, the model will also get updated on a real-time basis leading to the possibility of continuous auditing.

3.18 A preliminary model will encompass various insights thrown up by data analytics. However, there is a possibility that all factors may not have been considered, or were not available for data analytics, while the model was developed and further insights may be obtained when the model is deployed. The data model needs to be updated with these additional insights as well as with more relatable data sets that subsequently become available.

3.19 An important feature of the model is its reusability. A model, once created, can be used repeatedly by updating data. Hence, the utility of the model will depend on periodic updating of data. Therefore instead of treating the data collection process as a one-off exercise, a mechanism should be put in place to obtain data, annually/periodically. The data sets to be obtained from the data source/ auditable entity, including the data files or data tables should be clearly specified along with the mode of data transfer. If required, nodal officers should be identified for data handling. Data security issues should be adequately addressed to ensure complete security and to prevent any unauthorized access to the data sets. Obtaining endorsement at the senior levels of the audited entity will be essential for ensuring availability of data periodically.

3.20 As the number of data sets increase, the complexity of data management for the model will increase. It is preferable that the

model is not built directly on the restored data provided by the audited entities. Instead, relevant tables should be extracted and used, for creation of the model. Access to the model should be provided to users based on their access control profile (need to know/ need to use).

Documentation of data analytic process

3.21 Documentation of the analytic process facilitates planning, performance, and supervision of the analytic project. Documentation also aids review of the analytic process, including maintenance of data integrity during the process and providing suitable audit trail for data handling. Apart from supporting the auditors' results and findings, it assists future audit teams in repeating the analytic process. Documentation for data analytics should follow the Auditing Standards of IA&AD. All documentation should be signed by the auditor and countersigned by the supervising audit officer.

3.22 Documentation of the data analytic work should include the following aspects:

- Data identification
- Data collection
- Importing data into analytic software
- Analytic technique used
- Results of analysis
- Data Analytic Model
- Feedback from use in audit

Data repository

3.23 Evidence based approach to audit makes it imperative that various data sources are used to identify audit objectives. When risk

analysis through data analytics become part of the audit process, it is necessary that data is available readily to the audit team. This can be achieved through a systematic data collection and management system in IA&AD, which will ultimately lead to creation of a Data Repository for IA&AD. Such a Data Repository is envisaged both at the central level and at each field office level. A schematic diagram is provided below:

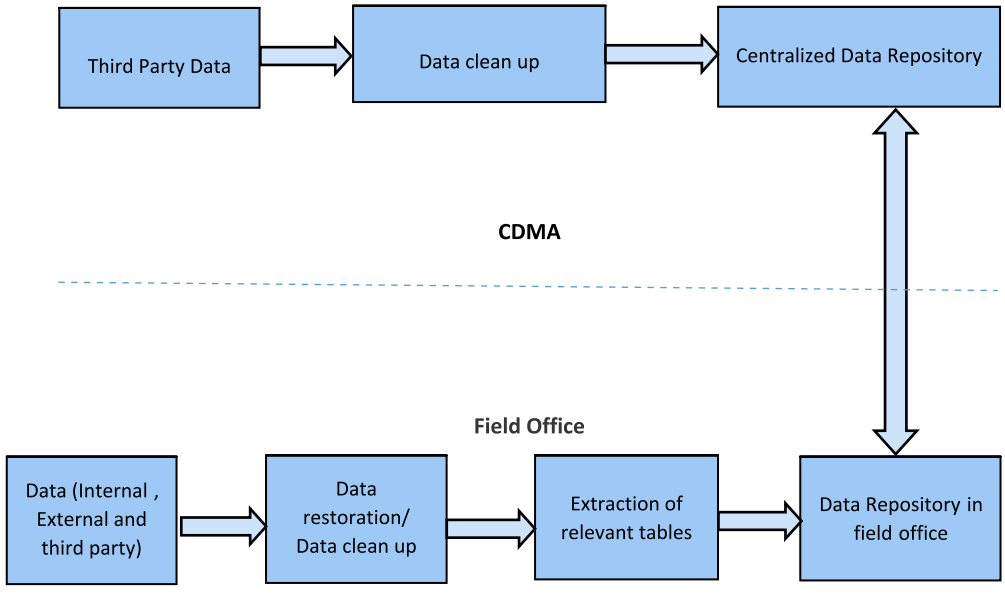

Figure 6 – Data Repository

Data Repository at field offices

3.24 Field audit offices are responsible for identification and collection of data falling within their domain. This will include internal data sources, data from audited entities and third party data, specific to their area of operation. A systematic way of managing multiple

data sources would be for field offices to build a Data Repository at their end.

3.25 Data collection for building a repository is not a one-time effort. Instead, it will be a continuous process over a period of time wherein data will be systematically identified, collected, prepared, organized, managed and stored to form the Data Repository. The following steps are to be followed for creation of Data Repository in each field audit office for managing data within their domain:

- **Data Identification** – The first step in A data management system is identifying data sources available in the environment. All field offices should identify data sources available within their jurisdiction. Data identification is a continuous process. Other than the data from the audited entities, field offices should also try to identify third party data relevant to their domain. The data sets collected from various data sources should be brought into the Data Repository maintained in the field offices.

- **Data mapping** – Once the data sources have been identified, the data should be mapped on a sectoral basis. Though data sources may be primarily designed for application in a particular sector, it can have utility in multiple sectors. A master data of utility of all data sources should be prepared in the following format.

Data Utility Master Table proforma

Name of data source	**Sector to which it is related** **Primary Sector**	**Other sectors where it can be used**
Mining MIS data	Mining	Commercial Tax, Transport
UDISE data	Education	
Census data		Education, Health etc.

As the data collected is used in data analytics, the knowledge gained through it should be used to update the Data Utility Master table, thereby establishing linkages between various data sources.

- **Data preparation -** Before storing the data sets in the repository, they should undergo data preparation stages to optimize the storage in data repository.

- **Data updation -** Field offices should establish a mechanism for getting the data sets periodically. Once the relevant data sets required for data analytic models are identified, data collection in subsequent years would be required for these data sets, unless their structure undergoes modification at data source.

- **Data storage -** While the relevant data sets extracted from various data dumps will go to the data repository, the data dumps collected should be systematically stored in external storage devices.

- **Metadata -** Proper metadata of the data sources, tables etc. needs to be maintained by those managing the data repository. Format for the metadata in the form of three interlinked tables is given below:

List of data sources

Name of data set	Name of data source	Name of audited entity	Sector	From (date/starting year)	Till (date/ latest year data available)	Number of Tables	Data Size	Tag/Key Words[20]

List of tables in data source

Name of data source	Table Name	Description	Number of Columns in table	Number of rows

List of fields in each table

Name of Data Source	Table Name	Field Name	Field type (char, int, etc.)	Field description	Remarks

3.26 Once the data has been prepared and stored in the Data Repository, the data will be stored permanently for future reference. Data Analytic Groups in field offices will be primarily responsible for all the stages mentioned above in developing and maintaining the Data Repository.

Central Data Repository

3.27 CDMA will establish a data repository for data which has applicability across multiple IA&AD offices. CDMA will be collecting such data sets for use in IA&AD. The field offices will be able to access the data sets available in central data repository as per

[20] Key words with which data sets can be identified or referenced while searching

defined access rights and protocol. Selected data analytic models will also be hosted in the Central Data Repository.

Ensuring continuity and availability

3.28 Continuity of data analytic activities in an office should be ensured by adhering to the business continuity management principles enunciated in the Information Systems Security Handbook for Indian Audit & Accounts Department (December 2003)[21].

3.29 Availability[22] of data and data analytic results/ models in an office should be ensured through adequate cataloguing and version control, apart from providing for adequate security.

[21] Para 8, Part II (Domain specific security instructions) of the Information Systems Security Handbook for Indian Audit & Accounts Department (December 2003).

[22] Availability means the characteristic of data, information and information systems being accessible and useable on a timely basis in the required manner, Part I (IT Security Policy) of the Information Systems Security Handbook for Indian Audit & Accounts Department (December 2003).

4. Use of Data Analytics in Audit

4.1 Data analytic results can be used at any stage of audit[23], be it planning, execution or reporting, to derive insights or evidence during the audit process. At the audit planning stage, identification of issues, unit planning and sample design can draw from the data analytic results. At the audit execution stage, data analytic results can identify exceptions, deviations or even describe an existing condition which can be used as audit evidence. At the audit reporting stage, data analytic results drawn at the execution stage can be presented for better appreciation of the audit findings.

4.2 The Auditing Standards, stipulate that auditors shall perform audit procedures that provide sufficient and appropriate audit evidence to support the audit report and that evidence shall be both sufficient (quantity) to persuade a knowledgeable person that the findings are reasonable, and appropriate (quality) – i.e. relevant, valid and reliable (Para 2.5.2.2 (a) of Auditing Standards, 2017). The Auditing Standards also mandate that auditors shall evaluate the audit evidence and draw conclusions. Data analytic results should be used as audit evidence only when they comply with the requirements of the Auditing Standards.

4.3 The specific audit processes where data analytic results can be employed while conducting performance, compliance and financial audits, have been summarised at **Annexure 4.** The degree of use of data analytic results, however, would depend upon the availability of data and the maturity of the field office in using data analytic techniques.

[23] Performance Audit/ Compliance Audit/ Financial Audit

Acquiring data for analysis

4.4 The first step for employing data analytics in audits conducted by IA&AD (Financial, Compliance, Performance audits) is to identify, collect and prepare relevant data for analysis. The auditor should identify all relatable data sets, - internal, external and third party, before finalizing the initial audit plan[24] of individual departments/ entities/ sectors. These datasets from various sources are to be linked and analysed which will result in utilities at various stages of audit.

4.5 It is possible that all relatable data sets are not identified before the start of audit. In such cases, the auditor should remain alert as to fresh data availability. As and when new data sets are identified and such data is accessed, it should be analysed to identify risk areas, areas of interest, exceptions or deviations that should be incorporated into the ongoing audit, as far as practicably possible.

4.6 Statistical tests are conducted with various underlying assumptions. At the same time, the data manifests different characteristics of statistical significance. It is essential to understand the data and assumptions or limitations of each technique/test, in order to derive valid interpretations. Hence, while applying specific statistical tests, the validity of the interpretations should be validated by the Nodal Statistical officer or the Statistical Advisor, in case the auditor wants to use the test results for audit conclusions.

Use of data analytics in audit planning

4.7 The offices in IA&AD adopt a risk based approach to audit planning. Data analytics lends support to the evidence based audit plan and aids in identification of high-risk entities in the audit universe as well as the risk areas with respect to the subject matter of audit.

[24] Financial Attest Auditing Guidelines, Compliance Auditing Guidelines and Performance Auditing Guidelines issued by the C&AG may be referred to for selection of units and sampling approaches in respective audits.

While it does not supplant the existing risk assessment practices in IA&AD, data analytics has the potential to strengthen them considerably. However, the extent of reliance to be placed on the results of data analytics is a matter of judgment of the auditor.

Annual audit planning

4.8 In an office with sufficient data analytic capability, it is envisaged that a data repository of the relevant internal, external and third party datasets and the analytic models will feed into the risk analysis process. Data analytic results can generate a holistic assessment of risks within the audit jurisdiction which should be utilised in preparation of the annual audit plan in addition to other risk assessment parameters. Data analytic models based on financial data (like Business Intelligence model on use of VLC data, PFMS, etc.) or other sector specific models will assist in this task and provide inputs to the process of annual audit planning.

Planning specific audits

4.9 Data analytics can be employed in risk analysis and identification of issues for specifc audits, including setting of audit objectives, drawing an evidence based sample for carrying out substantive audit checks as well as for unit level planning of audits.

Identifying risk areas leading to setting audit objectives

4.10 The Performance Audit Guidelines and Compliance Audit Guidelines refer to the necessity for understanding the entity before starting an audit. In financial audit, too, it is important to identify business processes and systems at the start of the audit. Where transactions are recorded electronically, data analytics facilitates insights from multiple data sets and enables identification of underlying concern areas for any type of audit (performance, compliance, financial). This will help in setting the

broad audit objectives, sub objectives and framing the audit design matrix.

Identifying sample units for substantive checks

4.11 Data analytics enables identification of risk areas within the audited entity, indicating data relationships, significant transactions and outliers thereby providing a more scientific and focused approach for selection of a sample of audit units for substantive checks. Arriving at a composite risk index for various audit units and ranking them on the basis of the weighted average scores of various risk parameters is an objective method of deciding the selection of a sample of audit units and determining the nature, extent and timing of substantive checks.

Unit level planning

4.12 Unit level planning refers to identification of the specific transactions for applying the substantive checks in the selected sample of audit units. Data analytic approaches will focus on identifying the deviations from the specific criteria within the sampled unit. Dynamic dashboards with drill down and filtering capabilities containing data analytic results may be developed for the peripatetic parties.

Use of data analytics in audit execution

4.13 During the audit execution phase, evidence is collected to substantiate the audit assertions identified during audit planning stage. Data analytics can be utilized during audit execution stages in the following ways:

- Peripatetic audit teams should be provided with dashboards/analytical results developed during audit planning stage. These dashboards will help in providing insights about the unit selected for audit apart from insights that led to the

development of audit plan and audit unit selection. The dashboards may also contain the list of transactions, which led to the selection of audit samples. At this stage, issues unique to the unit selected for audit can also be identified and exception reports generated. The insights equip the auditors to appreciate the risk patterns specific to the unit under audit along with its status *vis-à-vis* similar units to identify major deviations. This will help the auditor to focus on the specific sample selection for the unit, if not done already at the audit planning stage.

- Peripatetic audit teams would be able to apply the knowledge of data analytics in the audits, when they access electronic data during audits. Thus, the audit teams can utilise drill down techniques to substantiate their audit assertions. They should also apply data analytics to the subordinate data sets obtained during the audit process (if electronically available) which do not necessarily form part of the corporate or organisation level data.

Use of data analytics in audit reporting

4.14 The audit process involves preparing a report to communicate the results of audit to stakeholders, which should comply with the reporting requirements envisaged in the Auditing Standards.

4.15 A schematic representation of the salient uses of data analytics in audit planning, execution and reporting is given below.

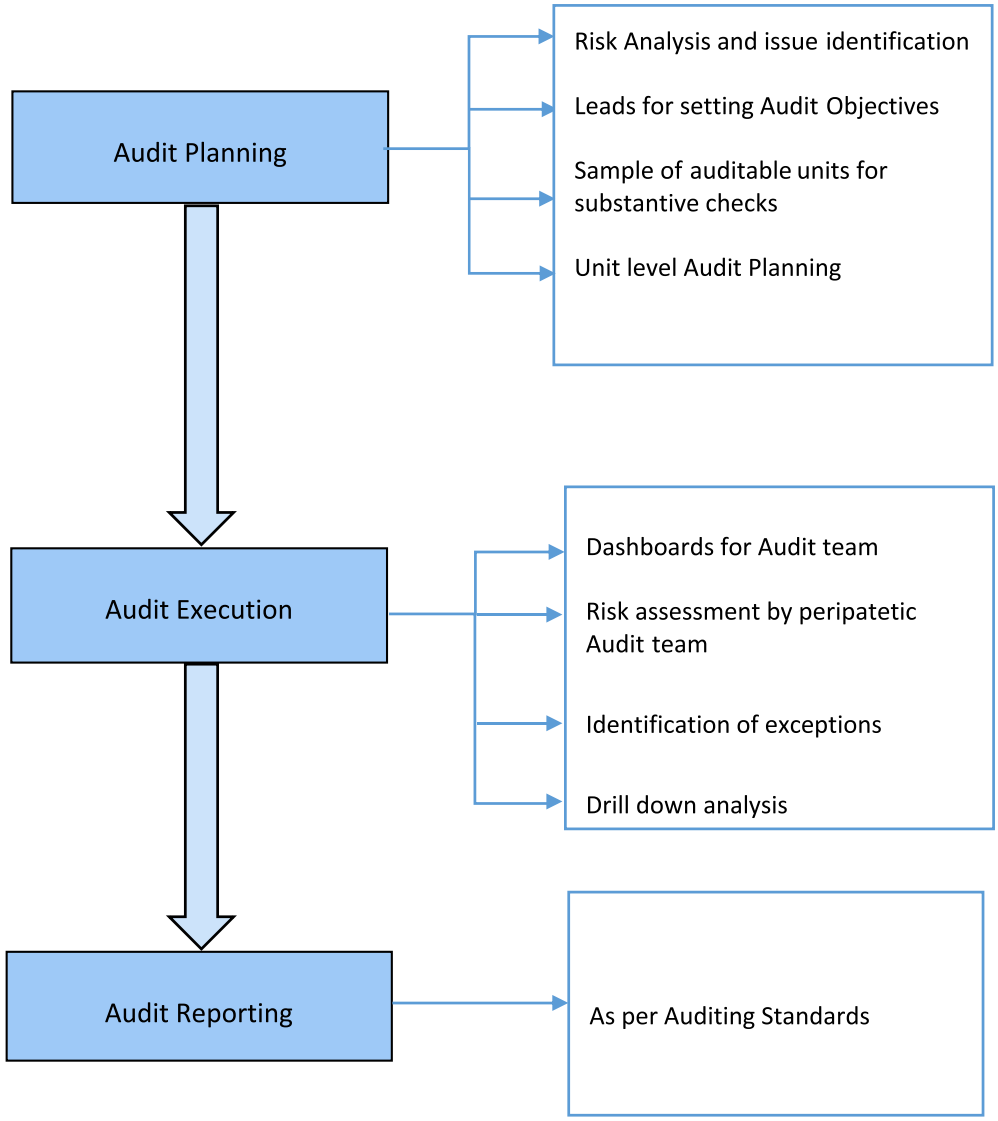

Figure 7- Uses of data analytics in different stages of the audit process

Annexures

Annexure 1 (Ref: para 1.10)

Roles and responsibilities for data analytic work

Roles and responsibilities of CDMA

CDMA will play an advisory and supporting role for the overall use of data analytics in IA&AD. CDMA will facilitate through capacity building, collecting third party data at the central level, identifying new software, assessing applicability of different analytic techniques/analytic models, and disseminating them in IA&AD. CDMA will provide technical support to the field offices in their data analytic efforts wherever necessary.

The Data Analytic models will be vetted and approved by CDMA, in consultation with functional wings in headquarters.

Roles and responsibilities in the field offices

The data analytic project is the responsibility of the functional group within the field office. The data analytic group will offer necessary technical assistance to the project.

Data Analytic Activity	Data Analytic group	Functional groups in the field office
1. Data identification		Primary Responsibility
2. Data Collection	Technical assistance in data collection process	Primary responsibility
3. Data Restoration	Primary Responsibility	
4. Data preparation	Primary Responsibility	To provide domain expertise in preparation
5. Data Analysis and Creation of model	Technical Assistance	Primary Responsibility
6. Model Deployment	Primary Responsibility	

Data Analytic Activity	Data Analytic group	Functional groups in the field office
7. Periodic collection of data for model	Technical Assistance	Primary responsibility
8. Data Repository and management	Primary Responsibility	
9. Documentation	To provide documentation on areas above where Data Analytic Group are having primary responsibility. To compile the documentation.	To provide documentation on areas above where Functional groups are having primary responsibility.
10. Infrastructure for Data Analytics	Identifying the technical requirements	Administration wing will have the primary responsibility.

Annexure 2 (Refer para 2.13)

Indicative Template of Certificate for completeness, consistency and integrity of data

(To be collected from audited entity while receiving data)

The data dump provided to the O/o _____ (name of audit office) in respect of _____ (name of database) for the period _____ to _____ maintained by Ministry/Department/ _____ (Name of entity providing data) on an external storage device/provided online duly marked as <XXXXX> (in case of external device) and signed/authorised by <XXXXX> (name and designation of nodal officer providing the data) on <date>.

It is certified that:-

(i) Officials are authorised by the _____ (name of audited entity) for sharing this data with audit and they understand relevant provision of the Information Technology Act 2008.

(ii) The data dump is full, complete and whole of actual data.

(iii) There is no erasure, tampering or overwriting of original data.

(iv) 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.

(v) There is no damage of data i.e. by destruction, alteration, modification, deletion or re-arrangement of any computer resources by any means.

Summary information on key parameters – total number of transactions, date and details of first and last transactions and hash totals of various numeric data fields is also provided to ensure the completeness, consistency and integrity of data.

(Name, designation, e-mail & signatures of authorised officials)

Date:

Place:

Annexure 3 (Refer para 3.13)

Deriving insights from data analytic results

Statistical analysis has been carried out on data regarding current tax demand and current tax collection for all zones. Correlation of demand and collection data has been worked out for each zone. The template for cataloguing and documenting statistical findings (correlation analysis) and arriving at insights from it is shown below.

Reference Figure/Table: (The visualisation image or the table showing Statistical results).	**Statistical Finding** (Statistical results).	**Insights generated** (Interpretation of the Statistical Finding).	**Focus areas for Audit**	**Whether the insight to be included in a model, if any?**
Table showing zone wise Correlation between Current demand and Current collection.	The figures of 0.533, 0.421, in zones xx, and yy, shows very poor correlation between current demand and current collection.	The collection of tax against current demand is not satisfactory.	Reasons for such poor collection against current demand should be explored.	✓

Annexure 4 (Ref: para 4.3)

Data analytics in various types of Audits

A. Performance Audit

As per chapter 4 of the Performance Audit Guidelines 2014, Planning individual Performance Audits, understanding the entity/ programme is the starting point of any individual Performance audit, which includes review of information in various forms and sources such as electronic databases, management information systems, MIS reports, information from website, etc. The chapter also envisages designing audit approach and methods, use of various analytical techniques, preparing an audit design matrix and deciding the data collection and analysis methods, which includes data analytics. The following steps mention the general methodology to be followed for using Data Analytics in Performance Audits.

- All relatable data sets should be identified at the planning stage for a Performance Audit. The basic information on format, size, mode of access, on the data sets should be collected from the audited entity. Data should be prepared as per need in accordance with the process prescribed for data preparation in Chapter 2 of these guidelines.
- Data Analytics may start on primary database/s of the audited entities. This can be the MIS database, transaction database of auditable entity. Analytics should start with data exploration using various visual and descriptive statistical techniques to classify data, understand geographic/administrative variations, variations over time (trend analysis) etc. Use of GIS maps will help to understand the spatial distribution of various parameters. For example, if it is a Performance Audit on health sector, analysis could be carried out to see
    - Whether health indicators are available

    - Whether health infrastructure analysis are being reported across various states/districts.
    - Variation in health indicators across various regions and over the years
- With this analysis, the auditor can have an understanding as to the nature of the entity/scheme and can identify focus areas of interest.
- Performance Audit primarily seeks to verify whether the scheme/programme is successful in meeting the desired outcomes or is being implemented efficiently. The auditor needs to identify the various government interventions and then see how it is affecting the outcomes. Hence, the next step would be to understand the relation between the outcomes/outputs and the various input factors /interventions of the government. This can be achieved by slicing and dicing of data and using analytic techniques like Scatter Plots, Correlation, regression etc. For example if it is a review on education, analysis can be carried out to identify factor(s), which improve outcomes such as enrolment, dropouts etc.
- At this stage, multiple data sources and data sets can be explored, which can be linked to the primary dataset. For measuring the outcomes, auditor need not rely only on the figures provided by the audited entity. Reliable third party data sets can be used. An increasing number of subject matters have readily accessible datasets (secondary data sets) that allow auditors to critically analyse relevant issues and address some of the compelling questions.
- With these preceding steps, auditor can understand the risks/interest areas in the organization that would have to be addressed in audit. The analysis of the various data sets not only provide a holistic perspective but also provides an evidence based approach for defining the audit objectives / sub –objectives. At the same time, auditor should be aware that there might be risk

parameters, which have not been captured by the existing data sets. Hence, previous understanding of the entity /information from pilot studies etc. should be used to strengthen the evidence based approach.

- Once the audit objectives are defined, next step will be to identify sample units for substantive checks based on the risk perception. Since multiple risk indicators will be identified through data analytics, composite score based on weighted average of various risk parameters can be made and sampling can be done on this score. A weighted risk score can also be assigned for factors which have not been identified through data analytics and incorporated into the composite score.
- Peripatetic teams can now take up unit level planning with the help of the dashboards provided to them. With dynamic drill down and filtering capacity available in modern data analytical tools, the audit teams can identify risks pertinent to each sample unit and can plan audit within the sub unit.
- At the reporting stage, the quality of presenting the audit findings to stakeholders can be improved through various visualisation techniques.

B. Compliance Audit

As per para 2.25 of Compliance Auditing Guidelines, Risk profiling of Auditable entities needs to be done to identify the high-risk areas/ activities of the organisation. The guidelines mentions taking advantage of Big Data and utilizing various data sources like socio-economic surveys, Budget/VLC and other data sources to identify the risk areas. Similarly, as per chapter 4 of the Compliance Auditing Guidelines, planning of individual Compliance Audit unit needs to be done.

In compliance audits, the primary question facing auditor is to identify cases where compliance to a law/rule has not been observed. From audit planning perspective, identification of units for substantive check where such non-compliance is likely to be observed will be crucial. Data Analytical

Models can assist in this task and systematically arrive at cases of such non-compliance.

Data analytics on the relevant data sets using various techniques will help to identify and rank/sort all units on various parameters. These risk elements will vary from sector to sector. The approach will be to identify multiple risk indicators pertaining to the sector. Some examples of risk parameters are:

- Expenditure
- Unusual variation in expenditure over previous years.
- Delay (in case of operations)
- Low tax to income/sales ratio in case of receipt audit

Once the risk parameters have been identified, risk scores can be assigned to each of audit units of the entity for each parameter. Composite scores based on weighted average of multiple parameters can be calculated and the sample units for substantive check can be selected based on this score. The next level of sampling will be of transactions to be selected in each of these units for substantive check. Such high-risk transactions can be identified by:

- Incorporating the rule position whose compliance is being sought through queries/dashboards such that exceptions/non-compliance cases can be identified.
- Visualisation techniques like Scatter Plots, Box Plots etc. to identify patterns, clusters or outliers.

With drill down and filtering capability available in most of the data analytic tools, a dashboard can be built for all units selected for substantive check. By ensuring periodic availability of data, the analysis can be repeated over the years leading to a Data Analytic Model. The analysis done for a Performance Audit can also be converted into a base model for Compliance Audits in the sector through suitable modifications.

C. Financial Audit

The purpose of an audit of financial statements is to enhance the degree of confidence of intended users in the financial statements. This is achieved through the expression of an opinion by the auditor as to whether the financial statements are prepared, in all material respects, in accordance with an applicable financial reporting framework, or – in the case of financial statements prepared in accordance with a fair presentation financial reporting framework – whether the financial statements are presented fairly, in all material respects, or give a true and fair view, in accordance with that framework.

In the audit of financial statements of an organisation or of the Union Government/State Governments, use of descriptive analytics and visualisation can assist the auditor in understanding and gaining insights into the various classes of transactions, account balances, specific grants or disclosures that potentially indicate an unexplained variation or abnormality. Samples for substantive checking can be drawn based on the insights drawn from data analytics.

The routine checks performed by the financial auditors can be automated and built into a model, which can be updated with data pertaining to subsequent years. The trend of financial transactions over years would assist in identifying any abnormal behaviour or pattern. Dashboards so prepared will assist the auditor in carrying out field audits.