A paper on

Big Data Analytics in SAI India

Presented by

SAI India

I Introduction

1. Data analytics, defined as the process of transforming individual data units into actionable information, has been the traditional strength of the Supreme Audit Institutions (SAIs) across the globe. Data analytics, including IT enabled analytics, largely performed on the data generated within the audited entities, has been at the core of audits carried out by the SAIs – compliance audits, performance audits, and financial audits.

2. Audited entities are transitioning into virtualized environment, generating and storing voluminous data and their operations are increasingly getting complex by the day. These apart, large amounts of meaningful and relevant data in disparate forms are being incessantly produced by the external eco systems in the form of surveys, performance statistics published by the Government, industry/domain specific data, comparable organizational data with benchmarking possibilities etc. When collated, they provide the contextual framework regarding the functioning of the audited entities. The rapidly evolving environment of entities makes it incumbent on the SAIs to prepare themselves for auditing in the big data environment.

II Big data and data analytics

3. Big data refers to extremely large, complex data sets that exceed the traditional processing capabilities of the IT infrastructure due to their size, format diversity and speed of generation. It refers to the large and complex data collated from all imaginable sources, which leverages information as the vital asset and includes structured and unstructured data, internal and external data and formal and informal communication. The three dimensions of big data that distinguishes it from conventional data are its features, processes and results.

4. Features:

   a. Volume - quantity, amount of data
   b. Variety – formats, data types, data from various sources
   c. Velocity – speed of data generated and processed
   d. Veracity – quality of data

5. Processes - includes capture, curation, storage, search, sharing and transfer of data

6. Results - visualization and analysis

7. Big data analytics refers to the process of analyzing big data to (i) provide deeper insights, (ii) discover patterns (correlation and causation) and (iii) throw up abnormal behaviour, red flags and outliers that are otherwise hidden, which would enable SAIs to:
    - Understand the audited entity better
    - Analyse, discover, predict and plan audits
    - Plan the nature, extent and timing of audit procedures
    - Establish benchmarks and
    - Assist in drawing conclusions and reporting outcomes

III Data sources for Big data analytics

8. Conventional data analytics, including IT enabled analytics undertaken by SAIs are primarily based on structured data generated within the audited entities. Technological and IT infrastructure limitations have thus far restricted the reach of auditors in many ways. Audit as a process was sustained with sampling techniques and extensive substantive procedures to address the perceived audit risks. The advent of big data and its analytics marks a paradigm shift, which by design envisages synthesizing and integration of relevant data not only from various sources but also in various formats– especially unstructured and textual data to transform data into actionable information and derive value by enhancing the efficiency and effectiveness of audits conducted by SAIs.

9. Given that the SAIs audit a variety of public sector entities, the volume and format diversity of data that is accessible to the SAIs are unparalleled. The data accessible to the SAIs, which can be leveraged for data analytics, can be categorized in three domains:
    - Data created by the SAIs
    - Data of audited entities
    - External data

Data created by the SAIs

10. Data under this category provides a greater flexibility for usage and comprises (a) internal databases of information collected by SAIs about audited entities (b) data of Audit Reports placed in the Parliament/Legislature (c) working papers relating to various audits conducted by the SAIs etc.

Data of audited entities

11. The data under this category is available with the SAIs in the professional capacity as Auditors and its usage involves sensitivities, which would have to be appropriately addressed by establishing and implementing a policy for data storage, access and usage. Such data comprises the (a) financial and non-financial data of audited entities (b) programme specific data including beneficiary databases (c) data pertaining to cases which are under litigation/arbitration/vigilance/sub-judice etc.

External data

12. This comprises data, which are available in the public domain and can further be categorised into:
    - ***Data published by the respective governments and other statutory authorities*** **–** which could include population data, data published by the various Ministries/Departments of Governments, annual budgets etc.
    - ***Other data available in the public domain*** **–** which could include surveys conducted by non-government institutions, information published by NGOs and other not for profit organizations, industry specific information published by various autonomous bodies/institutions, social media and other data available in public domain
13. While the data published by the respective governments and other statutory authorities are reliable, the accuracy of other data available in the public domain cannot be easily validated and the meaning and context of the data are not necessarily self-evident. The policy for data storage, access and usage would have to regulate usage of such data.
14. The volume and variety of data, however, provides immense opportunities for SAIs and at the same time casts responsibilities on the SAIs to be aware of the challenges of auditing in big data environment.

IV. Opportunities and Challenges for big data analytics

Opportunities for SAIs

15. The opportunities for the SAIs germinate from the technology explosion that has taken place in the auditing environment, the transformational impact that big

data analytics could have on SAIs and the consequential aid to governance.

16. **Technology explosion:** With their operations getting increasingly complex and voluminous, public sector entities are transitioning to virtualized environments and processes occurring real time. Proliferation of technology has ensured that capacities for storage, computing and networking are no longer limitations.

17. From the SAIs perspective as external auditors, technology explosion has transformed the limitations that existed in the conventional data analytical tools - their inability to combine data, incapacity to handle unstructured data and inadequacies in analyzing large and complex data into opportunities by enabling availability and accessibility of ubiquitous and cost effective tools, technology platforms and solutions.

18. **Transformational impact for SAIs:** Big data analytics leverages the evidence based approach and its deployment at the audit planning stage for a macro level analysis of almost the entire range of data, rather than on a small representative sample is in itself a paradigm shift from conventional analytical procedures. It enables moving away from the computer assisted audits to a more robust digital auditing to enhance the effectiveness of risk assessment by discovery of red flags, outliers, abnormal behavior, frauds, abuse and deeper insights. Thus, it underpins efficiency in determination of the nature, extent (including the extent of sampling) and timing of audit procedures, ultimately leading to providing a greater level of assurance in audits.

19. Further, it ushers in new age competencies of predictive analysis, advanced statistics and software usage for transformation of data into actionable information. Data visualization and big data analytics are the value added exploratory functions facilitated by big data, which enables discovery of relationships between variables and broader trends of risk and deepens the reach of auditors.

20. **Aid to governance:** It enables the SAIs to assume a proactive role in aiding governance by sharing insights with those charged with the governance of public sector entities to promote transparency, effective oversight and control.

SAI Challenges

21. To leverage the potential of big data analytics, the challenges that need to be addressed by the SAIs are to:

- **manage the people** - This involves transforming the mind set of auditors to adapt to digital auditing and auditing in the big data environment and capacity building
- **manage the data**- This involves defining the scope of data and addressing data sensitivities associated with access and usage of various sources of data such as third party databases, beneficiary databases, data pertaining to audited entities obtained by SAIs as a part of understanding the entities as well as data collected as evidence, addressing the veracity of data involving an assessment of strengths and weaknesses of various sources, privacy issues and compliance with legislative and regulatory requirements. This would also involve establishing data management protocols address the aspects of data quality, integrity and other processes for data analytics.
- **manage the infrastructure** – This involves augmentation of appropriate IT infrastructure and establishing the technical solutions.

IV Initiatives taken by SAI India

Fundamental research and Concept Paper on big data

22. SAI India started its efforts towards big data analytics by commissioning a fundamental research on big data by the Professional Practices Group of the SAI. A concept paper prepared after the research described the concept of big data and the opportunities and challenges pertaining to its use in SAI.

Experiments with visual analytical tools

23. The SAI also decided to explore the potential of visual analytical tools. Accordingly, limited versions of QlikView and Tableau visual analytical tools were purchased and few officers were trained on these software. The trained officers tested these software in three different audit engagements – a financial audit, a performance audit and a compliance audit involving forensic analysis.

24. The results from these initial experiments with analytical tools were presented at a workshop to the senior management and the encouraging results of these pilot experiments, motivated further work on big data.

Discussions with other SAIs

25. Meanwhile, during our bilateral programs with other SAIs e.g. SAI China and SAI Poland, mutual experiences on potential of big data analytics were shared and discussed.

Big data management policy

26. Based on further deliberations and brainstorming a Big Data Management Policy (BDMP) was adopted in February 2016 that sets out the broad contours of the framework for SAI India. Apart from building a common institutional understanding of concepts like big data, visualization etc., this policy covered important issues like:
    - **Identification of data sources:** The policy categorizes data sources in two parts i.e. internal which is created/maintained by SAI India and external which is either available with audited entities or is in public domain
    - **Establishing data management protocol:** The policy sets up a protocol to ensure that data being used has the essential characteristics of authenticity, integrity, relevance, usability and security. It also address the data access arrangements with external sources , data sensitivities , criteria for assessing veracity of data , privacy and confidentiality issues covering procedures for aggregation and anonymization as well as compliance with legislative and regulatory requirements.
    - **Digital auditing, data analytics and visualization strategy:** The policy calls for establishing a nodal authority which will develop guidelines and strategy for digital auditing, data analytics and visualization.
    - **Infrastructure, capacity building and change management**: The policy sets roles and responsibilities for development of infrastructure and capacity building within SAI India to mainstream data analytics.
    - **Monitoring:** A monitoring group for overseeing the implementation of this policy framework is envisaged in the policy.
27. As an important step, the policy identifies, the roles and responsibilities within the organisation for mainstreaming big data analytics. The policy envisages the establishment of a nodal authority in the department for big data analytics.

Task Force on Implementation of BDMP

28. A Task Force on Implementation of BDMP, has been set up by the SAI India, The Task Force has been entrusted with the:
    - Identification and listing of relevant databases
    - Identification of data analytic tools to be adopted

- Identification of data visualisation tools to be adopted
- Identification and capacity building of selected officials for working with the nodal authority
- Laying down of Standard Operating Procedure for the use of big data

29. The Task Force will shortly be laying down its recommendations on the above tasks.

V Pilot work on big data

30. Simultaneously, SAI India is also taking baby steps in some ongoing audits to test the waters with respect to big data analytics. Two such experiments, though still in nascent stage, are as follows:

Performance Audit of Social Security Scheme

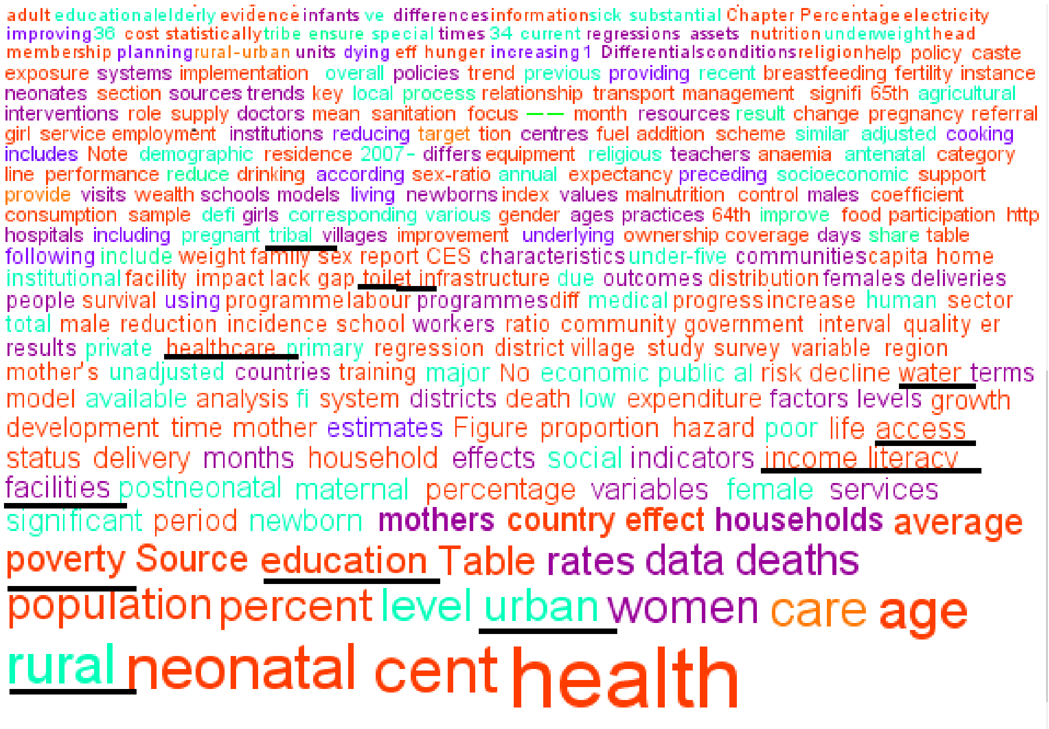
31. The Government of India's social security programmes aim to protect citizens from unforeseen contingencies and risks that cannot be handled individually. The programme is supposed to provide assistance to specifically targeted vulnerable populations including widows, individuals with disabilities, and the elderly. These programmes are often implemented by local self-governing institutions (LSGI) at the provincial level. A performance audit of three selected Social Security assistance programmes viz. Old Age Pension, Widow Pension and Disabled Pension schemes in one of the states was taken up. While the audit followed the regular framework of performance audit, an effort was made to increase the use of data and research evidence in the design and implementation of the performance audit.

32. Unlike the conventional approach, both primary data (from program documentation) and secondary data (from datasets available in public domain e.g. Census data, database of people below poverty line, expenditure database etc.) were used for analysis. The conventional approach would have only enabled detection of wrong inclusion of ineligible persons, through analysis and substantive tests of database of beneficiaries. The use of the external databases enabled audit to verify whether eligible beneficiaries got excluded (incorrect exclusion). This also facilitated survey of not only beneficiaries but also eligible people who got excluded from the schemes.

Analysis of Infant Mortality Rate

33. In this experiment, an important parameter to evaluate the improvement in health of children i.e. Infant Mortality Rate (IMR) was assessed using big data analytics tools. Conventionally, an audit of schemes targeted at IMR reduction would limit its analysis to implementation of the scheme. However, it would have been challenging for audit to comment on whether all significant factors affecting IMR were factored in correctly.

34. As a first step, a text analysis of the available literature on IMR (in electronic format) was done resulting in a word cloud. Around 30 documents which included research papers published in various journals, government reports , of about 1000 pages, with about a million words and about 1000 references in their respective bibliography were put for text analysis using a tool called KNIME and a word cloud was obtained.


adult educationalelderly evidence infants ve differencesinformationsick substantial Chapter Percentage electricity
improving36 cost statisticallytribe ensure special times 34 current regressions assets nutrition underweighthead
membership planningrural-urban units dying eff hunger increasing 1 Differentialsconditionsreligionhelp policy caste
exposure systems implementation overall policies trend previous providing recent breastfeeding fertility instance
neonates section sources trends key local process relationship transport management signifi 65th agricultural
interventions role supply doctors mean sanitation focus — — month resources result change pregnancy referral
girl service employment institutions reducing target tion centres fuel addition scheme similar adjusted cooking
includes Note demographic residence 2007- differs equipment religious teachers anaemia antenatal category
line performance reduce drinking according sex-ratio annual expectancy preceding socioeconomic support
provide visits wealth schools models living newborns index values malnutrition control males coefficient
consumption sample defi girls corresponding various gender ages practices 64th improve food participation http
hospitals including pregnant tribal villages improvement underlying ownership coverage days share table
following include weight family sex report CES characteristics under-five communitiescapita home
institutional facility impact lack gap toilet infrastructure due outcomes distribution females deliveries
people survival using programme labour programmesdiff medical progress increase human sector
total male reduction incidence school workers ratio community government interval quality er
results private healthcare primary regression district village study survey variable region
mother's unadjusted countries training major No economic public al risk decline water terms
model available analysis fi system districts death low expenditure factors levels growth
development time mother estimates Figure proportion hazard poor life access
status delivery months household effects social indicators income literacy
facilities postneonatal maternal percentage variables female services
significant period newborn mothers country effect households average
poverty Source education Table rates data deaths
population percent level urban women care age
rural neonatal cent health

35. Picking cues from the word cloud thus arrived; a list of various parameters possibly influencing IMR was drawn. (See the underlined words in the picture above). The words like availability of safe drinking **water**, **literacy**, access to **healthcare** facilities, **tribal** population (percentage of tribal/vulnerable sections of society), **toilet** (access to clean sanitation facility) etc. are predominant and throw intuitive ideas.

36. As an experimental study, the district level data pertaining to these variables from a sample was collected. The data sources included Census data and other State reports. As authoritative data was unavailable for many factors identified, the analysis was restricted to immediately available data.

37. Linear Correlation analysis between IMR and availability of good quality water showed almost a perfect correlation. Similar correlation was also found with other variables like percent of rural population, percentage of vulnerable population, percent of people resorting to open defecation, literacy rate, distance to Primary Health Centre etc. Yet, the analysis also threw some surprising insights, even going against conventional wisdom. For instance, conventionally, it is understood that building the medical infrastructure is extremely important to deal with social health issues. However, in this analysis, the distance to Primary Health Care Centers had a lesser correlation compared to other preventive variables.

38. Taking the analysis further, a multivariate regression analysis (using KNIME) was done on 30 of the 33 districts of the state, selected by means of random sampling in order to predict the figures for the remaining three districts where, IMR is the dependent variable and the other four variables were independent variables. The model was run a number of times on different samples of these districts.

39. It was observed that the prediction of figures for IMR came much closer to the actual figures most of the times. This strengthened the insight gained from the word cloud that these factors may have an influential role with respect to IMR.

40. With these insights, Audit would now be able to concentrate on public spending in these variables like water quality, prevention of open defecation etc. to assess the effectiveness of the government efforts to reduce the incidents of infant mortality.

VI Way forward

41. At SAI India’s level, the next steps would be to set up a nodal authority on big data. The other important task would be to follow up the recommendations of Task Force on implementation of big data. Separately, some more pilot projects could be taken up on big data analytics.

42. At the INTOSAI level, a recent survey commissioned by SAI India, as Chair of Knowledge Sharing Committee (KSC), revealed tremendous interest in the INTOSAI community in the field of big data as a cross cutting theme for research. We may consider taking up a research Project from KSC on the subject in near future.

VII Conclusion

43. Big data analytics helps in faster number/information crunching, understanding data linkages, and in providing new, deeper insights. This will in turn help enhanced risk assessment for better audit planning, analyzing possibly the entire population, discerning hidden linkages for arriving at more insightful audit findings and for presenting findings in a reader-friendly manner through visualization tools. Above all data analytics will help the SAI to focus audit resources on high risk areas.

44. However, to reap the benefits from big data, SAIs will have to address issues of data quality and its confidentiality, creation of infrastructure (hardware and software tools) for analysis and training manpower on big data. This would possibly require an overarching policy on big data management to be in place in the SAI.

45. More importantly, the extent of SAIs success in tapping the potential of big data depends upon the buy-in and support of top management. The progress achieved so far in SAI India on big data is largely due to the firm commitment and support of the top management.