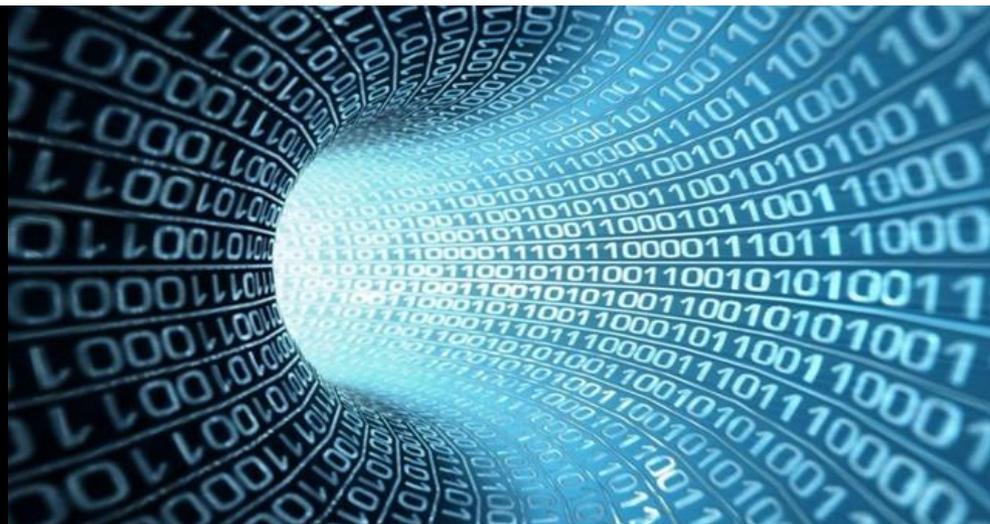




BIG DATA ANALYTICS IN SAI INDIA

SAI India





Overview

- Introduction
- SAI India initiatives
- Pilot work on big data
- Future Plans
- Conclusion



Introduction



- Winds of Change
 - transition to virtualized environment
 - generation of voluminous data
 - increasingly complex operations of entities
 - large amounts of meaningful and relevant data produced by external sources
- SAs need to prepare for audit in big data environment



Introduction - Big Data

- Extremely large, complex data sets that exceed traditional processing capabilities of IT infrastructure due to their
 - size
 - format diversity and
 - speed of generation



Introduction - Data Analytics

- Process of analyzing big data to provide
 - deeper insights
 - discover patterns (correlation and causation) and
 - throw up abnormal behaviour, red flags and outliers that are otherwise hidden



Introduction - Opportunities and Challenges

- **Opportunities**

- Technology explosion
- Transformational impact for SAI
- Aid to governance

- **Challenges**

- manage people
- manage data
- manage infrastructure



SAI India Initiatives



- Fundamental Research on Big Data
- Experiments with visual analytic tools
- National Workshop on big data
- Discussions at international bilateral forums
- Development of Big Data Management Policy (BDMP)
- Taskforce on implementation of BDMP



SAI India Initiatives

- Fundamental Research on Big Data
 - Research by Professional Practices Group
 - Concept paper on big data
- Experiments with visual analytic tools
 - QlikView and Tableau tools purchased
 - Selected officers trained
 - Tools tried in three audit engagements
 - National Workshop on pilot experiments
- Discussions at international bilateral forums



Big Data Management Policy

- Sets broad contours of framework for SAI India
- Policy covers:
 - Identification of data sources
 - Establishing data management protocol
 - Digital Auditing, Data analytics and Visualization strategy
 - Infrastructure, capacity building and change management
 - Monitoring



Big Data Management Policy

- Sets out broad contours of framework which SAI India intends to institutionalize
- Policy covers:
 - Identification of data sources
 - Establishing data management protocol
 - Digital Auditing, Data analytics and Visualization strategy
 - Infrastructure, capacity building and change management
 - Monitoring



Task Force on Implementation of BDMP

- Task Force entrusted with:
 - Identification and listing of relevant databases
 - Identification of data analytic tools
 - Identification of data visualisation tools
 - Identification and capacity building of selected officials for working with nodal authority
 - Laying down Standard Operating Procedure for use of big data
- Report expected shortly



Pilot work on big data

- Case 1: Performance Audit of Social Security Schemes
- Case 2 : Analysis of Infant Mortality Rate



PA of Social Security Schemes



- PA of three selected Social Security assistance programmes in a State
- Data used
 - Primary data (from audited agency) and
 - Secondary data (Census data, database of people below poverty line, expenditure database etc.)



PA of Social Security Schemes



- Conventional approach only focuses on wrong inclusion of ineligible persons
- Could verify whether eligible beneficiaries got excluded (incorrect exclusion)
- Survey of excluded people also



Infant Mortality Rate (IMR)

- Big data analytics tools used to identify most significant factors affecting IMR
- Started with text analysis of 30 documents of about million words using KNIME
- World Cloud built



Infant Mortality Rate (IMR)

adult education elderly evidence infants ve differences information sick substantial Chapter Percentage electricity improving 36 cost statistically tribe ensure special times 34 current regressions assets nutrition underweight head membership planning rural-urban units dying eff hunger increasing 1 Differentials conditions religion help 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 communities capita home institutional facility impact lack gap toilet infrastructure due outcomes distribution females deliveries people survival using programme labour programmes diff 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



Infant Mortality Rate (IMR)

- Linear Correlation analysis between IMR and
 - availability of good water quality
 - rural population
 - vulnerable population
 - open defecation
 - literacy rate
 - distance to Primary Health Centre etc.
- Almost perfect correlation between good quality water and IMR



Infant Mortality Rate (IMR)

- Multivariate regression in 30 of 33 districts in order to predict IMR figures for remaining 3 districts
- Four independent variables (good quality water, open defecation, rural population percent and literacy rate) analysed



Infant Mortality Rate (IMR)

District	Actual IMR	Predicted IMR
Baran	62	60.2
Nagaur	59	61.15
Sikar	56	56.65

- Close match between actual IMR and predicted IMR
- Strengthened insights gained from Word Cloud



Way forward



- SAI India
 - set up Nodal Authority on Big Data
 - consider recommendations of the Task Force and take appropriate follow-up action
 - more pilot projects on big data analytics



Way forward

- INTOSAI
 - KSC survey on cross-cutting research showed tremendous interest in big data
 - Future KSC cross-cutting research project on big data



Conclusion

- Big data offers several opportunities
- SAI needs to address issues of
 - Data quality
 - Training of manpower
 - Creation of infrastructure
- More importantly top management buy-in and support critical



Thank You