INTOSAI
Working Group on Big Data

RESEARCH PAPER ON INNOVATIVE AUDIT TECHNOLOGY

September
2022

FOREWORD

FOREWORD

The Fourth Industrial Revolution brings cutting-edge technologies that help governments achieve Sustainable Development Goals and increase data accessibility and openness. Innovative technologies, such as big data, artificial intelligence, data analytics, and machine learning, profoundly changed the mindset and ways of work for public sector auditing. In view of this, the INTOSAI Working Group on Big Data initiated a research project to investigate the innovative audit technologies and explore ways of utilizing those technologies in practice, aiming to help supreme audit institutions meet increasing expectations from stakeholders on more effective auditing and higher-quality audit results.

After three years of research by a team within the Working Group, we are now pleased to present the Research Paper on Innovative Audit Technology. The paper identifies big data and artificial intelligence as innovative technologies in auditing, surveys SAIs capacity in using the big data technology, and puts forward suggestions for SAIs to develop a big data governance framework and provide decision-makers with audit-based advice on strategic issues.

We would like to extend our gratitude to all the SAIs that contributed to this research. Special thanks go to SAI Indonesia for their tremendous efforts in leading and organizing the research. We also appreciate the SAIs of Austria, Brazil, Denmark, Ecuador, Norway, Russia, and the UK for their valuable input to the research.

We hope you find this paper useful.

侯凯

Hou Kai

Auditor General of the National Audit Office of China

Chair of INTOSAI Working Group on Big Data

ABOUT INTOSAI WORKING GROUP ON BIG DATA

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

Chair

China

Vice Chair

United States

Members

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

Observers

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

CONTENTS

1. Executive Summary

The purpose of this research paper is to find practical use of innovative technology in government auditing towards advanced technologies. Industrial Revolution 4.0 enables not only individuals but also institutions to take benefit of large amounts of data so that they can interact among others in a very efficient mean. One prominent technology in government auditing is Big Data. Many SAIs have started to use several technology platforms for dealing with Big Data. One of the major directions concluded in The Moscow Declaration is about how SAI uses advanced technology to enhance the process of public audit more efficiently.

In large part, the confluence of the 4th Industrial Revolution and the Moscow Declaration is a testament to the efforts of INTOSAI to remain at the forefront and to enhance stakeholder engagement. Indeed, utilizing advanced audit technology is itself another of the goals of the Moscow Declaration. While the advanced audit technology can be applied to SDGs and certainly promotes the availability and openness of data, we are currently seeing the advanced audit technology as a means to free auditors from data collection tasks to a more value-added effective, transparent, and informative input on accountability, seeking to provide decision-makers with audit-based advice on important and strategic issues.

Big Data and Artificial Intelligence are the emerging technologies that give an impact on how auditors in the future will interact and deal with the data and the algorithm. Between these technologies, there are practices that can be applied to any kind of data which auditors should deal with, i.e., data analytics and machine learning. These four pinpoints represent the journey of auditors to play around with the innovative audit technology.

This research found that most SAIs work on historical data which was collected on a periodical basis. Some SAIs obtain the data only during audit work while other SAIs have the advantage to get the data periodically regardless in an audit period or not. Also, most SAIs still deals with traditional data and conventional analytics process. The use of spreadsheet software and

Generalized Audit Software is still prominent in analyzing the data. Thus, auditors mostly use Computer-Assisted Audit Techniques (CAATs) during the audit work.

Due to the increasing need for implementing and using big data for enhancing the audit process and empowering the auditor, the development of the Big Data Governance Framework for SAI has become a necessity. The framework will guide the SAI for developing the Big Data Governance Framework. The framework will provide the SAI with the policies and standards regarding data management, data security, and data processing optimization.

2. Introduction

2.1 Background

The Moscow Declaration of INTOSAI, like so many other INTOSAI documents, is addressed and dedicated to SAIs for increasing their institutional capability. The declaration encourages SAIs to take the benefit of advanced technology. The Moscow Declaration is based on the main conclusions of INCOSAI XXIII. It outlines the major directions for the future development of public audit:

1. Ensuring independent external monitoring of the achievement of goals agreed upon on the national level as well as SDGs;
2. Efficiently using opportunities offered by technological progress;
3. Strengthening the role of SAIs' work with a view to improving transparency and accountability of public administration.

In line with the Moscow Declaration, Supreme Audit Institutions agreed to follow the ten basic principles:

1. Promoting more efficient, transparent, and informative accountability of the government for its outcomes;
2. Developing a comprehensive approach to audit for attaining national goals and priorities;
3. Preparing recommendations for the government on important strategic issues of public administration;
4. Promoting the principle of accessibility and openness of data, the openness of source code and algorithms;
5. Employing data analytics in audits;

6. Advancing the culture of experimental thinking;

7. Managing systemic risks in public administration;

8. Training auditors of the future;

9. Searching for ways to promote inclusiveness;

10. Establishing efficient collaboration with auditees and broadening the cooperation with the academia and the society as a whole.

In large part, the confluence of the 4th Industrial Revolution and the Moscow Declaration is a testament to the efforts of INTOSAI to remain at the forefront and to enhance stakeholder engagement. Indeed, utilizing advanced audit technology is itself another of the goals of the Moscow Declaration. While the advanced audit technology can be applied to SDGs and certainly promotes the availability and openness of data, we are currently seeing the advanced audit technology as a means to free auditors from data collection tasks to a more value-added effective, transparent, and informative input on accountability, seeking to provide decision-makers with audit-based advice on important and strategic issues.

In shifting this focus of auditing, the advanced audit technology is encouraging auditors to adopt a creative mindset stance to look at numbers, but to gain insight towards providing foresight. It is all of these elements of the Moscow Declaration, and perhaps more, that are either central or tangential to the new frontier of audit technology.

This research project was conducted in the context of the use of advanced audit technology. Although we found that SAIs do not have any special mechanisms for implementing advanced audit technology but rather follow general procedures, we highlight some specific features that affect the implementation of advanced technology on government auditing.

2.2 Project Scope and Methodology

This paper focuses on identifying emerging technology that can increase institutions' capacity. The project studied the purpose for which, and the manner in which, SAIs utilize the advanced audit technology for their audit activities in order to achieve a more efficient and effective audit process while fulfilling their maturity from Oversight to Insight, and furthermore to achieve the capacity for providing Foresight.

3. Big Data as an Innovative Technology in Auditing

3.1 Big Data

Big Data is the common term nowadays to describe a collection of huge amounts of data so that the use of traditional software is not relevant anymore (Warren et al., 2015). The growth of data follows Moore's law which means that the data grows exponentially. The growth of data is not only in terms of size that requires higher storage but also in terms of its variety and velocity. These attributes of Big Data are commonly referred to as the 3Vs, i.e., Volume, Variety, and Velocity (Doug Laney, 2001).

Volume in Big Data addresses the gigantic size of the data in which the size is approaching the Zettabyte. IP-based data traffic has increased significantly in the last five years. The presence of social media and the Internet of Things has contributed to the increasing size of data. Also, the information system owned by institutions has become the main source of information for data producers (Herschel & Miori, 2017).

Variety in Big Data refers to various data types. Until 2000, most institutions were dealing with structured data, data that consist of rows and columns in the tabular form stored in a database either a common database or relational database system. The variety in Big Data can be classified into two perspectives; the structure and the human-machine interaction. From the data structure perspective, Big Data consists of both Structured Data and Unstructured Data. While structured data refers to tabular data, unstructured data refers to the text of a speech, text in a document, images, voice, and video (Boyd & Crawford, 2012; Mayer-Schonberger & Cukier, 2013). From the human and machine interaction perspective, not only do humans produce the data but also machines can generate the data. For instance, location sensors from the GPS of millions of smartphones generate data in a real-time process, continuous, and granular. Humans, on another side, generate many unstructured data through social media and machine-to-machine connectivity.

Velocity in Big Data refers to the speed of data in and out. Data runs very fast. There have been several researchers measuring how fast the data pass over and over again on the Internet. The following picture illustrates the data traffic in a minute around the globe.

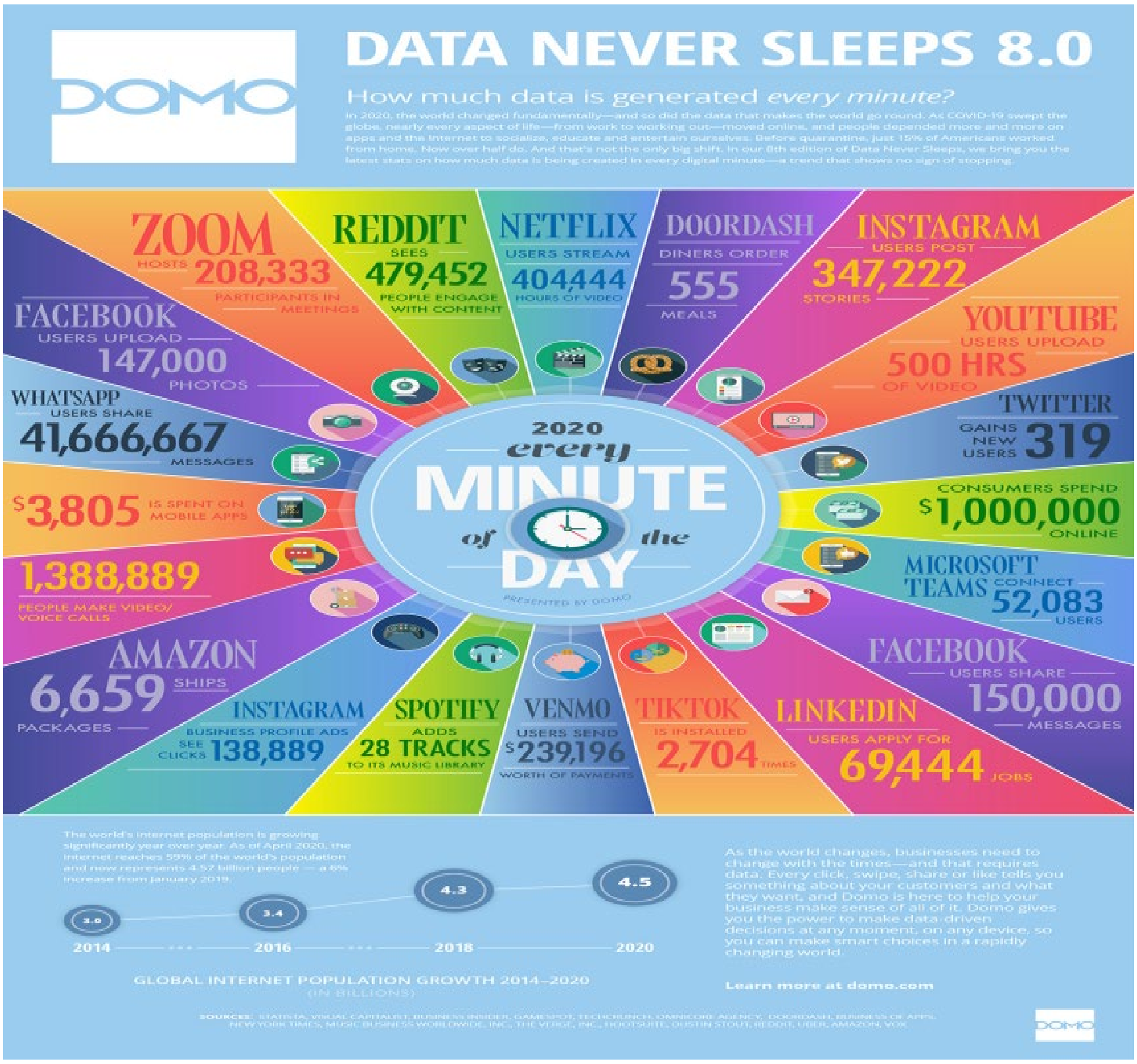


Figure 1. Data Traffic in a minute 2020[1]

The concept of 3Vs of Big Data has evolved to 4Vs and 5Vs. Veracity is the first additional of 3Vs. Veracity in Big Data highlights the inclusiveness of noise in the data, thus, validating the data for further analysis is essential. The next addition is Value. Value as an attribute of Big Data indicates the usefulness of gathered data. Data by itself, regardless of its volume, cannot tell the story, thus, useless. In order to be valuable, the data needs to be converted into insights or valuable information.

In auditing, Big Data technology enables SAI to expand the horizon of the data collection process while keeping such process is cost-effective. The data collection is, to some extent, a part of

1. https://www.visualcapitalist.com/every-minute-internet-2020/

collecting audit evidence. The audit standard states that there are three characteristics of audit evidence, i.e., sufficient, relevance, and reliability. Auditors can use Big Data to expand the scope of their data collection and draw patterns or comparisons over larger populations of data. Also, Big Data provides auditors with more than sufficient data to be converted into audit evidence.

Standard said that the auditor should have access to the audit entity’s information during the audit assignment as a means to collect audit evidence. However, the quality of audit evidence strongly depends on the audit entity’s information system and technology, the cost and benefit aspects, and interactions with the audit entity’s personnel. If such a worse case occurred, auditors may use alternative information from relevant external sources available in a big data platform. In this case, big data can support auditors in providing additional relevant information from other competent resources that are not available to the auditor. The use of external resources becomes essential for revealing the indication of fraud in a financial statement.

Obtaining audit evidence related to a fraud case is a challenge for auditors. The audit standard mentions that the auditor has a responsibility to plan and perform the audit to obtain reasonable assurance about whether the financial statements are free of material misstatement, whether caused by error or fraud (AICPA, 2002). As a consequence, the auditor should find either direct or indirect evidence that could reveal a person's motivation and rationalization for committing fraud. Motivation and rationalization of a person to commit fraud are usually reflected in their lifestyle, behavior, and morals. These aspects are usually not observed during a financial statement audit. To overcome this, auditors can collect data from social media to explore a person's lifestyle, motivation, and even rationalization, for example, his dissatisfaction with the institution where he/she works. Therefore, Big Data helps the auditor to minimize the risk of insufficient audit evidence to uncover irregularities.

The challenge in applying Big Data as complimentary audit evidence is reliability. On one side, Big Data is highly reliable since it is obtained from external sources and collected by auditors directly. Big Data is very difficult to be altered because of its vast volume. On another side, however, due to its characteristics, Big Data is not free from noise that can lead to false positives and data overload, thus, decreasing the reliability of the data (Yoon et al., 2015). To eliminate the noise, the auditor should apply a data preparation procedure to the content of Big Data before proceeding to further analysis (INTOSAI, 2019).

As for the relevance of audit evidence, the use of Big Data depends heavily on the professional judgment of auditors which is common when auditors deal with traditional data. Besides, using Big Data as audit evidence drives the concern about the verification of the accuracy of data as most of the data are from external sources and especially if the Big Data has a significant effect on financial statements (Appelbaum et al., 2017). Big Data evidence is considered sufficient due to the volume and variety of the data available in real-time. Appropriateness refers to the reliability and relevant evidence (Yoon et al., 2015).

Big Data is considered an evolutionary development in data technology. Big Data itself has no value, therefore, an evolutionary approach in analyzing Big Data is essential. The application of advanced analytics is important to extract potential value to help not only auditors but also decision-makers make sound and informed decisions (Alles & Gray, 2016). The analytics techniques mentioned above are called Big Data Analytics (BDA). BDA has plenty of definitions however in auditing it is defined as "the science and art of discovering and analyzing patterns, identifying anomalies, and extracting other useful information in data underlying or related to the subject matter of an audit through analysis, modeling, and visualization for the purpose of planning or performing the audit" (AICPA, 2017).

3.2 Artificial Intelligence

Initially, Artificial Intelligence (AI) was inspired by logic such as if something happened, then do that. AI is a computation program that enables a machine to work and think like human intelligence for making a decision, solving problems, and predicting the future (Elliot et al., 2020). Since AI enables the machine to think like a human and humans apply AI to extend the capacity of processing the information, AI is considered as the external intelligence (Arthur, 2017). Human creates AI by injecting several algorithms and loading the massive historical dataset to the machine. AI is an example of a human and machine interaction routine that is easy for humans but very difficult for a machine to do so.

AI was built on top of machine learning technology. Machine learning is the process of training the computer to think like a human by applying an appropriate algorithm and exploring a sufficient dataset (Goldberg & Holland, 1988). An algorithm is a set of rules and instructions that establish a formula to produce an output (Knuth, 1997).

Considering Big Data as a complement of audit evidence, auditors need a technology platform to enable them to identify, excavate, and assess the information contained in the Big Data, thus, significantly improving the audit quality. AI can be regarded as a technology platform that adds value to the area of improving audit quality. The ability of AI technology to efficiently deal with vast amounts of data has not only increased the value of a vast amount of data collected during the audit but also reduce the cost and time for processing such big data (Bizarro & Dorian, 2017).

3.3 Technology Platform

A big data technology platform is designed to handle the collection, processing, and analysis of data in which such processes are not relevant for traditional database systems due to its size, its variety, and its complexity. The end-to-end process from collecting to delivering the result of analysis typically involves one or more of the following types of workload:

1. Collecting the data in the form of either batch process or real-time process.
2. Big data exploration
3. Advanced Analytics of the data such as conducting predictive analytics and machine learning.

The following diagram shows the logical components that fit into the end-to-end big data technology platform. Every item in this diagram does not represent individual solutions. Some items may be combined together.

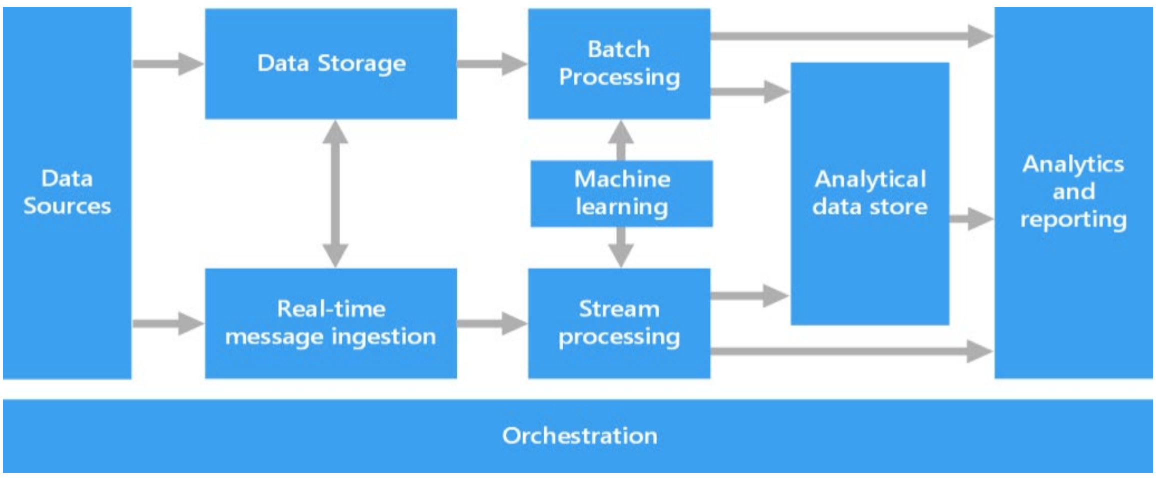


Figure 2. Components of a Big Data Architecture[2]

2. https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/

The following picture exposes the landscape of an end-to-end Big Data Technology Platform for software either an open-source or proprietary solution.

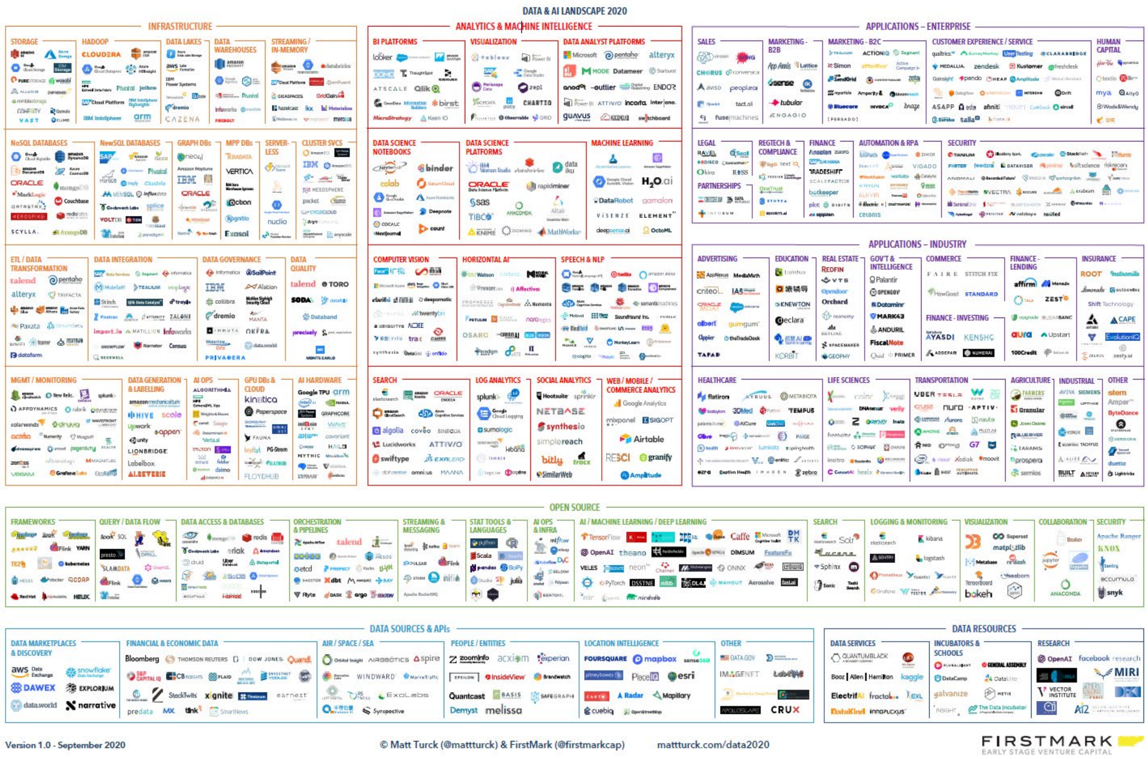


Figure 3. Big Data Landscape 2020[3]

As shown in the above figure, there are so many choices of technology for implementing Big Data including the capacity for analyzing these Big Data. A block in Figure 2 may use a single or a combination of the software shown in Figure 3.

Data Source Block in Figure 2 represents one or more origins of the data. This may include application-generated data such as data from the auditee's information system. The data source may also come from a static page such as a website. Furthermore, it may be sent by Internet-of-Thing (IoT) devices in a real-time process.

Data Storage Block provides the institutions with a single store for all of the raw data that anyone in an organization might need to analyze. In the context of Big Data, this scheme of storage refers to a Data Lake. The Data Lake, to some extent, is similar to what was previously known as

3. *https://mattturck.com/data2020/*

Data Warehouse. However, there is a stimulating distinction between the data lake and the data warehouse. The data lake stores raw data as it is and in whatever type the data source provides. It's up to the consumers of that data to make sense of that data for their own purposes. Data Warehouse, on another side, stores modified raw data as the data follow the pre-defined schema of the Data Warehouse (DataLake, n.d.). Also, there is a shifting of the data pipeline approach from ETL in Data Warehouse to ELT in Data Lake practices. In a data warehouse, the data source should be cleaned through the sequential procedure of Extract, Transform, and Load (ETL) from the data source to the data warehouse, however, the process is altered in which ETL is modified to ELT in a data lake. In the data lake, the data source is extracted and loaded directly to the data lake. The transformation process has proceeded whenever the data is needed for further analytics by the user and, in some cases, the transformed data will be stored in the data warehouse for repetitive analysis.

There has been a shift in the approaches on collecting the data from Valid-First to Collect-First. The Valid-First approach is concerned with the only validated data that can be collected and stored in the database while the Collect-First approach is concerned with the collectible data that can be directly collected and stored in the database.

The following picture depicts the difference between data warehouse and data lake.

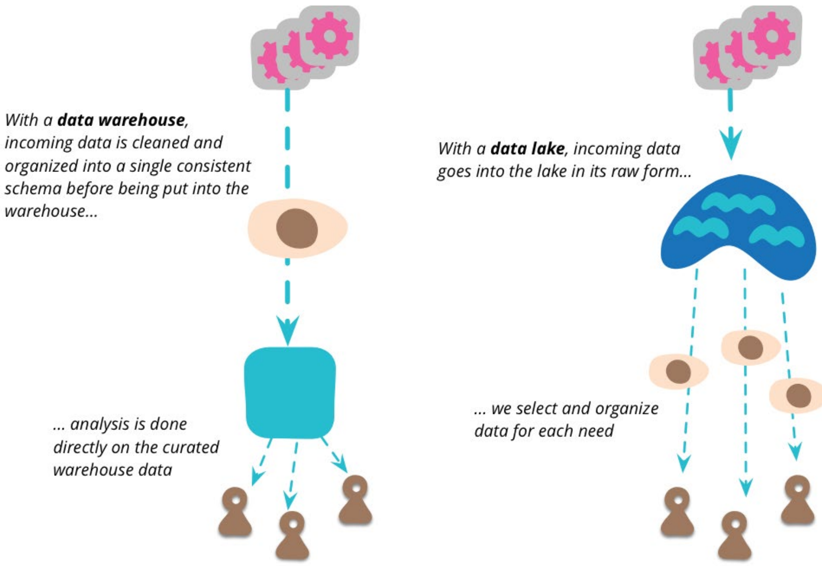


Figure 4. A distinction between the data lake and the data warehouse[4]

3. https://martinfowler.com/bliki/DataLake.html

Providing so many choices in Figure 3, options for implementing this storage include Azure Data Lake Store, Amazon S3, Cloudera Data Lake Service, and Apache HBase.

Batch Processing Block deals with a very large data set that must be processed using a long-running batch job to filter, aggregate, and other data processing for further analysis. Some options for implementing this block are Azure Data Lake Analytics, Cloudera Data Platform, Spark Cluster, and other software.

Real-time Message Ingestion Block deals with real-time data sources, thus, requiring stream processing to collect and store the real-time data. The component of a streaming block is commonly referred to as stream buffering. Some options of software to implement this block are Azure Event Hubs, Amazon MQ, and Apache Kafka.

Stream Processing Block deals with fast data income that must be processed using stream processing services to filter, aggregate, and other data processing for further analysis. Some options for implementing this block are Cloudera Data Flow, Storm, and Spark Streaming.

Analytical Data Store Block can be referred to as an Inmon-style relational data mart that is produced from the data warehouse, in this case, from the data lake. The schema stored in the Analytical Data Store is a group of data for the themed analytical process. Some options for implementing the Analytical Data Store are Azure Synapse Analytics, SQL Server 2019 Data Virtualization, and Spark SQL.

Analysis and Reporting Block has a role to amplify the results of the analytical process so that more users can access the analytical reports. This block provides the user with self-service analytics that enables them to design both their own analytics and their own reporting style, thus, supporting the democratization of data analytics. Some options for implementing the self-service analytics and reporting are Power BI, RStudio, Jupyter, and Cloudera Data Science Workbench.

Orchestration Block encapsulates the end-to-end Big Data Analytics Process including transforming the source data, moving the data between multiple sources, loading the processed data into an analytical data store, or pushing the results directly to a dashboard, a report, or even an API that can be consumed by Line-of-Business Application. Some options to automate these workflows are Azure Data Factory, Collibra, Talend, and Apache Oozie.

3.4 Selecting the Platform

This section will deliver a general concept so that the readers can relate their organization system to the generic of Big Data. Although the previous sections do not discuss the institutional strategy on Big Data, it is at the forefront of all IT vision, mission, objectives, hardware, and software. The institutional strategy on IT especially if Big Data is mentioned may state whether the organization will use Cloud Computing or On-Premise approach for the Big Data Platform. These approaches of building the Big Data guides the selections in different approach each other of selecting the technology platform. Therefore, the available strategy is an essential component of a Big Data System in an organization.

Prior to selecting the software as shown in Figure 3, Big Data at first will indicate what should be available and then create the plan to acquire the necessary hardware. Hardware is the core of a Big Data system. It deals with how much the capacity of the storage; whether the capacity is sufficient or not to store a huge amount of data, whether the process of retrieving the data is fast enough or not to start the analysis, and whether the analytical process is supported with adequate computation power (the amount of RAM and how many cores in the server) or not. Based on the organization strategy on Big Data, the hardware may incorporate not only branded computers but also some commodity computers. After selecting the size of the hardware, the next process is to select the operating system that will run on the hardware. The hardware and its operating system then will work with the selected system software later.

Selecting the operating system is not challenging since the choices are very limited in the Big Data System. These are either Linux-based or Windows operating systems. Since the hardware in Big Data can be categorized into four functionalities; as a collector, as a storage, as analytical processing, and as an end-user interface, the choice of software program for those functions is more challenging than that of an operating system.

The challenges in selecting the software program are as follows:

- **Open-Source vs Proprietary:**

Open-Source Software is software that has its source code open to the public. As a consequence, people are able to use and modify the software to fit their needs without any expense or near-free. Although the concept of open-source is about the freedom to modify, most people acknowledge open-source software as free software that we do not need to purchase. Proprietary software, on another side, is software developed by a company and the company retains and keeps the source code exclusively. Those who want to use the software should pay some amount of investment in the form of a license to use the software. In general, the buyer does not have the right to modify the software, thus, less flexible to fit the needs. Some examples of selecting between the open-source and proprietary are MongoDB vs Oracle Database Management System, Dagster vs Collibra, and RStudio vs SAS.

Perpetual vs Subscription Licensing

Perpetual licensing is the type of licensing that attaches to an account and does not have an expiration date. When a company buys software with a perpetual license, the company will be eligible to use the software with no time limit. In order to keep the software update, there is an additional cost to buy the service. On another hand, Subscription licensing is a type of licensing based on a certain period of time, usually one year. In order to keep using the software, consistent renewal should be taken place. This type of licensing is a common practice in cloud computing services.

Cloud vs On-Premise

In the era of Industry 4.0, Cloud Computing has become a buzzword in the way of developing the IT infrastructure. Some software for applying Big Data is only available on Cloud which means the user only has the option to subscribe to the services. Using Cloud technology, the subscriber does not buy the server nor the software. Some examples of software that is only available on the cloud are Azure Machine Learning and Dataiku. However, some technologies have still opted for the on-premise method in which the software is installed in our own server, for example, SQL Server Database Management System and Shiny Server.

The following chart illustrates a sample of a combination of open-source software for establishing a Big Data Technology Platform.

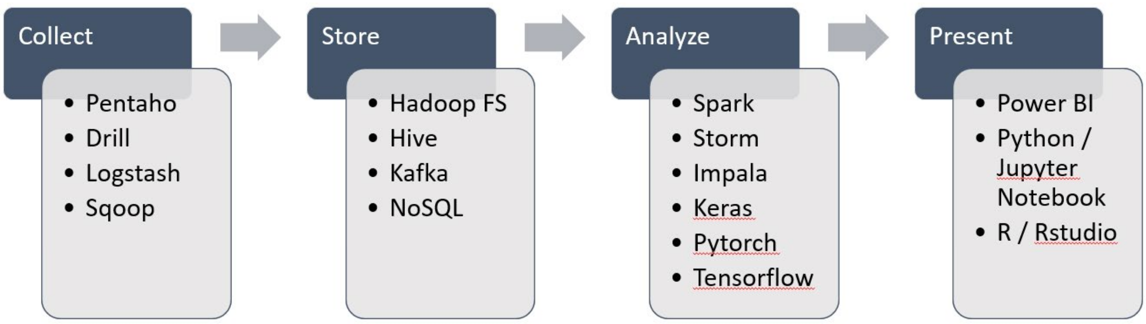


Figure 5. A sample of Big Data Platform by combining several Open-Source Software

In addition to these Big Data Technology Platforms, there has been also shifting in the infrastructure that leads to a more compact data center in which the Big Data Platform will reside. The state-of-art data center infrastructure is moving from monolithic, then virtualization, and now containerization. The following chart shows the development of data center infrastructure.

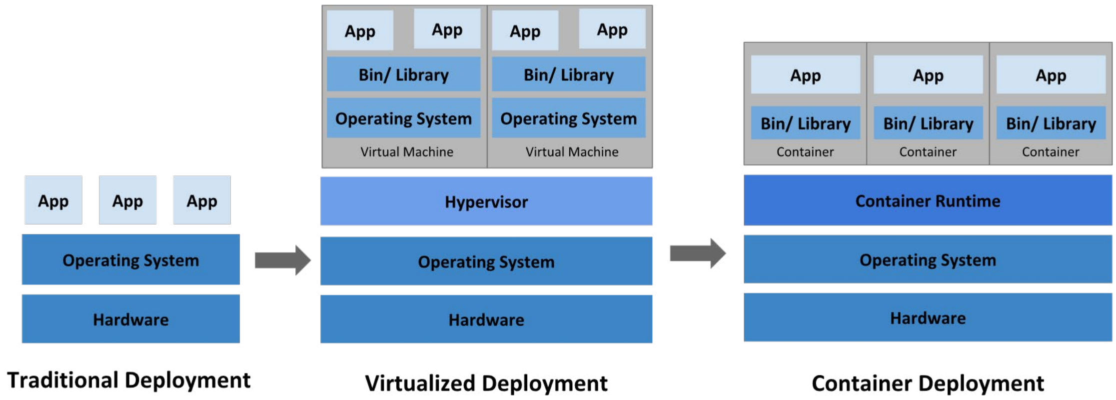


Figure 6. The trend of infrastructure deployment

An institution may consider the deployment type of the Big Data Platform whether to follow traditional, virtualization, or containerization approach. The type of deployment may impact the total cost of ownership of the platform.

4. A Survey on SAI's capacity on Utilizing the Big Data

4.1Data Collection

SAI Indonesia used a questionnaire as a method to collect the data in August 2020. The questionnaire consists of three parts, as follow:

1. Basic Information of SAI
2. Structure of IT Unit in SAI's Organization Structure
3. Basic IT Capacity of SAI
4. Big Data Identification
5. Data Collection: Technique and Procedure
6. Reporting and Data Visualization

At the end of the Data Collection phase, there were 24 SAIs replying to the questionnaire.

4.2 Big Data Platform

In dealing with Big Data, SAI needs a big data platform to store a vast amount of data with various types of data. On the question of the current use of database platforms, some SAIs use more than one database platform. The most popular database platform is Microsoft SQL Server. 14 SAIs use this relational database management system. The second and the next platforms are Oracle and MySQL Database which are used by 11 and 8 SAIs respectively, as shown on the following chart.

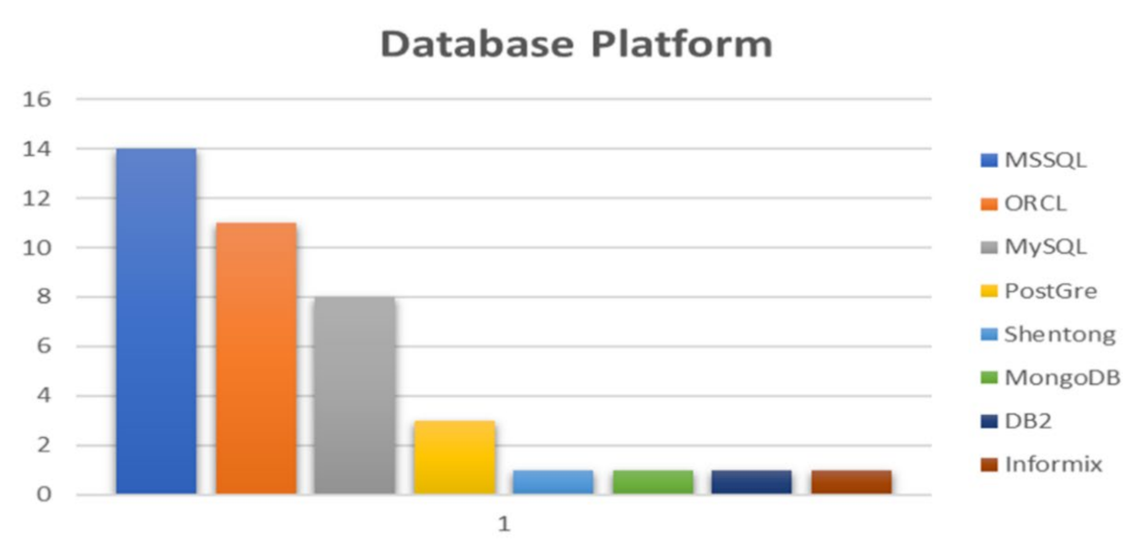


Figure 7. Data Platform used by SAI

These 24 SAIs have not used a specific platform for storing unstructured data such as Hadoop or other S3-Compliant File Systems. The content of their Big Data is mainly in the form of structured data. There is no sufficient evidence during this research to mention that some SAI have started to analyze text, image, and video.

4.3 The method of Data Collection

Most SAIs have authority, by law, to collect data from the auditee. However, the procedure of collecting the data is not explicitly stated in the regulation. However, there are some SAIs whose Audit Law states clearly that the auditee should submit the government financial transaction data electronically. From the survey, only 21% use the Internet as a means of transferring the data from the auditee remotely. A majority of the respondents collect the data using a conventional method such as manual collection (25%), Read-Only Access on auditee’s location (21%) and using removable media (21%). See the following chart.

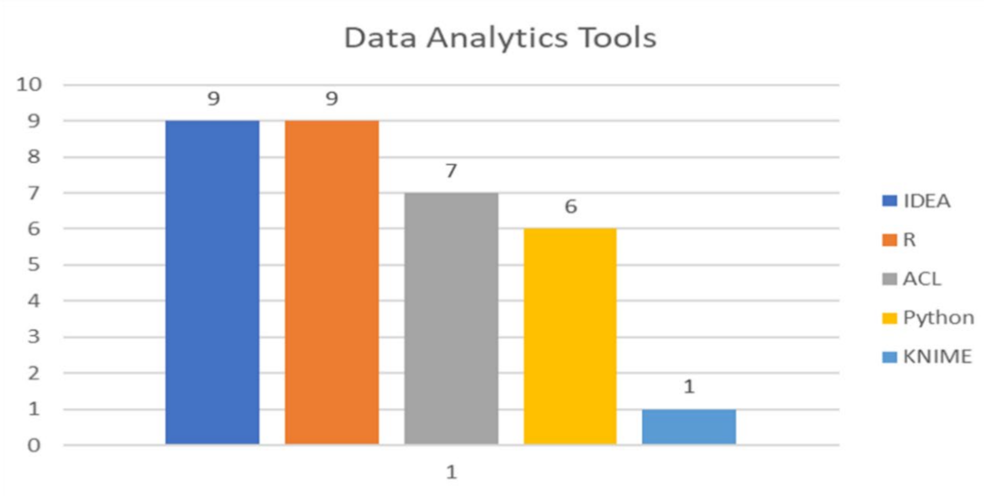


Figure 8. Method of collecting data from auditee

4.4 Big Data Analytics Platform

There is various software either open-source software or proprietary software for conducting advanced analytics on a huge amount of data. The survey shows that there has been a mix of both open source and proprietary software. The top three software for conducting data analytics are IDEA, R, and ACL. Most of the respondents are still using Generalized Audit Software (GAS) such as IDEA and ACL. The use of this software indicates that most SAIs are still mainly working on the structured data, which implies that the method of analytics remains prominent following the practices of Computer-Assisted Audit Techniques (CAATs). 15 SAIs have started to use analytics-specific programming languages such as R and Python. It also indicated that SAIs are aware of the need for advanced analytics on their data in which the use of GAS has not met the auditors' need.

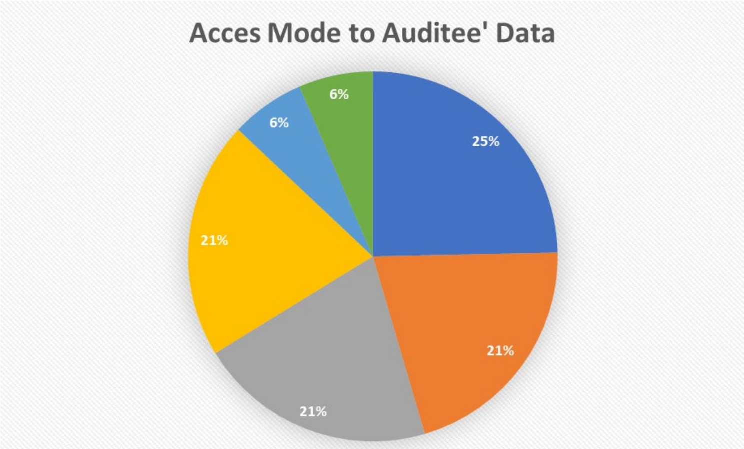


Figure 9. The use of Data Analytics Software

4.5 Big Data Analytical Model

Practically, there is five types of analytics model, i.e., Clustering, Classification, Association, Regression, and Prediction. These models are common in analyzing structured data. However, these five models do not fit to analyze unstructured data. The prior process should be taken for transforming the unstructured data to structured data, for instance, the text mining technique is used for transforming the unstructured data to structured. Some methods in Text Mining are Text Analysis, Social Network Analysis, and Sentiment Analysis. The survey result on the implementation of analytical models shows that the Classification method is the common practice in the audit process. Auditor uses classification in the form of doing cross-tabulate, summarizing, filtering, and grouping. Other than the classification method, the Clustering method and Regression are the prominent methods in analyzing the data. As shown on the following chart, 13 SAIs apply Clustering Analysis, 11 SAIs apply Regression Analysis, and 5 SAIs have applied Text Analysis for analyzing the data.

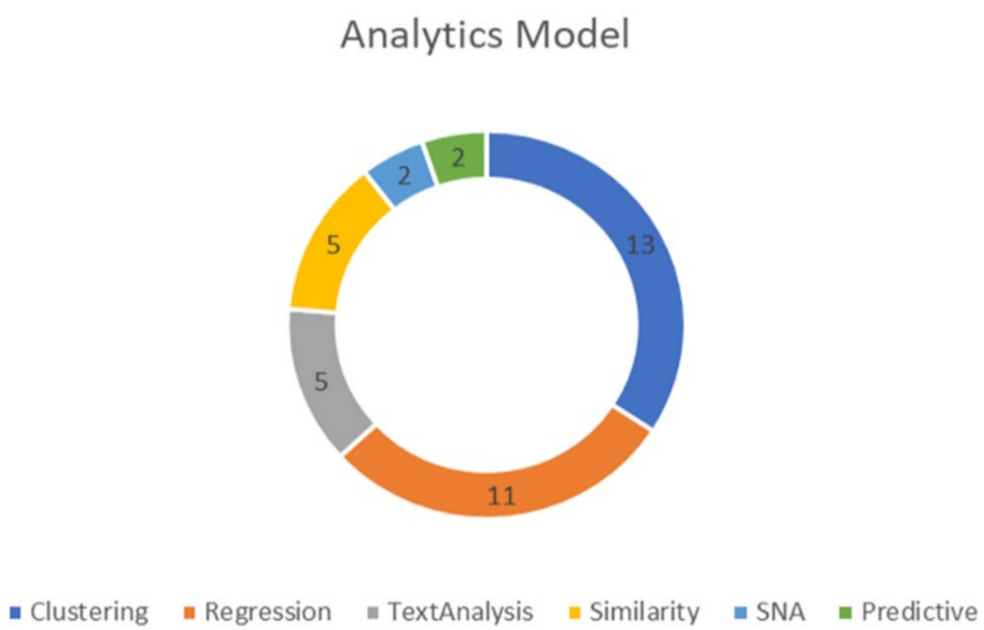


Figure 10. The use of the Analytics Model

4.6 Big Data Visualization Tool

Data Visualization has an important role in the implementation of Big Data Analytics. Data visualization could amplify the analytical result so that all auditors can have access equally. The research uncovers that Microsoft Excel is the most favorable software for visualizing Big Data. The result implies that most SAIs use structured data in their data analysis process. The result also opens the opportunity to enhance the SAI’ s capacity for storytelling with the data by using an astonishing data visualization.

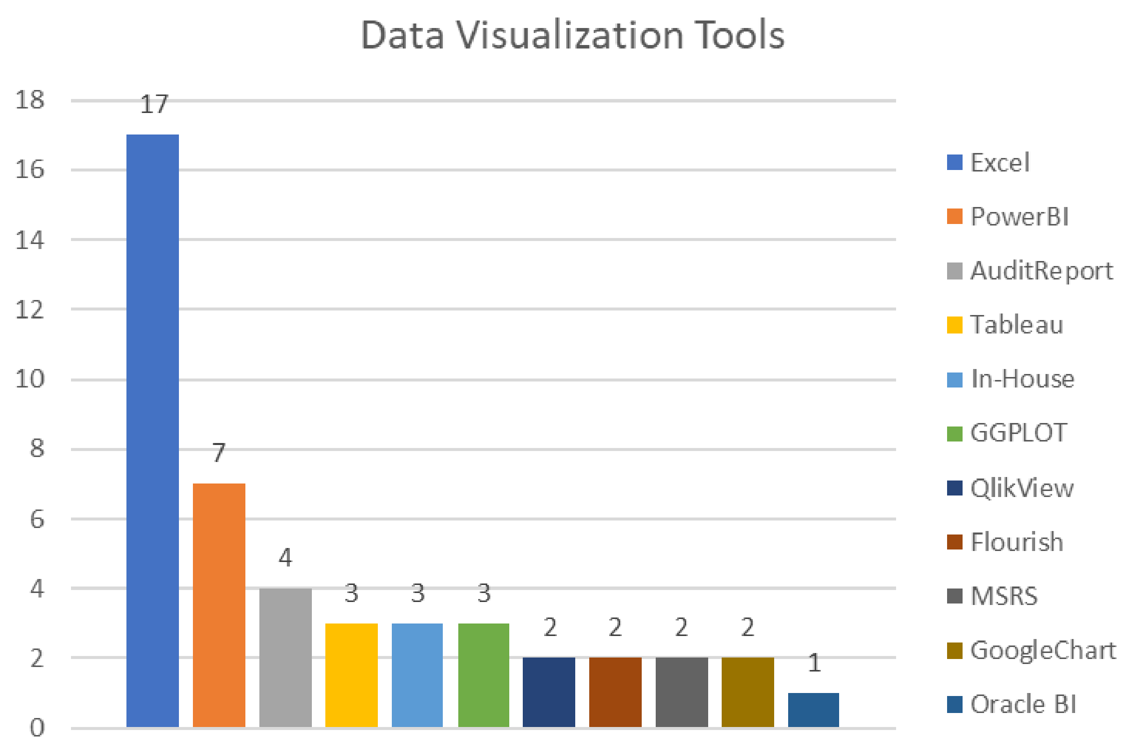


Figure 11. The use of Data Visualization Tools

5. Big Data Capacity Matrix

This section introduces the result of a questionnaire analysis on identifying patterns related to Big Data Capacity. The result portrays the Big Data capacity in SAIs based on the technology and the process of data analytics stated on the questionnaires. The following chart describes the position of each SAIs among others.

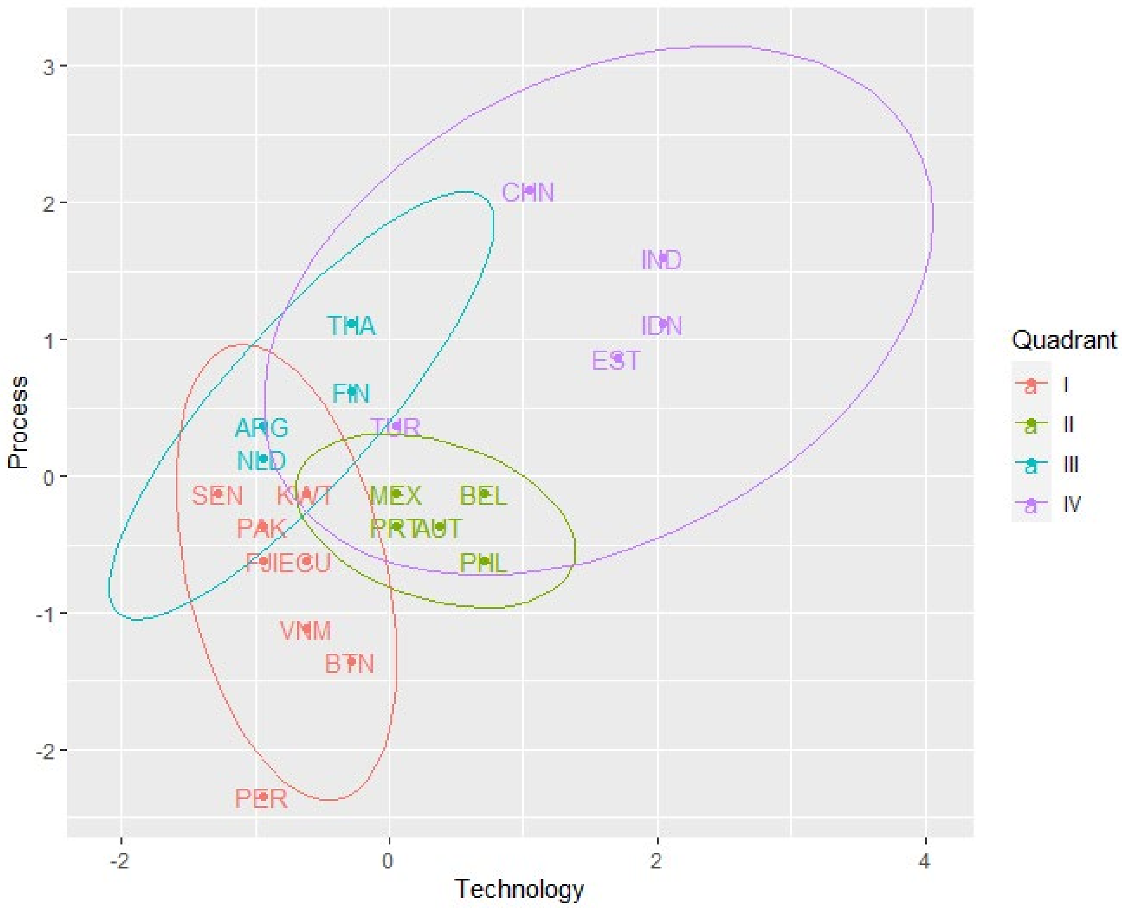


Figure 12. Big Data Analytics Capacity Matrix

For the analysis, two variables were defined;

◉ **Technology forms the axis of the chart. Three indicators from participant' s questionnaires formulated the Technology variable. These elements are;**

1. The tools used for analyzing the data
2. The tools used for visualizing the data
3. IT infrastructure capacity

◉ **The Process of Big Data Analytics comprises 8 indicators collected from the questionnaires. These elements are;**

1. Structure of IT Unit
2. The method of accessing the auditee' s data

3. The techniques in analyzing electronic data

The indicators were quantified and scaled. The output of the process was the score of Technology and Process. The scaled score was then mapped into the chart in Figure 10.

On the chart, there is a pattern that might be indicative of how an SAI implemented the Big Data Analytics Capacity. The SAIs in the chart are concentrated on quadrants I, II, and III. 30% of SAIs are still on quadrant I. This implies that most SAIs have started the Big Data Analytics Capacity Building by firstly developing the capacity of their technology or strengthening the procedure to conduct Big Data Analytics. Based on the country paper of 23 SAIs, SAIs in quadrant I are the SAIs with a very limited opportunity to implement Big Data Analytics. The limitation of their capacity on conducting Big Data Analytics is mainly organizational and legal issues. Some SAIs are able to obtain the data only during audit work. Some SAIs expressed their IT Capacity when they do not have a dedicated data center and they cannot use cloud service due to government regulation. Also, some SAIs are still working on CAATs in which their analytical style is still limited to summarizing and classification. They deal with traditional data and conventional analytics processes. The use of spreadsheet software and Generalized Audit Software is still prominent in analyzing the data. However, SAIs in quadrant IV are the SAIs that have experience in applying the Big Data Analytics approach in their audit assignment.

In all, the chart illustrates factors to consider for developing the Big Data Analytics Capacity in SAIs. Improving the Big Data Analytics capabilities of the institution is the first objective. Once there is a sufficient capacity for processing the huge amount of data with its variety, the next phase of Big Data Analytics Capacity Building is to strengthen the infrastructure capacity.

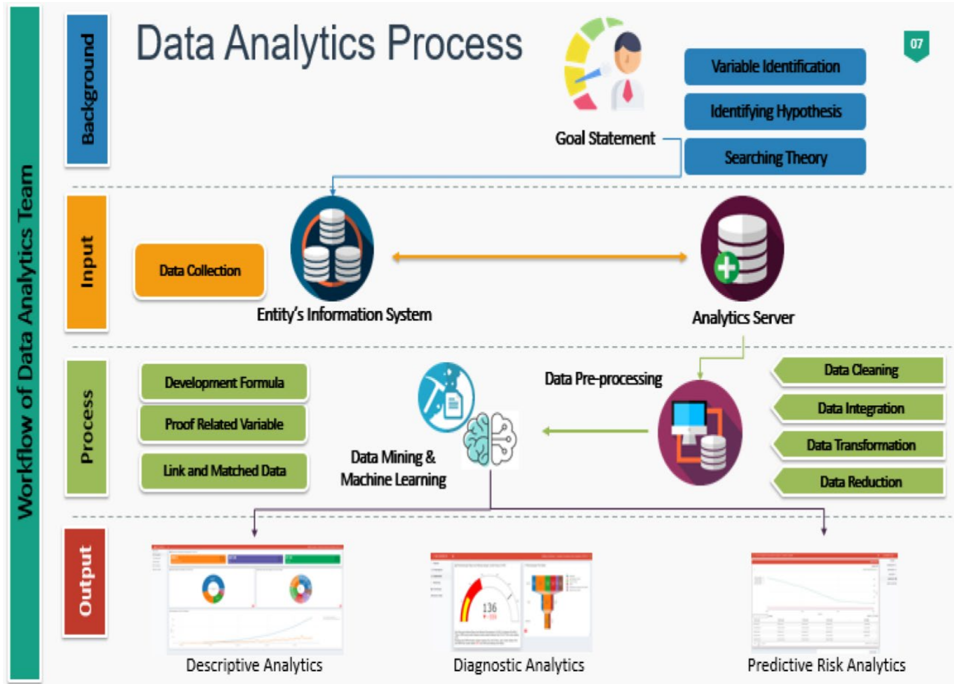
It was observed that some SAIs with enhanced Big Data Analytics capacity have many options for obtaining the data and have a wider horizon of analyzing the data. Therefore, they have a tool for increasing their institutional maturity level for having a foresight capability, which is beyond the oversight and insight capabilities.

6. SAI’ s experience in applying Big Data Analytics

6.1 Audit Board Office of the Republic of Indonesia (BPK RI)

BPK realizes that development in information technology provides an opportunity to collect, process, and analyze data from various sources, either obtained directly from audit entities or acquired from the public domain. Learning from BPK’ s experience in 2010-2014 in developing e-Audit, BPK revitalize the e-Audit with the Big Data Analytics concept. While in e-audit BPK applied the concept of validating the data first, then collecting them (Valid-First), in the Big Data Analytics approach, BPK applies the concept of Collect-First in which the data will be collected as soon as it is available, then validate whenever it is needed for further analysis. Big data analytics utilization is expected to improve audit quality, organization responsiveness, and information resources optimization. By pushing digital transformation, BPK is aiming to become a data-driven organization.

In the past, BPK has utilized Big Data Analytics for performance audits and special purpose audits. BPK recognizes that Big Data Analytics implementation for financial audit may further benefit the organization. As illustrated in the following diagram, BPK uses a goal-based approach for the Data Analytics implementation. In general, a Data Analytic process will start when a use case or an interesting question is proposed. Further discussion will be carried out to formulate the relevant data model and identify data that needs to be sourced. Sometimes the discussion is



held intensively that we call the process a boot camp. The final product is presented to users in a big data analytics dashboard. The following figure illustrates how the analytical process works.

Figure 13. Analytic Process Workflow

The following diagram illustrates the technology platform of BPK's Big Data Platform.

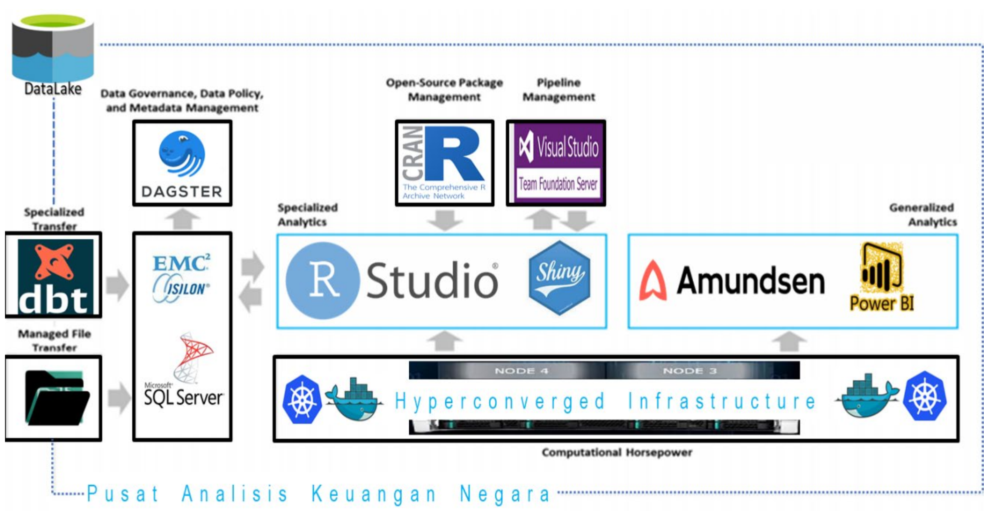


Figure 14. BPK's Technology Platform

BPK has developed analytics that is considered highly relevant for financial audit purposes, including the following:

- **Government revenue and expenditure**

The analytic features trend of central government revenue and expenditure for year-end and for each account for the last five financial years. In comparison, the dashboard also highlights the revenue and expenditure for significant ministries, bodies, and agencies during the same period.

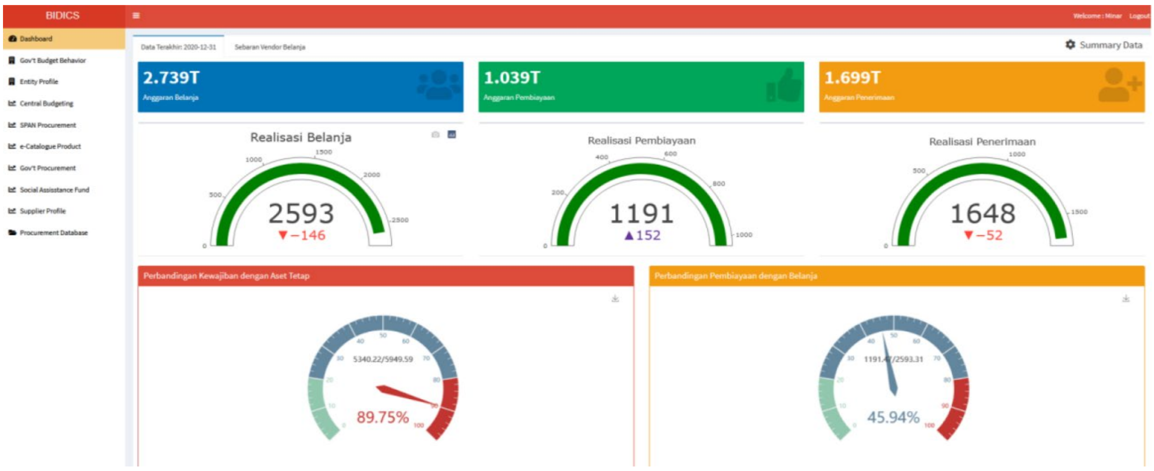


Figure 15. State Budget of Revenue and Expenditure

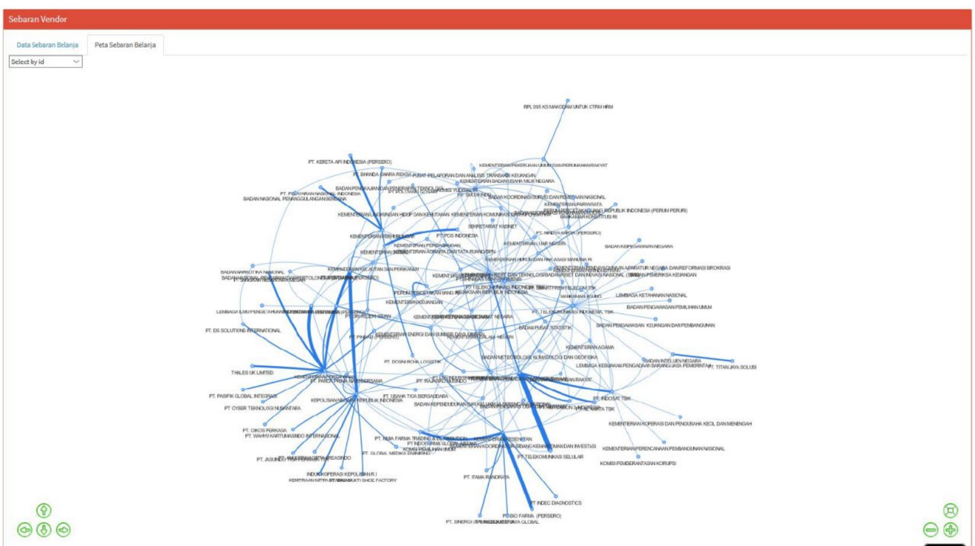


Figure 16. Network Graph of Top 50 Government's Vendors

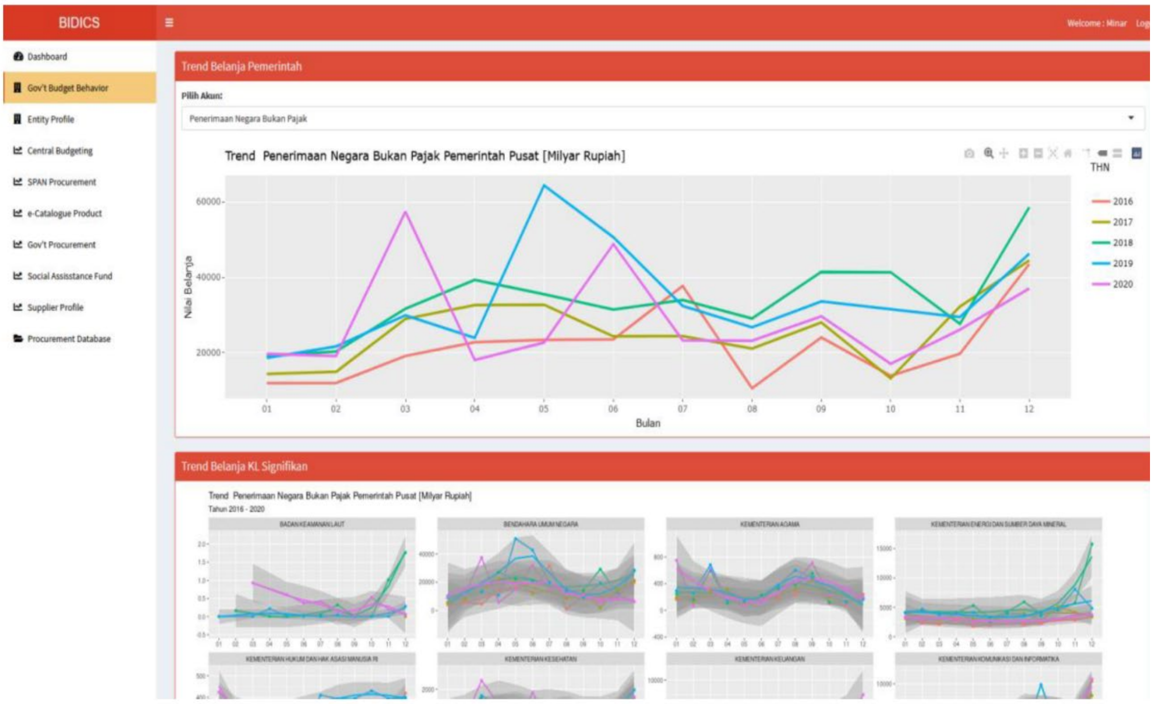


Figure 17. Central Government Revenue and Expenditure for Last 5 Years by Account

Further in the dashboard, the financial audit team is provided with a general profile of each ministry, agency, and body whereby they can access the entity's expenditure trend, budget profile, contractual payment profile, and fixed asset profile. In the fixed assets section, analytics are provided related to asset correction, reclassification, revaluation, and reconciliation.

- **Government procurement**

The Data Analytic dashboard provides two types of procurement data, one from the State Treasury and Budget System and the other one is procurement data from Electronic Procurement Service. The first data source provides yearly and monthly procurement spending by vendors and by ministries. While the latter provides a similar profile with a further granularity of procurement by tender and non-tender. Since it also provides details on procurement data, further analysis can be developed to detect abnormalities that should come to auditors' attention. The current procurement anomalies feature in the dashboard are a tender winner is not the lowest bidder, tender winner bid with a price higher than owner's estimate, tender winner name is different from the name in the actual procurement agreement. These anomaly indications dashboard is shown in Figure 18.

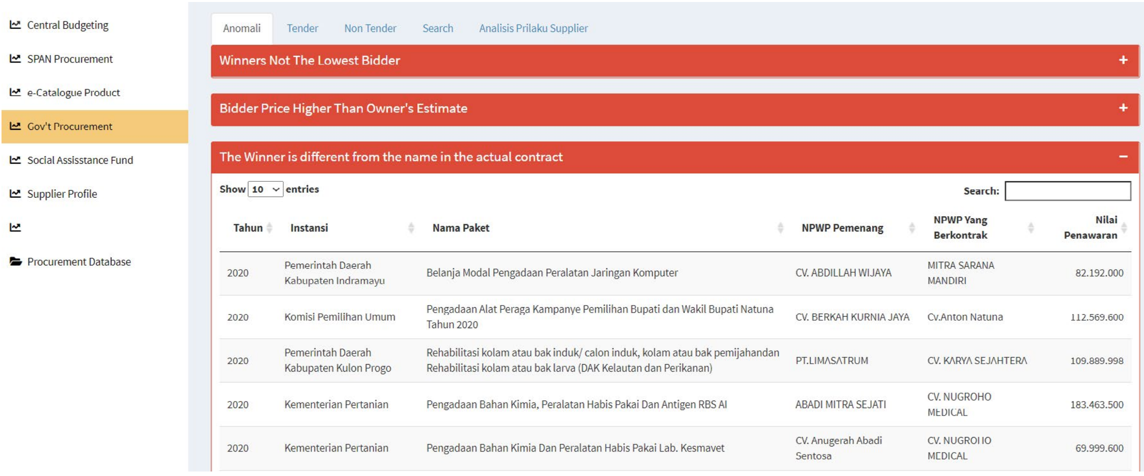
Social Assistance Distribution

The government launched many social assistance programs that targeted different groups. The general rule for the assistance is nobody registered for more than one social assistance program. The analytics help detect individuals whose receipt of social assistance does not comply with the rule. The dashboard of social assistance distribution is shown in Figure 19.

Supplier Profile

The Data Analytics dashboard provides the auditors with information on vendors whose projects are with a government institution. The information includes a historical project with government institutions, historical payment receipt from a government institution, and historical participation in government procurement; how many times the vendor applies the procurement, and how many times the vendor has been elected as the winner of the procurement. In addition, the analytic process enables the auditor to obtain information related to the owner and the manager of the vendor.

Figure 18. Dashboard of Social Assistance Distribution



Tahun	Instansi	Nama Paket	NPWP Pemenang	NPWP Yang Berkontrak	Nilai Penawaran
2020	Pemerintah Daerah Kabupaten Indramayu	Belanja Modal Pengadaan Peralatan Jaringan Komputer	CV. ABDILLAH WIJAYA	MITRA SARANA MANDIRI	82.192.000
2020	Komisi Pemilihan Umum	Pengadaan Alat Peraga Kampanye Pemilihan Bupati dan Wakil Bupati Natuna Tahun 2020	CV. BERKAH KURNIA JAYA	Cv.Anton Natuna	112.569.600
2020	Pemerintah Daerah Kabupaten Kulon Progo	Rehabilitasi kolam atau bak induk/ calon induk, kolam atau bak pemijahandan Rehabilitasi kolam atau bak larva (DAK Kelautan dan Perikanan)	PT.LIMASATRUM	CV. KARYA SEJAHTERA	109.889.998
2020	Kementerian Pertanian	Pengadaan Bahan Kimia, Peralatan Habis Pakai Dan Antigen RBS AI	ABADI MITRA SEJATI	CV. NUGROHO MEDICAL	183.463.500
2020	Kementerian Pertanian	Pengadaan Bahan Kimia Dan Peralatan Habis Pakai Lab. Kesmavet	CV. Anugerah Abadi Sentosa	CV. NUGROHO MEDICAL	69.999.600

Figure 19. Government Procurement Anomalies

6.2 United State Government Accountability Office (US-GAO)

GAO established the Innovation Lab to tackle signigicant oversight challenges that exist across the government through explorations of data science and emerging technologies. As a research and development entity, the Lab staff takes a hands-on approach to experienment, adapt, and scale capabilities across artificial intellienge, blockchain, cloud services, extended reality, and cyber-physical edge computing devices for use by auditors.

One way the Lab is elevating impacts of oversight is through intersections of data science and digital delivery. For example, the Lab pioneered an approach of using Monte Carlo simulation models to illustrate effects of policy tradeoff decisions that agency programs often face. This interactive web-based resource is a companion to a report on identity verification practices and allows users to toggle across various combinations of identity verification controls and see hypothetical impacts on the scope of improper payments. This calibrated approach recognizes important differences across federal programs and enables nuanced considerations on control implementations. In another example, the Lab collaborated with several GAO mission teams to help federal officials better understand and combat fraud that affects the federal government by launching the GAO Antifraud Resource. This web-accesible resource draws upon interrelated antifraud guidance from GAO's conceptual fraud model and utilizes a unique ontology model to organize diverse contents. The Antifraud Resource was nominated for the Association of Government Accountants' Relmond P. Van Daniker Government Transparency Award.

In addition, the Lab is supporting GAO's broader modernization efforts by strengthening capacity in data science, data literacy, and data governance. The Lab designed and deployed a state-of-the-art computational platform in the cloud—dubbed the Analytics Foundry—to sustain GAO's ability to conduct sophisticated advanced analytics work. In April 2022, the Lab's physical space at GAO Headquarters launched as a future-of-work concept that facilitates purposeful interactions across disciplines and engagement with prototype solutions.

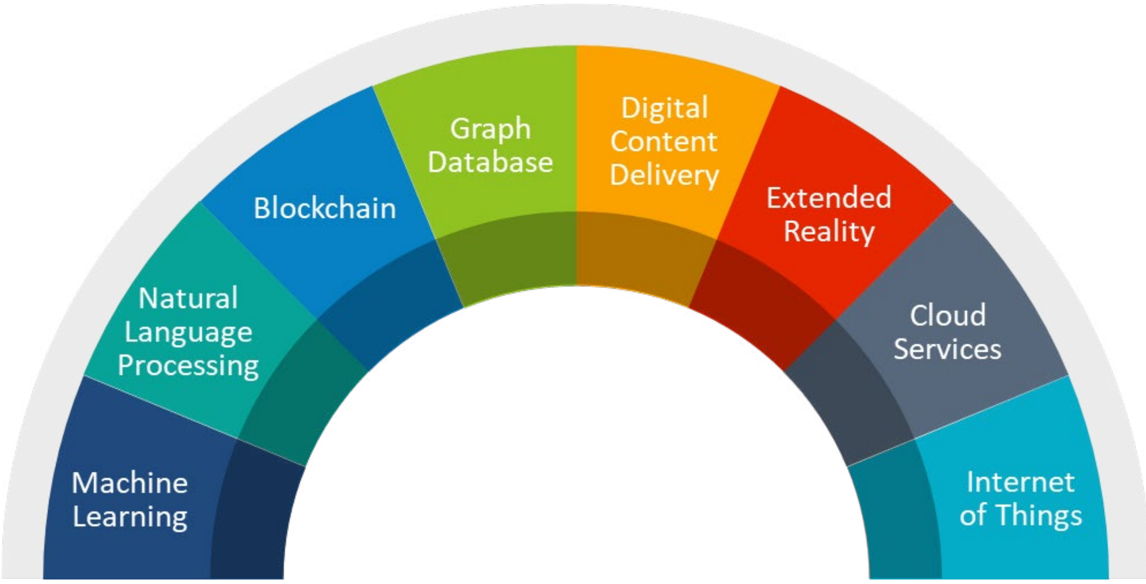


Figure 20. Innovative Technology Spectrum

7.Summary and Recommendation

Most SAIs work on historical data that was collected on a periodical basis. Some SAIs obtain the data only during audit work while other SAIs have the advantage to get the data periodically regardless in an audit period or not. This condition implies that SAI does not require a big data platform for handling real-time data for real-time analysis.

Most SAIs still deals with traditional data and conventional analytics process. The use of spreadsheet software and Generalized Audit Software is still prominent in analyzing the data. Thus, auditors mostly use Computer-Assisted Audit Techniques (CAATs) during the audit work. This condition indicates that CAATs are the basic competency of auditors in most SAIs.

Considering the variety of software, there are some considerations for designing the Big Data Solution. These considerations are as follow:

- **Understanding the data source and data user**

1. Whether the data should be collected online on a real-time or periodical basis.

2. Whether the data collected from Line-of-Business Application, IoT Devices, Social Network, Archives, or Public Web site.

3. Whether the data will be stored in RDBMS, Hadoop, or S3-Compliant File System

- **Discovering architecture drivers**

1. Whether the user needs to apply real-time or on-demand analytics.

2. Whether the solution will provide self-service analytics or not.

3. How big is the data needed for certain analytics?

- **Mapping the reference architecture to the technology stack**

Selecting the available software and applying them to the appropriate component on the architecture

- **Big Data technologies are evolving rapidly**

Some technologies, especially for data storage, are matured so that the system upgrade might be minimal while other technologies are relatively new so that the system upgrade might be more frequent.

Last but not least, an initiative of implementing the big data infrastructure requires more responsibility to manage and secure the data, to manage the data storage, to enforce policies and standards, and to optimize the data processing. These increasing responsibilities urge the development of the Big Data Governance Framework (BDGF)

APPENDIX A: QUESTIONNAIRE

Questionnaire on Big Data Analytics Practices in Supreme Audit Institutions

1. Basic Information of your SAI

- Country
- Name of your SAI in English
- Number of Personnel

1. Auditor

2. IT Specialist is those who are responsible for the daily operation of SAI's IT such as Database Administrator, Application Programmer, and Network Engineer.

3. If your SAI has a specific job title in IT other than IT Auditor, please mention and put the number of personnel.

- Please specify, if any, the IT-related international certifications, such as CISA, CISSP, and others, recognized in your SAI and how many personnel entitle these certifications.

2. Structure of IT Unit in your SAI

- Does your SAI have a specific unit within your organization in charge or has a role in managing and developing the following functions?

1. Line of Business Application [Y/N]

2. Network Infrastructure [Y/N]

3. Computer Servers [Y/N]

4. Database Administration [Y/N]

5. Data Analyst [Y/N]

6. Others…. Please mention.

7. If there is more than one [Y] answer, are those with [Y] answers in the same organization unit? Please draw the structure of that unit.

◉ Does your SAI have a specific unit within the IT Unit in charge of conducting IT Audit or Data Analytics? If yes, please explain the roles of that unit to support the audit assignments for your auditors.

3. Basic IT Capacity of your SAI

◉ The total capacity of your SAI's data storage: ………. [G/T/P bytes]

◉ The total computation power of your SAI's server: ………… Cores CPU

1. CPU Cores: ……….. Cores
2. Speed per Core : …… GHz
3. RAM: …… G [G/T/P bytes]
4. In acquisition process

◉ Your SAI has a dedicated Data Center [Y/N]

◉ Your SAI has a dedicated Secondary Data Center [Y/N]

◉ Please mention the databases used in your SAI, for example ORACLE, SQL Server, MySQL, Other Database Management System, please mention.

◉ Does your SAI use Cloud technology? [Y/N]:

1. If Yes, please mention the purposes or services you of the Cloud Technology for you SAI.
2. If No, please mention any reason, if any, of not using Cloud Technology.

◉ Does your SAI have an infrastructure for a Big Data? If yes, please describe the content and the purposes of such Big Data.

4. Big Data Identification

◉ Does your SAI collect electronic data from external such as from auditees or from public domain? [Y/N]

1. If yes, is it in a regular basis? [Y/N]
2. If your SAI collected data in a regular basis, please mention the kind of data and the mechanism of collecting such data.

◉ Among the data that reside on your SAI's Data Centre, how much is the biggest data among them in term of record count? What kind of data is that?

◉ If your SAI collected data from public domain, what kind of data does your SAI collected?

◉ Following is the common data type for analysing. Please fill in the blank with the "V" if it is true

Data Type	Available in your SAI	Analyzed in your SAI	Collected from Public Domain
Tabular			
Text			
Audio			
Video			
Image			

◉ When analysing a huge amount of data, does your SAI assign personnel with a specific expertise? What expertise is required to conduct such analysis?

◉ When analysing a huge amount of data, does your SAI use specific tools? What tools is used for conducting such analysis?

5. Data Collection: Technique and Procedures

◉ Your SAI collects data electronically from audit entity only when conducting an audit. [Y/N]

◉ Your SAI collects data electronically in a regular basis although not in an audit assignment. [Y/N]

◉ Among these methods of accessing electronic data, thus, collecting them from audit entities, please choose all that apply in your SAI.

Methods	Is Apply?	Only during an Audit Assignment	Type of Data Collected[5]
Manual Records by Auditors	☐	☐	[T,Tx,A,V,I]
Shared using a removable media	☐	☐	[T,Tx]
Read Only Access in Auditee's Information System	☐	☐	[T,Tx,A,V,I]
Electronic Transfer using Internet	☐	☐	[T,Tx,I]

Cloud Services online provided by Auditee	☐	☐	[T,Tx,A,V,I]
Real time data sharing	☐	☐	[T,Tx,A,V,I]

◉ Is there any arrangement between your SAI and Audit entities in term of collecting data? Such an arrangement is similar but not limited to a Memorandum of Understanding (MoU), Government Regulation, and Law.

◉ Is your SAI collected electronic data from public domain such Social Media Platform, Website? If yes, please mention some of these data including the source of the data and the purposes of collecting them.

6. Data Analytics: Technique and Procedures

◉ The followings are some techniques in analysing electronic data. Please choose all that apply, including the source data being analysed using the selected technique.

Technique	Is Apply	Data Source
Summarizing	☐	
Filtering	☐	
Clustering	☐	
Cross tabulation	☐	
Benford Analysis	☐	
Text Analysis	☐	
Regression Analysis	☐	

5. Choose all that apply.

T is Tabular (Row and Column data set)

Tx is Text information such as a report, description, and minutes of meeting. A PDF file can be regarded as a text Information.

A is Audio such as voice records of an interview, records of a meeting.

V is Video such as film records of an interview.

I is Image such as a photo, graph, and other scanned data. A PDF File can be regarded as an image, as well.

Similarity Analysis	☐	
Social Network Analysis	☐	

◉ The followings are some tools in analysing electronic data. Please choose all that apply, including the source data being analysed using the selected tools.

Tools	Is Apply	Data Source
ACL	☐	
IDEA	☐	
R Language	☐	
Python Language	☐	
Jupiter Notebook	☐	
Rapid Miner	☐	
Azure Analytics	☐	
KNIME	☐	
…		

◉ Does your SAI implement a system of Artificial Intelligence (AI) for your SAI, either for internal management or for government auditing purposes? If yes, please describe briefly the uniqueness and the advantage of the system for supporting you SAI operation.

7. Reporting and Data Visualization

◉ The followings are some tools in visualizing the data. Please choose all that apply, including the approach of updating the report.

---

6. Type of Update:

Near Real time -> The report will be updated automatically immediately once the data source was updated

Batch-> The report will be updated periodically

Tools	Is Apply	Type of Update[6]
Power BI	☐	
QlikView	☐	
Dundas	☐	
Tableau	☐	
In-House Application	☐	
Microsoft Reporting Service	☐	
Microsoft Excel	☐	
Google Chart	☐	
...		
...		

◉ Do you have an audit report that present some audit findings from the use of big data analytics in your organization? If Yes, please provide the summary of the report as attachment of this feedback.

References

AICPA. (2002). Consideration of Fraud in a Financial Statement Audit.

AICPA. (2017). Guide to Audit Data Analytics. Wiley.

Alles, M., & Gray, G. L. (2016). Incorporating big data in audits: Identifying inhibitors and a research agenda to address those inhibitors. International Journal of Accounting Information Systems, 22, 44–59. https://doi.org/10.1016/j.accinf.2016.07.004

Appelbaum, D., Kogan, A., & Vasarhelyi, M. A. (2017). Big data and analytics in the modern audit engagement: Research needs. In Auditing (Vol. 36, Issue 4, pp. 1–27). American Accounting Association. https://doi.org/10.2308/ajpt-51684

Arthur, W. B. (2017). Where is technology taking the economy? | McKinsey. McKinsey. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/where-is-technology-taking-the-economy

Bizarro, P. A., & Dorian, M. (2017). ARTIFICIAL INTELLIGENCE: THE FUTURE OF AUDITING - ProQuest. Internal Auditing. https://search.proquest.com/docview/1958593764/94F9BFB4178F4A88PQ/5?accountid=86413

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information Communication and Society, 15(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

DataLake. (n.d.). Retrieved May 16, 2021, from https://martinfowler.com/bliki/DataLake.html

Doug Laney. (2001). META Delta. Application Delivery Strategies, 949(February 2001), 4. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Elliot, V. H., Paananen, M., & Staron, M. (2020). Artificial intelligence for decision-makers. Journal of Emerging Technologies in Accounting, 17(1), 51–55. https://doi.org/10.2308/jeta-52666

Goldberg, D. E., & Holland, J. H. (1988). Genetic Algorithms and Machine Learning. In Machine Learning (Vol. 3, Issue 2, pp. 95–99). Springer. https://doi.org/10.1023/A:1022602019183

Herschel, R., & Miori, V. M. (2017). Ethics & Big Data. Technology in Society, 49, 31–36. https://doi.org/10.1016/j.techsoc.2017.03.003

INTOSAI. (2019). Data Analytics Guideline. In INTOSAI WDITA. https://doi.org/10.2307/j.ctvmd84wn.11

Knuth, D. E. (1997). The Art of Computer Programming, Vol. 1: Fundamental Algorithms, 3rd Edition: 9780201896831: Computer Science Books @ Amazon.com. In Addison-Wesley (3rd ed.). https://www.amazon.com/Art-Computer-Programming-Vol-Fundamental/dp/0201896834/ref=pd_lpo_14_t_0/130-1001902-8539403?_encoding=UTF8&pd_rd_i=0201896834&pd_rd_r=78d4e0b3-e067-4f58-ad30-6130e4450dc2&pd_rd_w=CmuJd&pd_rd_wg=OIJqh&pf_rd_p=a0d6e967-6561-454c-84f8

Mayer-Schonberger, V., & Cukier, K. (2013). Big Data: A Revolution that Will Transform how We Live, Work, and Think - Viktor Mayer-Schönberger, Kenneth Cukier - Google Books. Houghton Mifflin Harcourt. https://books.google.co.id/books?hl=en&lr=&id=uy4lh-WEhhIC&oi=fnd&pg=PP1&dq=big+data+schonberger&ots=JudbipJULL&sig=GsfnPAVEXiYbJcClAmG_yXiRlyQ&redir_esc=y#v=onepage&q=big data schonberger&f=false

Warren, J. D., Moffitt, K. C., & Byrnes, P. (2015). How big data will change accounting. Accounting Horizons, 29(2), 397–407. https://doi.org/10.2308/acch-51069

Yoon, K., Hoogduin, L., & Zhang, L. (2015). Big data as complementary audit evidence. Accounting Horizons, 29(2), 431–438. https://doi.org/10.2308/acch-51076

INTOSAI
Working Group on Big Data

Secretariat of INTOSAI Working Group on Big Data

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

Tel: + 86 (10) 50992782

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