25th Meeting of the Wiesbaden Group on Business Registers
- International Roundtable on Business Survey Frames

Tokyo, 8 – 11 November 2016

Siti Haslinda Mohd Din, Nur Aziha Mansor
Department of Statistics Malaysia, Malaysia
Session No. 2

Role of Business Registers

HOW BIG DATA DRIVE GREATER MSBR

Abstract

In today's digital age, the data owned by Department of Statistics Malaysia (DOSM) can be integrated to provide the fuel for new statistics. Meeting users' demands on brand new economic information and focusing on reducing respondent burden are the factors that encourage the utilization of Malaysia Statistical Business Registers (MSBR). The function of MSBR is now further expanded from providing a comprehensive businesses' and companies' frame with a set of stratification variables, towards integrating it with trade database. This integration provides a basis for policy makers to explore the firms that are engaged in global markets, and what are their characteristics. This micro-data linking of MSBR and trade is to gain more data insights that able to enrich the international trade statistics by providing closer views of traders. Since big data holds great potential for revolutionizing statistics, the big data platform has been used to integrate MSBR with trade database which is known as Trade by Enterprise Characteristics (TEC). This TEC project is one of the DOSM Big Data Analytics (StatsBDA) initiatives. The project is considered as big data initiative due to the data size, data variety and data veracity that beyond the ability of current hardware and software tools to process the data within a stipulated time. The exploration of big data for this TEC project can inform the decisions about the future of international trade performance by using predictive analytics. Extracting information from MSBR and existing trade data sets can determine the patterns and predict future trade outcomes and trends. Thus, the successful of this project is very much expected in order to produce new statistics without having additional surveys.

Keywords: Statistical Business Registers, Trade by Enterprise Characteristics, Big Data

1. INTRODUCTION

In recent years there are many discussions about big data and its potential to cure ills in official statistical production. The poor timeliness, unresponsive to emerging policy needs, high costs, burden on respondents and even inaccuracies become issues in official statistics. Everyone admits that official statistics have served policy in a credible and trusted manner, but with issues highlighted above the relevance of the statistics may be limited. Thus, the National Statistical Offices (NSOs) should evolve to stay relevant in the age of big data and gain benefit as much as possible from the emergence of big data.

Big data has potential to transform traditional statistical business process towards real-time business process. This modern architecture offers real-time data processing, real-time data ingest, fast analytics on fast data, and able to churn prediction. Big data should be explored with concern on data reliability, data representativeness, privacy and confidentiality, as well as the blurring lines between formal and informal data sources.

As a trusted source of official statistics in Malaysia, meeting user demands has become the priority of Department of Statistics Malaysia (DOSM). DOSM always keep abreast with technology trends in continuing pursuit efficiency. The statistical infrastructure which supports the operation of DOSM statistical system has been organised, improved and enhanced frequently to be in line with latest technology. The initial big data development work has put focus on Statistical Business Register (SBR) since it is a vital component in the core statistical infrastructure that supports collection of economic data and production of economic statistics. It is also considered as a backbone in the production of economic statistics.

In DOSM, the function of Malaysia Statistical Business Registers (MSBR) is now further expanded from providing a comprehensive businesses' and companies' frame with a set of stratification variables, towards integrating it with trade database. This integration provides a basis for policy makers to explore the firms that are engaged in global markets, and what are their characteristics. This

micro-data linking of MSBR and trade is to gain more data insights that able to enrich the international trade statistics by providing closer views of traders.

The purpose of the paper is to highlight how big data drives greater MSBR in the production of Trade by Enterprise Characteristics (TEC) statistics. The next section, Section 2, explains the TEC initiative in big data platform which includes methodology as well as TEC's output. Section 3 presents the potential predictive analytics from TEC and finally, the paper ends with a short conclusion.

2. TEC IN BIG DATA PLATFORM

Trade by enterprise characteristics takes a look at international trade statistics from a very specific point of view i.e. the characteristics of the enterprises actively engaged in exporting and importing. Traditional trade statistics record what types of goods are trading across borders between countries but they do not describe the characteristics of the enterprises that are behind of these trade flows. In order to know the actor actually engaged in cross border trade, trade data should be linked to the information of enterprises. This identification information can be obtained from SBR, such as name and address, main economic activity of businesses, type of ownership, employment size class, turnover etc. The linkage of trade statistics with business registers allows describing those who are engaged in global market, and what are their characteristics. On the other hand, TEC is to complement business data with detailed information on trade.

Due to the volume of trade data is generated at increasing rates and highly unstructured; DOSM decides to harness the massive amounts of trade data with big data platform. The integration of MSBR and trade database is beyond the ability of DOSM current hardware and software tools to process the data within a stipulated time. Thus, this TEC project has become one of the DOSM Big Data Analytics (StatsBDA) initiatives. The StatsBDA proposed architecture is shown in Chart 1.

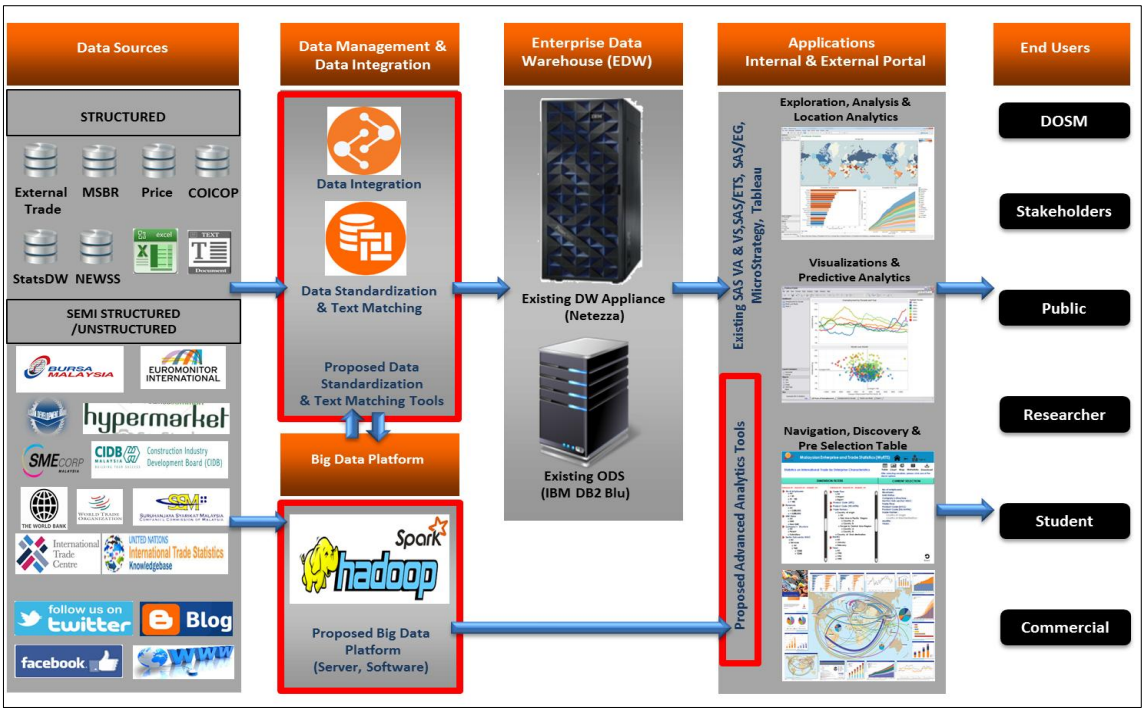

Chart 1: StatsBDA Proposed Architecture

2.1 MODERNIZE TEC SOLUTION

The identification of enterprise entities in both MSBR and Customs declaration is business registration number which is maintained by Companies Commission of Malaysia (CCM). CCM is a statutory body to serve as an agency to incorporate companies and register businesses as well as to regulate companies and businesses in Malaysia. This business registration number has been used as unique identifier for Malaysia's TEC project since it is the very reliable matching approach for the project is business registration number.

However, the free text fields in MSBR and Customs declaration forms; in order to provide flexibility during data entry as well as to facilitate trading activities, has tend to cause data quality problems particularly business registration number. For that reason, the matching technique is further improved by string matching algorithm approach. Instead of merely used business registration number in matching process, the businesses registered or trading name is used. Therefore, the element of data management has been considered in the development of StatsBDA

architecture which includes data profiling, data standardization, data clustering, data cleansing, text matching and data integration. The data management workflow is shown in Chart 2.

Chart 2: Data Management Workflow

The initial step is data profiling where the data available in MSBR and trade database is examined and analysed in a systematic process to come out with a brief and informative summary about that data. The purpose of the process is to evaluate the data with a methodical, repeatable, consistent, and metrics-based since the data is dynamic in nature especially trade data. The high correlation between known errors in the data can be detected once profiling effort is performed. For instance, the business registration number in Custom forms may be represented forwarding agent instead of exporters or importers. Thus, the proactive action can be done by investigating the declaration pattern done by a particular forwarding agent. In normal case, the traders usually employ the same forwarding agents to use the service.

The next step will be data standardizing to ensure data consistency (using common format) and clear (easily understood by those who are not involved with the data maintenance process). The standardization of companies' or businesses' name in MSBR is the critical process of bringing data into a common format to allow for string matching algorithm approach. In several cases, the data may contain several versions of similar word like 'Sdn. Bhd.' at the end of companies' name. 'Sdn. Bhd.' may appear in the data as 'SB', 'S/B', 'Sendirian Berhad' and 'Sdn Bhd'. By standardizing the output of the 'Sdn. Bhd.' rule, one can decide the standardization of these words to be 'Sdn. Bhd.', regardless whether it appeared as 'SB', 'S/B', 'Sendirian Berhad' and 'Sdn Bhd' in the original or input data. The same process is performed on trade data where the data is more diverse. The

standardized data does not overwrite the original data in the database; therefore the record with original format can be viewed.

There is a growing need for an effective approach to do data clustering due to multiple records of a similar establishment in MSBR (an establishment with multiple activities/ MSIC). The similar establishments will be clustered together to generate unique establishment record before matching process with trade data is done. The clustering process is also conducted to trade data. The data is partitions into groups based on similarity of exporters/ importers' name. It is a discovery tool that reveals associations, patterns, relationships, and structures in masses of trade data. The relationship between exporters/ importers and forwarding agents can be identified and the declaration patterns done by the agents can be established.

Subsequently, data cleansing which involved the correction of data content will be carried out when it falls below the accepted standards. The process includes of amending or removing data in a database that is incorrect, incomplete, improperly formatted, or duplicated within the cluster. Any data flaws will be systematically examine by using rules, algorithms and look-up tables. In the case of missing business registration number in both databases, the data cleansing tools are capable to do the correction by adding the missing number into the records. The approach can save a significant amount of time than fixing errors manually.

The final step will be data integration where the combination of MSBR and trade data can create the meaningful and valuable information. Thus, the integration process will involve text or string matching by finding strings in exporters' and importers' name that exactly or approximately match with business registration names in MSBR. The method is intended to find a 'closeness' score between the search string in trade database and the text in MSBR rather than a 'match' or 'non-match'.

2.2 MSBR QUALITY IMPROVEMENT WITH BIG DATA

Big data platform has paved a way to examines large amounts of trade data to discover hidden patterns, correlations and other insights. With big data's technology, it is possible to analyze and integrate the trade data with MSBR, in which the effort must be slower and less efficient if it is done with traditional business intelligence solutions.

Data integration of MSBR and trade database involves combining data from two different sources, which are stored using different technologies. The integration provides a unified view of the data where TEC can provide new information that would not exist in stand-alone statistical domains. The integrated datasets can indicate which enterprises are engaged in international trade as part of global value chains and measure the importance of those firms in the overall economy.

This TEC project with big data solution has provided an opportunity to improve the MSBR quality. Due to explosion in the number of administrative data sources in maintaining the national statistical register; the updating process can be done in short time with big data platform. Generally, the administrative data source comes from various agencies with various format and the data are collected for the various non-statistical purposes. As such, the records have a variety key identifier that maintain by different agencies. Advancements in big data technology will permit DOSM to overcome many limitations caused by processing large datasets with variety format and layout. Data quality solution offers by big data platform enable DOSM to do profiling, standardizing, clustering, cleansing, text matching, integrating and analyzing the data more quickly than before. Chart 3 shows the proposed TEC architecture in which data quality has become the focal point of TEC initiative.

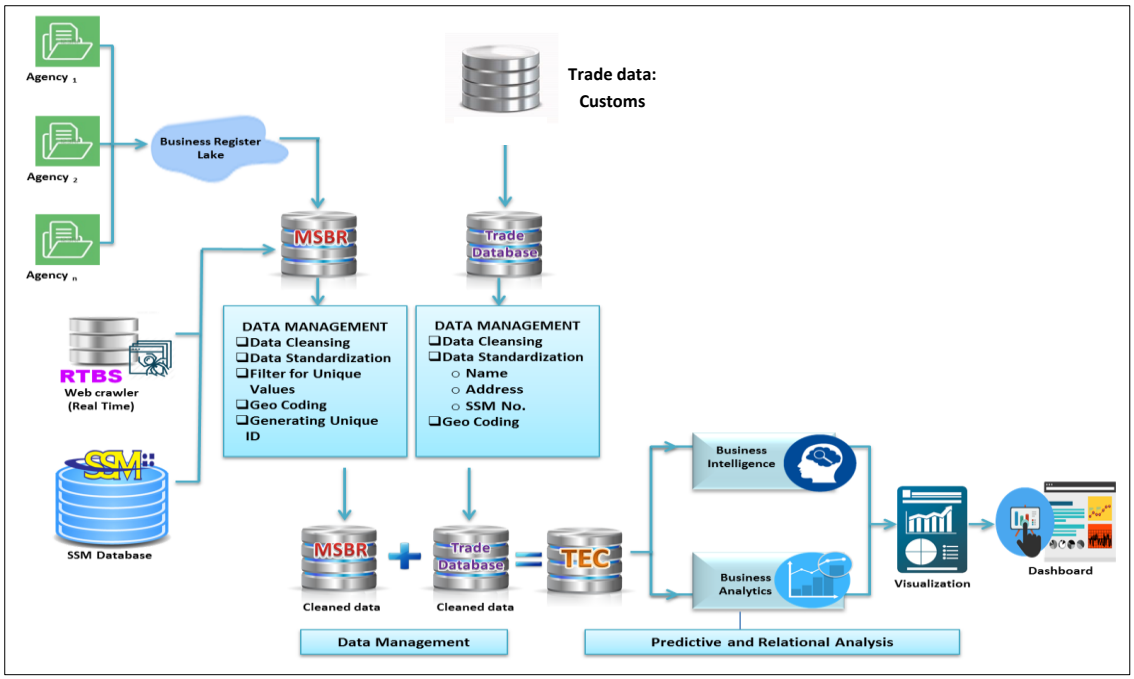

Chart 3: Proposed TEC Architecture

Before integrating MSBR with trade data, the MSBR need to be cleansed updated to ensure it portrays the current operation status of companies or businesses in Malaysia. Thus, all data management approach needs to be applied during MSBR updating process. In short, the development of TEC project in big data platform is concurrently assists DOSM to improve MSBR data quality.

3. TEC BUSINESS INTELLIGENCE & ANALYTICS

According to Pat Roche, Vice President of Engineering, Magnitude Software, business intelligence is needed to run the business while business analytics are needed to change the business. In TEC initiative, business intelligence represents a technology that transforms MSBR and trade data into more meaningful and useful information for analysis purposes to support better decision making.

The exploration of big data for this initiative can inform the decisions about the future of international trade performance as well by using predictive analytics. Extracting information from MSBR and trade data sets can determine the

patterns and predict future trade outcomes and trends. For instance, the sustainable of a particular product or exporter or importer for the next 5 years in global market can be identified by predictive analytics. It is also can help the government to discover industries that will emerge, evolve or disappear in international market. On top of that, predictive analytics can pinpoint the countries throughout the world that will create demand for Malaysian specific products. The information will assist government to formulate relevant policy by making better, smarter investment decisions and making them faster.

The big data technology automatically highlights relevant findings and significant discoveries by visually exploring all relevant TEC data, spot unknown patterns, identify key relationships and unearth hidden opportunities. The attractive visualizations will help DOSM's officers to quickly grasp what is the data all about. The data visualizations manipulate complex pools of data to visually display the data's patterns, trends, and correlations. It is the impactful ways for data analysts and scientists to communicate the findings through data visualizations.

4. CONCLUSION

In the nutshell, big data is a revolution that will transform how DOSM works in order to strengthen its' statistical delivery system. Since MSBR is an important DOSM's asset, the data quality is a highly significant consideration. Poor decision normally comes from 'garbage-in, garbage-out' data quality and poor data quality can ruin analytics in DOSM. With emerging trends of big data, DOSM takes the opportunity to develop the TEC project with big data platform. Instead of 'promised' potential benefits and TEC predictive analytics, data quality management offered by big data really helps DOSM to enhance the quality of MSBR. The appropriate steps to resolve data quality issues prior to integrating MSBR with trade data are extremely important. Without taking this step, the output can negatively impact the analytical systems in DOSM. Simply stated, the StatsBDA initiative able to drive greater MSBR.