2020 Asia–Pacific Statistics Week
A decade of action for the 2030 Agenda: Statistics that leaves no one and nowhere behind
15-19 JUNE 2020 | Bangkok, Thailand

# Statistical Business Process for Big Data Usage

Isnaeni Noviyanti*; Dwi Puspita Sari[1]; M. Hanif Fahyuananto[2]; Ade Koswara[3]

* BPS-Statistics Indonesia; isnaeni@bps.go.id
[1] BPS-Statistics Indonesia; dwi.sari@bps.go.id
[2] BPS-Statistics Indonesia; m.hanif.fahyuananto@bps.go.id
[3] BPS-Statistics Indonesia; adekoswara@bps.go.id

**Abstract:**

Nowadays, the utilization of big data as a new source to complement official statistics has become an opportunity for organizations focusing in statistics. The use of big data can lead to a more efficient data collection. Big data can be obtained through various sources such as social networks, traditional business systems, and internet of things (IoT). Up to now, there has not been any firm business process for big data collection and processing. Whereas, in order to fully take advantage of big data, maximum support is needed. To be specific, a firm procedure regarding the specified needs of data, data collection, processing, until dissemination, as well as the technology needed to run the process should be determined in a clear scheme.

Since 2017, BPS-Statistics Indonesia has initiated the implementation of big data platform in official statistics. There has been a preliminary study on its usage and also the technology itself. Some of the technologies that BPS owns are Hadoop, data warehouse, also Extract-Transform-Load (ETL) and exploration tool. However, technology adoptions alone cannot determine the success of big data utilization. It is widely known that big data utilization can be challenging, since there are issues regarding data access and quality, also the required skillsets. A firm business process can help related stakeholder understand how to overcome those challenges and ensure the use of big data can beneficially support official statistics.

This paper will propose a new business process specifically designed for big data usage, and also explain how the existing technology in BPS can be utilized to support it. The study for business process and technology related to big data usage is based on Generic Statistical Business Process Model (GSBPM) framework. With this proposed business process for big data and the support from big data platform architecture, hopefully statistical organization can be one-step ahead on utilizing big data and overcome the challenges that come with it.

**Keywords:** GSBPM; big data utilization; big data challenges; big data platform; Hadoop

## 1. Introduction:

According to law of Indonesia Number 16 year 1996 about statistics, BPS perform several ways of data collection, some of which are census, survey, administrative data compilation, and other ways in relevance to the development of science and technology. To date, data collection done through census, survey, and administrative data compilation have been regularly conducted. Whilst data collection using method relevance to the development of science and technology has started to be explored further, especially the utilization of big data as a new data source to support official statistics.

The use of big data is expected to reduce the time needed to deliver official statistics to public, respondent burden, and budget for data collection. However, big data doesn't only offer opportunities but also issues and challenges. In big data related project that has been conducted by several National Statistic Office (NSO), the persistence issues are related to data quality and accessibility, along with the required expertise (Cornelia, et.al, 2017).

BPS has designed Statistical Business Process Framework (SBFA) based on GSBPM as a basis for transformation. According to SBFA, there are eight main processes from specify needs up to evaluate (BPS, 2016). The utilization of big data as a new data sources for official statistics will surely need a formal business process as a guideline to ensure a more directed and organized process. Each phases in

the business process undoubtedly would need IT support. Beside its beneficial for process automation, IT can also support data governance, which in turn would minimalize data silo and realize one data policy in BPS. Thus, BPS initiates the development of Data Management System (DMS), which functions as single source of truth and big data platform for BPS.

This paper aims to explain the business process specifically designed to support the utilization of big data in official statistics, along with how the existing technology in BPS will support it. This paper is divided into several sections, section II will describe the methodology used in designing the business process, section III will explain the result, and section IV will discuss conclusions and recommendations.

## 2. Methodology:

Several steps are performed to design business process and technology support for the utilization which are identifying the problem, designing the big data architecture, designing road map implementation, also evaluation and improvement. Business process and supporting technology stated in SBFA is used as an input during literature study and business process design.

## 3. Result:

### A. Current Condition in BPS

One of the statistics transformation principle stated in SBFA is the use of paperless data collection. This principle combined with modernization strategy to conduct better acquisition and use of technology become a strong reason to enrich big data utilization. Since 2016, BPS has performed various preliminary studies and researches regarding the utilization of big data to support production of official statistics. Below are several researches and implementations of big data in BPS:

*Table 1. Big Data Utilization in BPS*

| No. | Big Data Source | Supported Official Statistics |
|---|---|---|
| 1. | Mobile Positioning Data (MPD) from major Mobile Network Operator (MNO) | 1. Tourism Statistics, Calculating Arrival of International Visitor through border<br>2. Migration Statistics, Measuring commuting pattern in major cities |
| 2. | Web scraping from accommodation websites | Tourism Statistics, Estimating Accommodation Occupancy Rate |
| 3. | Web scraping from market places | Price Statistics, Estimating Consumer Price Index (CPI) |
| 4. | Web scraping form Google trends and twitter | Employment Statistics, Estimating Unemployment Rate |

Those researches and implementations gave clear evidences that the use of big data as a data source indeed changes several aspect in statistical business process, starting from design until analysis and even dissemination phase. The most significance change is encountered in collect phase, where data collection has been proven to be more efficient in reducing the needed time, effort, and budget. This method can also ease the respondent burden. However, along with the advantages come the issues and challenges, especially regarding data accessibility, coverage, and quality. One of the research also mentioned the importance of capacity building to improve the required skillset, as BPS is relatively new and yet to be familiar with big data usage. Another issue is the process of utilizing big data has yet to be defined in a clear scheme. Each implementation can use unstandardized business process, which vary from one another.

The different business processes, which is not properly documented, would make finding solution for the existing challenges much more difficult. This problem needs immediate action, or else development of big data utilization would be slowed down when in the contrary it should be hastened in this digital era. Realizing this, BPS has started seriously design a framework for big data utilization, including the architecture, the business process, and the supporting technology platform.

In technology aspect, previous and current implementations often use different technologies causing data to scatter in various repository. The evaluation of these implementations suggests that a proper technology is needed to ensure an optimal big data processing and eliminate data silo by storing it in one place. To overcome this, BPS has started the initiative to develop big data platform in 2017. Below are the development phase:


*Figure 2. Big Data Platform Development Phase in BPS*

## B. Design of Big Data Architecture

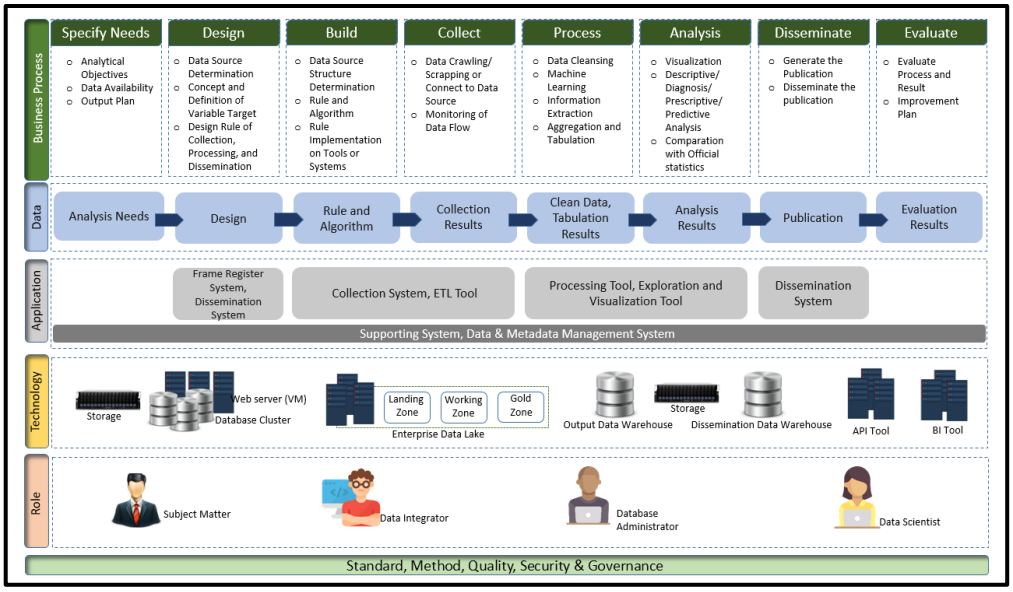


*Figure 3. Design of Big Data Architecture*

The proposed big data architecture, as shown in Figure 2, consists of four main layers, which are business, data, application, and technology. The rest is supporting layer that ensures the main layer would work smoothly. The explanation for each layers is as follows:

1) Business Layer

Business layer is the first thing that needs to be designed because this layer will decide the data flow, along with the applications and technologies required to improve business value. Considering the huge amount of potential data to be explored, some might prefer to collect as much data as possible without carefully consider the actual output they need. However, it would be wise to perform preliminary literature study and assessment to obtain clear objectives for data analysis. As more data does not necessarily mean more insights (Macfeely, 2018). When the objectives is clear, the needed data and expected output will be easily defined. According to UN-SQAF, big data does not need to be treated differently from other potential data sources. Big data potential as data source in producing statistics needs to be evaluated just like any other

data sources through assessment whether the collected data met the requirement and fit for purpose, and if it does, how to obtain and process them (UN, 2018).

In the proposed architecture, the existing processes are open to iteration, meaning it is possible to roll back into previous process when adjustment is deemed necessary. In big data utilization, it is important to consider whether the result would be published as official statistics, supporting phenomena, or even a mere research paper. As to publish an official statistics, a lot more things need to be properly validated, especially in methodology. Finally yet importantly, quality indicator needs to be defined to help monitoring the success in each phase.

2) Data Layer

Data layer represents data components that need to be considered in each process. Figure 2 shows that data layer encompasses main output in each process, which if to be detailed will result in many entities and attributes. For example, process phase requires data structure modelling for the expected result from data cleansing and information extraction to produce relevant clean data and tabulation. Data need to be linked to its metadata to help user understanding data. Across all business process, there would be input/output flow from/to data and metadata management systems–which will be referred as DMS and MMS.

3) Application Layer

Application layer describes the supporting system in each processes. How each systems contributes in business process is shown in Figure 3. Among all those systems, DMS and MMS will be an enabler systems that provide an over-arching support throughout all processes. Application Programming Interface (API) is also used every now and then as a main means of communication for all systems.

4) Technology Layer

Technology layer describes the infrastructure used in big data ecosystem. Figure 2 shows that enterprise data lake (EDL) component which contains several nodes is used to store and process large data. There are also database clusters, data warehouse, and other kind of storages each aimed for certain purposes.

5) Role

An appropriate division of role is very vital to ensure business process run smoothly. In utilizing big data, BPS takes the role of data consumer and if needed could coordinate with data provider and other NSOs who have conducted big data project before (Lestari, et.al., 2019). Beside that, each role must communicate with each other so that the result would be in concordance with the predefined output. There are several proposed main roles, which are Subject Matter, Data Integrator, Database Administrator, and Data Scientist. How each of this roles will contribute in the business process can be seen in Figure 3

6) Standard, Method, Quality, Security, and Governance
   a) Standard – that is the statistical standard implemented in BPS to ensure the comparability of the produced output, be it nationally or internationally.
   b) Method – that is the statistical method used in design phase and its implementation in data collection, processing, and analysis. For big data, there are several underlying issues, such as how to create a general and replicable model. With the uncertainty factor in big data, repetitive simulations is required to evaluate the model (Bühlmann, et.al., 2018).
   c) Quality – that is mechanism to ensure the quality of each performed business processes' outputs. Even if the big data project is a mere exploration, quality aspect should always be taken into account for optimal result. Each of NSO can develop their own Statistical Quality Assurance Framework (SCAF) as guideline for producing better quality statistics.
   d) Governance – that is guidelines consisting of policy and procedure that will be implemented throughout the running processes. This guidelines could be broken-down into technical details and later on applied in the application and technology that is being used. For example, data that will be collected must have and be linked to metadata.

## C. Design of Business Process and Infrastructure for Big Data Usage

The proposed business process along with the related roles and systems in each process can be seen in Figure 2. The design suggests that each process will produce metadata that later will be stored in MMS and used for the next process. With a well-documented metadata, its reusability may be possible. This will help divisions in BPS to learn and recreate previous big data project, be it the analysis purpose, rules and algorithms, collected variable, or even analysis result and its evaluation.

Figure 4 shows the process flow in DMS as *single source of truth* dan *big data platform*. Data source can be obtained through administrative product, census, survey, external web, and even internal systems. All data will be stored in working database and later then loaded into EDL. There several zones in EDL, which are landing (for newly loaded data with incomplete metadata), working (for data processing in EDL), and gold zone (for processing result and storing data with complete metadata). Data in gold zone will then be loaded into data warehouse for further access by internal system and dissemination system via API. BI Tool can then access data from EDL gold zone and data warehouse.

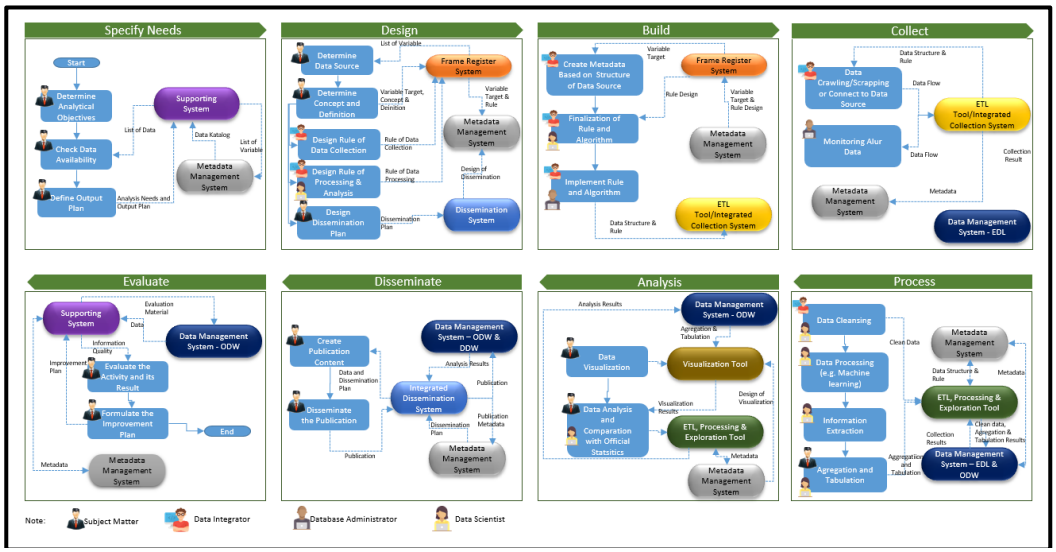


*Figure 4. Process Flow for Big Data Usage*

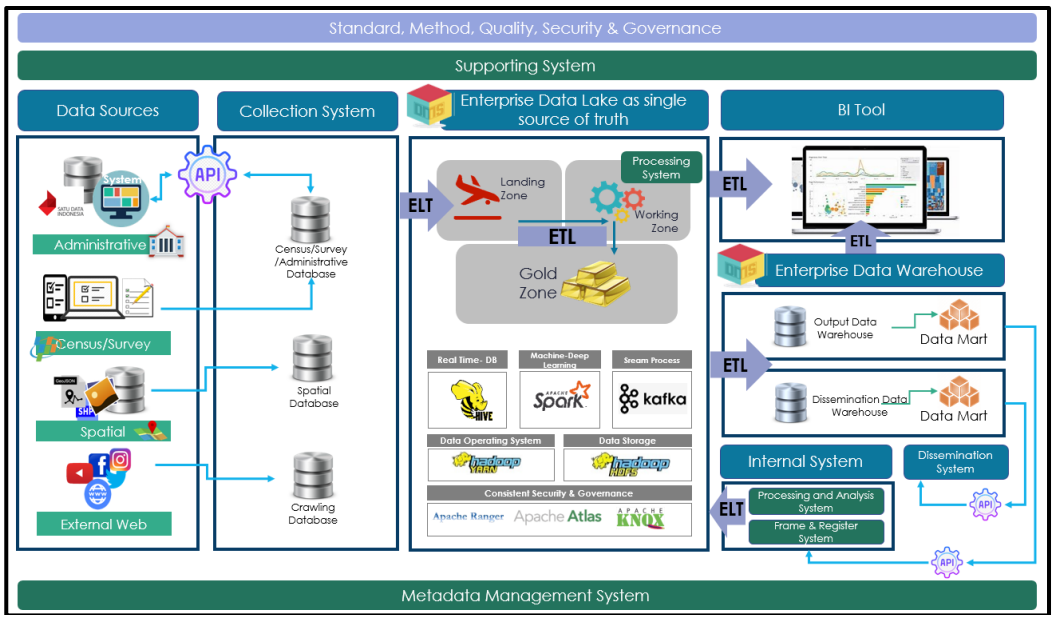


*Figure 5. Design of Big Data Platform*

### D. Road Map of Implementation

Below are the proposed roadmap for future big data implementation in BPS.

*Table 2. Roadmap for Big Data Implementation in BPS*

| Kegiatan | 2020 | 2021 | 2022 | 2023 | 2024 | Prerequisites |
|---|---|---|---|---|---|---|
| Forming a formal big data team | | | | | | A formally formed and legalized team dedicated for big data utilization |
| Conducting Pilot Project for Big Data Implementation | | | | | | Ideas for big data project using the available DMS components |
| Preparing DMS' infrastructure gradually | | | | | | Installed and ready-to-use hardware and software |
| Integrating DMS with related systems | | | | | | Partial integration has been performed since design and development |
| Enforcing governance regarding big data utilization and its platform (DMS) | | | | | | A living document draft for governance |
| Implementing DMS' infrastructure on BPS' business process | | | | | | A successful change management beforehand |
| Migrating data into DMS | | | | | | Initial migration since DMS component for storing data available |

## 4. Discussion, Conclusion and Recommendations:

The proposed business process is still a high-level design and will need adjustment according to each cases and NSOs characteristics. The implemented technology will surely help business process run more smoothly, especially during big data collecting and processing. The proposed business process and big data platform is expected to help overcoming the challenges of big data in the following ways:

1. Regarding quality, the use of metadata in each phase will serve as lesson learned which can help improvement in future projects. Besides that, by applying quality indicator, a firm business process, and governance, each processes and its output quality will be more controlled.
2. Regarding accessibility, after defining the possible data sources and its expected output, identifying the possible issues during specify needs and mitigating the risk, along with considering data coverage and confidentiality, during design phase will be easier. The supporting collection system can also help dealing with technical accessibility issue.
3. Regarding skillset, with clear division of role it would be easier to identify the required skillsets then form and train a team to meet the requirement.

Big data as a new data source doesn't necessarily need different treatment from other data sources, although each processes still need to be performed in caution to ensure data accessibility and quality. Like in census and survey, a firm business process is needed as a guideline in utilizing big data, starting from specifying needs up to evaluation. In the end, applying the right governance should be utmost concern to ensure the collected data won't turn into data swamp.

**References:**

1. Badan Pusat Statistik, (2016). Statistical Business Framework and Architecture version 4.5.
2. Bühlmann, P., Van de Geer, S., (2018). Statistics for Big Data: A Perspective, Statistics and Probability Letters vol. 136 pp: 37-41.
3. Macfeely, S., (2018). Big Data and Official Statistics, The United Nations Conference on Trade and Development (UNCTAD), Switzerland, pp: 25-54.
4. Lestari, T.K., Esko, S., (2019). Lessons for Effective Public-Private Partnerships (PPPs) from the Use of Mobile Phone Data in Indonesian Tourism Statistics, Asia-Pacific Economic Statistics Week 2019.
5. L., Cornelia, et.al, (2017). Big Data: Potential, Challenges, and Statistical Implications, IMF Staff Discussion Note.
6. United Nations, (2018). United Nations Statistical Quality Assurance Framework.