**Prepare your data warehouse for a Big Future, by including Big Data**

Gianpiero Bianchi, , Istituto Nazionale di Statistica, gianbia@istat.it

Antonio Laureti Palma[[1]](#footnote-1), Istituto Nazionale di Statistica, lauretip@istat.it

Sonia Quaresma1, Instituto Nacional de Estatistica, sonia.quaresma@ine.pt

**Abstract**

*National Statistical Institutes produce statistics based on consolidated statistical models according to the study-domain characteristics. Conversely, nowadays, Big Data (BD) opportunities force NSIs to deliver statistical products also based on a sequence of data analytic processing. In this way, NSIs must be able to combine different areas of expertise for the mapping of analytical constructs with statistical concepts. This is an epistemological process change that forces NSIs to meet some new basic requirements at the procedural, organizational and infrastructural levels. At the process level, the use of BD introduces a paradigm shift, typical of data mining, from theory to data driven models. At the Organizational level, it is necessary to guarantee and support the active participation of multidisciplinary experts in the process of knowledge extraction. At the infrastructural level, new and complex infrastructures are needed to support both analytical tools and multidimensional analysis.*

*In this work we aim to present a new generation S-DWH architecture able to sustain the required epistemological process change and facilitate a multidisciplinary approach. Two study cases of official statistical production processes involving Big Data sources will be discussed.*

**Keywords:** official statistic production; statistical data warehouse; big data; data analytics;

**1. Introduction**

The Big Data, originating from the digital breadcrumbs of human activities, sensed as a by-product of the technologies that we use for our daily activities, lets us observe the individual and collective behaviour of people with an unprecedented detail.

The generally accepted definition of Big Data (BD) are data sets, whose size or type, are beyond the ability of traditional relational databases to capture, manage, and process the data with low-latency. These data are generated on a very large scale from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media.

NSIs should carefully evaluate the impact and new opportunity that BD sources offer since this could revolutionize the way NSIs produce statistics, in order to improve official statistics quality and reduce costs.

Data Mining (DM) these BD allows analysts, researchers, and business users to produce better and faster decisions by applying advanced analytical techniques. In terms of IT, an analytical process could involve technics of Machine Learning (ML), which embodies the principles of DM, to enables [computer systems](https://en.wikipedia.org/wiki/Computer_systems) to learn behaviour from data in order to take autonomous decisions. The choice of an ML algorithm to apply largely depends on the user’s domain knowledge, the desired results and on the performance of computing platform.

From this perspective is evident that for implementing new official statistical procedures multidisciplinary approaches are crucial. Different experts must work together in a common data infrastructure and must be able to combine theoretical concepts with different areas of expertise for the mapping of analytical constructs.

This is an epistemological process change that forces NSIs to meet some basic requirements at the procedural, organizational and infrastructural levels. At the process level, the use of BD introduces a paradigm shift, typical of data mining [1], from theory to data driven models. At the Organizational level, it is necessary to guarantee and support the active participation of multidisciplinary experts [2] in the process of knowledge extraction. At the infrastructural level, new and complex data infrastructures are needed to support both analytical tools and multidimensional analysis.

In a NSI the data infrastructures should be the corporate Statistical Data Warehouse (S-DWH), which aims to support different users as well as multidisciplinary statisticians’ workflows to extract value from data [3].

Exploiting BD in a S-DWH means then dealing with new data types, new volumes, new data-quality levels and new performances. To this purpose, it is necessary to evolve from the classical S- DWH based on relational DBs to a new generation of S-DWH able to manage large amounts of data. This forces NSIs to adopt new data architecture models and new distributed-computing frameworks in a common performing statistical methodological frame.

In the current paper we show the impact and challenges that the BD sources present to NSIs and to their truth repositories, the S-DWH of new generation. In order to do that, we will first face the epistemological process change through the BD analytic process for official statistics production in chapter 2. This chapter illustrates why BD sources pose at the same time a challenge and offer new opportunities to official statistics. The answer to where the activities can and should take place is introduced in chapter 3 where the new generation of corporate S-DWH is shown. Finally chapters 4 and 5 illustrate how it can be performed through the description of two study cases based on text and sensor BD sources.

**2. Big Data in Official Statistics**

Big Data is a field dedicated to the analysis, processing, and storage of large collections of data that frequently originate from disparate sources. Specifically, Big Data addresses distinct requirements, such as the combining of multiple unrelated datasets, processing of large amounts of unstructured data and harvesting of hidden information in a time-sensitive manner. For example, BD can provide snapshots of the well-being of populations at high frequency, high degrees of granularity, and from a wide range of angles, narrowing both time and knowledge gaps.

Official statistics, meaning produced in the traditional way, will continue to generate relevant information, but the BD revolution presents a tremendous opportunity to gain richer, deeper insights into human experience that can complement the development of indicators that have already been collected.

From this point of view, the statistical community has recognized the potential for big data in improving accuracy and reducing costs for official statistics. BD is considered very interesting as an input for official statistics; either for use on its own, or in combination with more traditional data sources such as sample surveys and administrative registers. In 2014 Eurostat, EU Member States and EFTA countries launched the ESS Vision 2020 as a strategic response to the challenges that official statistics face. In Vision 2020 one of the key areas is the “Harness new data sources”, in which BD is central. Particular attention is paid to the potential use of BD with the aim of reducing the costs of statistical production, extending the range of statistical products and increasing the timeliness of official statistics [4].

Several factors have facilitated the advances in the use of big data, the most important of which are: advances in information technology that have reduced the costs of data collection, storage, and processing. Some examples of BD sources with an impact on official statistics production are: mobile phone data for Social Statistics or Tourism; web scraped enterprise data for Business Statistics; use of scanner data in the context of Consumer Price Index statistics; smart meters data for Energy Statistics, Housing Census or Household Costs.

*2.1. Big Data Analytics process*

Data analytics is a broader term that encompasses data analysis and includes the management of the complete data lifecycle. The definition of BD Analytics process is then advanced data analytics processes which operate on BD sets.

The implementation of BD analytics process is required by analysts, researchers and business users for investigating new high quality models or new usable information elements. Methodologically, in statistics, a BD analytics process introduces a epistemological change in the design of new statistical production processes.

In a traditional statistical production environment the data is first analysed to create a set of requirements which leads to a data model creation to load and process the data. On the contrary, when BD is involved, the dataflow requires that data is first loaded and then data structure is created to process data.

In the data life cycle of an analytical process is possible to distinguish three basic stages: *collect*; *discovery*; *analyse*. In the first stages, data is gathered and loaded into specific data store where a first structure is applied. Typically, data is gathered from different sources, like web sources (web scraping), accounting system (such as CDR of telecom networks) or from an API (like a social network application).

In the *discovery stage* the BD can be tagged and categorized to get basic insights for the first time and is converted to metrics. This stage can become complex to manage since data mining skills and high performing computational platforms are needed.

In the *analysis stage* knowledgeable users can extract the result sets from any elaboration step and use them alone or integrated with other conventional sources.

*2.2. Big Data Analytic engines*

Engineering a statistical production process which includes BD sources basically means including ML techniques to guarantee their automation. Moreover, in order to increase the level of automatization, which reduces the risk related to uncontrolled data sources, it is useful to define a calibration phase of ML algorithms that help set the typical algorithmic parameters required for processing.

In general, ML algorithms embody the principles of DM aspiring to emulate the human brain functions by closely studying available data, to arrive at conclusions or make future predictions. There are many different kinds of machine learning algorithms for discover patterns in big data that lead to actionable insights. At a high level, these different algorithms can be classified into three groups based on the way they “learn” about data to make predictions: *supervised learning*; *unsupervised learning*; *semi-supervised learning*. Supervised learning is when all data is labelled and the algorithms learn to predict the output from the input data. Unsupervised machine learning is more closely aligned with the idea that a computer can learn to identify complex processes and patterns without a human to provide guidance along the way. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. Finally, many commonly used machine learning algorithms actually fall into the category of semi-supervised learning where only some of the data is labelled.

Especially in the ML case, the quality of the model depends on the quality of the training data. In cases of supervised classification problems, a calibration phase strategy is to separate the training set into two parts, of which one is considered unknown [6]. This is called the test set and is used to evaluate the performance of the learned classifier. An improved version of this procedure is known as cross-validation. The cross-validation procedure can prevent the overfitting problem. Furthermore, a “grid-search” on algorithmic parameters using cross-validation is recommended in order to identify the best parameters.

**3. The corporate Statistical Data Warehouse**

In a NSI, where data and information are the core business, BD should be contained directly in the corporate S-DWH, which is the central data repository of all statistical data. Through an efficient S-DWH statisticians could manage their data in the different production phases at micro and macro granularity levels.

As central data repository, a S-DWH must support users to intensive use and re-use all available data sources, from surveys and administrative data to BD, facilitate the integration of data and processes and any new possible analytics activities, in order to maximize analyses and produce necessary statistics.

Typically, S-DWH users’ are experts, knowledgeable users, involved in designing of production processes aimed at creating data inter-relationships, which connect different data from common or different statistical domains. From this point of view, a production process can be seen as a unique procedure articulated in different basic-phases, from data collection to data dissemination, where each basic-phase collects input data and loads output data in the shared central data repository. Each basic-phase consists in one or more sub-process transformations which take place through an asynchronous process. In this way, an analytical process is a workflow of the separate activities performed by the different multidisciplinary experts involved.

Therefore, in building the new S-DWH model, we need to adopt a radical thinking like the pioneer data warehouse architects of the yore, where we will retain the fundamental definition of the data warehouse as stated by Bill Inmon, but we will be developing an architecture model that will not be constrained by the boundaries of a single technological platform. The reason to this is the high level of flexibility required in data processing, where the analytical process became the most critical requirement, and for solving the data integration challenge between the traditional data warehouse and the new BDs, leveraging all the investments on the current infrastructure.

To this aim a new S-DWH based on distributed systems and data virtualization was deployed. This has provided the opportunity to maintain a distributed workload effectively across the platforms adopting a scalable and flexible architecture, reusing existing infrastructures.

As schematized below in figure 1, the new S-DWH has been based on four logical layers where each level perform specific functions: *source layer*, *integration layer*, *interpretation layer* and *access layer*.

The *source layer* is the layer where all the activities related to storing and managing raw data sources are carried out. Considering BD sources, the source layer supports all activities related with the collect stage of an analytical process. This means gathering data from different DB sources and data loading into specialized distributed file system. The loading process can be seen as the breakdown of a large input into small chunks of files.

The *integration layer* is the layer where all main editing activities needed for the statistical production process are carried out. Basically, in the integration layer the new BD source is transformed and categorized in order to be made integrable. In this layer are carried out two basic analytical activity types: the initial discovery stage, aimed at identifying the best suitable data transformation, and its engineering.

The *interpretation layer* is the logical data store in which are organized all corporate information for the internal users, domain experts and methodologists. In this layer is made available the full corporate data warehouse, primary data and auxiliary sources, raw data in its native format including semi-structured data. To include semi-structured data in a data warehouse appears contradictory but it is a basic requirement of the new generation S-DWH. It comprises both predictive and prescriptive analytics to facilitate the development of new opportunities eliminating the need for the users to have to connect to multiple data bases.

To this aim, next to legacy data warehouse, the interpretation layer must be based on distributed data base system. These are [non-relational databases](https://en.wikipedia.org/wiki/Non-relational_database) that ensure a quick access to data over a large number of nodes, or [computer network](https://en.wikipedia.org/wiki/Computer_network). It allows a high level of computational performance by a computing distributed framework without the need for prior modelling and data preparation.

The *access layer* is the highest functional layer, designed to deliver statistical outputs to selected internal viewers and external users. Internal users can realize the last analysis stage of an structured analytical process or start the design of any new cross domain analytical process. External users can have access to controlled functionalities fundamentally related to the exercise of output systems.

 **Figure 1. Overview of the layered S-DWH architecture**

At the base of a S-DWH virtualization there is the implementation of two interfaces: a *virtualization interface* and an *abstraction interface*. The virtualization *interface* is at the bottom of the *access layer* and performs the function of the interface toward the interpretation layer. This allows access to all corporate data, in different contexts as well as the creation of new contexts.

The *abstraction interface* allows a conceptual generalization of the sources and of the integration outputs. From a logical point of view, the abstraction interface is on the bottom of the interpretation layer and acts as data interface towards the integration layer. The abstraction interface is based on a semantic middleware which supports context integration between multiple data sources and their relative metadata. The semantic middleware is based on a lexical interpreter and a semantic mapper. The lexical interpreter includes processing tokens and streams of text. This comprises a taxonomy definition process across different data context domains and internal dictionaries. The semantic mapper provides a schema for data visualization based on the integration of different data contexts. This includes a tag schema designed to maintain a relationship model between different data elements. In particular, data from different domains is paired and integrated on the bases of their semantics.

**4. Case study: population statistics from mobile phone traffic**

The case study focuses on an ISTAT project, called “Persons and Places” which produced a statistic evaluation comparing two approaches to mobility profile estimation, namely: estimation based on mobile phone data and estimation based on administrative archives. Mobile phones today represent an important source of information for studying people’s behaviour, for environmental monitoring, transportation, social networks and business.

The project compared mobile phone data with the origin/destination matrix of daily mobility at the municipality level [5]. Calling data was obtained from mobile phone CDRs (Call Detail Record). These are BD generated from voice connections and used in the billing system for charging purposes. We noticed that a major limitation of CDRs is that the localization of individuals occurs only during phone calls, which can lead to an incomplete view of their mobility or an underestimation of the real world.

The basic idea followed in this study case is that the user category of an individual within a specific municipality can be inferred by the temporal distribution of his/her presence in the area.

In order to process CDRs in the S-DWH environment we identify the following analytical process. In the ***source layer***, CDRs are loaded and structured into the distributed database. After this stage data are directly usable by advanced algorithms as well as by SQL procedures.

In the ***integration layer***, the data discovery stage is realized through an unsupervised learning procedure. At first, the Individual Call Profile (ICP) is built. This is a set of aggregated user’s calls computed by applying spatial and temporal rules. Then the centroids, which represent the behaviour of the population, are extracted using a clustering algorithm. Finally, a label propagation for automatically classifying all ICPs by labels is realized using a K- Nearest-Neighbour classification.

BD processed in the integration layer is made available for the upper ***interpretation layer***. In this phase output data are semantically linked with the OD Matrix obtained through admin survey, in order to compare results. The two outputs are joined together virtually and presented in the ***access layer*** for data validation.

**5. Case study: business statistics produced by web mining**

Another interesting case study is the experimental work realized in ISTAT to obtain a subset of the yearly estimates produced for the “Survey on ICT usage and e-Commerce in Enterprises”. The survey provides a wide and articulated set of indicators on Internet activities: connection used, e-Business, e-Commerce, ICT skills, e-Invoice. The objective of this experimental work was to use survey data as a ground truth to create a classification model enabling the prediction of the values of target variables. This was based on the scraped content of the websites accessed with the URL indicated in the survey questionnaires [7].

This model, applied to the whole Italian population of enterprises (nearly 190,000, with 70% owning a website), can make it possible to enrich the information in the Business Register and increase the quality of the estimates produced by the survey.

The experimental work realized an automatic detection of the key features and their conversion through a supervised classification. Each classified record was obtained from the automatic analysis of enterprise websites, using the presence or the absence of some phenomenon for this classification. For example, determining whether an enterprise website offers e-commerce facilities or online job application facilities or links to social media are interesting cases for this problem.

The analytical process starts with gathering useful information for the statistical estimation process. This stage is realized by automatic web scraping all information from an enterprise’s websites in a given website-list. From each website, all web pages and their tags are extracted and loaded into the ***data source layer*** as a single document characterized by a set of terms, like: the attributes of HTML elements, file image name, page keywords.

After data is collected the data discovery step starts, transforming the unstructured data into structured data in the ***integration layer***. Several analytical steps are performed to identify the relevant information parts of the websites and to exclude the noisy part as much as possible. The last step of the discovery stage is to classify synthetized records by using state-of-the-art classification algorithms. In particular, Support Vector Machines, Random Forest, Logistic classifiers are used. Each of these classifiers requires setting some algorithmic parameters, which greatly affect the result of the classification activity. The choice of parameters is performed in the training phase of the classifiers on the bases of best performance.

All classified data are then reconciled with survey data through the semantic layer and then loaded into the ***interpretation layer***. In this layer the output from the classification process is organized in data marts to allow an integrated view with survey data. Finally, in the ***access layer***, the last analysis stage is realized to verify and validate results. These are compared with estimated variables produced by the survey, by applying design based estimation methods.

**6. References**

[1] Hassani H., Saporta G., and Sirimal Silva E. (2014), Data mining and official statistics: the past, the present and the future, [Big Data](https://www.liebertpub.com/journal/big)[Vol. 2, No. 1](https://www.liebertpub.com/toc/big/2/1)

[2] Elragal A. and Klischewski R. (2017), Theory‑driven or process‑driven prediction? Epistemological challenges of big data analytics, [Journal of Big Data](https://link.springer.com/journal/40537) December 2017

[3] ESS Centre of Excellence on Data Warehousing , (2017), A corporate Statistical-Data Warehouse Architecture, NTTS 2017.

[4] ESS Vision 2020 (2015), Building the Future of European Statistics.

[5] Furletti B. et all (2014), Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility, In Proc. of SIS 2014.

[6] Bianchi G., Bruni R. (2015), Effective Classification using Binarization and Statistical Analysis. IEEE Trans. on Knowledge and Data Engineering Vol. 27.

[7] Barcaroli G. et all (2016), Machine learning and statistical inference: the case of Istat survey on ICT, Proceeding of 48th SIS 2016, Salerno, Italy.

1. Member of ESS Centre of Excellence on Data Warehousing. [↑](#footnote-ref-1)