Analysis of Big Data Survey 2015 on Skills, Training and Capacity Building

DRAFT Version 1.0

12 Oct 2015

By

UN Global Working Group on Big Data for Official Statistics

Task Team on Skills, Training and Capacity Building

Executive Summary

1. All countries from both OECD and non-OECD regions pointed out that guidance on "*Skills and Training for Big Data*" is a priority in order to embrace Big Data for official statistics. However, the emphasis is on the methods not the tools/technologies required. Sharing experiences and knowledge among statistical organizations is crucial, because many Big Data sources are available globally and shared the same characteristics.
2. The most importance cited skills needed to handle Big Data sets are as follows: a) M*ethodologist* on Big Data issues, b) *Data scientist* and c) *Mathematical/statistical modelling* specialist. A Big Data methodologist can leverage potential Big Data sources for the use in the compilation of specific official statistics, a data scientist can perform data exploration and possess knowledge of various IT and statistical tools, and a mathematical/statistical modelling can transform raw data into statistics while ensuring the quality.
3. Even though Big Data tools are widely available, however, majority of respondents used the traditional technologies and tools such as R or relational databases. This is a practical approach as learning Big Data tools may require time and investment. Furthermore, many organizations establish various partnerships in data providers or research institutes, so that the bulk works to collect, filter and process raw data are handled and managed by them. The smaller data sets then can be analysed by statistical organisations using existing tools and technologies.

Introduction

The survey on Skills, Training and Capacity Building is a part of the bigger survey on Big Data by the UN Global Working Group on Big Data for official statistics that was sent to all UN member states in July 2015. The survey was meant to assess the current situation with regard to engagement of statistical agencies with Big Data sources. The Survey included 3 questions on skills and training for dealing with Big Data, namely about the skills that are lacking in the Statistical Offices to deal with Big Data, hiring of data scientist and training of the existing staff on Big Data topics. Responses were received from 86 countries, and analysis of the results was divided into the OECD and non-OECD countries for comparison.

We can see from the results that OECD countries compare to non-OECD countries use different types of Big Data sources for analysis, they hire more data scientists and provide more training on Big Data, but there is not much difference in the methods and tools for the analysis of Big Data between the two sets of countries.

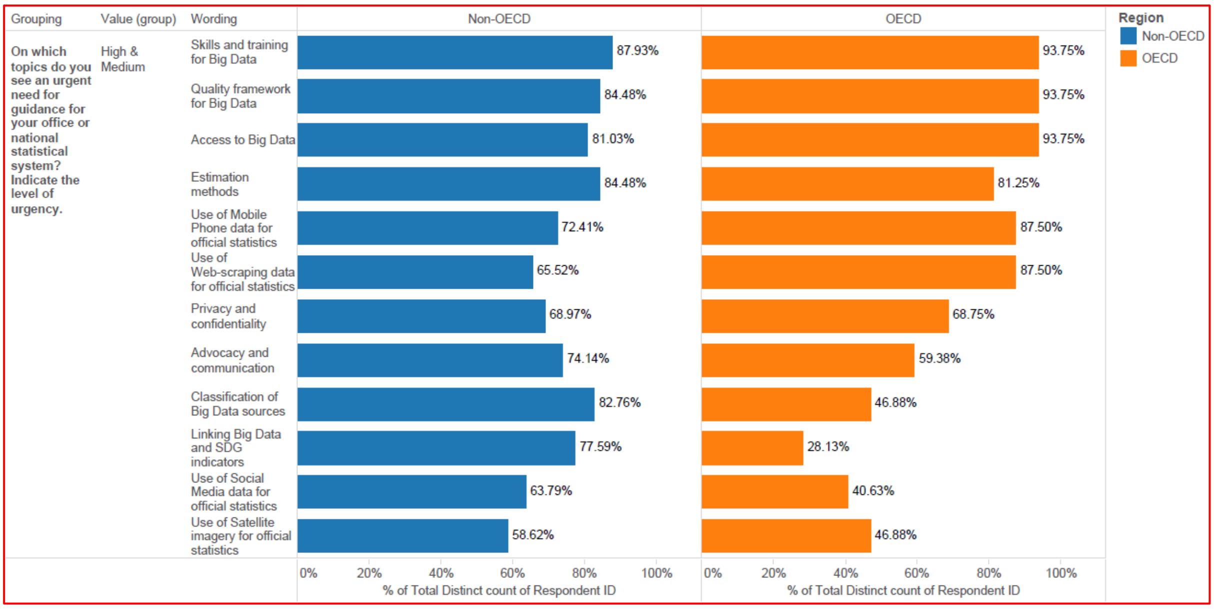
Results of this survey are similar to the results of the Survey on Skills necessary for people working with Big Data in the Statistical Organisations that was conducted by the UNECE in the autumn of 2014. The UNECE survey identified the following most important skills necessary for working with Big Data:

- IT skills: noSQL databses, SQL databases and Hadoop
- Statistic skills: Methodology and standards for processing Big Data, data mining
- Other skills: Creative problem solving, data governance and ethics

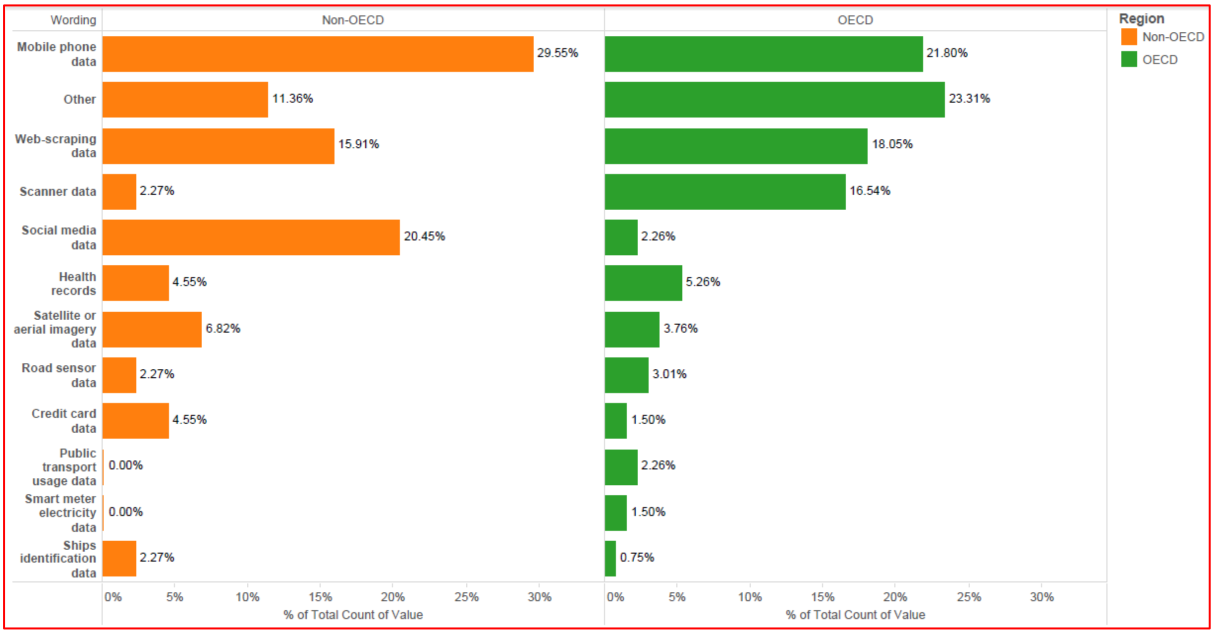
Based on the results of both surveys, we can say that at the moment there is a lot of demand for training in skills that were identified as very important for the work with Big Data, there is a need to share training materials, experiences and best practices among countries.

Results of the survey

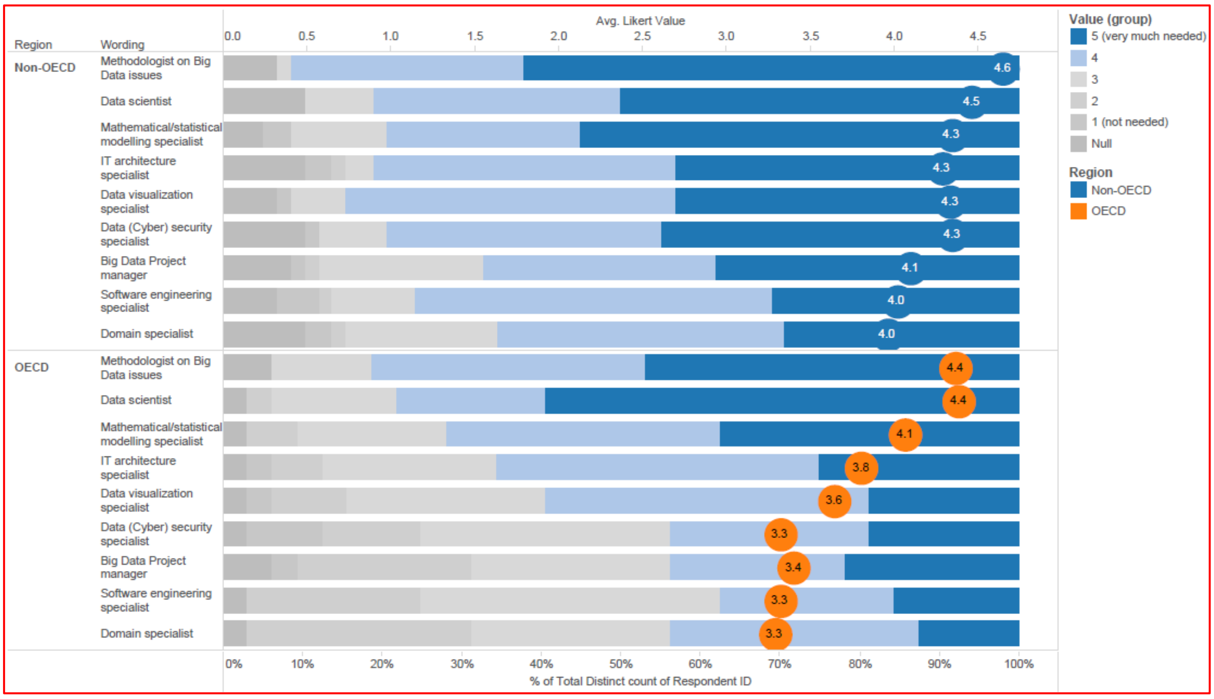
Respondents indicated that "*Skills and Training for Big Data*" topic as the highest priority to be addressed and guidance to be provided for national statistical system in both OECD and non-OECD countries (see below), followed by "*Quality framework for Big Data*" and "*Access to Big Data sources*". The use of emerging IT tools and techniques requires constant investment to improve IT capacities within the statistical office. Furthermore, topic "*Estimation methods*" which is closely related to skills to manipulate and manage Big Data sources is cited quite high in terms of priority. Nevertheless, non-OECD countries indicate the relative importance of "*Linking Big Data and SDG indicator* and *Use of Social Media data*" but not OECD countries.



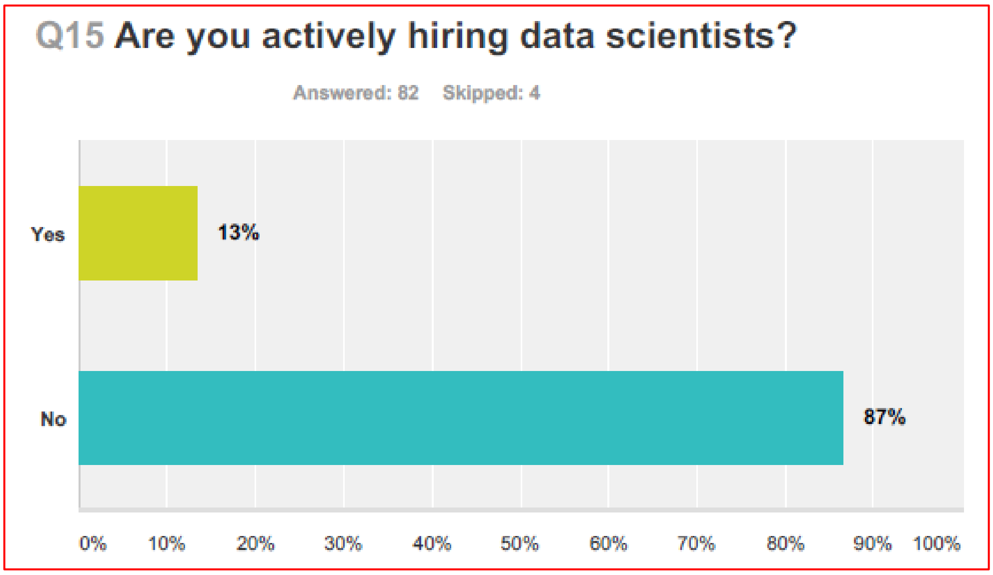
Question on Big Data sources ("*Which specific Big Data sources have you used or do you consider using?*") is relevant with the skills, training and capacity building task team because each data source has its own characteristics that require specific skills & technologies for data processing. **Mobile phone data**, **web scrapping data**, **scanner data** and **social media data** are the most widely considered/used. However, there is divergence between OECD and non-OECD countries: scanner data is significantly used in OECD countries, whereas social media data is quite popular in non-OECD countries. This correlates with the most cited importance topics to be addressed (see above). Other sources including online search data, administrative data (e.g., tax files, health records, education records), wikistat logs.



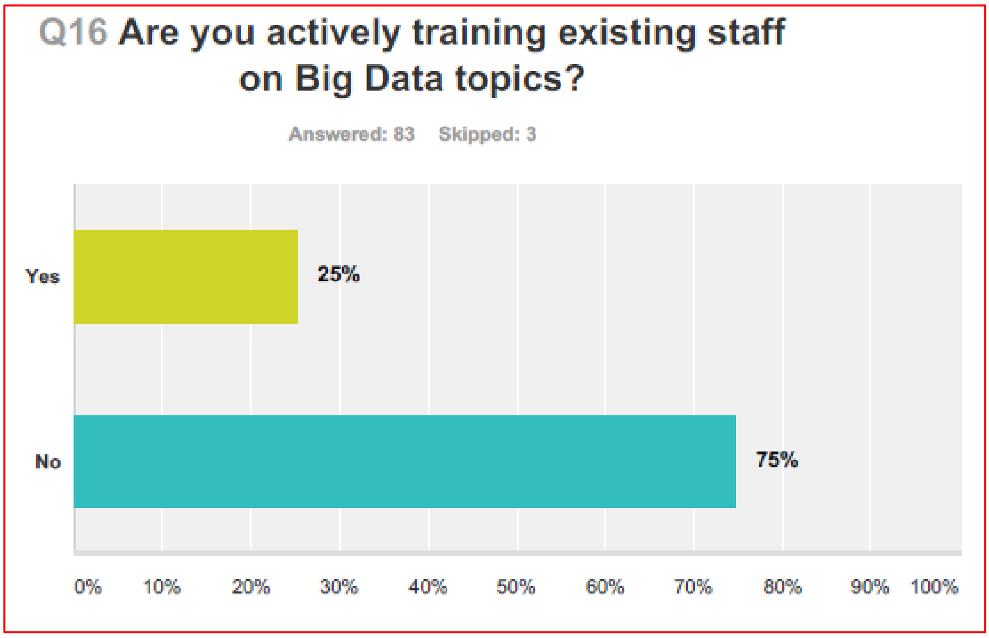
Question below is related directly with the skills needed to better deal with Big Data. *Methodologist* on Big Data subject matter, *Data scientist* and *mathematical/statistical modelling specialist* are the top three skills/occupations cited. However, non-OECD countries underline the relative importance of *IT architecture specialist*, *Data visualization specialist* and *Cyber security specialists*, but not OECD countries. Traditional skills/occupations that are normally available (and well established) in statistical offices comprise mathematical/statistical modelling specialists, IT occupations, and Domain specialists. Respondent defines that data scientist has a compilation of three skills sets of (Big Data) methodology, IT and subject-matter. In addition, Big Data project manager is defined as data scientist that possesses managerial proficiencies. It should be noted that the question asked about the needed skills and not their importance, therefore lower score does not mean that the skills are not needed (but instead, it implies that those skills are already possessed by the staff).



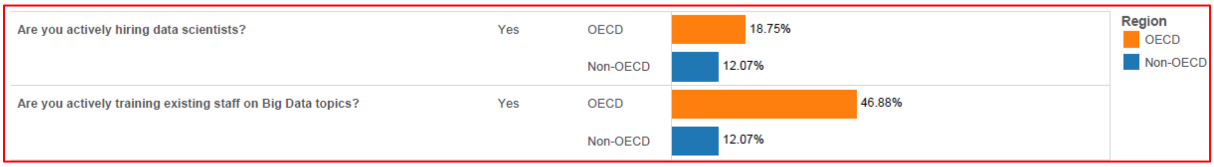
Even though data science is considered as much needed skill, the respondents do not actively hire data scientist. Some respondents noted that there is no specific position for data scientist, as it is still need to be defined (and formalised). Therefore, as an alternative, statistical offices look for candidates that poses both statistical and IT expertise.



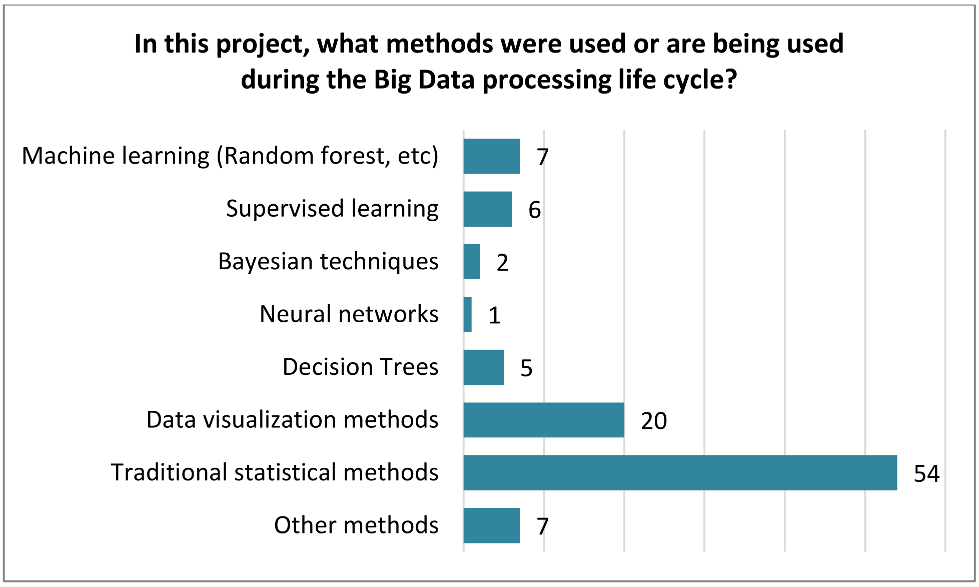
Due to limited resources, instead of hiring data scientist, the common approaches (to acquire those skills) are to train existing staff, to hire short term consultants, and to establish collaboration and partnership with universities and research institutes. Participation in Big Data conferences and workshops, UNECE Sandbox, inviting Big Data experts for in-house seminar and self-learning are cited as ways to improve the Big Data skills. However, the majority of respondents do not actively train existing staff on Big Data topics.

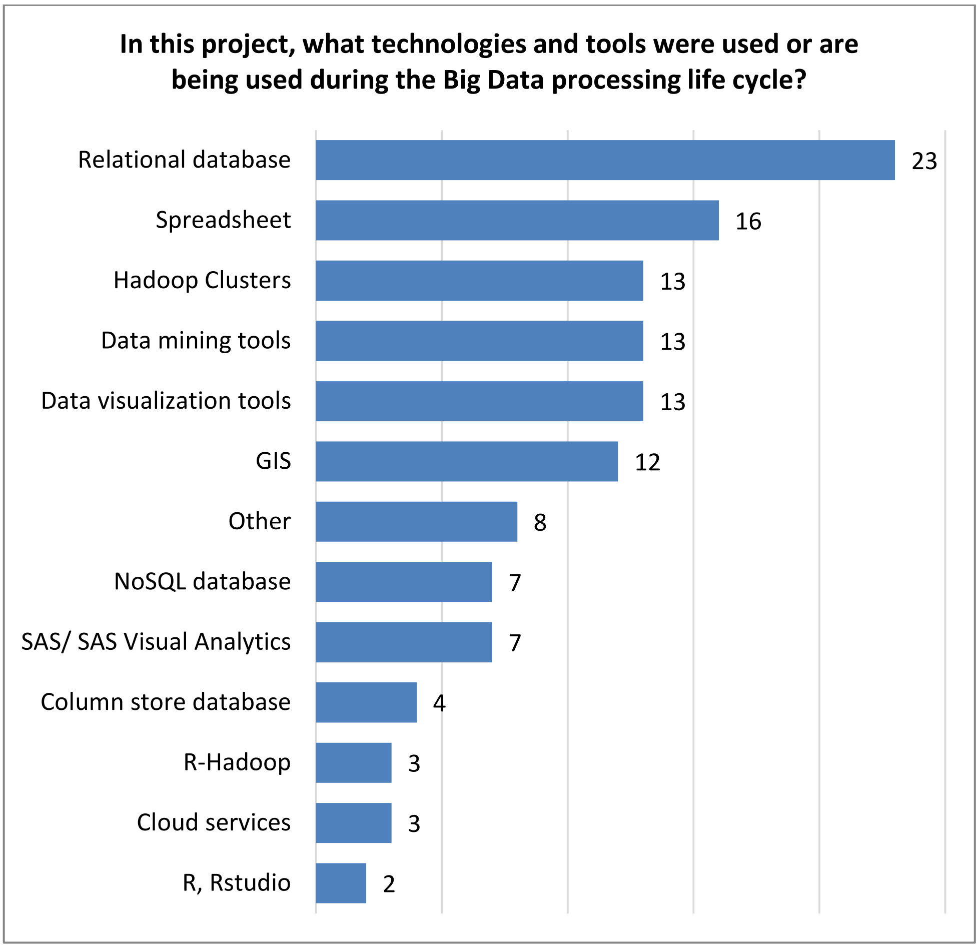


Looking at the regional breakdown between OECD and non-OECD countries, we found out that OECD countries are more actively hiring data scientists and providing training on various Big Data topics compared to non-OECD countries. This result shows that OECD countries are more aware about the importance of skill in Big Data era.

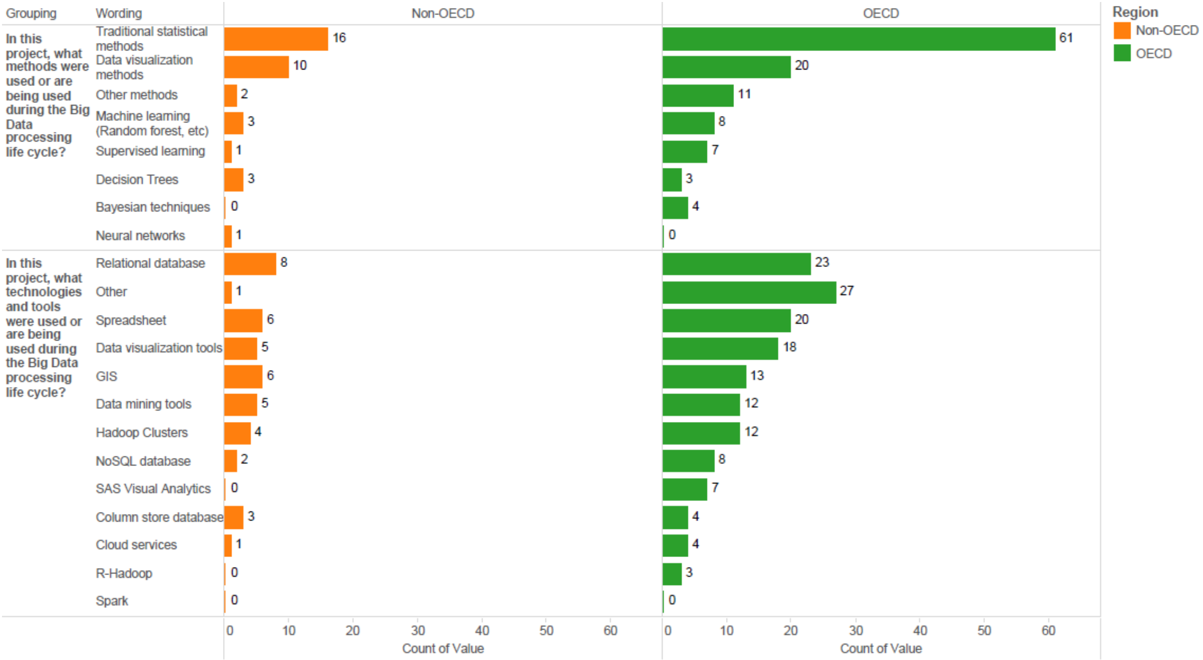

Are you actively hiring data scientists?
Yes
OECD
18.75%
Non-OECD
12.07%
Are you actively training existing staff on Big Data topics?
Yes
OECD
46.88%
Non-OECD
12.07%
Region
OECD
Non-OECD

In terms of methods and tools used in the Big Data project, most respondents indicated that traditional methods and technologies (e.g., relational database, spreadsheet) are currently being used or were used. This may indicated lack of access to and knowledge of Big Data technologies such as Hadoop Clusters or data mining tools. In addition, data visualization methods and tools are cited a quite a lot. This indicates that, presentation of Big Data result is quite an important factor.





Furthermore, there is no much difference between OECD and non-OECD countries in terms of methods, technologies and tools used in Big Data projects. Countries basically use traditional statistical methods using statistical software such as R and relational databases. This is probably due to lack of skills in or access to Big Data technologies.



Way forward

The importance of training and capacity building has been highlighted throughout the document. In many countries, using Big Data sources for official statistics is unthinkable, mostly due to lack of awareness of these new emerging data sources. Nevertheless, the principles of official statistics state that "… that data may be drawn from all types of sources to honor citizens' entitlement to public information (based on quality, timeliness, cost) …", therefore, it's an obligation for statistical organizations to explore and assess the possible use of Big Data sources.

The first step is to raise awareness on the use such data sources by sharing experiences, pilot projects and knowledge among countries. This can be aligned with the work of Big Data task team on advocacy. Furthermore, establishing web-based knowledgebase and document sharing platform is also crucial. The second step is to further identify which skills, both methods and tools, are relevant to deal with Big Data, including the project management skills, and the assessment whether traditional methods and tools are sufficient or not. Finally, to address the identified skills gap, training curriculum and courses can be developed.