United Nations

ECE/CES/2017/27

Economic and Social Council

Distr.: General
6 April 2017

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-fifth plenary session
Geneva, 19-21 June 2017
Item 4 of the provisional agenda
The next generation of statisticians and data scientists

Big data competences

Note by the Central Statistics Office of Ireland and the High-level Group for the Modernisation of Official Statistics

Summary

The issue of big data is a new challenge and new task for National Statistical Offices. This document outlines research undertaken by the High-level Group for the Modernisation of Official Statistics. This document discusses the steps carried out by the High-level Group Organisational Framework and Evaluation Committee, and the research project they undertook in order to define the skills set required within statistical organizations to respond to the demands of the big data revolution.

The document is presented to the Conference of European Statisticians' seminar on "The next generation of statisticians and data scientists" for discussion.

I. Background

1. The High-Level Group for the Modernisation of Official Statistics (HLG-MOS) was set up by the Conference of European Statisticians (CES) in 2010 to oversee and co-ordinate international work relating to statistical modernization. It promotes standards-based modernization of official statistics. The HLG-MOS oversees modernization projects and manages the models and tools needed to support modernization in statistical organizations. It aims to improve the efficiency of statistical production, and help statistical organizations produce outputs that meet user needs better.

2. The governance structure for the HLG-MOS included a number of modernization committees working on the modernization agenda. One of these committees, the Organisational Framework and Evaluation Committee (OFEC) was established to consider and make proposals about how best to bring about the organizational changes necessary to support modernization in statistical organizations. The overall governance structure also included a number of international groups collaborating on specific projects and one of these groups was established to examine how statistical organizations could use big data in statistical production.

3. In May 2014, as part of the collaborative working arrangements of all of these groups and projects, the OFEC received the following request from the Big Data Project group: "*The Big Data Project has requested assistance from this group to define the skills needed for statistical organizations to be able to make use of big data sources, assess the extent to which those skills are already present in statistical organizations, and propose training activities to fill any gaps".*

4. In response, the OFEC undertook a research project to define these skills.This document outlines the work of the HLG-MOS group in designing a research project with an outcome of a comprehensive competency framework ensuring statistical organizations have the required skill set in place in order to respond to the demands of the big data revolution. The competency framework acknowledged that one person could not encompass all the skills required but rather a team with a specific set of specialist skills would be best placed to meet future demands.

5. The outcomes of the project will assist National Statistical Offices to:

- Provide existing staff with good competency frameworks.
- Allow statistical organizations to implement focused training plans where gaps have been identified based on these competencies.
- Provide a framework for recruiting staff with appropriate skills and behaviours.

6. It is important that National Statistical Offices are agile with an ability to adapt to their external environment. In order to achieve this agility, skill gaps need to be identified and training interventions available to meet this demand. This document will discuss the work undertaken to achieve this and also the issue and challenges faced and recommendations for future research.

II. What has been done?

A. Survey design

7. As the OFEC group were unable to find comprehensive data available on this subject it was decided to issue a comprehensive survey to gather available information. It

was also decided that the target audience was human resources (HR), information technology (IT) managers and big data practitioners. These groups would be involved in the skills assessment, work force planning and systems infrastructure. The survey was designed in such a way as to gauge a personal viewpoint rather than an official response. It was structured in this way in order to achieve a maximum response rate and also to gather realistic and meaningful data, a copy of the questionnaire is presented in annex 1.

B. Survey distribution and response rate

8. The survey was sent in July 2014 to all member countries of the United Nations Economic Commission for Europe (UNECE) as well as other countries and organizations that participate in the work of the Conference of European Statisticians (CES). Responses were accepted on behalf of individuals and not organizations. A total of 137 responses were received. Due to some incomplete submissions, 107 responses were used for analysis.

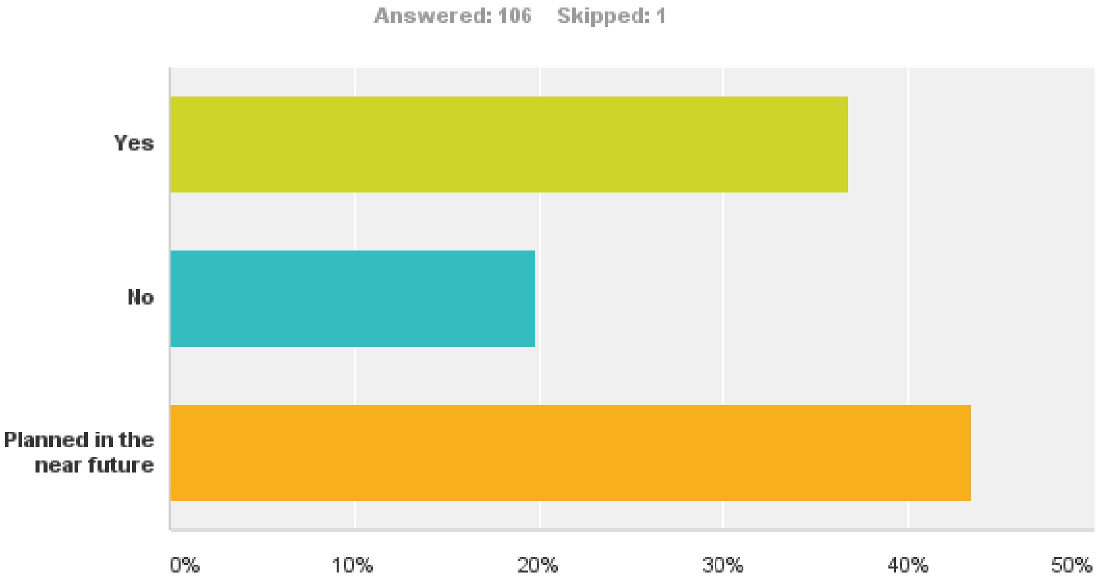
9. Responses were received mainly from UNECE member states with 77 per cent of respondents coming from Europe, 6 per cent coming from EECCA countries, 6 per cent from North and South America and 10 per cent from Asia-Pacific.

10. As the majority of responses were from Europe their country of origin was reviewed so as to assure unbiased data. 46 per cent of response from Europe originated from two countries, while the remaining 54 per cent consisted of 25 countries and 2 international organizations. There was a concern that their influence on overall results would affect the overall survey, for this reason they were analysed in two groups, group one all responses and group two responses excluding two countries that provided 46 per cent of the responses.

11. Most respondents came from IT departments, followed by statistics and HR/training departments.

12. Out of all respondents, almost 37 per cent answered that they already work with big data with 43.3 per cent stating that they will work with big data in the future (see figure 1).

Figure 1
Does your organization work with big data?



13. This shows that 80.4 per cent of all respondents would be using big data in the near future attesting to the importance of the development of the competency framework.

C. Key findings

14. Following analysis of the data, it was identified that the most important skills for working with big data could be grouped under the following three headings:

- IT skills
- Statistical skills
- Other skills

15. Each of the above will be discussed under their headings, summary tables outline in detail the particular skills required under each of these headings and the levels of these skills available in organization.

16. Skills which are absent or only present at the basic level in statistical organizations include IT skills, Hadoop[1], no SQL databases[2], statistical skills including methodology and standards for processing big data. Under the heading other skills most of the skills are present at advanced and intermediate levels. The research found there was also insufficient training as of (Oct 2014) in the skills that were identified as most important for people working with big data.

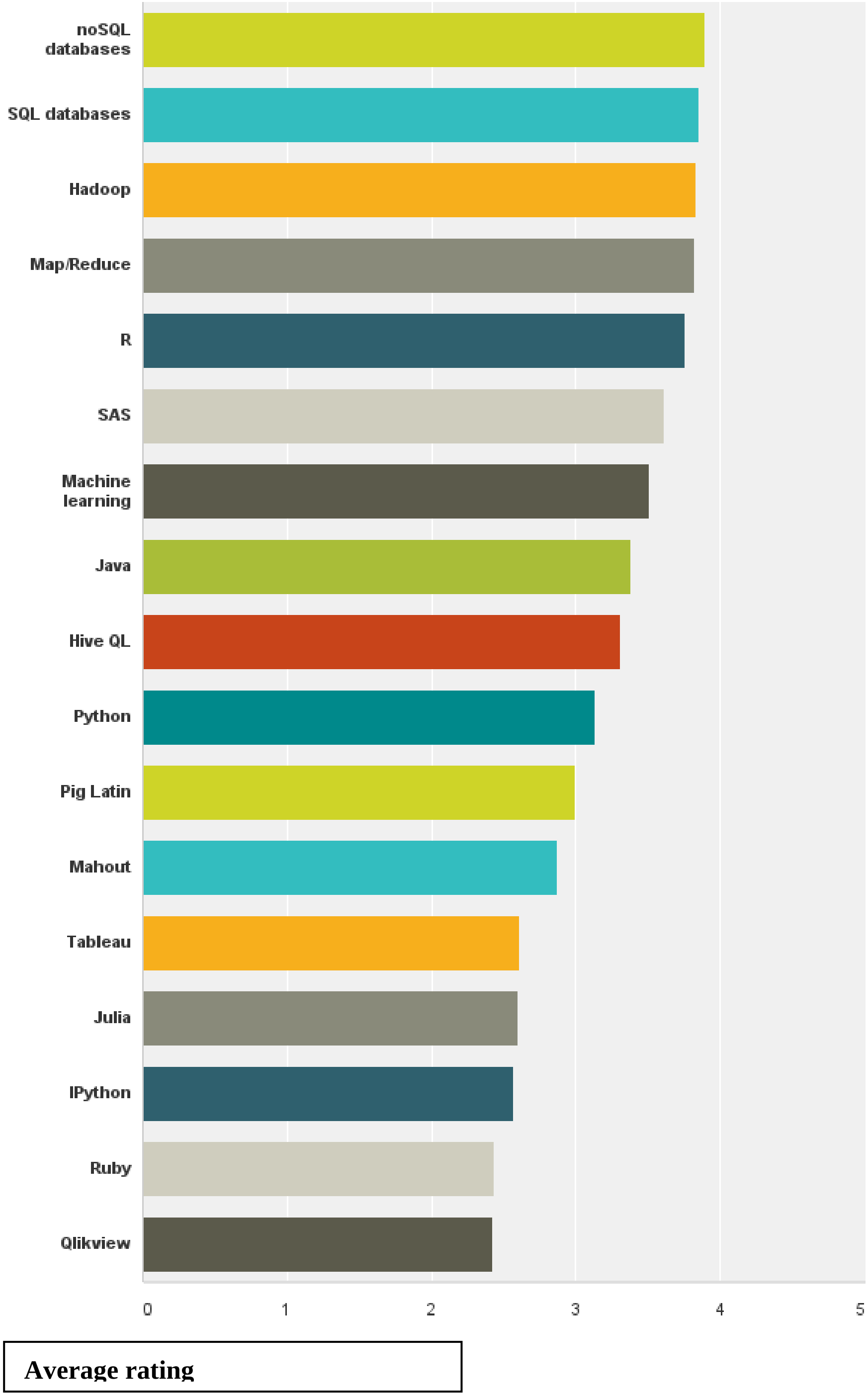
1. IT skills

17. Among all respondents the 6 most important IT skills for working with big data were identified as displayed in the infographic below (figure 2) rating from 1-5 the skills below were deemed most important.

[1] Hadoop is a software library and a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

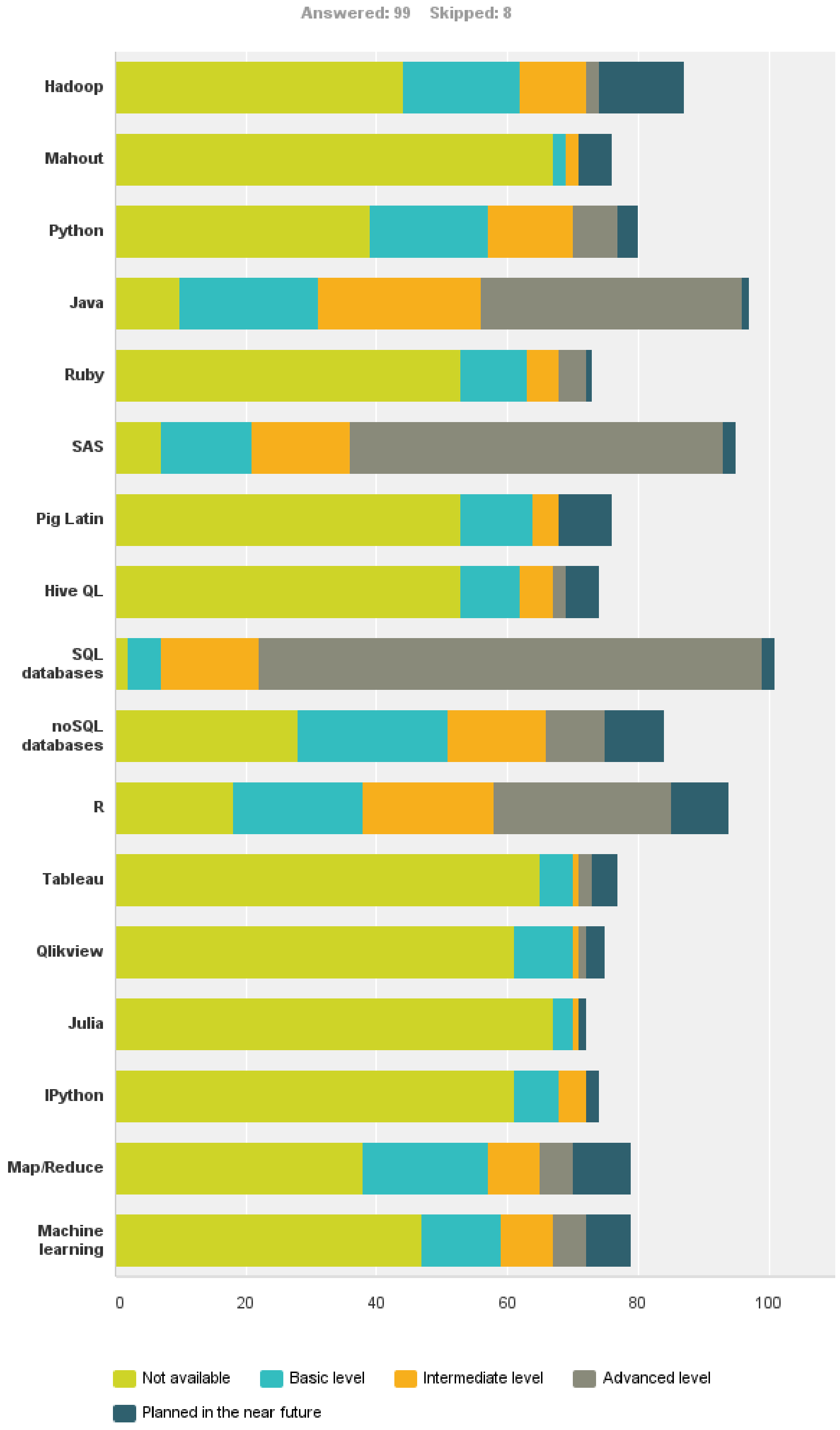
[2] No SQL (Structured Query Language) databases refer to databases with more than one storage mechanism and their features include: not using the relational database model, run well on clusters, mostly open source, built for the 21st century web estates and schema-less.

Figure 2
Rating of skills for working with big data



18. Comparing to what IT skills are already available the skills gaps were clear from the summary infographic below (figure 3)

Figure 3.
Which of the following skills you already have in your organization and at what level?

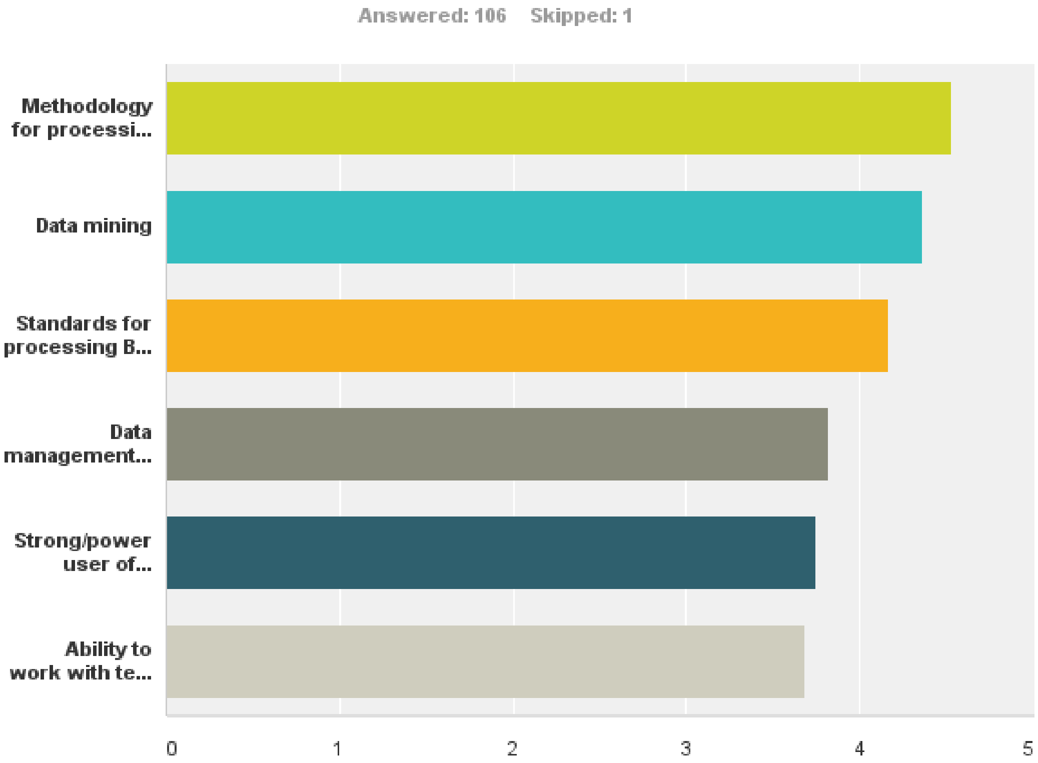


19. Figure 3 above shows that 30 per cent of organizations had no availability of no SQL database, over 40 per cent had no availability of Hadoop and over 35 per cent had no availability of Map/Reduce.

2. Statistical skills

20. Respondents reported on current statistical skills in their organizations see figure 4, below.

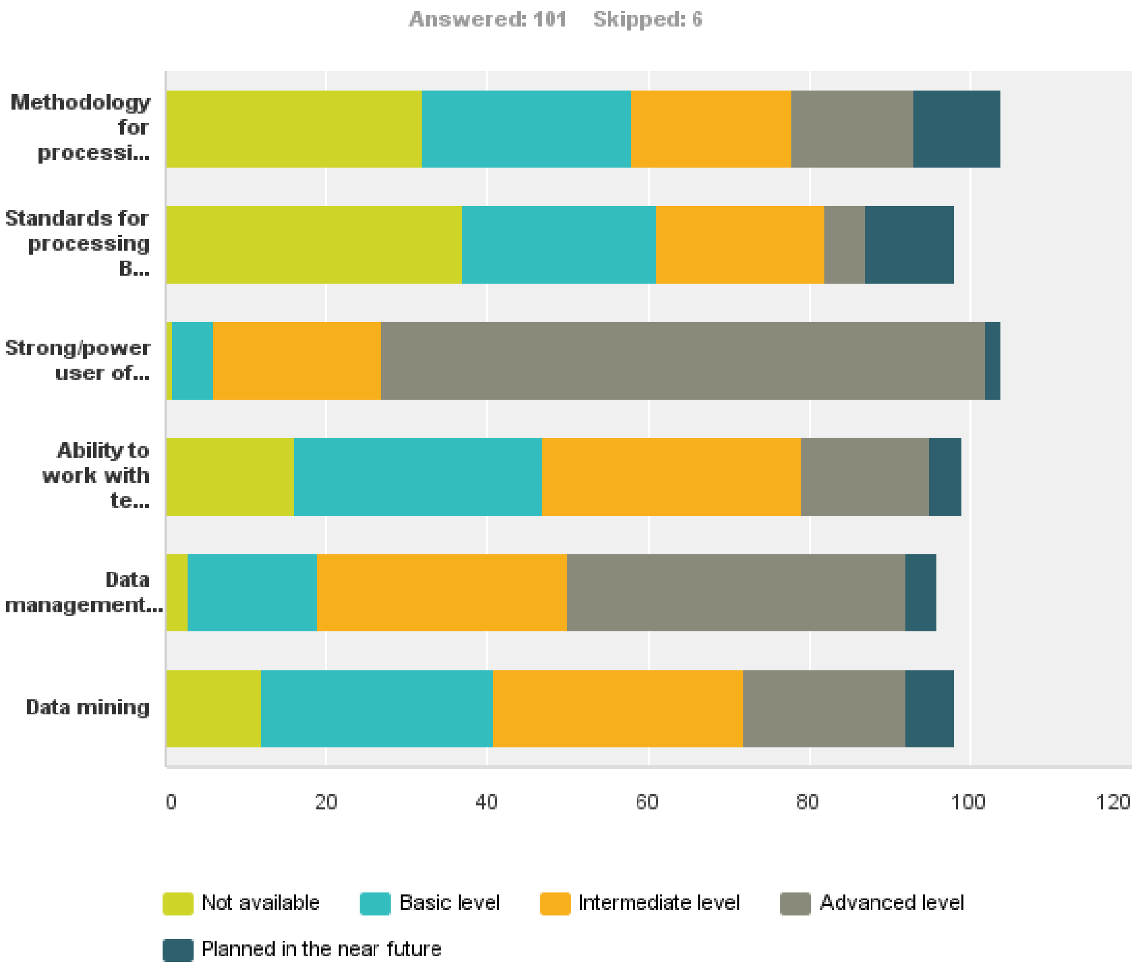
Figure 4
Statistical skills



Average rating

21. Statistical skills were also rated as per current organization skill level. Figure 5 summarizes the responses.

Figure 5
Current level of statistical skills

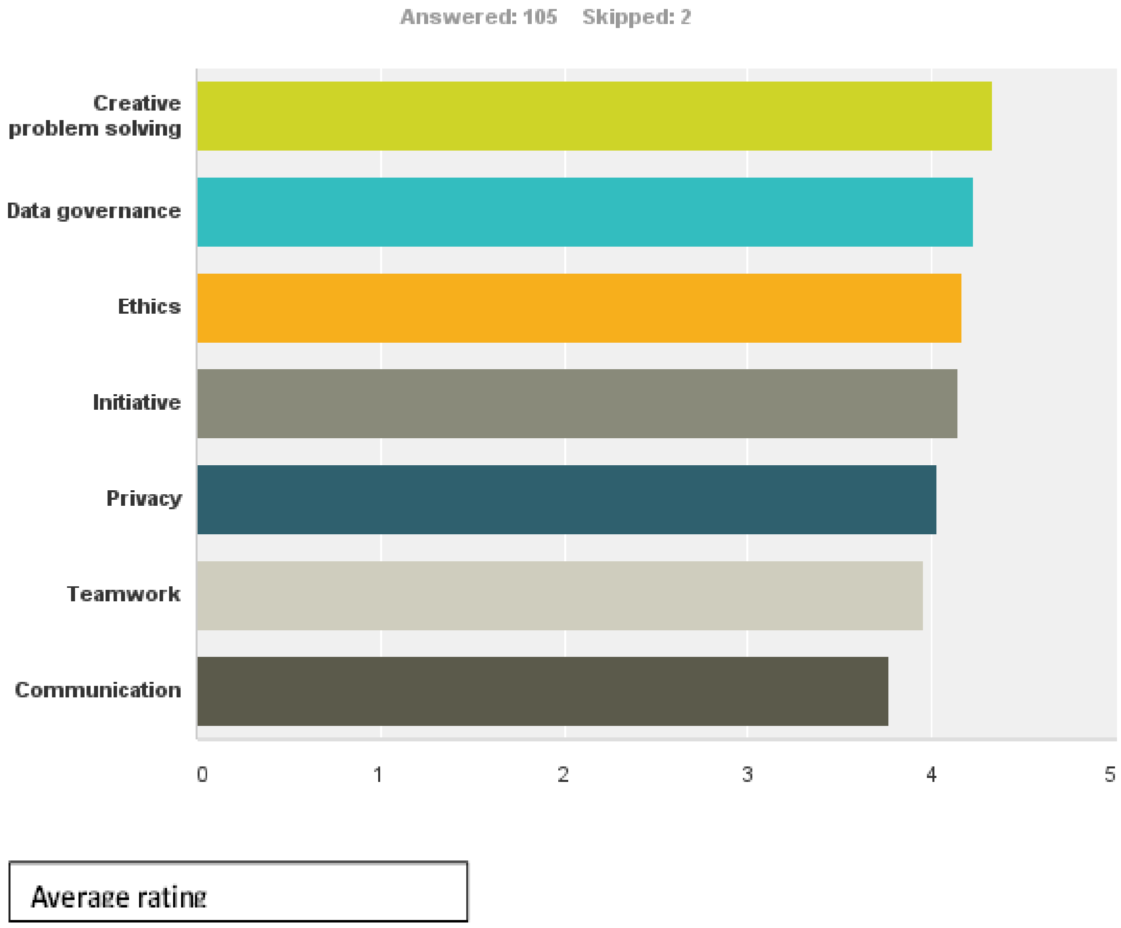


22. As per Figure 5, over 30 per cent of respondents did not have methodology for processing or standards for processing. Over 20 per cent of respondents were at a basic level in methodology for processing and standards for processing.

3. Other skills

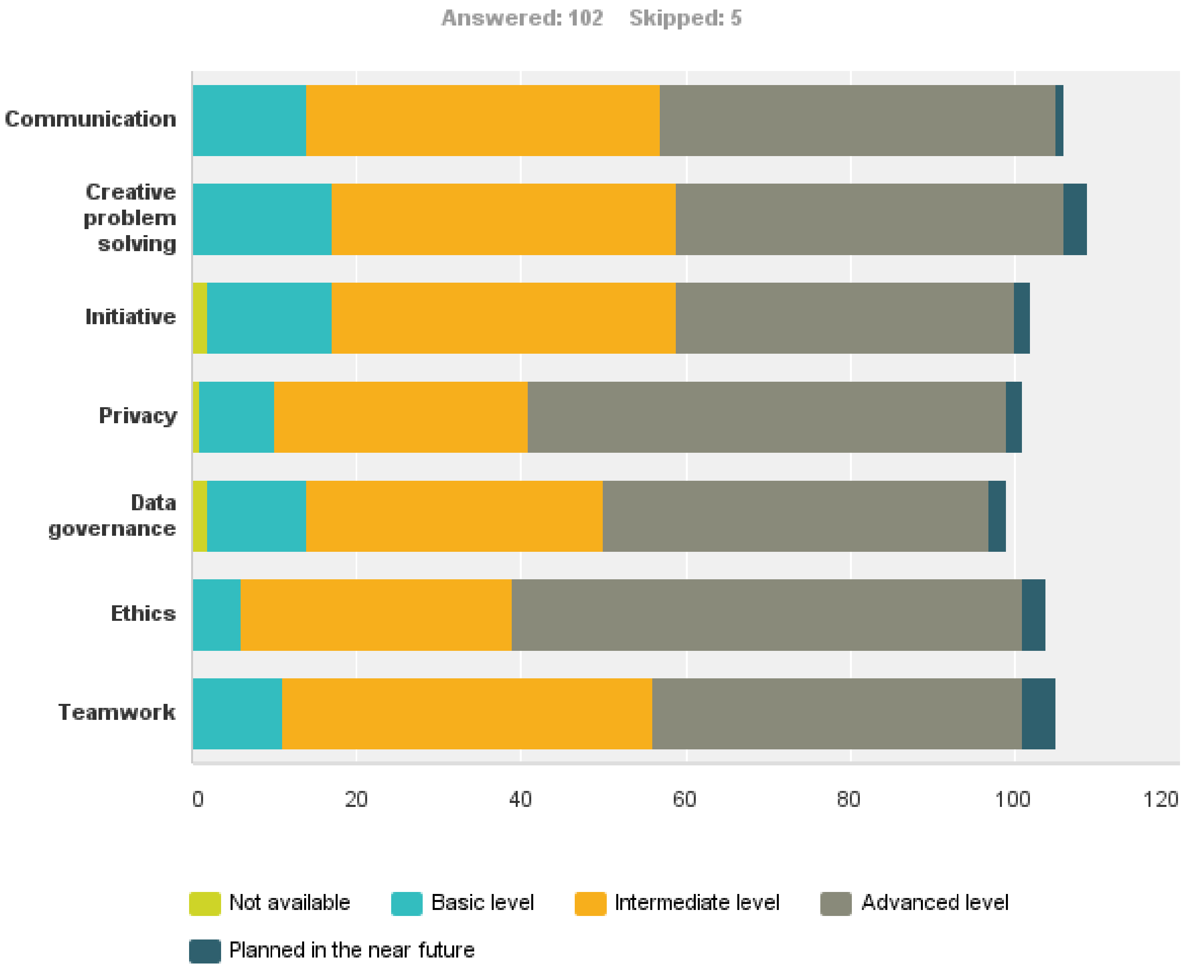
23. Respondents felt that creative/problem solving was very important followed by data governance and ethics (see figure 6).

Figure 6
Other skills



24. Over 30 per cent of respondents reported that other skills training was at an advanced level in the 7 topics (see figure 7).

Figure 7
Training of other skills



25. It is clear from the above research how important it is to identify skills related to big data. The survey data and key findings as presented above provided a comprehensive understanding of the skills required for working with big data. The working team decided the best way to use this data was to build a competency framework to be utilised for recruitment and training.

III. Competency framework overview

A. What is a competency framework?

26. Competencies provide organizations with a way to define in behavioural terms what people need to do to produce the results that the organization requires, in a way that is in line with its culture. These are the integrated knowledge, skills, judgment, and attributes that people need to perform a job effectively.

27. Having a defined set of competencies for each role or team in your business allows to identify the kind of behaviours the organization values and requires to help achieve its objectives. While developing the competencies for big data it was evident that working with big data required a range of specialist skills. The research team identified a number of roles in this team. All members of the team should have core Statistical/IT skills with individual team members having specialist skills. The group should consist of the following members:

(a) Big Data Team Leader;

(b) Methodologist specialist skills;

(c) Data analytic skills;

(d) ICT skills.

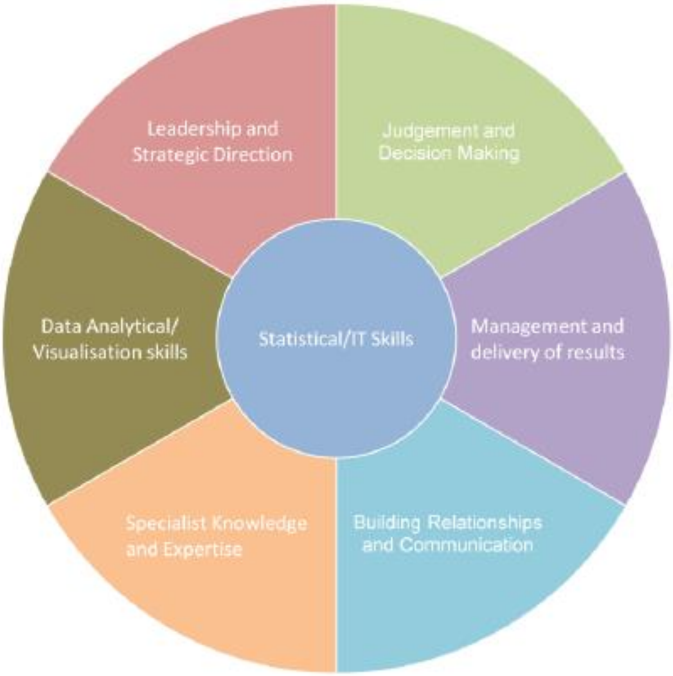
28. The competency framework as presented in picture 1 and picture 2 below sets out the competencies and performance indicators for this team. A detailed competency framework is presented in appendix 2.

29. The competency framework was divided into two levels for competency headings, Big Data Team and Big Data Team Leader. The Big Data Team Leader will have proven management competence with a record of influence, achievement, innovation and sound judgement. The team leader will have proven leadership skills operating at a high level in a pressurised environment where communication, liaison and negotiation skills are tested in a demanding work environment.

30. Picture 1 outlines the competencies required for Big Data Team Leader. The core skills requirement for the role of Big Data Team Leader is Statistical/IT skills which includes a detailed knowledge and understanding of statistical methodology and concepts and the IT skills relevant to statistical production and analysis. The outer parts depict the specific competencies required.

31. A key competency for a Big Data Team Leader is leadership skills with a strategic direction meaning that they should have an ability to inspire a sense of purpose and direction. They should also possess good judgement and decision-making, management and delivery of results and relationship building and communication. These skills should ensure that the Big Data Team Leader has the necessary skills to deliver results while building organizational capability, nurture both internal and external relationships and utilise intelligence decision making and communication clearly. In addition to the above skills, the Big Data Team Leader should also have the appropriate specialist knowledge to work effectively as part of the Big Data Team.

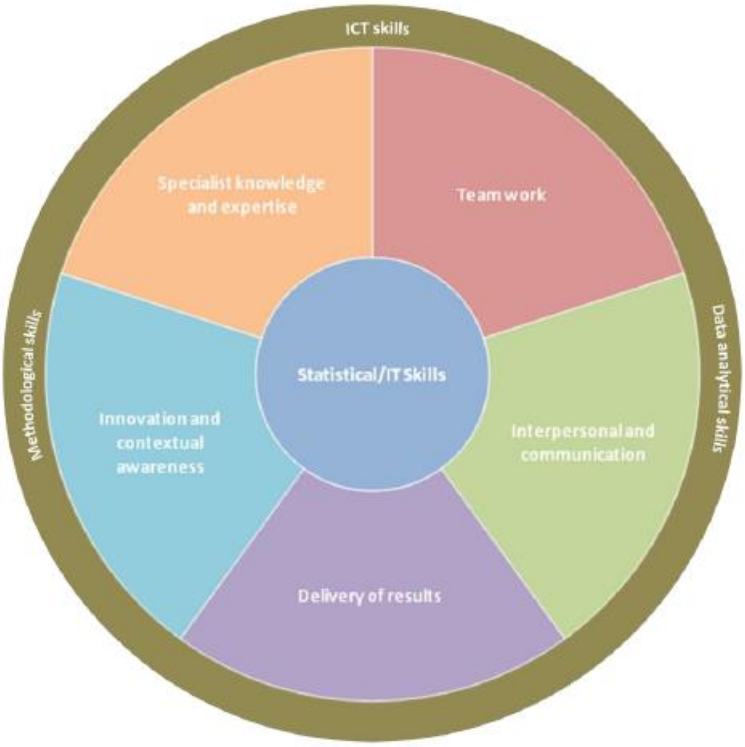
Picture 1
Big Data Team Leader competency:


Leadership and Strategic Direction
Judgement and Decision Making
Data Analytical/ Visualisation skills
Statistical/IT Skills
Management and delivery of results
Specialist Knowledge and Expertise
Building Relationships and Communication

32. Picture 2 shows the final competency framework which incorporates ICT/ Data analytic skills and methodology skills. During the research and survey findings it was decided that one person could not encompass the skill set required to complete a Big Data Team therefore ICT, Data analytic and methodology experts were necessary to complete the team profile.

33. Statistical and IT skills were at the core of the team level and each individual within the team should have those skills in order to work effectively with others within the team. Specifically, they refer to the team members’ ability to use core statistical skills for data analysis, use of programming/ scripting language such as R, SAS and SPSS, combining various data processing techniques to achieve given analytical tasks, good knowledge of data science methods and good evidence of keeping up to date with new trends in data techniques and technologies.

Picture 2
Big Data Team Competency



B. Issues and challenges

34. Working within an international team was both challenging and rewarding. Utilising web conferencing tools such as WebEx enabled the team to overcome these challenges and engage with one another on a regular basis. Working within this team dynamic meant that there was a varied skill level and knowledge of big data, enhancing and utilising the exiting knowledge within the team was a key challenge as it initially took some time working as part of the team to see what strengths team members brought to the working group.

35. The response rate of the survey was a key issue in terms of ensuring we had an unbiased return. It was mentioned under key findings that weighted averages had to be utilized as some responses from individual countries were higher than others. Acknowledging this and taking measures to avoid bias was key in producing a validated report.

36. Buy in from the wider statistical community was also a challenge and embedding the use of the competency framework tool in organizations. This has improved in recent times for example in Ireland we have adopted the framework and utilised it in our recent recruitment campaign (2016) for statisticians. This validates the findings as we identified under the competency framework for Big Data Team a number of key competencies that were assessed during our recruitment process for example, delivery of results, specialist knowledge and expertise and interpersonal and communications.

37. Another issue could be reviewing the type of questions that the survey addressed, were we asking the right type of questions did the questions illicit the right response. Would it be more beneficial to target a specific group or conduct a range of focus groups in

order to receive the right response and extract the correct skill requirement for working as part of a Big Data Team.

C. Recommendations and conclusion

38. A recommendation that is a key element to the development of these competencies is to revisit the competency framework on both levels (Big Data Team Leader, Big Data Team) to ensure that the competencies meet the demands of the role. As mentioned above a focus group may be best to re-test the framework selecting those who are working in big data for over one year.

39. Since the development of the framework there have been a number of big data advancements including the emerging smart data. In order to respond to the external environment a review of the competency framework in line with more recent IT, statistical and big data trends should be a priority.

40. Another recommendation would be to re-issue this survey and compare results. It would be interesting to see how the data would compare in terms of IT skills and training required. It might then prompt a review of the competency framework ensuring the framework is kept up to date.

Annex 1

Questionnaire about the skills necessary for people working with big data in statistical organizations

The target group for this questionnaire is the human resource and IT managers in national statistical offices. We are looking for your personal views rather than an official reply on behalf of your organization. For this reason we would also like to encourage you to forward this link to any of your colleagues that might be able to contribute.

We will appreciate if you can provide your answers by XXXX.

Results will be aggregated, and will not identify individual people, organizations or countries.

1. **Please provide your contact details:**

   Name:
   Organization:
   Field of work (HR, IT, Training, other):
   Contact e-mail:

2. **Does your organization work with big data?**

   Yes
   No
   Planned in the near future

   Comments:

3. **How important do you think are those skills for working with big data?**

   Please rate them from 1 (not important) to 5 (very important)

Skills	***Rating***
IT skills	
Ability to use big data technologies such as: • Hadoop • Mahout	
Ability to programme in: • Python • Java • Ruby • R • SAS • Pig Latin • Hive QL	
Strong user of: • SQL databases • noSQL databases	
Ability to use visualisation software such as: • R • Tableau • Qlikview • Julia	

• IPython	
Knowledge of : • Map/Reduce • Machine learning	
Other (please specify):	
Statistics skills	
• Methodology on statistical learning	
• Standards for processing big data	
• Strong/power user of software such as Excel, SAS, SPSS	
• Data management skills including documentation, registration, access control	
• Ability to work with text analytics	
• Data mining	
Other (please specify):	
Other skills	
Communication	
Creative problem solving	
Initiative	
Teamwork	
Data governance	
Ethics	
Privacy	
Other (please specify):	

4. Which of the following skills you already have in your organization and at what level?

Skills	*Not available*	*Level*			*Planned in the near future*
		basic	*intermediate*	*advanced*	
IT skills					
Ability to use big data technologies such as: • Hadoop • Mahout					
Ability to programme in: • Python • Java • Ruby • R • SAS • Pig Latin • Hive QL					
Strong user of: • SQL databases • noSQL databases					
Ability to use visualisation software such as: • R • Tableau • Qlikview • Julia					

• IPython					
Knowledge of : • Map/Reduce • Machine learning					
Statistics skills					
• Methodology for processing big data					
• Standards for processing big data					
• Strong/power user of software such as Excel, SAS, SPSS					
• Data management skills including documentation, registration, access control					
• Ability to work with text analytics					
• Data mining					
Other skills:					
Communication					
Creative problem solving					
Initiative					
Teamwork					
Data governance					
Ethics					
Privacy					
Other, please specify					

5. Please indicate in which areas you have training in your statistical organization and indicate if you have training materials that you can share or recommend?

(Training materials include: books, internet resources, training materials developed in the Statistical Organization, etc.)

Skills	*Training*	*Training materials that you can share or recommend*
IT skills		
Ability to use big data technologies such as: • Hadoop • Mahout		
Ability to programme in: • Python • Java • Ruby • R • SAS • Pig Latin • Hive QL		
Strong user of: • SQL databases • noSQL databases		
Ability to use visualisation software such as:		

• R • Tableau • Qlikview • Julia • IPython		
Knowledge of : • Map/Reduce • Machine learning		
Other, please specify		
Statistics skills		
• Methodology for processing big data		
• Standards for processing big data		
• Strong/power user of software such as Excel, SAS, SPSS		
• Data management skills including documentation, registration, access control		
• Ability to work with text analytics		
• Data mining		
Other, please specify		
Other skills:		
Communication		
Creative problem solving		
Initiative		
Teamwork		
Data governance		
Ethics		
Privacy		
Other, please specify		

6. **Please indicate top 5 priorities for training for your statistical organization across all areas: IT, statistics and other (by marking them 1-5, where 1 is the highest)**

Skills	***Training priorities***
IT skills	
Ability to use big data technologies such as: • Hadoop • Mahout	
Ability to programme in: • Python • Java • Ruby • R • SAS • Pig Latin • Hive QL	
Strong user of: • SQL databases • noSQL databases	
Ability to use visualisation software such as: • R	

- Tableau - Qlikview - Julia - IPython	
Knowledge of : - Map/Reduce - Machine learning	
Other (please specify):	
Statistics skills	
- Methodology on statistical learning	
- Standards for processing big data	
- Strong/power user of software such as Excel, SAS, SPSS	
- Data management skills including documentation, registration, access control	
- Ability to work with text analytics	
- Data mining	
Other (please specify):	
Other skills	
Communication	
Creative problem solving	
Initiative	
Teamwork	
Data governance	
Ethics	
Privacy	
Other (please specify):	

5. Other comments/suggestions