Big Data Architectures @ Istat

Keywords: big data, IT architectures, data science infrastructure

1. INTRODUCTION

Modern organizations have recognized the importance of having a defined and standard architecture that is able to guide the implementation of the organization's vision and to harness external drivers and changes. National Statistical Organizations (NSOs) are part of this trend and are more and more investing in defining architectures (business, application/information, IT) for their business. In the Big Data field, the need for defining appropriate architectures is, if possible, even more urgent, on the basis of the recent nature of the adoption of Big Data sources as additional sources for the Official Statistics production.

The Italian National Institute of Statistics (Istat) has been investing in Big Data as new sources for Official Statistics since 2013, when the "Scheveningen memorandum" acknowledged Big Data to represent new opportunities and challenges for Official Statistics for the European Statistical System and its partners.

In this paper, with respect to the current Istat's situation, we will:

- summarize the major investments made in the field of Big Data architectures till now and
- highlight the major open challenges that we see as important to be solved in the next future for a full-fledged production of Big Data-based statistics.

2. CURRENT BIG DATA INVESTMENTS

2.1. On-premise Big Data IT Infrastructure

Istat decided to set up an in-house Big Data platform completed in January 2015. The platform, based on a Cloudera Enterprise distribution, has the following (main) technical characteristics:

- 8-node Hadoop Cluster (32/16 Cores CPUs,128 Gb RAM per node, 20Gbit internal connection, 6 x 1.2Tb hard drives per node);
- Hue (Hadoop User Experience) as Analytics Workbench;
- Apache Impala massively parallel processing (MPP) SQL query engine for fast querying of large amount of data;
- Apache Spark for advanced analytics;

In addition to the Cloudera-based Big Data infrastructure, Istat has also recently acquired a data center of GPUs (Graphics Processing Units) to support Deep Learning tasks. The data center is based on Nutanix technologies and inludes advanced GPUs (Nvidia Tesla P40)

The Cloudera platform is currently used for the following functions:

- Historical data storing, the principal project being the Use of scanner data for the production of the consumer price index.

- Data staging, involving two main projects, namely Land use\Land Cover estimation from aerial images and Population estimation from mobile phone data.

- Large scale analysis, mainly involving again the Use of scanner data for the production of the consumer price index.

The Nutanix platform is still in a set up phase. However, it is has been tested on some deep learning tasks (a computer vision task and a natural language processing task) with the excellent performances shown in Figure 1.

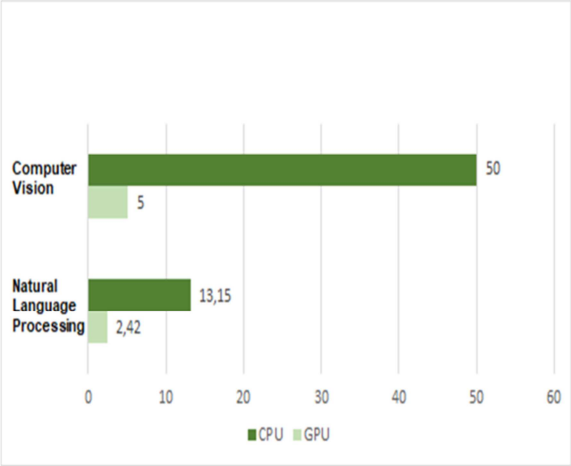

Figure 1. Example of performance gain of GPUs data center[1]

2.2. Organizational Structures

Istat does not currently have an organizational structure dedicated exclusively to Big Data management. Indeed, the group of people working on Big Data projects is composed by resources from three principal areas:

1. Methodological Directorate
2. IT Directorate
3. Production Directorates

The governance structure of the project is based on two levels:

- Strategic level: *Big Data Committee*, chaired by the Istat's President.
- Operational level: *Big Data Working Group*, *Big Data Initiatives*, grouped for monitoring purposes under a *Program Area*, and *Innovation Lab*.

All the three areas (Methodology, IT and Production) have Big Data Initiatives that are coordinated by the Big Data Working Group with respect to technical aspects. The outputs of the Big Data Initiatives are monitored by a dedicated Program Area. The

[1] More details at (in Italian): https://www.istat.it/img/poster2018/long/1.jpg

Istat's Innovation Lab is, instead, the place where the most cutting-edge projects are developed.

3. OPEN CHALLENGES

We identify the following five major challenges that are related to the IT infrastructure for Big Data usage within Istat.

Challenge 1: Access to the IT Infrastructure

There are different users of the IT infrastructure: system managers, IT application developers, data architect, data scientists. Defining the "boundaries" among such users can be quite difficult. For instance, given that Big Data technologies are rather new, the access to the lower levels of the platform may require to the IT developer to know the platform details, thus needing to do system management tasks. As a further example, the data scientist has often the necessity to play the data architect role in organizing data, thus working on database management tasks and not only on data analyses ones.

This situation makes hard for each category of users to have defined and clear functions to be granted to access.

Challenge 2: Maintenance of the IT Infrastructure and of deployed projects

The landscape of Big Data technologies evolve very fast, as well as the business models that Official Statisticians may adopt for Big Data. This situation makes very difficult the maintenance of infrastructures on-premise and also the maintenance of investments made on projects deployment. For instance, in the smart statistics evolution of Big Data management, data should reside more and more on producers' premises. This could mean for instance that the principal usage that we are doing of the Cloudera platform (historical data storing and data staging) could be downsized.

Challenge 3: Skill development

New methods and technologies, like those required by Big Data, need new skills, most of which are very much specific and hard to find. The challenge is not only to have training programs for internal personnel but also having access to private consultants to engage on a planned and regular basis.

Challenge 4: Infrastructure investments: on-premise vs cloud

The adoption of a cloud-based solution is of course an alternative but privacy issues are to be closely considered. Having strategies that even only occasionally permit the development of cloud-based solutions is also an option, e.g. for rapid prototyping activities.

Challenge 5: Communication and collaboration among people

From a people management perspective, Big Data teams are by definition multidisciplinary teams. If people are not under the same organizational structure, a competitiveness problem may arise and collaboration can get harder. Questions like which sectors lead which project and real commitments on operational projects are to be day-by-day discussed and answered.

4. CONCLUSIONS

The paper highlights the current status of Istat's investments on Big Data IT infrastructure and sketch the organizational structure that has been thought for Big Data projects.

We are now in the phase of elaborating a long-term strategy on how to progress from this state in the future. In particular, at this stage we can conclude that:

- The listed challenges draw a quite clear picture of what the strategy should address from an *internal* perspective.

- Some relevant *external* conditions are still in a definition phase, including strategies at ESS level (e.g. private-public partnerships, GDPR compliance by NSIs, etc.). However, Istat is currently involved in Eurostat-level groups and projects that are actively working in this respect.