Capacity Building for Big Data in Statistics: Experiences and Lessons Learned

Antonino Virgillito, Istat

About Me

- Senior IT Engineer at Istat
  - Focus on Big Data, Business Intelligence and Analytics
  - IT project leader of the project “Use of scanner data in the CPI survey”
- 15 years of experience in training on IT topics
- Former project manager of the UNECE Big Data Sandbox project
- Former adjunct professor and lecturer at University of Roma “Sapienza”

Learning to Work with Big Data

What are the main issues to face when defining a capacity building strategy for Big Data in Statistical Offices?

Learning to Work with Big Data: Issues

New topic - Everyone in the community is learning

- Cannot rely on other domains. Big Data in official statistics is a topic by itself
- The body of consolidated knowledge is continuously evolving and was built within the community
- Background information is required for everyone

Learning to Work with Big Data: Issues

Difficult to find representative data sets to work with

- Real (or realistic) data is difficult to get and normally is subject to privacy constraints.
- Datasets may be big and difficult to move and to treat.

Learning to Work with Big Data: Issues

Non-standard IT tools are required

- Broad range of different tools for different purposes
- Complex IT architectures, difficult (impossible) to set up during a normal training course
- Non familiar paradigms: statisticians are in general more confortable with desktop computing tools

Learning to Work with Big Data: Issues

Different needs for different skills

Statistician

What is special about Big Data?

How to get quick access to datasets?

IT

How to place big data tools in the IT architecture?

How to efficiently process large amounts of data?

Different Skills in Big Data

Data Analyst

Data Scientist

Data Engineer

Data Integrator

System Manager

R - SAS - SPSS

Pig

Map Reduce

ETL

Linux

BI and Visual Analytics

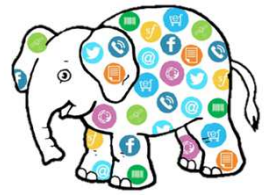
Java - Python

Excel

SQL

Difficult to organize a coherent capacity building strategy!

UNECE Big Data Project



2014 - Phase one of capacity building activity

- Training sessions in the the face-to-face meetings
- Training material available for all participants

2016 - The Sandbox as a consolidated training tool

- Access available on a subscription basis
- Use in training courses

Using the Sandbox for Training

The Sandbox is an effective platform for supporting training courses

It runs special software for high performance computing which **cannot be installed or run on standard computers**

Non-confidential demonstration datasets have been uploaded and shared

The software and the datasets are available from everywhere **only with a browser**

ESTP Courses on Big Data in 2016

Istat lead two courses on Big Data under the European Statistical Training Programme 2016

Introduction to Big Data and its Tools

Statisticians, managers, IT

Hands-on Immersion on Big Data Tools

IT

The Sandbox has been used in both courses to show and use the tools

The Sandbox in the “Hands-on” course

- All participants had a Sandbox account created once the course started
- No installation was required for tools in the Sandbox (about 70% of the course practical content) and sample datasets were immediately available
- The accounts were kept active after the course so that participants could experiment in their office

The Sandbox in the “Hands-on” course

- ✓ Quick access to browser-based tools with zero configuration
- ✓ Positive feedback from participants and quick adoption of new tools
- ✗ Tools that were not accessible through the browser were difficult to configure locally
- ✗ Concurrent use of the shared environment made things run slowly at times
- ✗ Difficult to prepare a realistic running example that used all the tools

Using Big Data Tools in a Real Project

- "Use of Scanner Data in the Consumer Price Survey"
  - First project at Istat using a real Big Data source in production
- Hybrid data architecture
  - Relational database holds current data while historical data is stored in a Hadoop cluster
  - Current size of the whole dataset around 1Tb – and growing
- Statisticians mainly used SAS for analysis
  - Problem: could not work easily on the data because of its size

Using Big Data Tools in a Real Project

- ✓ Big Data tools lead to a significant improvement of processing time on large data sets
- ✓ IT staff could easily switch on new tools (mainly thanks to use of SQL)
- ✗ Data access by statisticians requires IT intervention
- ✗ Direct SAS-Hadoop interface possible but with expensive licensing
- ✗ Difficult to convert statisticians to different ways of accessing data (SQL, BI)

Conclusions

- Capacity building for Big Data has been an important part of the community effort on this topics during last years
- The Sandbox has been effectively used as a supporting platform for training courses
  - Fast, convenient solution for testing complex tools
- It is crucial to recognize the differences in skills
  - Convergence toward the “Data Scientist” is difficult to achieve in practice…
- In particular, for real production processes, training about tools is not sufficient: an reasoned planning of the data access strategy is required first