

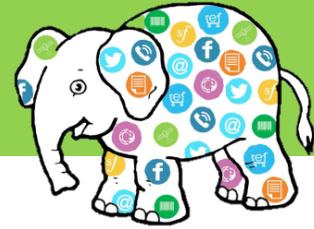


United Nations Economic Commission for Europe
Statistical Division

Workshop on the Modernisation of Statistical Production and Services

November 19-20, 2014

The Role of Big Data in the Modernisation of Statistical Production and Services



The Sandbox Task Team

Antonino Virgillito Project Consultant, UNECE

BACKGROUND

EXPERIMENTS

FINDINGS

DISCUSSION





Task Team Objectives

The Project Proposal defines 5 specific objectives for stating successful completion of the work

1 'Big Data' sources can be obtained, installed and manipulated with relative ease and efficiency on the chosen platform

2 The chosen sources can be processed to produce statistics (either mainstream or novel) which conform to appropriate quality criteria

3 The resulting statistics correspond in a systematic and predictable way with existing mainstream products

4 The chosen platforms, tools, methods and datasets can be used in similar ways to produce analogous statistics in different countries

5 The different participating countries can share tools, methods, datasets and results efficiently, operating on the principles established in the Common Statistical Production Architecture



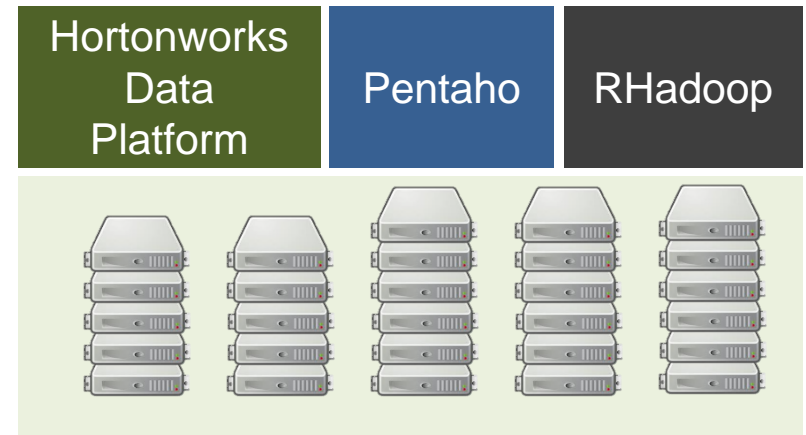
The Sandbox

Shared computation environment for the storage and the analysis of large-scale datasets

Used as a platform for collaboration across participating institutions

Created with support from:

- **CSO** Central Statistics Office of Ireland
- **ICHEC** Irish Centre for High-End Computing



Cluster of 28 machines
 Accessible through web and SSH
 Software: full Hadoop stack, visual analytics, R, RDBMS, NoSQL DB

Objectives

- Explore tools and methods
- Test feasibility of producing Big Data-derived statistics
- Replicate outputs across countries



Partnerships



ICHEC Assisted the task team for the testing and evaluation of Hadoop work-flows and associated data analysis application software



Hortonworks provided free support on their open source Hadoop platform (**Hortonworks Data Platform**)



Trial license of the **Pentaho Enterprise Platform** product for visual analytics were provided by Italian distributor **BNova**

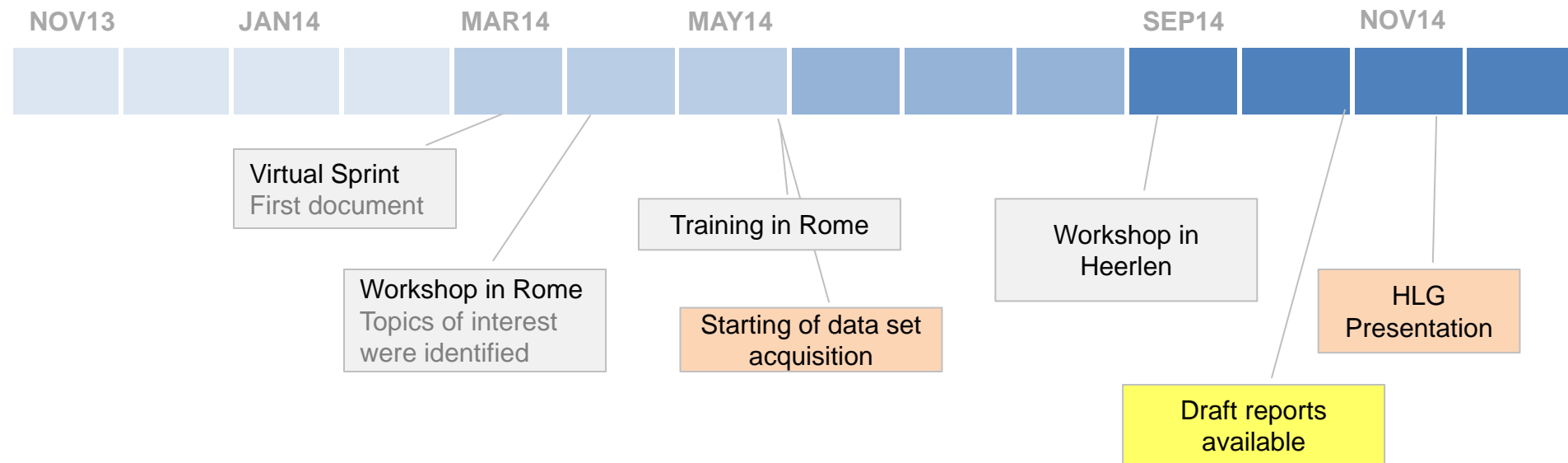




Task Team Overview

The Sandbox Task Team is composed of 38 participants from 18 among national institutes of statistics and international organizations.

The team is organized around a set of “experiment teams”, focusing on topics related to different statistical domains.



BACKGROUND

EXPERIMENTS

FINDINGS

DISCUSSION





Social Media



Job Vacancies Ads



Mobile Phones



Web Scraping



Prices







Traffic Loops



Smart Meters

Each experiment team produced a detailed report on its activity, already available in draft format on the wiki

A summary of the results is presented in the following

-  Positive indication
-  "Mixed" indication
-  Negative indication
-  More work needed / ongoing



Social Media: Mobility Studies

Countries



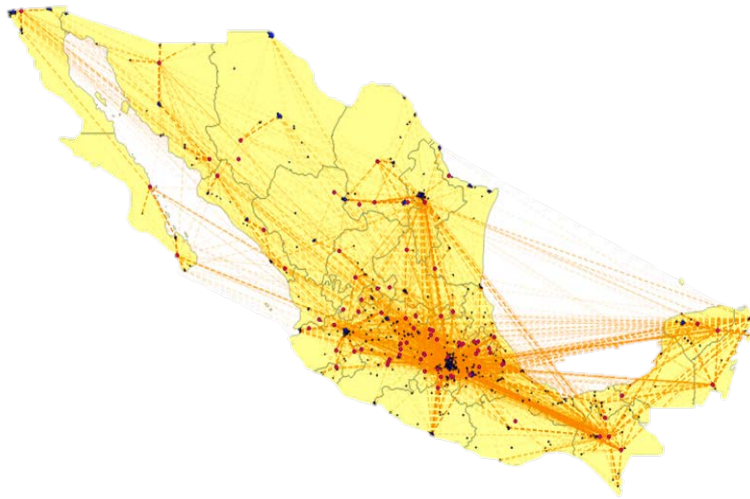
Dataset

records	42M
size	9.2Gb

Tweets generated in Mexico
Jan14/Jul14

Analysis of mobility starting from georeference data of single tweets

Patterns of mobility to touristic cities



Trans-border mobility



✓ Mobility statistics computed at detailed territorial level



Social Media: Sentiment Analysis

Countries



Dataset

records
42M

size
9.2Gb

Tweets generated
in Mexico
Jan14/Jul14

Derived sentiment indicator from analysis of Mexican tweets

Emoticons and media acronyms

- ✓ Statistic Nederlands applied its methodology to relate sentiment to consumer confidence
- ✗ Correlation is not as good as in previous study based on Dutch data
- ? More accurate, language-based computation of sentiment currently carried out in Mexico, based on partnership with university

Cross-country sharing of method

- Only emoticons were considered
- Dutch study also used Facebook as a source



Mobile Phones

Countries



Dataset

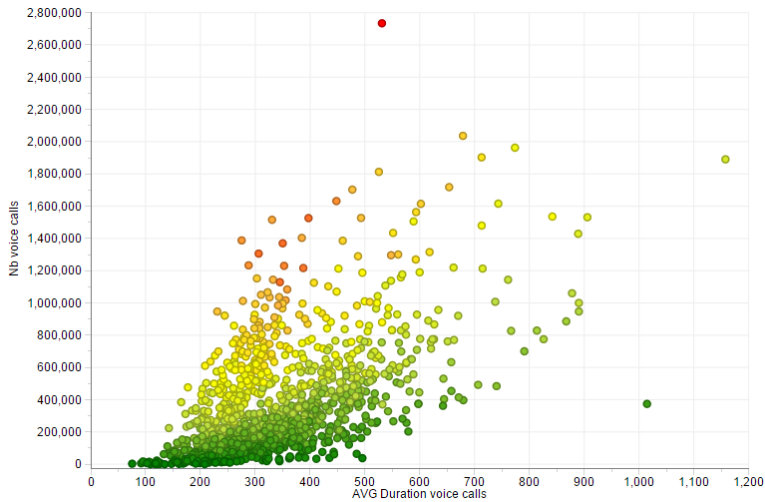
records
865M
size
31.4Gb

Four datasets from Orange. Call data from Ivory Coast.

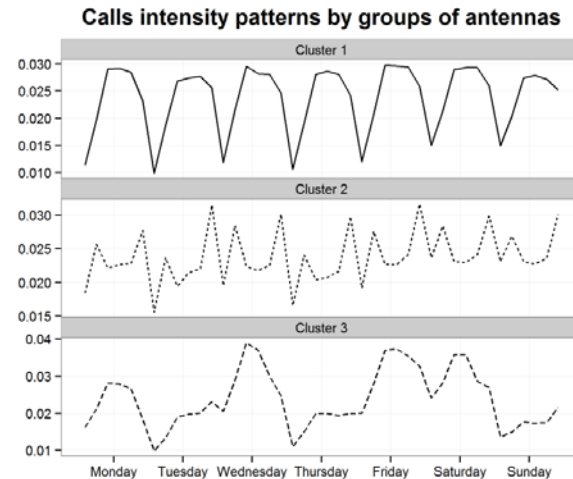
Experiments lagging behind because of long time required for data acquisition: 5 months!

Some preliminary output already in place as well as a plan of future experiments

Visual analysis of call location data



User categories from call intensity patterns





Mobile Phones: Findings

- ✓ Disaggregated data would allow to replicate mobility surveys with a higher level of detail
- ✗ Aggregated data only allows to produce parallel indicators that integrate traditional sources
- ? However, several methods have been proposed. Need more time for implementation
- ? Comparison with real data used in Slovenia will allow to give more responses about use in statistics and cross-country sharing



Consumer Price Index

Countries



Dataset

records
11G

size
260Gb

Synthetic scanner data

Test performance of big data technologies on big data sets through the computation of a simplified consumer price index on synthetic price data

Comparison between “traditional” and Big Data technologies

✓ Could write index computation script with one of the high-level languages part of Hadoop environment

✓ Big Data tools are necessary and achieve good scalability when data grow over tenth of Gb

Future work on methodology Work on scanner data is active in several NSIs. Data has same structure and methods can be shared. Novel statistics can be computed working on large scale data (no sampling)



Smart Meters

Countries



Dataset

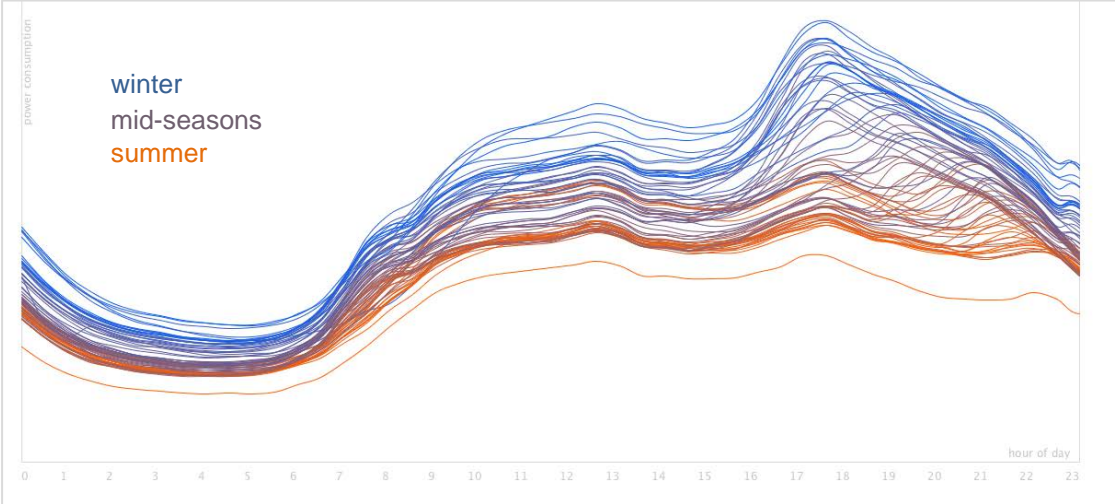
records
160M
size
2.5Gb

- Real data from Ireland
- Synthetic data from Canada

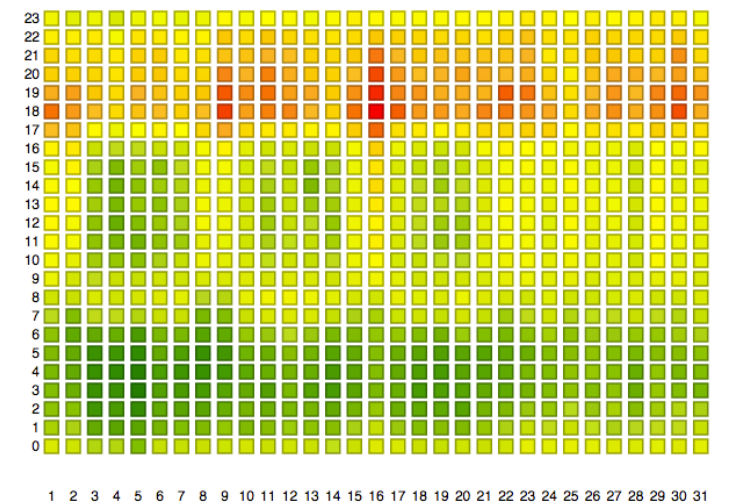
Test of aggregation using Big Data tools

Future work on sharing methods through the use of synthetic data sets

Weekly consumption per hour of day over a year (IE)



Hourly consumption per day (CAN)



Quickly wrote aggregation scripts that could be used on both datasets



Job Vacancies

Countries



Dataset

records
10K/day
size
2Mb/day

Collected data from job web portals

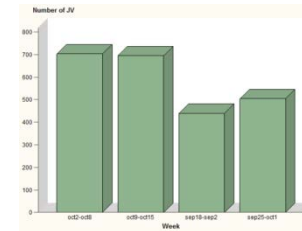
Set up continuous daily collection of data from job web portals to compute indices of statistics on job vacancies

Identified possible free and commercial data sources in different countries. Tested different techniques for data collection and methodologies for data cleaning

✓ **Timeliness.** Set up a process that collects and cleans data automatically. Computed the statistics on a weekly basis.

! **Coverage.** Collected sources were limited by the capability of the tools used and the structure of the web sites.

✗ **Coverage.** Sources could not guarantee all the variables that are necessary for computing the official job vacancy indicator.



Can be used for different - simplified - indicator, integration with other sources, benchmark.



Web Scraping

Countries



Dataset

records
-

size
?

Websites of Italian enterprises

Test of automated, massive datasets mining of text data extracted from the web

8,600 Italian websites, indicated by the 19,000 enterprises responding to ICT survey of year 2013, have been scraped and the acquired texts have been processed

✓ Sandbox approach resulted in significant performance improvement over the use of a single server

✓ A comparison of different solutions for extraction of data from the web, with recommendation about their use, has also been produced



Traffic Loops

Countries



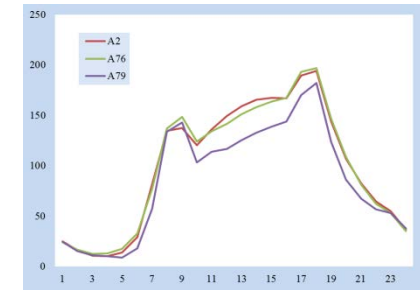
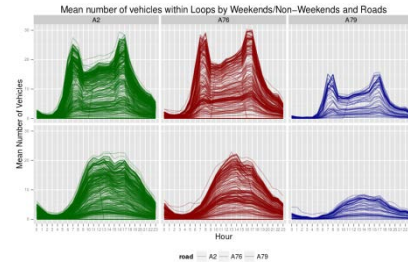
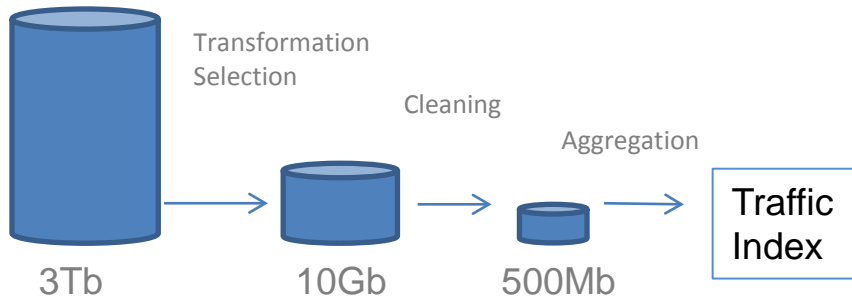
Dataset

records
156G
size
3Tb

Data from 20,000 traffic loops located on 3,000 km of speedway in the Netherlands

CBS will carry out the first test of the use of Sandbox for pre-production statistics

Experiments on aggregation, cleaning and imputation have been also conducted on a subset of data



The entire traffic dataset has been loaded in the Sandbox

- A disk had to be physically shipped in Ireland because dataset size did not allow network transfer

BACKGROUND

EXPERIMENTS

FINDINGS








DISCUSSION





Statistics








We showed some of the possible improvements that can be obtained using Big Data sources

							
Cheaper	✓	✓	✓	✓	✓	✓	✓
More timely	✓			✓	✓	✓	
Novel	✓	✓	✓	✓		✓	✓



Statistics

The results of the experiments were evaluated with respect to the initial objectives

							
SO1 Sources collection and manipulation	!	!	!	!	✓	✓	!
SO2 Production of quality statistics	!	✗	✓	✓	!	?	✓
SO3 Correspondence with existing products	✓	✓	✓	✓	!	?	✓
SO4 Cross-country sharing	!	?	✓	✓	✓	?	?
SO5 CSPA-based sharing	✓	✓	✓	✓	✓	✓	✓

Details on the evaluation can be found in the Project Summary Report



Skills

All available tools were used in the experiments by both researchers and technicians with no previous experience

The Sandbox can represent a capacity building platform for participating institutions

Projects in planning were less likely to use tools generally associated with “Big Data”. Often this decision was made due to a lack of familiarity with new tools or a deficit of secure “Big Data” infrastructure (e.g. parallel processing no-SQL data stores such as Hadoop).

UNSD Big Data Questionnaire

At present there is insufficient training in the skills that were identified as most important for people working with Big Data

Skills on Hadoop/NoSQL DBs indicated as “planned in the near future” by majority of organizations

UNECE Big Data Questionnaire

Crucial for building “data scientist” skills



Technology

- Big Data tools are necessary when dealing with data ranging from hundreds of Gb on
 - Effective starting from tenths of Gb
 - “Traditional” tools perform better with smaller datasets
- Researchers/technicians should be able to master different tools and be ready to deal with immature software
 - Highly dynamic situation with frequent updates and new tools spawning frequently
- Need strong IT skills for managing the tools
 - Support from software companies might be required in early phases



Acquisition

- 7 datasets were loaded
 - Initial project proposal required “one or more”
- Difficult to retrieve “interesting” (i.e., meaningful, disaggregated...) datasets
 - Privacy and size issues
- This also applies to web sources that are only apparently easy to retrieve
 - Issues with quality, in terms of coverage and representativeness



Sharing

- Naturally achieved sharing of methods and datasets according to CSPA principles
 - Big Data tools are de-facto standardized so not much need for CSPA-based integration
- Many data sets have the same form in all countries
 - Methods can be developed and tested in the shared environment and then applied to real counterparts within each NSI
- Privacy constraints on datasets limit the possibility of sharing
 - Can be partly bypassed through the use of synthetic data sets



Recommendations for Extension

- Continuation of experiments
 - Consolidated technical skills that now can be used more effectively in experiments
 - Some experiments started late. Need more time for full development
- More datasets can be loaded
 - Satellite data, transport data, ?...
- Possibility of testing new models of partnership
 - Moving data is too difficult. Why not trying to involve partners in running *our* programs on *their* data in *their* data centers?

The Role of Big Data in the Modernisation of Statistical Production and Services

BACKGROUND

EXPERIMENTS

FINDINGS

DISCUSSION

