



## 2<sup>nd</sup> Global Conference on Big Data for Official Statistics

20-22 October, 2015

# Big Data Sandbox

## Antonino Virgillito

Project manager “Big Data Project”  
UNECE

Head of Unit “Architectures for Business Intelligence, Mobile and Big Data”  
Istat



# BACKGROUND

## THE BIG DATA PROJECT

## THE SANDBOX INITIATIVE



# Big Data for Official Statistics

- The new opportunities for official statistics that are offered by Big Data are well known
  - New and more detailed statistics
  - New and faster ways of production
  - Additional sources of data
  - Very timely statistics
- Preliminary experiences carried out by the statistical community in this field highlighted several common issues
  - Difficulties to access to novel data sources
  - New methods required for handling non-traditional, non-structured sources
  - New tools for storing, accessing and processing datasets that reach levels of scale previously unseen in our organizations

# Challenges for Organizations

“We know we have to work on Big Data... but we don't know where to start!”

“We have data but... we do not have skills and/or backing IT infrastructure for properly handling them”

“We want to know what other organizations are doing in the field”

“We already have skills and experience and we want to share them”

# Big Data Technology

- The size of the datasets implied in Big Data scenarios requires using specialized IT platforms, based on distributed technology
  - Needed for processing datasets in the Tb order of magnitude
  - “Real” Big Data begin where your usual tools fail...
- Big Data infrastructures and tools are:
  - Complex to deploy and maintain for the IT sector
  - Costly to start with (order of 100k€, minimum)
  - Difficult to learn for statisticians

UN Global Survey on  
Big Data

**Statistical community** expressed  
**need for guidance** in the area of  
**skills and training**

# BACKGROUND

## THE BIG DATA PROJECT

### THE SANDBOX INITIATIVE



# The HLG Big Data Project

- Overseen by the High-Level Group for the Modernisation of Official Statistics (HLG-MOS)
- Objectives
  - to identify the main possibilities offered by Big Data to statistical organizations
  - to demonstrate the feasibility of efficient production of both novel products and 'mainstream' official statistics using Big Data sources
  - to explore new tools and new methods
- 75 participants from 20 Organizations
  - National Statistical Offices and International Organizations
- Two years of activity
  - 2014: Experiments on different data sources
  - 2015: Production of Multi-national statistics only basing on Big Data sources
- Focus on practical work on Big Data sources using a shared computing platform
  - the “Sandbox”

# What is the Sandbox

- Shared computing environment
  - developed in partnership with the Irish Central Statistics Office and the Irish Centre for High-End Computing (ICHEC)
  - hosted by ICHEC
- A unique platform where participating organisations can engage in collaborative research activities
- Open to all producers of official statistics



# What is the Sandbox

## Installed Software

- Hadoop (Hortonworks Data Platform)
- R – Rstudio
- RHadoop
- Spark
- ElasticSearch

*and growing...*



### *4 Data/Compute nodes*

- 2 x 10 core Intel Xeon CPUs
- 128 GB RAM
- 4 x 4TB disks
- 56 Gbit InfiniBand network

### *2 Service/login nodes*

- Similar hw as data nodes
- 10Gbit connection to Internet

# The Sandbox – Web interface

The screenshot displays the Big Data Sandbox web interface, which is divided into two main sections: a File Browser and a My Scripts editor.

**File Browser:** This section shows a directory structure under the path `/datasets`. It includes a search bar and various action buttons such as `Rename`, `Move`, `Copy`, `Change Permissions`, `Download`, `Delete`, `New`, and `Upload`. A table lists the contents of the directory:

Type	Name	Size	User	Group	Permissions	Date
Folder	.		hdfs	hdfs	drwxrwxrwt	October 28, 2014 05:59 pm
Folder	..		hdfs	hdfs	drwxrwxr-x	February 05, 2015 04:01 pm
Folder	job_portals		toni	hdfs	drwxr-xr-x	February 04, 2015 11:30 pm
Folder	mobile_phones		toni	mobilephones	drwxr-xr-x	October 15, 2014 03:58 pm
Folder	prices uk					
Folder	prices_scannerdata					
Folder	smart_meters					
Folder	smart_meters_can					
Folder	traffic_loops					
Folder	traffic_loops_full					
Folder	tweets mx					

**My Scripts:** This section allows users to create and manage Pig scripts. It features a `My scripts` list with a `NEW SCRIPT` button. The list includes scripts such as `batting`, `index_test`, `jp_base`, `laspeyres_index`, `nyse test`, `price it test`, `price uk clean`, `sm_can_aggr_dayhour`, `sm_can_base`, `sm_ie_aggr`, `sm_ie_aggr_daystat`, `sm_ie_aggr_weekhour`, and `sm_ie_count`. Below the list are `Settings` for email notifications and a section for `USER-DEFINED FUNCTIONS` with an `Upload UDF Jar` button.

The `My Scripts` editor shows a script titled `sm_ie_aggr_dayst` with the following Pig script:

```

1 REGISTER datafu-1.2.0.jar
2
3 define VAR datafu,pig,stats,VAR();
4
5 sm = load '/datasets/smart_meters/electricity/data' using PigStorage(' ')
6 as (meter: chararray, timestamp: chararray, reading: double);
7
8 daygrp = group sm by SUBSTRING(timestamp, 0, 3);
9
10 daystat = foreach daygrp generate group as day, SUM(sm.reading) as sumcon
11 , AVG(sm.reading) as avgcon, VAR(sm.reading) as varcon;
12
13
14 store daystat into '/user/toni/sm_ie_daystat';
  
```

At the bottom of the editor, there are buttons for `Save`, `Execute`, `Explain`, and `Syntax check`.

# The Sandbox – RStudio web access

The screenshot displays the RStudio web interface. The main editor shows R code for a MapReduce job. The Environment pane is empty. The Files pane shows a directory structure with files like .Rhistory, rhadoop test.Rproj, rhadoop-ex1.R, and rhadoop-ex2.R.

```

7 rmr.options("backend.parameters")
8
9
10
11 input.file = '/user/toni/Batting-noh.csv'
12 output.file = '/user/toni/Batting-totalGames3.csv'
13
14 btt.map=function(k,v) keyval(v[,1], v[,6])
15
16 btt.reduce=function(kk,vv) keyval(kk,sum(vv))
17
18 mr.result=mapreduce(
19   input=input.file,
20   input.format=make.input.format("csv",sep=","),
21   output=output.file,
22   output.format=make.output.format("csv",sep=","),
23   map=btt.map,
24   reduce=btt.reduce)
  
```

Console output:

```

~/rstudio/rhadoop test/
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
  
```

Name	Size	Modified
..		
.Rhistory	0 B	Oct 6, 2015, 11:20 PM
rhadoop test.Rproj	205 B	Oct 6, 2015, 11:18 PM
rhadoop-ex1.R	268 B	Jun 8, 2015, 7:38 AM
rhadoop-ex2.R	706 B	Sep 1, 2015, 12:47 PM

## Scope of the Sandbox

The initial focus was to use the Sandbox to understand the importance of Big Data for official statistics

...however, it soon became obvious that it has many more uses

# BACKGROUND

## THE BIG DATA PROJECT

### THE SANDBOX INITIATIVE



# Beyond the Big Data Project

- Current initiative to extend access to the Sandbox beyond the current project (December 2015)
  - Based on strong interest from a number of statistical organisations
- ICHEC is willing to continue to provide the sandbox as a service to the international statistical community, on a non-profit basis
- Users will be required to pay an annual subscription to cover the costs of technical support, hardware upgrades and installation of software

# The Sandbox: Use Cases

## Running experiments and pilots

The sandbox can be used for experiments involving creating and evaluating **new software programmes**, developing **new methodologies** and exploring the potential of **new data sources**

This use case extends the current role of the sandbox beyond Big Data, and encompasses **all types of data sources**

# The Sandbox: Use Cases

## Testing



**Setting up and testing** of statistical pre-production processes is also possible in the sandbox, including simulating complete workflows and process interactions



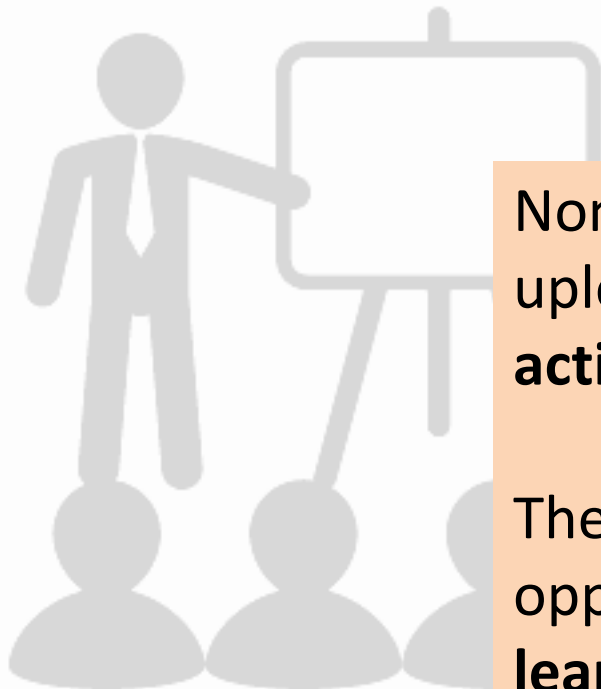
# The Sandbox: Use Cases

## Training

The sandbox can be used as a platform for supporting training courses. It can run special software for high performance computing which **cannot be installed or run on standard computers**

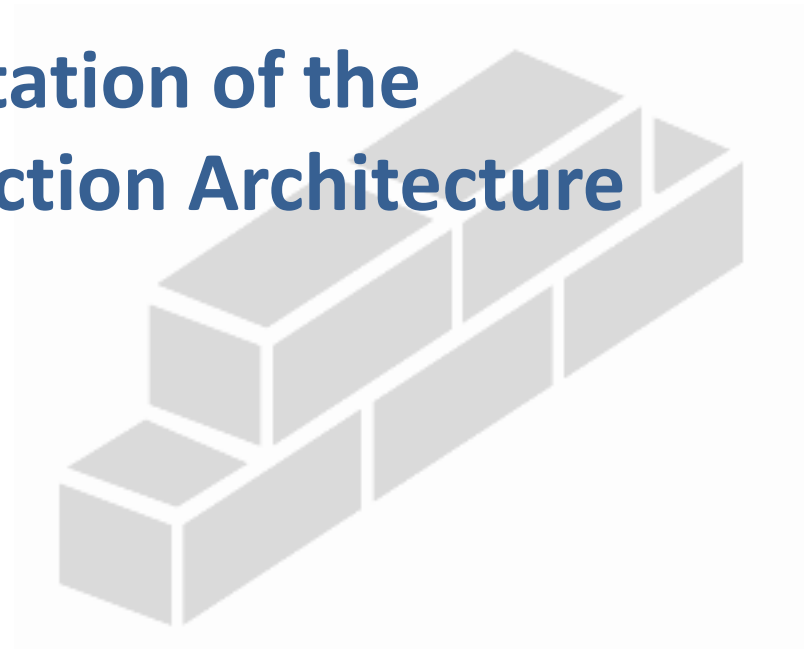
Non-confidential demonstration datasets can be uploaded and shared, facilitating **shared training activities** across organisations

The sandbox environment also allows statisticians opportunities for self-learning, e-learning and **learning by doing**



# The Sandbox: Use Cases

## Supporting the implementation of the Common Statistical Production Architecture (CSPA)



The sandbox can be used as a statistical laboratory where researchers can **jointly develop and test** new CSPA-compliant software

# The Sandbox: Use Cases

## Data Hub



The sandbox also provides a **shared data repository** (subject to confidentiality constraints)

It can be used to **share non-confidential data sets** that cover multiple countries, as well as public-use micro-data sets

## What you pay

Subscription fee **10k€ per year**

## What you get

- Access to the Sandbox, shared tools, datasets and other resources for your staff
- Access to international collaboration projects and opportunities
- Technical support to ensure the Sandbox infrastructure is kept up to date and relevant to your needs

# How to Subscribe

- To subscribe, please send an expression of interest (a simple e-mail is sufficient) to [support.stat@unece.org](mailto:support.stat@unece.org)
- Each subscriber will be given a seat on the Strategic Advisory Board, a new group which will oversee Sandbox operations
  - collectively decide, in consultation with ICHEC, on subscription levels and priorities for expenditure on software and hardware
- Any organisation producing official statistics can subscribe to the Sandbox
  - Other organisations may be considered on a case-by-case basis subject to the approval of the Strategic Advisory Board
- Because the Sandbox activities are closely linked to the work of the HLG-MOS, the UNECE has been asked to facilitate contacts between the official statistical community and ICHEC, and support the functioning of the Strategic Advisory Board

# Conclusions

- The Sandbox represents **the fastest route** available to statistical organizations **for starting with Big Data**
- It offers several features that facilitate the approach to Big Data and data science
  - An infrastructure for big data processing ready for being used at a low subscription cost
  - Software already installed and proved/tested
  - Shared datasets instantly available
  - Tools and material for capacity building
- It is driven by the community
  - Worldwide network of organizations, with a catalogue of initiatives
  - A place for collaboration on methods and products, not only related to Big Data