Italian Experience and Perspective of Using Big Data to Estimate Inflation

Federico Polidoro (*Istat, polidoro@istat.it*)
Antonino Virgillito (*Istat, virgilli@istat.it*)

CPI Expert group meeting
Geneve, 2-4 May 2016

Outline of the Presentation

- The main features of the Italian consumer price survey re-design
- Scanner data as new data source: main statistical issues
- Web scraping as new technique to capture the data: statistical issues
- The new IT architecture supporting the collection and usage of scanner data
- Future steps
- Conclusions and aims

The Re-Design of the Italian Consumer Price Survey: Main Features

The present situation of the survey (2016)

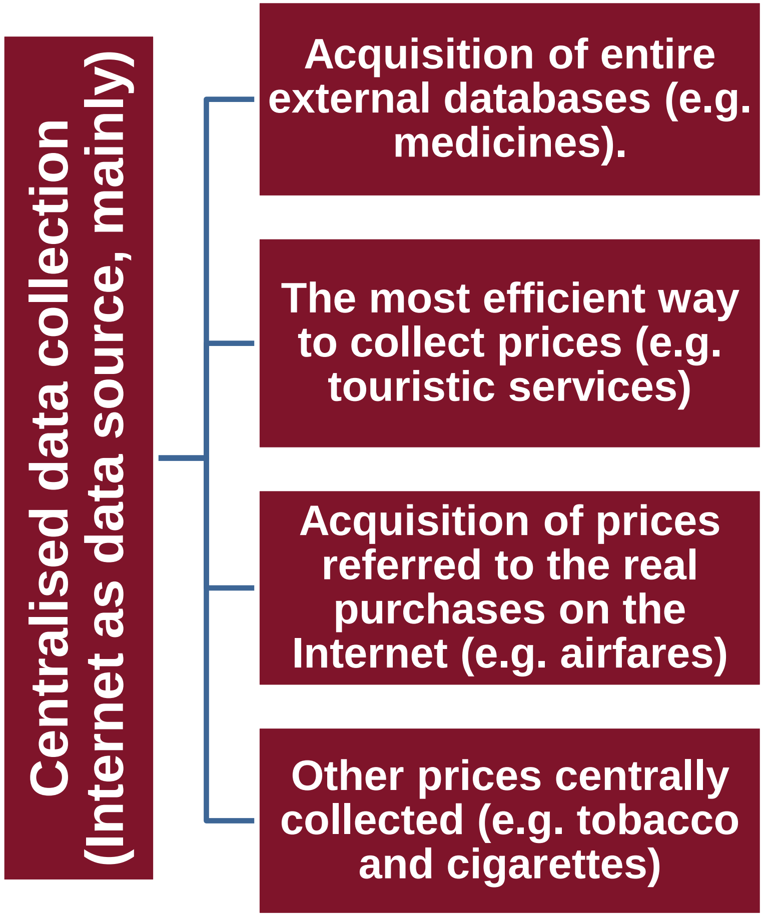
Two ways to collect the data to estimate Italian inflation

Traditional way: 350 data collectors in 80 province chief towns. 76.4% of the basket of products in terms of weights

Centralised data collection (Istat): 23.6% of the basket of products in terms of weights. Mainly internet as data source

The Re-Design of the Italian Consumer Price Survey: Main Features

The present situation of the survey (2016)



Centralised data collection (Internet as data source, mainly)

Acquisition of entire external databases (e.g. medicines).

The most efficient way to collect prices (e.g. touristic services)

Acquisition of prices referred to the real purchases on the Internet (e.g. airfares)

Other prices centrally collected (e.g. tobacco and cigarettes)

The Re-Design of the Italian Consumer Price Survey: Main Features

The aims of the re-design

- ✓ Reducing to around 50% the weight of the traditional data collection through
  1. Expanding the use of Internet as data source widening the use of web scraping techniques
  2. Using scanner data as new source
  3. Enlarging the use of administrative data
- ✓ Reviewing the sample design of the survey moving towards a probabilistic approach
  1. Exploiting the potentialities of scanner data information
  2. Enhancing the use of web scraping techniques as a tool to access big data

Scanner Data as New Data Source:
Main Statistical Issues

Accessing the data: partnership with large scale retailers

- ✓ The agreement (reached through GS1 – Italy) with ADM (Association of Modern Distribution) the main representative of large scale retailers
- ✓ Istat sent (in 2014) the request of the data (with formal and legal meaning) to the most important chains of the retail trade distribution, asking them for scanner data for their outlets
- ✓ The chains authorized Nielsen Italy to send the data to Istat
- ✓ In parallel Istat reached an agreement with Nielsen
- ✓ A partnership with large scale retailers was the key to access scanner data in a partially fragmented market

Scanner Data as New Data Source: Main Statistical Issues

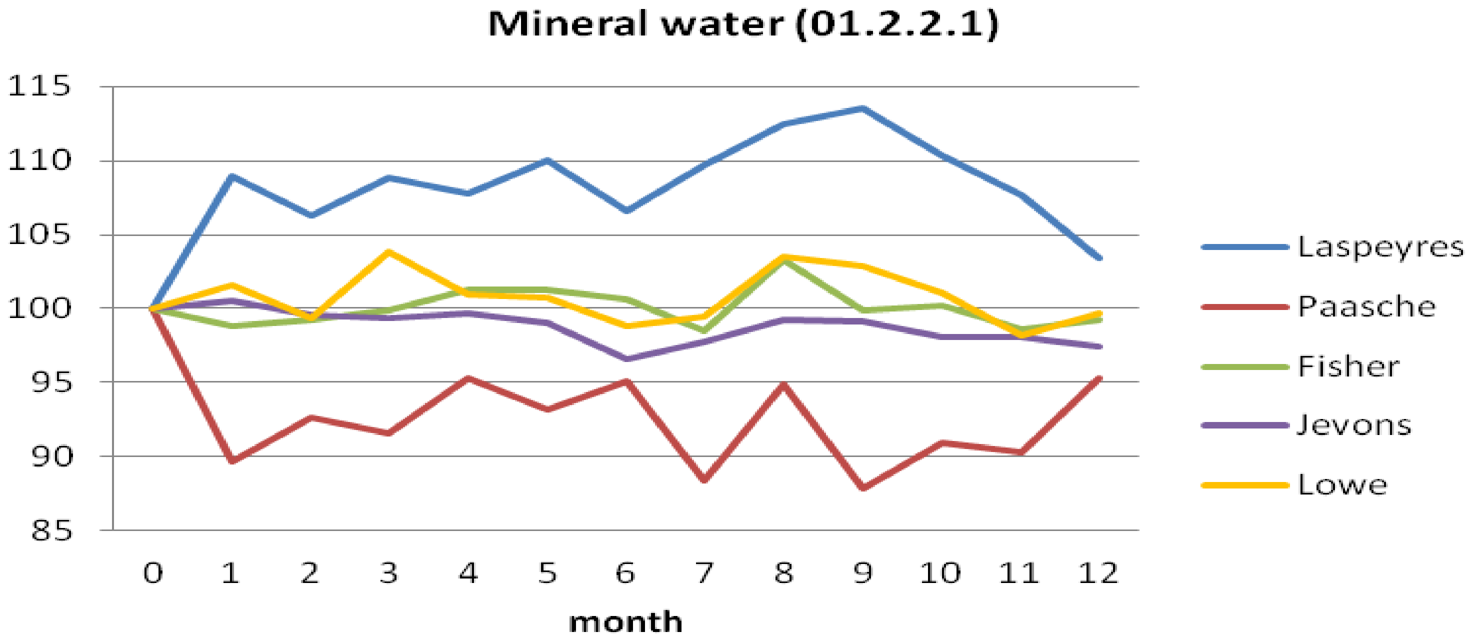
Current scanner data set available

- ✓ Data of Coop Italia, Conad, Selex, Esselunga, Auchan, Carrefour (in total almost 57% of the turnover of modern distribution)
- ✓ Nielsen sent and is going on to send to Istat, data concerning grocery products, for 37 Italian provinces, time series of the last 24 months with weekly data (turnover and quantity) referred to each outlet
- ✓ Now 1 billion and three hundred millions records
- ✓ All the ongoing activities are monitored within the framework of the partnership with ADM, that participated in the workshop on scanner data, organized by Istat and Eurostat (1-2 October 2016)

Scanner Data as New Data Source: Main Statistical Issues

The main methodological issues

- ✓ Scanner data bring detailed information about quantity and turnover. Is it possible to overcome the use of Jevons formula to aggregate elementary indices referred to each EAN code ?

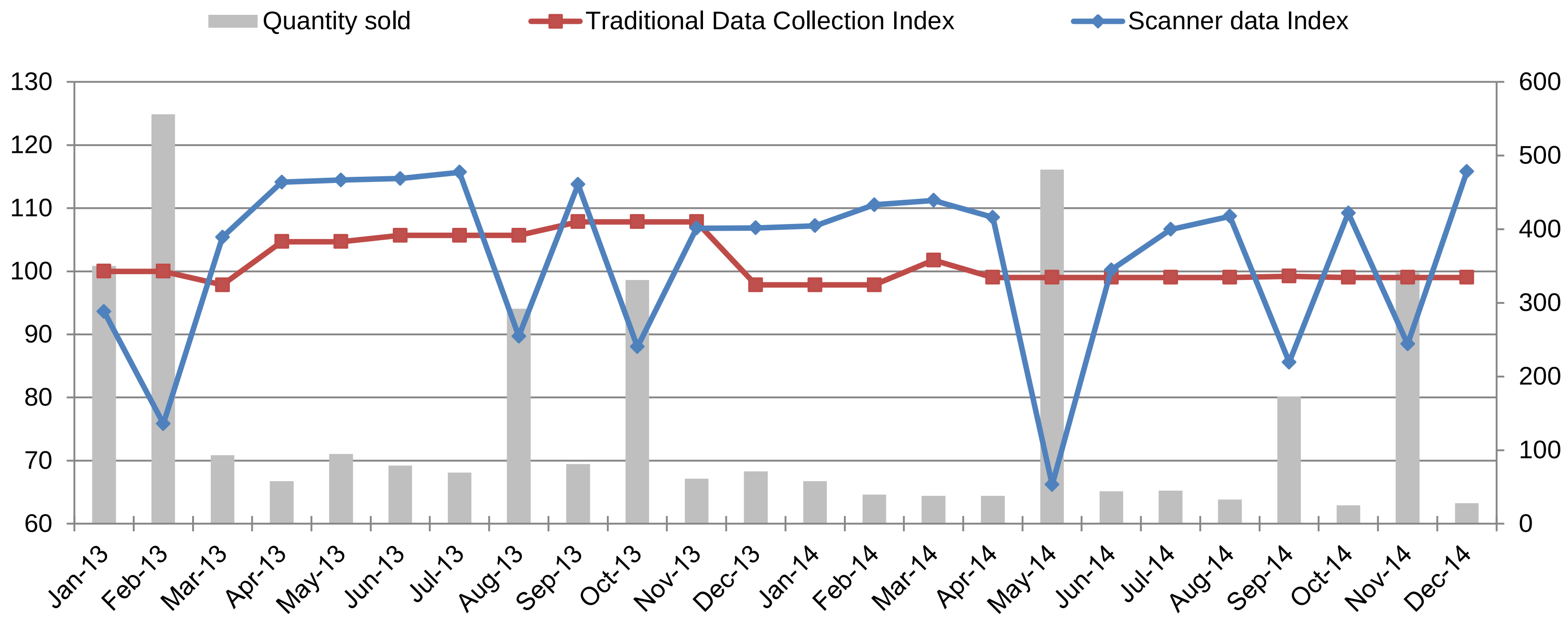


Scanner Data as New Data Source: Main Statistical Issues

The main methodological issues

✓ What will the impact be on the national CPI, taking into account the reactivity of the demand to price temporary reductions ?

Micro indices referred to one EAN code of roasted Coffee in one outlet (Torino)



Web Scraping as a New Technique to Capture Data: Statistical Issues

The preliminary experiences in the field

- ✓ In a first stage, experimental activities in the field of consumer electronics products and air fares
- ✓ The aim was to replace the activity of data collection carried out by Istat team with tools that reproduce it automatically
  - same statistical rules, the same calendar of data collection, the same pre-defined sample of elementary items
- ✓ Results obtained:

For consumer electronics: reduction of working time, improved accuracy, more elementary quotes

For air fares: uncertain results and limited improvements

Web Scraping as a New Technique to Capture Data: Statistical Issues

Towards a big data perspective

- ✓ Web scraping techniques to change the survey design with reference to the web channel of commercial distribution in a logic of big data
- ✓ Starting the test of robots able to download big amount of data, better covering the universe of commercial transactions

Web Scraping as a New Technique to Capture Data: Statistical Issues

Towards a big data perspective

- ✓ In the coming months, test of robots collecting prices of air fares on a daily basis for a sample of routes, on one side keeping the same temporal distance from a moving departure date and on the other side shortening day by day this distance keeping fix the date of departure
- ✓ Expected results: reducing the present temporal volatility of air fares indices, that appears to be mainly due to calendar reasons (in addiction to the seasonal ones)
- ✓ Technical challenges
  - Automatic injection of parameters in web sites (i.e. the tool can query different dates in a same run)
  - Unassisted recognition of prices information on the page (adaptation to changes in web sites without manual intervention)

The New IT Architecture Supporting the Collection and Usage of Scanner Data

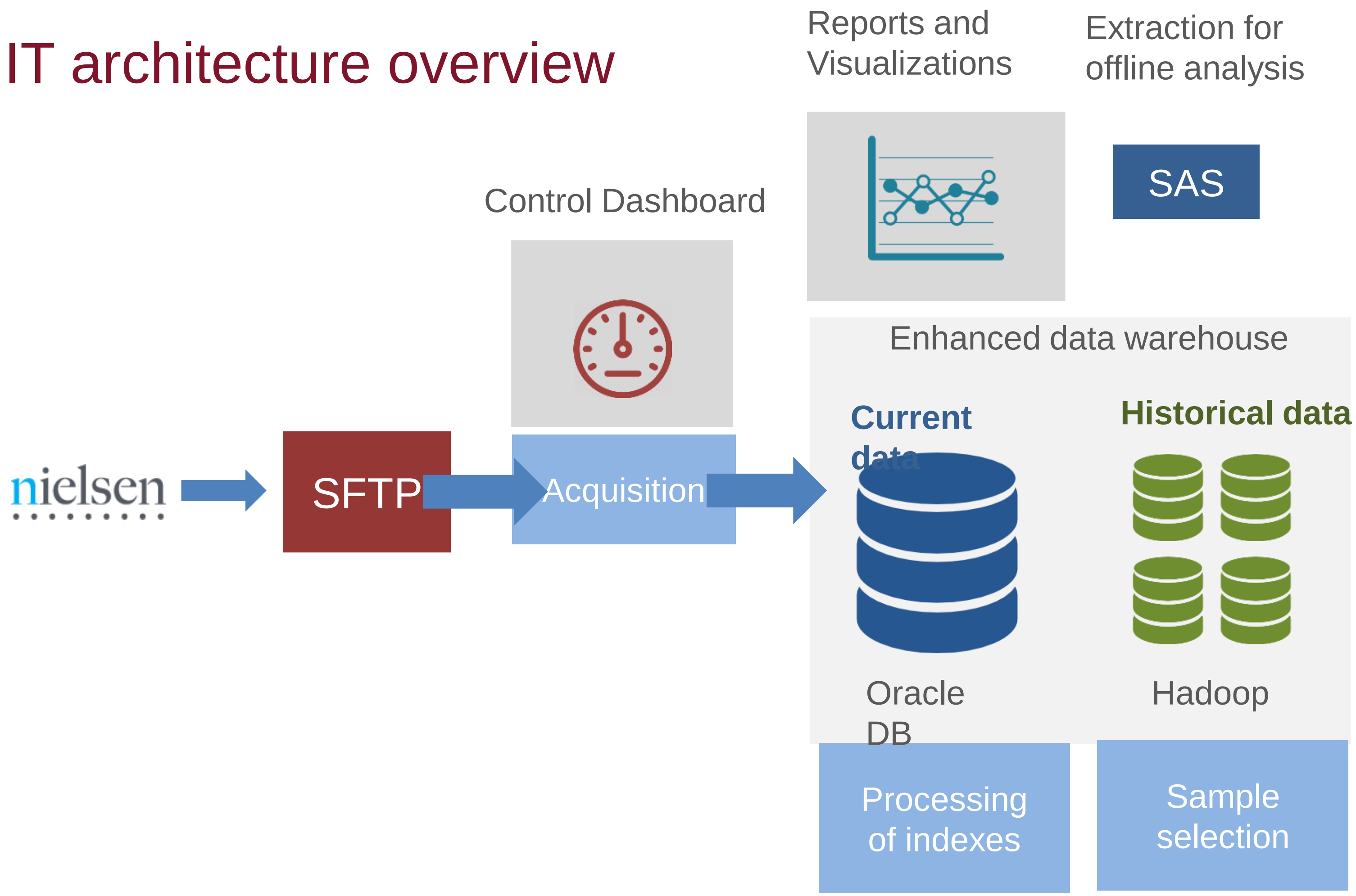
IT architecture

- ✓ The continuous flow of data generates a large dataset whose global size will constantly grow in an unpredictable way, creating a number of problems
  - ✓ Difficult to analyze and process with traditional methods
  - ✓ Impossible to determine an upper bound for the disk space required

- ✓ An “hybrid” data architecture has been set up, mixing different kinds of tools
  - ✓ Relational database
  - ✓ Statistical software
  - ✓ Business intelligence - visualization
  - ✓ Ad-hoc development

The New IT Architecture Supporting the Collection and Usage of Scanner Data

IT architecture overview



nielsen

SFTP

Control Dashboard

Acquisition

Reports and Visualizations

Extraction for offline analysis

SAS

Enhanced data warehouse

Current data

Historical data

Oracle DB

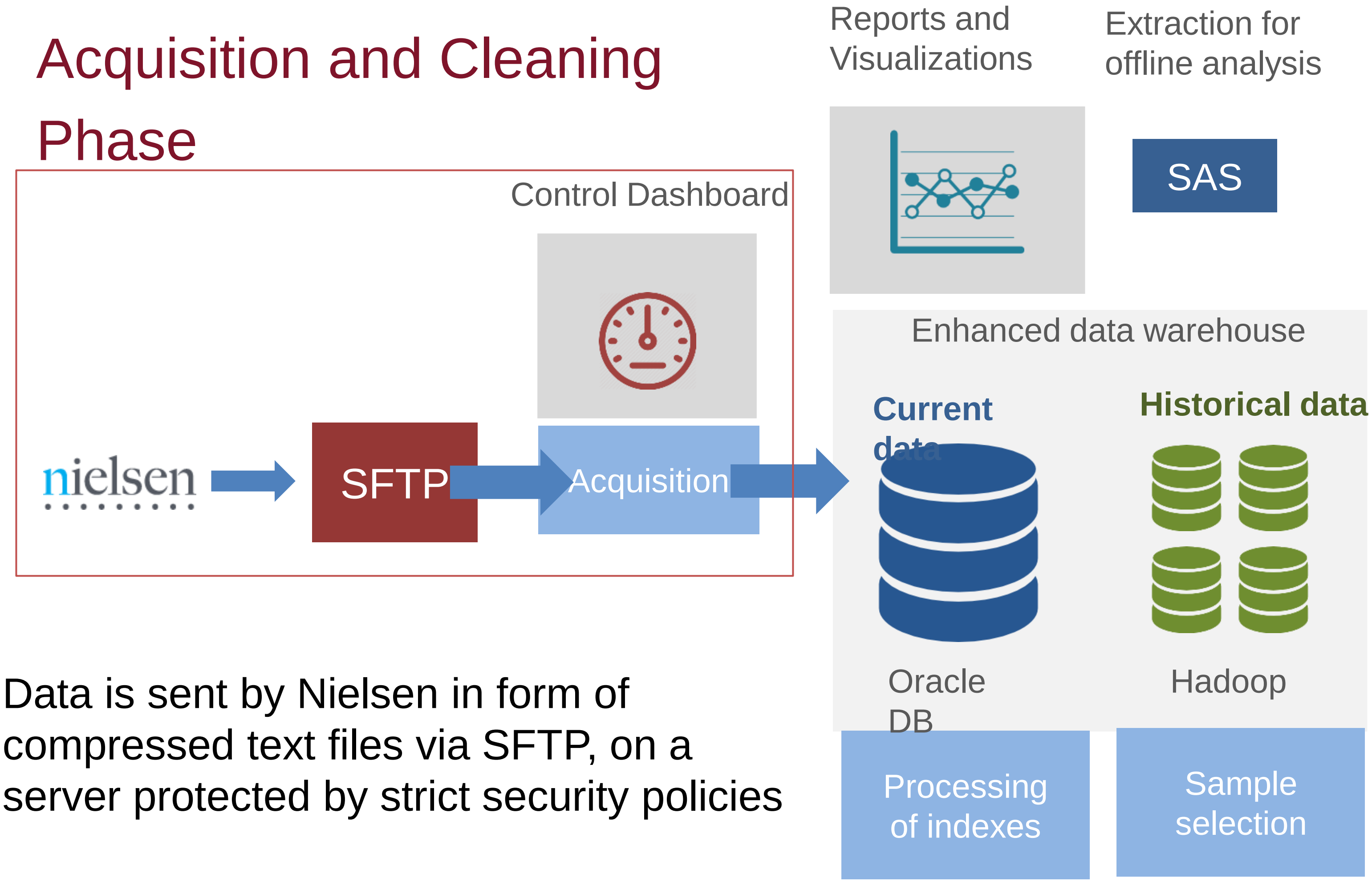
Hadoop

Processing of indexes

Sample selection

The New IT Architecture Supporting the Collection and Usage of Scanner Data

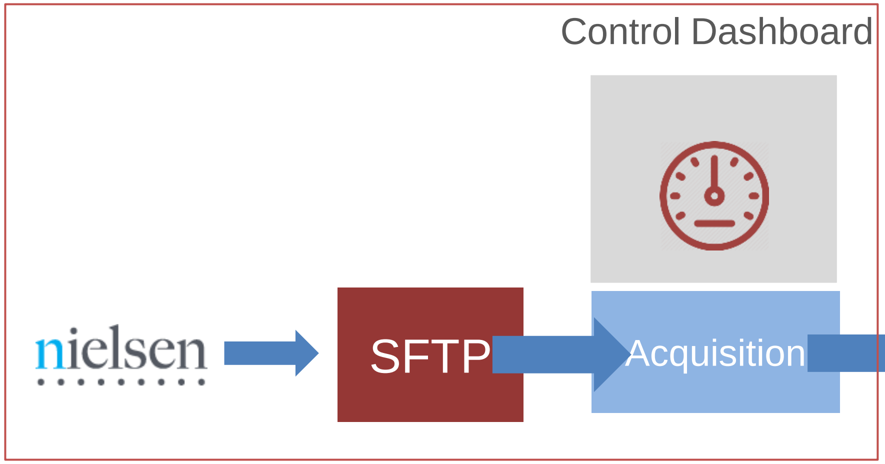
Acquisition and Cleaning Phase

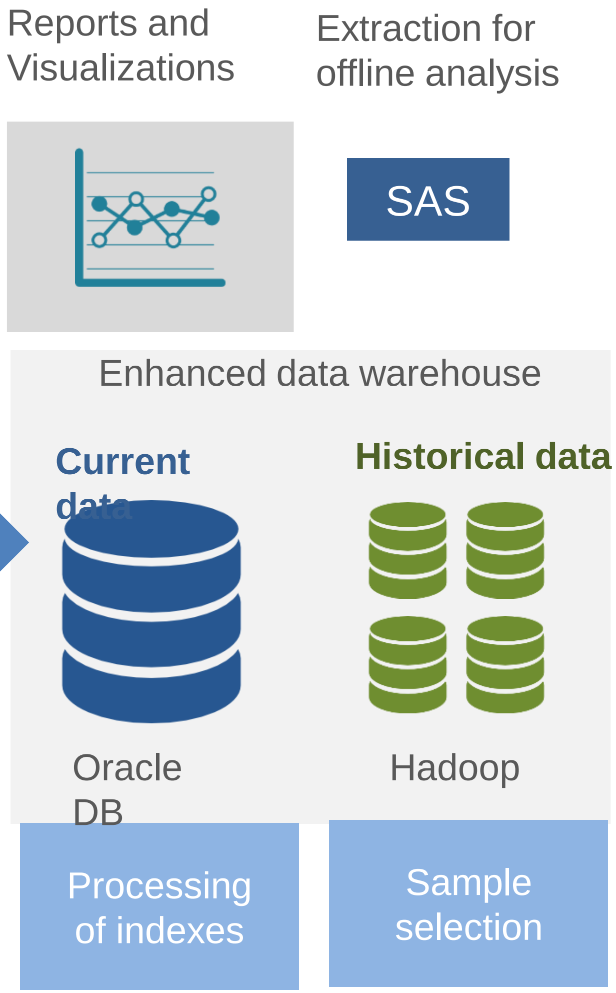


Data is sent by Nielsen in form of compressed text files via SFTP, on a server protected by strict security policies

The New IT Architecture Supporting the Collection and Usage of Scanner Data

Acquisition and Cleaning Phase





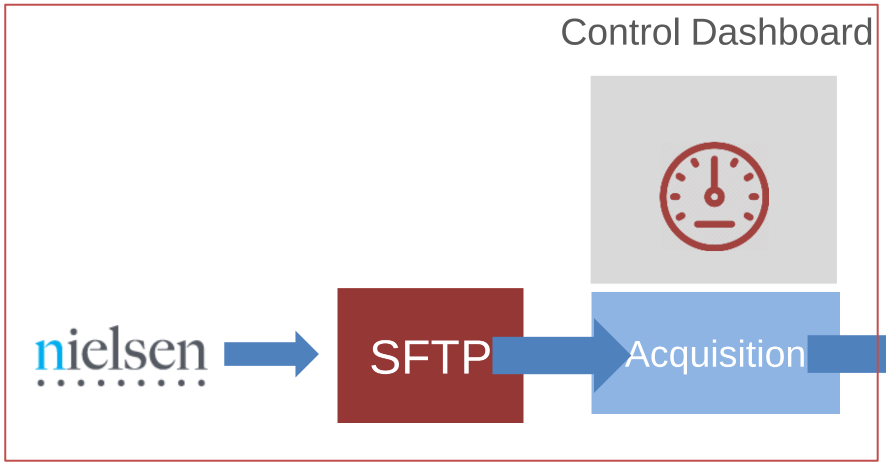
Data is pre-processed through programs written in Java

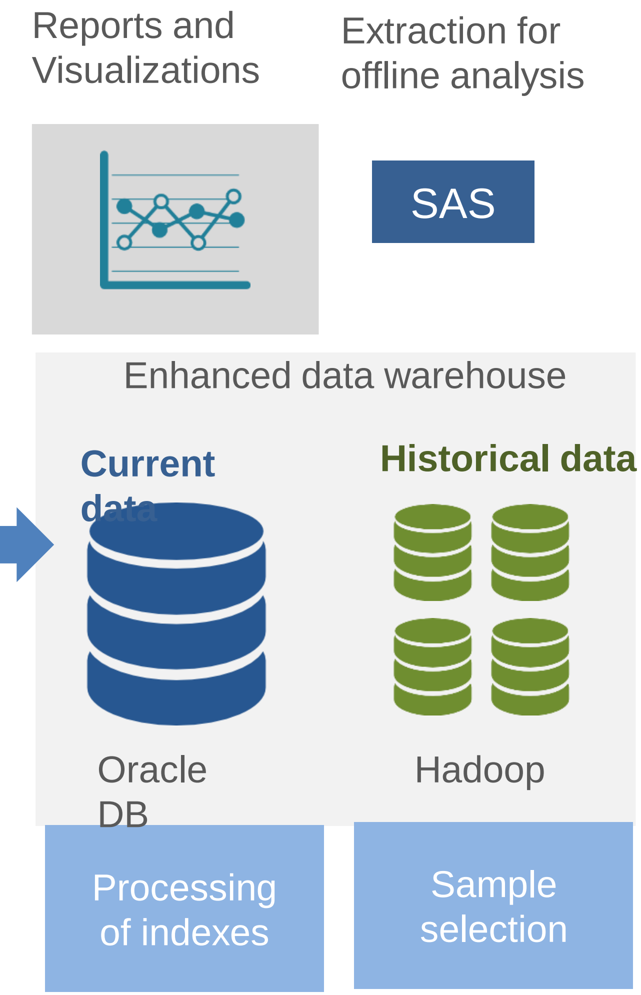
- Integrity checks for files
- Quality checks at record level, dirty data is discarded

The whole acquisition process is controlled by a web dashboard

The New IT Architecture Supporting the Collection and Usage of Scanner Data

Acquisition and Cleaning Phase





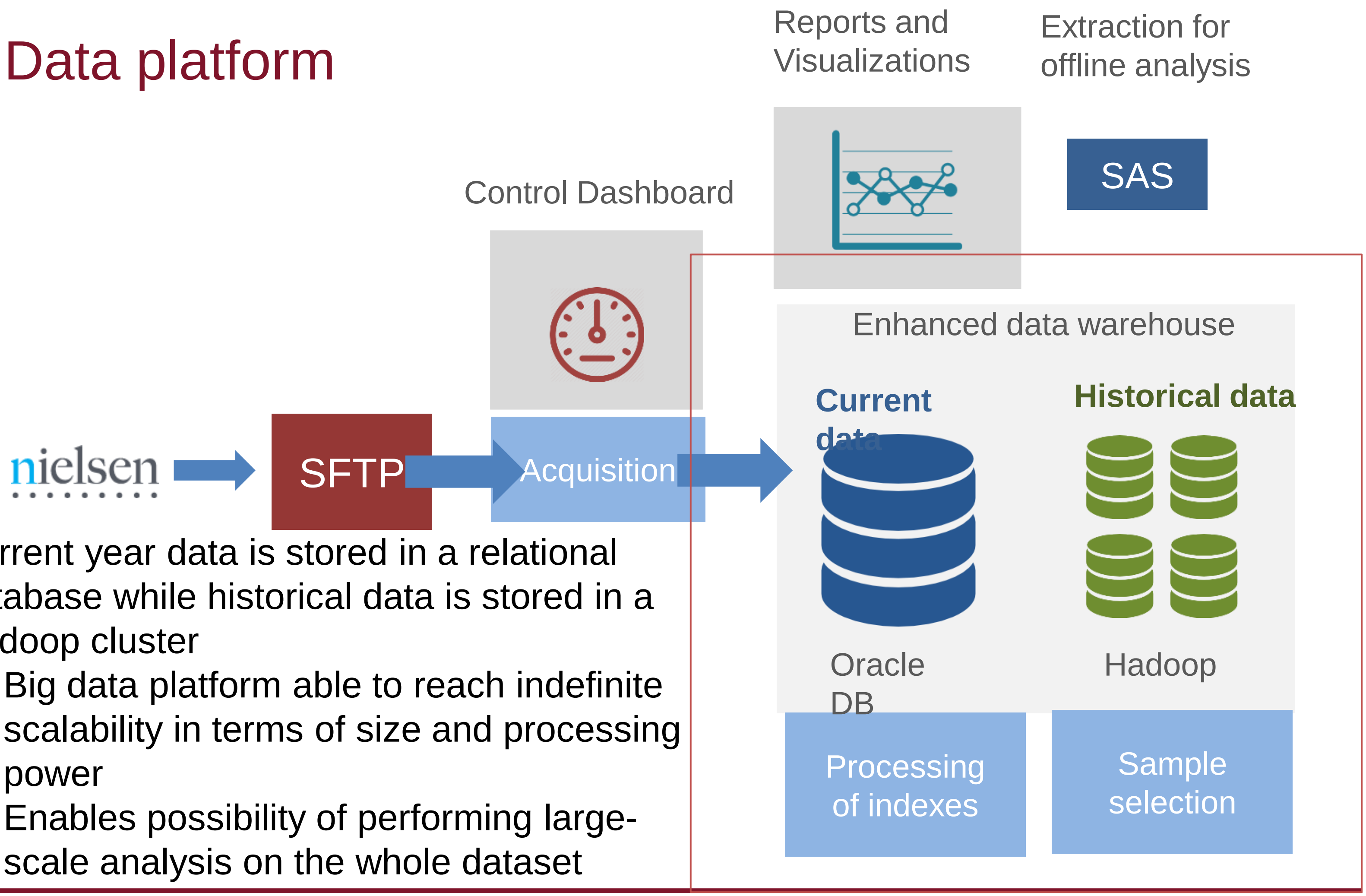
Data is pre-processed through programs written in Java

- Integrity checks for files
- Quality checks at record level, dirty data is discarded

The whole acquisition process is controlled by a web dashboard

The New IT Architecture Supporting the Collection and Usage of Scanner Data

Data platform

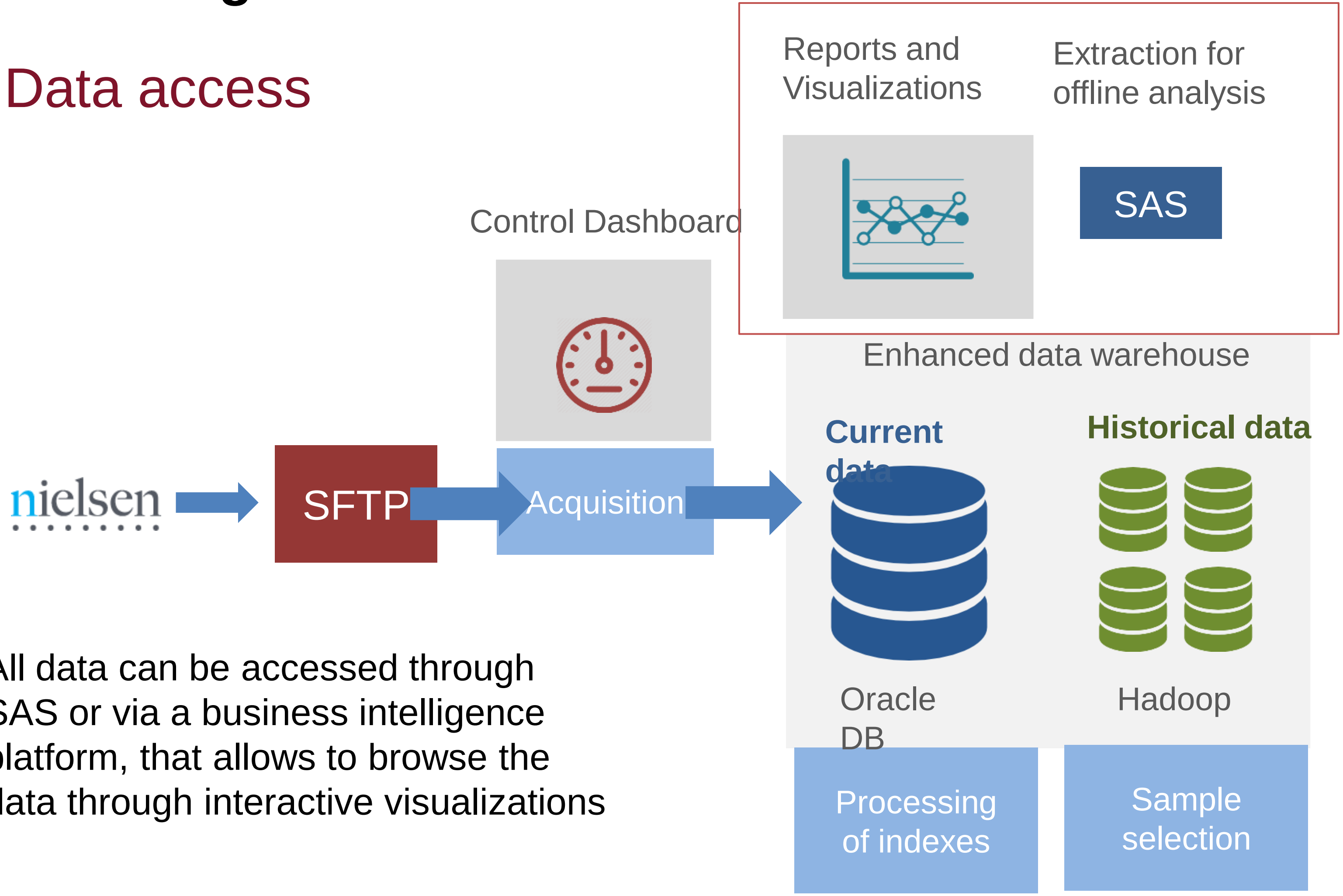


Current year data is stored in a relational database while historical data is stored in a Hadoop cluster

- Big data platform able to reach indefinite scalability in terms of size and processing power
- Enables possibility of performing large-scale analysis on the whole dataset

The New IT Architecture Supporting the Collection and Usage of Scanner Data

Data access



All data can be accessed through SAS or via a business intelligence platform, that allows to browse the data through interactive visualizations

Future Steps

Statistical and organisational features

- Enlarging the coverage of scanner data to the entire national territory
- Starting the managing of big amount of data obtained through web scraping
- Enhancing the specialisation of the data collection stage within the new Istat organization that has set up a Direction dedicated to the data collection management
- Implementing dedicated tools and procedures to carry out the treatment and analysis of scanner and web scraped data

Future steps

IT features

- Using Big Data technology to improve the efficiency of data processing
  - Test of different sampling schemes is currently made using SAS on a small portion of the data set, taking several hours for its execution
  - We will experiment an alternative implementation using Hadoop, testing the execution over the whole available data set

- Construction of a global historical data warehouse of price data

Conclusions and aims

- In 2018 starting the use of Big data (scanner and web scraped data) in the regular production process of compiling CPI
- Re-designing Italian survey on consumer prices as a mix of different approaches to sampling and data collection
- Finding the best methodological solutions to combine together these different approaches
- Improving and enhancing the entire IT and organisational architecture of the CPI allowing both specialisation (data acquisition/data treatment/CPI compilation/analysis) and integration

Thanks for the attention