European Commission

Big Data

Effective Processing and Analysis
of Very Large and Unstructured data
for Official Statistics.

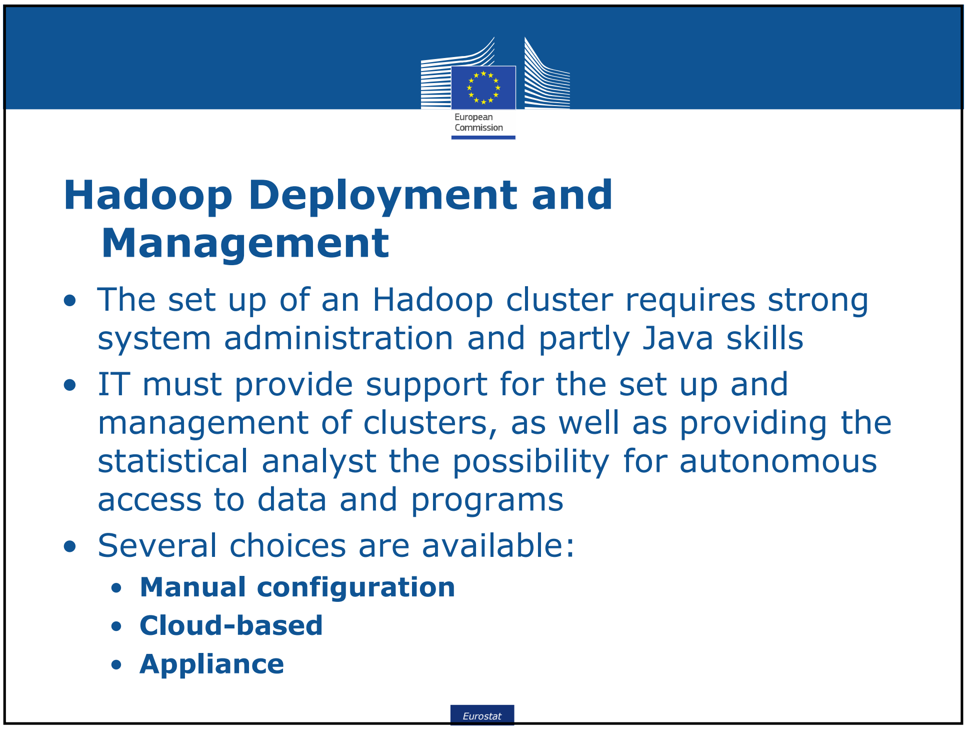
IT issues in using Big Data for Official Statistics

Antonino Virgillito
Istat (virgilli@istat.it)

European Commission

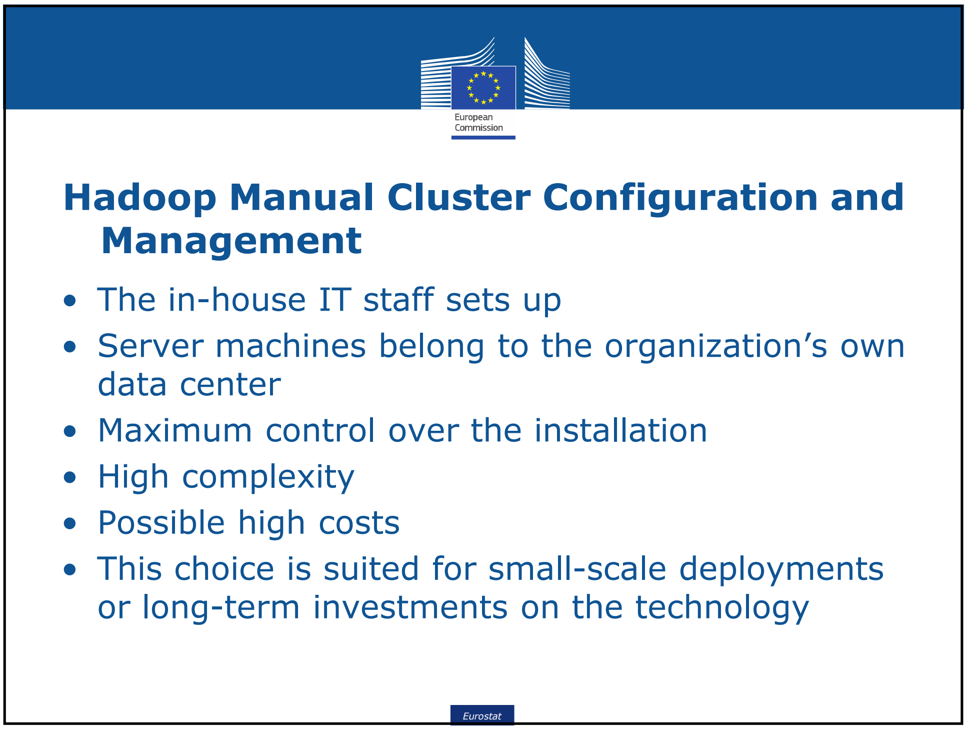
Perspectives

- The Statistical Analyst
  - I want to use my own tools and methods and don’t care about this distributed stuff
- The IT
  - I don’t want to write programs for every analysis
  - But I can set up and manage the infrastructure



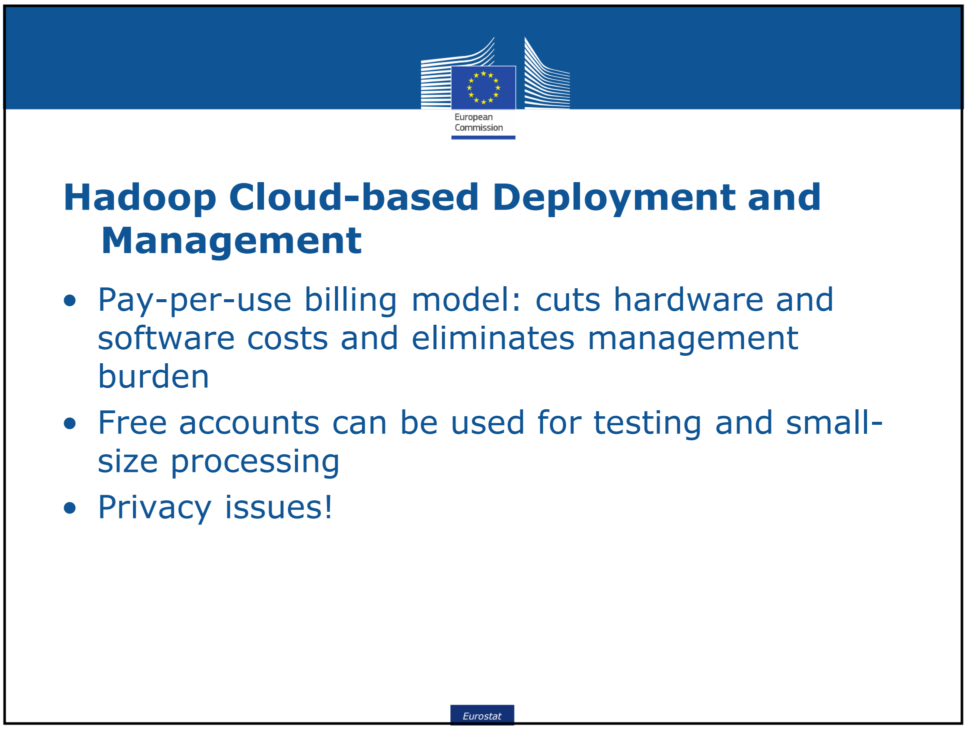
Hadoop Deployment and Management

- The set up of an Hadoop cluster requires strong system administration and partly Java skills
- IT must provide support for the set up and management of clusters, as well as providing the statistical analyst the possibility for autonomous access to data and programs
- Several choices are available:
  - **Manual configuration**
  - **Cloud-based**
  - **Appliance**



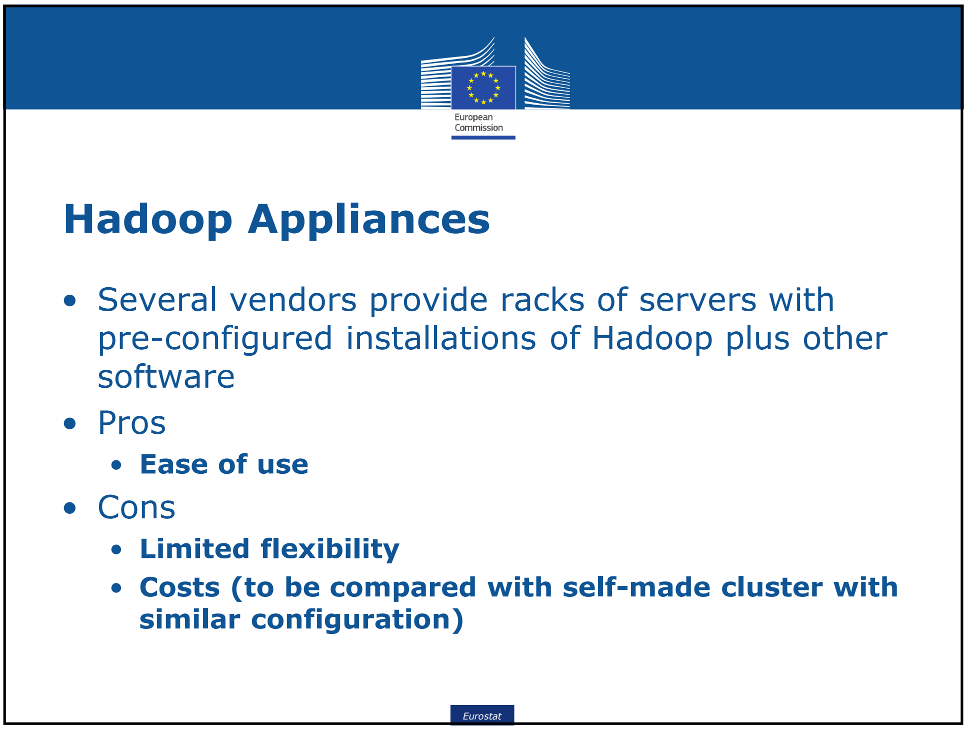
Hadoop Manual Cluster Configuration and Management

- The in-house IT staff sets up
- Server machines belong to the organization’s own data center
- Maximum control over the installation
- High complexity
- Possible high costs
- This choice is suited for small-scale deployments or long-term investments on the technology



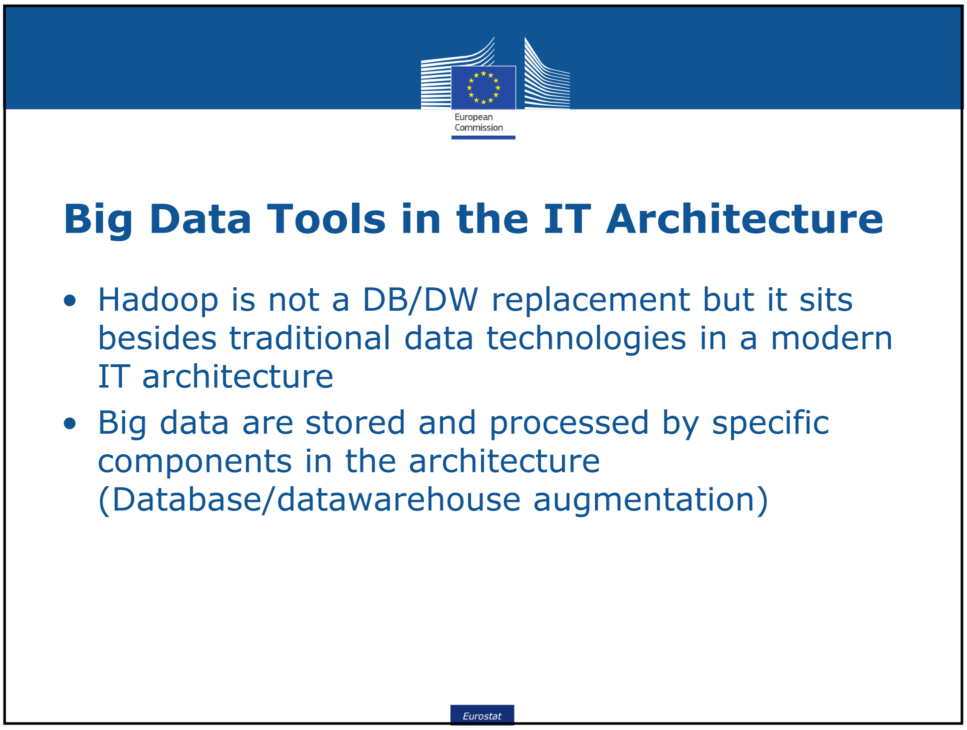
Hadoop Cloud-based Deployment and Management

- Pay-per-use billing model: cuts hardware and software costs and eliminates management burden
- Free accounts can be used for testing and small-size processing
- Privacy issues!



Hadoop Appliances

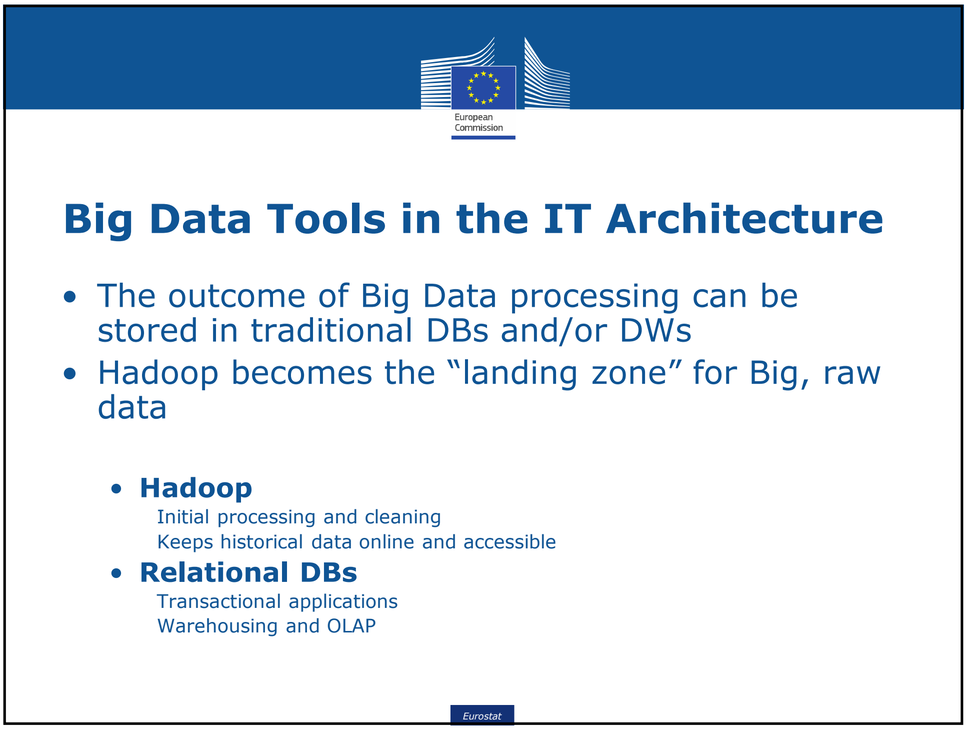
- Several vendors provide racks of servers with pre-configured installations of Hadoop plus other software
- Pros
  - **Ease of use**
- Cons
  - **Limited flexibility**
  - **Costs (to be compared with self-made cluster with similar configuration)**



Big Data Tools in the IT Architecture

- Hadoop is not a DB/DW replacement but it sits besides traditional data technologies in a modern IT architecture
- Big data are stored and processed by specific components in the architecture (Database/datawarehouse augmentation)

Eurostat



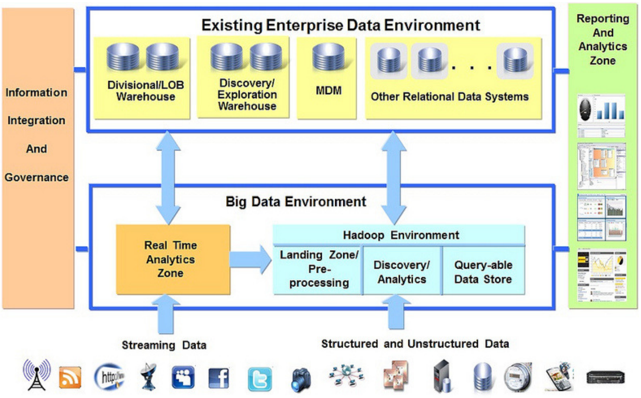
Big Data Tools in the IT Architecture

- The outcome of Big Data processing can be stored in traditional DBs and/or DWs
- Hadoop becomes the “landing zone” for Big, raw data

  - **Hadoop**
    Initial processing and cleaning
    Keeps historical data online and accessible
  - **Relational DBs**
    Transactional applications
    Warehousing and OLAP

Eurostat

Data Warehouse Augmentation



Existing Enterprise Data Environment

Divisional/LOB Warehouse

Discovery/ Exploration Warehouse

MDM

Other Relational Data Systems

Information Integration And Governance

Reporting And Analytics Zone

Big Data Environment

Real Time Analytics Zone

Hadoop Environment

Landing Zone/ Pre-processing

Discovery/ Analytics

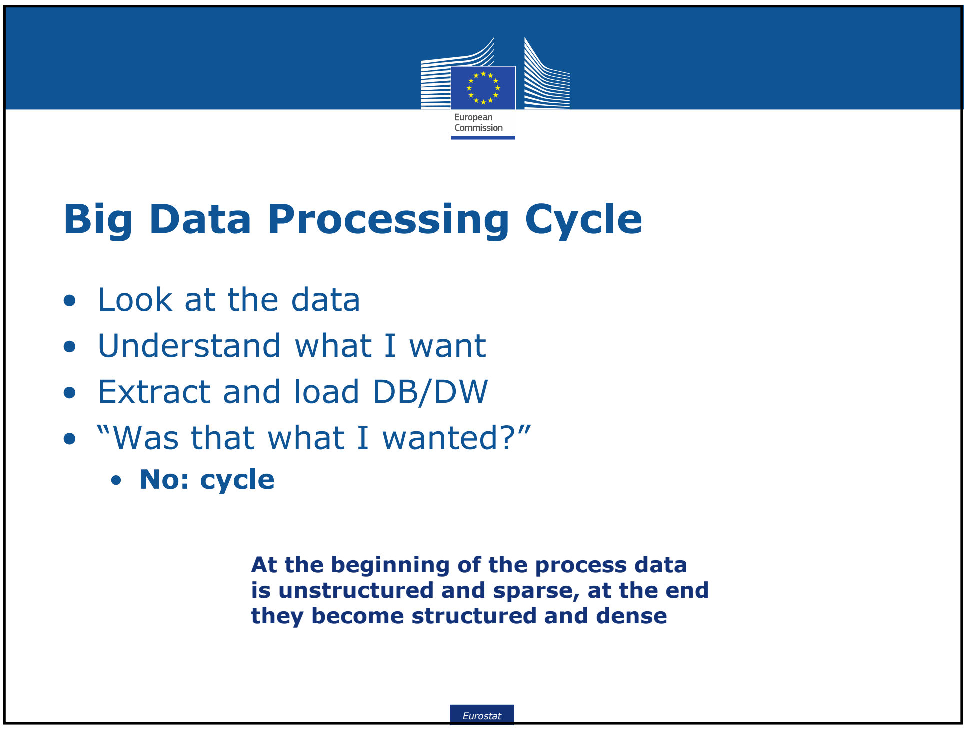
Query-able Data Store

Streaming Data

Structured and Unstructured Data

Big Data and Visual Analytics

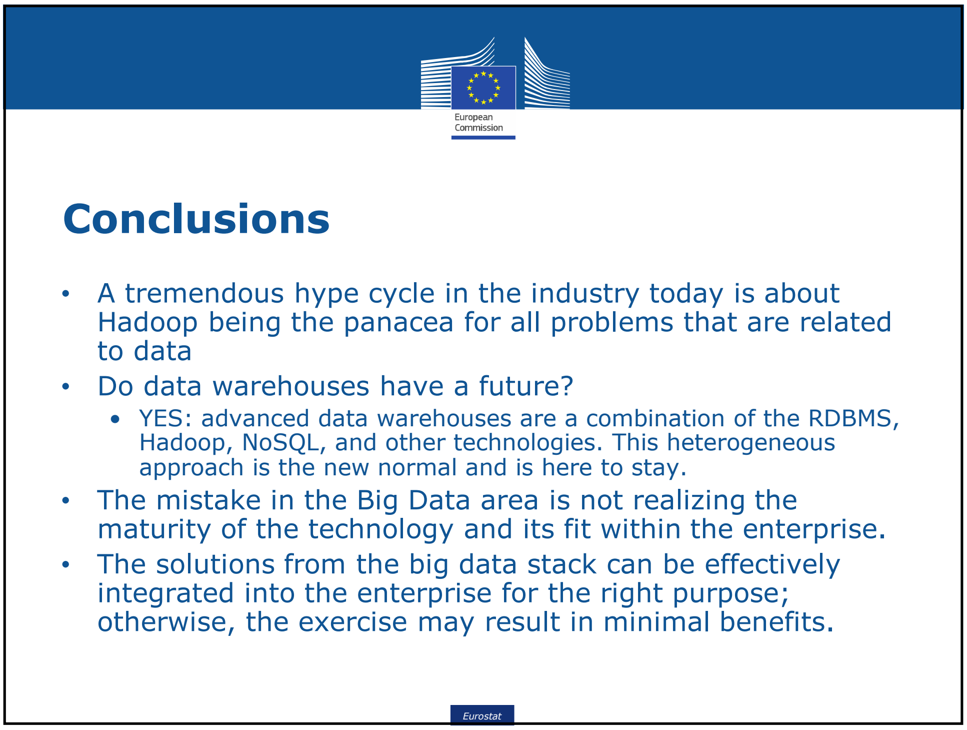
- Modern (visual) analytics tools can integrate both kinds of data sources
- Connect to heterogeneous data sources, including HDFS
- Create joins between them
- Create visualizations with integrated sources



Big Data Processing Cycle

- Look at the data
- Understand what I want
- Extract and load DB/DW
- “Was that what I wanted?”
  - **No: cycle**

At the beginning of the process data is unstructured and sparse, at the end they become structured and dense



Conclusions

- A tremendous hype cycle in the industry today is about Hadoop being the panacea for all problems that are related to data
- Do data warehouses have a future?
  - YES: advanced data warehouses are a combination of the RDBMS, Hadoop, NoSQL, and other technologies. This heterogeneous approach is the new normal and is here to stay.
- The mistake in the Big Data area is not realizing the maturity of the technology and its fit within the enterprise.
- The solutions from the big data stack can be effectively integrated into the enterprise for the right purpose; otherwise, the exercise may result in minimal benefits.