IT infrastructure for Big Data and Data Science: Challenges at Statistics Netherlands

**Keywords:** IT infrastructure, Big Data, Data Science, Tools

# Introduction

Until a couple of years ago, processing data at national statistical institutes (NSI’s) did not differ that much as processing data at an administrative environment. In case of surveys, one needed a database comprising all companies or persons in the country, take a sample of this population, and, when the questionnaire was sent back, process the (relatively) small amounts of data to calculate the wanted estimates. In case of register data, the amount of data was higher, but still manageable for relational database management systems (RDBMS) or even in the form of comma separated values. Most of the data sets would fit in memory of small dimensioned desktops. The role of statistical methodology was simple: Given one or several datasets, define an algorithm that executes a certain method for estimating the target variables. Processes as data cleaning could be rule based or could encompass an iterative process to find the “right” values. Most of the times, these algorithms did not need to be efficient, since inefficient algorithms would maybe run for a couple of hours once every three months or once a year.

In the age of big data and data science, this is not the case anymore. Nowadays, many questions can be asked and answered by using more timely data and it is expected from modern statisticians to be data scientists, who use modern technology to be able to quickly answer current questions. Transitions of the used systems is not an easy task. The administrative landscape is based on low data throughput and relatively small latencies can be relatively easy realized. For data science infrastructures, this is another story. Small latencies are not easy to realize. In certain cases, a high data throughput needs to be realized and this cannot be realized with more traditional infrastructures. Whereas in traditional infrastructures, vertical scaling, or scaling up, was normally done when more compute power was needed, horizontal scaling, or scaling out, is used in data intensive infrastructures. More servers are coupled to create clusters, which scale easily with the data [1]. In this paper, we will discuss what the transition from “classical” statistics to modern data science means and what challenges lie ahead when implementing.

# Data science and the use of big data

When doing data science, the major activities are data exploration and data modelling. These activities are discussed in the following subsections and the technology are mentioned which can be used when carrying out these activities.

## Data exploration

The working field of data scientists deviates from that of statisticians. Whereas datasets used by statisticians are most of the time predefined, data scientists search actively for new data. Most of the time, this involves open data (like satellite data, road sensor data, open street map, public data on social media), but also sensitive data (like mobile phone data, AIS data, data involving companies and persons) can be used by data scientists. If this sensitive data is not available within the NSI, one can import the data into a secure environment or use secure data sharing to realize it.

Data exploration one of the most common activities of data scientists and one can say that 80% of the time of a data scientist is spend during data exploration. For performing data exploration, the data scientist uses advanced visualisation techniques, but also techniques like dimension reduction, clustering, filtering and regression analysis are common to data exploration. In some cases, when the datasets are small, these analysis can be done in R or Python, but sometimes, when the datasets are too large, big data tools like Spark are needed. However, there is not a one size fits all solution. In some cases, solutions like column stores, key value stores or graph databases are needed to process the data more efficiently.

Most of the time, these methods are already implemented by peer data scientists and are available in external repositories like github, CRAN (for R), PyPi (for Python) or Maven (for Java or Scala). In these cases, the data scientist needs to be able to install this software quickly. Note that this is very different from the classical work of a statistician. Since the data was predefined the tools to be used could also be predefined.

Most software runs very well on an open source platform like Linux, and henceforth Linux is the dominant operating system of choice for data scientists. Another reason for using Linux is that the standard tools in Linux are well suited for dealing with data science problems. Whereas it is really easy to access the first lines of a really big file in Linux, this is not that easy in Microsoft Windows.

In certain cases the data scientist realizes that other data sets are also needed for their study. For that reason, easy access to new data sets is necessary. This means at one side that access to open data should be easy (i.e. connection to the internet via API’s, but also by using webscraping technology) and that data, available at the NSI’s, should be easy to access. It should be noted that this has not only implications for the used infrastructure, but also for data governance. When, for instance, data is collected for a certain statistics and the data source is part of a stove pipe for creating this statistics, every secondary use of that data source is negotiated with the data owner in that stove pipe. This means that data is not readily available for a data scientist. This asks for a data lake environment and a centralized data governance model.

## Data modelling

Another task in doing data science is data modelling. Here, the data scientist will try to create a model which can answer, in case of NSI’s, policy questions. These kinds of models involve processing large amounts of data and in many cases big data solutions are needed, including solutions for data storage. Furthermore, models need to be applied in an efficient fashion. These models often involve iterative algorithms and the infrastructure needs to be ready for this. For these kinds of problems, Apache Spark and multicore programming or GPU programming are often used. Also here, it is hard to find a one size fits all solutions and data scientist need to resolve to new technology developed, hence a versatile infrastructure which can run many different environments is needed.

## Requirements

In summary, Data scientists require:

* GPU’s for executing certain kinds of models (like deep neural networks)
* Graphical workstations, preferably with GPU’s (also for performance computing)
* Linux as an operating system
* Easy deployment of different big data solutions like Spark.
* Easy deployment of other data solutions like column stores, key value stores and graph databases
* Availability of repositories like CRAN, Pypi and Maven
* Data lake environment
* Secure data sharing
* Easy data access for internal as well as external (open) data.

# requirements for an infrastructure for big data and data science.

From the previous section, it is clear that a data scientist needs a versatile environment for processing various amounts of data. This can involve larger desktops (like workstations) to a large server infrastructure, depending on the size and type of data. Roughly four important issues need to be resolved in an administrative infrastructure to be suited for the use of Data Science:

1. Access to the internet
2. Realize performance computing using horizontal scaling techniques (like Spark) and using GPGPU’s
3. Easy to deploy new infrastructures
4. Easy access to other data sources

## Direct access to the internet

Directly accessing the internet is hard to accomplish. Opening the internet for everyone means that the data available at the NSI is also easier to access, and one should take extra measures need to be taken to prevent data leakage.

Most institutes having highly sensitive data solved the security issue by blocking any kind of internet access and it is hard to find a solution that can work. This needs time to investigate. Opening the internet will mean that data scientists will be able to download open data sets and can access external available repositories for installing new software or modules.

As long as no internet access can be realized within such institutes, one can:

1. Import repositories and synchronize them on a daily basis
2. Define a fast lane which makes it easy to import data needed for research

This is what we are implementing at statistics Netherlands for the time being.

## Performance computing

Whereas vertical scaling asks servers to be big, horizontal scaling can be realized by coupling smaller servers together. This, however, puts extra stress on the network connectivity and creates another demand on the available hardware within the institute. Hence, it should be easy to place extra servers in these environments.

The use of GPGPU’s is a problem that we solved by placing a workstation with a sophisticated GPU at the office. However, solving it with local workstations, like we did for now, makes it hard to share the infrastructure between different researchers or research groups.

The use of GPU’s is twofold: first, it gives the researcher the ability to train much more complex models than he was able to do before. Secondly, however, the researcher can also use this technology for visualization purposes, boosting his ability to create much more complex visualisations.

## Deploying new infrastructures

Deploying new infrastructures involves using virtualization techniques like container technology and virtual environments. These kinds of environments help us to create a versatile environment. However when using virtualization techniques, enough hardware should be available to host all the initiatives.

For the time being, we are using a private cloud solution for this scalability issue.

## Data governance

Data governance needs to be changed for data scientists to access data available at different groups within NSI’s. A governance model which is built around the idea that one data source is collected for serving one statistics will throw unnecessary hurdles for data scientists to perform their job.

Initiatives like a data lake, secure data sharing and direct access to open data are necessary to implement the data policy needed for data science. At statistics Netherlands, a data lake is implemented and, together with universities, we are working at secure data sharing.

# Conclusion

As one can see, Statistics Netherlands is on its way to realize an infrastructure that is versatile, has direct access to open data, can use secure data sharing protocols, uses Performance computing components and is able to use state of the art libraries and tools which are available from the outside.

Since the infrastructure was designed in a time that security meant complete decoupling from the internet and statistics were made mainly on the basis of registers and surveys, the necessity was not high to have other tooling available. The transition from this old environment to a new, more versatile environment takes time.

# References

# [1] D. Singh, C. Reddy. A survey on Platforms for Big Data Analytics, Journal of Big Data (2015) 2(1): 8.

# <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4505391/>