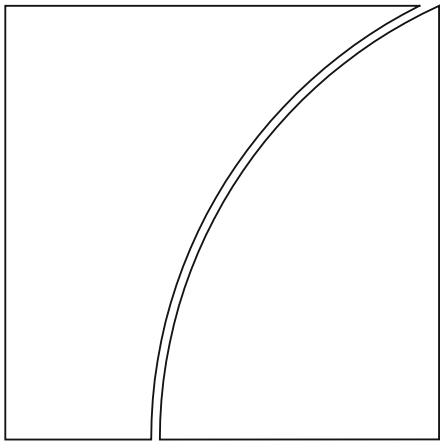
Irving Fisher Committee on Central Bank Statistics



IFC Report

No 11

Computing platforms for big data analytics and artificial intelligence

April 2020

BANK FOR INTERNATIONAL SETTLEMENTS

Contributors to the IFC Report[1]

Bank of Italy	Giuseppe Bruno
BIS	Hiren Jani
	Rafael Schmidt
IFC Secretariat	Bruno Tissot

This report has been prepared on the basis of the workshop on "Computing platforms for big data and machine learning" organised in Rome on 15 January 2019 by the Bank of Italy with the support of the Irving Fisher Committee on Central Bank Statistics (IFC) of the Bank for International Settlements (BIS); see www.bis.org/ifc/events/boibis_jan19/programme.pdf.

The views expressed are those of the authors and do not necessarily reflect those of the IFC, its members, the Bank of Italy, the BIS or the institutions represented at the meeting.

This publication is available on the BIS website (www.bis.org/).

ISSN 1991-7511 (online)

ISBN 978-92-9259-362-9 (online)

[1] Respectively, Head of Support for economics and statistics at the Bank of Italy (Giuseppe.Bruno@bancaditalia.it); Head of Platform Engineering in Information Technology and Services at the Bank for International Settlements (BIS) (Hiren.Jani@bis.org); Head of IT, BIS Monetary and Economic Department (Rafael.Schmidt@bis.org); and Head of the Secretariat of the Irving Fisher Committee on Central Bank Statistics (IFC) and Head of Statistics & Research Support, BIS (Bruno.Tissot@bis.org).

We thank Giulio Cornelli, Edward Lambe and Xavier Sosnovsky for their helpful comments and suggestions, and Giulio Cornelli again for excellent research assistance.

Contents

Computing platforms for big data analytics and artificial intelligence – central banks' experience[2]

Executive Summary

Public authorities, and central banks in particular, are increasingly realising the **potential of big data sets and analytics** – with the development of artificial intelligence (AI) and machine learning (ML) techniques – to provide new, complementary statistical information (Hammer et al (2017)). Yet the question remains: how should institutions organise themselves to benefit the most from these opportunities? Two areas appear particularly important for central banks. The first is how to organise their statistical information in relation to their IT infrastructure. The second is to think strategically as to how to use appropriate techniques to further process and analyse the new information collected.

Like many national statistical offices (NSOs) and international organisations, central banks have already **launched numerous initiatives** to explore these issues and exchange on their experience, in particular through the cooperative activities organised by the Irving Fisher Committee on Central Bank Statistics (IFC) of the Bank for International Settlements (BIS). The BIS itself has developed its new medium-term strategy, Innovation BIS 2025, which relies on important investment in next-generation technology to build a resilient and future-ready digital workplace for the organisation (BIS (2019)).

A key feature of these various initiatives is that many central banks are currently setting up, or envisaging implementing, **big data platforms** to facilitate the storage and processing of very large data sets. They are also developing high-performance computing (HPC) infrastructure that enables faster processing, in-depth statistical analysis and complex data simulations. However, these initiatives face important organisational challenges, as central banks trade off factors such as technology trends, system complexity, cost, performance, reliability, operating model and security.

In view of this experience, it is essential to **carefully assess the options available** before selecting a technology and architecture to set up big data / HPC platforms. Among the various issues to be considered, attention should primarily focus on the hardware selection, the choice between proprietary and open source technology, the decision to develop the solution in-house or in the cloud, and the type of information to be handled.

Once the main options are selected, the actual implementation of the related technologies is often a **long and multiform journey**. From a project development perspective, success will depend on the approach for conducting the project, the types of workload to be supported, the data architecture envisaged and the use of software development best practices. Having a **broader, institution-level**

[2] This report draws on the proceedings of the workshop on "Computing platforms for big data and machine learning" organised by the Bank of Italy with the support of the BIS and the IFC in Rome on 15 January 2019.

perspective is also key, so as to adequately take into account the full range of business requirements as well as resources and security constraints. It may therefore be recommended to develop a comprehensive information strategy for the institution, with a high-level roadmap for the adoption of continuously changing technologies to manage data and respond to users' needs. Last but not least, knowledge-sharing can be instrumental, and can be facilitated by the cooperative activities promoted by the BIS and the IFC.

1. Big data and high-performance computing platforms

1.1 Background

Big data and ML and AI use cases are becoming omnipresent in business and academia as well as in public institutions (IFC (2019a)), and addressing them effectively calls for continuous technological innovation (Signorini (2019)). Efforts have particularly focused on the implementation of so-called big data and/or high-performance computing (HPC) platforms. Big data platforms enable the storage and processing of very large data sets, and are increasingly requested to deal with information of a semi-structured or unstructured nature.[3] HPC platforms enable fast data processing and complex statistical analysis/simulations, and are typically well suited for dealing with medium-size structured data sets. Certainly, this distinction is a bit artificial, and the boundary between big data and HPC platforms is becoming blurry with the rapid development of IT technologies that are able to serve both types of user needs.

Against this backdrop, the Bank of Italy, with the support of the BIS and the IFC, organised a workshop on "Computing platforms for big data and machine learning". Convening in Rome in January 2019, participants from about 30 organisations including central banks, NSOs and international organisations took this opportunity to present on the status of their current big data initiatives and exchange experiences.

This IFC report draws on the examples reviewed on that occasion and presents the **main big data / HPC technologies and platforms currently considered or already implemented** by public authorities such as central banks. It also gives insights on how to deal with the related challenges and get the most of the opportunities provided by these new infrastructures to support policymaking. The rest of this section reviews the key factors driving the demand for big data and the related technology, by drawing on selected use cases. Section 2 derives some lessons from the projects already set up and discusses the main opportunities and challenges faced. Section 3 reviews the main technical options to consider when choosing the related technology. Section 4 provides generic guidance for the actual implementation of big data infrastructure projects and the building-up of the

[3] "Semi-structured or unstructured" means that such data sources have no predefined data model and often do not fit into a conventional relational database. It should be noted that big data platforms are also in demand for those very large data sets that are still well structured, in particular in order to handle the constraints faced in terms of storage/processing. This is particularly the case for the type of "financial big data sets" central banks have been increasingly dealing with since the Great Financial Crisis of 2007–09 (Tissot (2019)).

underlying computing platforms. Section 5 draws some general conclusions looking forward.

1.2 Increasing interest in big data information…

The **term big data** is relatively vague and refers in general to the huge proliferation of information brought by the so-called data revolution and generated by social media, web-based activities, machine sensors, financial, administrative or business operations, etc. One well known characterisation of this concept relates to the "Vs" of big data, such as volume (ie number of records and attributes), velocity (speed of data production eg tick data) and variety (eg structure and format of the data set) (Laney (2001)).

The **financial big data sets** central banks and international financial institutions more specifically are facing typically comprise four main sorts of sources: internet-based indicators, commercial records, financial market indicators and administrative registers. Their analysis can provide many opportunities, such as: the filling of existing data gaps; the collection of new data that can flexibly complement "traditional" survey- and census-based official statistics; the availability of almost real-time data, opening up the possibility of getting more timely economic signals and bridging time lags in statistical publication; and the possibility to apply brand new methodologies to extract original knowledge (Bholat (2015)). These opportunities can greatly facilitate the conduct of central bank policies, for instance by enhancing economic forecasts, assisting micro supervision tasks, conducting new types of financial stability assessments and obtaining more rapid feedback on policy decisions (IFC (2017)).

In addition to capturing structured transactional data that sit at the core of central banks' work, **various and innovative big data sets** – such as granular data collected from the web (eg page views, online ads, search indicators), sentiment indicators derived from social media (eg Facebook, LinkedIn, Twitter), payment transactions and traditional textual information processed with new big data tools (eg policy announcements, speeches, emails, press articles) – can be mobilised in practice to facilitate central banks' day-to-day work in various areas, including the production of statistical information, economic analysis and research supporting monetary and financial stability policies, and supervisory tasks (Tissot (2019)).[4]

As regards **statistical work**, big data can enhance the quality of existing, "more conventional" statistics, or even complement them with new types of indicators. One good example, particularly important in view of central banks' price stability mandates, relates to the measurement and analysis of inflation patterns. The use of new big data-type collection methods has expanded significantly, allowing more detailed price information, greater product variability and a larger number of

[4] Compared with counterpart institutions like NSOs, central banks' interest has so far been rather limited for the wider range of unstructured data comprising geospatial data (eg mobile phones, satellite images), smart grid data (eg information derived on the users connected to electricity networks) and more generally all types of sensor data that result from the recording of some type of input from the environment (eg temperature); see Meeting of the Expert Group on International Statistical Classifications (2015) for a general review of the wide range of what can constitute a big data source. However, central banks' interest may well expand for these types of data too as new technologies/ideas emerge.

transactions to be captured (Eiglsperger (2019)). In particular, the direct web scraping of online retailers' price data can complement the coverage of "traditional" survey-based collection of prices in physical stores, enhance the measuring of specific volatile components of inflation, provide real-time and/or high-frequency estimates of the consumer price index (CPI) and fill existing data collection gaps in some regions (see the Billion Prices Project developed by the MIT Sloan School of Management; Cavallo and Rigobon (2016)). Public authorities are also trying to make use of the rapidly expanding volumes of scanner data sets,[5] which provide important potential, for instance to track expenditure patterns in a granular way or to increase the accuracy of CPI figures (Chessa et al (2017)).[6]

1.3 … and big data analytics

Faced with vast and rapidly expanding big data, public authorities need to extract relevant information and make sense of it so as to take informed, evidence-based policy decisions – putting a premium on transforming data into information and then information into knowledge (Drozdova (2017)). The way to do so is to use so-called **big data analytics**, by which the various data sets are analysed to uncover information such as specific patterns, correlations, trends etc. Many statistical techniques, including those supporting ML and AI (as well as more traditional ones too) can be used on these data sets to support pattern recognition, knowledge discovery, data mining etc (IFC (2019a)).

As regards **central banks' research and analytical work supporting policy**, big data analytics are being applied for a wide range of purposes, especially in the area of economic forecasting (eg for indicators such as inflation, housing prices, unemployment, GDP and industrial production, retail sales, external sector developments) and business cycle analysis (eg compilation of sentiment indicators and use of nowcasting techniques to "forecast" the present), which are key ingredients supporting the conduct of monetary policy. Big data sources and analytics are also increasingly useful for their financial stability work, for instance to construct risk indicators, assess the behaviour of market participants, identify credit and market risk, and monitor financial transactions and capital flows.

Lastly, big data information and tools are also playing an increasing role in supporting financial risk assessment and surveillance exercises that sit at the core of the mandates of **supervisory authorities** – including those central banks that are directly tasked with supervising financial institutions such as commercial banks as well as the functioning of the payment systems. In particular, IT innovation has opened promising avenues for using the vast amounts of information entailed in granular data sets – with the increased role played by "suptech" (Broeders and Prenio (2018)). This can help to, for instance, detect market abuse through the use of text mining

[5] Data on sales of consumer goods obtained by scanning the bar codes of individual products in retail outlets. The data can provide detailed information about quantities, characteristics and values of goods sold as well as their prices.

[6] The 2020 Covid-19 pandemic provided another example of the usefulness of these "non-standard" sources. A number of NSOs were able to compensate for the inability to measure price data in closed stores by collecting prices on the internet and using scanner data for certain consumption segments to compute CPIs (for example, see in the case of France the INSEE communication at www.insee.fr/en/information/4471928).

techniques flagging misconduct or insider trading; spot odd patterns in the data signalling the build-up of possible idiosyncratic vulnerabilities; and identify network effects supporting the assessment of system-wide risks (Wibisono et al (2019)).

Reflecting these various opportunities, a key priority for central banks is to **invest in big data infrastructure**. Faced with an expected significant increase in use cases and related computing/storage demands, they have already been conducting extensive preparatory work in terms of IT infrastructure, skill sets, resources, processes etc. They are also mobilising important financial and staff resources. This trend appears very likely to continue in the foreseeable future, as the multiple insights that can be gained to support policy will become increasingly apparent, and because IT technology is rapidly evolving.

1.4 High-performance computing (HPC)

Public authorities are also focusing on **developing HPC platforms** to handle exponentially growing data sets and, in particular, run multifaceted algorithms and analytical tools. While big data analytics use cases mainly relate to dealing with large and/or complex (eg unstructured) data sets, the impetus for developing HPC comes from the need to have more computing power compared with a typical desktop computer, workstation or standard server environment. The aim is therefore to have enough resources to solve, simulate or analyse complex statistical problems, and not exclusively to handle a large amount of data. In practice, however, these two types of needs are often interrelated (see below).

Central banks' experience shows that HPC platforms are primarily developed to **ensure that computing resources are used in the most efficient way**, so that analytical processes can be completed as rapidly as possible. The solution is to aggregate IT capacity, with HPC platforms being made as clusters of individual computers, referred to as nodes – a common cluster size being between 16 and 64 nodes, representing from 64 to 256 processing units ("cores");[7] see eg Ho and Uddin (2019) for the specifications of the computing platform set up at the Bank of Canada. Given the high cost of buying and operating such platforms, the solutions implemented are typically shared among many users at the departmental or institutional level. To do so, subsets of the computing nodes are commonly allocated to dedicated "jobs" to address specific user requests, depending on the IT resources available. Hence, the key role of an HPC platform is to define the priorities of the different jobs to be executed by the computing nodes, while ensuring that the resources are not overloaded. Typically, one has to implement batch processing (see below) and schedulers – for example the YARN (Yet Another Resource Negotiator) scheduler for using the Hadoop library developed by the open source Apache software foundation – when submitting jobs, inquiring about their status and modifying them without end user interaction.

One recent feature among central banks already running some type of HPC platform is to extend this infrastructure to integrate other forms of computing, in

[7] A processor core (or "core") is a single processing unit. Today's computers – or CPUs (central processing units) – have multiple processing units, with each of these cores able to focus on a different task. Depending on the analytical or statistical problem at hand, clusters of GPUs (graphics processing units, which have a highly parallel structure and were initially designed for efficient image processing) might also be embedded in computers, for instance to support mass calculations.

particular big data analytical tools, so as to distribute the processing of large data sets across clusters of computers. They build a so-called **big data cluster**, made of connected computers and designed specifically for storing and analysing huge amounts of structured/unstructured data in a distributed way. With this trend expected to continue in the future, the next-generation HPC platforms will be increasingly able to deal with both large-scale and high-performance computing. For instance, the Apache Spark project provides a unified analytical engine for large-scale data processing and allows various types of big data file systems – such as the Hadoop Distributed File System (HDFS) to be flexibly dealt with. It also supports various programming interfaces such as Java, Python, R or Scala, allowing for computing complex calculations – see Condello (2019) for the characteristics of the Spark platform developed at the Bank of Italy. Another expected development is the push for implementing those IT solutions in a standalone mode, ie without requiring heavy additional infrastructure (eg no need to have a specific scheduler as well). Such a standalone feature appears to greatly facilitate the processing of HPC jobs.

Yet another avenue, which some institutions, such as the Bank of Canada, have already experimented with, is to explore **quantum computing** for next-generation data analytic environment (Collignon (2019)). A quantum computer is able to generate and manipulate quantum bits (or qubits, which can be 0, 1, or 0 and 1 at the same time); its computing is intrinsically parallel, and its power can thus grow exponentially with the number of qubits. For tasks that would currently require hundreds of years to be solved by standard computers, it is expected that a quantum computer could reduce the time frame very significantly, possibly to a few minutes.

2. Lessons already learned

2.1 Implementing the new platforms

General feedback from the projects already implemented by central banks is that **new big data and HPC platforms can serve multiple purposes**. One is to confirm existing, preliminary analyses based on more traditional econometric and statistical methods and/or on smaller data sets. A second is to provide complementary insights, compared in particular with conventional data sources such as surveys and censuses. And a third contribution is to identify unforeseen issues, in turn raising the need for additional information and/or research. To address these various use cases, it is important that the platforms being set up provide a common data repository, so that users can find a host of different information that can be combined and analysed according to their specific needs.

Another key lesson is that the approach followed is intrinsically a **learning-by-doing** one, with projects constantly adapted as experience progresses. The preferred method has been to develop new projects in a gradual way, by starting with individual research initiatives and implementing preliminary use cases (Marcucci (2019)); this progressive approach has proved particularly helpful given the various obstacles faced when actually seeking access to useful big data sources. From this perspective, the platforms being set up should allow for some flexibility as regards the provisioning and customisation of hardware and software resources, in particular to provide enough space for innovation and experimentation. In addition, this flexibility

is key to incorporate unforeseen developments due to evolving users' needs as well as to rapid technological changes. This is not always easy to ensure in central banks' environment, which is often characterised by multiple daily operations, limited acceptance for failures not least because of credibility and reputation aspects, and constrained public resources.

A third insight gained from these experiments is that **cooperation and knowledge-sharing** can add significant value, by preventing the "NIH (not invented here) syndrome", when everybody is reinventing the wheel in parallel. Indeed, central banks have a strong interest in cooperating closely with peers and counterparts, for instance with NSOs. There is also a great appetite for international cooperative approaches, such as the one underlying the activities of the BIS/IFC in supporting the global community of central bankers and financial supervisors – in particular through regular meetings and the sharing of experience from pilot projects (the Basel Process; see BIS (2019)). Such collaboration and cooperation is important not only to exchange on specific projects and experience gained, but also to concretely share data as well as IT tools – in particular software codes, for example through secure online software repositories (eg based on the free, open source system Git).[8] This can be particularly attractive for those institutions with limited capacity to invest in costly big data / HPC platforms.

Yet perhaps the most important lesson of recent initiatives is that, rather than being an isolated IT project, they **involve the whole institution**. First, the analysis and processing of increasingly large and diverse volumes of data sets calls for significant financial and staff resources that need to be committed at the upper level of the organisation. In addition, the technology choices have a fundamental impact on the full range of business activities. In particular, they call for a review of the institution's data strategy and a revamping of its organisational structures and data management frameworks – see McHugh (2019) for the IMF experience in addressing these issues.

2.2 Significant challenges

Yet, at the same time, the various projects implemented in recent years show that the **deployment and utilisation of big data/HPC platforms face a number of challenges**. The three main ones are: (i) the important resources to be mobilised: (ii) the difficulties of dealing with new types of information; and (iii) the process for implementing the related IT solutions.

As regards the first challenge, the building and effective use of the new platforms is often **constrained by the available resources**. These constraints reflect large IT costs, with the investments needed being very expensive. They are also due to human capital bottlenecks when setting up multidisciplinary teams. On the technical side, there is a need for staff with specialised statistical skills, an analytical mindset, strong IT skills and a good sense for extracting valuable new insights from data (ie "data scientists"). But additional profiles (eg economists, statisticians, computer scientists, mathematicians, lawyers) are also required when dealing with big data information

[8] It should be noted that the international statistical community including central banks is already actively engaged in public sharing of such IT tools in the context of the SDMX (Statistical Data and Metadata eXchange) standard for structuring and exchanging statistics (IFC (2016)).

and tools, and the necessary skill sets may not be available in-house. To make things worse, public institutions may be ill equipped to compete with the private industry and face the "war for talent" resulting from a limited supply of graduates. Reflecting these various constraints, a vast majority of central banks are just at the beginning of regular production of statistical and economic information based on big data sources and/or analytical tools.

The second challenge results from **the new nature of the information generated by the ongoing data revolution**. Until recently, the work of public authorities mainly relied on "traditional" statistical exercises (eg registers, surveys, censuses), with data collected, managed, compiled, disseminated and analysed according to well defined principles (United Nations (2013)). In contrast, the lack of transparency in methodologies and the poor quality of some big data sources represent important impediments for central banks, as policymaking is increasingly expected to be based on evidence and reputation is a key ingredient supporting the credibility of public decisions. These issues are compounded by the fact that a number of the new big data sets, because of their high granularity, pose important risks in terms of security, privacy and confidentiality. These risks have to be adequately addressed when setting up the distributed applications to deal with these data, requiring an explicit and comprehensive data governance framework – for instance by having strong security protection measures, sound frameworks in place, and high public transparency standards (IFC (2020)).

A third difficulty relates to the **actual implementation of the IT solutions**. In general, the various projects aim for IT infrastructure of a distributed nature, comprising:

- a centralised software layer providing a resilient environment;
- the decentralised processing on different nodes for various tasks; and
- the possibility to use different programming environments, considering in particular the variety of the use cases as well as the increasing demand for open source tools (eg Julia, Python, R).

But the choices made in practice often reflect legacy issues (eg existing platforms and established IT solutions being reused) and the limited resources available (financial but also in terms of human skills) resulting in an **heterogeneous IT environment**. It is also important to stress that the big data technology space is overwhelmed with numerous products, solutions, libraries and frameworks and is rapidly and constantly evolving, both in the open source and proprietary application spaces. Moreover, there has been huge investment in big data-related firms in recent years and the trend appears to be continuing (Graph 1); continuous uncertainty as to when the technology will actually mature raises the risk of investing too much / too early.

In any case, the variety of tools available, always to solve similar types of problems, makes the **selection process challenging**. As a result, it is nearly impossible to choose a technology that would have all the features needed by the users and would also address the key requirements in terms of IT infrastructure in terms of resilience, reliability, scalability and cost effectiveness. As a consequence, the actual selection of the technologies and their implementation often requires to make choices and trade-off different priorities. This may be particularly important for

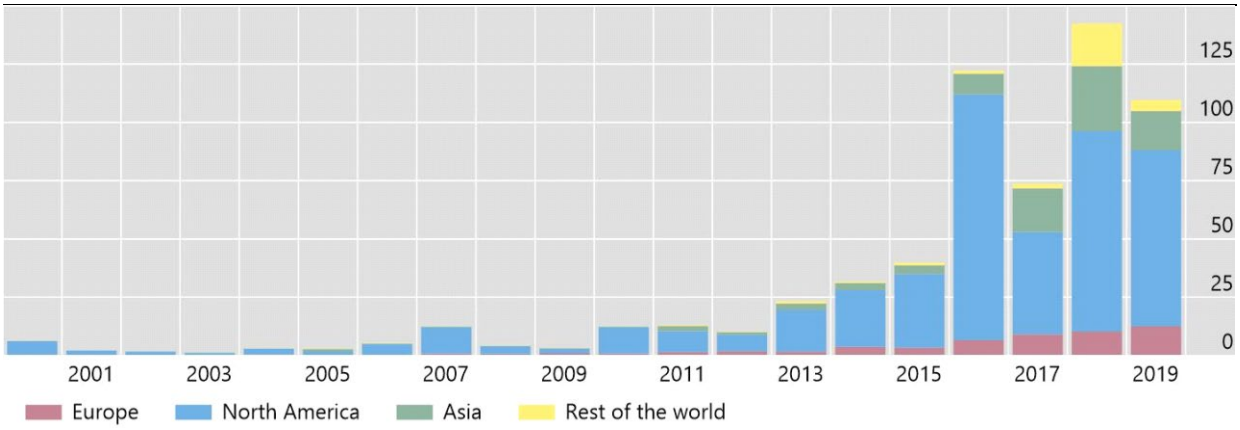
central bank users, whose requirements can be limited in particular by strong security constraints.

In view of these challenges, the remainder of this report discuss the main elements to consider to facilitate the developments of big data / HPC platforms, first when choosing the infrastructure and then when implementing it.

Capital raised by companies working in the area of big data, AI and ML

In billions of US dollars

Graph 1



Private Equity (PE), Venture Capital and Mergers & Acquisitions (M&A) capital raised by big data and artificial intelligence & machine learning companies. M&A excludes buyouts to avoid double-counting from PE.

Source: PitchBook Data Inc.

3. Main options to consider when choosing big data infrastructure

What are the options for central banks when selecting a technology and architecture to set up big data / HPC platforms? Four main issues need to be considered: the hardware selection; the choice between proprietary and open source technology; the decision to develop the solution in-house or in the cloud; and the type of information to be handled.

3.1 Hardware choice (for on-premise implementation)

Experience shows that central banks are primarily working on developing big data solutions on their own premises (see Section 3.3 for a discussion of the motivations behind this trend). The hardware to be implemented on-site thus should offer a highly scalable infrastructure to respond to users' increasing needs for parallel processing and large data storage – see Bruno et al (2019) on how these factors have been instrumental in defining the features of the new platform developed at the Bank of Italy. Broadly speaking, **four types of hardware options** are available: commodity

servers; enterprise servers; converged data infrastructure; and proprietary big data appliances.

The first option is to work with **commodity servers**, ie simple, standardised computer systems commonly referred to as "bare metal servers". These are typically made of a large number of low specification machines, instead of a few high-spec machines. A key advantage is their relatively low cost, as well as their ease of replacement. However, they require significant maintenance costs to oversee the IT operations and network teams, so they may not be the most resource-efficient solution. Moreover, they provide limited fault tolerance, with, for instance, an imperfect ability to continue working properly in the event of the failure of the entire data centre. Lastly, they are difficult to scale up, especially when the types of workloads are diverse.

A second option is to use **enterprise grade servers with virtualisation**. These servers are powerful physical machines placed in the institution's data centre and enabled with a virtualisation layer that offers flexibility in managing those machines – in practice, they are typically configured to run as multiple virtual machines (VMs) on the physical machines. These servers present many benefits in terms of cost efficiency and functionality features (eg computing power, resiliency against hardware or even data centre failure). The management of the compute nodes of big data / HPC platforms is therefore usually installed on such types of hardware.

A third option is the highly virtualised **converged data infrastructure**, ie an IT system that groups multiple components – eg hardware and servers, operating system, software for infrastructure management including VM management (hypervisor), as well as firmware and middleware. Such infrastructure provides important resiliency benefits compared with the other options, reflecting the fact that the various IT components are better integrated and optimised. However, it requires significant resource overhead to manage the virtualised environment.

The last option relates to **proprietary big data appliances**, which can be bought or rented from major hardware and software vendors and installed in the organisation's data centre. In most cases, the solution proposed by the vendor includes specific big data software and services, as well as the possibility to outsource the management of the infrastructure. The advantage in terms of ease of use, however, should be weighed against the risks posed by vendor and technology dependency, constraining the options for future technology upgrades (eg if enhancements developed by competitors cannot be provided by the vendor itself).

Choosing between these various options described above to set up the optimal hardware for the institution is not straightforward and will often depend primarily on cost aspects and actual use cases. A number of other factors are also important to consider, such as:

- Required computing intensity, in terms of the relative number of computing operations compared with I/O (input/output) communication managed by the information system
- Required storage capacity, including the need to provide hardware and software redundancy when the system is fully virtualised
- Disaster recovery requirements
- Compatibility with existing data centre technology

- In-house skills and competencies for managing network and infrastructure
- Users' needs in terms of flexibility and scalability

3.2 Proprietary versus open source technology

Most popular big data tools (including those in the Hadoop ecosystem) are available as **open source software**.[9] This solution presents a number of advantages compared with proprietary systems (ie those owned by a specific company), especially in terms of innovation (with the possibility to constantly add new features developed by the open source community), compatibility between systems and reduced dependency on service providers and IT suppliers (eg vendor risk). In addition, a key benefit is the low cost of ownership, reflecting the absence of license fees; in contrast, proprietary big data software solutions are not freely licensed, and in some cases may also require additional specialised (and costly) hardware to run those solutions.

However, the benefits listed above should be balanced against the significant **investment required for the institution to manage open source technology**. In reality, similar to proprietary systems, it will be necessary to purchase external support (consultancy services) and/or invest significant resources in training in-house staff. These costs should be factored in when deciding on adopting open source applications, despite the fact that they are available "for free". Another important risk is that the institution could end up with a very heterogeneous IT landscape – composed of software/technologies acquired from various sources, with different life cycles and ages,[10] and that are not jointly developed and tested. This can lead to system incompatibilities, performance limitations, and even security risks. Because of the way they are structured/assembled, proprietary solutions are typically better suited to address these issues. Yet this does not mean that open source technology cannot be considered too: indeed, a number of vendors offering data platform solutions (eg Cloudera, MapR) also provide packaged open source software distributions; in addition, some of them can offer their proprietary software technologies for "free" but complement this with a commercial support offering.

3.3 On-premise versus cloud-based solution

A key factor to be considered when setting up any big data architecture is whether the platform will be run in the cloud (ie with data centres made available to many users over the internet) or in the institution's own data centre (ie "on premise"). **Cloud-based solutions** provide important benefits, including their adaptability to users' increased demand ("scalability") and ease of management. In addition, most

[9] That is, available for free and distributed with its source code available for modification.

[10] These issues are reinforced by the difficulty to identify the products that will be present in the longer term at the time of the initial investment decision. This uncertainty is illustrated by the so-called hype cycle (Gartner (2018)), which basically shows that the adoption of even a successful technology follows a life cycle with different stages, from initial conception to inflated expectations, followed by growing disillusion, before the maturity stage characterised by widespread adoption. Whenever possible, it is therefore recommended to assess the maturity and long-term perspective of an open source solution before adopting it.

cloud vendors are investing heavily in AI and ML techniques, and constantly add new big data functionalities to their offering (in form of SaaS, PaaS or even IaaS).[11]

However, public authorities are paying increasing **attention to data security**, especially for confidentiality and reputation reasons. Many of them will thus tend to prefer keeping sensitive data on premise, so as to limit the potential risk of data disclosure and also for legal reasons – eg to avoid that third parties could access their data stored on a foreign server. This appears to be a key reason driving central banks' investment in on-premise data solutions.

Nevertheless, a number of central banks are also gradually considering a **"hybrid" approach** to run part of their big data applications in the cloud (and the other part on premise). Such decisions are made based on a number of considerations – eg in terms of external dependency, vendor risk, legal risk, compliance, security risk – that are carefully analysed as part of a well defined "cloud strategy".

3.4 Which types of information to handle?

The de facto standard for storing data since decades has been to use **so-called relational databases**, which rely on a specific way to organise data in tables (or "relations"). By predefining the structure (the "schema") of the databases, one can use a programming language to read, store, query and manage them – for instance with the widely used Structured Query Language (SQL). Since the big data of interest to central banks can be made of well structured information (eg large registers of financial transactions), should big data technology be used to organise these data sets? This may not be necessary, since traditional data warehousing technologies can be successfully kept – with the condition that the infrastructure allows for enough storage and processing capacities.

However, there is a growing interest in investing in big data technologies irrespective of the data format, because of their powerful analytical features, for instance to support AI/ML calculations. Moreover, these technologies are a natural fit for dealing with unstructured information (eg "not only SQL" (NoSQL) databases),[12] not least in order to handle the complexity of managing the related complex data warehouse requirements. As a result, a growing number of central banks are interested in using non-traditional database structures. Their key advantage is that they allow data to be stored without an exact predefined structure, for instance by using search engines that can retrieve information in a flexible way based on specific attributes, even if it is not stored in predefined format; this is called the "schema-on-read" approach. Another key benefit of non-traditional database structures is the concept of **schema evolution**. The information stored as is (ie in raw format) can be first represented using a standard data structure, which can then be relatively easily

[11] SaaS, PaaS, and IaaS are three ways to describe how the cloud can be used to organise the platform. SaaS (software-as-a-service): the software is available via a third party over the internet; PaaS (platform-as-a-service): hardware and software tools available over the internet; IaaS (infrastructure-as-a-service): pay-as-you-go for the whole range of cloud-based services such as storage, networking and virtualisation.

[12] See Fournier (2016) for the related initiative developed at the Bank of France to store heterogeneous data and also IFC (2019b).

changed, potentially allowing for faster / more complex analyses. In contrast, schema evolution is much harder to achieve with relational databases, built under a traditional "schema-on-write" approach.

Yet the benefits of the schema-on-read approach obviously depend on the cleanliness of the underlying data. In particular, the quality of some big data sets can be quite poor (eg missing, wrong or conflicting attributes), thereby limiting their usability. Fortunately, a number of data cleaning techniques, in particular based on AI/ML techniques, are available to help with this matter.[13]

All in all, a number of central banks are trying to build data lakes that can **deal with both SQL and NoSQL databases**, allowing different technologies to be applied in a complementary way, and depending on users' needs. An important feature is that they can continue to perform their "traditional" data management tasks, while exploring novel capabilities that are in increasing demand in their institutions. Another benefit of this exploratory approach can be parsimony, especially in case of important constraints in terms of IT resources and skills available in the institution. A third one is flexibility: the new big data technologies may not be mature enough and/or may not (yet) be ready to fully replace other existing tools, at least at a reasonable cost. Moreover, they are still evolving rapidly, with new functionality being added constantly, and are thus much less mature in comparison with traditional database technologies.

4. Implementing big data / HPC technologies

Once the main options are selected, how should one actually implement the related technologies to set up the platform? Four areas deserve specific attention: how to proceed with the implementation; the types of workload the institution wants to support; the data architecture envisaged; and the way to develop the IT solution.

4.1 Implementation approach

There are **two different main approaches** to consider when implementing big data technologies: the top-down, or "deductive" approach, and the bottom-up, or "inductive" approach.

The starting point of the **top-down approach** relates to when users have developed a certain theory based on some hypothesis. For instance, one will assume that analysing cross-border credit card transactions can help to assess tourism activity in a certain country. All this granular information will be collected and tested to know whether it can provide a good fit compared with the existing macroeconomic indicators available. The issue is therefore to use the existing data to validate (or not) the underlying model, by testing its suitability to factual observations. Such a deductive approach is typically associated with users dealing with repetitive, similar types of questions: since the questions are known, the structure of the results can be

[13] For a discussion of these quality issues for very large, well structured data on derivatives trades collected by trade repositories, see IFC (2018).

more easily guessed. The analytical top-down approach can thus easily be implemented by following a "structure, ingest and analyse" data processing cycle.

The **bottom-up approach** has an inverse process. The starting point is to collect data without having a prior idea of a specific theory; for instance, in the preceding example, one would collect all credit card operations, not just the cross-border ones. Once a substantial amount of data has been collected, one can look for patterns in that data, and then form a theory that could explain those patterns. Typically this approach will be suited to answering non-repetitive, ad hoc analytical questions, and the speed of data processing will be more important than exactness – because the calculations will rely on repetitive interactions with the user, following an "ingest, analyse and structure" data processing cycle.

In practice, big data technologies have tended to be more useful for the inductive, bottom-up approach; in contrast, relational database technologies can produce superior results with use cases that require a deductive, top-down approach. Yet, while these approaches may appear very different, they are frequently **applied in parallel**. Moreover, with database technologies and data science practices constantly evolving, implementation approaches will continue to be adapted in the foreseeable future. Hence, modern data architecture should provide for the possibility to combine the different technologies and approaches for managing data.

4.2 Handling big data through various types of workload

There are basically four main types of workload that can be envisaged when dealing with high-volume data: batch processing, interactive (ad hoc) analytics, streaming and operational workload.

With **batch processing**, the workload can be done by running different jobs as a group ("batch") without requiring end user intervention. A key advantage is that the workload can be scheduled to run depending on the resources available at a given point in time. Most batch processing is thus configured to run automatically based on some schedule or trigger. This will particularly be suited for complex data processing requiring important resources and/or time. For this type of workload, the data is typically stored in a specific way – eg the Hadoop Distributed File System (HDFS), the Apache Parquet data storage format. It is often processed in a distributed way, ie by using the different networked computers and a programming construct to manage this distribution (with programs written in languages such as Scala, Pyspark or SparkR, relying on a software layer such as Spark). The results are usually available through user-friendly interfaces – such as Apache Hive or Impala, for summarising, analysing and querying large data systems in the Hadoop ecosystem.

Turning to **interactive workload**, big data architecture is particularly suitable for use cases requiring exploratory or research work. In general, the end user does not have a fully predefined model in mind: it is through the data analysis work that new questions arise, leading to further data exploration in an interactive and repetitive way. In this type of workload, an important requirement is that the time needed to process user requests ("network latency" plus processing time) be minimal. This is key to keep users fully engaged. Another important feature is that – apart from the high requirements in terms of data storage and processing technology (a common consideration for every type of big data workload) – there is a need for adequate user

interfaces to support users' analysis work and their exploratory interaction with the system.

Increased importance has been put in this context on the use of **business intelligence (BI) tools**. The concept of BI is usually understood as encompassing the technology-driven methods and techniques (eg IT applications and practices) mobilised to collect, manage and analyse data in order to inform business decisions (IFC (2019b)). BI tools can greatly facilitate the interactive discovery process for big data, using various functions such as:

- drill-down capabilities, ie functions allowing users to move from the general representation of the data to a more granular view; this function is particularly useful for managing multidimensional databases – or "cubes" made of several "dimensions" that can be flexibly queried by users to produce a customised view of the data they are interested in, eg by using online analytical processing (OLAP) technology
- drill-across/-through capabilities, ie functions allowing access to parallel information that may be relevant to the data being analysed
- dashboards, ie reports with key indicators/analytics presented as a table/graphical user interface updated as the data come in
- interfaces allowing users to easily query and retrieve data from the system, such as SQL-enabled query interfaces

With **streaming**, the end user receives a constantly updated presentation of the data, instead of having to download the data and then process/analyse them. Streams are thus basically made of infinitely long, never-ending tables allowing data to be queried as they flow in. The technology used for this processing has developed rapidly in recent years, with tools like the open source stream-processing software platforms Apache Kafka, Spark streaming, Flink, Storm, Samza and others. Central banks are increasingly interested in this type of workload in order to analyse data in real time, for instance to monitor their internal IT network to detect cyber attacks or to supervise the functioning of financial transactions or payment systems when they are tasked with doing so (eg to detect fraudulent activities or risks of disruptions in the financial system). While most central banks have not yet developed operational data streaming capabilities, it is expected that this will become a standard requirement in the near future.

Operational workload is often implemented with streaming technologies to provide real-time insights on business transactions that are split across multiple systems. Experience shows that a technology such as Kafka is often used to support data transfer between the different operational systems, in particular to provide a comprehensive, consistent and accurate view of all the data processed (eg batch data, online data). It is also useful to track the various steps involved in analysing the data, from their root origin to the end product (ie "data lineage"). Another important system supporting time-critical operational workloads is online transaction processing (OLTP), which manages applications that modify the database (ie performing "database transactions"). Typically, OLTP will facilitate the immediate response to users' requests – except for the more complex queries that would rather be dealt with using the OLAP technology (see above).

The above list is of course a schematic presentation. In fact, there is often a **combination of these different types of workload**. An important technology in this

context is the Lambda architecture,[14] which can be used to handle massive quantities of data that can be processed under different types of workloads, in particular through batch and real-time streaming methods.

4.3 Architecture for storing, processing and querying big data

The various big data projects implemented by central banks show that there is no one-size-fits-all data architecture that can meet all types of needs in terms of workload, ability to compute operations concurrently instead of sequentially ("concurrency"), requirements for data storage, access and processing, and latency. This calls for implementing a "**polyglot architecture**", capable of addressing heterogeneous requirements. For instance, the Bank of Italy–wide big data platform covers multiple user requirements – eg data ingestion, storage and processing; use of data science, machine learning and BI applications – that require the combined use of multiple IT tools (Cariello and Quarta (2019)). Hence, the specific nature of these requirements inherently drives the development of the supporting architecture, often in a progressive and interactive way.

In practice, the big data environment architecture implemented by most organisations aims to set up a **data warehouse** (or "enterprise data warehouse" (EDW)), which is a central repository where data from disparate sources are stored, analysed and queried. This warehouse comprises both current and historical observations, with the data sourced (or "reported") from an operational data store (ODS). This "data lake" comprises all the possible information needed for feeding the warehouse (Lacroix (2019)). It covers all data irrespective of their format, their actual inclusion (or not) in the EDW, the use cases they support, and their degree of processing – with users able to access the "raw", original data and not just the "transformed" ones. For instance, the ECB big data platform (DISC) comprises a central data store including all the data of interest to the institution and a single data dictionary supporting the combination of the various data sets (Sánchez and Trzeciok (2019)).

4.4 Big data software development

Generally speaking, **big data software is typically *manually* put together** and deployed by specialised IT engineers. However, manual deployments and the related management of the (often various) big data technology tools pose significant risks, including the lack of automated testing, misconfigurations and possibly severe system outages. A more structured approach is clearly needed, and "traditional" DevOps[15] practices that clearly prevail in other domains should ideally also be applied to support an automated deployment of the big data platform. To do so, an infrastructure-as-a-code (IaaC) approach is required, using technologies such as Puppet or Ansible that can support the automation of deploying highly distributed IT infrastructure. Moreover, proper change management practices should be applied to

[14] Lambda architecture is based on serverless computing services that are run in response to events and automatically manage the underlying computing resources.

[15] DevOps is a set of practices that combines software development (Dev) and information technology operations (Ops) which aims to shorten the system's development life cycle and provide continuous delivery with high software quality.

deal with segregated environments for development, system integration, user acceptance and production usages in a comprehensive way. Central banks may also consider extending DevOps to DataOps practices, with a view to implementing automation for data management and analytical processing.

However, experience shows that **use of DevOps practices for developing big data software is still in an early adoption phase**. Surely, this a complex task due to the many moving technological parts involved in big data / HPC platforms. An additional layer of complexity comes from the distributed nature of the software, as it involves multiple machines and different configuration options. But a consequence often observed is that many projects fail to go beyond the pilot phase due to a lack of IT governance including suitable change management and DevOps processes.

Central banks are nevertheless **making progress** to tackle these issues. Successful initiatives show that, once adequate automated testing and deployment processes are set up for building the big data cluster, the IT and analytical teams are rewarded with agility, ease of reproducibility, ease of upgrades and scalability possibilities.

5. Looking forward with a comprehensive information strategy

With the exponential growth in the data central banks collect, more and more analytical needs arise that require the use of big data analytics and high-performing tools. These technologies are evolving rapidly and can **deliver significant value** to deal with the large amounts of data and apply complex AI and ML calculations. In addition, while traditional data analytics were limited to structured data, the use of big data technologies enables central banks to use alternate data sources for their various use cases.

When deciding on a project to set up a big data / HPC computing platform, **many dimensions have to be carefully taken into account**, including business requirements (use cases), technological needs, system complexity, resource constraints (eg financial costs, in-house competencies in terms of number of staff and skill mix), performance, reliability, operating model and security – see Lambe et al (2019) for a review of the key drivers influencing the platform architecture (IDEA) developed by the BIS. Moreover, these aspects should not be seen in isolation, as they need to be comprehensively addressed in the context of the organisation's **information strategy**. This strategy should basically aim at providing a high-level roadmap for the adoption of continuously changing technologies to manage data and respond to users' needs – for the situation at the Bank of England, see Vaughan and Willis (2019).

Central banks' experience with the **actual implementation** of big data related platforms and technologies shows that attention should focus on the following **main considerations to ensure success**:

- *Medium- to long-term business objectives:* Knowing what business objective to accomplish is of utmost importance when choosing the right big data technology.

- *Incremental growth:* Starting small, learning from early success, periodically monitoring the progress and reviewing the plan are key.
- *Business and IT work together:* Business and IT leaders should work together to promote a data-driven culture throughout the entire organisation (eg by well recognised "data stewards").
- *Data governance:* A data governance framework is vital for delivering continuous business benefits with a big data platform.

Lastly, **international cooperation** and knowledge-sharing between central banks can add significant value, not least considering the heterogeneous and fast-changing technological landscape. The way forward is to set up regular initiatives to ensure the sharing of experience, cross-fertilisation and the promotion of synergies among the various projects undertaken by central banks and the public community more generally.

References

Bank for International Settlements (2019): *Annual Report 2018/19*.

Bholat, D (2015): "Big data and central banks", Bank of England, *Quarterly Bulletin*, March.

Broeders, D and J Prenio (2018): "Innovative technology in financial supervision (suptech) – the experience of early users", *FSI Insights on policy implementation*, no 9, July.

Bruno, G, D Condello and A Luciuani (2019): Big Data processing: a framework suitable for Economists and Statisticians, presentation given at a workshop on computing platforms for big data and machine learning, Bank of Italy, Rome, 15 January.

Cariello, P and F Quarta (2019): BI wide big data platform, presentation given at a workshop on computing platforms for big data and machine learning, Bank of Italy, Rome, 15 January.

Cavallo, A and R Rigobon (2016): "The Billion Prices Project: Using online prices for measurement and research", *Journal of Economic Perspectives*, spring 2016, vol 30, no 2, pp 151–78.

Chessa, A, Verburg, J and L Willenborg (2017): "A comparison of price index methods for scanner data", presented at the 15th Meeting of the Ottawa Group, 10–12 May, Eltville am Rhein, Germany.

Collignon, B (2019): BOC's Analytic Environment: a leap into the future, presentation given at a workshop on computing platforms for big data and machine learning, Bank of Italy, Rome, 15 January.

Condello, D (2019): The Spark platform at ECS, presentation given at a workshop on computing platforms for big data and machine learning, Bank of Italy, Rome, 15 January.

Drozdova, A (2017): "Modern informational technologies for data analysis: from business analytics to data visualization", *IFC Bulletin*, no 43, March.

Eiglsperger, M (2019): "New features in the Harmonised Index of Consumer Prices: analytical groups, scanner data and web-scraping", *ECB Economic Bulletin*, issue 2.

Fournier, J (2016): "Regulatory reporting à la française: the Banque de France Data Lake", presentation at the European Institute of Financial Regulation Workshop, 20 September.

Gartner (2018): "Hype cycle for open-source software", *Gartner Research*, www.gartner.com/en/documents/3891628/hype-cycle-for-open-source-software-2018.

Hammer, C, D Kostroch, G Quirós and staff of the IMF Statistics Department (STA) Internal Group (2017): "Big data: potential, challenges, and statistical implications", *IMF Staff Discussion Notes*, no 17/06, September.

Ho, A and G Uddin (2019): Arrears and Credit Utilization of Canadian Households, presentation given at a workshop on computing platforms for big data and machine learning, Bank of Italy, Rome, 15 January.

Irving Fisher Committee on Central Bank Statistics (2016): "Central banks' use of the SDMX standard", *IFC Report*, March.

——— (2017): "Big data", *IFC Bulletin*, no 44, September.

——— (2018): "Central banks and trade repositories derivatives data", *IFC Report*, October.

——— (2019a): "The use of big data analytics and artificial intelligence in central banking", *IFC Bulletin*, no 50, May.

——— (2019b): "Business intelligence systems and central bank statistics", *IFC Report*, October.

——— (2020): "Current issues in data governance", *IFC Bulletin*, no 53, forthcoming.

Lacroix, R (2019): "The Bank of France datalake", *IFC Bulletin*, no 50, May.

Lambe, E, D Micic and X Sosnovsky (2019): BIS initiative to build a big data Platform, presentation given at a workshop on computing platforms for big data and machine learning, Bank of Italy, Rome, 15 January.

Laney, D (2001): "3D data management: controlling data volume, velocity, and variety", META Group (now Gartner).

McHugh, J (2019): IMF experience with Big Data, presentation given at a workshop on computing platforms for big data and machine learning, Bank of Italy, Rome, 15 January.

Marcucci, J (2019): Recent research on Big Data and Machine Learning at the Bank of Italy, presentation given at a workshop on computing platforms for big data and machine learning, Bank of Italy, Rome, 15 January.

Meeting of the Expert Group on International Statistical Classifications (2015): *Classification of Types of Big Data*, United Nations Department of Economic and Social Affairs, ESA/STAT/AC.289/26, May.

Sánchez, J-A and M Trzeciok (2019): Big data and Machine Learning initiatives at the ECB, presentation given at a workshop on computing platforms for big data and machine learning, Bank of Italy, Rome, 15 January.

Signorini, L F (2019): Opening Remarks delivered at a workshop on computing platforms for big data and machine learning, Bank of Italy, Rome, 15 January.

Tissot, B (2019): "Making the most of big data for financial stability purposes", in S Strydom and M Strydom (eds), *Big data governance and perspectives in knowledge management*, IGI Global, pp 1–24.

United Nations (2013): "Fundamental Principles of Official Statistics", Resolution adopted by the Economic and Social Council, E/RES/2013/21, 28 October.

Vaughan, N and B Willis (2019): The Bank of England's Big Data Journey, presentation given at a workshop on computing platforms for big data and machine learning, Bank of Italy, Rome, 15 January.

Wibisono, O, H D Ari, A Widjanarti, A A Zulen and B Tissot (2019): "The use of big data analytics and artificial intelligence in central banking", *The Capco Institute Journal of Financial Transformation*, Data analytics, 50th edition, November.