Comité Irving Fisher de Estadísticas del Banco Central

Informe IFC No 11

**Plataformas informáticas para análisis de big data e inteligencia artificial**

Abril de 2020

Contribuyentes al Informe IFC[[1]](#footnote-1)

Bank of Italy Giuseppe Bruno

BIS HirenJani Rafael Schmidt

IFC Secretariat Bruno Tissot

Este informe se ha elaborado sobre la base del taller sobre “Plataformas informáticas para big data y aprendizaje automático” organizado en Roma el 15 de enero de 2019 por el Banco de Italia con el apoyo del Comité Irving Fisher de Estadísticas del Banco Central (IFC) del Banco de Pagos Internacionales (BPI);

Ver [www.bis.org/ifc/events/boibis\_jan19/programme.pdf](http://www.bis.org/ifc/events/boibis_jan19/programme.pdf).

Las opiniones expresadas son las de los autores y no reflejan necesariamente las de la CFI, sus miembros, el Banco de Italia, el BPI o las instituciones representadas en la reunión.

Esta publicación está disponible en el sitio web del BPI (www.bis.org/). © Banco de Pagos Internacionales 2020. Todos los derechos reservados. Se pueden reproducir o traducir breves extractos siempre que se indique la fuente.

ISSN 1991-7511 (en línea)

ISBN 978-92-9259-362-9 (en línea)

**Plataformas informáticas para análisis de macrodatos e inteligencia artificial: experiencia de los bancos centrales[[2]](#footnote-2)**

**Resumen ejecutivo**

Las autoridades públicas, y los bancos centrales en particular, se están dando cuenta cada vez más del potencial de los grandes conjuntos de datos y análisis, con el desarrollo de técnicas de inteligencia artificial (IA) y aprendizaje automático (ML), para proporcionar información estadística nueva y complementaria (Hammer et al ( 2017)). Sin embargo, la pregunta sigue siendo: ¿cómo deberían organizarse las instituciones para beneficiarse al máximo de estas oportunidades? Dos áreas parecen particularmente importantes para los bancos centrales. El primero es cómo organizar su información estadística en relación con su infraestructura de TI. El segundo es pensar estratégicamente sobre cómo utilizar las técnicas apropiadas para procesar y analizar la nueva información recopilada.

Al igual que muchas oficinas nacionales de estadística (ONE) y organizaciones internacionales, los bancos centrales ya han lanzado numerosas iniciativas para explorar estos temas e intercambiar su experiencia, en particular a través de las actividades de cooperación organizadas por el Comité Irving Fisher de Estadísticas del Banco Central (IFC) de la Banco de Pagos Internacionales (BIS). El propio BIS ha desarrollado su nueva estrategia a mediano plazo, Innovation BIS 2025, que se basa en una importante inversión en tecnología de próxima generación para construir un lugar de trabajo digital resistente y preparado para el futuro para la organización (BIS (2019)).

Una característica clave de estas diversas iniciativas es que muchos bancos centrales están creando, o previendo implementar, plataformas de big data para facilitar el almacenamiento y procesamiento de conjuntos de datos muy grandes. También están desarrollando una infraestructura informática de alto rendimiento (HPC) que permite un procesamiento más rápido, análisis estadístico en profundidad y simulaciones de datos complejas. Sin embargo, estas iniciativas enfrentan importantes desafíos organizacionales, ya que los bancos centrales compensan factores como las tendencias tecnológicas, la complejidad del sistema, el costo, el rendimiento, la confiabilidad, el modelo operativo y la seguridad.

En vista de esta experiencia, es esencial evaluar cuidadosamente las opciones disponibles antes de seleccionar una tecnología y arquitectura para configurar plataformas de big data / HPC. Entre las diversas cuestiones a considerar, la atención debe centrarse principalmente en la selección del hardware, la elección entre tecnología patentada y de código abierto, la decisión de desarrollar la solución internamente o en la nube y el tipo de información que se manejará.

Una vez que se seleccionan las opciones principales, la implementación real de las tecnologías relacionadas suele ser un viaje largo y multiforme. Desde la perspectiva del desarrollo de un proyecto, el éxito dependerá del enfoque para llevar a cabo el proyecto, los tipos de carga de trabajo que se admitirán, la arquitectura de datos prevista y el uso de las mejores prácticas de desarrollo de software. Tener un nivel institucional más amplio.

La perspectiva también es clave, a fin de tener en cuenta adecuadamente la gama completa de requisitos comerciales, así como los recursos y las limitaciones de seguridad. Por lo tanto, se puede recomendar desarrollar una estrategia de información integral para la institución, con una hoja de ruta de alto nivel para la adopción de tecnologías en constante cambio para administrar los datos y responder a las necesidades de los usuarios. Por último, pero no menos importante, el intercambio de conocimientos puede ser fundamental y puede ser facilitado por las actividades de cooperación promovidas por el BPI y la CFI.

**1. Big data y plataformas informáticas de alto rendimiento**

**1.1 Antecedentes**

Los casos de uso de big data y ML e IA se están volviendo omnipresentes en los negocios y el mundo académico, así como en las instituciones públicas (IFC (2019a)), y abordarlos de manera efectiva requiere una innovación tecnológica continua (Signorini (2019)). Los esfuerzos se han centrado especialmente en la implementación de las llamadas plataformas de big data y / o computación de alto rendimiento (HPC). Las plataformas de big data permiten el almacenamiento y procesamiento de conjuntos de datos muy grandes, y cada vez se solicitan más para tratar información de naturaleza semiestructurada o no estructurada.[[3]](#footnote-3) Las plataformas HPC permiten un procesamiento rápido de datos y análisis / simulaciones estadísticas complejas, y generalmente son muy adecuadas para tratar con conjuntos de datos estructurados de tamaño mediano. Ciertamente, esta distinción es un poco artificial, y el límite entre las plataformas de big data y HPC se está volviendo borroso con el rápido desarrollo de las tecnologías de TI que pueden satisfacer ambos tipos de necesidades de los usuarios.

En este contexto, el Banco de Italia, con el apoyo del BIS y la CFI, organizó un taller sobre “Plataformas informáticas para big data y aprendizaje automático”. En una reunión en Roma en enero de 2019, los participantes de unas 30 organizaciones, incluidos bancos centrales, OSN y organizaciones internacionales, aprovecharon esta oportunidad para presentar el estado de sus iniciativas actuales de big data e intercambiar experiencias.

Este informe de la CFI se basa en los ejemplos revisados ​​en esa ocasión y presenta las principales tecnologías y plataformas de big data / HPC actualmente consideradas o ya implementadas por autoridades públicas como los bancos centrales. También brinda información sobre cómo abordar los desafíos relacionados y aprovechar al máximo las oportunidades que brindan estas nuevas infraestructuras para respaldar la formulación de políticas. El resto de esta sección revisa los factores clave que impulsan la demanda de big data y la tecnología relacionada, basándose en casos de uso seleccionados. La sección 2 extrae algunas lecciones de los proyectos ya establecidos y analiza las principales oportunidades y desafíos enfrentados. La Sección 3 revisa las principales opciones técnicas a considerar al elegir la tecnología relacionada. La Sección 4 proporciona una guía genérica para la implementación real de proyectos de infraestructura de big data y la construcción de las plataformas informáticas subyacentes. La sección 5 extrae algunas conclusiones generales de cara al futuro.

**1.2 Mayor interés en la información de macrodatos...**

El término big data es relativamente vago y se refiere en general a la enorme proliferación de información que trajo la llamada revolución de los datos y generada por las redes sociales, actividades basadas en la web, sensores de máquinas, operaciones financieras, administrativas o comerciales, etc. La caracterización conocida de este concepto se relaciona con las "V" de los grandes datos, como el volumen (es decir, el número de registros y atributos), la velocidad (velocidad de producción de datos, p. ej., datos de ticks) y la variedad (p. ej., estructura y formato del conjunto de datos) (Laney (2001)).

Los conjuntos de macrodatos financieros a los que se enfrentan los bancos centrales y las instituciones financieras internacionales más específicamente comprenden normalmente cuatro tipos principales de fuentes: indicadores basados ​​en Internet, registros comerciales, indicadores del mercado financiero y registros administrativos. Su análisis puede brindar muchas oportunidades, tales como: llenar los vacíos de datos existentes; la recopilación de nuevos datos que puedan complementar de manera flexible las estadísticas oficiales basadas en encuestas y censos “tradicionales”; la disponibilidad de datos casi en tiempo real, lo que abre la posibilidad de recibir señales económicas más oportunas y reducir los retrasos en la publicación de estadísticas; y la posibilidad de aplicar nuevas metodologías para extraer conocimiento original (Bholat (2015)). Estas oportunidades pueden facilitar en gran medida la conducción de las políticas del banco central, por ejemplo, mejorando los pronósticos económicos, ayudando a las tareas de micro supervisión, realizando nuevos tipos de evaluaciones de estabilidad financiera y obteniendo una retroalimentación más rápida sobre las decisiones de política (IFC (2017)).

Además de capturar datos transaccionales estructurados que se encuentran en el núcleo del trabajo de los bancos centrales, varios e innovadores conjuntos de big data, como datos granulares recopilados de la web (por ejemplo, visitas a páginas, anuncios en línea, indicadores de búsqueda), indicadores de sentimiento derivados de redes sociales. medios de comunicación (por ejemplo, Facebook, LinkedIn, Twitter), transacciones de pago e información textual tradicional procesada con nuevas herramientas de big data (por ejemplo, anuncios de políticas, discursos, correos electrónicos, artículos de prensa) - se pueden movilizar en la práctica para facilitar el día a día de los bancos centrales trabajar en diversas áreas, incluida la producción de información estadística, el análisis económico y la investigación en apoyo de las políticas de estabilidad monetaria y financiera, y tareas de supervisión (Tissot (2019)) .[[4]](#footnote-4)

En lo que respecta al trabajo estadístico, los macrodatos pueden mejorar la calidad de las estadísticas “más convencionales” existentes o incluso complementarlas con nuevos tipos de indicadores. Un buen ejemplo, particularmente importante en vista de los mandatos de estabilidad de precios de los bancos centrales, se relaciona con la medición y el análisis de los patrones de inflación. El uso de nuevos métodos de recopilación del tipo de big data se ha expandido significativamente, lo que permite obtener información de precios más detallada, una mayor variabilidad del producto y una mayor cantidad de transacciones para capturar (Eiglsperger (2019)). En particular, el raspado web directo de los datos de precios de los minoristas en línea puede complementar la cobertura de la recopilación de precios "tradicional" basada en encuestas en las tiendas físicas, mejorar la medición de componentes volátiles específicos de la inflación, proporcionar estimaciones en tiempo real y / o de alta frecuencia del índice de precios al consumidor (IPC) y llenar los vacíos de recopilación de datos existentes en algunas regiones (ver el Proyecto de mil millones de precios desarrollado por MIT Sloan School of Management; Cavallo y Rigobon (2016)). Las autoridades públicas también están tratando de hacer uso de los volúmenes en rápida expansión de conjuntos de datos de escáner,[[5]](#footnote-5) que brindan un potencial importante, por ejemplo, para rastrear patrones de gasto de manera granular o para aumentar la precisión de las cifras del IPC (Chessa et al (2017)).[[6]](#footnote-6)

**1.3 ... y análisis de big data**

Frente a una gran cantidad de macrodatos que se expande rápidamente, las autoridades públicas deben extraer información relevante y darle sentido para poder tomar decisiones políticas informadas y basadas en evidencia, dando prioridad a la transformación de los datos en información y luego de la información en conocimiento (Drozdova (2017)). )). La forma de hacerlo es utilizar el llamado análisis de big data, mediante el cual se analizan los diversos conjuntos de datos para descubrir información como patrones específicos, correlaciones, tendencias, etc. Muchas técnicas estadísticas, incluidas las que admiten ML e IA (así como más tradicionales también) se pueden utilizar en estos conjuntos de datos para respaldar el reconocimiento de patrones, el descubrimiento de conocimientos, la minería de datos, etc. (IFC (2019a)).

En lo que respecta a la política de apoyo al trabajo analítico y de investigación de los bancos centrales, el análisis de macrodatos se está aplicando para una amplia gama de propósitos, especialmente en el área de la previsión económica (por ejemplo, para indicadores como la inflación, los precios de la vivienda, el desempleo, el PIB y la producción industrial ventas minoristas, desarrollos del sector externo) y análisis del ciclo económico (por ejemplo, compilación de indicadores de sentimiento y uso de técnicas de predicción inmediata para “pronosticar” el presente), que son ingredientes clave que respaldan la conducción de la política monetaria. Las fuentes de macrodatos y el análisis también son cada vez más útiles para su trabajo de estabilidad financiera, por ejemplo, para construir indicadores de riesgo, evaluar el comportamiento de los participantes del mercado, identificar el riesgo crediticio y de mercado y monitorear las transacciones financieras y los flujos de capital.

Por último, la información y las herramientas de big data también están desempeñando un papel cada vez más importante en el apoyo a los ejercicios de evaluación y vigilancia de riesgos financieros que se encuentran en el centro de los mandatos de las autoridades supervisoras, incluidos los bancos centrales que tienen la tarea directa de supervisar instituciones financieras como los bancos comerciales así como el funcionamiento de los sistemas de pago. En particular, la innovación de TI ha abierto vías prometedoras para utilizar la gran cantidad de información que conllevan los conjuntos de datos granulares, con el papel cada vez mayor que desempeña la “suptech” (Broeders y Prenio (2018)). Esto puede ayudar, por ejemplo, a detectar el abuso del mercado mediante el uso de técnicas de minería de texto que señalan la mala conducta o el uso de información privilegiada; detectar patrones extraños en los datos que indiquen la acumulación de posibles vulnerabilidades idiosincrásicas; e identificar los efectos de la red que respaldan la evaluación de los riesgos de todo el sistema (Wibisono et al (2019)).

Como reflejo de estas diversas oportunidades, una prioridad clave para los bancos centrales es invertir en infraestructura de big data. Ante un aumento significativo esperado en los casos de uso y las demandas de computación / almacenamiento relacionadas, ya han estado realizando un extenso trabajo preparatorio en términos de infraestructura de TI, conjuntos de habilidades, recursos, procesos, etc. También están movilizando importantes recursos financieros y de personal. Es muy probable que esta tendencia continúe en el futuro previsible, ya que los múltiples conocimientos que se pueden obtener para respaldar las políticas se harán cada vez más evidentes y porque la tecnología de TI está evolucionando rápidamente.

**1.4 Computación de alto rendimiento (HPC)**

Las autoridades públicas también se están centrando en el desarrollo de plataformas HPC para manejar conjuntos de datos en crecimiento exponencial y, en particular, ejecutar algoritmos y herramientas analíticas multifacéticas. Si bien los casos de uso de análisis de big data se relacionan principalmente con el manejo de conjuntos de datos grandes y / o complejos (por ejemplo, no estructurados), el ímpetu para desarrollar HPC proviene de la necesidad de tener más potencia informática en comparación con una computadora de escritorio típica, una estación de trabajo o un entorno de servidor estándar. El objetivo, por tanto, es disponer de recursos suficientes para resolver, simular o analizar problemas estadísticos complejos, y no exclusivamente para manejar una gran cantidad de datos. En la práctica, sin embargo, estos dos tipos de necesidades a menudo están interrelacionados (ver más abajo).

La experiencia de los bancos centrales muestra que las plataformas HPC se desarrollan principalmente para garantizar que los recursos informáticos se utilicen de la manera más eficiente, de modo que los procesos analíticos puedan completarse lo más rápidamente posible. La solución es agregar la capacidad de TI, con plataformas HPC que se hacen como grupos de computadoras individuales, denominadas nodos - un tamaño de grupo común es de entre 16 y 64 nodos, lo que representa de 64 a 256 unidades de procesamiento ("núcleos");[[7]](#footnote-7) ver, por ejemplo, Ho y Uddin (2019) para las especificaciones de la plataforma informática establecida en el Banco de Canadá. Dado el alto costo de comprar y operar tales plataformas, las soluciones implementadas generalmente se comparten entre muchos usuarios a nivel departamental o institucional. Para hacerlo, los subconjuntos de los nodos informáticos se asignan comúnmente a "trabajos" dedicados para atender las solicitudes específicas de los usuarios, según los recursos de TI disponibles. Por lo tanto, el papel clave de una plataforma HPC es definir las prioridades de los diferentes trabajos a ejecutar por los nodos informáticos, al tiempo que se asegura que los recursos no se sobrecarguen. Por lo general, uno tiene que implementar el procesamiento por lotes (ver más abajo) y los programadores, por ejemplo, el programador YARN (Yet Another Resource Negotiator) para usar la biblioteca Hadoop desarrollada por la base de software Apache de código abierto al enviar trabajos, preguntar sobre su estado y modificarlos. Sin interacción del usuario final.

Una característica reciente entre los bancos centrales que ya ejecutan algún tipo de plataforma HPC es extender esta infraestructura para integrar otras formas de computación, en particular herramientas analíticas de big data, para distribuir el procesamiento de grandes conjuntos de datos entre grupos de computadoras. Construyen un denominado clúster de big data, formado por ordenadores conectados y diseñado específicamente para almacenar y analizar grandes cantidades de datos estructurados / no estructurados de forma distribuida. Con esta tendencia que se espera que continúe en el futuro, las plataformas HPC de próxima generación estarán cada vez más capacitadas para hacer frente a la informática a gran escala y de alto rendimiento. Por ejemplo, el proyecto Apache Spark proporciona un motor analítico unificado para el procesamiento de datos a gran escala y permite que varios tipos de sistemas de archivos de big data, como el sistema de archivos distribuido de Hadoop (HDFS), se manejen de manera flexible. También es compatible con varias interfaces de programación como Java, Python, R o Scala, lo que permite realizar cálculos complejos; consulte Condello (2019) para conocer las características de la plataforma Spark desarrollada en el Banco de Italia. Otro desarrollo esperado es el impulso para implementar esas soluciones de TI en un modo autónomo, es decir, sin requerir una infraestructura adicional pesada (por ejemplo, sin necesidad de tener un programador específico también). Esta característica independiente parece facilitar enormemente el procesamiento de trabajos de HPC.

Otra vía más, con la que algunas instituciones, como el Banco de Canadá, ya han experimentado, es explorar la computación cuántica para el entorno analítico de datos de próxima generación (Collignon (2019)). Una computadora cuántica es capaz de generar y manipular bits cuánticos (o qubits, que pueden ser 0, 1 o 0 y 1 al mismo tiempo); su computación es intrínsecamente paralela y, por lo tanto, su potencia puede crecer exponencialmente con el número de qubits. Para las tareas que actualmente requerirían cientos de años para ser resueltas por computadoras estándar, se espera que una computadora cuántica pueda reducir el marco de tiempo de manera muy significativa, posiblemente a unos pocos minutos.

**2. Lecciones ya aprendidas**

**2.1 Implementación de las nuevas plataformas**

La retroalimentación general de los proyectos ya implementados por los bancos centrales es que las nuevas plataformas de Big Data y HPC pueden servir para múltiples propósitos. Una es confirmar los análisis preliminares existentes basados ​​en métodos econométricos y estadísticos más tradicionales y / o en conjuntos de datos más pequeños. El segundo es proporcionar información complementaria, en particular en comparación con fuentes de datos convencionales, como encuestas y censos. Y una tercera contribución es identificar problemas imprevistos, lo que a su vez plantea la necesidad de información y / o investigación adicional. Para abordar estos diversos casos de uso, es importante que las plataformas que se están configurando proporcionen un repositorio de datos común, de modo que los usuarios puedan encontrar una gran cantidad de información diferente que se pueda combinar y analizar de acuerdo con sus necesidades específicas.

Otra lección clave es que el enfoque seguido es intrínsecamente de aprender haciendo, con proyectos que se adaptan constantemente a medida que avanza la experiencia. El método preferido ha sido desarrollar nuevos proyectos de manera gradual, comenzando con iniciativas de investigación individuales e implementando casos de uso preliminares (Marcucci (2019)); Este enfoque progresivo ha resultado particularmente útil dados los diversos obstáculos que se enfrentan cuando se busca acceder a fuentes útiles de macrodatos. Desde esta perspectiva, las plataformas que se están estableciendo deberían permitir cierta flexibilidad en lo que respecta al aprovisionamiento y la personalización de los recursos de hardware y software, en particular para proporcionar suficiente espacio para la innovación y la experimentación. Además, esta flexibilidad es clave para incorporar desarrollos imprevistos debido a la evolución de las necesidades de los usuarios, así como a los rápidos cambios tecnológicos. Esto no siempre es fácil de garantizar en el entorno de los bancos centrales, que a menudo se caracteriza por múltiples operaciones diarias, una aceptación limitada de las fallas, sobre todo por aspectos de credibilidad y reputación, y recursos públicos limitados.

Una tercera percepción obtenida de estos experimentos es que la cooperación y el intercambio de conocimientos pueden agregar un valor significativo al prevenir el “síndrome de los NIH (no inventado aquí)”, cuando todos están reinventando la rueda en paralelo. De hecho, los bancos centrales tienen un gran interés en cooperar estrechamente con sus pares y contrapartes, por ejemplo, con las OSN. También existe un gran apetito por los enfoques cooperativos internacionales, como el que subyace a las actividades del BPI / IFC en el apoyo a la comunidad mundial de banqueros centrales y supervisores financieros, en particular a través de reuniones periódicas y el intercambio de experiencias de proyectos piloto (el Proceso de Basilea; véase BPI (2019)). Dicha colaboración y cooperación es importante no solo para intercambiar proyectos específicos y la experiencia adquirida, sino también para compartir de manera concreta datos y herramientas de TI, en particular códigos de software, por ejemplo, a través de repositorios de software seguros en línea (por ejemplo, basados ​​en el código abierto y gratuito). Sistema Git).[[8]](#footnote-8) Esto puede ser particularmente atractivo para aquellas instituciones con capacidad limitada para invertir en costosas plataformas de Big Data / HPC.

Sin embargo, quizás la lección más importante de las iniciativas recientes es que, en lugar de ser un proyecto de TI aislado, involucran a toda la institución. Primero, el análisis y procesamiento de volúmenes cada vez más grandes y diversos de conjuntos de datos requiere importantes recursos financieros y de personal que deben comprometerse en el nivel superior de la organización. Además, las opciones tecnológicas tienen un impacto fundamental en toda la gama de actividades comerciales. En particular, exigen una revisión de la estrategia de datos de la institución y una renovación de sus estructuras organizativas y marcos de gestión de datos; véase McHugh (2019) para conocer la experiencia del FMI al abordar estos problemas.

**2.2 Desafíos importantes**

Sin embargo, al mismo tiempo, los diversos proyectos implementados en los últimos años muestran que la implementación y la utilización de plataformas de big data / HPC enfrentan una serie de desafíos. Los tres principales son: (i) los importantes recursos a movilizar; (ii) las dificultades para manejar nuevos tipos de información; y (iii) el proceso de implementación de las soluciones de TI relacionadas.

En cuanto al primer desafío, la construcción y el uso eficaz de las nuevas plataformas a menudo se ven limitados por los recursos disponibles. Estas limitaciones reflejan grandes costos de TI, y las inversiones necesarias son muy caras. También se deben a los cuellos de botella del capital humano a la hora de configurar equipos multidisciplinares. En el aspecto técnico, existe la necesidad de personal con habilidades estadísticas especializadas, una mentalidad analítica, sólidas habilidades de TI y un buen sentido para extraer nuevos conocimientos valiosos de los datos (es decir, "científicos de datos"). Pero también se requieren perfiles adicionales (por ejemplo, economistas, estadísticos, informáticos, matemáticos, abogados) cuando se trata de información y herramientas de big data, y es posible que los conjuntos de habilidades necesarios no estén disponibles internamente. Para empeorar las cosas, las instituciones públicas pueden estar mal equipadas para competir con la industria privada y enfrentar la “guerra por el talento” resultante de una oferta limitada de graduados. Como reflejo de estas diversas limitaciones, la gran mayoría de los bancos centrales se encuentran apenas en el comienzo de la producción regular de información estadística y económica basada en fuentes de macrodatos y / o herramientas analíticas.

El segundo desafío es el resultado de la nueva naturaleza de la información generada por la revolución de datos en curso. Hasta hace poco, el trabajo de las autoridades públicas se basaba principalmente en ejercicios estadísticos “tradicionales” (por ejemplo, registros, encuestas, censos), con datos recopilados, gestionados, compilados, difundidos y analizados de acuerdo con principios bien definidos (Naciones Unidas (2013)). Por el contrario, la falta de transparencia en las metodologías y la mala calidad de algunas fuentes de macrodatos representan impedimentos importantes para los bancos centrales, ya que se espera cada vez más que la formulación de políticas se base en pruebas y la reputación es un ingrediente clave que respalda la credibilidad de las decisiones públicas. Estos problemas se ven agravados por el hecho de que varios de los nuevos conjuntos de big data, debido a su alta granularidad, plantean riesgos importantes en términos de seguridad, privacidad y confidencialidad. Estos riesgos deben abordarse de manera adecuada al configurar las aplicaciones distribuidas para manejar estos datos, lo que requiere un marco de gobernanza de datos explícito y completo, por ejemplo, teniendo fuertes medidas de protección de seguridad, marcos sólidos establecidos y altos estándares de transparencia pública (IFC ( 2020)).

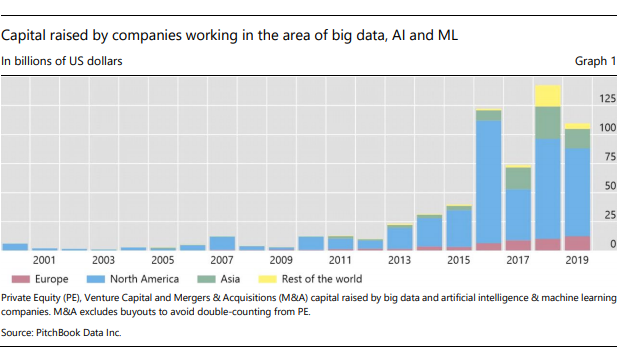
Una tercera dificultad se relaciona con la implementación real de las soluciones de TI. En general, los distintos proyectos tienen como objetivo una infraestructura informática de carácter distribuido, que comprende:

* Una capa de software centralizada que proporciona un entorno resistente;
* El procesamiento descentralizado en diferentes nodos para diversas tareas; y
* La posibilidad de utilizar diferentes entornos de programación, considerando en particular la variedad de casos de uso, así como la creciente demanda de herramientas de código abierto (por ejemplo, Julia, Python, R).

Pero las decisiones tomadas en la práctica a menudo reflejan problemas heredados (por ejemplo, plataformas existentes y soluciones de TI establecidas que se reutilizan) y los recursos limitados disponibles (financieros pero también en términos de habilidades humanas) que dan como resultado un entorno de TI heterogéneo. También es importante enfatizar que el espacio de la tecnología de big data está abrumado con numerosos productos, soluciones, bibliotecas y marcos y está en rápida y constante evolución, tanto en los espacios de aplicaciones de código abierto como patentadas. Además, ha habido una gran inversión en empresas relacionadas con big data en los últimos años y la tendencia parece continuar (Gráfico 1); La continua incertidumbre sobre cuándo madurará realmente la tecnología aumenta el riesgo de invertir demasiado o demasiado pronto.

En cualquier caso, la variedad de herramientas disponibles, siempre para resolver problemas similares, hace que el proceso de selección sea un desafío. Como resultado, es casi imposible elegir una tecnología que tenga todas las características necesarias para los usuarios y que también aborde los requisitos clave en términos de infraestructura de TI en términos de resiliencia, confiabilidad, escalabilidad y rentabilidad. Como consecuencia, la selección real de las tecnologías y su implementación a menudo requiere tomar decisiones y compensar diferentes prioridades. Esto puede ser particularmente importante para los usuarios del banco central, cuyos requisitos pueden verse limitados, en particular, por fuertes restricciones de seguridad.

Ante estos desafíos, en el resto de este informe se analizan los principales elementos a considerar para facilitar los desarrollos de plataformas de big data / HPC, primero a la hora de elegir la infraestructura y luego al implementarla.



**3. Principales opciones a considerar al elegir la infraestructura de big data**

¿Cuáles son las opciones para los bancos centrales al seleccionar una tecnología y arquitectura para configurar plataformas de big data / HPC? Deben considerarse cuatro cuestiones principales: la selección del hardware; la elección entre tecnología patentada y de código abierto; la decisión de desarrollar la solución internamente o en la nube; y el tipo de información a manejar.

**3.1 Elección de hardware (para implementación local)**

La experiencia muestra que los bancos centrales están trabajando principalmente en el desarrollo de soluciones de macrodatos en sus propias instalaciones (consulte la Sección 3.3 para una discusión de las motivaciones detrás de esta tendencia). Por lo tanto, el hardware que se implementará en el sitio debe ofrecer una infraestructura altamente escalable para responder a las crecientes necesidades de los usuarios de procesamiento paralelo y almacenamiento de datos de gran tamaño; consulte Bruno et al (2019) sobre cómo estos factores han sido fundamentales para definir las características de las nueva plataforma desarrollada en el Banco de Italia. En términos generales, hay cuatro tipos de opciones de hardware disponibles: servidores básicos; servidores empresariales; infraestructura de datos convergentes; y dispositivos de big data patentados.

La primera opción es trabajar con servidores básicos, es decir, sistemas informáticos simples y estandarizados comúnmente denominados "servidores bare metal". Por lo general, se fabrican con una gran cantidad de máquinas de baja especificación, en lugar de unas pocas máquinas de alta especificación. Una ventaja clave es su costo relativamente bajo, así como su facilidad de reemplazo. Sin embargo, requieren costos de mantenimiento significativos para supervisar las operaciones de TI y los equipos de red, por lo que pueden no ser la solución más eficiente en recursos. Además, brindan tolerancia a fallas limitada, con, por ejemplo, una capacidad imperfecta para continuar funcionando correctamente en caso de falla de todo el centro de datos. Por último, son difíciles de ampliar, especialmente cuando los tipos de cargas de trabajo son diversos.

Una segunda opción es utilizar servidores de nivel empresarial con virtualización. Estos servidores son potentes máquinas físicas ubicadas en el centro de datos de la institución y habilitadas con una capa de virtualización que ofrece flexibilidad en la gestión de esas máquinas; en la práctica, normalmente están configuradas para ejecutarse como múltiples máquinas virtuales (VM) en las máquinas físicas. Estos servidores presentan muchos beneficios en términos de rentabilidad y características de funcionalidad (por ejemplo, potencia de cómputo, resistencia al hardware o incluso fallas del centro de datos). Por lo tanto, la gestión de los nodos informáticos de las plataformas de big data / HPC suele instalarse en este tipo de hardware.

Una tercera opción es la infraestructura de datos convergentes altamente virtualizada, es decir, un sistema de TI que agrupa varios componentes, por ejemplo, hardware y servidores, sistema operativo, software para la gestión de la infraestructura, incluida la gestión de VM (hipervisor), así como firmware y middleware. Dicha infraestructura proporciona importantes beneficios de resiliencia en comparación con las otras opciones, lo que refleja el hecho de que los diversos componentes de TI están mejor integrados y optimizados. Sin embargo, requiere una sobrecarga de recursos significativa para administrar el entorno virtualizado.

La última opción se relaciona con los dispositivos de big data patentados, que pueden comprarse o alquilarse a los principales proveedores de hardware y software e instalarse en el centro de datos de la organización. En la mayoría de los casos, la solución propuesta por el proveedor incluye software y servicios específicos de big data, así como la posibilidad de subcontratar la gestión de la infraestructura. Sin embargo, la ventaja en términos de facilidad de uso debe sopesarse con los riesgos planteados por el proveedor y la dependencia de la tecnología, lo que limita las opciones para futuras actualizaciones de tecnología (por ejemplo, si el proveedor no puede proporcionar las mejoras desarrolladas por los competidores).

La elección entre estas diversas opciones descritas anteriormente para configurar el hardware óptimo para la institución no es sencillo y, a menudo, dependerá principalmente de los aspectos de costos y los casos de uso reales. También es importante tener en cuenta otros factores, como:

* Intensidad de computación requerida, en términos del número relativo de operaciones de computación en comparación con la comunicación de E / S (entrada / salida) administrada por el sistema de información
* Capacidad de almacenamiento requerida, incluida la necesidad de proporcionar redundancia de hardware y software cuando el sistema está completamente virtualizado
* Requisitos de recuperación ante desastres
* Compatibilidad con la tecnología existente del centro de datos
* Habilidades y competencias internas para administrar la red y la infraestructura.
* Necesidades de los usuarios en términos de flexibilidad y escalabilidad.

**3.2 Tecnología patentada versus tecnología de código abierto**

Las herramientas de big data más populares (incluidas las del ecosistema de Hadoop) están disponibles como software de código abierto.[[9]](#footnote-9) Esta solución presenta una serie de ventajas en comparación con los sistemas propietarios (es decir, los que pertenecen a una empresa específica), especialmente en términos de innovación (con el posibilidad de agregar constantemente nuevas características desarrolladas por la comunidad de código abierto), compatibilidad entre sistemas y menor dependencia de los proveedores de servicios y proveedores de TI (por ejemplo, riesgo de proveedor). Además, un beneficio clave es el bajo costo de propiedad, que refleja la ausencia de derechos de licencia; por el contrario, las soluciones de software de big data patentadas no tienen licencia gratuita y, en algunos casos, también pueden requerir hardware adicional especializado (y costoso) para ejecutar esas soluciones.

Sin embargo, los beneficios enumerados anteriormente deben equilibrarse con la importante inversión requerida para que la institución administre la tecnología de código abierto. En realidad, al igual que en los sistemas propietarios, será necesario adquirir apoyo externo (servicios de consultoría) y / o invertir importantes recursos en la formación del personal interno. Estos costos deben tenerse en cuenta a la hora de decidir la adopción de aplicaciones de código abierto, a pesar de que están disponibles “gratis”. Otro riesgo importante es que la institución podría terminar con un panorama de TI muy heterogéneo, compuesto por software / tecnologías adquiridos de diversas fuentes, con diferentes ciclos de vida y edades,[[10]](#footnote-10) y que no se desarrollan y prueban conjuntamente. Esto puede provocar incompatibilidades del sistema, limitaciones de rendimiento e incluso riesgos de seguridad. Debido a la forma en que están estructuradas / ensambladas, las soluciones propietarias suelen ser más adecuadas para abordar estos problemas. Sin embargo, esto no significa que no se pueda considerar también la tecnología de fuente abierta: de hecho, varios proveedores que ofrecen soluciones de plataforma de datos (por ejemplo, Cloudera, MapR) también ofrecen distribuciones de software de fuente abierta empaquetadas; Además, algunos de ellos pueden ofrecer sus tecnologías de software patentado de forma “gratuita”, pero complementarlo con una oferta de soporte comercial.

**3.3 Solución local versus solución basada en la nube**

Un factor clave a considerar al configurar cualquier arquitectura de big data es si la plataforma se ejecutará en la nube (es decir, con centros de datos disponibles para muchos usuarios a través de Internet) o en el propio centro de datos de la institución (es decir, "en las instalaciones"). ). Las soluciones basadas en la nube brindan importantes beneficios, incluida su adaptabilidad al aumento de la demanda de los usuarios ("escalabilidad") y la facilidad de administración. Además, la mayoría de los proveedores de la nube están invirtiendo fuertemente en técnicas de IA y ML, y constantemente agregan nuevas funcionalidades de big data a su oferta (en forma de SaaS, PaaS o incluso IaaS).[[11]](#footnote-11)

Sin embargo, las autoridades públicas están prestando cada vez más atención a la seguridad de los datos, especialmente por razones de confidencialidad y reputación. Por lo tanto, muchos de ellos tenderán a preferir mantener los datos confidenciales en las instalaciones, para limitar el riesgo potencial de divulgación de datos y también por razones legales, por ejemplo, para evitar que terceros puedan acceder a sus datos almacenados en un servidor externo. Esta parece ser una razón clave que impulsa la inversión de los bancos centrales en soluciones de datos locales.

No obstante, varios bancos centrales también están considerando gradualmente un enfoque "híbrido" para ejecutar parte de sus aplicaciones de big data en la nube (y la otra parte en las instalaciones). Dichas decisiones se toman en función de una serie de consideraciones, por ejemplo, en términos de dependencia externa, riesgo del proveedor, riesgo legal, cumplimiento, riesgo de seguridad, que se analizan cuidadosamente como parte de una “estrategia de nube” bien definida.

**3.4 ¿Qué tipo de información manejar?**

El estándar de facto para almacenar datos desde hace décadas ha sido el uso de las llamadas bases de datos relacionales, que se basan en una forma específica de organizar los datos en tablas (o "relaciones"). Al predefinir la estructura (el “esquema”) de las bases de datos, se puede usar un lenguaje de programación para leerlas, almacenarlas, consultarlas y administrarlas, por ejemplo, con el lenguaje de consulta estructurado (SQL) ampliamente utilizado. Dado que los macrodatos de interés para los bancos centrales pueden estar compuestos de información bien estructurada (por ejemplo, grandes registros de transacciones financieras), ¿debería utilizarse la tecnología de macrodatos para organizar estos conjuntos de datos? Esto puede no ser necesario, ya que las tecnologías tradicionales de almacenamiento de datos se pueden mantener con éxito, con la condición de que la infraestructura permita suficientes capacidades de almacenamiento y procesamiento.

Sin embargo, existe un interés creciente en invertir en tecnologías de big data independientemente del formato de los datos, debido a sus potentes funciones analíticas, por ejemplo, para admitir cálculos de IA / ML. Además, estas tecnologías son ideales para tratar con información no estructurada (por ejemplo, bases de datos “no solo SQL” (NoSQL)),[[12]](#footnote-12) sobre todo para manejar la complejidad de administrar los complejos requisitos del almacén de datos relacionados. Como resultado, un número creciente de bancos centrales está interesado en utilizar estructuras de bases de datos no tradicionales. Su principal ventaja es que permiten que los datos se almacenen sin una estructura predefinida exacta, por ejemplo, mediante el uso de motores de búsqueda que pueden recuperar información de forma flexible en función de atributos específicos, incluso si no se almacena en un formato predefinido; esto se llama el enfoque de "esquema de lectura". Otro beneficio clave de las estructuras de bases de datos no tradicionales es el concepto de evolución del esquema. La información almacenada tal cual (es decir, en formato sin procesar) se puede representar primero utilizando una estructura de datos estándar, que luego se puede cambiar con relativa facilidad, lo que potencialmente permite análisis más rápidos / más complejos. En contraste, la evolución del esquema es mucho más difícil de lograr con las bases de datos relacionales, construidas bajo un enfoque tradicional de "esquema sobre escritura".

Sin embargo, los beneficios del enfoque de esquema en lectura obviamente dependen de la limpieza de los datos subyacentes. En particular, la calidad de algunos grandes conjuntos de datos puede ser bastante deficiente (por ejemplo, atributos faltantes, incorrectos o en conflicto), lo que limita su usabilidad. Afortunadamente, hay una serie de técnicas de limpieza de datos, en particular basadas en técnicas de IA / ML, disponibles para ayudar con este asunto.[[13]](#footnote-13)

En general, varios bancos centrales están tratando de construir lagos de datos que puedan manejar bases de datos SQL y NoSQL, permitiendo que se apliquen diferentes tecnologías de manera complementaria y en función de las necesidades de los usuarios. Una característica importante es que pueden continuar realizando sus tareas de administración de datos "tradicionales", mientras exploran capacidades novedosas que tienen una demanda creciente en sus instituciones. Otro beneficio de este enfoque exploratorio puede ser la parsimonia, especialmente en caso de limitaciones importantes en términos de recursos y habilidades de TI disponibles en la institución. Un tercero es la flexibilidad: es posible que las nuevas tecnologías de big data no estén lo suficientemente maduras y / o es posible que (todavía) no estén listas para reemplazar completamente otras herramientas existentes, al menos a un costo razonable. Además, todavía están evolucionando rápidamente, con nuevas funcionalidades que se agregan constantemente y, por lo tanto, son mucho menos maduras en comparación con las tecnologías de bases de datos tradicionales.

**4. Implementación de tecnologías de macrodatos / HPC**

Una vez que se seleccionan las opciones principales, ¿cómo se deben implementar realmente las tecnologías relacionadas para configurar la plataforma? Cuatro áreas merecen atención específica: cómo proceder con la implementación; los tipos de carga de trabajo que la institución quiere soportar; la arquitectura de datos prevista; y la forma de desarrollar la solución de TI.

**4.1 Enfoque de implementación**

Hay dos enfoques principales diferentes a considerar al implementar tecnologías de big data: el enfoque de arriba hacia abajo o "deductivo" y el enfoque de abajo hacia arriba o "inductivo".

El punto de partida del enfoque de arriba hacia abajo se relaciona con cuando los usuarios han desarrollado una cierta teoría basada en alguna hipótesis. Por ejemplo, se supondrá que el análisis de las transacciones transfronterizas de tarjetas de crédito puede ayudar a evaluar la actividad turística en un determinado país. Toda esta información granular será recopilada y probada para saber si puede proporcionar un buen ajuste en comparación con los indicadores macroeconómicos existentes disponibles. Por lo tanto, la cuestión es utilizar los datos existentes para validar (o no) el modelo subyacente, probando su idoneidad para las observaciones fácticas. Este enfoque deductivo se asocia típicamente con usuarios que tratan con tipos de preguntas repetitivas y similares: dado que las preguntas son conocidas, la estructura de los resultados se puede adivinar más fácilmente. El enfoque analítico de arriba hacia abajo se puede implementar fácilmente siguiendo un ciclo de procesamiento de datos de “estructura, ingesta y análisis”.

El enfoque de abajo hacia arriba tiene un proceso inverso. El punto de partida es recolectar datos sin tener una idea previa de una teoría específica; por ejemplo, en el ejemplo anterior, se cobrarían todas las operaciones con tarjeta de crédito, no solo las transfronterizas. Una vez que se ha recopilado una cantidad sustancial de datos, uno puede buscar patrones en esos datos y luego formar una teoría que pueda explicar esos patrones. Por lo general, este enfoque será adecuado para responder preguntas analíticas ad hoc no repetitivas, y la velocidad del procesamiento de datos será más importante que la exactitud, porque los cálculos se basarán en interacciones repetitivas con el usuario, siguiendo una "ingesta, análisis y estructura” Ciclo de procesamiento de datos. En la práctica, las tecnologías de macrodatos han tendido a ser más útiles para el enfoque inductivo de abajo hacia arriba; por el contrario, las tecnologías de bases de datos relacionales pueden producir resultados superiores con casos de uso que requieren un enfoque deductivo de arriba hacia abajo. Sin embargo, aunque estos enfoques pueden parecer muy diferentes, con frecuencia se aplican en paralelo. Además, con las tecnologías de bases de datos y las prácticas de ciencia de datos en constante evolución, los enfoques de implementación seguirán adaptándose en el futuro previsible. Por tanto, la arquitectura de datos moderna debería ofrecer la posibilidad de combinar las diferentes tecnologías y enfoques para la gestión de datos.

**4.2 Manejo de macrodatos a través de varios tipos de cargas de trabajo**

Básicamente, hay cuatro tipos principales de carga de trabajo que se pueden considerar cuando se trata de datos de gran volumen: procesamiento por lotes, análisis interactivo (ad hoc), transmisión y carga de trabajo operativa.

Con el procesamiento por lotes, la carga de trabajo se puede realizar ejecutando diferentes trabajos como un grupo ("lote") sin requerir la intervención del usuario final. Una ventaja clave es que la carga de trabajo se puede programar para que se ejecute en función de los recursos disponibles en un momento determinado. Por lo tanto, la mayoría del procesamiento por lotes está configurado para ejecutarse automáticamente en función de algún programa o desencadenante. Esto será particularmente adecuado para el procesamiento de datos complejos que requieren importantes recursos y / o tiempo. Para este tipo de carga de trabajo, los datos generalmente se almacenan de una manera específica, por ejemplo, el sistema de archivos distribuido de Hadoop (HDFS), el formato de almacenamiento de datos Apache Parquet. A menudo se procesa de forma distribuida, es decir, utilizando las diferentes computadoras en red y una construcción de programación para administrar esta distribución (con programas escritos en lenguajes como Scala, Pyspark o SparkR, apoyándose en una capa de software como Spark). Los resultados suelen estar disponibles a través de interfaces fáciles de usar, como Apache Hive o Impala, para resumir, analizar y consultar grandes sistemas de datos en el ecosistema Hadoop.

En cuanto a la carga de trabajo interactiva, la arquitectura de big data es particularmente adecuada para casos de uso que requieren trabajo exploratorio o de investigación. En general, el usuario final no tiene en mente un modelo completamente predefinido: es a través del trabajo de análisis de datos que surgen nuevas preguntas, lo que lleva a una mayor exploración de datos de forma interactiva y repetitiva. En este tipo de carga de trabajo, un requisito importante es que el tiempo necesario para procesar las solicitudes de los usuarios ("latencia de red" más tiempo de procesamiento) sea mínimo. Esto es clave para mantener a los usuarios completamente comprometidos. Otra característica importante es que, además de los altos requisitos en términos de tecnología de procesamiento y almacenamiento de datos (una consideración común para cada tipo de carga de trabajo de big data), existe la necesidad de interfaces de usuario adecuadas para respaldar el trabajo de análisis de los usuarios y su interacción exploratoria. con el sistema.

En este contexto, se ha dado mayor importancia al uso de herramientas de inteligencia empresarial (BI). Por lo general, se entiende que el concepto de BI abarca los métodos y técnicas impulsados ​​por la tecnología (por ejemplo, aplicaciones y prácticas de TI) movilizados para recopilar, administrar y analizar datos con el fin de informar las decisiones comerciales (IFC (2019b)). Las herramientas de BI pueden facilitar enormemente el proceso de descubrimiento interactivo de big data, utilizando varias funciones como:

* Capacidades de desglose, es decir, funciones que permiten a los usuarios pasar de la representación general de los datos a una vista más granular; esta función es particularmente útil para administrar bases de datos multidimensionales, o "cubos" hechos de varias "dimensiones" que los usuarios pueden consultar de manera flexible para producir una vista personalizada de los datos que les interesan, por ejemplo, mediante el uso de tecnología de procesamiento analítico en línea (OLAP)
* Capacidades de obtención de detalles, es decir, funciones que permiten el acceso a información paralela que puede ser relevante para los datos que se analizan.
* Paneles, es decir, informes con indicadores clave / análisis presentados como una tabla / interfaz gráfica de usuario actualizada a medida que ingresan los datos
* Interfaces que permiten a los usuarios consultar y recuperar datos fácilmente del sistema, como interfaces de consulta habilitadas para SQL

Con la transmisión, el usuario final recibe una presentación constantemente actualizada de los datos, en lugar de tener que descargar los datos y luego procesarlos / analizarlos. Por lo tanto, los flujos se componen básicamente de tablas infinitamente largas e interminables que permiten consultar los datos a medida que fluyen. La tecnología utilizada para este procesamiento se ha desarrollado rápidamente en los últimos años, con herramientas como las plataformas de software de procesamiento de flujo de código abierto Apache Kafka, Spark streaming, Flink, Storm, Samza y otros. Las centrales anks están cada vez más interesadas en este tipo de carga de trabajo para analizar datos en tiempo real, por ejemplo, para monitorear su red de TI interna para detectar ataques cibernéticos o para supervisar el funcionamiento de las transacciones financieras o los sistemas de pago cuando tienen la tarea de hacerlo (por ejemplo, para detectar actividades fraudulentas o riesgos de perturbaciones en el sistema financiero). Si bien la mayoría de los bancos centrales aún no han desarrollado capacidades de transmisión de datos operativos, se espera que esto se convierta en un requisito estándar en un futuro próximo.

La carga de trabajo operativa a menudo se implementa con tecnologías de transmisión para proporcionar información en tiempo real sobre las transacciones comerciales que se dividen en varios sistemas. La experiencia muestra que una tecnología como Kafka se utiliza a menudo para respaldar la transferencia de datos entre los diferentes sistemas operativos, en particular para proporcionar una vista completa, coherente y precisa de todos los datos procesados ​​(por ejemplo, datos por lotes, datos en línea). También es útil realizar un seguimiento de los diversos pasos involucrados en el análisis de los datos, desde su origen raíz hasta el producto final (es decir, "linaje de datos"). Otro sistema importante que soporta cargas de trabajo operativas en las que el tiempo es crítico es el procesamiento de transacciones en línea (OLTP), que administra aplicaciones que modifican la base de datos (es decir, realizan "transacciones de base de datos"). Por lo general, OLTP facilitará la respuesta inmediata a las solicitudes de los usuarios, a excepción de las consultas más complejas que prefieren tratarse con la tecnología OLAP (ver más arriba).

La lista anterior es, por supuesto, una presentación esquemática. De hecho, a menudo existe una combinación de estos diferentes tipos de carga de trabajo. Una tecnología importante en este contexto es la arquitectura Lambda,[[14]](#footnote-14) que se puede utilizar para manejar cantidades masivas de datos que se pueden procesar bajo diferentes tipos de cargas de trabajo, en particular a través de métodos de transmisión por lotes y en tiempo real.

**4.3 Arquitectura para almacenar, procesar y consultar big data**

Los diversos proyectos de macrodatos implementados por los bancos centrales muestran que no existe una arquitectura de datos única para todos que pueda satisfacer todos los tipos de necesidades en términos de carga de trabajo, capacidad para calcular operaciones al mismo tiempo en lugar de secuencialmente ("concurrencia"), requisitos para almacenamiento, acceso y procesamiento de datos y latencia. Esto requiere implementar una “arquitectura políglota”, capaz de abordar requisitos heterogéneos. Por ejemplo, la plataforma de big data del Banco de Italia cubre múltiples requisitos de los usuarios, por ejemplo, ingesta, almacenamiento y procesamiento de datos; uso de ciencia de datos, aprendizaje automático y aplicaciones de BI, que requieren el uso combinado de múltiples herramientas de TI (Cariello y Quarta (2019)). Por tanto, la naturaleza específica de estos requisitos impulsa inherentemente el desarrollo de la arquitectura de soporte, a menudo de forma progresiva e interactiva.

En la práctica, la arquitectura del entorno de big data implementada por la mayoría de las organizaciones tiene como objetivo configurar un almacén de datos (o "almacén de datos empresarial" (EDW)), que es un repositorio central donde los datos de fuentes dispares se almacenan, analizan y consultan. Este almacén comprende observaciones tanto actuales como históricas, con los datos obtenidos (o "informados") de un almacén de datos operativos (ODS). Este “lago de datos” comprende toda la información posible necesaria para alimentar el almacén (Lacroix (2019)). Cubre todos los datos independientemente de su formato, su inclusión real (o no) en la EDW, los casos de uso que admiten y su grado de procesamiento, con usuarios capaces de acceder a los datos originales “sin procesar” y no solo a los “transformados”. Por ejemplo, la plataforma de big data (DISC) del BCE comprende un almacén de datos central que incluye todos los datos de interés para la institución y un diccionario de datos único que respalda la combinación de los distintos conjuntos de datos (Sánchez y Trzeciok (2019)).

**4.4 Desarrollo de software de big data**

En términos generales, el software de big data suele ser ensamblado e implementado manualmente por ingenieros de TI especializados. Sin embargo, las implementaciones manuales y la administración relacionada de las (a menudo varias) herramientas de tecnología de big data plantean riesgos significativos, incluida la falta de pruebas automatizadas, configuraciones incorrectas y posiblemente cortes graves del sistema. Es evidente que se necesita un enfoque más estructurado, y las prácticas DevOps[[15]](#footnote-15) “tradicionales” que prevalecen claramente en otros dominios también deberían aplicarse idealmente para respaldar una implementación automatizada de la plataforma de big data. Para hacerlo, se requiere un enfoque de infraestructura como código (IaaC), utilizando tecnologías como Puppet o Ansible que pueden respaldar la automatización de la implementación de infraestructura de TI altamente distribuida. Además, se deben aplicar prácticas adecuadas de gestión del cambio para tratar con entornos separados para el desarrollo, la integración de sistemas, la aceptación del usuario y los usos de producción de una manera integral. Los bancos centrales también pueden considerar extender DevOps a las prácticas de DataOps, con miras a implementar la automatización para la gestión de datos y el procesamiento analítico.

Sin embargo, la experiencia muestra que el uso de prácticas de DevOps para desarrollar software de big data aún se encuentra en una fase de adopción temprana. Sin duda, esta es una tarea compleja debido a las muchas partes tecnológicas en movimiento involucradas en las plataformas de big data / HPC. Una capa adicional de complejidad proviene de la naturaleza distribuida del software, ya que involucra múltiples máquinas y diferentes opciones de configuración. Pero una consecuencia que a menudo se observa es que muchos proyectos no pasan de la fase piloto debido a la falta de gobernanza de TI, incluida la gestión de cambios adecuada y los procesos DevOps.

No obstante, los bancos centrales están avanzando para abordar estos problemas. Las iniciativas exitosas muestran que, una vez que se configuran los procesos de implementación y pruebas automatizados adecuados para construir el clúster de big data, los equipos analíticos y de TI son recompensados ​​con agilidad, facilidad de reproducción, facilidad de actualizaciones y posibilidades de escalabilidad.

**5. Mirando hacia el futuro con una estrategia de información integral**

Con el crecimiento exponencial de los datos que recopilan los bancos centrales, surgen cada vez más necesidades analíticas que requieren el uso de análisis de big data y herramientas de alto rendimiento. Estas tecnologías están evolucionando rápidamente y pueden ofrecer un valor significativo para lidiar con grandes cantidades de datos y aplicar cálculos complejos de IA y ML. Además, mientras que el análisis de datos tradicional se limitó a datos estructurados, el uso de tecnologías de big data permite a los bancos centrales utilizar fuentes de datos alternativas para sus diversos casos de uso.

Al decidir sobre un proyecto para configurar una plataforma informática de big data / HPC, se deben tener en cuenta muchas dimensiones, incluidos los requisitos comerciales (casos de uso), las necesidades tecnológicas, la complejidad del sistema, las limitaciones de recursos (por ejemplo, costos financieros, competencias internas en términos de cantidad de personal y combinación de habilidades), rendimiento, confiabilidad, modelo operativo y seguridad; consulte Lambe et al (2019) para obtener una revisión de los factores clave que influyen en la arquitectura de la plataforma (IDEA) desarrollada por el BIS. Además, estos aspectos no deben considerarse de forma aislada, ya que deben abordarse de manera integral en el contexto de la estrategia de información de la organización. Básicamente, esta estrategia debe tener como objetivo proporcionar una hoja de ruta de alto nivel para la adopción de tecnologías en constante cambio para administrar los datos y responder a las necesidades de los usuarios; para la situación en el Banco de Inglaterra, consulte Vaughan y Willis (2019).

La experiencia de los bancos centrales con la implementación real de plataformas y tecnologías relacionadas con big data muestra que la atención debe centrarse en las siguientes consideraciones principales para garantizar el éxito:

* Objetivos comerciales a mediano y largo plazo: saber qué objetivo comercial lograr es de suma importancia al elegir la tecnología de big data adecuada.
* Crecimiento incremental: comenzar poco a poco, aprender del éxito temprano, monitorear periódicamente el progreso y revisar el plan son claves.
* El negocio y la TI trabajan juntos: los líderes empresariales y de TI deben trabajar juntos para promover una cultura basada en datos en toda la organización (por ejemplo, mediante “administradores de datos” reconocidos).
* Gobernanza de datos: un marco de gobernanza de datos es vital para ofrecer beneficios empresariales continuos con una plataforma de big data.

Por último, la cooperación internacional y el intercambio de conocimientos entre los bancos centrales pueden agregar un valor significativo, sobre todo considerando el panorama tecnológico heterogéneo y cambiante. El camino a seguir es establecer iniciativas periódicas para garantizar el intercambio de experiencias, la fertilización cruzada y la promoción de sinergias entre los diversos proyectos emprendidos por los bancos centrales y la comunidad pública en general.

**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-software2018.

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

1. Respectivamente, Jefe de Apoyo a la economía y las estadísticas del Banco de Italia (Giuseppe.Bruno@bancaditalia.it); Jefe de Ingeniería de Plataformas en Tecnología y Servicios de la Información en el Banco de Pagos Internacionales (BIS) (Hiren.Jani@bis.org); Jefe de TI, Departamento Monetario y Económico del BPI (Rafael.Schmidt@bis.org); y Jefe de la Secretaría del Comité Irving Fisher de Estadísticas del Banco Central (CFI) y Jefe de Apoyo a la Investigación y Estadísticas, BIS (Bruno.Tissot@bis.org).

   Agradecemos a Giulio Cornelli, Edward Lambe y Xavier Sosnovsky por sus útiles comentarios y sugerencias, y a Giulio Cornelli nuevamente por su excelente asistencia en la investigación. [↑](#footnote-ref-1)
2. Este informe se basa en las actas del taller sobre "Plataformas informáticas para big data y aprendizaje automático" organizado por el Banco de Italia con el apoyo del BIS y la CFI en Roma el 15 de enero de 2019. [↑](#footnote-ref-2)
3. “Semiestructurada o no estructurada” significa que dichas fuentes de datos no tienen un modelo de datos predefinido y, a menudo, no encajan en una base de datos relacional convencional. Cabe señalar que las plataformas de big data también tienen demanda para aquellos conjuntos de datos muy grandes que aún están bien estructurados, en particular para manejar las limitaciones que enfrentan en términos de almacenamiento / procesamiento. Este es particularmente el caso del tipo de "grandes conjuntos de datos financieros" con los que los bancos centrales se han enfrentado cada vez más desde la Gran Crisis Financiera de 2007-09 (Tissot (2019)) [↑](#footnote-ref-3)
4. En comparación con instituciones homólogas como las ONE, el interés de los bancos centrales hasta ahora ha sido bastante limitado por la gama más amplia de datos no estructurados que comprenden datos geoespaciales (por ejemplo, teléfonos móviles, imágenes de satélite), datos de redes inteligentes (por ejemplo, información derivada de los usuarios conectados a las redes eléctricas). ) y, de manera más general, todos los tipos de datos de sensores que resultan del registro de algún tipo de entrada del medio ambiente (por ejemplo, temperatura); ver Reunión del Grupo de Expertos en Clasificaciones Estadísticas Internacionales (2015) para una revisión general de la amplia gama de lo que puede constituir una fuente de macrodatos. Sin embargo, el interés de los bancos centrales también puede expandirse por este tipo de datos a medida que surjan nuevas tecnologías / ideas. [↑](#footnote-ref-4)
5. Datos sobre las ventas de bienes de consumo obtenidos mediante el escaneo de los códigos de barras de los productos individuales en los puntos de venta. Los datos pueden proporcionar información detallada sobre cantidades, características y valores de los bienes vendidos, así como sus precios. [↑](#footnote-ref-5)
6. La pandemia de Covid-19 de 2020 proporcionó otro ejemplo de la utilidad de estas fuentes "no estándar". Varias OSN pudieron compensar la incapacidad de medir los datos de precios en tiendas cerradas mediante la recopilación de precios en Internet y el uso de datos de escáner para ciertos segmentos de consumo para calcular los IPC (por ejemplo, ver en el caso de Francia la comunicación del INSEE en at [www.insee.fr/en/information/4471928](http://www.insee.fr/en/information/4471928)). [↑](#footnote-ref-6)
7. Un núcleo de procesador (o "núcleo") es una sola unidad de procesamiento. Las computadoras de hoy, o CPU (unidades centrales de procesamiento), tienen múltiples unidades de procesamiento, y cada uno de estos núcleos puede enfocarse en una tarea diferente. Dependiendo del problema analítico o estadístico en cuestión, los grupos de GPU (unidades de procesamiento de gráficos, que tienen una estructura muy paralela y fueron diseñadas inicialmente para un procesamiento de imágenes eficiente) también podrían integrarse en computadoras, por ejemplo, para admitir cálculos masivos. [↑](#footnote-ref-7)
8. Cabe señalar que la comunidad estadística internacional, incluidos los bancos centrales, ya participa activamente en el intercambio público de tales herramientas de TI en el contexto del estándar SDMX (Statistical Data and Metadata eXchange) para estructurar e intercambiar estadísticas (IFC (2016)). [↑](#footnote-ref-8)
9. Es decir, disponible de forma gratuita y distribuida con su código fuente disponible para su modificación. [↑](#footnote-ref-9)
10. Estos problemas se ven reforzados por la dificultad de identificar los productos que estarán presentes a más largo plazo en el momento de la decisión de inversión inicial. Esta incertidumbre está ilustrada por el llamado ciclo de la exageración (Gartner (2018)), que básicamente muestra que la adopción de incluso una tecnología exitosa sigue un ciclo de vida con diferentes etapas, desde la concepción inicial hasta las expectativas infladas, seguido de una creciente desilusión, antes la etapa de madurez caracterizada por una adopción generalizada. Siempre que sea posible, se recomienda evaluar la madurez y la perspectiva a largo plazo de una solución de código abierto antes de adoptarla. [↑](#footnote-ref-10)
11. SaaS, PaaS e IaaS son tres formas de describir cómo se puede utilizar la nube para organizar la plataforma. SaaS (software como servicio): el software está disponible a través de un tercero a través de Internet; PaaS (plataforma como servicio): herramientas de hardware y software disponibles en Internet; IaaS (infraestructura como servicio): pago por uso para toda la gama de servicios basados ​​en la nube, como almacenamiento, redes y virtualización. [↑](#footnote-ref-11)
12. Ver Fournier (2016) para la iniciativa relacionada desarrollada en el Banco de Francia para almacenar datos heterogéneos y también IFC (2019b). [↑](#footnote-ref-12)
13. Para una discusión de estos problemas de calidad para datos muy grandes y bien estructurados sobre operaciones de derivados recopilados por registros de operaciones, consulte IFC (2018). [↑](#footnote-ref-13)
14. La arquitectura Lambda se basa en servicios informáticos sin servidor que se ejecutan en respuesta a eventos y administran automáticamente el recurso informático subyacente. [↑](#footnote-ref-14)
15. DevOps es un conjunto de prácticas que combina el desarrollo de software (Dev) y las operaciones de tecnología de la información (Ops) cuyo objetivo es acortar el ciclo de vida de desarrollo del sistema y proporcionar una entrega continua con software de alta calidad. [↑](#footnote-ref-15)