IFC-Bank of Italy Workshop on "Data Science in Central Banking: Applications and tools"

14-17 February 2022

Modern computing platforms as key technology for central banks, financial supervisors, and regulators[1]

John Ashley and Jochen Papenbrock,
NVIDIA

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Modern Computing Platforms as Key Technology for Central Banks, Financial Supervisors, and Regulators

Dr. Jochen Papenbrock and Dr. John Ashley[1]

Abstract

A team from NVIDIA examined the leading and most challenging use cases for data science and the corresponding IT/software tools that central banks, financial supervisors, and financial regulators are currently discussing and implementing around the world. NVIDIA is a leading global provider of accelerated computing platforms for AI, Generative AI (GenAI), data processing, high-performance computing (HPC), and the industrial metaverse across multiple industries including financial services and public sector. Accelerated computing reduces time to results and cost, and is significantly more energy efficient, allowing existing data centers to be more sustainable.

There are two main observations that can be drawn from the team's analysis. First, central banks and supervisors are embracing Big Data, AI, GenAI, ML (machine learning) and accelerated simulation methods, and the open-source software ecosystem to support them. Second, cloud-native, full stack AI & simulation factories have started to be on the radar of central banks, financial supervisors, and regulators as key technology. However, the common workflows and tools used by many today exhibit computational bottlenecks due to CPU-only, non-accelerated computing. This creates severe roadblocks in scalability, productivity, agility, and time-to-value - unnecessarily reducing return on investment and driving unnecessary levels of cost and energy consumption. Also, the tools that are available for model optimization during training, inferencing and MLOps are rarely used.

Accelerated computing platforms using specialized hardware and software stacks based on massively parallel execution, integrated with industry standard software and open-source frameworks can help address these issues. GPUs – graphics processing units – are the most broadly available, supported, and performant foundations for such platforms. The accelerated hardware layers are just one building block of an accelerated computing platform. The other equally important building block is the complementary software stack with corresponding application frameworks. Those symbiotic hardware and software stacks are the basis for building entire 'factories' for AI and simulation to solve the most challenging problems, to predict the future, and thus are a highly important technology for central banks, financial supervisors, and regulators.

Such modern accelerated computing platforms can operate in any environment. Because GPUs are ubiquitous, these platforms can be deployed in public or private clouds, or in-house/on-prem data centers and colocation facilities. They can even be used in hybrid, cloud-native set-ups. They leverage popular open-source data science and engineering tools and Python packages. These characteristic of modern accelerated computing platforms are very important for broad acceptance in data science and IT departments and in global developer communities.

This paper discusses a wide range of central bank use cases and outlines how to implement these in an effective and efficient way utilizing accelerated computing platforms. These technologies also help implementing AI governance, AI model risk management and building trustworthy, transparent and explainable AI that will further increase the confidence of supervisors, regulated entities, and the public. These same

[1] Respectively, Head of Financial Technology EMEA and Lead Developer Relations Manager Banking Global, NVIDIA, Germany (jpapenbrock@nvidia.com); Chief Architect, AI Nations & Director, NVIDIA AI Technology Centers, NVIDIA Corp., USA (jashley@nvidia.com)

platforms support the technical side of the transition to building and supervising a more sustainable and resilient financial system, taking economic and climate-related change into account.

The paper includes somewhat more technical content to highlight the ready availability of GPU-accelerated tools and application frameworks that enable and accelerate a range of data science operations that are key solutions building blocks for many current and future problems in central banking and macroprudential analysis. This includes areas such as large graph analysis, graph neural networks, conversational/speech AI, deep learning, GenAI and foundation models, NLP and Large Languages Models (LLMs), geospatial AI, quantum computing simulation, simulated digital environments, digital twins, physics-informed AI/ML and climate intelligence. Leveraging full stack accelerated computing platforms delivers the scalability, productivity and time to insight central banks, supervisors, and regulators need to accelerate and enhance their mission, mandate and policies.

Keywords: central banks, data processing, explainable AI, trustworthy AI, RAPIDS, ESG, accelerated computing, climate intelligence, Generative AI, LLM, foundation model, financial supervision, financial regulation, RegTech, SupTech, AI computing

JEL classification: Q58; Q57; Q56; E58; G28; C31; C81

Contents

Accelerated Computing Platform for Enterprise Data Science and AI

A team from NVIDIA examined the use cases for data science and the IT/software tools that central banks are currently discussing and implementing around the world.

We drew on several sources such as IFC/BIS publications[2] and attended central bank conferences such as the "International Conference on Statistics for Sustainable Finance"[3] and "Data Science in Central Banking, Part 1: Machine Learning Applications" [4]. We held discussions and were involved in projects with some of the leading central banks. One project we would like to highlight that involved many European central banks and regulators is the EU Horizon2020 project FIN-TECH[5].

There are two main observations that can be drawn from these sources:

First, central banks and supervisors are embracing both Big Data and AI/ML (artificial intelligence and machine learning) methods and the open-source software ecosystem to support them. The establishment of a transformed supervisory model can be observed that leverages supervisory technology, sustainable finance technology and climate science to

- digest the vast volumes of structured and unstructured data,
- improve timeliness ("real time") of identification, monitoring, and early warning/intervention of (emerging) risks.
- support both the "big picture" and a granular view at different zoom levels

It was possible to identify typical workflows and tools that are very useful and that we have observed and discovered in many other industries dealing with Big Data and AI/ML.

A typical workflow involves 3 steps: data preparation, model training and visualization/explanation. Different types of databases are used, and the analysis sometimes involves graph/network analysis. Many open-source and Python tools are used.

---

[2] "Computing platforms for big data analytics and artificial intelligence"

"Big data and machine learning in central banking"

"Big data for central banks"

"The suptech generations"

"The use of big data analytics and artificial intelligence in central banking"

"Central Bank Communications: information extraction and semantic analysis"

[3] Conference link: https://www.banque-france.fr/en/international-conference-statistics-sustainable-finance; our contribution is titled "Accelerated Data Science, AI and GeoAI for Sustainable Finance in Central Banking and Supervision" and can be found in the second video at 7:13:10 – 7:29:10 [Q&A 8:10:33 – 8:14:50].

[4] https://www.bis.org/ifc/events/211019_ifc_bdi.htm

[5] This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 825215 (topic ICT-35-2018, Type of action: CSA). The content reflects only the author's vie and the Commission is not responsible for any use that may be made of the information it contains.

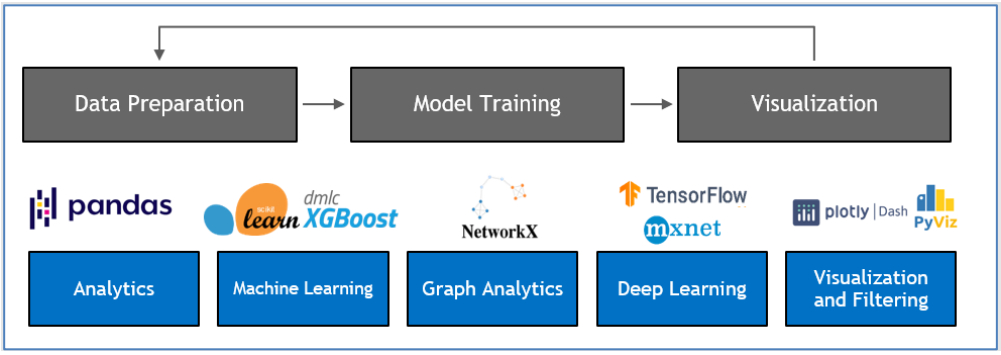


Figure 1: typical workflow and exemplary Python tools in many central bank projects utilizing big data and AI. Open-source projects and Python tools have democratized data science. However, the approach exhibits the typical computational bottlenecks due to CPU-only processing.

At this point in time, less focus is put on industrial tools to deploy the models built and scale them across the organisation in the inferencing (production) step.

Second, in all data science steps mentioned above there are computational bottlenecks with CPU-only processing that are made worse by forcing general purpose infrastructure to serve AI powered workflows.

This has negative impacts:

- inhibits developer productivity growth
- delays time to insight
- limits model scalability
- increases total cost of ownership and carbon emissions
- increases infrastructure complexity.

Also, software tools for model optimization during training and inferencing and for MLOps are rarely used.

In some important and growing areas - such as initiatives to greening the financial system and to develop climate stress tests and scenarios for banks and insurance companies - technologies such as AI, HPC, high-performance data analytics (HPDA) and scientific visualisation still appear rarely to be used.

The Challenge: Enabling the AI Transformation

It all starts with the key business drivers. A bank will be focused on improving fraud detection, enabling virtual assistants, and creating recommenders to produce next best actions. An insurance company will also want to automate claims processing and identify fraudulent claims. Regardless of the type of financial institution surveyed, they all experience traditional AI infrastructure challenges: AI is complex, it is hard to deliver scalability and reliability at the same time and needs to operate within the budget constraints of the bank.

At the same time, fraudsters are already using Generative AI to create "deepfake" calls to bankers to get the bank itself to steal money from customers on the fraudster's orders. New identities will be backstopped by AI generated histories. Money will be laundered using flows planned and eventually executed by AI to escape detection.

The transition to being an AI enabled bank cannot be leisurely. It is urgent that banks build a native capability for AI – leading the pack is not required for survival but being solidly in the pack is. Leading firms have months of head start with generative AI and in some cases years with pre-ChatGPT AI. No firm can afford to take a "wait and see"

position – ramping up too late, or even too slowly, will expose customers to massive risks and is arguably a failure of fiduciary duty.

The Solution: Accelerated Computing Platform and Enterprise AI Software Stack

Accelerated computing platforms using specialized hardware integrated with industry standard software can help address many of these issues and challenges. GPUs – graphics processing units – are the most universally available, supported, and performant foundations for such accelerated platforms.

According to the IFC-BIS publication "Computing platforms for big data analytics and artificial intelligence" (see Bruno et al. 2020), "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. [...] A processor core (or 'core') is a single processing unit. Today's computers – or CPUs – 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."

Today the superb computing power of GPU (Graphics Processing Unit) clusters are widely used in research where many of the most capable supercomputers are powered with GPUs. Industrial firms and other research organizations have long since adopted GPU computing to address high-performance computing requirements.

Here are some examples of GPU-accelerated computing:

- Cambridge-1 is the fastest supercomputer in the UK, boosting Covid-19 research.
- Large Language Models like GPT-3 with billions of parameters are usually trained on a GPU infrastructure like Selene, one of the fastest supercomputers, built and managed by NVIDIA. NVIDIA technologies power many systems on the Top500 and Green500 lists. [6]
- MLPerf[7] is a benchmark produced by a consortium of AI leaders from academia, research labs, and industry whose mission is to 'build fair and useful benchmarks' that provide unbiased evaluations of training and inference performance for hardware, software, and services—all conducted under prescribed conditions. To stay on the forefront of industry trends, MLPerf continues to evolve, holding new tests at regular intervals and adding new workloads that represent the state of the art in AI. Systems powered by NVIDIA technology deliver leading performance across all MLPerf tests for training, both per chip and at scale. In inferencing, NVIDIA accelerated systems continued to deliver exceptional performance across the full range of MLPerf tests.
- Meta/Facebook established a Research SuperCluster (RSC) for AI research based on GPUs.[8]

The accelerated hardware layers (GPU servers and high-speed networking) are just one building block of an accelerated computing platform. The other equally important building block is the optimized software stack. It helps to program the hardware, give access to numerous tasks in an accelerated way and to further optimize the compute performance.

Modern AI platforms must work in a similar manner across supercomputers and public cloud all the way to on-prem data centers and edge. They also need to leverage existing

[6] https://www.top500.org/

[7] https://mlcommons.org/en/

[8] https://ai.facebook.com/blog/ai-rsc/

open-source tools and Python packages. These characteristics of modern accelerated computing platforms are important for broad acceptance in data science and IT departments and in global developer communities.

Modern computing platforms support the execution of data science and high-performance computing (HPC) workloads as well as building/deploying AI models at scale. Those platforms are flexible, powerful stacks of hardware and software that are orchestrated to reduce performance bottle necks. These platforms provide access to accelerated data science and complex AI model building for a wide range of users, while addressing a larger number of use cases. Mass calculations and accelerations based on GPUs are crucial as they directly translate into several benefits:

- Accelerated training allows building larger and more accurate models
- Accelerated inferencing allows real-time utilization of trained models, e.g., for detecting fraud and cyber attacks
- Accelerated generation of synthetic data (e.g., Generative Adversarial Networks - GANs) amplifies existing data at large scale and simplifies collaborations of central banks with SupTech and RegTech startups
- Accelerated simulation enables more realistic and complex simulation
- Accelerated computing allows data scientists to better assess, validate, audit, and explain AI models
- People with strong data science skills are a limited and expensive resource globally. Accelerated computing brings leverage to that resource, this reduces time-to-insight and improves productivity.

Another major advantage of modern computing platforms is their leveraging of open-source software. This leverages the innovation potential from global developer and data science communities. Here are some examples:

- Open-source packages like TensorFlow and PyTorch are the de-facto standard frameworks for building and tuning large language models like BERT, ChatGPT, and LLAMA2. GPU support is a standard part of these frameworks.
- There are open-source Python-based projects for enabling end-to-end data science and analytics pipelines entirely on GPUs. This aims to accelerate the entire data science pipeline including data loading, ETL, model training, and inference. In-memory analytics help to scale up and out with accelerated data science. This enables more productive, interactive, and exploratory workflows.
- Open-source systems for automating deployment, scaling, and management of containerized applications as well as virtualization technologies enable business continuity and workload balancing, resource sharing and improved utilization of the existing resources in modern computing platforms.

The Economics and Benefits of Accelerated Computing

GPU accelerated computing moves compute intense and embarrassingly parallel parts of the application to the GPU. This enables dramatic performance improvements, with process that used to take days taking hours, or those that took hours to minutes.

GPU accelerated computing has two main benefits. By accelerating the compute, firms can choose what to optimize for: time to results, more compute intensive (higher fidelity) models, exploring more of the solution space, processing more data, or using a smaller system for the same workload. Because accelerated computing is also more energy efficient, the same results can be had faster and with lower energy consumption, which depending on the energy source, may also mean reduced CO2 emissions.

In addition to the savings in computer systems and energy, improved productivity of data scientists and engineers can have a significant impact on the top and bottom lines.

These performance improvements and energy savings are reflected in industry standard benchmarks like MLPerf and the Green 500 list.

For DL workloads, GPU-based computing platforms have set records for the MLPerf[9] benchmark (an industry-standard set of benchmarks across a variety of AI modelling tasks), handily surpassing all other commercially available systems (see Mattson et al. (2020)). Comparable results are documented with respect to the STAC-A2™ Benchmark suite[10] which is the industry standard for testing technology stacks used for compute-intensive analytic workloads involved in pricing and risk management.

GPU-accelerated Deep Learning (DL) frameworks[11] offer building blocks for designing, training, and validating deep neural networks through a high-level programming interface. Widely used DL frameworks, such as MXNet, PyTorch, TensorFlow, and others rely on GPU-accelerated libraries to deliver high performance, multi-GPU accelerated training. These optimized DL framework containers are performance-tuned for GPUs. This eliminates the need to manage packages and dependencies or build DL frameworks from source. Containerized DL frameworks, with all dependencies included, can be used to develop common applications, such as conversational AI, natural language understanding (NLU), recommenders, and computer vision.

GPU-acceleration translates into faster training and can be used for scaling hyper-parameter optimization and simply training a larger variety of models and framework which is a way to find even better models. Also, GPU acceleration can be used to produce synthetic data to enhance the real data set which can amplify and improve training results.

Enterprise Platform for Building Accelerated Production AI

Accelerated Computing is a key element of Enterprise AI which is a concept including an end-to-end, cloud-native suite of AI and data analytics software that's optimized to enable any organization to use AI and benefits from accelerated computing. It can be deployed from the enterprise data center to the public cloud. It has development tools and frameworks for the AI practitioner and reliable management, orchestration, and virtualization layers for the IT professional to ensure performance, high availability, and security[12].

An accelerated production AI environment is enabled by an enterprise AI system which is characterized by the following capabilities:

- Supporting workflows for Data Science, AI, GenAI/LLM
- SDKs, application frameworks and pre-trained models
- deployable anywhere – from on-prem to cloud
- cloud-native and supporting hybrid set-ups
- supporting open-source software (OSS) frameworks and reducing development complexity
- secure and scalable environments and infrastructure
- certifications and service levels

[9] https://mlcommons.org/en/

[10] https://stacresearch.com/

[11] https://developer.nvidia.com/deep-learning-frameworks

[12] https://www.nvidia.com/en-us/data-center/products/ai-enterprise/

Such an enterprise AI stack is the foundation for delivering secure, trustworthy, and scalable AI. It helps addressing the following solution requirements:

- **Acceleration**
    - Faster time to value, high perf and low cost, massive scale
    - putting AI workloads on virtualized solutions preserves the performance while adding the benefits of virtualization, such as ease of management and enterprise-grade security
- **Choice/Flexibility/Sovereignty**
    - Sovereignty of data, operations, and software
    - Range of AI software ecosystem (AI Enterprise, OSS, ISV), vendor-agnostic
    - cloud-native operating model to address multi cloud, hybrid cloud, on-prem environments
    - optimizing your workload placement strategy with cloud-native architecture
    - 'Mobility of compute' to address data gravity (and privacy)
    - balance workload placement based on cost and performance
- **Customization**
    - to improve accuracy, leverage enterprise data, increase trustworthiness (e.g. reduce hallucination and Factual but out-of-context answers in Generative AI)
- **Operations**
    - Easy to deploy, boosts developer productivity, simple to operate and future proof hybrid cloud strategy
    - AI lifecycle management
- **Performance**
    - Satisfying scaling infrastructure demands regarding data and compute (data prep, queries, testing, real-time inferencing)
- **Privacy & Sovereignty**
    - Enterprise data and IP are private and critical - this data needs to be controlled and protected to prevent leakage outside the organizational boundary.
    - securely run your private corporate data to do fine-tuning, run inferencing, and, in some cases, even training in-house.
- **Safety and Security**
    - Intrinsic security at every layer of the stack
    - integrated security and management
    - Guardrails are topical, and for safety and security
- **Cost and Sustainability**
    - easy to deploy, boosts developer productivity, Simple to operate and future proof hybrid cloud strategy
- **Compliance and AI Governance**
    - compliance needs that enterprise solutions, including AI, must meet. Access control and audit readiness

The entire enterprise AI stack has the following components:

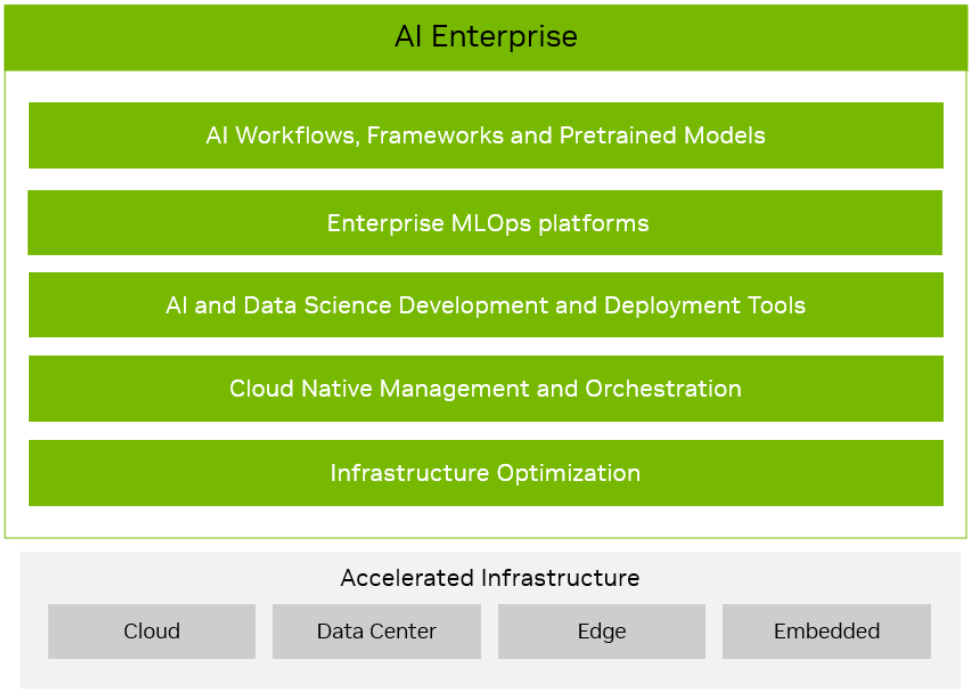


Figure 2: Components of an entire enterprise AI stack.

The technological basis is the optimized, accelerated infrastructure plus cloud-native management and orchestration layers. On top of this layer are tools for development and deployment of AI and Data Science. An enterprise MLOps platform can be added. This end-to-end stack can be used in any environment and established workflows based on this platform are portable across public and private clouds, can be used in a multi-cloud environment and in a hybrid way (like cloud plus on-prem).

An example of such a cloud native stack can be found in a a repository of tools, models, and sample workflows called NGC (NVIDIA GPU Cloud)[13]. This repository also contains GPU-activated tools for model optimization and model inference serving. It is a cloud native repository for many of the previously discussed SDKs, applications frameworks, and deployment tools. The content simplifies building, customizing and the integration of GPU-optimized software into workflows, accelerating the time to solutions for users.

- **Containers** package software applications, libraries, dependencies, and run time compilers in a self-contained environment so they can be easily deployed across various compute environments. They enable software portability.
- **Models and Resources:** the NGC Catalog offers pre-trained models for a wide range of common AI tasks. The pre-trained models can be used for inference or fine-tuned with transfer learning, saving data scientists and developers' valuable time. Resources provide reference neural network architectures across all domains and popular frameworks with the state-of-the-art accuracy to enable reproducibility as well as documentation and code samples which make it easy to get started with deep learning.
- **Helm Charts:** Kubernetes is a container orchestrator that facilitates the deployment and management of containerized applications and microservices. A Helm chart is a package manager that allows DevOps to configure, deploy and update applications across Kubernetes environments more easily. The

[13] https://www.nvidia.com/en-us/gpu-cloud/

NGC Catalog provides Helm charts for the deployment of GPU-optimized applications and SDKs.

- **Software Development Kits:** SDKs deliver all the tooling users need to build and deploy AI applications across domains such as recommendation systems, conversational AI or video analytics. They include annotation tools for data labelling, pre-trained models for customization with transfer learning and SDKs that enable deployment across the cloud, the data center, or the edge for low-latency inference.

End-to-end accelerated Data Science and AI

Data science workflows have traditionally been slow and cumbersome, relying on CPUs to load, filter, and manipulate data and train and deploy models. GPUs reduce infrastructure costs and provide superior performance for end-to-end data science workflows using RAPIDS[14] open-source software libraries. GPU-accelerated data science and AI workloads is available regardless of the location where GPUs are deployed, whether in the laptop, the workstation, in the data center, at the edge or in the public cloud.

The NGC repository has containers for RAPIDS which is used for GPU Open Data Science. It is a data science framework that is designed to have a familiar look and feel for data scientists working in Python. It also relies on and connects to many more open-source projects like Apache Arrow. RAPIDS is a suite of open-source software libraries and APIs for executing data science pipelines entirely on GPUs—and can reduce training times from days to minutes, also reducing energy consumption. RAPIDS unites years of development in graphics, machine learning, deep learning, HPC. It can run entire data science workflows with high-speed GPU compute and parallelize data loading, data manipulation, and machine learning for much faster end-to-end data science pipelines.

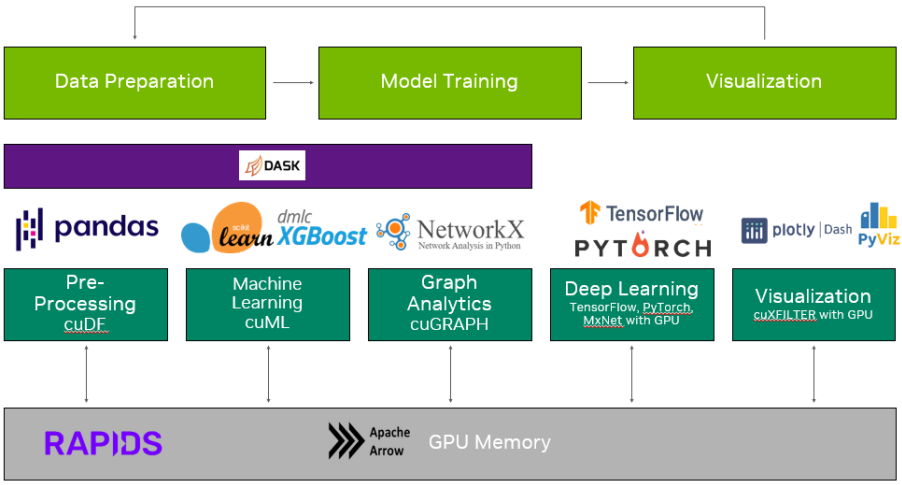


Figure 3: The same workflow as in Figure 1 but with GPU acceleration activated for the entire end-to-end data science workflow including AI, ML, ETL/pre-processing, network/graph analysis and filtering/visualization. There is a minimal change of the involved Python code using RAPIDS components like cuDF, cuML, cuGraph, cuXFilter as well as GPU-accelerated DL Python libraries like PyTorch and TensorFlow. Processing times due to GPU-acceleration can reduce training from days to minutes and there are also MLOPs-related tools to enable real-time inferencing.

RAPIDS can execute end-to-end data science and analytics pipelines entirely on GPUs. Integration into existing workflows normally requires only a few lines of code because

[14] https://rapids.ai/

its API was deliberately designed to be consistent with existing data science utilities (e.g., Pandas DataFrame, SciKit Learn). RAPIDS is incubated based on extensive hardware and data science experience from many contributors. It exposes GPU parallelism and high-bandwidth memory speed through user-friendly Python interfaces. RAPIDS focuses on common data preparation tasks for analytics and data science. This includes a familiar data frame that integrates with a variety of machine learning algorithms for end-to-end pipeline accelerations without paying typical serialization costs. It also includes support for multi-node, multi-GPU deployments, enabling vastly accelerated processing and training on much larger dataset sizes.

These libraries democratize the power of GPU accelerated data science with observed accelerations from CPU to GPU that can range from a factor of 10x to 1000x in some cases.

Parallel Data Processing with Spark

Given the parallel nature of many data processing tasks, the massively parallel architecture of a GPU is also able to parallelize and accelerate Apache Spark data processing queries, in the same way that a GPU accelerates deep learning (DL) in artificial intelligence (AI). There has been implemented a GPU acceleration through the release of Spark 3.0 and the open-source RAPIDS Accelerator for Spark[15]. The RAPIDS Accelerator for Apache Spark uses GPUs to:

- Accelerate end-to-end data preparation and model training on the same Spark cluster.
- Accelerate Spark SQL and DataFrame operations without requiring any code changes.
- Accelerate data transfer performance across nodes (Spark shuffles).

As ML and DL are increasingly applied to larger datasets, Spark has become a commonly used vehicle for the data pre-processing and feature engineering needed to prepare raw input data for the learning phase. The Apache Spark community has been focused on bringing both phases of this end-to-end pipeline together, so that data scientists can work with a single Spark cluster and avoid the performance penalty of moving data between Spark based systems for data preparation and PyTorch or TensorFlow based systems for Deep Learning. Apache Spark 3.0 represents a key milestone, as Spark can now schedule GPU-accelerated ML and DL applications on Spark clusters with GPUs, removing bottlenecks, increasing performance, and simplifying clusters. In Apache Spark 3.0 there is is now a single pipeline, from data ingest to data preparation to model training on a GPU powered cluster.

[15] Further reading:

- https://nvidia.github.io/spark-rapids/
- https://www.nvidia.com/en-us/deep-learning-ai/solutions/data-science/apache-spark-3/
- https://developer.nvidia.com/blog/gpus-for-etl-run-faster-less-costly-workloads-with-nvidia-rapids-accelerator-for-apache-spark-and-databricks/
- https://developer.nvidia.com/blog/accelerated-data-analytics-machine-learning-with-gpu-accelerated-pandas-and-scikit-learn/

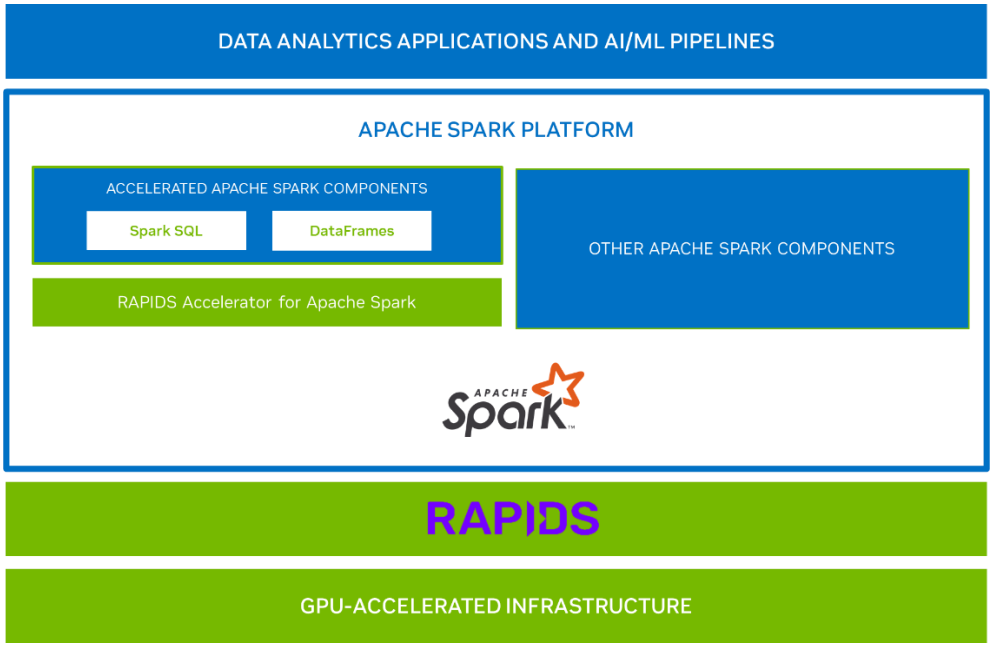


Figure 4: Overview of the Spark architecture leveraging the RAPIDS accelerator for data analytics applications and AI/ML pipelines.

Trustworthy, Explainable AI models

Another role of accelerated computing platforms in creating trustworthy, explainable AI models and in implementing automated AI model audit and evaluation. This goes far beyond MLOps which is a set of best practices for institutions to run AI successfully. Recent regulatory efforts like the draft AI Act released by the European Commission require transparent and trustworthy AI for certain more consumer-facing applications but central banks might want to adopt these principles, too.

There are a number of tasks to be addressed in AI assurance like post-hoc explainability layers (e.g. SHAP values based on cooperative game theory), testing for unwanted bias and adverse model reactions, visualizing large amounts of data and model outcomes, repeatable audits, impact assessments, unit tests, model drift detection, as well as techniques to support more human-centric and data-centric model building. High-performance computing is key in addressing all these dimensions of AI Assurance. As more AI/ML models are deployed around the globe it becomes clearer that the community has strong needs for tools and approaches that help to maintain trust and transparency in the models used. Researchers, developers, engineers, and architects are currently developing approaches and MLOps tools that support the implementation of AI quality, governance, trustworthiness, explainability and other approaches to ensure compliance with upcoming AI regulation and to assure the quality of AI, especially in high-risk AI applications. Modern computing platforms play a key role here as they can support building high quality of AI models and to test, benchmark, validate and certify them on an ongoing basis. Those platforms support internal teams for model risk management and model validation but also the entire TIC (test, inspect, certify) industry and the growing ecosystem of ethical AI startups.

RAPIDS also supports explainability of ML models. Model Interpretability aids developers and other stakeholders to understand model characteristics and the underlying reasons for the decisions, thus making the process more transparent. Being able to interpret models can help data scientists explain the reasons for decisions made

by their models, adding value and trust to the model. There are six main reasons that justify the need for model interoperability in machine learning[16]:

- Understanding fairness issues in the model
- Precise understanding of the objectives
- Creating robust models
- Debugging models
- Explaining outcomes
- Enabling auditing

RAPIDS provides GPU-accelerated model explainability through Kernel Explainer and Permutation Explainer. Kernel SHAP is the most versatile and commonly used black box explainer of SHAP. It uses weighted linear regression to estimate the SHAP values, making it a computationally efficient method to approximate the values.

Using a specialized tree based SHAP, a single GPU can provide explanations 20x faster than a 40-core CPU node for moderate-sized tree models, with even further acceleration possible for explanations of feature interactions.

An example in a bank could be to accelerate explainable AI implementation in managing risk in a loan book or credit portfolio. Credit risk management is a critical task for financial institutions, as it involves assessing the likelihood of borrowers defaulting on their loans. Traditional methods of credit risk management rely on statistical models and machine learning algorithms, which can be time-consuming and resource intensive. AI can be used to improve credit risk management by analyzing large amounts of data, identifying patterns, and making predictions about borrower behaviour. GPUs can accelerate data processing and AI workloads, making it possible to train models faster and more efficiently. Trustworthiness in those AI systems is important and can achieved through techniques such as data validation, model interpretability, and explainability. GPUs can support this implementation e.g., by GPU-accelerated SHAP computations. The entire workflow from data preparation, model training, model inferencing and model explanation is accelerated by GPUs so real-live data sets for real-life portfolio sizes can be processed in minutes or a few hours, which is crucial in production.[17] Visual dashboards and ad-hoc analytics tools for large data sets support the implementation of trust into the data science and AI workflow.[18]

---

[16] https://developer.nvidia.com/blog/model-interpretability-using-rapids-implementation-of-shap-on-microsoft-azure/

[17] The entire use case is presented here including code release: https://developer.nvidia.com/blog/accelerating-trustworthy-ai-for-credit-risk-management/

[18] See example dashboard here: https://dash-demo.plotly.host/nvidia-xai/practical-XAI/loan-default-datase

Figure 5: Interactive Plotly dashboard with focus in explainable and trustworthy AI for large scale data sets and models.

Graph Analytics and Graph Neural Networks

Learning from graph and relational data plays a major role in many applications. In the last few years, Graph Neural Networks (GNNs) have emerged as a promising new machine learning framework, capable of bringing the power of deep representation learning to graph and relational data. This ever-growing body of research has shown that GNNs achieve state-of-the-art performance for problems such as link prediction, fraud detection, target-ligand binding activity prediction, knowledge-graph completion, and product recommendations. The Deep Graph Library (DGL) is a DL library with efficient implementations for GNNs and demonstrate the speedup for GNN training and inference on GPUs. GPU-accelerated DGL containers will enable developers to work more efficiently in an integrated, GPU-accelerated environment that combines DGL and PyTorch.

cuGraph is paving the way in the graph world with multi-GPU graph analytics, allowing users to scale graphs with billions of nodes and edges. Accelerated algorithms for many common graph analytics tasks exist, across areas like centrality, community, link analysis, link prediction, and other traversal methods.

An example of leveraging this technology is the optimization of fraud detection based on financial data.[19] Another example is detecting fraud with generative models, network analysis and synthetic data.[20]

[19] A related blog can be found here: https://developer.nvidia.com/blog/optimizing-fraud-detection-in-financial-services-with-graph-neural-networks-and-nvidia-gpus/

[20] A related blog can be found here: https://developer.nvidia.com/blog/detecting-financial-fraud-using-gans-at-swedbank-with-hopsworks-and-gpus/

Generative AI and Large Language Models in Central Banks

The acceleration of deep learning ignited the big bang of AI. ChatGPT, a large language model powered by a DGX AI supercomputer[21], reached 100 million users in just two months. Its magical capabilities have captured the world's imagination. Generative AI is a new computing platform, like the PC, internet, and mobile cloud. Accelerated computing and AI have fully arrived.

The past months of developments in GenAI/LLM have dramatically pushed the boundaries in this area towards and artificial general intelligence. These models are increasingly complex and trained on an increasingly large text corpus. There is empirical evidence for Power Law scaling in multibillion parameter models.[22]

These kinds of models can create anything with a text structure, and not just human language text. It can also automatically generate text summarizations and has potential to automate tasks that require language understanding and technical sophistication. Examples show that it can interpret complex documents, launch actions, create alerts, or generate code.

And those models are not only about natural language but also other languages, images, video – even in a multi-modal way.

With massive size comes massive generalization ability: those models are competitive in many benchmarks without even tuning on the target task and the model still scales smoothly in performance instead of plateauing, implying that still-larger models would perform even better.

The generalization capabilities have a very meaningful impact and how we produce models and on the cost of model production. On can use a pre-trained large model and adapt it to new domains and task with the help of a few shots and prompts. The amount of labelled training data can thus be reduced, and this is very beneficial because many labelled data sets are expensive and can also contain labelling errors.

What does that mean Central Banks and Financial Supervisors? Generative AI (GenAI) and Large Language Model (LLM) systems will certainly be used to create better applications, like digital assistants and intelligent chatbots but the true power will come from its ability to ingest a wide variety of unstructured data and then to synthesize answers to natural language queries. GenAI is a smart language-powered interface to complex data, and other analytical and AI capabilities. Theoretically, every employee and staff member can become a researcher, knowledge worker or coder. Companies, organizations, and institutions will build AI factories for intelligence production, leveraging GenAI/LLM frameworks to train, customize, validate and deploy such models.

Here are some examples how Central Banks and Financial Supervisors can leverage those technologies and models, like building digital assistants including powering Avatars[23], enterprise search, summarization, translation, and report generation:

- predicting inflation and analyzing sentiment, using alternative data like text, images, videos, etc.
- document management: summarization and report generation will optimize middle/back office workflows

---

[21] https://www.nvidia.com/en-us/data-center/dgx-platform/

[22] Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017). Deep Learning Scaling is Predictable, Empirically. arXiv preprint arXiv:1712.00409.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling Laws for Neural Language Models. arXiv preprint arXiv:2001.08361.

[23] Like https://developer.nvidia.com/omniverse/ace

- research and surveillance activities in/of financial systems, markets and institutions based on huge amounts of unstructured data like text
- search and Q&A: optimized information retrieval by evaluating multiple sources, and summarizing results
- Transaction fraud: improves accuracy and generates reports, reducing investigations and compliance risk
- analysing sustainability and climate risk of the financial system but also the Central Banks' own initiatives like greening the financial system
- summarizing news feeds and market sentiment
- understanding the impact of the Central Banks' own activities, programs, projects and policies

Due to the disruptive, innovative, and transformative nature of GenAI/LLM, some of the following topics will be discuss, or are already being discussed in the communities of Central Banks and Financial Supervisors:

- Which use cases in FSI will be addressed and at which impact? How will this change financial service, financial markets, and financial centers? How will different segments of FSI, like banking, insurance, investments adopt?
- How will sustainable finance, ESG, climate risk, physical risk and biodiversity loss be addressed using GenAI and foundation models?
- Which other foundation models will we see in FSI besides natural language, like foundation models built on payments data or geospatial data?
- How will LLMs and GenAI be implemented, customized and how will adoption, flexibility, agility, and model performance be increased and how will cost, efficiency, and latency be decreased?
- Will we see more models in the cloud or rather in hybrid, or on-prem systems? What role will customization play? How will the optimal infrastructure, platforms, stacks, data processing technology and MLOps platforms look like?
- How will Trustworthiness, Security, Safety, Guardrails, Explainability and AI Governance be addressed? How will regulatory sandboxes look like? How will automated GenAI assessments and certification processes be established?
- How will large models be supervised and monitored? How will financial supervisors, regulators and central banks leverage those technologies for themselves and how will they monitor/supervise FSI activities?
- How will Avatars and Digital Assistants look like that are powered by LLMs?
- How will talents and startups develop and how will job profile change?

Challenges Of Developing and Deploying GenAI in Central Banks

There are some specific requirements for leveraging GenAI and LLMs in central banks like:

1. leveraging data that can only reside on-prem due to data privacy reasons
2. need for customization to better fit the the specific task, leverage proprietary data and for application trustworthiness.
3. safety and security reasons

And there are some challenges using foundation models without further customization:

Lack of Domain or Enterprise-Specific Knowledge:

1. Foundation models are trained on general datasets and may not possess the specialized knowledge required for specific domains or enterprises.
2. This limitation restricts their ability to provide accurate and relevant information in context-specific scenarios.
3. Without domain-specific knowledge, the models may struggle to understand and generate content that aligns with specific industry jargon, terminology, or practices.

Limited Adaptability to Changing Requirements:

1. Foundation models are static and do not have the inherent capability to adapt and evolve with evolving requirements.
2. As new trends, technologies, or business needs emerge, the models may become outdated and fail to provide up-to-date information.
3. Without the ability to continuously learn and integrate new knowledge, the models may lose their effectiveness over time.

Generation of Inaccurate or Undesired Information:

1. Foundation models, when used as is, can occasionally generate content that is inaccurate, misleading, or irrelevant to the user's needs.
2. This phenomenon, known as hallucination, occurs when the models generate information that appears coherent but lacks factual accuracy.
3. The generation of undesired or irrelevant information hampers the reliability and usefulness of the models for specific tasks or applications.

Risk of Bias and Toxic Information:

1. Foundation models may inadvertently reflect biases present in the training data, leading to biased outputs.
2. This bias can manifest in various forms, including gender, racial, or cultural biases, which can perpetuate unfair or discriminatory information generation.
3. Additionally, the models may inadvertently generate toxic or harmful content, such as hate speech or misinformation, which can have negative consequences if not addressed.

These challenges highlight the need for customization techniques to address the limitations of foundation models. Users can overcome these challenges and create more reliable, accurate, and tailored large language models that are specific to their domains or enterprises using specific customization methods.

Beyond the aspects of customizing foundation models there are other challenges of building and using them:

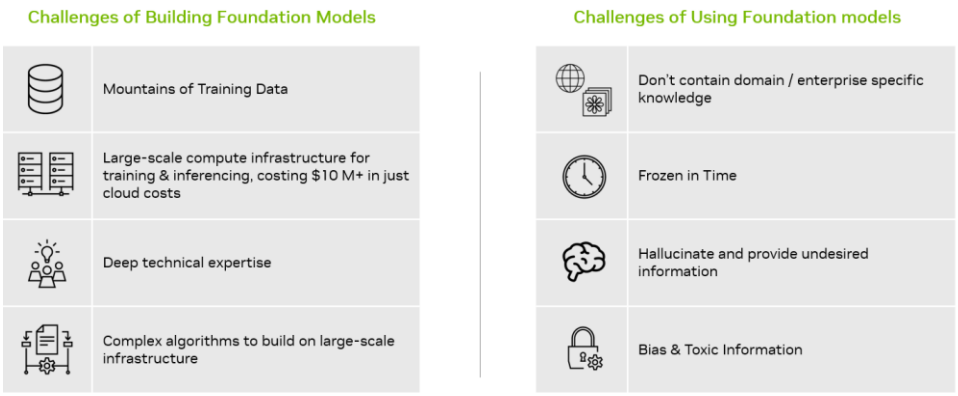


Figure 6: Challenges of building and using foundation models.

The solution is a flexible framework and a flexible, accelerated compute infrastructure plus some specific techniques and tools, e.g., to guardrail the models. [24]

Training such large models is an engineering challenge that has been solved by frameworks for efficiently training the world's largest transformer-based language models, based on Pytorch. The framework is for building, training, and fine-tuning GPU-accelerated speech, and natural language understanding (NLU) models with a simple Python interface and can be used to build models for real-time automated speech recognition (ASR), natural language processing (NLP), and text-to-speech (TTS) applications such as video call transcriptions, intelligent video assistants, and automated call center support.

Data curation tools and several engineering approaches to accelerated training like tensor and pipeline parallelism, sequence parallelism and selective activation recomputation are involved:

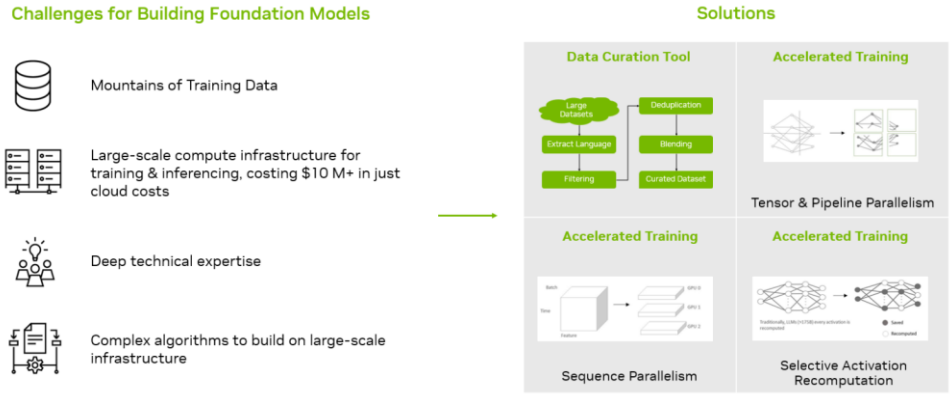


Figure 7: Solutions to meet the challenges of building foundation models.

Several customization techniques help to overcome the challenges using foundation models:

[24] https://www.nvidia.com/en-us/ai-data-science/generative-ai/

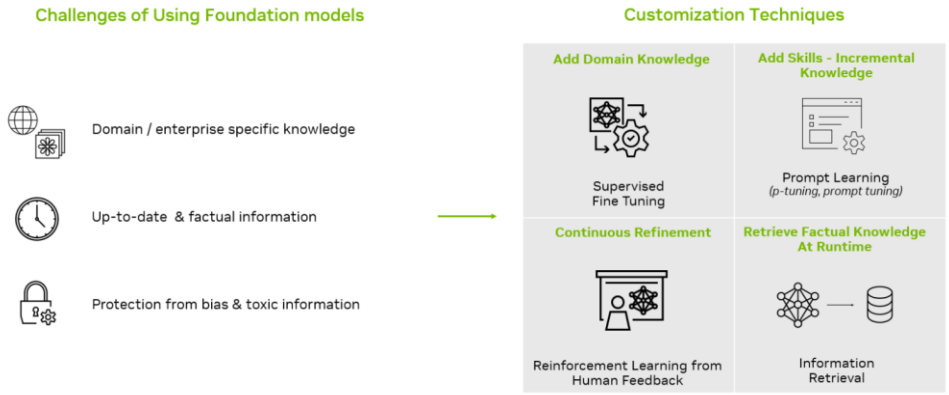


Figure 8: Solutions to meet the challenges of using foundation models.

Increasing the value of generative AI and foundation models in specific business use cases, institutions will increasingly customize pretrained models by fine-tuning them with their own data—unlocking new performance frontiers.

Extensive customization means building models from scratch or performing extensive fine-tuning.

There are a diverse range of customization techniques for generative AI models. These methods range from zero modification to the model's weights, all the way to substantial fine-tuning of every single parameter.

- Prompting: The simplest method of customization is by prompting, where no weights are changed within the model. This is a process of finding the right prompt to produce the desired output. As the original model's weights are not altered, the effectiveness of this method relies heavily on the ability to craft an appropriate prompt. It is somewhat of an art and a science, which involves understanding the tendencies of the model and crafting the prompts to leverage these tendencies for a specific task or output.
- P-Tuning: This technique stands for Prompt Tuning, where the weights of an additional small model, called the prompt encoder, are fine-tuned. In this case, we keep the generative model's weights fixed and only change the parameters in the external prompt encoder. This allows for better control and adaptability than basic prompting without modifying the primary model. It is considered a form of parameter-efficient fine-tuning, as the tuning happens only on a tiny external component, keeping the resource utilization significantly lower compared to traditional fine-tuning.
- Parameter-Efficient Fine-Tuning (PEFT): This technique aims to strike a balance between computational resources and customization. It involves tuning a small fraction, typically less than 1% of the total number of weights. This method includes strategies such as P-Tuning, Adapters, and LoRA (Low rank Adaptation). For example, in the adapter method, small adapter layers are inserted between the pre-trained layers of the model, and only these adapter parameters are trained. These methods aim to deliver high performance with less computational cost, making it an attractive option for many tasks.
- Fine-tuning: This is the most comprehensive method for model customization, where all model weights are adjusted. Two instances of fine-tuning are the Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) stages. Here, 100% of the model's weights are tuned to adapt to the specific task at hand. It is often used when the task is substantially different from the original pre-training task or when maximum performance is critical.

The choice of the method for customization depends on the task at hand, the available resources, and the level of performance required. Each of these methods has its trade-

offs and advantages, and a deeper understanding of these can help us make an informed decision. So, generative AI can be thought of as a spectrum, with the needs and solutions varying depending on where an enterprise customer falls on this spectrum. Understanding their specific requirements will help us determine the appropriate type of engagement.

Here is an overview of all customization techniques:

- Using domain knowledge with Supervised Fine Tuning:
    1. Fine-tuning the model with domain-specific data to incorporate relevant knowledge.
    2. Enhancing the model's understanding and generation capabilities within a specific domain.
- Adding incremental knowledge with prompt learning:
    1. Gradually training the model with additional knowledge through prompt-based learning.
    2. Expanding the model's understanding of new concepts and information over time.
- Reinforcement Learning from Human Feedback for continuous knowledge:
    1. Utilizing human feedback to train the model further and improve its performance.
    2. Reinforcing positive behaviors and refining the model's responses based on user input.
- Information retrieval to execute for runtime knowledge to answer proprietary information.
    1. Implementing information retrieval techniques to enable the model to access real-time knowledge.
    2. Allowing the model to dynamically retrieve and incorporate up-to-date information during runtime.

These customization techniques empower users to adapt large language models to their specific needs, enhancing their accuracy, relevance, and reliability by leveraging domain knowledge, incremental learning, reinforcement from human feedback, and runtime information retrieval.

Another important topic is guard railing models for building trustworthy, safe, and secure LLM conversational systems.[25] Enterprise use cases require guardrails to exclude everything outside functional domain, eliminate bias and toxicity, and to align to enterprise goals.[26]

---

[25] https://developer.nvidia.com/blog/nvidia-enables-trustworthy-safe-and-secure-large-language-model-conversational-systems/

[26] https://github.com/NVIDIA/NeMo-Guardrails. NeMo Guardrails is an open-source toolkit for easily adding programmable guardrails to LLM-based conversational systems. It has programmable rails for LLMs with the following functionalities: steering the LLM towards producing outputs that accurately and effectively meet user intent, align the LLMs with the business goals of the enterprise, and prevent the model from generating undesirable, biased, or harmful content. The following guardrails have been implemented to date:

- Topical guardrails: prevent apps from veering off into undesired areas. For example, they keep customer service assistants from answering questions about the weather.
- Safety guardrails - ensure apps respond with accurate, appropriate information. They can filter out unwanted language and enforce that references are made only to credible sources.
- Security guardrails - restrict apps to making connections only to external third party applications known to be safe

Candidate Use Cases at Central Banks for Accelerated Computing

The mandate of central banks and financial supervisors has become challenging. Pandemics, economic instability, cyber risks, climate change, and sustainable investing are just a few examples of complex, emerging risks. Also, the digital, technological disruption of the financial system, as well as the speed of change and the growing complexity of the industry and its unbundling, increase the likelihood of supervisory blind spots.

For this reason, central banks, supervisors, and regulators will need to significantly adapt their operating model over the next decade to include technology and collaboration with (fin)tech firms. Many regulators are beginning to move in this direction with a variety of initiatives. Big Data Analytics and AI/ML will play a critical role, for example, in improving the quality and timeliness of risk identification and monitoring. Central banks already have access to vast amounts of valuable data, drawn from traditional, structured, and unstructured sources; streaming, complex, multi-layered, multi-modal, and alternative data to provide a holistic picture. Using advanced data collection and analytics techniques, certain areas of supervision will be able to use real-time monitoring of emerging risks and generate much earlier warning signals. Central bankers will be able to near cast and even predict the developments and provide a holistic picture of issues but also will be able to drill down into more granular micro developments and trends.

Considering these developments, data science and AI/ML models are becoming important tools used by central banks to fulfil their difficult mandate. Vast amounts of data are becoming available and numerous use cases in central banking are currently being developed. Data Science helps central banks model and analyse the economy and financial system -- AI/ML can support economic forecasting (e.g., for indicators such as inflation, housing prices, unemployment, GDP and industrial production, retail sales, external sector developments) and in business cycle analysis (e.g., compilation of sentiment indicators and use of nowcasting techniques to "forecast" the present).

General Uses of AI and ML

AI/ML can identify risk indicators, assess the behaviour of market participants, identify credit, and market risk, and monitor financial transactions and capital flows, detecting fraud, greenwashing, and assessing climate and economic risk.

AI/ML also supports financial risk assessment and surveillance exercises (functioning of the payment systems). It can detect market abuse with text mining 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.

Using detailed data from the "bottom up" can lead to a substantial improvement in inflation forecasts so central banks are increasingly using more granular data and greater computing power to produce their forecasts[27].

Using large-scale accounting data for financial statement audits can help to monitor trustworthiness of financial statements and detect potential misstatements by applying neural networks to learn representations of accounting data that constitute a representative audit sample and using these systems to detect potential anomalies.

[27] https://econpapers.repec.org/paper/boeboeewp/0915.htm

Neural Networks for Simulating Markets & Analytics

Neural networks for exotic options and risk: complex derivative modelling can be improved by neural nets. The unavailability of analytical solutions, the higher dimensionality for complex interest-rate and foreign-exchange products (volatility surfaces, volatility smiles, multiple curves), and the derivative requirements used to hedge, pose unique challenges. The oracles (traditional modelling) can do complex computations, but they are expensive and slow, and are traditionally done on computer grids overnight. A modern GPU accelerated computing platform can obtain accurate valuations in well under a second.

The application of generative adversarial networks (GANs) and transformers to simulate market data for predictive analytics is another technique to benefit regulators and central banks. Generative adversarial networks are one tool for developing synthetic financial datasets that can be used as training data across many classes of machine learning models. GANs can complement and augment the value of Monte Carlo simulations and replicate regime-specific conditions to better prepare models for more robust predictive analytics. Using a generator and discriminator, with care, the model will transition the simulated data so that it converges to an empirical distribution in a Nash or Quasi-Nash process, preserving more of the real-world temporal characteristics consistent with the targeted market regime.

Systemic Risk Modelling

A special challenge is Systemic Risk Monitoring (SRM) and monitoring 'at scale.' The scaling problem in SRM is the analytical dimension as the number of analyses, indicators, processes, and aggregations increases exponentially:

- Complexity & size: the financial system is inherently complex, interconnected and adaptive
- Datasets exhibit very different formats, granularities and origins but must be cleaned up and combined; several layers of aggregation are needed (micro, meso, macro). There are layers and a taxonomy from areas to countries to sectors to groups to corporations; data quality issues will be prevailing.
- Speed: as markets shift rapidly, analytical outputs must be produced at a faster pace
- Besides economics and financial shocks and risks there are also climate risks (physical and transitional) including stranded assets.

A scalable data model is needed to model the entire financial system at any level as a dynamic multilayer network (multigraph). There need to be filters / aggregations / modifications preserving the data model enabling scalable operations on the data.
This is naturally modelled as a network with a hierarchical taxonomy of nodes representing different nested aggregation layers. The links between the networks are direct transactions or relationships in terms of contracts, ownerships etc. with different weights and intensities.
This graph is a quasi-knowledge graph or multilayered knowledge base that uses a graph-structured data model or topology to integrate data.

Once the graph is established the analytical part can be executed. The graph contains the complex nested relationships and graph/network analysis can be done looking for clusters, communities and spreading nodes, e.g. to understand contagion in such an interconnected system. Shocks can be simulated and links can be predicted.

To allow the use of knowledge graphs in various machine learning tasks, several methods for deriving latent feature representations of entities and relations have been devised. In recent years Graph Neural Networks (GNN)[28] have become popular.

[28] https://developer.nvidia.com/gnn-frameworks

Another attempt to better understand the graph is by visualization in interactive dashboard to understand the large structures but also to deep dive into very detailed local structures.

All mentioned operations in building, analyzing and visualizing the graph require heavy compute workloads, especially when the graph structure is rich and the data set is complex and large. Traditional CPU-only systems quickly encounter bottlenecks making an accelerated infrastructure for storing, analyzing and visualizing large graphs highly desirable.

There are many other network structures in financial supervision like payment/transaction networks that can be analysed and simulated with graph/network technologies. Propagation of shocks, risks, defaults, and fraudulent activity can be better understood with such models and technologies which helps to monitor systems and design policies. NLP (Natural Language Processing) can be used to recognize entities, link structures and types of relationships.

Outlier and Anomaly Detection

A classical application for AI / ML is automatic data validation and outlier control, e.g., of loan and securities microdata. An example is daily transactions in foreign exchange derivatives and interest rates executed in OTC market by financial intermediaries.

The outcomes of the models must be validated and checked for accuracy and plausibility. Results must be interpretable and controllable, and there must be the possibility of automatic selection of informative features. These jobs can be supported by machines as humans would not be able to reliably analyse such large and diverse datasets in reasonable time and quality.

A popular approach is XGBoost which provides a regularizing gradient boosting framework in many programming languages and frameworks. It often achieves higher accuracy than a single decision tree. There are frameworks to tune the hyperparameters and to extract information on feature importance and interaction, e.g., based on Shapley values.

Anomaly detection can also be applied in credit register data to detect reporting gaps and strange patterns. The evolutions of reporting can be overseen, and structural changes and inconsistencies can be identified. The quality deviations can be ranked according to severity. Data assessment processes can thus be much more effective and efficient.

Some modelers use several different AI/ML methods at the same time to build a robust ensemble of models. Typical methods are isolation forests, distance/density based KNN and autoencoders which can all be GPU-accelerated.

Another classical outlier detection application is in time series data, e.g., using unsupervised representation learning and clustering like DBSCAN or network filtering like minimal spanning trees using distance measures like Gower. GPU-accelerated packages for these procedures exist as well as for the compute intensive pre-processing for time series as well as for unstructured data like written reports. Deep learning approaches like LSTM (Long Short Term Memory) are one of the more recent approaches to validation of (financial market) time series.

Outlier and anomaly detection in economic dataset can identify sudden changes in data sets and identify deviations from trends. Deep analytical capabilities and complex data visualization help to gain greater familiarity with large data set and a deeper understanding of data and inherent complex patterns and relationships.

An example for ML for anomaly detection in datasets with categorical variables is the Money Market Statistical Reporting (MMSR) data including additional data sets that are usually added.

Anomalies can be detected in both labelled and unlabelled data, and they can be organized into multiple categories regardless of whether the original data was labelled.

Natural Language Processing (NLP) and Large Languages Models (LLMs)

NLP/LLM is a reliable powerhouse for Central Banks when it comes to processing of natural and non-natural language. Many different tasks can be executed with NLP like sentiment analysis, entity recognition, text summarization, question answering, topic modelling, translation, speech recognition, aspect mining and natural language generation.

NLP is a tool to translate unstructured information like text and voice into structured representations and also generating and responses with text and speech in much the same way as humans do.

An example is using NLP for the identification of topics in FOMC Transcripts from the Federal Open Market Committee Meeting Minutes. It could assist researchers at central banks and institutions to determine topic priorities.

Another example is an NLP tool to understand and assess information provided in text form, such as annual reports and audit reports or capital and liquidity assessments. Findings can be classified, plausibility can be checked, and reference could be provided to corresponding regulatory literature.

Another popular NLP area is analyzing central bank speeches and if they predict financial market turbulence[29]. NLP can also improve central bank accountability and policy[30].

A classical and especially useful NLP task is sentiment analysis for economic forecasting. Increasingly, central banks have been relying on timelier indicators to assess the near
term developments of the economy in advance of the release of official statistics. The news sentiment measures move with the fluctuations in economic conditions and can provide information in nowcasting movements in the less timely, quarterly survey-based business sentiment. News sentiment can help to forecast or at least nowcast private investment, CPI, inflation, and employment vulnerability. Data sources can also be alternative data from social media like Twitter. Adding explainability and visualization is important to understand how data and models come to conclusions.

Our chapter 'Generative AI and Large Language Models in Central Banks' described the topic and frameworks in-depth.

Computational Trustworthy AI, AI Governance, Trustworthiness and Explainability

A key issue in financial big data will be interpretable modelling. This is because to make evidence-based policy decisions central banks need to identify specific explanatory causes or factors which they can take action to influence. Furthermore, transparency regarding the information produced by big data analysis is essential to ensuring that its quality can be checked and that public decisions can be made on a sound, clearly communicated basis. Lastly, there are important legal constraints that reduce central banks' leeway when using private and confidential data; interpretable modelling helps address all these issues.

---

[29] https://www.sciencedirect.com/science/article/pii/S1303070121000329

[30] https://www.cato.org/cato-journal/spring/summer-2020/how-natural-language-processing-will-improve-central-bank

The focus of ML/AI models in central banking and supervision cannot be just predictive accuracy. Models must be trustworthy, interpretable, explainable, interactive, fair, robust, accountable, and secure. Proper risk management, data/AI governance, and compliance must be in place.

A bank, commercial bank or supervisor using AI in production needs to overcome the explainability gap to produce transparent, appropriate governance, risk management, and controls over AI. The publication "Financial Risk Management and Explainable, Trustworthy, Responsible AI" (Fritz-Morgenthal et al. 2021) discusses further details.

More specific use cases can be found, moving the discussion from the realm of the general to the specific. One related use case based on SHAP explanation values can be found in Bussmann et al. (2020). The use case had been selected as the best AI case in the EU Horizon2020 project FIN-TECH (www.fintech-ho2020.eu) by the European financial services community including the European supervisors. Other related Explainable AI (XAI) use cases can be found in Jaeger et al. (2021) and Papenbrock et al. (2021). The developed approaches can help to implement Explainable AI using techniques like Shapley values (a local, and global variable importance method with mathematical footings in co-operative game theory) even for large and complex models. For classical datasets, these methods can already substantially improve the transparency of portfolio allocation processes. They also enable the visualization of the variables and their influences of the entire data set in a single analysis. Clustering and network analysis of the variables and their influences are often used to find overall model structure and connections. Real-time monitoring of model drift in continuous learning machines is applied. Simulations and perturbations to test the robustness of the model can be run at large scale. Iterative and evolutionary approaches are now able to create and evaluate millions of models, allowing supervisors to select those that best balance prudential goals. It will also be necessary to meet the upcoming requirements from the European AI Act, especially the technical and auditing requirements for High-Risk AI[31]:

- Creating and maintaining a risk management system for the entire lifecycle of the system.
- Testing the system to identify risks and determine appropriate mitigation measures, and to validate that the system runs consistently for the intended purpose, with tests made against prior metrics and validated against probabilistic thresholds.
- Establishing appropriate data governance controls, including the requirement that all training, validation, and testing datasets be as complete, error-free, and representative as possible.
- Detailed technical documentation, including around system architecture, algorithmic design, and model specifications.
- Automatic logging of events while the system is running, with the recording conforming to recognized standards.
- Designed with sufficient transparency to allow users to interpret the system's output.
- Designed to always maintain human oversight and prevent or minimize risks to health and safety or fundamental rights, including an override or off-switch capability.

In summary, meeting the overall combination of the supervisory, legal, diverse stakeholder, and technical requirements will drive model development, deployment, monitoring, and retirement process that features enhanced auditability, transparency, and explainability. The ever-increasing data volume, velocity, and variety – across structured and unstructured sources – combined with the rapid pace of AI development will drive an overall system architecture that is scalable, flexible, and secure. To be efficient with both people's time and energy, the system must strongly adopt lessons learned in the leading HPC and AI supercomputers of today, leveraging

[31] See https://datainnovation.org/2021/05/the-artificial-intelligence-act-a-quick-explainer/

GPU accelerated compute and networking that can accelerate workloads across the diverse, end-to-end data science use cases of today and tomorrow.[32]

[32] For additional information, see "Computing Platforms for Big Data Analytics and Artificial Intelligence" (Bruno et al. (2020)), which highlights the experiences of central banks with respect to HPC platforms.

Appendix: Selected accelerated libraries and frameworks

Topic	Accelerated libraries and frameworks
Machine Learning	Rapids[33] is an open source suite of accelerated Python libraries including XGBoost, which is also available through the open source XGBoost[34].
AI Development and Training	Performance optimized containers and code for PyTorch[35], TensorFlow[36] and others. Enterprise Support[37] available.
Graph Models	Accelerated libraries for graph neural networks (GNN[38]) and graph analytics(cuGraph)[39].
Language Models	Support for development (NeMo[40]) and open source repositories (HuggingFace[41]) plus runtime support via Triton inference serving[42].
Speech and Transcription Models	Customizable, multi-cloud speech to text, text to speech, and translation (RIVA[43]).
AI/ML Models in Production	CPU and GPU support for latency or throughput optimized production deployment of AI and tree based models(Triton).
Recommendation Engines	Support a variety of AI powered models for product or next best action recommendations (Merlin[44]).
Quantum Simulation & Integration	Support for quantum simulation (cuQuantum[45]) and integrating

[33] https://rapids.ai/

[34] https://github.com/dmlc/xgboost

[35] https://catalog.ngc.nvidia.com/orgs/nvidia/containers/pytorch

[36] https://catalog.ngc.nvidia.com/orgs/nvidia/containers/tensorflow

[37] https://docs.nvidia.com/ai-enterprise/index.html

[38] https://developer.nvidia.com/gnn-frameworks

[39] https://github.com/rapidsai/cugraph

[40] https://github.com/NVIDIA/NeMo and https://catalog.ngc.nvidia.com/orgs/nvidia/containers/nemo

[41] https://huggingface.co/nvidia

[42] https://github.com/triton-inference-server/server

[43] https://developer.nvidia.com/riva

[44] https://github.com/NVIDIA-Merlin/Merlin

[45] https://github.com/NVIDIA/cuQuantum

	traditional and quantum computers(CUDA Quantum[46]).
Spark and ETL / Data Processing	User transparent Spark plug-ins[47] and Python libraries (Rapids).
Federated Learning	Federated learning using homomorphic encryption. (NVFlare[48])
Visualization	Plotly Dash, cuxFilter (part of Rapids), and other librarie[49]s enable accelerated visualization and interaction with large amounts of data.
Mathematical & Scientific Computing	Comprehensive support via the HPC-SDK[50], cuNumeric[51] for open source Python computing.

Acknowledgements

This work partially relies on the support from the European Union's Horizon 2020 research and innovation program "FIN-TECH: A Financial supervision and Technology compliance training programme" under the grant agreement No 825215 (Topic: ICT-35-2018, Type of action: CSA).

The authors are grateful to Kevin Levitt and Marc Staempfli for valuable advice concerning this paper.

[46] https://github.com/NVIDIA/cuda-quantum

[47] https://nvidia.github.io/spark-rapids/

[48] https://github.com/NVIDIA/NVFlare

[49] https://docs.rapids.ai/visualization

[50] https://developer.nvidia.com/hpc-sdk

[51] https://github.com/nv-legate/cunumeric

References

Ashley, John, Papenbrock, Jochen and Schwendner, Peter, (2022), "Accelerated Data Science, AI and GeoAI for sustainable finance in central banking and supervision" in Settlements, Bank for International eds., Statistics for Sustainable Finance, vol. 56, Bank for International Settlements, https://EconPapers.repec.org/RePEc:bis:bisifc:56-23.

Bruno, Giuseppe, Hiren Jani, Rafael Schmidt, and Bruno Tissot. 2020. "Computing platforms for big data analytics and artificial intelligence." IFC Reports 11. Bank for International Settlements. https://ideas.repec.org/p/bis/bisifr/11.html.

Bussmann, Niklas, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2020. "Explainable Machine Learning in Credit Risk Management." Computational Economics, September. https://doi.org/10.1007/s10614-020-10042-0.

Fritz-Morgenthal, Sebastian and Hein, Bernhard and Papenbrock, Jochen, Financial Risk Management and Explainable Trustworthy Responsible AI (June 25, 2021). Available at http://dx.doi.org/10.2139/ssrn.3873768

Jaeger, Markus, Stephan Krügel, Dimitri Marinelli, Jochen Papenbrock, and Peter Schwendner. 2021. "Interpretable Machine Learning for Diversified Portfolio Construction." The Journal of Financial Data Science 3(3). https://doi.org/10.3905/jfds.2021.1.066

Li, A., S. L. Song, J. Chen, J. Li, X. Liu, N. R. Tallent, and K. J. Barker. 2020. "Evaluating Modern Gpu Interconnect: PCIe, Nvlink, Nv-Sli, Nvswitch and Gpudirect." IEEE Transactions on Parallel and Distributed Systems 31 (1): 94–110. https://doi.org/10.1109/TPDS.2019.2928289.

Mattson, Peter, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David Patterson, Hanlin Tang, et al. 2020. "MLPerf Training Benchmark." http://arxiv.org/abs/1910.01500.

Papenbrock, Jochen, Peter Schwendner, Markus Jaeger, and Stephan Krügel. 2021. "Matrix Evolutions: Synthetic Correlations and Explainable Machine Learning for Constructing Robust Investment Portfolios." The Journal of Financial Data Science, March, jfds.2021.1.056. https://doi.org/10.3905/jfds.2021.1.056.

Radhakrishnan, Ramesh, Yogesh Varma, and Uday Kurkure. 2019. "Evaluating Gpu Performance for Deep Learning Workloads in Virtualized Environment." In 2019 International Conference on High Performance Computing & Simulation (Hpcs), 904–8. IEEE.

RAPIDS-Spark. 2021. "RAPIDS Accelerator for Apache Spark." GitHub Repository. https://github.com/NVIDIA/spark-rapids; GitHub.

Serena, Jose Maria, Bruno Tissot, Sebastian Doerr, and Leonardo Gambacorta. 2021. "Use of big data sources and applications at central banks." IFC Reports 13. Bank for International Settlements. https://ideas.repec.org/p/bis/bisifr/13.html.

Tissot, Bruno, Timur Hulagu, Per Nymand-Andersen, and Laura Comino Suarez. 2015. "Central Banks' Use of and Interest in "Big Data"." IFC Reports. Bank for International Settlements. https://EconPapers.repec.org/RePEc:bis:bisifr:3.

Tissot, Bruno. 2018. "Big Data for Central Banks." In International Workshop on Big Data for Central Bank Policies–Bali, 23:25.

Zeranski, Stefan, and Ibrahim Ethem Sancak. 2020. "Digitalisation of Financial Supervision with Supervisory Technology (Suptech)." J. Intl. Banking L. & Reg., no. 8. https://ssrn.com/abstract=3632053.

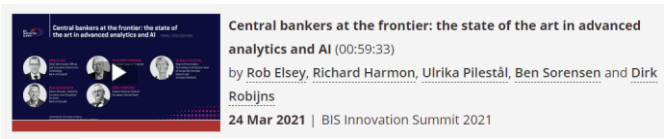
MODERN COMPUTING PLATFORMS AS KEY CENTRAL BANKING TECHNOLOGY

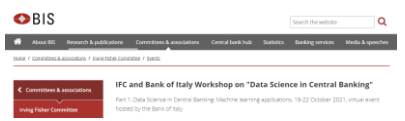
BY DR. JOCHEN PAPENBROCK, FINANCIAL SERVICES AND TECHNOLOGY DEVELOPER RELATIONSHIP LEAD EMEA

jpapenbrock@nvidia.com

DATA SCIENCE AT CENTRAL BANKS

We have analysed (and contributed to) the data science use cases and IT/software tools the central banks are currently establishing.





Sources:

- IFC/BIS publications:
  - "Computing platforms for big data analytics and artificial intelligence"
  - "Big data and machine learning in central banking"
  - "Big data for central banks"
  - "The suptech generations"
  - "The use of big data analytics and artificial intelligence in central banking"
  - "Central Bank Communications: information extraction and semantic analysis“
- Own projects and interactions with some of the leading central banks globally

Observation
Central banks embrace Big Data and AI with typical tools and workflows but acceleration is not yet realized.

TYPICAL WORKFLOW AND TOOLS

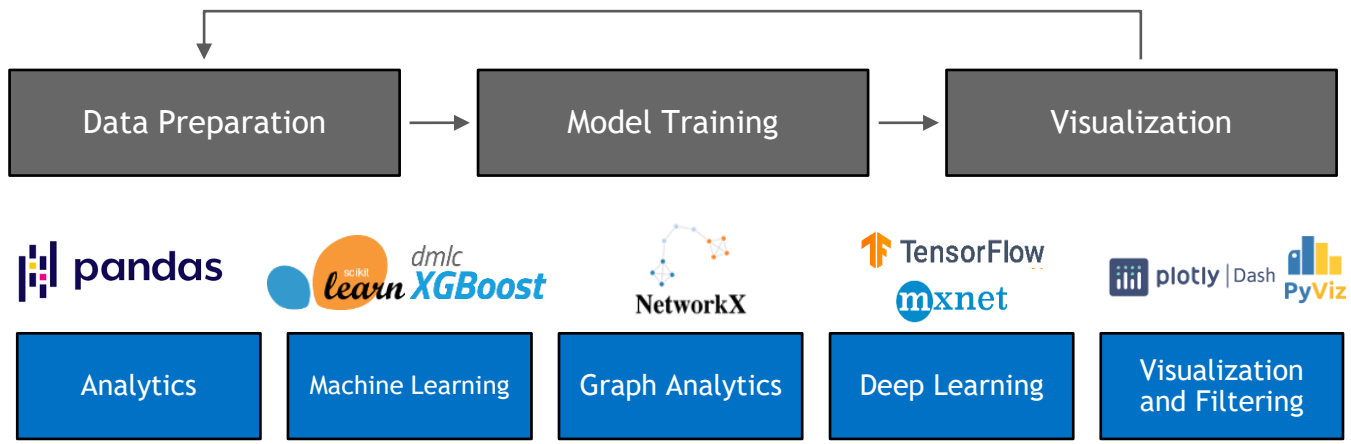
Open-source projects and Python tools have democratized data science

SQL and graph data base

PyData

Spark

NumPy



There are computational bottlenecks with CPU-only processing

GPU SUPPORTS PARALLEL MASS CALCULATIONS

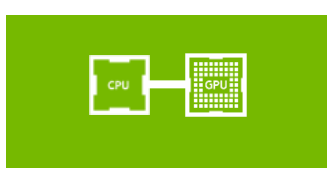
IFC Report
No 11

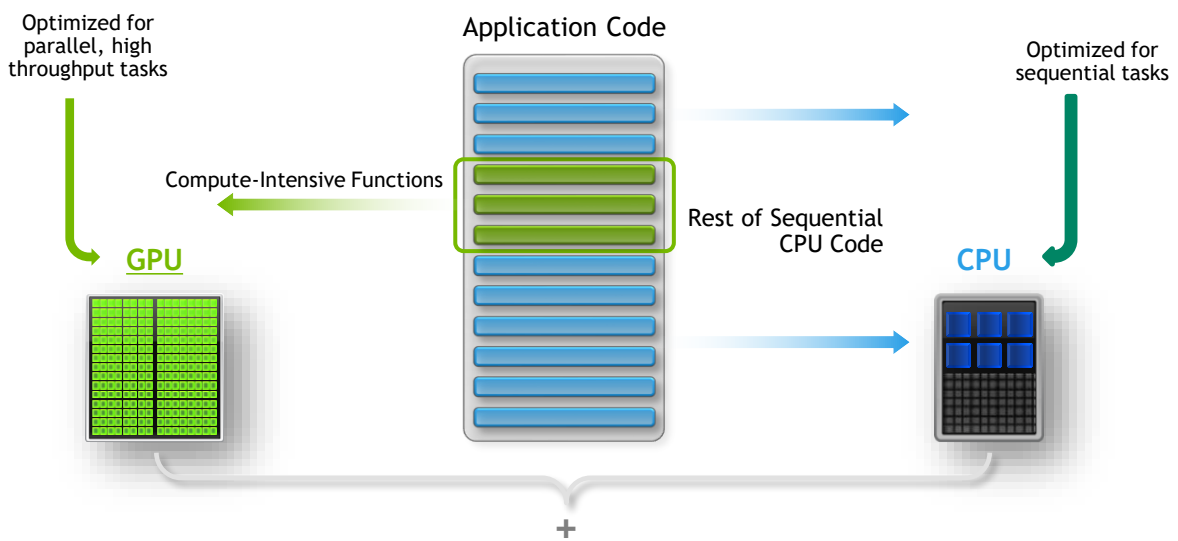
Computing platforms for big data analytics and artificial intelligence

April 2020

BANK FOR INTERNATIONAL SETTLEMENTS

"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."





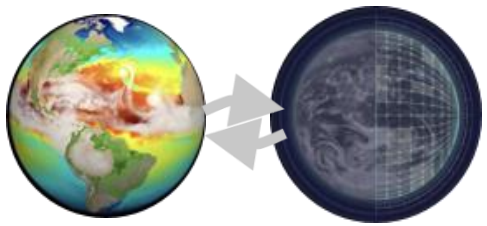
WHAT HAPPENS WHEN WE EMBED GPU?

CAMBRIDGE-1

Boosting COVID-19 research

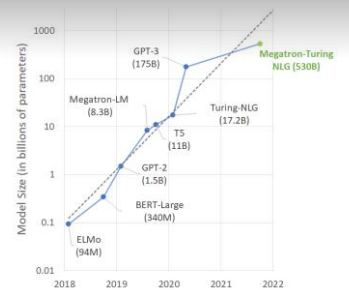
SUSTAINABLE FINANCE

Earth digital twin to forecast climate change



LARGE LANGUAGE MODELS

Of the size of GPT-3 and Megatron-Turing (530B)



AI RESEARCH

Research SuperCluster (RSC) with Meta / Facebook

Accelerated Computing puts Big Data Processing, AI, Simulation, and Visualization to a new level

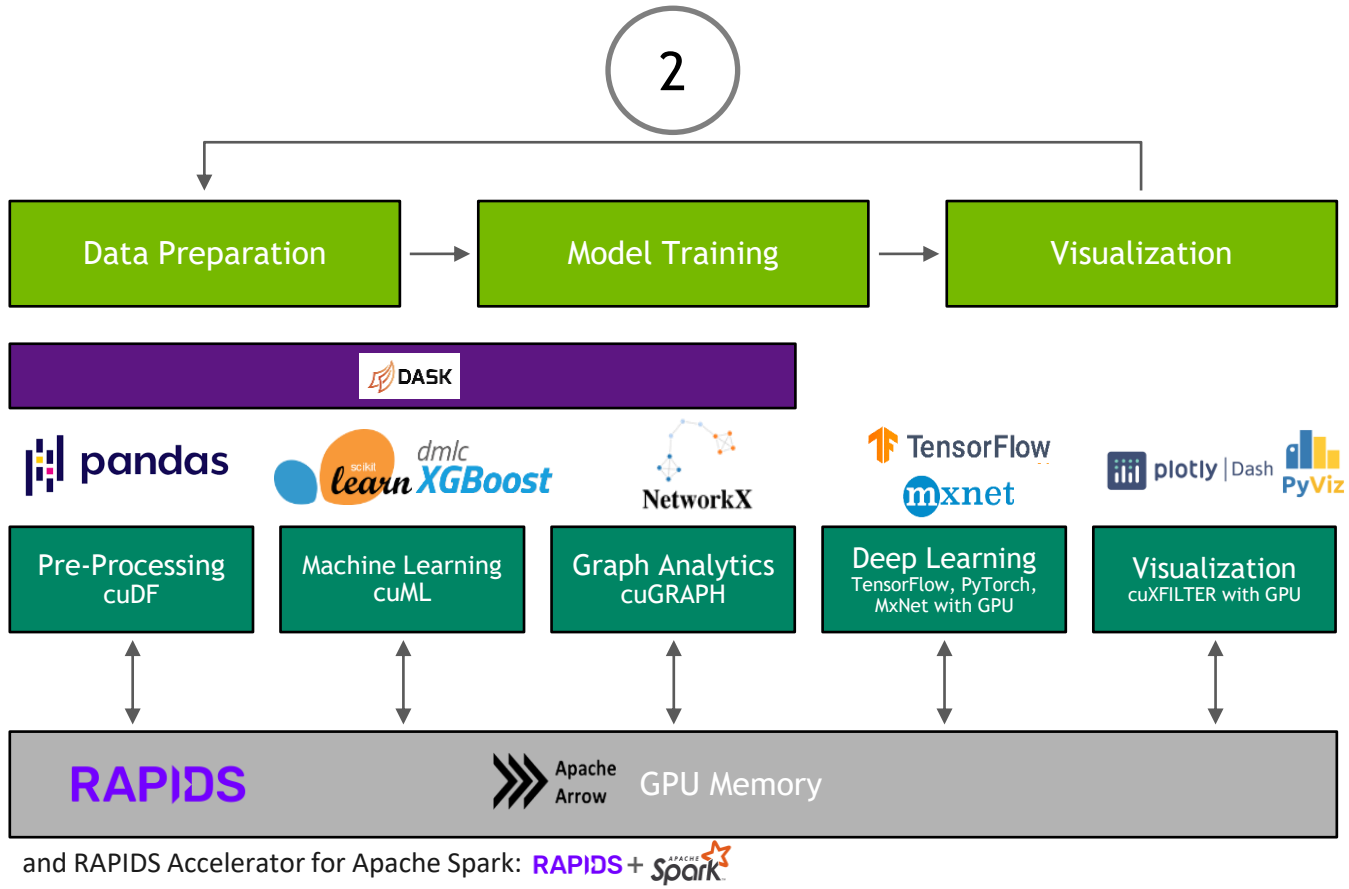
EXAMPLES OF ACCELERATED APPLICATIONS

1

SDK	NLP Task
NeMo / Megatron	Named Entity Recognition Topic Modeling Intent Identification Relation Extraction Sentiment Analysis Language Translation Text Summarization
RIVA	Speech to Text
TAO	Model Training (Computer Vision & NLP)
TensorRT	Model Optimization
Triton	Inference Serving

Accelerated NLP to address ESG needs

2

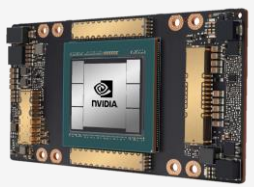


and RAPIDS Accelerator for Apache Spark: RAPIDS + Apache Spark

Accelerated end-to-end data science

ACCELERATED COMPUTING PLATFORM

1. Chips & Systems
2. Platform Software
3. Application Frameworks



Platform Symbiosis
Hardware & Software

2.5M
Developers

30M
CUDA Downloads

2,500
GPU-Accelerated Applications

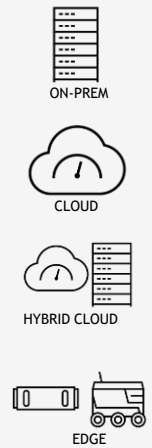
9,000
AI Startups

Ecosystem
Developers & Partners

#1

MLPerf

Productivity
World Records



Available Everywhere
Every Cloud and OEM

150 SDKs (often pythonic)
Ready Application Frameworks

OUTCOMES

- **increases developer productivity and scalability**
- **reduces TCO, time to insight and infrastructure complexity**

AI TRUSTWORTHINESS AND AI GOVERNANCE

- We are engaged in numerous projects, webinar series and software projects

- Development of new XAI workflows in investment management with Munich Re

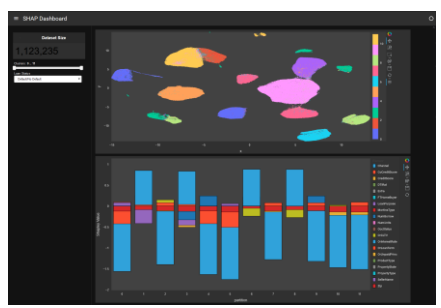
- FAIC (Financial AI Cluster)

  Stimulating an ecosystem around a collaborative technology platform for automating compliance with AI regulation, AI governance, AI assurance

gaia-x

FINTECH RISK MANAGEMENT

European Commission



Computation of Explainable AI based on an arbitrary AI black box model
Interactive Exploration of the entire model

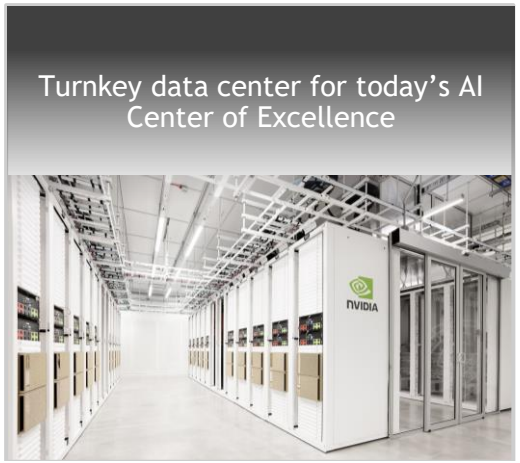
*) This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 825215 (topic ICT-35-2018, Type of action: CSA)The content reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

NVIDIA

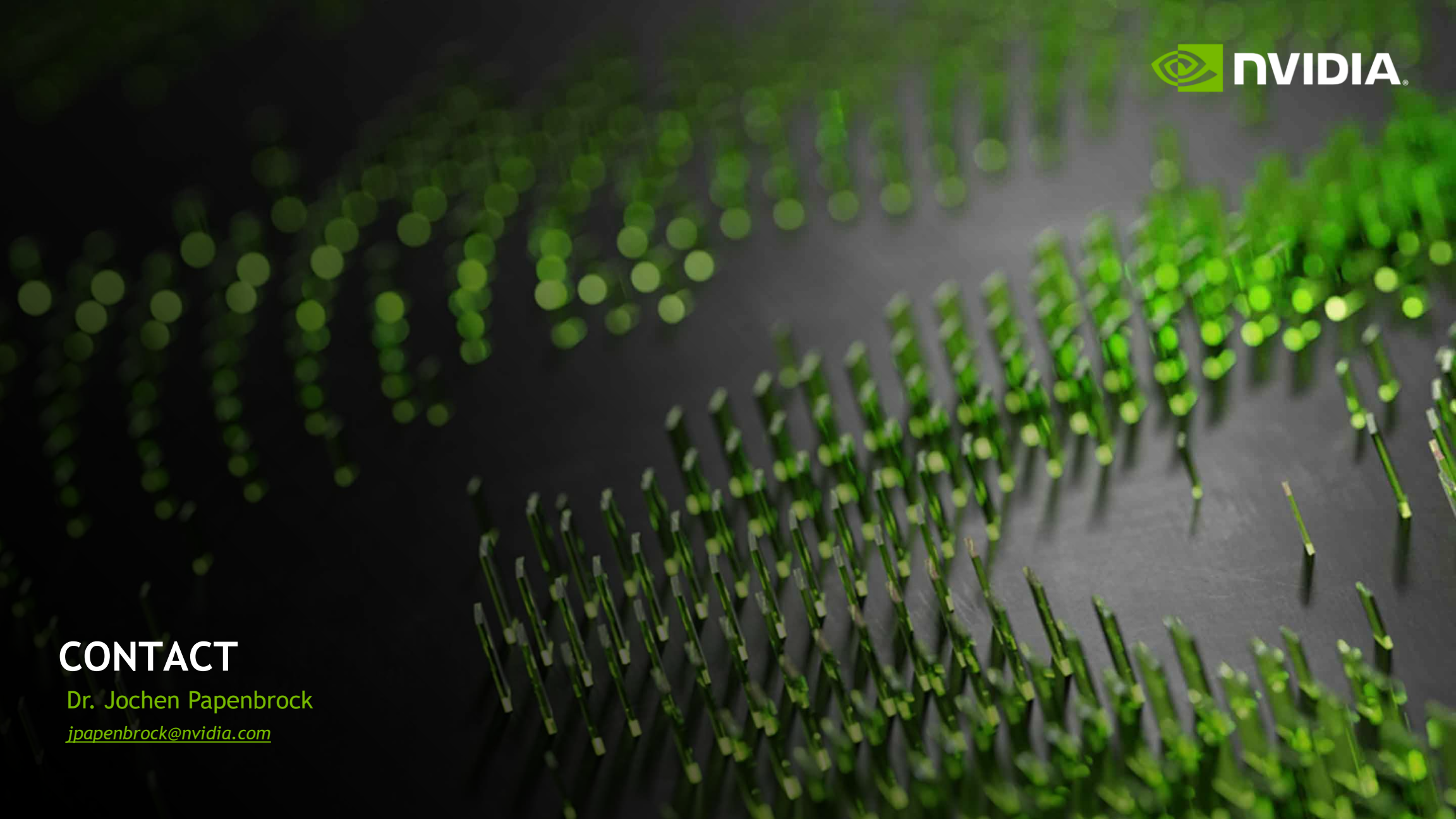
SUMMARY AND OUTLOOK

- There is a rise of AI/HPC workload in enterprises and organizations around the globe
- There is a need for a new generation of computing platforms
- Many organizations build their AI Center of Excellence including the appropriate infrastructure
- We help with hardware, software, ecosystem and know-how

With many developers and technology companies presenting, registration is free



NVIDIA. DEEP LEARNING INSTITUTE



CONTACT

Dr. Jochen Papenbrock

jpapenbrock@nvidia.com