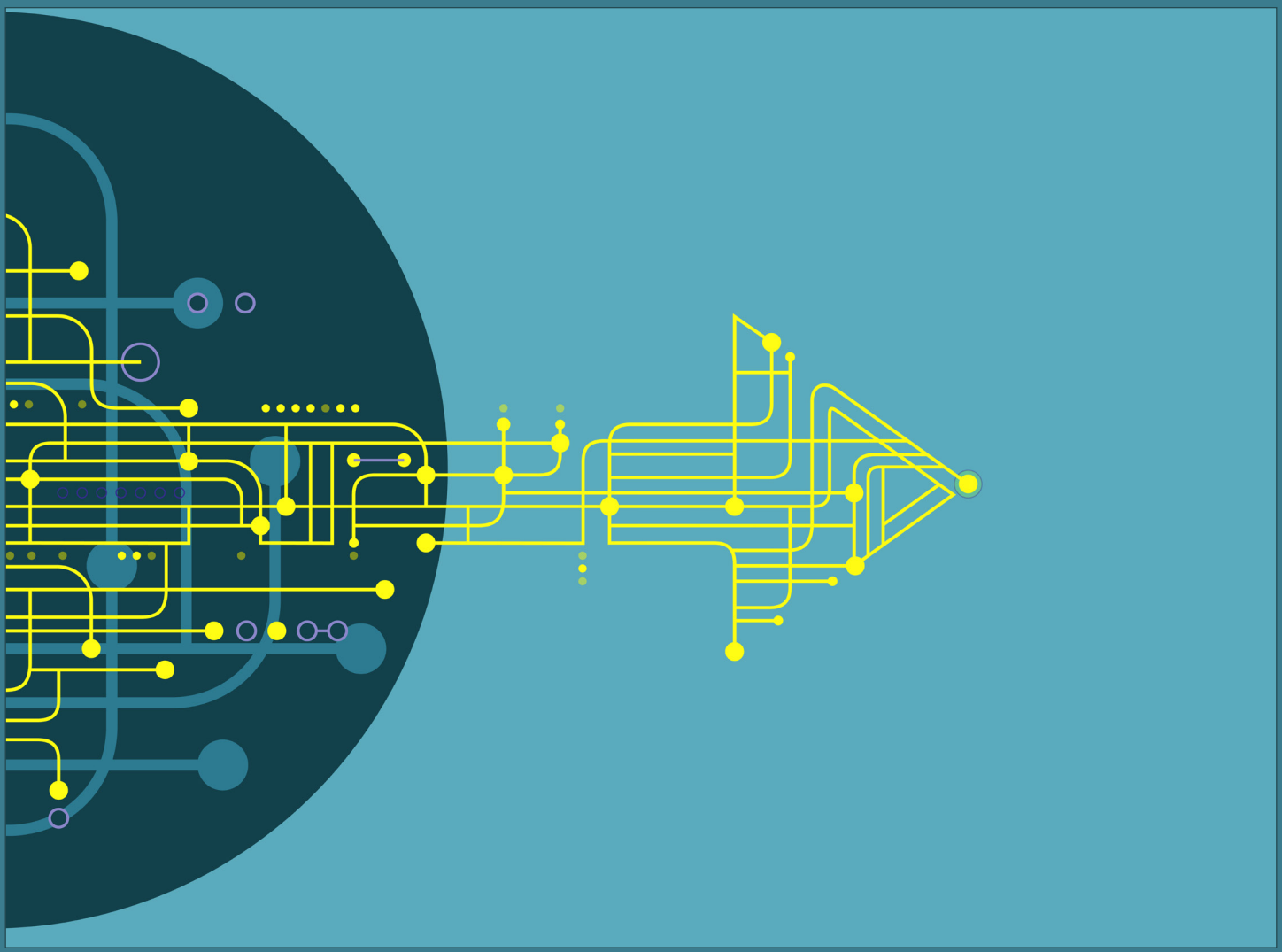
Risk publications

Data to anchor a new age of risk management

The growth of AI and real-time compliance drives new requirements in data and analytics

White paper

January 2022



Risk.net

databricks

Contents

2 Introduction

3 Speed, scale and transparency

4 Testing and recalibration

5 Dynamic data

6 Data sovereignty and ethical AI

7 New transformations

About Databricks

Databricks is the data + artificial intelligence (AI) company. With origins in academia and the open source community, Databricks was founded in 2013 by the original creators of Apache Spark™, Delta Lake and MLflow.

As the world's first and only lakehouse platform in the cloud, Databricks combines the best of data warehouses and data lakes to offer an open and unified platform for data and AI.

www.databricks.com

databricks

Cover image Amtitus/Getty

Published by Infopro Digital

Introduction

Model risk management (MRM) – the supervising of models with a goal of reducing the likelihood of an adverse outcome from a misused or poorly performing model – first found a larger home following the global financial crisis that began in 2007–08. This led firms to discover – often in retrospect – the dangers of a laissez-faire approach to desks governing their own siloed development methods and output.

In the years since, the cost of risk and compliance management has skyrocketed, stretching across a series of regulatory changes and new mandates. The Fundamental Review of the Trading Book (FRTB), which is the new mandate going live in January 2023, the Comprehensive Capital Analysis and Review (CCAR), General Data Protection Regulation, Current Expected Credit Losses (CECL) methodology and the Anti-Money Laundering Act 2020 are but a few. And that is to say nothing of the financial losses and productivity consequences accrued via account fraud and money laundering itself, especially when models fail.

Over the horizon, many risk and compliance officers could also envision a reality in which data would reign and compute power could be commoditised, with cloud infrastructure gaining acceptance. Today, new platforms for these frameworks deliver far faster and more prolific model production. But, without governance layered in, those capabilities can also prove to be a curse – as the model landscape becomes larger, more fragmented and trickier to harness.

A wave of seismic shifts, both within and beyond the technology stack, has also seen stakeholders across financial services – from clients and counterparties to regulators, industry groups, council and board members – ponder where risk decisions come from. Equally important is how to document and defend them.

Model risk is now centre of attention and a guiding principle for chief risk officers across banks and asset management firms. When properly implemented with data governance as its foundation, a new MRM estate is transformative and can influence how decisions are made from seemingly mundane know-your-customer (KYC) checks to algorithms governing trillions of dollars in trade netting or collateral movements. It likewise reduces operational costs and lends a solid footing to effective risk management.

Industry observers say today most firms lie somewhere between: they have a modicum of dexterity around their model governance and infrastructure, where it sits and an ambition to improve, while still re-engineering their way into the future.

The hard work lies at touchpoints where artificial intelligence (AI) techniques are being applied directly to outcomes – such as customer due diligence and trade compliance or where new model flavours iterating upon unstructured data across different desks and borders must be validated and effectively managed. This white paper takes a closer look at each of these drivers, as the challenges posed by data science and new platform solutions move to the fore.

Contributors

Antoine Amend, Technical Director, Databricks

Brian McConnell, Senior Solutions Architect, Databricks

Lourenco Miranda, Managing Director and Head of Model Risk, Americas, Societe Generale

Speed, scale and transparency

Today, modern enterprises must tackle unstructured data, semi-structured data, and data with high variety, velocity and volume. But current data systems for compliance cannot perform the requisite advanced analytics that require scale.

Data quality is also of critical importance. If your data isn't well governed, how can you ensure your downstream compliance and risk models are accurate? Equally as important is transparency. Risk teams must demonstrate to internal stakeholders and regulators alike how their models were built, what data they used and why they came to their conclusions. The unprecedented Covid-19 pandemic-era volatility illustrated that risk and compliance models – and their testing – need ongoing recalibration. In these dynamic times, data quality and transparency cannot be an afterthought.

To meet modern compliance requirements, financial services institutions (FSIs) must report on growing volumes of data stretching years into the past. Risk calculations that were run weekly or daily must now be run several times per day – in many cases, in real time as new data comes in.

Additionally, regulations such as CCAR, CECL and FRTB require risk teams to scale simulations for thousands if not millions of scenarios in parallel. The volume of data, reporting frequency and scale of calculations require massive compute power that far outstrips the capabilities of legacy on-premise analytics platforms.

For instance, the Financial Industry Regulatory Authority (Finra), a US self-regulatory organisation that regulates member brokerage firms and exchange markets, is taking advantage of a next-generation analytics platform to swiftly iterate on machine learning models and scale detection efforts up to hundreds of billions of market events per day.

Finra deters misconduct by enforcing rules, detecting and preventing wrongdoing in US markets, and monitoring millions of transactions daily to meet this task. Modern data analytics and AI platforms, such as the Databricks Lakehouse Platform, enable democratisation of data and bring previously siloed teams together, cutting down overall time to market, increasing reusability of feature libraries and improving operational efficiency.

As a result, Finra is understood to have significantly improved fraud prevention and provided better protection of the securities market, leading to a safer financial future for US investors.

Today, the ability to deliver real-time insights on streaming data is critical to protecting businesses and customers. Whether for compliance violation alerts or anomaly detection in fraud and other risk-related activities, traditional data warehouses often don't have the agility to both ingest and analyse streaming data at speed.

AI tools in legacy systems are often disconnected from data processing, which results in data needing to be replicated – leading to slow time to insight on these platforms. The detection of a threat or risk event can occur days or weeks after the event takes place, generating significant losses and operational risk for FSIs.

There is a need to pull in and analyse intraday data to quickly assess risks and respond. Coins.ph – a digital payments platform in the Philippines with a customer base of 10 million – performs anomaly detection in real time to prevent fraud. A tier-one US bank also uses real-time anomaly detection for fraud and AML to rapidly analyse and deter financial crime.

With AI increasingly being brought to bear in many risk and compliance activities surrounding AML and fraud prevention, FSIs are encouraged to undertake a complete rethink of their data management and infrastructure, aligning data ingestion with analytical layers in an effective way.

Testing and recalibration

But how to cope from a model risk perspective? To start, the issue is not one of sheer complexity; it lies in AI's operational role alongside risk managers and the way AI translates and retranslates information. For these reasons, any AI problem is, above all, a data problem.

"Five years ago, everyone was talking about an AI strategy as a 'centre of excellence'," says Antoine Amend, former director of data science at Barclays and now technical director at Databricks. "The idea was: if we put some PhDs in a room, AI innovation will just happen. Today, how many of those models are still in production? Instead, the centre of excellence has become a centre of enablement; it is less about the smartest model and more about the foundation to do this right. Now it's really about moving from an AI strategy to a data strategy."

Lourenco Miranda, managing director and head of model risk, Americas, at Societe Generale says today's model validation is a process that essentially turns the data inside out – and for good reason.

"The overarching policy from SR 11-7 [supervisory guidance on MRM from the US Federal Reserve] still works; it's a sturdy piece of regulation," Miranda adds. "The biggest change between a traditional statistical and a machine learning model is the process of development, and then that impacts model validation and changes in two aspects. The first is related to 'feature engineering', which is how you transform initial data into something more useful for modelling. For traditional statistical regression, you may perform some transformation or combinations of explanatory variables, whereas in a machine learning or AI model this feature engineering is part of the structure, especially if there is some automation built into a decision. You have to almost reopen the box and scrutinise what the model does to the original data."

Dynamic data

Next is the changing nature of the data itself. Many applications that use AI, such as financial crime detection, are feeding in new data types and interpreting unstructured pieces of information, such as news, voice or an image that requires translation into something the machine can understand. Miranda refers to this as 'vectorisation', using elements of computational linguistics called natural language processing.

"The vectorisation transforms the data to a language that the machine understands, which is still binary. For the AI, you still use linear algebra, optimisation or minimisation algorithms like any other AI model," he explains. For instance, using computational linguistics in AI comes with lots of complications, teaching not only lexicon or vocabulary to the machine but also semantics – meaning. Thereafter, humans must be able to understand the output, so it must be translated back into meaningful context. "That requires a lot of engineering in the data, preparation for the model to understand and ultimately impacts its validation," Miranda adds.

These aspirations rely in large part on the promise of the cloud, but not exclusively. As Amend puts it, KYC and AML projects – dynamically scraping a picture of a customer's house or verifying a potential fake address – once required six months to build the dedicated hardware nodes out alone. Today, they can spin it up in hours. Therefore exogenous datasets of 2021 – from new pandemic hotspots to natural catastrophes – can be layered on top.

"Today, more than ever, we've seen that customer checks and trade compliance work within imperfect rules; often they can get swamped by transitory effects or a single market move, and a new AI engine may not draw the right conclusion on a given client check or a trade," Databricks' senior solutions architect Brian McConnell explains, pointing to such cases as the Suez Canal blockage in March 2021 and its effect on commodities traders, or a data breach in a particular country causing a higher likelihood of account fraud.

He adds that traditional financial models would need to be recalibrated and backtested in these circumstances. "However, since there is a reliance on historic data to calibrate the models, it is difficult to adjust properly for these kinds of changes where there has been no precedent. With AI/machine learning, you can incorporate new information and data quickly into the model and properly account for it," McConnell says.

A newer and more effective data management architecture that has emerged at firms over the past few years is the data lakehouse – fast becoming an all-encompassing solution. A data lakehouse architecture enables efficient and secure AI and business intelligence directly on vast amounts of data stored in data lakes.

The lakehouse architecture also allows organisations to bring in new data from any source at short notice and process it even if it is less mature, says McConnell. "One would still need robust model governance to pick up these effects, retrain to improve the model performance and deliver this seamlessly into production. The results from these models can help inform and augment traditional models, allowing risk managers greater insight when managing the risk," he adds.

Moving deeper into AI, compliance processes can introduce more voluminous and less structured data, such as tracking ship locations or a retail customer's social networking activity – but, for older data management approaches, that creates a challenge for AI's learning. "With a data warehouse, you're taking a lot of data and making it smaller to provide the user information, as opposed to training an algorithm that needs to process all of the data in bulk and build a model," adds McConnell.

McConnell explains that, with a data lakehouse, an organisation can have data strongly managed, beautifully mapped and structured, or live with it and work on it even if it's raw. "The key is that you can choose how to deal with your data based on its maturity and structure."

Data sovereignty and ethical AI

Aside from effective data management, there are two emerging sets of questions posed by regulation: first, how far apart models and the underlying data can physically sit; and second, how validation can take proper account of potential biases that are coded into or mistakenly learned by AI over time. The latter is a concept known as 'ethical AI' and part of the broader mandate around model explainability.

Challenges have always existed here as firms build a "mesh" architecture that is "both together and separate", says McConnell. "Knowing every country has its own regulations used to require warehouses scattered all over, with different versions of the code and bespoke solutions that get out of step with each other. This leads to inconsistent views of the same data that then produce different models when used for training. For model risk, it's often a logistical nightmare."

Meanwhile, there are rising expectations for ethical AI, Miranda says. "When we talk about explainability, that really comes from an opacity in the way decisions based on the output of the AI could be made.

"As customers and humans – for instance, in a potential credit decision – we look at potential bias in the initial data annotation used to train the AI, which could come from past natural human perception, by first always asking: 'Why this outcome?', and then, in a contrastive way, ask: 'How can I, as a consumer, fix it?' When validating a model, that's a big discussion – can the modelling process itself create a bias in the AI? How do we show consumers and regulators that it hasn't?," he adds.

The question relates back to interrogation of the data itself. "Explainability, transparency and balance in line with regulatory requirements are all a challenge, and there remains a lot of grey area," Amend says.

In practice, McConnell says, data governance remains critical. "Ultimately identifying these issues means shortening the testing lifecycle; we should understand it today as a business problem, not as an AI problem," he says. "You don't train one model at a time when you could run 100 in parallel and decide against your criteria which performs best. Proving a model against stressed conditions for bias or balance is even harder when having to augment or conduct evaluation outside your environment, with moving parts and data in motion. So much can change in the months it may take to validate a model and fix its flaws in this fashion; often it is here that mistakes or oversights can enter the process."

New transformations

Today, most senior risk officers and data scientists view technology transformation as the pinnacle of a broader sociological exercise at their companies. This further reduces operational friction, in addition to decreasing spin-up time by an order of magnitude, freeing up regulatory capital and reducing lost opportunities to serve clients.

One large banking client, McConnell observes, was able to reshape its entire operating model away from cost silos and pass-throughs by more collectively distributing model ownership. “We saw that change where they were able to fully federate their model creation and validation processes and reduce the time to deliver models down from nine months – start to finish – to a matter of weeks. Not only because of smarter AI, but because of that new distribution of responsibility,” he says.

There is also a constant effort to educate at each and every step, and this is why model risk managers such as Miranda choreograph model validation at leading banks in the way they do.

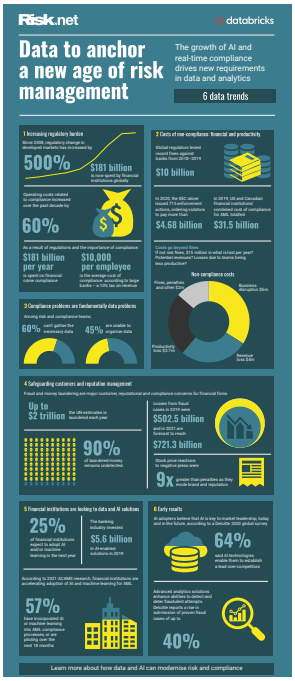
“We’ve acknowledged that new models and more powerful AI algorithms will always be arriving,” Miranda says, “However, the future is really about understanding how to deploy models in bigger and more transparent data and for wider applications that could have a positive impact in the environment and the society. To do this successfully, everything starts with that: the right infrastructure, annotation and data quality.”

The shape and speed of contemporary regulation is one great reason, but the best way to persuade that internal talent is with a data platform already designed for what’s coming next.

Read more

Learn how financial institutions are modernising risk and compliance with the Databricks Lakehouse Platform at www.databricks.com/discover/smarter-risk-compliance-with-data-ai

Six data trends



Risk.net and Databricks have identified six data trends in the growth of AI and real-time compliance to drive new data governance:

1 Increasing regulatory burden

2 Costs of non-compliance: financial and productivity

3 Compliance problems are fundamentally data problems

4 Safeguarding customers and reputation management

5 Financial institutions are looking to data and AI solutions

6 Early results

View the infographic taking a deeper dive into these trends at www.risk.net/7897056


Risk.net


databricks