unravel™

WHITE PAPER

WINNING THE AI INNOVATION RACE

How AI Helps Optimize Speed to Market and Cost Inefficiencies of AI Innovation

Abstract

In this paper you will learn:

- Top pitfalls that impede speed and ROI for running AI and data pipelines in the cloud
- How the answers to addressing these impediments can be found at the code level
- How AI is paramount for optimization of cloud data workloads
- How Unravel helps

AI Needs AI

Business leaders from every industry now find themselves under the gun to somehow, someway leverage AI into an actual product that companies (and individuals) can use. Yet, an estimated 70%-85% of artificial intelligence (AI) and machine learning (ML) projects fail[1]. According to IDC, the second-biggest reason why AI projects fail is lack of production-ready data pipelines for diverse data sources (31%, behind "AI technology didn't perform as expected or as promised" at 35%).[2]

The complexity of the AI pipelines have data engineering and operations teams' hands full with troubleshooting broken pipelines—from identifying the problems to determining how to fix them—which dramatically impacts MTTR.

And as companies rely on cloud data platforms to run their AI and other data products, there is a direct correlation of performance to cost for AI and data pipelines in the cloud. While the cloud offers great elasticity and scalability with infinite resources and compute, using those infinite resources can also come with an infinite bill.

The secret weapon to resolve these problems is to use AI to help create your AI. Managing data pipelines—for cost and performance—at the scale of today's data estates has become virtually impossible for humans alone. But it's perfectly suited for automation and AI. AI assistance enables and empowers organizations to "get it right" the first time while also enabling a wider skills base in organizations to create a compound effect of more people doing more work much faster.

Impediments to Speed and ROI when Running AI and Data Pipelines in the Cloud

Speed

If we look at the vast amounts of data needed to build models, it's continuously growing—both the number of sources and the sheer volume of continuous data inputs. With more data and more data sources increasingly being added, data pipelines need to be able to scale without significant performance degradation. Data pipelines break for various reasons, such as schema changes, inconsistent data formats, upstream data changes, or even due to network issues.

Building performant data pipelines is tricky for even the most skilled data scientist or ML engineer, let alone for those less experienced. In today's ML/AI data teams at even the largest enterprises, the range of experience and expertise ranges from PhDs to interns—and every level in between. And often these folks are dispersed across the business, from marketing to risk to finance to product and supply chain. Of those who build models in some capacity, many lack the training and expertise to run these models efficiently. There are hundreds, if not thousands, of people running data jobs, but even the largest enterprise could count on two hands the number of experts able to spot code problems quickly and accurately.

Ask any data engineering team about broken pipelines and more often than not, they'll tell you it's "drop everything and get it fixed ASAP." If it's a critical app, like a fraud detection app for a bank, even a few minutes of downtime can cost millions. The toilsome process for fixing broken pipelines involves a myriad of time-consuming steps of checking logs, reviewing configuration, debugging code, and more.

In the cloud, performance = costs

Enterprises today may have 10,000+ pipelines running at any given time. Running sub-optimized pipelines in the cloud often causes costs to quickly spin out of control. And the enormity of datasets being used for AI pipelines exacerbates the problem.

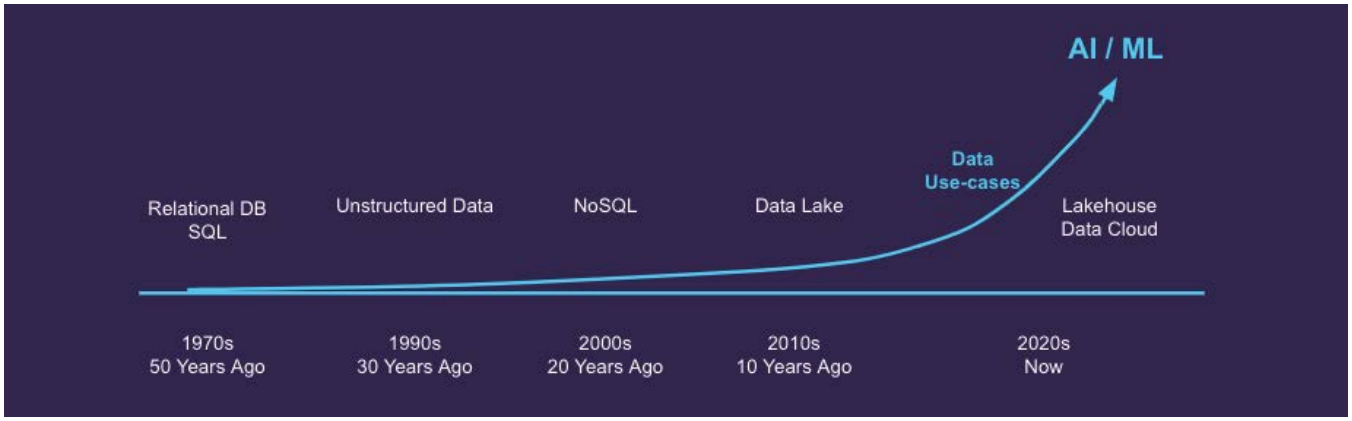

Figure 1. The speed, scale and complexity of data is increasing exponentially

The most obvious culprit is oversized infrastructure, where users are simply guessing how much, how many, what kind of resources they need rather than basing their decisions on actual usage requirements. Storage cost is another big factor. Teams may be using more expensive options than necessary for huge amounts of data that they rarely use. Not all data needs to be on hand immediately. No point in keeping seldom-used data in expensive "hot" storage rather than more cost-effective cold storage. But knowing which data is hot, warm, or cold is the key question.

Higher, unnecessary cost as a symptom of performance inefficiencies becomes amplified with AI projects. There's more data to process. More people running more, and bigger, workloads in Databricks, Snowflake, BigQuery, Amazon EMR. The scale and scope of performance inefficiencies unwittingly introduced by the 1000s of users across the business grows exponentially when running AI.

Cloud data costs account for about 39% of the cloud bill and is the fastest growing part of the bill. With the mad dash for AI, it's about to increase exponentially. Adding insult to injury, many estimate that wasted cloud data spend is about 30%. But without the right tools, identifying what's waste and what's reasonable (or even better, optimized) cost takes a lot of time and a lot of experts. With all the attention on skyrocketing cloud costs. GenAI pioneer, industry thought leader, and former chief cloud strategy officer at Deloitte David Linthicum argues that we should even start asking ourselves whether (or when) we should go back on-prem.[3] Given the cloud's elasticity and scalability, the benefits are tremendous. It will just take some smarter measures to keep costs under control. Much of the same measures for keeping costs under control help keep the AI pipelines performant.

Pricing structures and cost levers

Every cloud data platform is different. First, there's a dizzying array of pricing structures and options—pay-as-you-go, prepaid/fixed subscriptions, on-demand or reserved instances, savings plans, spot instances. Then the way you're billed for cloud usage varies from vendor to vendor: in Databricks, you're thinking DBUs and workspaces; in Snowflake, queries and warehouses; in BigQuery, slots and queries.

Consequently, the various cloud providers and data platforms all have their own idiosyncrasies and nuances when it comes to optimizing for performance and cost. For instance, it may be more cost-efficient to consolidate warehouses in Snowflake, but just the opposite in Databricks (spinning up individual job clusters in lieu of sharing an interactive cluster). In Databricks you're dealing with all the intricacies of Spark; in BigQuery and Snowflake, you're looking more at SQL queries.

Key takeaways:

Every day 1000s of individual users are running data workloads—pulling on cloud data cost levers—that impact the speed and ROI of data and AI pipelines running in the cloud.

- Oversized infrastructure (number, size, type) inadvertently causes monthly cloud data bills to soar.
- Storage costs for AI and other massive data analytics projects are rapidly rising.
- Identifying inefficiency and waste is a time-consuming slog, robbing time and effort from innovation.
- Optimizing for performance and cost happens down at the code level.

Code: The Secret Lever to Keeping AI and Data Pipelines Performant and Cost Efficient

We've all heard a story about a query that ran in the cloud for a full weekend and cost $200,000. One query. Or read about how Shopify found themselves ambushed by a single $1 million query.[4] As the Shopify example shows, even the most data-forward company with extremely skilled developers can commit inefficient code. Most enterprises have thousands of developers committing code, but only a handful of data engineers to make sure everything runs reliably and efficiently.

Code runs everything: resource allocation, dataset processing, pipeline orchestration, job configuration, etc. It's down at the code level where potential reliability issues are first introduced and where most cost inefficiencies lie. As shown in the composite chart below, most performance/cost inefficiencies are buried below the surface, deep "in the weeds." How individuals are requesting resources or using data tables or configuring jobs or writing SQL queries causes workloads to take longer to run than they should, cost more than necessary, even cause pipelines to fail outright.

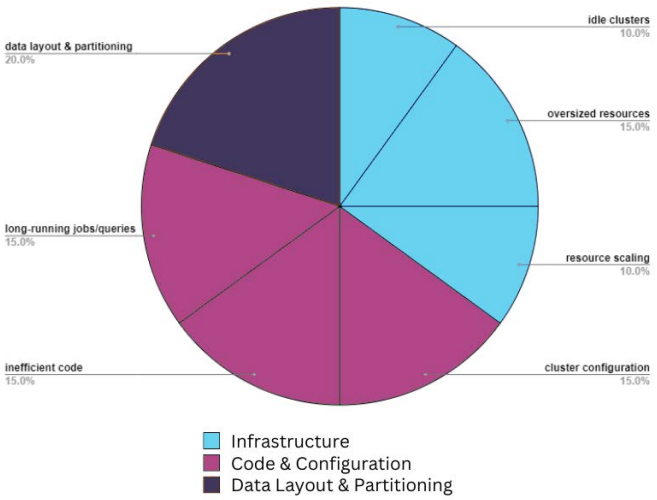

Figure 2. Breakdown of cost and performance inefficiencies

Inefficient code affects performance and time to value and costs: data schema, data skew and load imbalances, idle time, and a rash of other code-related issues that make data pipelines run inefficiently. Knowing how and when to pull in just one column with the least amount of data for a model or using partitions or Select* can make a huge difference in the reliability of the pipelines and the cost to run them.

Key takeaways:

- Code inefficiencies derail the reliability and speed (time to value) and drive up the cost of AI and data pipelines running in the cloud.
- These inefficiencies are introduced, and costs incurred, deep "under the hood," down at the individual user level.
- Code inefficiency manifests itself in a variety of ways:
  - Data skew
  - Inefficient joins
  - Resource contention for CPU or in driver
  - Out-of-memory errors
  - Load imbalance
  - Excess data used or inefficient data pull
  - Slow tasks and long-running jobs
  - Excessive idle time
  - etc.

Data Engineers: Star or Pressure Cooker?

Data and AI pipeline performance problems fall on the shoulders of data engineering teams. These teams are small by design and have to drop whatever they're working on in order to resolve the issue. Often the people tasked with fixing the problem don't have the information they need at their fingertips, so they spend hours (maybe even days) manually hunting down and cobbling together the details that will allow them to fix things. Or they simply don't know exactly what to do to resolve the problem—these pipelines are highly complex, and it's rare for one individual to understand all the ins and outs of the different (and constantly changing) technologies in play.

Many data observability tools surface the equivalent of a "check engine" light, but knowing that something is wrong is completely different from understanding how to fix it—or even how to avoid it in the first place—because they do not connect to the code. And an organization's experts are already buried under an avalanche of trouble tickets, troubleshooting mission-critical applications/pipelines, feeling like they're shoveling sand against the tide.

With millions of lines of code running daily, It is humanly impossible to validate and optimize everything with traditional approaches like code reviews.

AI can help

All the performance and cost details you need to understand what's going on in your data estate are out there, hidden in plain sight. What's needed is to (1) corral all that telemetry and metadata from across the various cloud data platforms, application systems, and cloud vendors, (2) synthesize and correlate everything in context for the task at hand, and (3) throw some highly trained algorithms at the correlated data model to automate prescriptive AI-driven recommendations on exactly what to do, in plain English.

AI that connects to the code level and continuously scans and monitors the code to capture the granular telemetry and performance details goes beyond a "check engine" light to detail diagnostics of where the issue lies—be it with inefficient joins, resource contention, excess data, etc.

Knowing what is failing and why is good, but with the velocity of data pipeline running on any given day, you cannot effectively manage for speed and costs without having AI-driven insights and recommendations for optimizing infrastructure, storage, and code for speed, scale and costs. Think hundreds of Shopify examples running every day—individually they may not be that extreme on a cost level, but collectively they add up to significant impact on performance and cost.

AI can root out these thousands of inefficiencies. The people actually running AI and data pipelines in the cloud need some kind of automated intelligence to collect and correlate the thousands of details from logs, traces, metrics, events, and other granular details from dozens of different components in their pipelines—and then make sense of it all in order to optimize their jobs. It's become more than humans can handle by themselves—especially at enterprise scale—at the speed demanded by the business.

The ability to really dig down into usage and spend by individual job or user—who is spending what on which project—yields invaluable insights. You might find that the most expensive jobs are not the ones that are making your company millions. You may find that you're paying way more for exploration than for data models that will be put to good use. Or you may find that the same group of users are responsible for the jobs with the biggest spend and the lowest ROI (in which case, it might be time to tighten up on some processes).

Unravel: Data Observability, FinOps, and AI-driven Optimization

Unravel's data observability + FinOps platform provides AI-driven insights and recommendations for optimizing pipelines for performance, reliability, and costs according to the specific cloud data platform's cost structure (e.g., DBUs for Databricks).

Because Unravel connects to the code level, job level, project level, and user level and combines AI and ML algorithms, Unravel is able to provide real-time cost visibility, predictive spend forecasting, and performance insights for workloads to accelerate time to value.

At Unravel's core is an AI-powered Insights Engine, which is purpose-built and trained to understand all the intricacies and complexities of specific modern cloud data platforms. Unravel provides platform-specific solutions for Databricks, Snowflake, Google Cloud BigQuery, and Amazon EMR. The Engine has been built to ingest and interpret information from the continuous millions of ongoing data streams to provide real-time insights into application and system performance, and recommendations to optimize costs, including right-sizing instances and applying code recommendations for efficiencies.

Most of the data teams that use Unravel mitigate performance and cost issues by identifying and fixing them in dev environments before they hit production. This empowers developers (or anyone running AI and data pipelines in the cloud) with self-service processes for cost and performance optimization by putting Unravel right in the front of the development environment for immediate feedback and optimization prior to moving apps and jobs to production.

In Conclusion

Companies are going to realize that the exponential efficiency of data analytics in the cloud rests in optimizing the code, not just the infrastructure. How code is written often determines how much compute is used. So having deep visibility into costs of data pipelines, and recommendations for how to optimize the code to lower costs and enable more workloads for the same costs will be paramount.

Organizations that have adopted a FinOps approach will begin to tackle the more difficult nut to crack: cloud data costs. If they're not already, cloud data costs will become the top expense in an enterprise IT budget. Companies will bring the same financial discipline and collaborative decision-making to running AI and data pipelines in the cloud that they have begun with other cloud operations across the business.

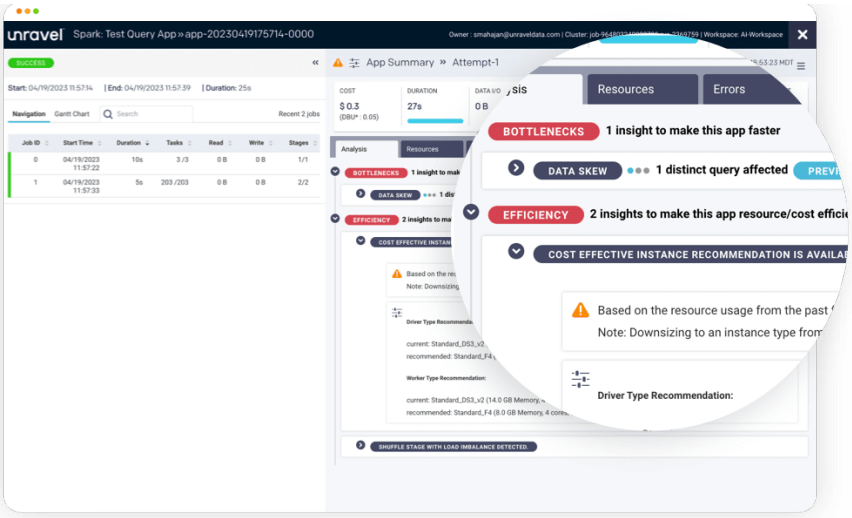

Figure 3. Unravel's AI automatically analyzes everything you have running and provides actionable recommendations for improvement down to the individual user, job, line of code—in real time.

References:

1. Harvard Business Review, Keep Your AI Projects on Track (2023)
2. IDC Research, Create More Business Value from Your Organizational Data (2023)
3. David Linthicum, Why We're Seeing an Evolution of Cloud Computing with the Focus Away from "the Cloud" (2023)
4. Shopify, Reducing BigQuery Cost: How We Fixed a $1 Million Query (2022)

About Unravel Data

Unravel Data radically transforms the way businesses understand and optimize the performance and cost of their modern data applications – and the complex data pipelines that power those applications. Unravel's market-leading data observability and FinOps platform with purpose-built AI for each data platform, provides actionable recommendations needed for cost and performance data and AI pipeline efficiencies. A recent winner of the Best Data Tool & Platform of 2023 as part of the annual SIIA CODiE Awards, some of the world's most recognized brands like Adobe, Maersk, Mastercard, Equifax, and Deutsche Bank rely on Unravel Data to unlock data-driven insights and deliver new innovations to market. To learn more, visit www.unraveldata.com.