



The Data Lake Graveyard Petabytes In, No Insights Out

Dr.A.Shaji George

Independent Researcher, Chennai, Tamil Nadu, India.

Abstract – The promise of enterprise data lakes was that they would be the powerhouse of competitive advantage. Boards bought into them, CIOs backed them, and vendors sold them with machine learning at scale, real-time analytics and a single source of truth in tow. The balance sheet is a different story 4-5 years later. Data is constantly coming in from ERP systems, CRM platforms, IoT sensors and third-party feeds, in petabytes. The amount of flow out as decisions moving revenue, cost or risk is very small. This article explores why most enterprise data platforms, such as on Snowflake, Databricks, AWS, Azure Synapse, Google BigQuery and on-premise Hadoop distributions, have turned into a graveyard for expensive storage and not a decision engine. It demystifies the consumption-based vendor playbook that keeps compute meters running, explores twenty real-world consumption patterns that result in financial leakage and estimates the annual leakage, which typically starts at SAR 8 million and goes up to SAR 25 million in large organisations. The analysis then provides a structured recovery roadmap, which includes a FinOps baselining, decommissioning, rationalization and governance. The main thesis is straightforward. There is no correlation between the amount of data ingested and data value. It's dependent on the quantity of decisions enabled. Businesses that don't know the difference pay a monthly bill to maintain a graveyard.

Keywords: Data lake cost, Snowflake FinOps, Databricks waste, Cloud data ROI, Hadoop TCO, Data platform leakage, Zombie dashboards, Data lake governance.

1. INTRODUCTION

1.1 The Library Nobody Reads

There is a data lake somewhere in the middle of nearly all big companies. It was started 3-5 years ago with a bold presentation that was something like "Becoming a Data-Driven Organisation. The slides offered machine learning, predictive analytics, dashboards in real time, views of customers in 360 degrees and a continuous competitive advantage. The executive sponsorship was obtained. Capital was allocated. A team of platform was established. Pipelines were constructed. Data flowed.

Today, the same data lake contains four to 10 petabytes of ingested data. It costs from SAR 8 million to SAR 25 million per year to run on Snowflake, Databricks, AWS, Azure, Google Cloud or on-premise Hadoop infrastructure. And it has an impact on less than 10 real business decisions per month. The dashboards exist. The models exist. The pipelines are currently on schedule. Invoices are received in a timely fashion. However, the business decisions are still made the same way as 2018 using 15 Excel reports that are maintained by an analyst who pulls out the data with ad hoc queries.

This is the Data Lake Graveyard. It is the most costly storage system in use in the modern enterprise and it has the lowest insight to cost ratio of any technology investment that is on the books. Data enters continuously. Nothing meaningful exits. The meter for computing activity is constantly running. A helpful analogy to make the picture come to life. Let's say a family constructed a large private library in their house. All books, newspapers, magazines and pamphlets in print are delivered daily in trucks. The lights are

operated by electricity. The shelves are kept clean by cleaners. The arrivals are catalogued by librarians. After five years, the collection comprises of 10 million volumes. But when a child requires a book for a school essay, that child turns to Google, since it's easier to look for something on the open Web than in a family library. The trucks continue to roll in. The librarians continue to catalogue. The electric meter just continues to rise. Even by the most reasonable measure, the library has failed and every operational measure says it's running smoothly.

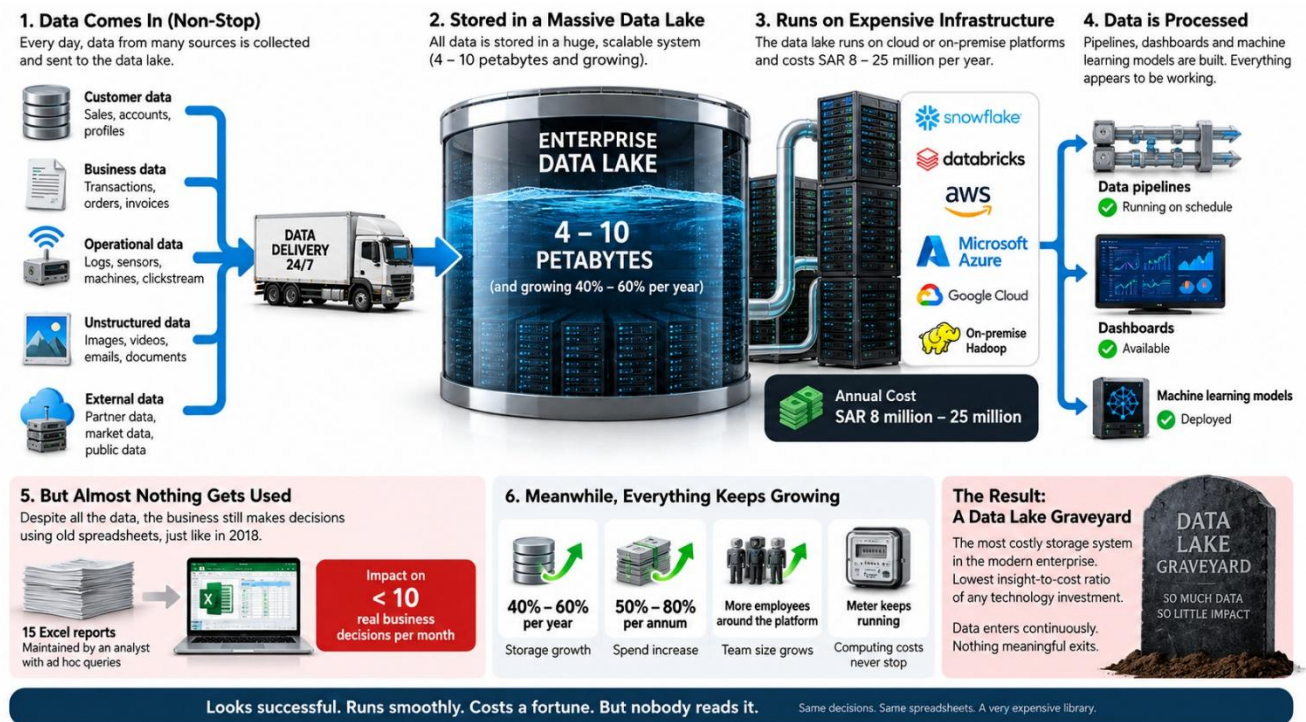


Fig -1: Enterprise Data Lake

That's the current situation of enterprise data lakes in 2026. The storage is increasing at a rate of 40% to 60% per year. Spend is increasing at 50-80% per annum. The number of employees increases in the vicinity of the platform. Business wise, operation wise it's still on the same spreadsheets that it was on before the platform was in existence.

2. OBJECTIVES

The goals of this study are intentionally practical and geared to the finance executive, the technology executive and the board member who signs the checks.

- First, the article seeks to set a straightforward baseline for enterprise data platforms, as to what they actually provide and what they were contracted to provide.
- Second, it aims to unravel the economic logic of the new data platforms that are sold based on consumption, and reveal the choices that are made in their design that benefit the vendor and not the customer.

- Third, it estimates, in Saudi Riyals, the annual amount of the top twenty most common types of financial leakage seen in large scale deployments of data lakes in the Gulf.
- Fourth, it offers a decision making framework for deciding whether a data platform is worth saving or decommissioning not all underperforming data lakes need to be modernized and not all modernization efforts yield investment.
- Fifth, it provides a step-by-step recovery plan starting of FinOps intervention, and extending to multi-year rationalization.
- Finally, it suggests a list of metrics that boards and CFOs should insist on before renewing, committing to more credit, or upselling to more vector databases.

The study is designed to help inform decisions and not to justify past decisions and, sunk cost is not relevant to the forward question of value.

3. HISTORICAL CONTEXT HOW THE INDUSTRY ARRIVED HERE

The graveyard can only be understood in the light of the three different waves of architecture which preceded it. Every wave was touted as the remedy for the problems of the preceding wave, and each wave left a residue of unsolved costs which added up to today's bill.

From Warehouses to Data Lake Graveyards

Three waves of technology, one massive amount of data. Each wave promised a solution, but also left behind new costs. Here's how we got here, and how the technology works.

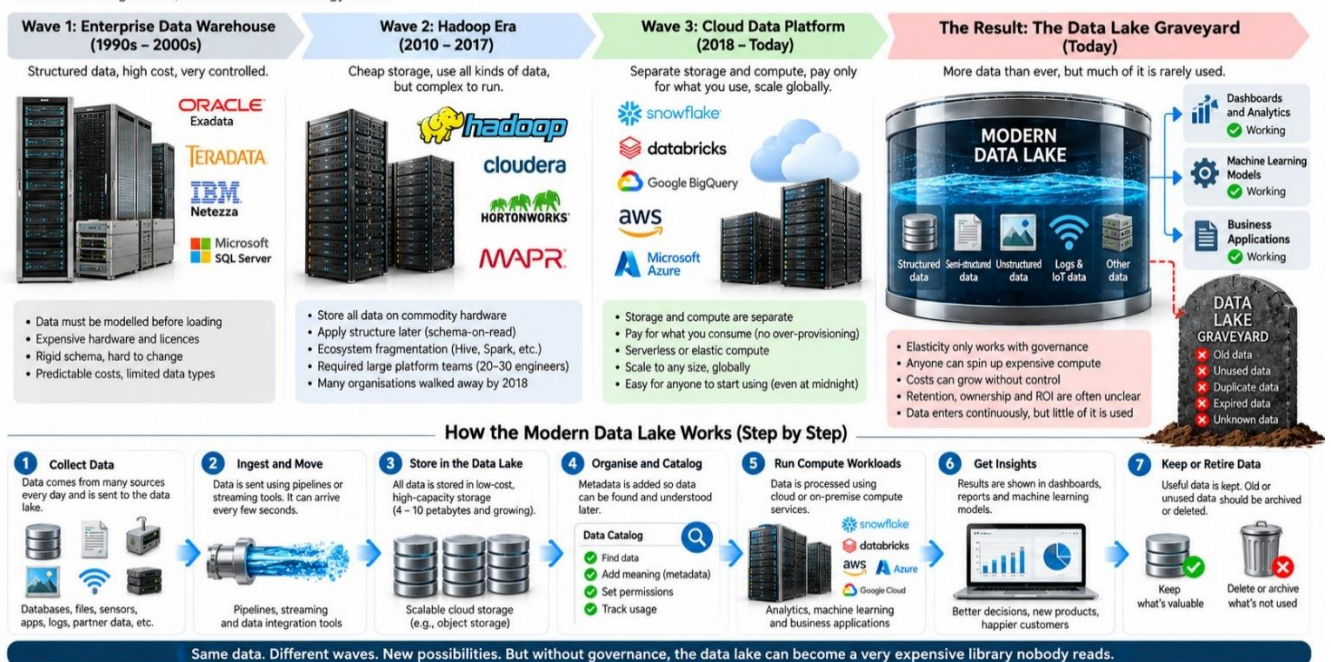


Fig -2: From Warehouse to Data Lake Graveyards

The first wave was the enterprise data warehouse era, which was ruled by Teradata, Oracle Exadata, IBM Netezza and Microsoft SQL Server throughout the late 1990s and 2000s. These were costly, inflexible and schema-rigid but they were very clear about their limitations. The data needed to be modelled prior to



loading. Business users had knowledge about the contents of the warehouse. There were few costs that were variable and unpredictable.

The second wave was the Hadoop years (2010–2017). The idea that the pitch was revolutionary is quite correct. Keep all of this on commodity hardware, and have low prices for it. Apply structure later. Unlock unstructured data. There was a significant investment in Cloudera, Hortonworks and MapR distributions. There were under deliveries in most of these deployments. The ecosystem fragmented. Spark replaced MapReduce. Hive started to be a problem. Platform teams grew to twenty or thirty engineers, just to maintain clusters. By 2018, many organisations had silently decided that on-premise Hadoop was too expensive to be worth the investment, and business use had not been as high as expected.

The third wave that began around 2018 was the cloud data platform era. Snowflake split the storage and compute, and enabled consumption-based pricing. Databricks made Spark a commercial product and introduced the concept of a lakehouse. Google BigQuery has launched serverless, planetary scale querying. Then, AWS Redshift, Azure Synapse and more. This time the pitch was a stretchy one. Don't pay for services that you don't consume. Scale infinitely. Never over-provision again.

The pitch was on the money, but it was not on the money. Elasticity resolves over provisioning only in case of governed consumption. In most enterprises consumption was not regulated. With the switch from capital expenditure to operating expenditure, the natural brake that hardware budgets always provided was eliminated. During the days of warehouses, a CIO had to make the case for a new server. With the advent of the cloud, an analyst can start up a compute cluster at midnight, without permission, and incur a five-figure charge before breakfast. The architectural change moved risk from the vendor to the customer and most customers failed to put in place the governance machinery to manage the risk. This is what makes up the graveyard. Three waves of investment, each built upon the previous one, and retention policies, ownership and ROI systems that never got a handle on the technology.

4. CURRENT TRENDS

4.1 What the 2026 Enterprise Data Landscape Looks Like

There are several patterns that characterize the situation of enterprise data platforms today, and it's important to understand these patterns to see why the graveyard problem is worsening instead of getting better.

The first one is the generative AI overlay. After the period of rapid growth of large language model adoption in 2023 and 2024, practically every major data platform vendor introduced an AI add-on. New platform stacks were added to existing ones, including Snowflake Cortex, Databricks Mosaic AI, Google Vertex AI, AWS Bedrock integrations and standalone vector databases like Pinecone and Weaviate. These additions were approved by the enterprise-level AI mandates in most enterprises, and sizing was done based on the anticipated enterprise-wide deployment. The utilization rate has always been less than 10% of the capacity. The infrastructure is still there, the bills are still there and the per-user cost is oftentimes higher than SAR 10,000 per year for an application that can be delivered on-demand for a fraction of that cost.

The second trend is towards the convergence of lakehouse architectures. The line between data warehouses and data lakes has become more indistinct, with Databricks championing the idea of the lakehouse, Snowflake adding support for Iceberg tables and open table formats like Apache Iceberg and Delta Lake becoming popular. This meeting is a good thing for architecture. It's also business friendly for vendors, as it will allow them to bid on workloads that they would not have been able to host before, and to

sell additional capabilities. Very few businesses have taken advantage of the convergence to make things easier. The majority have deployed it to run parallel platforms and Snowflake **warehousing**, Databricks **data science** duplicate storage, governance and operational overhead.

What the 2026 Enterprise Data Landscape Looks Like

More technology. More possibilities. Also more complexity, cost and risk. Here's a simple, clear view of what's happening.



Fig -3: The 2026 Enterprise Data Landscape

The third trend is that FinOps is becoming a formal discipline. More than 10,000 people are now members of the FinOps Foundation. Native vendor cost managers, Apptio Cloudability, CloudHealth and Vantage have all come of age. But FinOps is not even in data platform scenarios. Many organisations have cloud FinOps functions but are not applying it to the Snowflake credits or Databricks DBUs as these bills are outside of the regular cloud billing stream.

The fourth trend is a regulatory crackdown. The Personal Data Protection Law (PDPL), which is controlled by Saudi Arabia's National Cybersecurity Authority (NCA), the Saudi Central Bank (SAMA) cyber framework and the Zakat, Tax and Customs Authority (ZATCA) e-invoicing mandates all have data lineage, retention, and audit requirements. These regulations are leading to investments in governance tooling, but in most cases that tooling is implemented without the operating model, and so is a high cost shelfware and not a real governance uplift.

The fifth trend and perhaps most significant is the unobtrusive change in boardroom discussions. From 2024 to 2025, audit committee questions shifted from **What is our AI strategy.** to **What is the measurable return on our existing data investments.** Finance leaders will no longer sign off on renewals of platforms based on ingestion metrics or the number of models. They demand business outcome measurement and the majority of data organisations are not structured to provide it.

5. THE VENDOR PLAYBOOK HOW THE METER KEEPS RUNNING

Today's data platform vendors aren't the bad guys. They are rational commercial entities with a pricing mechanism which is advantageous to them if the customers are consuming without governance. That model and how it works is crucial for any finance leader trying to right-size the spend.

The illusion that is paid for as you go is the first. Consumption-based pricing is a selling point as it is cost efficient in that the customer only pays for what they use. In fact, if **use** is not regulated, the consumption goes up without any apparent limit. Each query, each ingestion job, each machine learning experiment, each dashboard refresh will cost you money. No automatic limit, no default alarm, no accountability process from the engineer running the query to the invoice that gets to the CFO's desk.

The second one is the separation of storage and compute. It's a beautiful design. Commercially it is two meters that are operated in parallel. The storage continues to grow indefinitely, as deletion is weak and governed by governance. Compute has an issue of growing per query, as analysts continually run unoptimized SQL against the same tables day after day.

The third one is schema-on-read doctrine. There is always a push to take all the data and determine structure later, from the vendors. **Later** rarely arrives. The data is stored in raw tables and is paid for, but is never deemed important enough to do the modelling work to make it analytically useful.

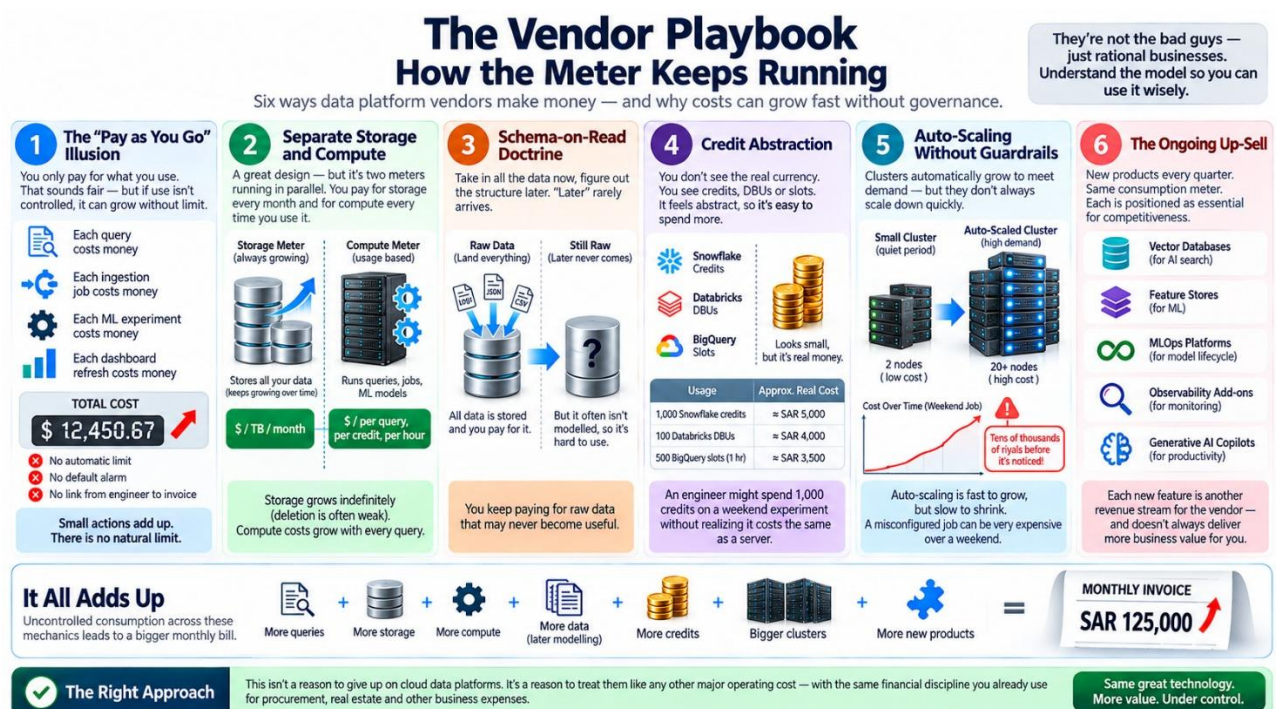


Fig -4: The Vendor Playbook

The fourth is called credit abstraction. Snowflake credits, Databricks DBUs, and BigQuery slots are abstract units, which take the developers away from the real currency cost of their actions. An engineer that would refuse to purchase a server for SAR 5,000, will eat up 1000 credits on a Saturday night experiment without realizing that he has just spent the same price on a server.

The fifth one is auto-scaling, without guardrails. Automatic scaling up of compute clusters to meet demand. They are not usually as aggressive when it comes to reducing in size. A wrongly set up job can rack up tens of thousands of riyals over a weekend, before it's spotted when the monthly bill is sent out.

The 6th mechanic is the ongoing up-sell. New product category added each quarter, with the same consumption meter. Vector databases. Feature stores. MLOps platforms. Observability add-ons. Generative AI copilots. Each is viewed as table stakes for competitiveness and each is another revenue stream for the vendor without necessarily having a corresponding increase in the business results for the customer.

Understanding these dynamics does not mean that cloud data platforms should be given up. It's an argument for treating them like any other big operating cost that businesses already use financial discipline on, such as procurement, real estate.

6. THE TWENTY SCENARIOS OF FINANCIAL LEAKAGE

The remainder of this article will take a detailed look at 20 specific patterns of financial leakage that have been seen in enterprise data platform deployments. Every pattern contains a description of a setting, what actually happens in the setting, the measured leaks and the organisational process which maintains the pattern.



Fig -5: The Twenty Scenarios of Financial Leakage

6.1 Scenario One- Ingestion Without Purpose

In 2022, a regional retail bank created a Snowflake data lake and loaded data from forty-seven source systems, such as core banking, cards, loans, CRM, call centre transcripts, mobile app telemetry, ATM activity feeds, branch footfall data, social media mentions and credit bureau feeds. In 2026, just 11 of those 47 feeds



have a dashboard or model that a business user uses. Thirty-six pipelines are in operation every day, using compute and storage and they are based on the **just in case** governance principle.

The leakage is easily calculated. With 36 unused pipelines, at an average cost of SAR 18000 per pipeline each year, for ingestion compute, storage of landed data and pipeline maintenance overhead, the cost of data that nobody queries is SAR 648,000 per year. This is consistent since data onboarded is the measurement for the data engineering function, not data consumed. There's no auditability of the relationship between an ingestion job and a downstream dashboard and pipelines just pile up like unopened mail.

6.2 Scenario Two-The Zombie Dashboard Farm

The business intelligence team built three hundred and forty Tableau or Power BI dashboards over four years. Each dashboard is backed by scheduled queries that refresh underlying data. Usage telemetry reveals that two hundred and twenty of these dashboards receive fewer than five views per month, and eighty of them have zero views in the past ninety days. The scheduled queries behind them continue to execute, warming compute clusters daily. Two hundred and twenty zombie dashboards at SAR 2,400 per year each in refresh costs equals SAR 528,000 annually spent maintaining visibility that no one consumes. The dynamic that sustains this is the absence of lifecycle management for analytic assets. Analysts leave the organisation and their dashboards remain. Nobody has the authority or incentive to decommission.

6.3 Scenario Three-The Over-Provisioned Warehouse

When the platform was launched, the default Snowflake warehouse size was Large, which used eight Snowflake credits per hour, to guarantee good performance as the platform was rolled out. The default was never to be looked upon again. 85% of queries run against this warehouse would run at the same speed against a Small warehouse, for two credits per hour. Large warehouses are spun up frequently and are used to run 10 row lookups. Typically, this over-sizing wastes 40 per cent or about SAR 264,000 per year in compute that is four times the size needed for an enterprise that is spending sixty thousand credits per year, at SAR 11 per credit. This goes on because there is no central control over warehouse sizes, devs optimize for query performance and no one brings cost-per-query metrics to the attention of the devs writing the queries.

6.4 Scenario Four-The Auto-Suspend That Does Not

The main way to control idle compute cost is to have Snowflake warehouses automatically suspend when they are not used, after a short period of idle time. In most deployments, a small number of warehouses (usually 10-20) are configured to not auto-suspend or to suspend them every 60 minutes or longer, under the notion that a batch job might come in. Those warehouses are warm and are working 24/7. Fourteen warehouses with average idling time of twenty hours per day, three hundred and sixty five days per year, three credits per hour and SAR 11 per credit, consume about SAR 2.2 million per year. If 50% of this can be recovered by proper suspend policies, SAR 1.1 million will be freed. This situation remains unchanged, as the configuration of warehouses is done by individual engineers and not reviewed centrally.

6.5 Scenario Five-The Databricks Cluster That Never Dies

A data science team is using an all-purpose cluster on Databricks with 20 worker nodes for running machine learning experiments. The cluster has an auto-termination policy in place, however it is set for 2 hours. In real life, it is manually lengthened by the engineers before they go for lunch, meetings, or night experiments. Audit logs were examined over an eleven month period and revealed that the cluster was operating 22 hours a day, on average, and that the notebooks were being used 3-4 hours a day. With 18



idle hours per day on 20 nodes at SAR 4.5 per node-hour, there are some 330 active days per year, leading to idle compute of about SAR 534,600 per year. The situation is cultural. Data scientists don't like having to wait for their clusters to start up and finance doesn't see the daily charges for clusters.

6.6 Scenario Six-Storage That Grows Forever

The data lake was created in 2019, and the raw transactional data was ingested, with the policy of storing everything, and determining retention at a later stage. The lake will contain 6.8 petabytes of data, with 55% annual growth, by 2026. Query log analysis reveals that 92% of queries are for data under 90 days old and only 1% of queries are for data older than 2 years is queried more than once every quarter. About 4.2 petabytes of effectively cold data is stored on hot **active** storage at about SAR 92 per terabyte per month, or SAR 4.6 million per year. This data can be transferred to an archive or cold tier, and the storage cost will be reduced by 85 per cent, freeing up SAR 3.9 million per year. This is the pattern because the ownership of the retention policy is unknown, legal counsel is **keep everything** and there is no query pattern analysis against the retention tiers.

6.7 Scenario Seven-Development Environment Sprawl

The data engineering team consists of 24 data engineers, each with their own personal Snowflake account schema to use for development. Full clones of production data are stored in each development environment for realistic testing and clones are rebuilt every week. Development warehouses are open during business hours but may not be in operation. Twenty-four environments at SAR 3,800 per month **Compute and Storage** = SAR 1,094,400 per year for environments that are used, say, 20% of the time. This trend continues because it's easier for developers and it was sold as free storage **zero copy cloning**, but not free compute to query it.

6.8 Scenario Eight-The Machine Learning Model That Never Shipped

The data science team took 18 months to develop 14 machine learning models for Customer Churn Prediction, Fraud Scoring, Next Best Offers, Inventory Forecasting, and other models. The total amount of SAR 2.8 million was spent on training models on platforms. There were fourteen models of which eleven never went into production. They are stored in Databricks notebooks and are technically valid but are not currently being used. Those three that have shipped have only marginally measurable business value, and none has an associated ROI measurement. SAR 2.2m of model development compute went unused, and there is a cost of SAR 300k or more per year for unused storage of abandoned training datasets, feature stores, and experiment logs. The pattern continues since the measure of the performance of data science is the creation of models, not the deployment of models and the measurement of their performance.

6.9 Scenario Nine-The Query That Scans Everything

An analyst issued a query that returns all the data from a transactions table with 4 billion rows of data, where the customer_id is equal to a particular value. There is no cluster partition key on the table, so each execution scans the entire table. The query is used around 1200 times a month in the dashboards and ad-hoc analyses. The monthly price of this single query pattern is SAR 102,000 (at SAR 85 per full-table scan) or SAR 1.2 million per year. With a proper clustering key, this cost can be reduced by about ninety five per cent. This continues as developers still do not have the visibility of costs at the query level and there is no standard query performance review.

6.10 Scenario Ten-The Vendor Lock-In Premium

The organisation signed a three-year contract with Snowflake, with an annual consumption of eighty thousand credits, at SAR 11 per credit. The actual consumption ended up at 55,000 credits per year. Pre-



purchased credits expire or roll over at a decreasing rate. The cost for similar workloads on BigQuery or managed Databricks is normally 25–35% less. The total cost of unused SAR 11 credits for 25,000 per year is SAR 275,000 per year of commitment waste. The above-market premium of SAR 3 per credit (on fifty-five thousand actual credits) is another SAR 165,000 per year. The total leakage for the two years is SAR 440,000. This is because CIOs are often forced to enter multi-year contracts by vendor pressure to lock in rates and downside consumption scenarios are not often modelled prior to signing.

6.11 Scenario Eleven–The Data Quality Tax

The data lake takes data as-is from source systems and data quality will be done downstream by analysts or data engineers. Some 60% of data engineering time is spent on cleaning, deduplicating and reconciling data which could have been corrected at source. There are several pipelines for the same quality problems to be solved repeatedly. SAR 480,000 per year, 12 data engineers at 60% effort and SAR 3.5 million per year in human effort to fix the same problems. The cost of a compute for cleansing jobs is an additional SAR 400,000 or more. This is the case because the data lake fails to enforce any data quality requirements and the source system owners are not held accountable for any data quality issues.

6.12 Scenario Twelve–The Shadow Analytics Problem

Even with the investment of SAR 15 million in Snowflake and Databricks, around 40% of business decisions are still based on Excel spreadsheets that are managed by analysts and have to be queried to get the data. The data lake has turned into a query engine for exports to Excel and an expensive one at that. Often, curated semantic layers and ML pipelines are skipped, as well as governed data products. All finance, sales operations and marketing is done in Excel. The value proposition of the platform which is self-service analytics, single source of truth and real-time insights has not come about. Let's say that an enterprise is spending SAR 15 million a year, and that SAR 10 million is unrealized value, not leakage, but the impact on the economy is the same. This continues to be the case because business user change management was never funded. The building of the platform. Adoption was not.

6.13 Scenario Thirteen–The Egress Charge Shock

The data lake is located within AWS. Data downstream applications, data science workbenches, and reporting tools can access the data in the lake through Azure, Google Cloud, and on-premise systems. AWS data egress fees are stealthy, as they are occurring on hundreds of cross-cloud moves, and the amount of data sent in egress is well over 450 terabytes per year. Egress fees are currently around SAR 0.34 per gigabyte, which equates to around SAR 156,700 per year a cost line not included in the original platform business case. This pattern continues with the fact that cross-cloud architecture decisions were taken without considering cost modelling and the egress costs hit network cost centres, which are outside the data platform owner's budget.

6.14 Scenario Fourteen–The Catalogue and Governance Shelfware

The organisation acquired a data catalogue and governance platform, which usually cost SAR 1.2 million per annum, either Collibra, Alation or Informatica. Of the 1500 people who are potential users, less than 50 log in every month. 70% of the data domains do not have data stewards. The catalogue only has auto-crawled technical data, no business context, descriptions, or ownership tags. What this means is SAR 1.2 million a year for a governance tool that provides no governance and no adoption. The pattern continues as the tool was bought prior to the governance operating model being established and there was a belief that technology would influence behaviour. It never does.

6.15 Scenario Fifteen–Real-Time Streaming That Is Not

The combined infrastructure cost for reference architecture is SAR 1.8 million per year, which consists of the real-time ingestion **Kafka, Kinesis, or Event Hubs** and processing **Spark Structured Streaming**. Only two streams are consumed in true real-time usually fraud detection and mobile app events, out of all the streams that are configured. The rest are aggregated downstream into hourly or daily buckets, so streaming pipelines don't offer any additional benefit over less expensive batch pipelines for these types of workloads. About 80% of streaming infrastructure spend is for use cases that could be implemented on batch for 15% of the cost, totaling SAR 1.2 million per year, for over engineered real time plumbing for batch tolerable workloads. This is because real-time was considered as an architectural principle at the beginning and has never been revisited based on the cadence needs of the business.

6.16 Scenario Sixteen–The Open–Source Fallacy

The on-premise Hadoop cluster, typically Cloudera CDP, was seen as a cheap open-source alternative to the cloud data warehouses and placed at procurement. The total cost of ownership which includes hardware refresh cycles, data centre space, networking, Cloudera subscription fees, a fourteen person platform team, patching and security operations is more than SAR 9 million per year, whereas the workloads would cost SAR 4 to 5 million per year on managed cloud platforms. The open source alternative turned out to be the costliest. The size of the cluster is for peak demand, with an average utilisation of twenty-five per cent. Cloud elasticity can save on-premise overheads of approximately SAR 4 million per year. That trend continues with sunk cost reasoning, hardware depreciation accounting, the desire for the platform team to maintain itself and the lack of a new TCO baseline since the initial decision in 2020.

6.17 Scenario Seventeen–The Vector Database and Generative AI Overlay

After the generative AI craze of 2023 and 2024, the organisation deployed vector database infrastructure **Pinecone, Weaviate or pgvector** and then deployed large language model capabilities **Databricks Mosaic AI or Snowflake Cortex**. Three prototypes of applications were developed. One is an in-house document search system, with about eighty users, and is in limited production. The vector database infrastructure and LLM were designed for enterprise-wide deployment which didn't happen. The cost of over-provisioned GenAI infrastructure **per year** is SAR 1.4 million, and serves less than one hundred users, resulting in an effective cost of SAR 14,000 per user per year, as opposed to right-sized or on-demand infrastructure at approximately SAR 1,200 per user per year. That's because the GenAI investment was forced on board and it would be politically unpalatable to reduce the investment.

6.18 Scenario Eighteen–The Compliance and Audit Burden

Detailed Data Lineage, Access Logs and Audit Trails are required by regulatory obligations, such as SAMA, NCA, PDPL and ZATCA. These are automatically created by the data platform. Around 180 terabytes per year are needed for storing audit logs. Lineage tracking jobs are always running and are consuming compute. Compliance infrastructure accounts for about 12% of the overall platform cost but isn't a line item. Around SAR 600,000 of the SAR 1.8 million per year spend on compliance could be optimised by implementing audit log tiering, log consolidation, and improving the efficiency of the scheduling of the lineage. This is because compliance is never questioned its cost is never discussed and **we have to keep everything** becomes a dogma and not a design parameter.

6.19 Scenario Nineteen–Platform Team Oversizing

Typically, a large enterprise has 29 FTEs (one Head of Data, three Data Architects, twelve Data Engineers, six Data Scientists, four BI Developers and three DataOps or Platform Engineers) running the data lake, with an average fully-loaded cost of SAR 480,000 per FTE per year, resulting in people cost of SAR 13.9 million per

year. The cost per real business decision for the data organisation for an enterprise with less than 10 real business decisions per month is about SAR 115,000. Even if you add in day-to-day dashboards and operational analytics, the ROI is very negative. It isn't the cost of technology that is leaking here. It is an organisational cost that is based on a platform that doesn't provide proportional value. There would be a typical SAR of 4 to 6 million per year in people cost savings from right sizing to the actual demand. It is the same with data teams, which grow with the platforms, and no one asks the question why should the platform grow.

6.20 Scenario Twenty–The ROI That Was Never Measured

The original business case estimated SAR 60 million worth of benefits over 5 years, which would be achieved by optimizing inventory, operational efficiency, reducing churn, better fraud detection and cross-sell. After four years, nobody has taken any measurements to determine if any of these advantages have been realised. The business case document is stored in a shared drive and has not been opened since go live. The platform has met one of the go-live KPIs and no business outcome KPIs. The total run cost for the entire year is SAR 15–20 million, which is not accounted for against the original benefits. Even if only half of the anticipated benefits were realized, it represents an opportunity cost of SAR 6 million a year to the enterprise, in addition to the direct cost of the platform. This is because nobody wants to measure something that they have been a big advocate for, the sponsoring CIO is usually either gone or in another role by now, and the CFO has quietly reclassified the spend as the cost of being modern.

7. THE CONSOLIDATED FINANCIAL PICTURE

The 20 scenarios can be combined to create a defendable leakage range for a typical large enterprise in Saudi Arabia with a mature cloud data platform.

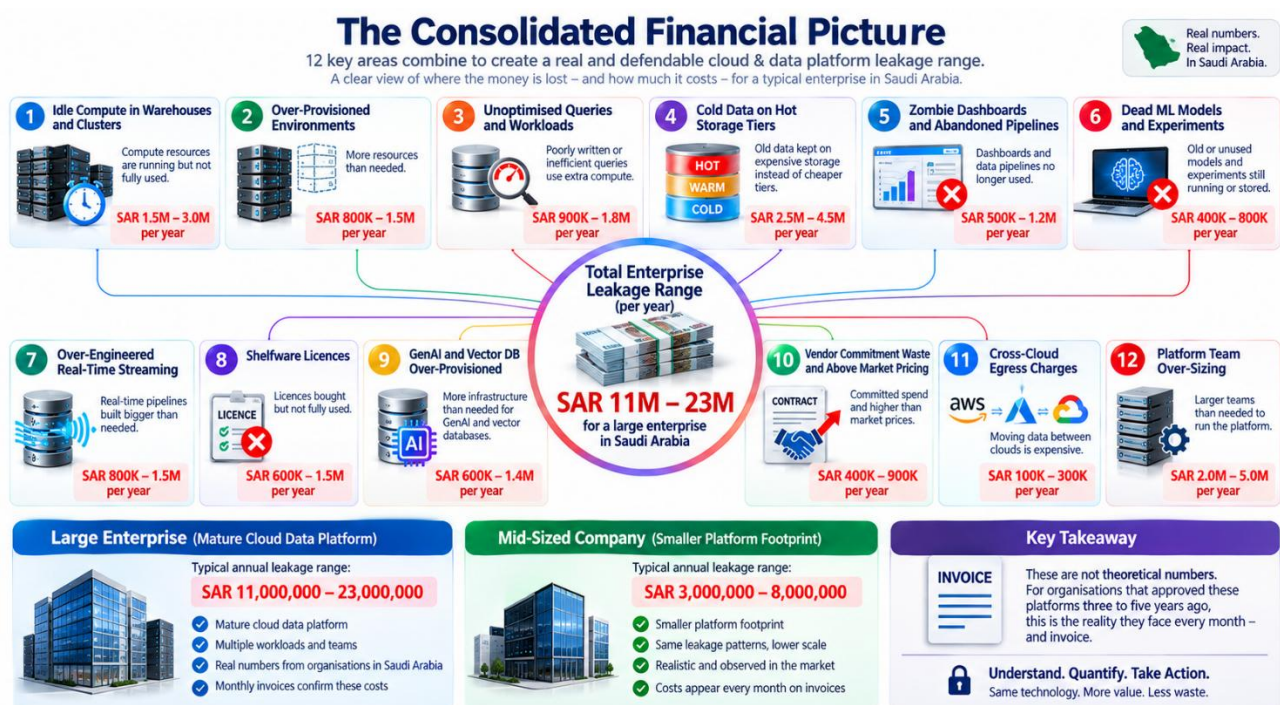


Fig –6: The Consolidated Financial Picture

Typically, idle compute in warehouses and clusters is between SAR 1.5 million and 3.0 million annually. Over-provisioned environments provide SAR 800,000 1.5 million. There are SAR 900,000 to 1.8 million for unoptimized queries and workloads. Cold Data stored on hot storage tiers count towards SAR 2.5m to 4.5m. Zombie dashboards and abandoned pipelines cost SAR 500,000 to 1.2 million. Dead machine learning models and experiments make SAR 400,000 to 800,000. Real-time streaming is over-engineered, and costs SAR 800,000 to 1.5 million. Shelfware licenses are worth SAR 600,000 to 1.5 million. GenAI and vector database infrastructure is over-provisioned and costs SAR 600,000 1.4 million. The contribution from vendor commitment waste and above market pricing is SAR 400,000-900,000. Cross-cloud egress charges add SAR 100,000 to 300,000. Platform team over-sizing gives SAR 2.0 million-5.0 million. A big company's leakage range per year is between SAR 11,000,000 and SAR 23,000,000. The leakage in mid-sized companies with smaller platform footprint ranges between SAR 3 million and SAR 8 million. These are no theoretical numbers. For organisations that approved these platforms three to five years ago, they are the reality that they face every month and invoice.

8. CASE STUDIES IN SUCCESSFUL DATA PLATFORM DEPLOYMENT

While it's important to consider failed deployments to understand enterprise data platform economics, it's equally important to consider those organisations that have derived significant and measurable value from the same technologies that were discussed elsewhere in this article. In all sectors and regions, there are three trends that can be seen in successful deployments that are not seen in graveyard deployments. These trends can be seen in global financial services companies, energy giants, digital native companies and tier one banks throughout the Gulf region.

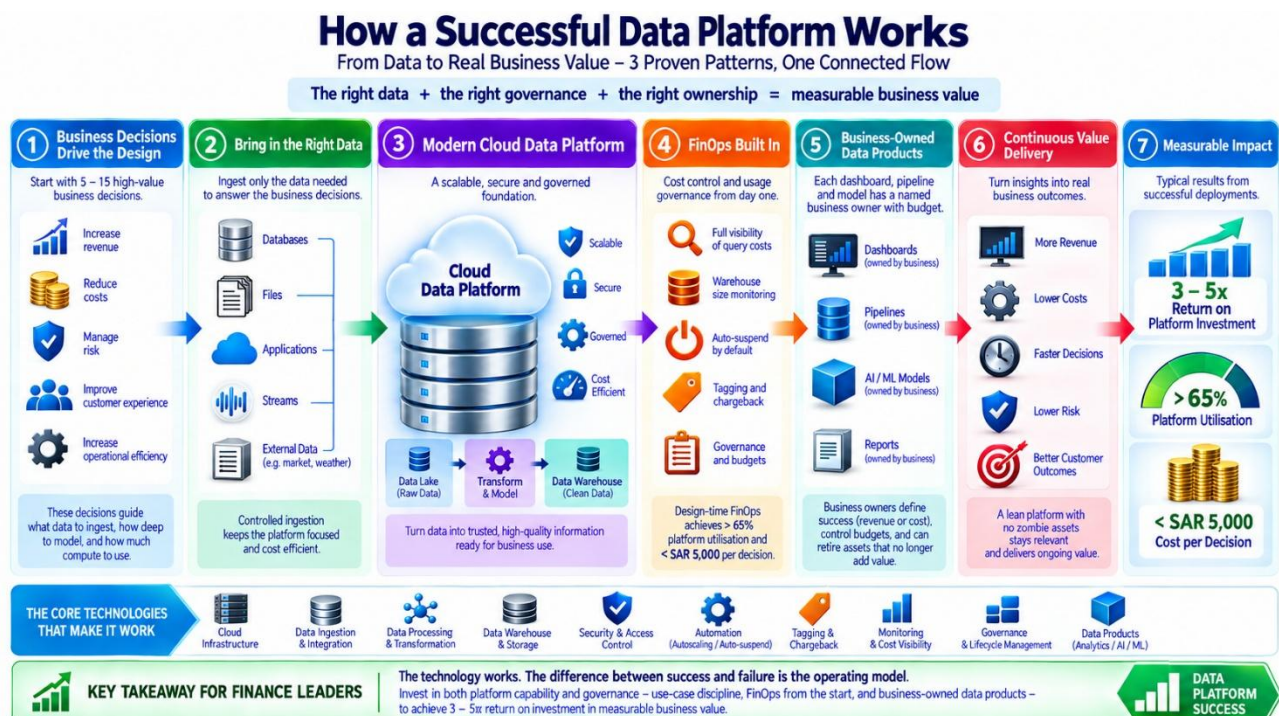


Fig -7: Successful Data Platform Works



Pattern One: Use Case Led Architecture. The first distinguishing pattern is that of architectural discipline based on stated business decisions. Successful businesses start with a small number of high value business decisions **typically 5 to 15** and work their way backwards in designing the data platform. The platform is designed to support certain decisions that limit the ingestion scope, modelling depth, warehouse sizing and compute allocation. Expansion is only done when there is a need for new capacity due to a new decision. This use case-driven approach ensures that it does not become an underperforming deployment with uncontrolled ingestion of data that outpaces the need for analysis.

Pattern Two: FinOps Embedded from the start. The second pattern is to incorporate consumption governance as part of the basic architectural requirement, not as a reactive measure when costs are going out of control. Organisations that build FinOps capability at platform start-up consider query cost visibility, warehouse sizing defaults, auto-suspend, tagging taxonomies and chargeback mechanics to be design time decisions, not operational afterthoughts. Such businesses usually achieve platform utilisation rates above sixty-five per cent, have cost per decision less than SAR 5,000 and do not have any structural waste that occurs when financial discipline is added on after scaling. It's not that FinOps tooling is enough, it's that the FinOps mindset of governance must be there before the first production workload.

Pattern Three: Data Products in the Business. The third pattern is the business owner(s) not the technology owner(s) are assigned to each dashboard, pipeline and model. Successful deployments have the following characteristics. The business owner has budget control, they have a revenue or cost definition of success, and has standing authority to decommission assets that don't meet their stated purpose. This governance framework helps to ensure that there are no **zombie assets** in which ownership is fragmented, success is not clearly measured and no one is empowered to end suboptimal investments. If the ownership of the business is well defined, there is no need to add unproductive assets, as they can be retired at the same time productive assets are added, thus keeping the platform lean.

Finance Leaders' implications. All of these patterns suggest that the data platform itself is seldom the cause for value erosion. It is the operating model associated with the platform that makes the difference. Companies that invest as much in platform capability as they do in platform governance regularly deliver 3-5x the value of platform investment in measurable business outcomes in line with the best in class benchmarks mentioned earlier in this article. The takeaway for finance leaders isn't to cut back on investments in modern data platforms, but to make it a prerequisite for further investment in platform spending that they are able to demonstrate the maturity of their governance, such as use-case discipline, FinOps embedded in the platform, and business-owned data products.

9. CHALLENGES IN ADDRESSING THE PROBLEM

Even when the numbers are in sight, the leakage remains elusive for enterprises, due to several structural problems.

The first challenge is of an organisational nature. Data platforms are usually owned by infrastructure teams, data engineering teams, business intelligence teams and data science teams. There is no one accountable executive who is responsible for the end-to-end cost-to-value ratio. When it is not owned, waste builds up as no one is being specifically rewarded for the removal of waste.

Challenge 2 is cultural. The volume and sophistication of data is a measure of prestige within data organisations. Consuming more sources, creating more pipelines, training more models, deploying more streaming infrastructure all sound like good performance on performance reviews. Reducing footprint, de-

commissioning assets and not building what the business doesn't need are not often hailed as successes and are often punished.

The third challenge is political one. The platform was approved by top executives who took it to board. It's not easy to admit that it under delivered, professionally. The standard answer is to postpone the reckoning, throw in another ability, and cross your fingers that the next initiative is worth the previous.

The fourth difficulty lies in the technical aspect. To be able to allocate cost to the business outcomes of the data, you need end-to-end lineage from the data being ingested through the various layers of transformation to the downstream dashboards and models. Not many organisations have this legacy in the form of operations. There is no lineage and it is hard to do a rigorous cost benefit analysis, and so it's a risky decision to decommission.



Fig -8: Inside the Technology From Data Source to Business value

The fifth challenge is contractual. Multi-year contracts with Snowflake, Databricks or hyperscaler marketplaces also require minimum spend. If consumption is less than the commitment level, then the consumption is not reduced from the invoice. This can be a short term disincentive to optimised, despite the fact that it may be beneficial to the enterprise in the long term.

10. SOLUTIONS THE RECOVERY ROADMAP

For a large enterprise, it's realistic to expect SAR 4 million to 12 million per year to be recovered in a disciplined recovery programme, in 6 to 9 months with no disruption to business users. The roadmap is broken up into five phases.

10.1 Phase One–Baseline the Reality

Export consumption report from Snowflake, Databricks, and cloud providers giving the consumption of the last 12 months at warehouse, user, query, and workload level. Map each ingestion pipeline to an ingestion consumer downstream. Correlate all dashboard to real usage telemetry for more than 90 days. Develop data-to-decision end-to-end lineage. This phase results in one document that contains all the data assets, the cost of the data asset per month, the data consumer, and the measurable business outcome.

10.2 Phase Two–Kill, Archive, Optimise

Remove zombie dashboards **with no views**, abandoned pipelines **with no downstream consumers**, dead ML experiments and unused development environments in 90 days. Transfer infrequently accessed and cold data (**data that is older than 12 months**) to glacier or archive storage tiers for a 60% to 70% cost savings. Optimize the size of warehouses and clusters. Enforce aggressive auto-suspend policies. Implement Developer Level Query Cost Visibility. Cluster top 20 biggest tables and partition the table.

10.3 Phase Three–Govern Consumption

Implement a FinOps platform, such as Apptio Cloudability, CloudHealth, Vantage or a native vendor cost manager. Set budgets per team and limit the consumption per month. Implement chargeback or show back per business unit to make cost transparent where decisions on consumption are made. Require jobs over a certain threshold to do query cost review. Have a monthly Data Platform Review Board meeting with finance and technology business representation.

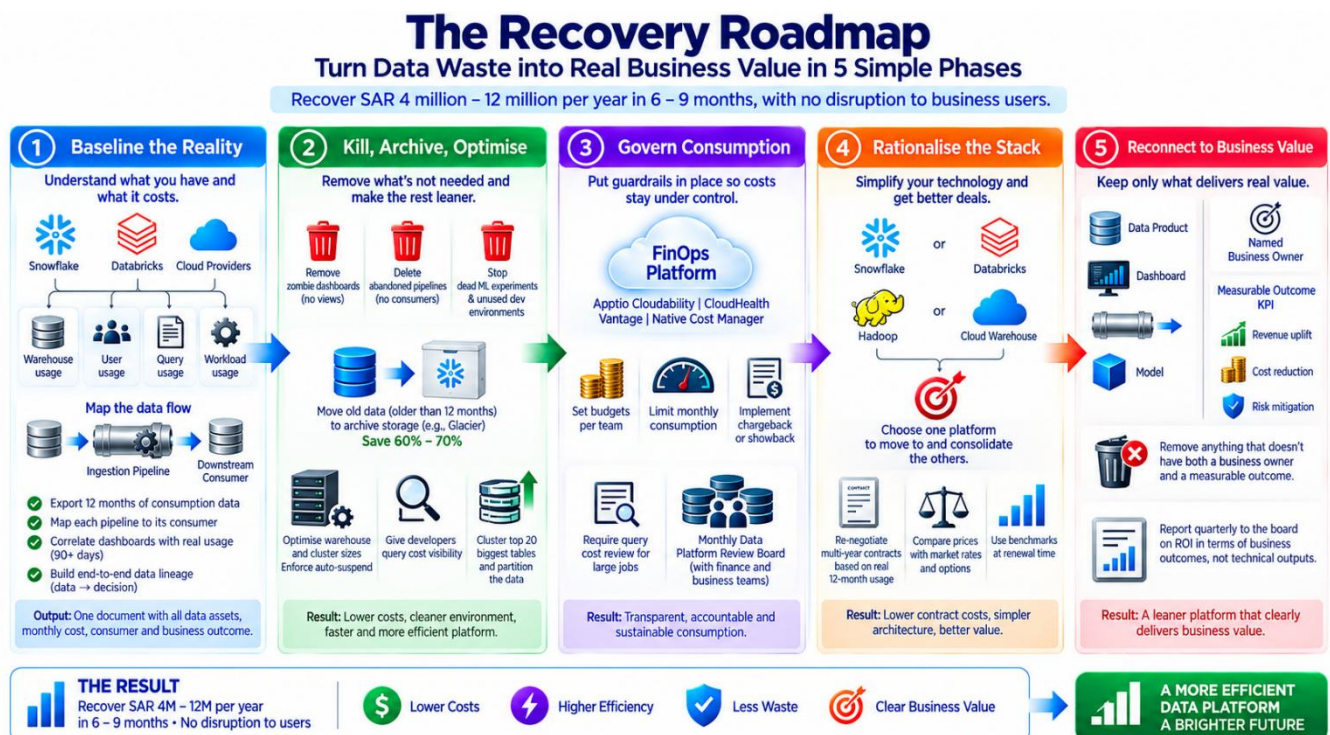


Fig –9: The Recovery Roadmap

10.4 Phase Four–Rationalise the Stack

If there are parallel platforms, like Snowflake and Databricks, or both Hadoop and cloud warehouses, choose one platform to move to and move the others. Re-negotiate multi-year contracts based on the actual 12-month consumption and not on the aspirational consumption. Compare prices with other market rates and options. Leverage the benchmarks at renewal time.

10.5 Phase Five–Reconnect to Business Value

For each of the data products, dashboard, pipeline or model that still exists, specify a named business owner and a measurable outcome KPI in the form of revenue uplift, cost reduction, or risk mitigation. Throw out anything that doesn't have both. Quarterly report to the board ROI of the platform in terms of business outcomes and not technical outputs.

11. BENCHMARK METRICS EVERY CFO SHOULD DEMAND

There are a few key metrics that can be requested during the quarterly business review, which shift the focus from cost of the platform to the value of the platform.

The best indicator of cost is cost per active business decision. The best in class enterprises have a SAR 2,000 to 8,000 per decision. Dysfunctional platforms are SAR 50 000 or greater per decision. For a healthy data organisation, the percentage of platform costs in the total IT spend is in the range of 8-12%. If the ratio is 20% or more but the business output is not proportionate, then it is a case of over-investment. The average utilisation **warehouse or cluster usage** should be in the 65-75% range. Below 40% is considered to be a severe over provisioning.

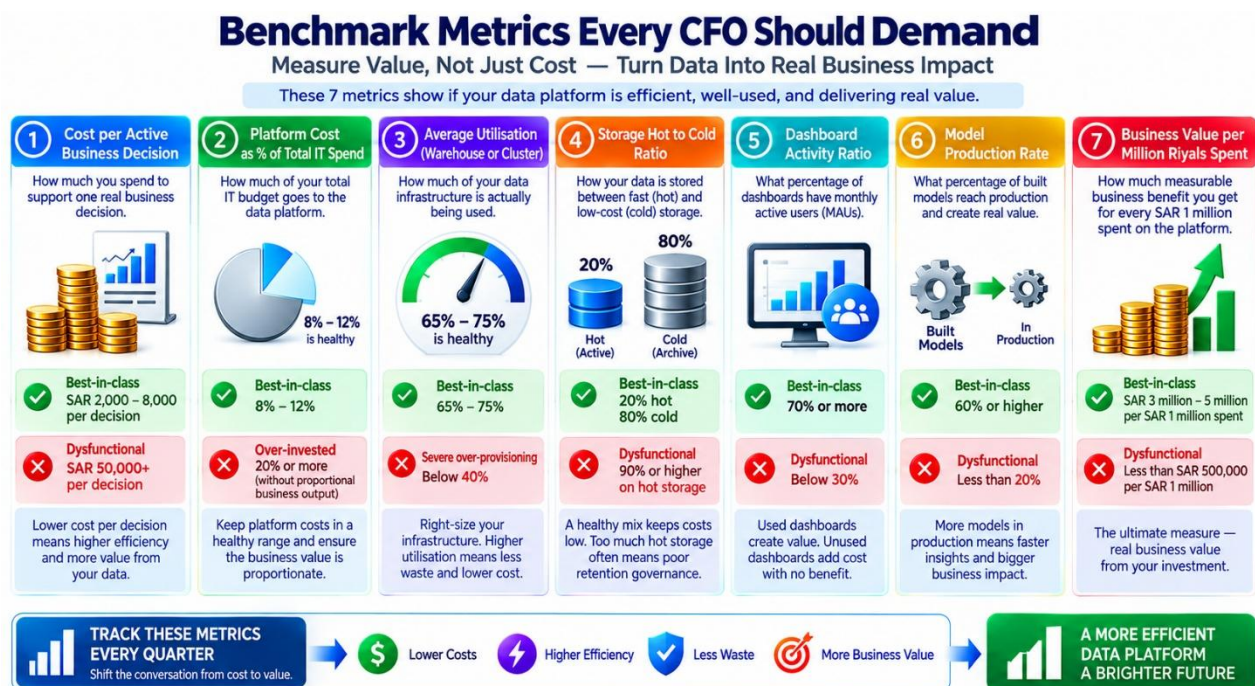


Fig -10: Benchmark Metrics Every CFO Should Demand

The amount of storage hot to cold should be around 20% hot and 80% archive or cold. The inverse ratio 90% or higher on hot active storage is a good indicator of lacking retention governance. The percentage of

dashboards that have monthly active users **MAUs** is known as dashboard activity ratio. Organisations with a score of 70% or more are best in class. Below 30% are dysfunctional organisations. Model production rate is the percentage of built model which reaches production. The best in class is defined as 60% or higher. Dysfunctional is less than 20%. The final ratio, business value per million riyals spent, is a measure of the ultimate ratio. Best-in-class is able to provide SAR 3 million to 5 million of measurable business benefit for every SAR 1 million that is spent on the platform. Dysfunctional provides less than SAR 500,000 on every SAR 1 million.

12. THE QUICK-WIN RECOVERY PLAN

A sequenced and phased plan yields predictable and measurable results for enterprises that need to see financial recovery in the short term within a quarter. It is intentionally designed to provide value in waves starting with the highest impact, lowest risk interventions, and then progressing to interventions that involve more coordination with engineering and procurement. The following sequence is broken down into six phases of execution which can be done in sequence over a quarter or be compressed and parallelised based on organisational capacity.

Phase 1 Compute Baseline and Auto-Suspend Correction. The initial phase is to create a baseline on the fact of consumption. Download 90 days' worth of warehouse and cluster activity and determine the top 10 cost drivers and warehouse(s) that have disabled or extended auto-suspend settings. This stage of corrections are at the configuration level and are relatively low risk. The annualized recovery is estimated to be SAR 500,000– SAR 1 million.



Fig –11: The Quick Win Recovery Plan



Phase 2 Dashboard Suspension in Dormant. The second phase is directed towards waste on the consumption side. Find dashboards that have not had any views for the last 90 days and pause the refresh for these dashboards. Business owners should be informed, but not consulted, as the fact that it is being used is justification enough. Annualized recovery is expected to be SAR 200,000 – SAR 400,000.

Phase 3 Top Cost Drivers Query Optimisation. The third phase deals with the efficiency at an engineering level. Find the top 20 most costly recurring queries and hand them over to data engineering for optimisation, using one of the following techniques: clustering, partitioning or result materialization. This stage needs technical ability, and provides compounding returns. The recovery is estimated to be between SAR 400,000 and SAR 800,000 per year.

Phase 4 Data Retention and Archival Review. The fourth phase is on rationalization of storage. Archive or cold store data older than 24 months and infrequently accessed **less than once every quarter**. Data recovery is usually the biggest lever available, although regulatory retention requirements will need to be confirmed prior to moving any data. Annual recovery expected SAR 1million–SAR 2million.

Phase 5 SaaS Tooling & Shelfware Audit. The fifth phase is directed towards the surrounding ecosystem. Audit all SaaS Data Tooling such as catalogues, governance platforms and observability products. Eliminate underused shelfware **less than 20% adoption** and integrate redundant functionality as appropriate. SAR 500,000 – SAR 1.2 million is expected to be recovered annually.

Phase 6 Vendor Commitment Redo. The last stage is related to the commercial layer. Re-negotiate vendor commitments to a 12-month actual consumption basis, rather than forecasted peaks and seek reserved capacity, committed-use discounts or multi-year pricing where consumption is predictable. The annualized recovery is expected to be between SAR 300,000 and SAR 800,000.

Expected Cumulative Outcome. For a large enterprise, the total amount of expected recovery over the entire recovery cycle is SAR 3M–SAR 6M per annum. The plan aims to achieve these outcomes without disrupting business users, without making compromises to existing analytical capability, and without having to purchase new technology. Equally significant is the sequencing, which institutionalizes cost discipline well beyond the initial recovery cycle, and sets up a repeatable governance rhythm.

13. FUTURE PROSPECTS WHERE THE DATA PLATFORM MARKET IS HEADING

There are three key directions that will influence the next round of enterprise data platform investments and each one has an impact on how finance leaders should make their current investment decisions.

The first change is that of open tables and interoperability. Technical cost of migration from one platform to another is being reduced by Apache Iceberg, Delta Lake and Apache Hudi. In practice, this means that it diminishes the vendor lock-in and boosts enterprise negotiating leverage when it comes time to renew. Finance leaders must demand new workloads to be open format and that the workloads are portable as a procurement requirement.

The second change is that of adding LLMs features directly into the SQL query experience. Natural language query over governed data is becoming a reality with tools like Snowflake Cortex, Databricks Genie, and others. Adoption, if successful, could help to close the gap in the shadow analytics problem outlined in Scenario Twelve. If the adoption is not successful, then the GenAI overlay cost outlined in Scenario Seventeen will remain. The direction will be more evident through 2026 and 2027 and enterprise-wide commitments should be held back until there is evidence of usage.



Fig -12: Future Prospects Data Platform Market

Third, the FinOps in data organisations is slowly moving toward professionalism. Snowflake and Databricks cost management are expected to be the models created by the FinOps Foundation's data specific working groups. The cost tooling is improving in vendor native. Data platform FinOps will be a regular discipline, similar to how cloud FinOps was established over the past five years from 2018 to 2023. Adopting early will give them a long-lasting cost advantage.

A fourth, and more speculative, change is the place of data platforms in Vision 2030 and other national transformation programmes. The divide between enterprises that are able to make the most of their data platforms and those that are not will become a moat as Saudi Arabia and the Gulf region undergo true data-driven transformation in the financial services, energy, logistics and public services sectors. Over the next year, the organisations that optimised and manage their platforms will be in a better position to compete in the next phase, with cost models that cannot be matched.

14. THE STRATEGIC QUESTION FOR THE BOARD

The question boards should ask is not is the data lake working. The pertinent question is Which business decisions would cease to be taken if the data lake were turned off for 30 days. If the answer is less than 10, then there is no data platform problem in the enterprise. It's a business decisioning issue, masquerading as a data platform issue. More compute, more storage, more ML tooling and more AI copilots will NOT solve a decisioning gap. They will just charge it to the monthly bill.

There's a common thread among organisations that make the most out of their data platforms. They began by taking a few key high-value decisions and backward engineered to the data they needed to



make those decisions successfully. Organisations stuck in data lakes did just the opposite. They ate it all up, and hoped that decisions would come out of the bulk. Choices don't come from quantity. They are born out of clarity, modelling and governance of well-chosen data.

15. CONCLUSION

Most enterprises are not willing to admit that they have a Data Lake Graveyard, which is the most costly technology failure. Doesn't go out with a bang. It does not crash. It's not a big news story. It sucks up SAR 10 million to SAR 25 million annually without a lot of fanfare, and spits out a little drips of information that might have been generated from a well-designed excel model and a competent analyst. The way forward is not to have more data engineers, more vendors, or more AI add-ons. It is comprised of 3 disciplined moves. The platform needs to be reduced to what really matters when it comes to decisions. Consumption needs to be controlled as any other P&L line would be. Reporting of outcomes not technical outputs must be done and measured in terms of business outcomes. SAR 4 million to SAR 12 million per year can be recovered realistically from a large enterprise in 6 months, and the quality of decisions that the platform supports can be improved, through a focused FinOps baselining, decommissioning, rationalization, and governance programme. There is no better data than clean, governed and well modelled data than a petabyte-sized data swamp. In today's day and age, where every riyal spent on running the business is under fire and real data-led transformation is a national goal, the enterprise that right-sizes its data platform will outperform the enterprise that continues to grow its data platform. What's in question is whether the board has received a report that shows that the data platform cost is correlated with business results. The answer is never and the data lake is not an asset. It's a cemetery and it costs \$100 a month.

ACKNOWLEDGEMENT

This article is based only on first hand observations of recurrent patterns of financial leakage, governance deficiencies, and underutilized analytical capability from direct professional experience gathered throughout enterprise data platform engagements. The suggested solutions have been influenced by what has been shown to work in operational contexts, and the scenarios, cost structures, and recovery routes described here represent real world situations observed in the field rather than theoretical creations. The author expresses gratitude to the several platform architects, data engineering teams, financial leaders, and governance professionals whose open discussions and common struggles have influenced this work.

REFERENCES

- [1] Gartner. (2025). Lack of AI-Ready data puts AI projects at risk. <https://www.gartner.com/en/newsroom/press-releases/2025-02-26-lack-of-ai-ready-data-puts-ai-projects-at-risk>.
- [2] Abraham, R., Schneider, J., & Brocke, J. V. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, 49, 424–438. <https://doi.org/10.1016/j.ijinfomgt.2019.07.008>
- [3] Al-Ruithe, M., Benkhelifa, E., & Hameed, K. (2018). A systematic literature review of data governance and cloud data governance. *Personal and Ubiquitous Computing*, 23(5–6), 839–859. <https://doi.org/10.1007/s00779-017-1104-3>
- [4] Armbrust, M., Ghodsi, A., Xin, R., Databricks, UC Berkeley, & Stanford University. (2021). Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. In 11th Annual Conference on Innovative Data Systems Research (CIDR '21). http://cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf



- [5] Ataei, P., Thorpe, S., Regula, S., & Staegemann, D. (n.d.-a). Why big data projects fail? A systematic literature review. AIS Electronic Library (AISeL). <https://aisel.aisnet.org/acis2024/35>
- [6] Ataei, P., Thorpe, S., Regula, S., & Staegemann, D. (n.d.-b). Why big data projects fail? A systematic literature review. AIS Electronic Library (AISeL). <https://aisel.aisnet.org/acis2024/35>
- [7] Biagi, V., Patriarca, R., & Di Gravio, G. (2021). Business intelligence for IT governance of a technology company. *Data*, 7(1), 2. <https://doi.org/10.3390/data7010002>
- [8] Checking your browser – reCAPTCHA. (n.d.). <https://pmc.ncbi.nlm.nih.gov/articles/PMC9442782/>
- [9] Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From big data to Big impact. *MIS Quarterly*, 36(4), 1165–1188. <https://doi.org/10.2307/41703503>
- [10] Ciancarini, P., Poggi, F., & Russo, D. (2016). Big Data Quality: A roadmap for open data. *IEEE International Conference on Big Data Computing Service and Applications (BigDataService)*, 210–215. <https://doi.org/10.1109/bigdataservice.2016.37>
- [11] George, D. (2026b). Digital Inclusion vs. Fiscal Revenue Assessing India's Per-Gigabyte Data Tax. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.20922073>
- [12] Dageville, B., & Cruanes, T. (2024, July 1). Industry Benchmarks and Competing with Integrity. Snowflake. <https://www.snowflake.com/en/blog/industry-benchmarks-and-competing-with-integrity/>
- [13] Databricks vs. Snowflake | Databricks. (n.d.). Databricks. <https://www.databricks.com/databricks-vs-snowflake>
- [14] George, D. (2026a). Digital dividend data taxation's potential to transform India's economy and redefine fiscal policy in the mobile era. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.19021895>
- [15] Gahi, Y. (2026). What Went Wrong with Data Lakes? A 15-Year Reality Check from the Field. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2606.08266>
- [16] George, D. (2026h). From Billable Hours to Billable Outcomes-How Artificial intelligence is restructuring pricing models in the global IT services industry. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.22678525>
- [17] Grover, V., Chiang, R. H., Liang, T., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems*, 35(2), 388–423. <https://doi.org/10.1080/07421222.2018.1451951>
- [18] George, D., George, A., Dr.T.Baskar, & Pandey, D. (2026). Digital Dependency and Temporal Distortion: A Critical review of smartphone use, cognitive impact, and Behavioral intervention in Post-Pandemic Society. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.19864918>
- [19] Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. *Information & Management*, 53(8), 1049–1064. <https://doi.org/10.1016/j.im.2016.07.004>
- [20] Hai, R., Koutras, C., Quix, C., & Jarke, M. (2023). Data Lakes: A survey of functions and systems. *IEEE Transactions on Knowledge and Data Engineering*, 35(12), 12571–12590. <https://doi.org/10.1109/tkde.2023.3270101>
- [21] George, D., & Dr.T.Baskar. (2026). Degrees, skills, and the Architecture of Human Capital: Reconceptualising India's path to inclusive Economic development. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.20843906>
- [22] Hasan, M. M. U. (2026a). Cost Optimization (FINOPS) for Large-Scale public Cloud adoption: budgeting controls, Chargeback/Showback, and usage forecasting (Public Sector and Enterprises). *Global Academic Journal of Economics and Business*, 8(01), 43–50. <https://doi.org/10.36348/gajeb.2026.v08i01.006>
- [23] George, D. (2026i). Managers who build and managers who break: Abusive supervision, psychological safety, and Team Learning-A critical review of the evidence and its limits. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.22914731>
- [24] Inakollu, M. V. (2024). FinOPS-Driven cloud optimization models for enterprise applications. *Digital Commons - University of South Florida (University of South Florida)*. <https://doi.org/10.46121/pspc.52.2.10>
- [25] George, D. (2026d). When Technology works but adoption fails: Leadership capability gaps in enterprise Artificial intelligence transformation. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.20759384>
- [26] Joshi, D. & Ernst & Young. (2024). From Quality to Value: How Robust data Governance Drives Analytics-Led Business performance. *JOURNAL OF INTERNATIONAL CRISIS AND RISK COMMUNICATION RESEARCH*, S4, 403–405. <https://distantreader.org/stacks/journals-ojs/jicrcr/jicrcr-3528.pdf>



- [27] George, D. (2026c). Employee retention in the modern workplace: the role of organizational culture, leadership, and career growth beyond compensation. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.20819049>
- [28] Khine, P. P., & Wang, Z. S. (2018). Data lake: a new ideology in big data era. ITM Web of Conferences, 17, 03025. <https://doi.org/10.1051/itmconf/20181703025>
- [29] George, D. (2026e). Big data, behavioral biometrics, and the era of continuous authentication – a review. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.22070510>
- [30] McKnight Consulting Group. (2026). Comparative performance and cost analysis of modern analytical data platforms. In McKnight Consulting Group. <https://mktgsite.enterprisedb.com/sites/default/files/pdf/Comparative-Performance-and-Cost-Analysis-of-Modern-Analytical-Data-Platforms.pdf>
- [31] George, D. (2026g). Cost to Company and Revenue to Company-Reading Labour Value from Both Sides in an Age of Rented Machines and Metered Software Agents. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.22924283>
- [32] Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2017). Big data analytics capabilities: a systematic literature review and research agenda. *Information Systems and e-Business Management*, 16(3), 547–578. George, D. (2026f). Killer acquisitions and the blind spot of merger control: economic logic, contested evidence, and implications for innovation policy. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.22250776>
- [33] <https://doi.org/10.1007/s10257-017-0362-y>
- [34] Nambiar, A., & Mundra, D. (2022). An overview of data warehouse and data lake in modern enterprise data management. *Big Data and Cognitive Computing*, 6(4), 132. <https://doi.org/10.3390/bdcc6040132>
- [35] Ogunwale, O., Onukwulu, E. C., Joel, M. O., Adaga, E. M., & Achumie, G. O. (2023). Strategic Roadmaps for AI-Driven Data Governance: Aligning Business Intelligence with Organizational Goals. *International Journal of Management and Organizational Research*, 2(1), 151–160. <https://doi.org/10.54660/ijmor.2023.2.1.151-160>
- [36] Otto, B. (2011). Data governance. *WIRTSCHAFTSINFORMATIK*, 53(4), 235–238. <https://doi.org/10.1007/s11576-011-0275-1>
- [37] Pashikanti, S. (2024). Building Scalable Data Lakes with Databricks and Snowflake: An Enterprise Approach. *Journal of Artificial Intelligence Machine Learning and Data Science*, 2(2), 2064–2067. <https://doi.org/10.51219/jaimld/santosh-pashikanti/453>
- [38] Richman, J. (2026, May 4). Best Data Management Tools in 2026: An Honest Comparison for real teams. *Estuary*. <https://estuary.dev/blog/data-management-tools/>
- [39] Shah, R., Mukherjee, K., Tyagi, A., Karnam, S. K., Joshi, D., Bhosale, S. P., & Mitra, S. (2023a). R2D2: reducing redundancy and duplication in data lakes. *Proceedings of the ACM on Management of Data*, 1(4), 1–25. <https://doi.org/10.1145/3626762>
- [40] Snowflake. (2026, September 8). Beyond the benchmark: the real world total cost, security and agility of enterprise data platforms. Snowflake. <https://www.snowflake.com/en/blog/engineering/snowflake-vs-databricks/>
- [41] Somajohassula, D. K. (2025a). Financial cloud cost optimization: A FinOps framework for modern financial institutions. *World Journal of Advanced Research and Reviews*, 26(1), 2620–2631. <https://doi.org/10.30574/wjarr.2025.26.1.1323>
- [42] Sreejindeth, P., & Padmavathy, T. (2026). Eco-FinOps: a causal-agentic framework for energy-efficient and explainable cloud cost optimization. *Frontiers in Artificial Intelligence*, 9, 1861514. <https://doi.org/10.3389/frai.2026.1861514>
- [43] Sura, R. (2025a). Measuring ROI of data and analytics programs: A framework for enterprise impact. *World Journal of Advanced Engineering Technology and Sciences*, 16(3), 265–276. <https://doi.org/10.30574/wjaets.2025.16.3.1343>
- [44] Sura, R. (2025b). Measuring ROI of data and analytics programs: A framework for enterprise impact. *World Journal of Advanced Engineering Technology and Sciences*, 16(3), 265–276. <https://doi.org/10.30574/wjaets.2025.16.3.1343>
- [45] Van Renen, A., & Leis, V. (2023). Cloud Analytics Benchmark. *Proceedings of the VLDB Endowment*, 16(6), 1413–1425. <https://doi.org/10.14778/3583140.3583156>
- [46] Van Renen, A., Leis, V., Friedrich-Alexander-Universität Erlangen-Nürnberg, & Technische Universität München. (2023). Cloud Analytics Benchmark. *PVLDB*, 16–6, 1413–1425. <https://doi.org/10.14778/3583140.3583156>



- [47] Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J., Dubey, R., & Childe, S. J. (2016). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, 70, 356–365. <https://doi.org/10.1016/j.jbusres.2016.08.009>
- [48] Wavestone. (2026, September 22). Wavestone | Management & Technology Consulting Firm. <https://wavestone.com/>
- [49] Why most big data analytics projects fail. (2019). [Dataset]. <https://doi.org/10.1287/orms.2019.06.08>
- [50] Wieder, P., & Nolte, H. (2022). Toward data lakes as central building blocks for data management and analysis. *Frontiers in Big Data*, 5, 945720. <https://doi.org/10.3389/fdata.2022.945720>
- [51] Xin, R., & Mokhtar, M. (2021, November 2). DataBricks sets official data warehousing performance record. Databricks. <https://www.databricks.com/blog/2021/11/02/databricks-sets-official-data-warehousing-performance-record.html>