Unlocking the value of future-ready cloud through data observability

Data governance knowledge series – Topic 6

October 2024

01 Cloud: The new era of data management

The evolution of cloud data platforms has been significant, moving from expensive, constrained and maintenance-heavy on-premises systems to dynamic, scalable and reasonably priced cloud-based solutions. As enterprises navigate digital transformation, the adoption of cloud data platforms is accelerating at an unforeseen rate. Gartner® predicts that more than 50% of enterprises will use industry cloud platforms by 2028 to accelerate their business initiatives.[1] Moreover, worldwide end user spending on public cloud services is expected to grow by 20.4% to USD 675.4 billion in 2024, up from USD 561 billion in 2023, according to the latest forecast.[2]

These organisations are increasingly adopting microservices architecture over traditional monolithic systems, driving significant traction to cloud data platforms due to the increased flexibility, scalability, and resilience of the architecture as it breaks down applications into smaller, loosely coupled services that can be individually developed, deployed and scaled.

Further, strategic cloud data platforms are being rapidly adopted by organisations due to their unmatched scalability, allowing for dynamic resource adjustments to meet demand and prevent inefficiencies. They spur innovation and agility by facilitating quicker application deployment and shortening time-to-market. According to Gartner, the growth in end user spending of public cloud services is expected to increase from 17.3% in 2023 to 22.1% in 2025.[2]

While the benefits of cloud data platforms are plenty, evolving cloud environments present significant challenges. Data integration is complex, as integrating disparate data sources and legacy systems into a cohesive ecosystem can be difficult and time consuming. Also, although ensuring data quality has been recognised as essential, businesses expect more. Rather than address emerging issues, businesses want to stop them from arising before they affect the enterprise applications.

[1] Gartner Press Release, Gartner says cloud will become a business necessity by 2028, 29 November 2023 - https://www.gartner.com/en/newsroom/press-releases/2023-11-29-gartner-says-cloud-will-become-a-business-necessity-by-2028

[2] Gartner Press Release, Gartner forecasts worldwide public cloud end-user spending to surpass USD 675 billion in 2024, 20 May 2024 - https://www.gartner.com/en/newsroom/press-releases/2024-05-20-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-surpass-675-billion-in-2024
GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

Further, the need to meet regulatory compliance and governance requirements quickly makes it critical to have robust end-to-end visibility on data flow, especially with evolving regulations like Digital Personal Data Protection (DPDP) Act 2023, automated data flow (ADF)/centralised information and management system (CIMS), Draft National Data Governance Framework Policy in 2022, International Financial Reporting Standards (IFRS) 17, Insurance Regulatory and Development Authority – Business Analytics Project (IRDAI-BAP) reports, etc.

We, now, are aware that the cloud is the backbone of any modern organisation, but its sustainability is contingent upon the implementation of effective cloud management. The nature of cloud environments, where data flows across disparate systems and sometimes several cloud providers, makes it difficult to maintain a unified view of data health. This nature makes it more difficult to discover and diagnose issues due to gaps in the monitoring coverage. Moreover, the dynamic and elastic nature of cloud resources add to this difficulty, as instances can be spun up or down based on demand, requiring real-time updates to monitoring configurations.

One of the lesser-known downsides of the cloud computing industry is that a significant portion of cloud spending is often squandered because of factors such as resource overprovisioning, inefficient use and a lack of cost-saving measures. Due to this, many firms have significant problems with cost optimisation. To reap the full benefits of cloud adoption, effective data platform modernisation and migration is a must.

The concept of data observability has become increasingly important in this setting. At a time when cloud data platforms are becoming the foundation of organisations, data observability is not just 'nice-to-have' but a business imperative. It empowers modern organisations with effective operation management, significant cost control and high trust in data, making it essential for business success. In this article, we will explore how organisations can navigate the modernisation of cloud data platforms with data observability to drive sustained business success.

02 Understanding data observability

Data observability is an organisation's ability to comprehend the complete state of its data and data systems at any given time. Data observability is a critical component of any standardised cloud-enabled data platform where organisations can take metrics being tracked across multiple systems and integrate them into a single pane of glass. As described below, data observability is fundamentally a function of data quality, data pipeline surveillance, end-to-end data traceability and expenditure insights. When integrated with an organisation's cloud data platform, it can maximise efficiency, reduce downtime and maintain a competitive edge in the data-driven market.

Core components of data observability

Data quality

Data quality pertains to the state where data defined by factors like accuracy, completeness, reliability and relevance meets the needs of its intended use. In today's data-driven world, data quality is essential for competitiveness to make informed decisions and promote innovation.

Strong data observability practices promote high trust, improved decision-making, enhanced operational efficiency, and proactive anomaly detection and resolution in real time before they impact business.

Pipeline surveillance

Pipeline surveillance is the continuous monitoring of data pipelines to detect, notify and address issues, guaranteeing steady and reliable data flow from data's source to data's destination. In this era of big data, ensuring data pipeline integrity is crucial for making informed decisions, innovation and competitiveness.

Data observability practices offer real-time insights to monitor pipeline execution timings, spot bottlenecks and identify issues that could delay data delivery.

Expenditure insights

Expenditure insights provide thorough examination and illustration of spending trends to maximise budget allotments and make informed financial decisions, thereby empowering businesses. Organisations using cloud platforms often risk overspending due to a lack of visibility with respect to usage and costs of resources.

Through data observability practices, organisations can attain a holistic view of their data operations expenditure, efficiently distribute resources and ensure sustained data operations.

Data lineage

Data lineage provides transparency and traceability by following the path and transformations of data from its source through numerous transformations till its destination. Understanding the effects of alterations made to the path of data, as well as how the alterations affect downstream systems, is essential in today’s increasingly complex data environments.

By implementing data observability practices, organisations can guarantee visibility, traceability and management of data processes, resulting in data systems that are more dependable.

03 Fitment of data observability across the modern stack

Having established an understanding of what constitutes data observability, let us now consider specific measures that must be applied across a data platform to ensure that the four components of data observability deliver the desired outcomes.

1. Data profiling, validation and anomaly detection

Organisations can get started with data observability by applying data profiling to understand the structure and quality of their data. This can be done by examining source systems and/or systems that host data to understand the structure and quality of the data before applying any transformations on it. Through data profiling, statistics and summaries on the data are collected and created to provide an understanding of the data. By doing this, process inconsistencies, missing values and anomalies can be identified. Post this, data validation and cleansing are done to ensure data accuracy and reliability, conformity to predefined rules, and correction of inaccuracies. Equipping data management systems with anomaly detection helps monitor the performance of the systems and facilitates quick identification and remediation of issues.

Within cloud data platforms with data observability, data profiling is to be conducted during data ingestion to understand the structure and quality of data coming in from various sources. Validation and cleansing during data integration and extract, transform and load (ETL) processes ensure that data is accurate and reliable, while the seamless integration of anomaly detection helps to monitor and handle unexpected data patterns, thereby providing robust and reliable data pipelines.

2. Data freshness and pipeline health

Having fresh data helps organisations to react quickly to market changes, customer needs and emerging trends. By monitoring data update timeliness and guaranteeing timely ingestion, processing and availability of data, organisations can ensure the availability of fresh data. Regular assessments of pipeline health, by monitoring metrics like pipelines' throughput (assessing processed data volume), data latency (data movement time) and disruptions in the pipelines, is a must.

During data ingestion, the latest data should be quickly imported and integrated, and data processing resources should preserve data freshness in real time. ETL operations should allow for quick integration and transformation along with utilising elastic storage capabilities for data to be available in a timely manner, which also helps maintain pipeline health.

3. End-to-end traceability and lineage

End-to-end data traceability can be achieved through data lineage, which increases data reliability while ensuring that data comes from a trusted source, has been transformed correctly and loaded to the specified destination. There are many data lineage tools that help document and visualise this movement and transformation of data through the platform, enabling impact analysis to assess changes and outlining the impact of changes on its quality.

To ensure this traceability across the pipeline, data movement should be documented and visualised during and post data integration and ETL processes. When data is being processed, complete visibility will be available by tracking compute resources from the point of data ingress to data egress. Having thorough records of sources and changes made to data further improves traceability and governance.

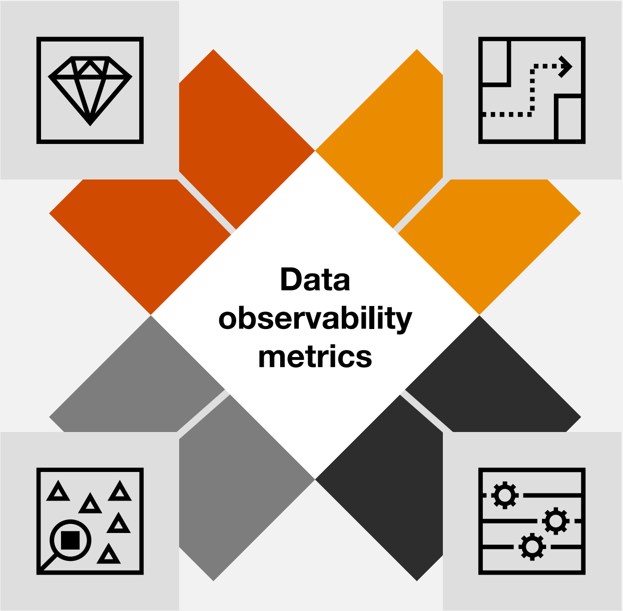
4. Spend categorisation and trend analysis

Cloud spending can be organised by classifying and analysing expenses to find areas that can be optimised or reduced. Determining what contributes to the cost requires a thorough analysis of expenditure categories like computing, storage and networking. Thus, to evaluate the performance of these categories and compare them against standards or benchmarks, patterns and areas for improvement must be found.

Expenditures on all types of storage solutions should be grouped to identify areas for saving costs, and costs related to computational resources and data transformation frameworks must be tracked diligently. Here, business intelligence (BI) and analytics tools can be leveraged to get visual insights into expenditure patterns and hidden cost optimisation opportunities. Furthermore, spends on secure data sharing and collaborative projects must be watched closely to ensure comprehensive financial monitoring.

Figure 1 highlights the quantifiable measures that help businesses evaluate their success in the above areas. By systematically tracking these metrics, organisations can gain valuable insights, make data-informed decisions, and drive continuous improvement.

Figure 1: Key metrics in data observability



Data quality (DQ)

- DQ dimension indices
- Performance of systems
- DQ issues resolution metrics

Pipeline surveillance

- Data freshness lag
- Pipeline status
- Pipeline recovery
- Data volume

Expenditure insights

- Resource utilisation efficiency
- Data platform costs

Data lineage

- Defined lineages
- Lineage accuracy rate
- Impact analysis Average time

04 Data observability in action

Alert fatigue to precision monitoring

A global ride-hailing organisation has its software architecture relying on thousands of microservices, enabling teams to iterate quickly and support global expansion. The organisation needed a scalable system to quickly detect, mitigate and notify engineers of service issues. The challenge was to create a robust metric and alerting pipeline that could strike a balance between alerting on the smallest scope critical issues and avoiding alert fatigue during larger outages. The organisation's observability team developed two in-data centre alerting systems for easy alert management (including infrastructure alerts), flexible actions (paging, emails, chat notifications, automated mitigations) and handling high cardinality to address small-scope issues. The new alerting systems and de-duplication platform scaled the organisation's systems by identifying even the smallest problems and presenting the user with only the relevant information while suppressing unnecessary notifications. This provided a generic solution to the organisation as it grew to operate many different product lines across hundreds of cities.

Data chaos to clarity

A leading e-commerce marketing platform encountered increasing data pipeline complexity as their operations scaled, making it hard to maintain data completeness and reliability. An erroneous data point in one of their instances generated six times more rows than expected, leading to overconsumption charges. Additionally, the business applications team needed to replace a crucial, outdated field in their dashboards, risking data integrity. The organisation established a data observability platform to address these challenges. This data observability platform helped track data completeness, lineage and quality, ensuring data reliability. For the field replacement challenge, the organisation utilised the data observability platform's data lineage capabilities to prepare in advance and mitigate disruptions. Because of this, the marketing platform was able to handle the erroneous data points effectively and replaced critical fields, therefore improving data reliability and operational efficiency, and supporting their global expansion and acquisitions.

Conclusion

The digital era’s cloud data platforms are revolutionising data management by offering unmatched scalability, flexibility, and cost-efficiency, thus, transforming data management. Robust cloud ecosystem strategies focused on data observability are essential, ensuring reliable data ecosystems through insights into data quality, pipeline health, traceability and cost management. Methods like data profiling, validation, anomaly detection and real-time monitoring ensure the integrity and accuracy of data, while lineage tracking enhances traceability and spend and trend analysis to optimise expenditures. To ensure these techniques cater to both current and future challenges, defining the data observability requirements is essential, following which the best solution can be analysed and identified. Further, a structured process must be created to manage data observability, which includes defining roles and responsibilities, as well as establishing an escalation and interaction model for seamless solution implementation. Industry leaders have demonstrated the monumental impact of data observability on scalability and operational efficiency, setting benchmarks for its successful implementation.

The future of data observability and data management is driven by automation, real-time analytics and cloud solutions – with artificial intelligence and machine learning enabling real-time monitoring and improving anomaly detection to enhance data quality and reliability. By embracing these capabilities, organisations can unlock new heights of efficiency, scalability and trust, thus paving the way for a data-driven future.

About PwC

At PwC, our purpose is to build trust in society and solve important problems. We're a network of firms in 151 countries with over 360,000 people who are committed to delivering quality in assurance, advisory and tax services. Find out more and tell us what matters to you by visiting us at www.pwc.com.

PwC refers to the PwC network and/or one or more of its member firms, each of which is a separate legal entity. Please see www.pwc.com/structure for further details.

Acknowledgements

This knowledge series has been reviewed by Abhishek Chaurasia and authored by Tanvi Yerrapragada and Nikhil Sawant.

Contact us

Mukesh Deshpande
Partner
One Consulting
mukesh.deshpande@pwc.com

Abhishek Chaurasia
Director
One Consulting
abhishek.chaurasia@pwc.com

Nishu Jain
Managing Director
One Consulting
nishu.jain@pwc.com

Prakash Suman
Director
One Consulting
prakash.suman@pwc.com

pwc.in

Data Classification: DC0 (Public)

In this document, PwC refers to PricewaterhouseCoopers Private Limited (a limited liability company in India having Corporate Identity Number or CIN : U74140WB1983PTC036093), which is a member firm of PricewaterhouseCoopers International Limited (PwCIL), each member firm of which is a separate legal entity.

This document does not constitute professional advice. The information in this document has been obtained or derived from sources believed by PricewaterhouseCoopers Private Limited (PwCPL) to be reliable but PwCPL does not represent that this information is accurate or complete. Any opinions or estimates contained in this document represent the judgment of PwCPL at this time and are subject to change without notice. Readers of this publication are advised to seek their own professional advice before taking any course of action or decision, for which they are entirely responsible, based on the contents of this publication. PwCPL neither accepts or assumes any responsibility or liability to any reader of this publication in respect of the information contained within it or for any decisions readers may take or decide not to or fail to take.

AW/October 2024 - M&C 41161