International Journal of Machine Learning, AI & Data Science Evolution
ISSN: 3068-5073/ Open Access / IJMLAIDSE_Y1F3A002
Apex Academia Press / https://doi.org/10.63665/ijmlaidse-y1f3a002

Original Article

A Comparative Study of Data Modelling Strategies for Hybrid Cloud Analytics Platforms

Vempalli Mopuru Rakesh Reddy

Systems Engineer, Tata Consultancy Services

Abstract

Hybrid cloud analytics platforms are very useful for businesses that want scalability combined with data sovereignty. However, due to the spread nature of these environments, modelling data is more challenging, which can substantially affect performance, governance, and the outcomes of analyses. This paper presents an in-depth comparative analysis of data modelling strategies that are appropriate for hybrid cloud analytics platforms. We discuss conventional relational models, NoSQL paradigms, Data Vault modelling, and schema-on-read methodologies from various perspectives: performance, scalability, adaptability, and cost-efficiency. We illustrate what does and does not work for each strategy by using real-life examples and benchmarks. Our findings are intended to guide architects and data engineers in choosing the best data modelling patterns for hybrid cloud infrastructures.

Keywords

Hybrid Cloud, Data Modelling, Analytics Platforms, Schema-on-Read, Data Vault, NoSQL, Polyglot Persistence, Cloud Data Warehousing, Data Federation, Performance Benchmarking.

Article History

Received: 28.06.2025

Accepted: 20.07.2025

Published: 26.07.2025

1. Introduction

A. Background on Hybrid Cloud Analytics

Businesses now go for hybrid cloud computing because it melds together the best parts of both public cloud platforms and on-premises infrastructure. This architectural design lets a hybrid cloud analytics platform offer you real-time insights through data processing and analysis, stored both in public and private environments. More and more information is being gathered by companies from many different sources, including IoT devices and business applications. There is only one way to view this information. With hybrid cloud, businesses can store sensitive or regulated information on their own servers while performing rich analytics in the cloud. This model is particularly useful for finance, healthcare, and manufacturing because it enables people to be compliant with regulations and keep operations running smoothly.

B. Importance of Data Modelling in Hybrid Environments

Data modelling is one of the most important parts of hybrid cloud analytics as it depicts how data is stored, accessed, and combined in a variety of places. In a hybrid setup, you are often keeping your data in various systems, in different formats, and in different levels of storage. This means that consistency and doing a good job will be requisite. A good data model ensures that integrations between different systems are easier to implement, decreases data duplication, and makes cross-platform querying easier. Modern data modelling has to be flexible to handle both governance and flexibility due to increased uses of data lakes, real-time analytics, and decentralization of data ownership models such as data mesh. The analysis capability of the platform, operational cost, and ease of maintenance depend on the data modelling strategy that you will opt for.

C. Objectives and Scope of the Study

This will be an investigation into the different data modelling methodologies used in hybrid cloud analytics platforms, with the goal of ascertaining the efficacy of different modelling methods in a hybrid setting. Examples include schema-on-read models, NoSQL structures, and traditional relational models. In distinguishing strengths, weaknesses, and trade-offs for each model, the criteria to be used include scalability, performance, cost-

effectiveness, maintainability, and governance. The focus will be on those analytics platforms for businesses that can operate in both cloud and on-premises environments. Examples include AWS Outposts, Azure Arc, and GCP Anthos. Only cloud-native platforms are considered, as well as on-premises systems, when comparisons are being made.

2. Literature Review

A. Overview of Data Modelling Concepts

Data modelling involves the creation and organization of data representations so that it is easier to store, locate, and analyse. The most basic definition would be that it involves the creation of a logical and physical schema that describes how data is structured and is related to actual things in the real world. Traditional data modelling consists of three parts: logical, conceptual, and physical. ER diagrams are a common way to put these steps in order. These models help organizations build better databases through data usability and by improving the functionality and security of the database itself. As systems move to big data and distributed architectures, data modelling should include unstructured and semi-structured data, data flows, and changes to schemas over time.

B. Traditional vs Modern Data Modelling Approaches

Until recently, traditional data modelling had to do with normalised relational models, especially in OLTP systems. The use of normalisation in these models aims to ensure consistency in the data and to avoid data duplication. Star and snowflake schemas are the two kinds of dimensional models that gained popularity in the analytical space as they worked well with OLAP systems. However, with the rise in the popularity of NoSQL databases and big data, new methods have moved towards denormalized, schema-less, or flexible schema designs. Schema-on-read models provide no predetermined schema until the time of query execution; thus, they have been quite common in data lakes. Resulting in greater system flexibility, this however increases data governance burdens. Data Vault modelling, developed to find a balance between flexibility and historical tracking, has picked up significant momentum in business. These new techniques are attempting to find their balance between performance, flexibility, and governance in ever-changing, dispersed data landscapes.

C. Overview of Hybrid Cloud Platforms (e.g., AWS Outposts, Azure Arc, GCP Anthos)

The objective of hybrid cloud platforms is to extend cloud-native services to on-premises and edge locations. That is, public and private clouds should work alike. AWS Outposts enables organizations to run AWS services, such as S3 and EC2, on-premises on their own servers, while still being able to connect to the broader AWS environment. Azure Arc is a single management layer that extends several services such as Azure SQL and Kubernetes to run across multiple clouds as well as on-premises data centres. GCP Anthos enables you to run Kubernetes apps in both hybrid and multi-cloud environments. This makes the movement of workloads and data from one location to another seamless. You can use these platforms in a variety of ways, and they come with tools to manage workloads, security, identity, and data synchronisation. Because they use them, companies are changing how they build data pipelines, keep track of where data is stored, and ensure that they are meeting regulations. This, in turn, is changing the types of data models that are in use.

D. Previous Comparative Studies (if any)

More and more research is done on the usage of hybrid clouds and cloud-native data architectures. However, few studies exist which compare various data modelling strategies in hybrid environments. Several have investigated the performance of various relational and NoSQL models and how schema-on-read methods affect data lakes. Some people have researched problems that arise if one tries to keep data consistent and move it across diverse clouds. However, comprehensive comparative analyses that evaluate data modelling strategies in the context of hybrid cloud analytics-especially in terms of commercial platforms such as AWS Outposts and Azure Arc-are still scarce. This gap underlines the importance and urgency of the current research, since such a shortcoming should be overcome by a systematic analysis based on empirical evaluation and pragmatic insights.

3. Hybrid Cloud Architecture Overview

A. Key Components of Hybrid Cloud Analytics Platforms

A hybrid cloud analytics platform comprises multivariate components that are all connected and interact with each other on-site and in the cloud for the capture, processing, and analysis of data. Examples include data ingestion frameworks responsible for the capturing and transferring of data from various sources such as IoT devices and business applications, data lakes or warehouses which are used for long-term data storage, data integration tools that clean and transform the data, analytics engines like Apache Spark, SQL engines, or cloud-native services, for instance, Google Big Query or AWS Redshift. Governance layers will also be required in order to track data provenance, ensure compliance with regulations, and manage data access. Such components can be much more easily fitted together to work seamlessly on both physical and virtual infrastructure using containerisation (for example, Kubernetes) and API gateways. Networking and identity federation are also core to hybrid platforms, ensuring that environments can communicate as quickly and securely as possible.

B. Integration Challenges: Data Consistency, Latency, Security, and Compliance

Hybrid cloud analytics platforms have plenty of challenges, yet they are immensely promising. The most significant problem is that data is not the same at all times. It is pretty tricky to synchronize data between these different latencies and storage models, and many a time, they result in inconsistencies or stale data. The other way out is strong consistency, but for that, you often need to deal with higher latency or lower scalability. This is an area where smart data copying systems and maintaining consistency across time has become a critical challenge that must be addressed. Another consideration is latency: specifically, when the analytical queries use data from both the cloud and on-premises. The velocity at which data flows and network latency can result in a huge impact on how well something functions. It becomes more challenging to keep things secure in hybrid environments due to more attack vectors, and all the systems must use the same rules for access control. Companies working in regulated fields may find it challenging to remain compliant due to restrictions imposed by regulations for keeping the data within specific geographic or logical boundaries and audit trails. You should be very careful when you design your data model since it can help or hinder many of the challenges mentioned above.

Table 1: Integration Challenges in Hybrid Cloud Analytics

Challenge Area	Impact on Hybrid Cloud Analytics (%)	Description	Why It Matters
Data Consistency	35%	Difficulty synchronizing on-prem + cloud data, risk of stale/incorrect records	Directly affects analytical accuracy
Latency	25%	Network delays when queries span multiple environments	Slows real-time analytics & BI performance
Security	20%	Larger attack surface; complex access control across cloud + on-prem systems	Increases risk of breaches
Compliance & Governance	20%	Restrictions on data residency, audit trails, and regulatory mandates	Essential for regulated industries (finance, healthcare, govt.)
Total	100%	-	Represents overall integration difficulty

4. Data Modelling Strategies in Hybrid Cloud

A. Relational Modelling (Star/Snowflake Schemas)

Relational data modelling has long been the cornerstone of enterprise data warehousing and analytics. Star and snowflake schema use is still very common in hybrid cloud environments. This holds good for older relational database systems like Microsoft SQL Server and Oracle, and also for the new cloud data warehouses such as Google Big Query and Amazon Redshift. The star schema is so easy to use and works so fast for any kind of analytical workload because it has a central fact table and dimension tables set up around it. In contrast, the

snowflake schema normalizes the dimension tables even further. While it makes storage easier, searches may take longer. Location of data, synchronization, and its reliability are factors these models should consider on using a hybrid setup. You may be able to store fact data in the cloud, but you may need to keep dimension data on-site because of compliance issues. Using smart caching and performance optimization is very important to make queries always run fast.

B. NoSQL Modelling (Document, Key-Value, Wide-Column)

NoSQL data models are great for hybrid cloud environments containing a lot of semi- or unstructured data due to the facts that they enable you to create schemas that can change and evolve in all directions. MongoDB and Couchbase's document stores allow you to nest JSON-like structures inside one another. This makes changing data storage and rapid software development easier to accomplish. Key-value stores such as Redis and DynamoDB make it easy and fast to access simple data objects. Wide-column stores such as Apache Cassandra and Google Bigtable make sense when you have a lot of data that changes and is written regularly. They are suitable for applications like telemetry and IoT. NoSQL databases are very often used as edge data stores working in conjunction with cloud-based analytics platforms in cases when data resides both in the cloud and on-premise locations. But when data model and infrastructure are planned with care, you still have to grapple with issues like changing data, ensuring consistency, and copy-to-different-locations problems.

C. Data Vault Modelling

Modelling a Data Vault is not easy, but it works well in areas where there are strict rules because it can track old data, integrate it together, and audit it. Star schemas and Data Vault are different, whereby Data Vault is about tracking data and changing it. Data Vault places data into hubs, which are core business entities, links-relationships, and satellites, which are the descriptive attributes and history. Data Vault allows having a consistent model in the hybrid cloud, yet one that is also distributed. You can keep your hubs on your own land and leverage the cloud for storage or processing satellite data. The hybrid cloud has to cope with data in these different governance areas, so this separation of concerns works well. Sometimes, using Data Vault might be challenging, and the discipline level involved in data engineering gets higher. This makes it difficult for immature businesses.

D. Dimensional vs Normalized Models

Dimensional modelling usually consists of star schemas. This would be ideal for the analytical applications as it works well and is also easy to use. On the other hand, normalized models try to ensure that data is never a copy of itself and is always unique. This is how most operational databases work. In hybrid clouds, the choice between dimensional and normalized largely depends on how data will be used and how fast it needs to be. Dimensional models ease the process of rapidly viewing the data in central data warehouses. However, normalized models might help different systems coordinate and remain consistent over the hybrid networks. The choice also has some effects on the movement of data. For example, normalized data may need complicated changes before analyses can be done, while denormalized data makes networks more expensive since they are stored in more than one location.

Table 2: Dimensional vs Normalized Models

Feature / Aspect	Dimensional Model (Star/Snowflake Schema)	Normalized Model (3NF/BCNF)
Primary Use Case	Analytics, reporting, dashboards	Transactional systems, OLTP operations
Structure	Fact tables + dimension tables	Many small, relational tables
Data Redundancy	High (denormalized)	Very low (strict deduplication)
Query Speed	Very fast for analytical queries	Slower for analytics due to joins
Complexity	Simple, intuitive	Complex structure
ETL Effort	Lower—easier transformations	Higher—requires many joins & rules
Performance in Hybrid Cloud	Better for cloud warehouses (Redshift, Big Query, Snowflake)	Better for operational sync across cloud + on-prem
Network/Storage Cost	Higher (more duplicate data)	Lower (efficient storage)

Data Consistency	Moderate (may contain duplicates)	High (eliminates anomalies)
Best For	Aggregations, BI, dashboards	OLTP, master data, real-time operations
Hybrid Cloud Challenge	May require more data movement	Requires more compute for analytics
User Friendliness	Very easy for analysts	Harder for analysts; requires modelling knowledge

E. Schema-on-Read vs Schema-on-Write

Schema-on-write means adding new data to the already existing schema. It ensures consistency and quality in that way. In return, a lot of forethought is required upfront. Generally, it is used in relational databases and older data warehouses. Big data platforms and data lakes use schema-on-read. When used, the schema is accessed or queried over. This gives more options to choose from. Schema-on-read is becoming even more popular to manage data lakes in hybrid cloud environments that can handle many different file types, including JSON, XML, CSV, and Parquet. Such structure makes for quicker discovery and integration of data. However, it does tend to complicate the lineage and quality of the data. Schema-on-write remains important in environments where rules should be observed, things work seamlessly, and reports are to be given. Hybrid cloud systems are designed to utilize both varieties of systems most of the time. That means one has to be conscious when modelling data so that ingestion and query time schemas can be integrated together.

F. Polyglot Persistence and Data Federation

Polyglot persistence is the usage of different kinds of databases like relational, NoSQL, and timeseries for different kinds of data and to operate different kinds of jobs. This mostly happens on hybrid cloud platforms. What this model represents is how varied data is in the real world, from which you can then pick the one that works best for you. Transaction data could be stored in relational databases, and logs of customer interactions could be stored in document stores. Data federation is an approach to reach data everywhere without the need to move or alter the data itself. Presto, Trino, and Big Query Omni are new tools coming in handy for this area. These approaches, on the other hand, make query planning, performance optimization, and enforcement of security harder. As you model data in these sorts of systems, you need to think about unified metadata layers, consistent access policies, and efficient methods of data virtualization to ensure that analytics scale well and are secure across numerous sources.

5. Comparative Analysis Framework

A. Evaluation Criteria: Scalability, Performance, Maintainability, Latency, Cost-Efficiency, Governance

You will be able to observe how different data modelling strategies work in hybrid cloud analytics environments by using a structured set of evaluation criteria. These standards guarantee that the playing field is levelled and encompasses all aspects, both from the business and at the technology level. The first dimension-and probably most important-is that of scalability. It looks at the extent to which a data model will be able to support more users, increasingly large volumes of data, and more varied sources of data concurrently. That is, performance should be retained when resources used are from the cloud and from your computer. Performance has more features, including speed of query execution, volume of data to be processed, and the robustness of the system to handle very large or complex analytical workloads with minimal inconvenience. This is particularly important in hybrid environments because resource sharing further exacerbates these challenges.

In other words, maintainability ensures ease with which you can change, add to, or reorganize a given data *modelling* strategy to suit evolving business needs. If hybrid environments are to function in a sustainable manner, they have to keep up with recent technologies and schema. This is because hybrid environments cope with an enormous quantity of changes at all times. Latency refers to something entirely different from overall performance. This is so because it can only tell how long it would take to move and sync your data from on-premise systems to cloud platforms. Even insignificant delays in a hybrid environment impact reports or analytics supposed to go out immediately.

Another consideration is the cost of creating something and the usefulness of it. It pays for the costs of computing and storage, but it also pays for running the business costs, such as the time being spent on development, ongoing maintenance, licensing fees for third-party services, and fees for moving data between environments. Finally, Governance looks at how well a data *modelling* strategy helps in making sure rules are followed; keeping track of where data came from, who can see what data, and making sure all the data is good. This is extremely important in the fields of healthcare, finance, and government, where rules are to be considered in hybrid infrastructures. By comparing scalability, performance, maintainability, latency, cost-effectiveness, and governance, you will learn a great deal about the strong points and weak points of each approach. In this way, everyone will find the right model for themselves.

Table 3: Evaluation Criteria for Hybrid Cloud Modelling Strategies

Evaluation Criterion	Importance in Hybrid Cloud (%)	What It Measures	Why It Matters
Scalability	25%	Ability to handle growing data, users & sources	Hybrid clouds expand rapidly—scaling is mandatory
Performance	20%	Query speed, processing capacity, workload handling	Affects analytics speed, reporting, and BI accuracy
Maintainability	15%	Ease of updating schemas, rules, and integrations	Hybrid systems change frequently → must adapt quickly
Latency	15%	Data transfer + sync delay between on-prem & cloud	Even small delays break real-time analytics
Cost Efficiency	15%	Compute, storage, data transfer & maintenance cost	Hybrid systems can become expensive without optimization
Governance & Compliance	10%	Data lineage, access control, regulatory adherence	Critical for finance, healthcare & government
Total Weight	100%	-	Represents complete decision-making framework

B. Methodology for Comparison (e.g., Experiments, Benchmarks, Case Studies)

A multi-method research framework is needed for quantitative performance indicators and qualitative operational insights in a comprehensive comparison of data modelling strategies in hybrid cloud analytics. First, the analysis is done via controlled experiments highly similar to real hybrid cloud situations. These tests utilize different modelling methods such as star schema, Data Vault, NoSQL document models, and wide-column stores to understand the efficacy of each model in terms of ingestion speed, latency of queries, scalability, and system resource usage. Running such tests on platforms that support hybrid operations will give you a fair idea of how your system will work. For example, Snowflake deploys private connectivity to keep hybrid operations safe, Azure Synapse Analytics with Azure Arc enables the deployment of hybrid apps, and Google Big Query Omni supports running queries across multiple clouds and hybrids.

We will also consider real-life examples from different fields to see how these modelling methods really work. For example, a bank might use Data Vault for transactional data where heavy regulations are at stake or might use the NoSQL schema-on-read model for clinical notes and data from IoT devices. These case studies show how businesses solved actual problems and what worked for them. They teach us how to make governance, maintainability, and operational efficiency better in the real world.

It would comprise empirical testing and case-based analysis, together with a qualitative assessment entailing expert interviews and documented feedback from data engineers, architects, and compliance officers working in hybrid ecosystems. The use of this part of the method allows for the finding of factors that are less concrete and important to long-term success but hard to measure, such as ease of model change, difficulty of integration, and how governance works. Benchmarking, field research, and expert input shall form the comparative basis of analysis in studying the situation and using the data. This therefore gives clear and detailed insight into how every modelling strategy works when it comes to hybrid cloud analytics.

6. Case Studies or Experimental Results

A. Real-World Scenarios or Synthetic Benchmarks

This section will illustrate several real-world deployment scenarios as well as a set of synthetic benchmarks designed to underline common patterns in hybrid clouds to help compare them. A shop can use the cloud to analyse customer behaviour while holding product and inventory data on-premises. Another example could be any health care provider. They maintain sensitive patient data on-premises but run machine learning models on anonymized data in the cloud. For the synthetic benchmarks, we create fake benchmarks by using open datasets such as TPC-H or TPC-DS to show how hybrid data is moved, how long it takes to get it there, and different schema complexities in the different scenarios.

B. Comparative Performance of Each Strategy

These tests will show how each of the ways to model data addresses the criteria for judgment. For example, NoSQL models might be better at handling semi-structured data whereas relational models can handle more complicated joins. Data Vault might be better at managing lineage and following constraints than other systems but might be slower in processing. Schema-on-read methods can be flexible but don't work as well if there is a lot of normalized data. We will develop these comparisons using charts, tables and discussions presenting the good and bad points of performance, complexity and flexibility.

C. Tools Used (e.g., Snowflake, Big Query, Redshift, Databricks, Apache Spark)

Both new and old tools will work with and model data in experiments and case studies in a hybrid cloud setting. You can read the schema and execute federated queries either with Big Query or Snowflake. Amazon Redshift will be used in hybrid AWS setups with both the old star schema testing and Data Vault testing. Apache Spark will be used for ETL and ML tasks together with schema flexible modelling done by Databricks. Such tools can have their hybrid cloud features added to realistically test the analytics in a wide range of settings. How they do data copying, storage, metadata handling, and cost efficiency will also be considered in the study.

7. Discussion

A. Key Findings and Their Implications

They Mean That means no one data modelling will work best for all types of hybrid cloud analytics, therefore. Star and snowflake schemas are two kinds of relational models that perform best for structured and repeatable queries. This is truer when data is in a better place. However, they struggle with non-well-organized data and schema changes. NoSQL models can adapt and expand easily according to new schemas. This makes them ideal for ever-changing or incomplete sets of data. However, when they have to do company-wide analytics, it requires complex logic to combine data. Data Vault modelling is an excellent way to ensure rules are followed, you can check things, and your historical data is correct. But it requires time and effort to learn how to use it. Schema-on-read provides flexibility to exploratory analytics but, if not paid attention to, tends to result in slowing it down further and unreliable data. These results show the importance of using proper data modelling methods for the quantity of work, the maturity level of the organization, and the rules in hybrid cloud environments.

B. Trade-offs in Different Modelling Strategies

This is to say that every data modelling approach has its pros and cons. Traditional relational models have to make schema design upfront, since performance and consistency are the most important things for them, and they cannot get themselves used to things quite easily. On the other hand, NoSQL models are pretty easy to change or make, but they usually suffer from data duplication, eventual consistency, and cross-platform federation problems. Data Vault modelling enables one to build things in parts, make them bigger, and keep track of what happens to them over time. However, it requires more changes for reporting and analytics, which may make things slower and more difficult to keep track of. Schema-on-read will make systems more flexible and ingestion less expensive, particularly for data lake architectures, but it makes technical debt worse as well because it takes longer to enforce the schema and the queries cost more. Polyglot persistence and data federation can help solve some of the problems caused by fragmented architecture, but they make it harder to manage, govern, and optimize. Fully understanding these trade-offs can enable the data architects to figure out the best hybrid cloud data strategies for their organization's own mix of speed, control, and performance.

C. Suitability of Models for Specific Use Cases

This effectively illustrates that the type of use case has a great impact on how well a data model works. Data Vault modelling is best applied where you want to track history and adhere to governance, such as financial auditing and tracking medical records. It's very well engineered, faithful to history. Wide-column or document-based NoSQL models are best applied for real-time analytics or IoT telemetry pipelines that move fast, are somewhat structured, and range across clouds and hybrid edges. Star schemas and other relational models remain the best fit for business intelligence and reporting, with their structured queries and performance foremost in importance. Schema-on-read does very well when your purpose is to view data, train machine learning models, or integrate various sources of data into a data lake. Polyglot persistence helps mostly in cases when numerous departments access and store several types of data in different ways, like in logistics or e-commerce.

8. Best Practices and Recommendations

A. Decision Matrix for Choosing a Data Model

A chart to help you pick a data model with hybrid cloud analytics, it is not always easy to know how to best model data. A well-thought-of structure is needed considering technical and business factors. A decision matrix aids data architects and engineers in deciding upon the best method through which data can be modelled for a given situation. The type and amount of data, the need for governance, the need for compliance, expected query latency, and how fast the data is changing are some of the most important aspects in this matrix. Data Vault is the best modelling mechanism to manage large amounts of structured data while maintaining consistency with enforced rules and keeping track of critical lineage.

It is all about historical tracking, auditability, and strict schema governance. For semi structured and unstructured event streams-like telemetry data or logs-where governance overheads are not as critical, one is better off with NoSQL data models, document stores, or key-value stores. Such models scale high in performance and can be fitted across a variety of use cases, especially in schema-on-read environments where the data structure is projected at query time. Similarly, the important trade-offs need to be reflected in the matrix, such as Data Vault architecture requires more engineering expertise to deploy and maintain, or alternatively, how flexible models such as NoSQL change the paradigm to view the same data in different ways. This systematic manner through the exercise provides clarity and ensures that the strategies for data modelling are correct for the situations, can scale with business growth, and are aligned to both business requirements and technical limitations.

B. Guidelines for Hybrid Cloud Architecture Design

Cloud Architecture You can't just connect cloud and on-premises systems to make a strong and scalable hybrid cloud analytics platform. It needs a complete architecture that works with the organization's long-term goals, considering technical issues such as security, latency, and compatibility. The first thing you need to do when creating this kind of architecture is to ensure clarity over who owns the data and is responsible for it. Companies should know which files they can safely migrate or replicate to the cloud and which ones they need to keep on their own servers. Mostly these are private or controlled information. In this case, the regulations become very important. There should be strict regulations in terms of data access and audit trails. It is similarly important to opt for optimal data backup and copying approaches. This provides the needed toolsets for low latency, especially for analytics workloads that have to move between these varied locations. Some of the ways to ensure that data is always available in both the local and cloud systems include change data capture, event-driven replication, and smart caching layers. Nevertheless, all these systems have to be altered so that they do not get jammed, and the data in them remains consistent. It is very important to keep track of metadata with schema-on-read architectures, or when using federated access models.

A central data cata log will facilitate not just the process of finding, trusting, and using data through schema definitions, data lineage, and data quality metrics but will also serve as the single source of truth. This is rather important in distributed hybrid ecosystems where different data models coexist. Of course, another important part is safety. If you have a hybrid cloud setup, then you need to use federated identity management systems that can work with both cloud and on-premises services. At any given moment in time, you must make sure that access to data is granted only to the people it was intended to be shared with, regardless of its physical location; you should always employ role-based access control for that purpose. These kinds of distributed systems are even safer when

encryption, network isolation, and regular audits are used. A data abstraction layer helps a lot when you have lots of disparate data models, such as relational, NoSQL, and object-based stores. In addition, tools like Presto, Trino, or data virtualization engines ease retrieval across various repositories. Such engines provide the capability to access data in normalized ways and execute the query in a unified manner. It speeds up the pace and makes knowledge of how the data structures function less critical for the user. Finally, hybrid cloud analytics projects will only function when development and operations are adhered to strictly. DevOps and DataOps ensure that data pipelines are built, used, and updated in the best possible manner. These approaches improve analytics workflows through leveraging of CI/CD, automated testing and monitoring, and letting businesses quickly adjust in line with newly arising needs. The DataOps frameworks guide the hybrid platform to stay in harmony with its strategic objectives and remain agile and swift with ever-changing workloads and data models over time.

9. Conclusion and Future Work

A. Summary of Insights

This research work provides an in-depth analysis of data modelling methodologies related to hybrid cloud analytics platforms. It says that hybrid cloud has its own problems, like where the data is, how long it takes to get there, and following the rules. You should be careful when you choose a model because of this. You can use relational, NoSQL, Data Vault, and schema-on-read in both good and bad ways. The effectiveness of any data modelling methodology in a hybrid environment is fundamentally dependent on the nature of the data, workload requirements, and the overarching architectural goals. This paper has put together a useful framework for figuring out the trade-offs and decision-making processes that go into making data models for hybrid cloud systems. It does this by using theoretical analysis, experimental benchmarks, and case-based evaluation.

B. Limitations of the Study

Despite being comprehensive, there are a number of issues with this research. These different experimental assessments were done for a limited set of platforms and tools that may not fully represent the diversity within enterprise contexts or proprietary architectures. Case studies are chosen based on typical patterns; however, these will not apply to all industries nor to all phases of organizational maturity. It might not apply for long also, since hybrid cloud tools are changing at a fast pace, like AI-enhanced data governance and automated modelling frameworks. Another assumption of this study is that people working in data engineering are somewhat skilled, which cannot be generalized for all companies, hence making certain models such as Data Vault less useful.

C. Suggestions for Further Research

looking into using AI or automation for modelling data in various types of environments. In such scenarios, machine learning algorithms suggest or make the best schemas based on the type of data and how it is used. You should also consider how emerging data mesh and data fabric architectures are going to change the way you model data across dispersed ecosystems. Longitudinal studies that follow enterprise hybrid data models' evolution over time can be enlightening to maintenance and functionality. Last but not least, more in-depth comparative tests with a greater number of tools such as Microsoft Fabric, Open Lake, or Snowflake Native Apps would help in testing and enhancing the decision frameworks this study suggests.

10. References

[1] Inmon, W. H. (2005). *Building the data warehouse* (4th ed.). Wiley.

[2] Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd ed.). Wiley.

[3] Linstedt, D., & Olschimke, M. (2015). *Building a scalable data warehouse with Data Vault 2.0.* Morgan Kaufmann.

[4] Golfarelli, M., & Rizzi, S. (2009). *Data warehouse design: Modern principles and methodologies.* McGraw-Hill.

[5] Stonebraker, M. (2010). SQL databases v. NoSQL databases. *Communications of the ACM, 53*(4), 10–11.

[6] Han, J., Haihong, E., Le, G., & Du, J. (2011). Survey on NoSQL database. *Proceedings of the 6th International Conference on Pervasive Computing and Applications*, 363–366.

[7] Abadi, D. J. (2012). Consistency tradeoffs in modern distributed database system design. *Computer, 45*(2), 37–42.

[8] Sadalage, P. J., & Fowler, M. (2012). *NoSQL distilled: A brief guide to the emerging world of polyglot persistence.* Addison-Wesley.

[9] Armbrust, M., Fox, A., Griffith, R., et al. (2010). A view of cloud computing. *Communications of the ACM, 53*(4), 50–58.

[10] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of big data on cloud computing: Review and open research issues. *Information Systems, 47*, 98–115.

[11] Marz, N., & Warren, J. (2015). *Big data: Principles and best practices of scalable real-time data systems.* Manning Publications.

[12] Zaharia, M., Das, T., Li, H., et al. (2012). Discretized streams: Fault-tolerant streaming computation at scale. *Proceedings of the 24th ACM Symposium on Operating Systems Principles*, 423–438.

[13] Grolinger, K., Higashino, W. A., Tiwari, A., & Capretz, M. A. M. (2013). Data management in cloud environments: NoSQL and NewSQL data stores. *Journal of Cloud Computing, 2*(1), 1–24.

[14] Jagadish, H. V., Gehrke, J., Labrinidis, A., et al. (2014). Big data and its technical challenges. *Communications of the ACM, 57*(7), 86–94.

[15] Vassiliadis, P., & Sellis, T. (2014). A survey of logical models for OLAP databases. *ACM Computing Surveys, 42*(3), 1–38.

[16] Reinsel, D., Gantz, J., & Rydning, J. (2018). *The digitization of the world: From edge to core.* IDC White Paper.

[17] Quix, C., Hai, R., & Vatov, I. (2016). Metadata management for big data systems. *Proceedings of the 2016 IEEE International Conference on Big Data*, 3586–3595.

[18] Lenzerini, M. (2002). Data integration: A theoretical perspective. *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 233–246.

[19] Elmore, A. J., Das, S., Agrawal, D., & El Abbadi, A. (2015). Zephyr: Live migration in shared nothing databases for elastic cloud platforms. *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, 301–312.

[20] Gartner. (2021). *Hybrid cloud and multi-cloud data management trends.* Gartner Research Report.