Chapter -11

Multi Cloud and Federated Data Management Frameworks

Mr. Rakesh Verma
Assistant Professor (Department of ECE)
Email Id - rakesh005004@gmail.com
Hindu College of Engineering Sonipat

Ms. Seema
Assistant Professor (Department of Applied Science)
Email I'd- seemahmr9@gmail.com
HMR Institute of Technology and Management

Mr.VIGNESHWAR RANGINI
Data Warehouse Solution Architect (IT Department)
Capgemini America, USA
Email: ranginivigneshwar@gmail.com

Dr. Samta Suman Lodhi
Assistant Professor (Department of Computer Science and Engg.)
Global Institute of Information Technology, Greater Noida
Email ID: samtasuman@gmail.com

Abstract:- Organizations are rapidly implementing multi-cloud strategies to take advantage of various cloud service companies for improved flexibility, effectiveness, and stability in an era of dispersed computation and accelerating expansion of data. This chapter examines how federated data administration frameworks, which facilitate smooth data processing and transfer across several systems without centralizing data storage, interact with several-cloud settings. We start by outlining multi-cloud architectures and their essential elements, such as connectors for periphery & hybrid computation. After that, the topic of federated data management is covered, with a focus on collaboration, data sovereignty, including preserving privacy. The functions of important technologies like Apache Arrow, Kubernetes Federation, and new standards from groups notably the Cloud Native Computing Foundation (CNCF) in coordinating processing of data are examined. Benefits like greater capacity and adherence to laws like GDPR are balanced against drawbacks like lock-in of vendors, latency, and threats to security. Practical implementations are demonstrated through actual-life instances from sectors such as healthcare and banking. The chapter ends with suggestions for practitioners and a summary of upcoming developments, such as AI-driven federated learning. This thorough review gives readers the skills they need to create and implement reliable multi-cloud federated platforms.

Introduction:-The way businesses manage handling data has completely changed as a result of the exponential expansion of data generation and the growing complexity of their IT infrastructures. These days, businesses work in heterogeneous settings where data is spread across several cloud providers, on-site headquarters, and edge computing equipment. Modern data ecosystems are scattered, which has led to the development of complex mixed-cloud and federated repository frameworks that can smoothly coordinate data across several platforms while preserving speed, safety, and reliability.

Due to a number of beneficial factors, such as minimizing vendor lock-in, optimizing costs, guaranteeing business continuity, meeting regulatory compliance requirements, and utilizing best-of-breed services from various providers, multi-cloud plans have become a fundamental component of contemporary enterprise architecture.

Most big businesses today perform in multi-cloud settings, with many using three or more cloud platforms at once, according to industry analyses. The handling of information, regulation, emancipation and security now face previously unheard-of difficulties as a result of this paradigms change.

Organisations can maintain uniform access and governance over dispersed data sources without the need for physical data consolidation thanks to federated data management, a complementary strategy. Its design gives users and programs a virtualised, integrated perspective while allowing data to stay in its original location. Shorter data migration, lower latency, better data sovereignty compliance, and increased flexibility in adapting to shifting business needs are all major benefits of the confederation architecture.

The architectural concepts, technical frameworks, execution techniques, and new developments in multi-cloud and federated data management are all covered in this chapter. We look at how businesses may efficiently create and implement these frameworks to optimise the value of their remote information collections while reducing risks to operations and administration.

The Multi-Cloud Data Management Landscape

Evolution and Drivers-

Adoption of multi-cloud has been slow but unstoppable. Early cloud adopters usually started with solely one cloud provider, concentrating on particular workloads like disaster restoration or testing and developing environments. Enterprises increased their cloud footprint as cloud technologies developed and organisational trust increased. They frequently adopted numerous providers informally across various company departments or acquired companies with distinct cloud strategies.

The adoption of multi-cloud has been accelerated by several important factors.

First, many organisations continue to place a high priority on avoiding vendor lock-in. Reliance on a single cloud provider reduces negotiating power, raises strategic risk, and may limit technology possibilities. Organisations preserve options and lessen reliance on the roadmap, price structure, or company resilience of any one vendor by spreading applications over several providers.

Second, several cloud service providers have unique advantages and specialised offerings. While specialised providers offer distinctive capabilities in areas like edge computing or industry-specific solutions, Microsoft Azure supplies deep integration with enterprise software and mixed-cloud capabilities, whereas Google Cloud Platform leads in data analytics and machine learning services, and Amazon Web Services shines in the breadth of solutions and robust ecosystem. Organisations can utilise the finest services from all providers for particular use cases by implementing a multi-cloud plan.

Third, geographic sharing of information is becoming more and more necessary due to legislative and data independence needs. Certain types of data must stay inside national boundaries or in particular geographical areas, according to many countries. By allowing businesses to assign workload to suppliers who retain a suitable regional presence while preserving reliability throughout the larger systems, multi-cloud techniques aid in adherence.

Fourth, the widespread utilisation of multi-cloud is driven by concerns about business continuity and risk reduction. Distributed designs lessen vulnerability to security events, provider-specific outages, and service deteriorations. By deploying active-active architectures across several clouds, organisations may guarantee maximum availability and tolerance even in the event of serious outages to individual providers.

Challenges in Multi-Cloud Data Management

In order to take advantage of advantages like redundancy, cost optimisation, and adaptability, multi-cloud administration entails storing, processing, and analysing data across several cloud providers (such as AWS, Azure, and Google Cloud). But it adds complexity that goes beyond single-cloud configurations. I list the main obstacles below, along with industry insights from publications like Gartner and Forrester that show how these problems affect businesses implementing numerous cloud solutions.

Key Challenges

Data Consistency and Synchronization

Different APIs, data formats, and inter-cloud transfer latency make it challenging to guarantee real-time or nearly instantaneous information consistency across different clouds. Different databases or conflicts could arise, for instance, if the database's transaction in AWS does not immediately appear in Azure.

According to Gartner, 70% of multi-cloud adopters have trouble synchronising data, which frequently leads to errors or operational deficiencies in programs that depend on uniform data models.

Security and Compliance Risks

Every cloud provider has its own encryption regulations, security models, and compliance frameworks (including GDPR and HIPAA). Administering audits, information encryption, and login rules across clouds makes you more susceptible to security lapses or noncompliance.

Multi-cloud systems increase risks by two to three times contrasted to single clouds, according to Forrester's 2023 Cloud Security Report. Global expansions are made more difficult by data residency requirements, such as European Union data residing in Europeans.

Vendor Lock-In and Interoperability

One provider's exclusive applications and services could not work well with another, leading to dependencies and impeding portability. It can be costly and tedious to move workloads or data amongst clouds.

According to IDC's 2022 poll, 60% of businesses have interoperability issues, which increase the expense of switching and limit their ability to implement superior goods.

Architectural Patterns for Multi-Cloud Data Management

Architectural Patterns for Multi-Cloud Data Management

Data Fabric Architecture

A crucial architectural design for multi-cloud management of information is the Data Fabric, which effectively breaks down silos for uniform, online data access and statistical analysis by establishing a single, intelligent layer to access and govern data across various cloud and on-premise sources using metadata, AI, and automation. It ensures that data is controlled, safe, and accessible in current circumstances across hybrid circumstances, avoiding exclusivity between vendors and facilitating agile insights by managing the complete data journey, from identification and quality to access through consume.

A central control plane oversees dispersed information networks across several clouds in data fabric designs, which generally use a hub-and-spoke approach. This control plane coordinates data transportation and transformation procedures, enforces accountability principles, keeps an exhaustive inventory of all data assets, and records lineage and relationships. Although they are subject to centralised control and control, individual clouds behave as spokes, housing real data stores and analysing power.

Data Mesh Architecture

Four fundamental ideas form the foundation of the data mesh design. Domain teams that comprehend company circumstances and requirements are given explicit accountability for data products under subject-oriented ownership. Handling data resources with the same discipline and quality requirements as consumer products is emphasised by the concept of "data as a product." The cloud-based features offered by self-serve data platforms allow domain teams to independently create and manage their data assets. Collaborative computational governance strikes a compromise between enterprise-wide regulations and guidelines and domain flexibility.

Through federated oversight, the data mesh allows enterprises to embrace platform variety in different clouds while preserving coherence. Through standardised interfaces and protocols, various domains can participate in global data exchange while optimising their cloud technology selections depending on particular demands.

Hybrid and Distributed Data Architecture

Many businesses use hybrid systems, which create intricate distributed data environments by combining on-premises technology with several public clouds. Complex methods for data insertion, replication, synchronisation, and access are needed for these designs.

Thermal data travels to primary storage on the cloud, cold data archives to low-cost cloud storage tiers, as well as hot data that is requiring low latency access stays on-premises or in edge locations in hybrid data architectures. Policies for handling the data life cycle automatically transfer data around echelons according to its economic importance, age, and consumption practices.

Techniques for data synchronisation and replication preserve consistency between dispersed places. Businesses can use a variety of consistency models, from strong consistency for transactional systems to ultimate coherence for insignificant data. Multiple modifications to duplicated data are handled using conflict resolution techniques, which differ according to the needs of the application and the degree of contradiction endurance.

This transmission is extended to thousands of edge sites via computing edge designs, posing hitherto unheard-of scale issues for data management. While integrating with centralised cloud computing facilities, edge repository approaches must manage sporadic connectivity, constrained computational capabilities, and their own autonomy.

Federated Data Management Frameworks

Without physically relocating or recreating data, federated data management frameworks allow organisations to combine and query several independent, divergent data sources as a single virtual database. These systems enable scalability in intricate, dispersed contexts like data meshes by striking a compromise between centralised policies and decentralised execution. Businesses managing big data between clouds, premises-based systems, and different DBMS types like MySQL, Oracle, or Hadoop platforms will find them very helpful.

Popular Frameworks and Tools

- **Apache Atlas:** Offering tag-oriented regulations, lineage surveillance, and Hive/HBase connections, it is competent for Hadoop environments. The
- **Presto:** A SQL query engine with robust analytics optimisation for federated access over Hadoop, S3, or RDBMS.
- **Egeria:** Metadata interaction across instruments for replicated catalogues that is independent of vendors.
- **Amundsen/Open Metadata/DataHub:** In big data arrangements, data catalogues facilitate federated search and administration.

Benefits and Use Cases

These frameworks scale for global initiatives like retail analytics across Oracle and BigQuery, decrease silos, and improve compliance (e.g., GDPR/CCPA via local modifications). Ensuring appropriate governance and optimising queries across diversity are challenges. They integrate with data mesh for domain-developed "data as product" concepts, making them suitable for input-intensive industries.

Data Governance in Multi-Cloud and Federated Environments

Unified Governance Frameworks

In federated and multi-cloud contexts, effective data governance becomes exponentially more difficult. Establishing governance frameworks that respect provider-specific capabilities and limitations while spanning numerous platforms is crucial for organisations. Data standards, confidentiality, security, legal compliance, and management through the lifecycle must all be uniformly addressed by such structures throughout the whole data ecosystem.

Centralised policy design and administration capabilities that span several clouds and on-premises systems are offered by unified governance solutions. By adjusting implementation to platform-specific mechanisms while preserving consistency in intent, these kinds of networks allow organisations to establish policies on governance once and apply them universally.

Data catalogues, which offer thorough visibility into data assets across all platforms, are the cornerstone of governance in dispersed systems. Cloud environments are regularly scanned by automated discovery processes, which find new data assets and catalogue their attributes. Sensitive material that needs particular handling, including personally identifiable information or authorised data, is quickly identified by machine artificial intelligence classifying.

Privacy and Compliance Management

Data processing must adhere to strict privacy standards, including the Consumer Privacy Act of California, the General Privacy Regulation, and several industry-specific regulations. Because data may be located in several jurisdictions with disparate regulatory requirements, multi-cloud setups make compliance more difficult.

Tracking data lineage across system borders, implementing data subject rights at scale, enforcing purpose constraints on data usage, and proving compliance through thorough audit trails are all necessary for the handling of privacy within federated systems.

Consent management procedures that track customer privacy preferences and enforce them across all data processing operations, independent of the system, must be put in place by organisations.

Security in Multi-Cloud Data Environments

For businesses using several cloud providers to guarantee data security, compliance, and efficiency, security in multiple-cloud data structures is crucial. The following are important factors and recommended behaviours:

Data Protection and Encryption:

Comprehensive encryption techniques that cover data in use, in transit, and at rest are necessary for data protection. Organisations frequently choose client-side encryption for sensitive data in order to retain control over secret keys and lower confidence needs in cloud providers, even if cloud-native encryption tools offer encryption on the server for the information stored.

Identity and Access Management (IAM):

Managing user rights on several platforms raises the possibility of insider threats or incorrect configurations.

Statutory and Accountability Concerns:

Different supplier options make it more difficult to meet requirements like SOC 2 or ISO 27001.

Supplier Dependent and Lock-In:

Safety hazards may arise from relying on third-party services in the event that a supplier is exploited.

Implementation Strategies and Best Practices

Assessment and Planning

Implementing multi-cloud and federated data management successfully starts with a thorough evaluation of the current situation and a precise statement of the desired results. Businesses should inventory all of their present systems, data possessions, and application software across every platform, recording dependencies, performance traits, and management needs.

Multi-cloud information structures should be in line with corporate goals and capabilities through strategic planning. When choosing the best implementation strategies, organisations must assess their technical maturity, talent set, and ability to manage change. Ambitious big-bang changes usually don't work as well as phased deployments that build organisational capabilities and provide steady value.

Design and Architecture

Modularity, abstraction, and standardisation should be prioritised in architectural design. Future flexibility is made possible by clearly specified interfaces between components, which also avoid lock-in to certain suppliers or platforms. While guaranteeing uniformity throughout the company, reference architectures and design patterns speed up execution.

Data organisation, user patterns, efficiency needs, compliance restrictions, and cost optimisation should all be taken into account when selecting choices about data architecture. Organisations need to decide on proper management of life cycle rules, queuing strategies, replicating and synchronisation techniques, and data placement choices. Contrary to conceptual structural perfection, these choices should be based on real-world commercial needs.

Operations and Optimization

Strong monitoring, notifications, and handling of incidents tools across several clouds are necessary for successful operation. Companies should put in place unified observability platforms that compile metrics, logs, and traces from every environment to give a thorough understanding of the performance and health of the system. In multi-cloud settings, where usage-based pricing models may result in unforeseen costs, controlling expenses comes essential. Companies should put in place cost monitoring and optimisation processes that monitor cloud spending, find opportunities for optimisation, and enact cost leadership guidelines. While preserving performance, automated cost optimisation techniques can reduce waste, use less expensive solutions, and appropriately scale capabilities.

Conclusion:-Frameworks for multiple clouds and federated data management are crucial tools for contemporary businesses negotiating more complicated and dispersed data environments. These frameworks allow businesses to maintain consistent governance and control over the information they hold while utilising the distinct advantages of several cloud providers, avoiding vendor lock-in, meeting regulatory requirements, and optimising costs.

Careful consideration of design, safety, authority, and operation is necessary for successful implementation. In order to effectively manage heterogeneity, organisations must strike a balance between the advantages of multi-cloud freedom and the degree of complexity it adds by establishing the proper levels of abstraction and standardisation. By providing unified access to dispersed data without necessitating expensive and disruptive combining, syndicated storage and management techniques enhance multi-cloud solutions.

With new developments like edge computing, artificial intelligence integration, quantum computing, and ecological concerns, the field is still changing quickly. In order to secure investments in current capabilities while adapting to evolving technologies and requirements, organisations must retain architectural agility and continuous improvement attitudes.

In the future, automation, creativity, and standardisation will make multi-cloud or federated data management more advanced. Platform- particular complex will be reduced by industry efforts to create universal standards and protocols, and more autonomous and self-optimizing information systems will be made possible by artificial intelligence. Businesses that make significant investments in developing strong multi-cloud and federated data management skills set ourselves up for success in a future with is becoming more distributed and analytics-driven.

References

- Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50-58.
- Bernstein, P. A., Haas, L. M., & Halevy, A. Y. (2011). Full-disjunctions: Polynomial-delay iterators in action. Proceedings of the VLDB Endowment, 4(11), 1064-1075.
- Buyya, R., Vecchiola, C., & Selvi, S. T. (2013). Mastering cloud computing: Foundations and applications programming. Morgan Kaufmann.
- Dehghani, Z. (2019). How to move beyond a monolithic data lake to a distributed data mesh. Martin Fowler Blog.
- Fahmideh, M., Daneshgar, F., Rabhi, F. A., & Beydoun, G. (2019). A generic cloud migration process model. European Journal of Information Systems, 28(3), 233-255.