International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 1, January - February 2023 |

DOI:10.15662/IJEETR.2023.0501008

Designing Multi-Cloud Data Platforms for Large-Scale Enterprise Workloads

Karthikeyan Selvarajan

Senior Staff Software Engineer, San Francisco, California, United States

karthik.selvarajan83@gmail.com

ABSTRACT: The first-time workloads of large-scale businesses also demand increasing amounts of data platforms that can support elastic scalability, capacity to support high availability, interoperability, and resilience to heterogeneous clouds. The design of such platforms to have a multi-cloud environment presents significant architectural problems related to data integration, location of workloads, distributed processing, security, governance, performance, and cost optimization. This study introduces an architectural framework for crafting multi-cloud data platforms incorporating services and capabilities throughout Amazon Web Services (AWS) and Google Cloud Platform (GCP). The given architecture constitutes one data architecture that has ingestion, storage, processing, orchestration, metadata management, analytics, security, and monitoring layers. It can pool distributed data processing capabilities and can policy-driven workload placement to in addition to scaling to enterprise-scale apps, reduce the reliance on a single cloud provider. The architecture is also interoperable as standardized interface is offered to provide automated governance and centralized observability across boundaries of cloud. The key evaluation aspects to ensure that the proposed platform is efficient are the performance, scalability, availability, and resource usage. The paper explains how a well-planned multi-cloud environment can be used to help power and agile enterprise data solutions and deal with the distributed infrastructure challenges. The proposed architecture provides a powerful foundation where interested companies seek to implement transforming enterprise workloads, on scalable, regulated and cloud independent data system bases.

KEYWORDS: Multi-Cloud, AWS, GCP, Data Platforms, Cloud Architecture, Distributed Systems

I. INTRODUCTION

The need to develop digital services, enterprise applications, Internet of Things (IoT) systems, artificial intelligence, and data-intensive business processes has increased the importance of data as a company's asset. Contemporary firms produce huge amounts of structured, semi-structured, and unstructured data in transaction systems, customer applications, sensors, web services, and analytical systems. Computing infrastructures are required to process and handle these heterogeneous data and offer scalability, availability, flexibility, and effective utilization of such data. This has in turn led to the incorporation of cloud computing as an important source of data management within the enterprise based on on-demand computing, storage, networking, and analytics services. The reliance on a single cloud provider, however, may come with certain problems, which are associated with the lock-in with the vendors, craft of geographic coverage, fluctuations in prices, data sovereignty and relying on the technologies developed by the provider.

Multi-cloud computing is an approach where organizations are utilizing the service of multiple independent cloud providers depending on workload, performance, cost, availability and regulatory requirements. Businesses are not wedded to a single infrastructure but can spread out applications, and data-processing loads in other set ups in the clouds. Distributed Processing Cloud computing service providers like Amazon Web Services (AWS) and Google Cloud Services (GCP) offer a massive number of options to store data, carry out distributed processing, analytics and orchestrate and vanish the infrastructure. Workloads on two or more clouds cannot just be utilized by someone as an effective multi-cloud data platform. The variations in service models, APIs, storage mechanisms, networking mechanisms, security controls, and pricing arrangements and operational interfaces result in great architectural complexity.

Multi-clouds design of large scale data systems is as a distributed system problem. Enterprise workloads can interact in some of the areas such as real-time data loading, batch-loading, stream analytics, and data transformation, machine learning, reporting, and archival. The computational and storage needs of these workloads vary and they might need to dynamically move between cloud environments. An appropriate architecture should therefore offer the coordination of the workloads, data placement, distributed processing, service interoperability, fault tolerance, security and monitoring.

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 1, January - February 2023 |

DOI:10.15662/IJEETR.2023.0501008

This latency also makes performing measurement hard and overhead of latency, throughput, availability, resource and data-transfer overhead could differ between providers and physical sites.

This is the problem which has been discussed with help of some existing works. Measurement scales of cloud, fog and edge environments, which include latency, throughput, scalability, use over resources, reliability and quality of service have been come up in performance evaluation studies [1]. Research on new clouds has also demonstrated the increasing importance of distributed smartness, automation, as well as heterogeneous computing environments [2]. Distributed stream-processing has been developed to work with the large data sets on distributed systems that are scalable [3]. Similarly, cloud-fog architectural studies have examined the distributed computation and storage process over the heterogeneous layers of infrastructural layers [4]. The works are helpful to consider how to construct distributed data-processing systems; however, it is not very specific to the architecture of multi-cloud systems as a business.

The other acute issue is the resource management. The workloads distributed should be allocated to the respective computing facilities and performance-based, cost-based, availability-based as well as the nature of the workloads. The approach of dividing and planning of the resource to run in a non-homogeneous environment has been discussed in the literature on the distribution of cloud to things resource management [5]. Kubernetes, and, newest of all, cloud-native applications slowed down the containment of the applications within the heterogeneous infrastructure by providing support in deploying and managing containerized applications [6]. This kind of technologies can enhance the portability of applications, automation; however, it is not concerned with data consistency, data movement, control, security, or the dependency on provider specific services.

The continuous increase in distributed artificial intelligence adds to the necessity to make the flexible data platforms yet more prominent. The fact that distributed forms of learning have been deployed illustrates that it can be possible to transfer computational loads between cloud and edge resources when the abilities are accessible [7]. On the same note, cloud simulation systems have offered the means to test the provisioning of resources, scheduling, and behaviors of workloads prior to implementation [8]. Simulators of cloud-to-fog are compared with simulators, emphasizing the role of experimental evaluation in layouts of distributed infrastructures [9], whilst experiment-driven designs are used to reproducibly assess distributed computing surroundings [10]. These researches suggest that, in general, successful distributed platforms must imply the concept of architectural abstraction, along with methodological performance testing.

In spite of these, error still exists in terms of covering an integrated architecture specifically built to support large-scale enterprise data loads that are spread out amongst multiple public clouds. The area-specific solutions to the problem include resource scheduling, stream processing, cloud-native deployment or performance evaluation. The capabilities described here must be united into a multi-cloud data platform, which must provide the following features: (a) interoperability of data across cloud providers and (b) not moving data in vain.

This paper therefore provides a multi-cloud data platform architecture design to both AWS and GCP as a representative of cloud platforms. The proposed approach takes into account the consumption of data and data location, data processing, orchestration, metadata management, security, governance, observability, and workload location as layers in the architecture all of which are interconnected. The idea is to establish a malleable platform that will be able to support a multiplicity of enterprise workloads without reliance on provider-specific infrastructure. The architecture concentrates on standard interfaces, containerized services, policy-orchestration, distributed processing and centralized monitoring. The parameters may be utilized to assess its performance in regards to scalability, latency, throughput, availability, resource utilization, data-transfer overhead and operational efficiency. The proposed design will offer feasible architecture platform to any company who may need the deployment of trustworthy and scaling data platforms on cloud-agnostic platform that would preclude interoperability and ability to increase workloads to meet dynamic scale demands of extensive enterprise workloads.

II. RELATED WORK

Enterprise applications have then added the strain of scalable distributed interoperable performance conscious data platform to take up cloud computing. The studies that have been conducted to date have evaluated cloud, fog and edge environments respectively, with regards to various views, namely, performance review, resource management, distributed processing, orchestration, simulation and cloud-native deployment. The relevance of these works is that they have offered such long-awaited arguments on which to develop multi-cloud data platforms with the capacity to serve the massive loads of enterprises.

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 1, January - February 2023 |

DOI:10.15662/IJEETR.2023.0501008

A thorough taxonomy of cloud, fog, and edge computing performance assessment measures developed a systematic foundation on which the distributed computing landscapes will be measured [1]. Latency, throughput and resource utilization, energy consumption, reliability, scale and quality of service are some of the dimensions, which were the focus of the study. They are specifically applied in multi-clouds where the workloads of the enterprise can be transparently distributed to heterogeneous cloud operator. Nevertheless, the traditional evaluation methods basically assume single or hieratical cloud-edge scenarios instead of explicitly targeting cross-provider information platforms, with many public clouds.

The other direction of research has been a smarter and more diverse evolution of cloud computing which takes into account the introduction of new technologies i.e., IoT, blockchain and artificial intelligence [2]. The article has highlighted the effects of the shift of cloud platforms to a more distributed and intelligent computing ecosystems, rather than centralized infrastructure. These trends can be applied to multi-cloud systems because enterprise data systems have a tendency to increase their requirements in terms of automation, smart resource control, enhanced security, and distributed analytics. However, the adoption of these technologies creates more complexities in architecture that needs to be resolved with a single governance and interoperability platforms.

Enterprise applications of large scale are turning out to be the source of live time flow of data, which needs to be processed in real time. A study on stream-processing systems on a distributed computing system that processes large amounts of data had been conducted in which stream processing architectures and techniques that played a role in processing large amounts of data streams on a distributed computing system had been studied [3]. It was unique by its scalability, low processing latency and failure resistant, distributed execution. All these features can be directly transferred towards multi-cloud data platforms where data ingestion and analytics workloads could be spread out over cloud resources that are geographically spread. Nevertheless, the stream-processing architecture is not scalable to accommodate the multi-cloud architecture like cross-cloud storage, work load mobility and single-management.

Common accounts of the architectural vicinance of cloud and fog and mutually obligatory computing domain have also been discussed [4]. As shown in the current cloud-fog architecture, one can allocate the computational and storage resources to various layers of the infrastructure so as to bring the latency to a minimum extent and a higher scalability. Standards, tools and application models have also been used to further assist interoperability even in the heterogeneous environments. The principles provide solid architectural principles to multi-clouds particularly in the provision of distributed resource consumption and location of workload. In this instance, however, the cloud-fog architecture is very locality-sensitive to the end devices, although with multi-clouds platforms of the enterprise at stake, a structured control of autonomous cloud providers, and cloud services niche.

Another factor of research is resource management. Overviews of resource-management strategies upstream and down resource-allocation techniques to the distributed world computing resource-allocation were presented [5]. These issues as dynamic workloads, heterogeneous resource, scalability, resource efficient utilization were identified in the research. These issues are exacerbated in the multi-cloud environment since the workloads may even be distributed among providers based on performance, cost, availability guarantees and data-location guarantees. Consequently, resource-management mechanisms would have to be altered to shift to policy-based cross-cloud workload orchestration that is at the infrastructural-level optimization.

The cloud native is very useful especially with regards to portability of the enterprise workloads. Kubernetes can also be considered as a breakthrough in cloud-native computing given that it offers container orchestrating capabilities, auto deploying applications, scaling and managing distributed applications [6]. Even the level of abstraction using containers can reduce dependence on infrastructure specific to a particular provider and can be deployed on diverse cloud environments. However, portability of the orchestration is not necessarily what brings a regular data administration, cross cloud uniformity, administration and effective information transfer.

Distribution of machine-learning workloads input to heterogeneous resource coordination of computation has also been demonstrated [7]. Studies of distributed deep learning and cloud and edge resources show how a single computation process can be decentralized to different edges across the infrastructure to enhance efficiency in the processing. The methods can be utilized on the enterprise data platforms with AI and analytics loads. However, distributed computation has to be incorporated with data lifecycle management, storage architecture, security and workload orchestration to be a complete multi-cloud platform.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 1, January - February 2023 |

DOI:10.15662/IJEETR.2023.0501008

The research on cloud architecture has also been of great help due to the simulation based research. CloudSim is able to simulate clouds [8]. Likewise, the dexterity of choosing suitable experimental scenes of evaluating the distributed architectures was revealed in comparative case-studies of cloud-to-fog simulators [9]. These plans can be used to analyze multi-cloud design before its implementation by modeling the work load, performance, and resource utilization and scalability. All the real life issues associated with cross-cloud networking, provider specific services, governance and complexity of operation would also not be available in simulation.

The other way of assessing the distributed computing systems is through experiment based research frames [10]. The test on distributed infrastructures can be repeated on these architectures and can be used to test workload allocation, resource usage, and application performance. Divergent infrastructure and service properties are also a major consideration in the multi-cloud enterprise platforms; particularly the reproducibility is an important attribute of cloud providers.

In general, the backgrounds of the present research are adequately defined in terms of the sphere of distributed processing, resource management, orchestration of the cloud-native services, or experimental validation and measurement of the performance [1]–[10]. However, a significant portion of their work has been done on unique facets of distributed cloud computing other than to provide a complementary architecture on AWS-GCP multi-cloud data structures. The work, therefore, targets unifying cross-cloud data ingestion, distributed storage, processing, and orchestration, governance, security, observability, and positioning workloads in enterprise levels all on a single architecture.

III. FEDERATED DATA FABRIC AND POLICY-DRIVEN CLOUD EXECUTION ARCHITECTURE

The concept architecture is developed as a cloud-federated data platform which will consolidate various heterogeneous services on the Amazon Web Services (AWS) and Google Cloud Platform (GCP) and provide an overlying architectural and governance layer. The aim is not to ensure a complete reflection of the components in both providers, but to ensure ingestion storage processing, orchestration, analytics, and security and governance and observability are all parts that operate to form logical data architecture. This approach will remove the dependency on provider-specific implementations and allow organizations to investigate the scalability and specialization, which are available in different cloud environments.

The architecture is layered in the sense of which its enterprise data process passes through heterogeneity with the help of a service of control in analytics. Cross-cloud control plane coordinates the workload placement, synchronization of the metadata's, execution of a policy, monitoring and control of resources. The data plane, however, contains of the actual ingestion, storage, processing and analytical workloads that are utilized on the AWS and GCP. This isolation enables the control policies to be relatively independent of any particular cloud services and the selection of the execution resources dynamically based on the workload requests.

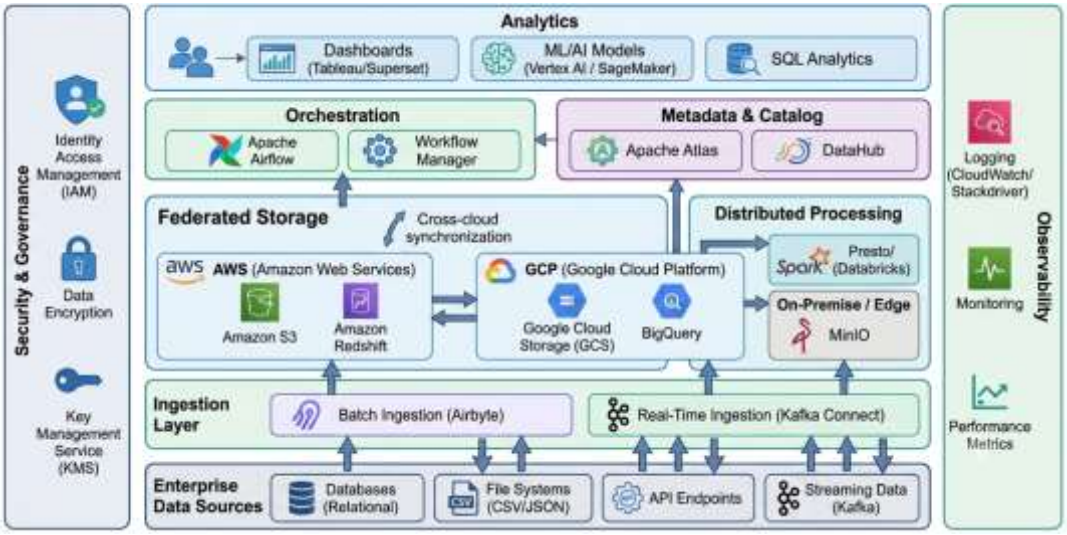

Figure 1- Federated Multi-Cloud Data Platform Architecture

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 1, January - February 2023 |

DOI:10.15662/IJEETR.2023.0501008

3.1 Multi-Source Data Ingestion Layer

Top level is adopted with the aim of having a single access to semi structured, unstructured and structured enterprise data. Relational databases, enterprise applications, transaction system, IoT devices, application logs are the most common, streaming platform, file and external data services. Ingestion subsystem enables the batch and flow of data acquisition.

Incoming data is given metadata to specify the source, format, time, sensitivity, ownership and computation needs. A routing system determines the appropriate cloud endpoint, based on the location of the data, what to do with it, and what its security boundaries are, as well as what resources it is currently using. One such resource may be the high-volume streaming data, which can be presented to an environment which is competent to carry out a sustained workload (e.g. a constant workload-optimized environment), or periodic enterprise records, which can be presented using a batch ingestion pipeline.

The validation, schema checking, duplicates checking and primary quality checking is also done by the ingestion layer. This can be done through the various different data sources by using standardized interfaces to communicate with the platform without necessarily having to use AWS- or GCP-specific application logic.

3.2 Federated Storage and Data Management Layer

The second level is distributed storage both in AWS and GCP. The proposed architecture is an abstraction of the logical or federated storage and not just the end-product storage over the network as it is considered that the cloud storage systems are storing the actual product or service. Storing physical data could be done using cloud-specific object stores, cloud-specific databases or repositories of analysis, but metadata was a single description of its location and properties.

The model facilitates the delivery of the raw, processed and curated areas of data. The raw zone contains information of source data, the processing zone contains information about the data undergoing processing and the curated zone contains information of data which has been verified to be valid and can be used to analyze the applications. There is a policy of data lifecycle, which describes data retention, data archival, data replication and data deletion.

The geometries of data placement depend on the frequency of access to that data, regulatory needs, regionality of processing and availability goal, as well as transfer overhead. The frequently accessed information can be stored in the location where the accessibility is required and the least used information can be transferred to the cheaper lower storage costs. Any decision on replicating in reference to storage and network-transfer overheads must also enhance resiliency, including replicating the datasets of choice to clouds.

3.3 Distributed Processing and Compute Layer

Processing layer: This offers batch analytics, stream processing, transformation, machine learning and also the ability to execute large-scale data engineering distributed execution. The processing workloads could be deployed in either AWS or GCP from where they depend on how they are computed and needed depending on their data needs.

The disaggregation of the processing workloads of underlying cloud infrastructure is one of the most significant factors of the proposed framework. Providers may age out of dependency on providers by using containerized offering, a generalized API and infrastructure-as-code. As a result, the identical logic processing pipeline can be executed in a variety of execution environments with minimal alteration.

The processing layer may also be used to classify the workloads. Jobs may be classified based on their latency sensitivity, strength of computational capabilities, data locality, the resource demand and target of availability. This classification will give the input to the workload-placement mechanism as detailed below.

3.4 Cross-Cloud Orchestration and Policy-Driven Placement

The orchestration provision has the decision making section of the structure. It bridges data pipelines, data processing activities, data dependencies, resource assignments and deployment of functional states between AWS and GCP. Placing the workloads on a single provider is not permanently done as witnessed in a framework where policy-based placing is used.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 1, January - February 2023 |

DOI:10.15662/IJEETR.2023.0501008

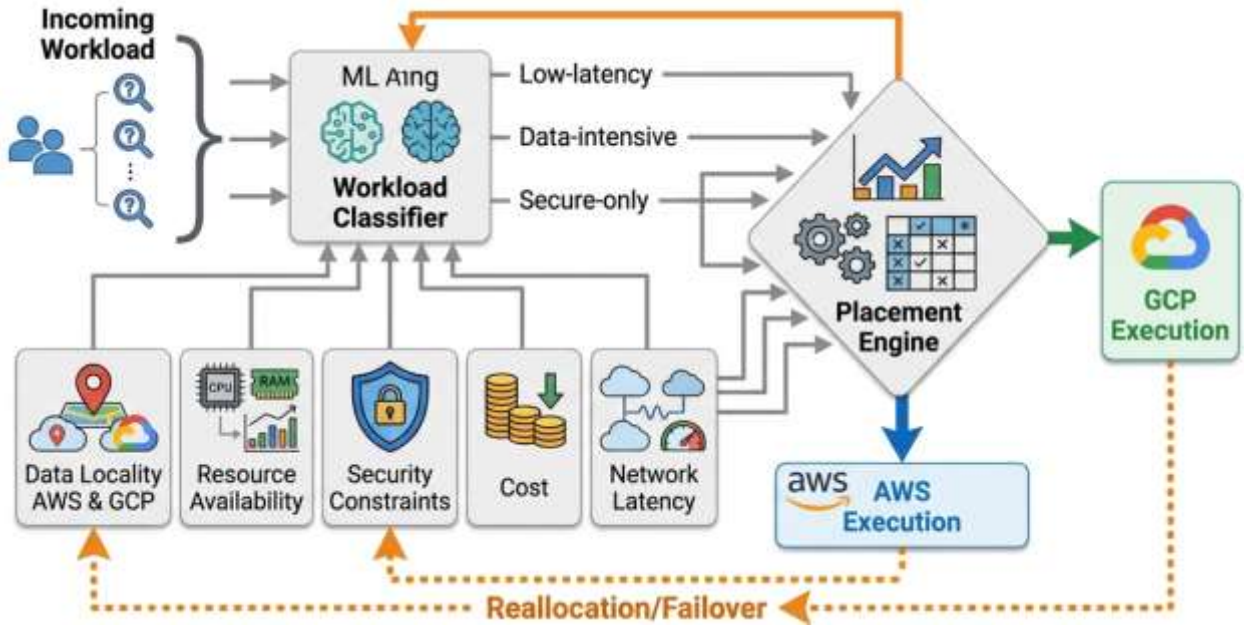

Figure 2 – Policy-Driven Workload Placement Model

An example of such a scenario is that the workload can be rated less on the placement, where the required dataset already exists within an optimal cloud, and thus will never be distributed between clouds. Likewise a workload with any governance constraints, which cannot be executed in any ambiance that does not meet any policy, cannot be deployed.

The orchestration layer is thus a policy-oriented (enterprise-policies) layer between the enterprise applications and the cloud infrastructure. It will also be in a position to play the role of fallback execution in the event of unavailability of a chosen cloud service.

3.5 Metadata, Catalog, and Semantic Management

There is a consistent description of datasets, processing jobs, services, ownership data at Schemas, lineage, quality signals, and storage locations are described by a cross-cloud metadata layer. This layer is essential as the physical distribution makes enterprise data very difficult to locate, as well as to control.

The catalog holds logical keys of data groupings, and gives identifications with their physical location. Metadata synchronization processes ensure that schema, location update, ownership, lifecycle status changes are synchronized across the environment of the federation.

The information lineage also records the relationships between the source data, transformational activities, smaller data structures and analysis end products. This enhances traceability and auditing, troubleshooting, impact analysis, and regulatory needs.

3.6 Unified Analytics and Intelligence Layer

Analytics layer will deliver access to business intelligence, statistical analysis, and machine learning, predictive analytics and enterprise reporting. Data accessed by analytical applications is accessed via logical interfaces instead of being directly dependant on physical location of individual datasets.

This abstraction enables analytical workloads to access data that is distributed to AWS and GCP. By query federation of data and virtualization of the data, minimization of unnecessary move of datasets where not necessary can be realized. In the case of computationally intensive workloads, the orchestration layer may choose which execution environment to employ depending on the processing resources available, and how much the workload is local to the data available.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 1, January - February 2023 |

DOI:10.15662/IJEETR.2023.0501008

The identical architecture is relied upon to run not only real-time dashboards but also pipelines featuring long-running pipelines alongside analytical and machine-learning. By doing so, analytical services are not linked to the infrastructure.

3.7 Security, Governance, and Compliance Plane

Security is carried out as cross-cutting plane within all the architectural layers, as well. Authentication and authorization are part of identity and access management, and role-based, or attribute-based policies regulate access to datasets and services. Both the transit and the rest of the data is encrypted, and key-management systems are implemented.

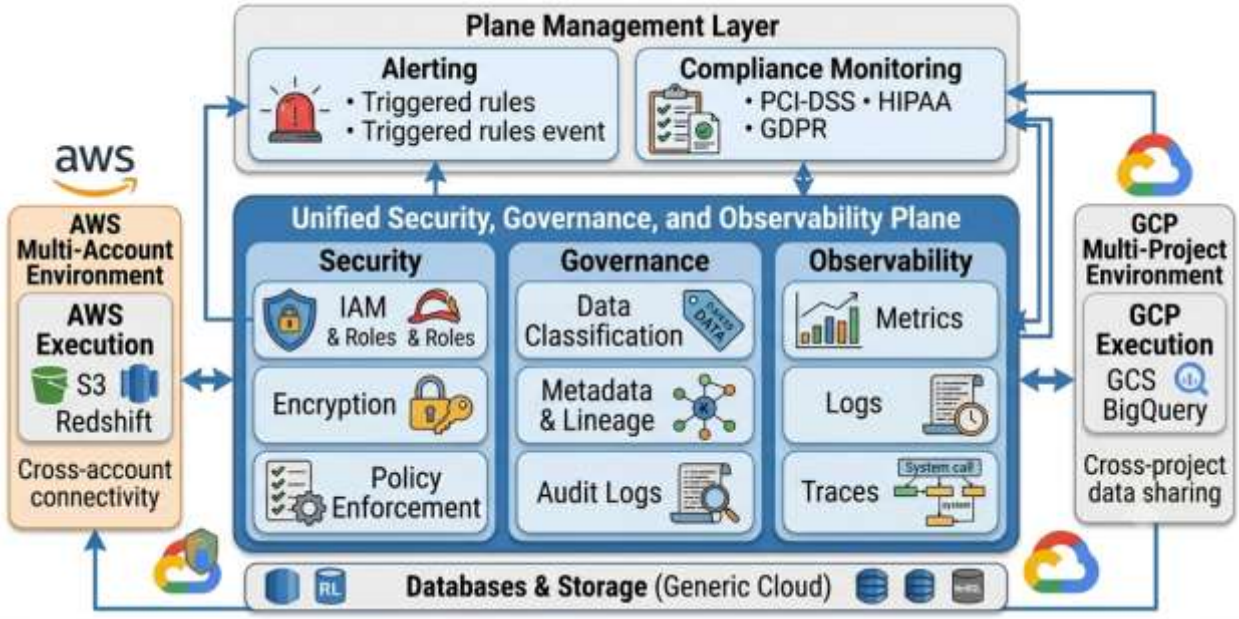

Figure 3– Unified Security, Governance and Observability Plane

Governance sub system determines the policies regarding the classification of data, retention, geographic limits, access control, lineage on workloads and the location of workloads. It is then possible to restrict sensitive ones to authorized execution environments. The policy-enforcement sites are segmented into ingestion and storage, processing and analytical areas.

Automated governance minimizes the need to engage in manual intervention and all violations of the policies are possible to be spotted during the implementation of workload and not just at each periodic audit score.

3.8 Observability and Operational Intelligence

A single monitoring method is also afforded by observability layer on both AWS and GCP. The single unified monitoring frameworks gather and link resource utilization, workload performance measures, security events, logs, measures, traces, pipeline status, and logs.

Cross-cloud observability is more imperative since the ultimate point of failure may be in the infrastructure, networking, data pipelines, processing services or in the orchestration components. Telemetry correlation can help the operators to distinguish between cross-cloud and on-site services breakages.

The framework also helps in the production of alerts with reference to set thresholds and abnormal behavior. Historical data can then be used to optimize telemetry to place workloads, allocate resources, and plan capacity.

3.9 Cross-Cloud Data and Control Flow

This all begins with enterprise data entering into the ingestion layer. This new information is transported with the achievement metadata affixed and governance rules and the information is then routed to its location. Evaluation of the

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 1, January - February 2023 |

DOI:10.15662/IJEETR.2023.0501008

workload requirements and processes on which transformation or analytical processing is to occur is done by the orchestration engine. Components are handled and run on AWS or GCP, or both.

Metadata, the lineage, the performance measurements and operational events are all constantly transferred to the control plane in execution. The data and observability checks have a governance policy that regulates all the lifecycle of the data and checks the service condition and resource consumption. Generally, workload migration, redistribution or failover can be triggered by a pre-defined policy in case of any fluctuations in the workload conditions orchestration layer can instigate the migration.

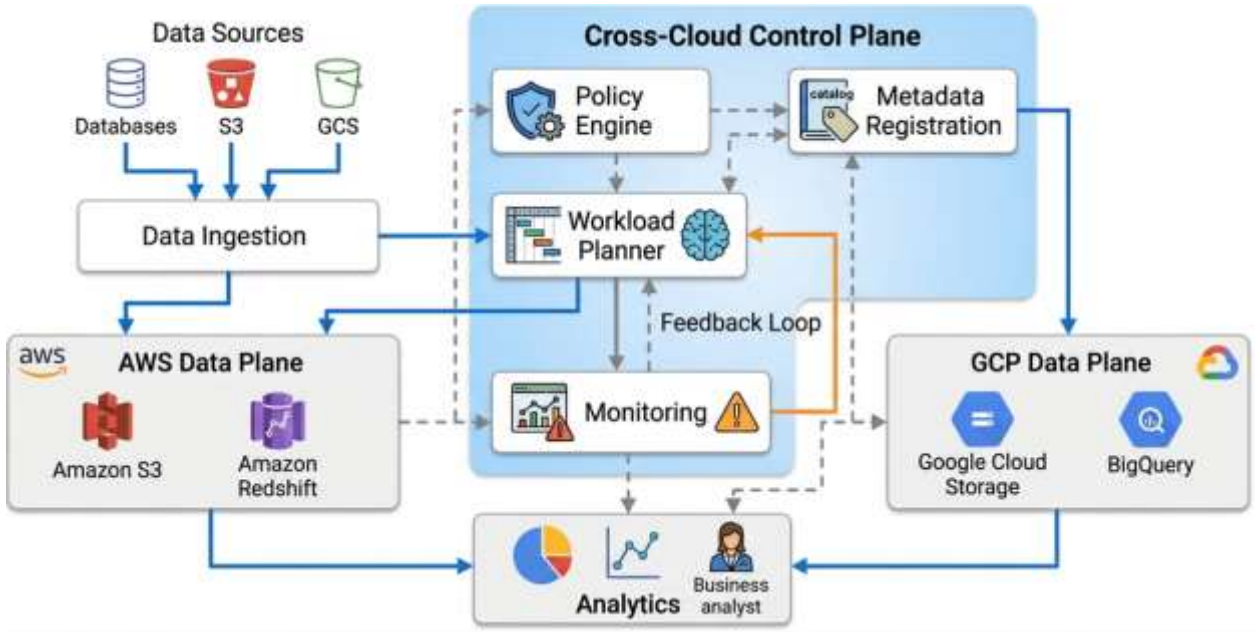

Figure 4 – Cross-Cloud Data and Control Flow

The given framework will therefore lead to a federated, but not replicated multi-cloud architecture. The required abstraction to handle AWS and GCP as a unitary, homogenous enterprise data platform is in the shape of standardized interfaces and centralized metadata, policy-orchestrated execution, security controls and single visible observability. The architecture offers a foundation to be used in considering scalability, availability, interoperability, performance and resource efficiency in the section of the framework below titled evaluation.

IV. FEDERATED ASSURANCE AND CROSS-CLOUD PERFORMANCE CHARACTERIZATION

The suggested multi-cloud data platform will go through evaluation to establish that it is effective in enabling the scalable, resilient, and interoperable and governed enterprise data and cross-AWS and GCP. Since the structure adopted integrates other architectures in a single structure instead of illustrating a single computational algorithm, system-level properties are considered in the process of evaluation. It is measured based on performance, scalability, availability, interoperability, resource utilization, cost efficiency and overhead. It is possible to run representative batch, streaming, and workloads at different levels of data and cloud-resource conditions to investigate the behavior of the proposed architecture.

4.1 Performance and Processing Efficiency

The performance processing metrics are the execution-time, throughput, the rate of response latency and the rate of data-processing. The workloads are implemented in both the background into their respective cloud backgrounds and when your call is needed to both AWS and GCP. Comparison helps to establish the impact of distributed execution and cross-cloud communications. The time on data-transfer latency is viewed as particularly special since the transmissions of mass data between providers can be accompanied with a further processing load. The proposed placement mechanism will lead to a minimum of unnecessary data flow since it will handle the data locality and then choose an

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 1, January - February 2023 |

DOI:10.15662/IJEETR.2023.0501008

execution environment. Other platform efficiency measures, which are required in cases where the workload is analytical, include query response time, and processing throughput.

4.2 Scalability and Elasticity

Experiments with scaling are provided by incremental additions of more data and parallel workloads and computational demands. Small data, medium data, and big data can be performed on the platform and the throughput, processing latency, CPU used, and memory consumption and scaling time can be tracked. The layered architecture allows independent scaling of the resources to ingestion, storage, processing and analytics. This character is appreciated in particular with the business loads in which the demand may vary significantly with time. Elasticity of AWS to GCP also, can be put to test in terms of workloads that cannot be served by a single work execution environment even though it can be enhanced to a superior level using more resources without the unnecessary disruption of services.

4.3 Availability and Fault Tolerance

Availability testing looks at the ways the platform may experience infrastructural and service outages. The failure of a processing-node, failure of a storage-service, network failures or the temporary unavailability of a single cloud environment are some of the controlled failure modes. The availability of the services, recovery time, and successful redirection of work and data consistency are some of the key measures. The federated design provides an ability to migrate those workloads of choice, to another cloud in a designated recovery state. However, metadata co-ordination, a compatible set up in interfaces, data replication and recovery policy set up is required to have a successful failover.

4.4 Interoperability and Portability

The test on interoperability is conducted under the performance of AWS workloads and GCP workloads, datasets, metadata and processing pipelines. This ability is contributed by the use of common metadata models, containerized services, standard interfaces, data formats that are portable, as well as infrastructure abstraction. The initiative to take a workload to another provider and the level of change that should be cloud specific can be a good measure. Limitation of modification work is translated into better portability and dependency on proprietary cloud service. Metadata consistency across providers can also be measured in terms of synchronization accuracy, and propagation latency.

4.5 Resource Utilization and Cost Efficiency

CPU, network bandwidth and compute utilization are some of the measures used to get the resource efficiency. The work to know whether it is efficiently utilizing the resources, the workloads could be linked to the fixed work pinned working policy, which would be dynamically set. This costing should cover compute, storage and network transmission costs, monitoring and orchestration costs. The cross-cloud system can add flexibility to the set up and could have other costs of data-transmission and data-management; therefore the cost of the whole operation must also be taken into account rather than the cost of the compute.

4.6 Governance and Operational Overhead

The last consideration of the evaluations is effectiveness of centralized governance and observability. Some of the measures may be policy-enforcement coverage, synchronization time (metadata), completeness of the lineage, monitoring coverage, alert-generation latency, and administrative effort. Unified observability also assists operators to visualize heterogeneous cloud resources based on a common visualization of their activities. Overall, the analysis will end with how the proposed architecture will enable it to strike a balance between its performance, resiliency, interoperability, governance, and cost-effectiveness without sacrificing flexibility of operations expected of a modern multi-cloud data platform.

V. FUTURE HORIZONS: TOWARD AUTONOMOUS AND ADAPTIVE MULTI-CLOUD DATA PLATFORMS

The provided architecture provides a foundation of a scalable and controlled multi-cloud data management, although, it is possible to see several opportunities to deploy the framework to more adapting, intelligent, and autonomous data platforms.

5.1 AI-Driven Workload Placement

In the future, it is possible to incorporate the machine learning models in the workload-placement engine, to estimate resource demand, execution time, network performance, and costs of the cloud-service. The platform might not have to be based on pre-defined policies largely because it might get to know the past working patterns of the workloads and propose or automatically select the appropriate execution environments. The seamless maximization where the

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 1, January - February 2023 |

DOI:10.15662/IJEETR.2023.0501008

decisions are achieved in the absence of the various workload and infrastructure condition could also be carried over through reinforcement learning.

5.2 Autonomous Data Management

A study example could be the creation of a self-managed data platform that has the capability of automating storage, replication, migration and lifecycle policies. Smart systems would be able to forecast data-access patterns, and would automatically put datasets that are heavily utilized in proximity to the data processing requests. Recovery and replication as is the case with resilience would enable to reduce unnecessary storage and network costs.

5.3 Cross-Cloud Federated Analytics and AI

It is possible to extend the architecture in the future with federated analytics and distributed artificial intelligence. Instead, of transferring complete datasets between AWS and GCP, analytical models can be executed with closer proximity to the data using only the results of the required intermediate calculations or changing model parameters. This would help minimize data traffic and offer further security to privacy-sensitive enterprise applications.

5.4 Intelligent Security and Governance

The advantages of AI-enhanced governance plane are as follows: the presence of an AI-based anomaly detector, policy checker, and ongoing compliance checking. The systems of the future would be able to dynamically detect an abnormal response by the data-access as well as policy violation and suggest corrective actions. A combination of metadata, lineage, identity, and operational telemetry can have a comprehensive security posture across the cloud boundaries.

5.5 Sustainable Multi-Cloud Optimization

The concept of energy use and carbon intensity can also be included in work place-location choices in the future. This cloud decision might then take into account the performance, cost, availability and the effects on the environment at the same time. The expansion would also contribute towards achieving sustainable, autonomous, and policy-sensitive multi-clouds data systems that can transform continuously because of the constantly evolving requirements of an enterprise.

VI. CONCLUSION

The accelerated usage of heterogeneous cloud infrastructure has led to the demand for a data platform that can offer scalability, interoperability, resilience, governance, and efficient utilization of resources among different providers. As shown in this paper, federated multi-cloud data platform architecture would traverse AWS and GCP on a hybrid notion of managing enterprise data. It offers a coordinated collection of data ingestion, federated storage, distributed processing, orchestration, metadata management, analytics, security, governance, and cross-cloud observability.

The significant value in the framework is its policy-based mechanism of workload placement that enables workloads to be distributed according to the factors of locality of data, computational needs, availability, security needs, and the state of the resources. The control and implementation plane separation further enhances portability and minimizes reliance on provider-specific implementations. Centralized metadata and observability are also enablers of data lineage and operational monitoring, fault detection, and cross-cloud governance.

The provisional evaluation process assumes that processing performance, processing scalability, availability, interoperability, resource use, cost effectiveness, and operational overhead are important measures of evaluation. The architecture may offer a justified located patronage to the enterprise that aims at decentralizing enterprise loads without causing centralization to the leadership of such institutions and their attitudes towards processes. However, cross-cloud data transfer, complexity of synchronization, provider-specific services, and other management overheads are still significant issues.

Future possibilities include AI-based workload optimization, autonomous data management lifecycle, federated analytics, intelligent security, and sustainability-conscious resource location. Broadly, the provided architecture provides insight into how a quality federated architectural manner can bring a skeleton of scaling, high-quality, controlled, and scalable multi-cloud-based business data frameworks.

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 5, Issue 1, January - February 2023 |

DOI:10.15662/IJEETR.2023.0501008

REFERENCES

[1] M. S. Aslanpour *et al.*, "Performance evaluation metrics for cloud, fog and edge computing: A review, taxonomy, benchmarks and standards for future research," *Internet of Things*, vol. 12, 2020.
[2] S. S. Gill *et al.*, "Transformative effects of IoT, blockchain and artificial intelligence on cloud computing: Evolution, vision, trends and open challenges," *Internet of Things*, 2019.
[3] A. H. Ali *et al.*, "Recent trends in distributed online stream processing platform for big data: Survey."
[4] A. A. Alli *et al.*, "The fog cloud of things: A survey on concepts, architecture, standards, tools, and applications," *Internet of Things*, 2020.
[5] M. Bendechache *et al.*, "Simulating resource management across the cloud-to-thing continuum: A survey and future directions," *Future Internet*, 2020.
[6] E. A. Brewer, "Kubernetes and the path to cloud native."
[7] Y. Chen *et al.*, "Exploring the use of synthetic gradients for distributed deep learning across cloud and edge resources."
[8] R. N. Calheiros *et al.*, "CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Software: Practice and Experience*, 2011.
[9] D. P. Abreu *et al.*, "A comparative analysis of simulators for the cloud to fog continuum," *Simulation Modelling Practice and Theory*, 2019.
[10] R. A. Cherrueau *et al.*, "Enoslib: A library for experiment-driven research in distributed computing," *IEEE Transactions on Parallel and Distributed Systems*, 2021.