Redesigning a Unified Data Analytics Architecture for a Rapidly Growing Payment Processing Company

GROUP PROJECT

GROUP 1

Group Members:

Shubham

Ankita Sirurmath

Anusha Dasaraju

Jagjeevan Kaur

Jai Raj Yadav Jakkula

Sai Rakshith Chithreddy

Topic - Revamping Data Analytics Architecture for Payment Processing Companies

Problem Statement

The current data analytics architecture in our payment acquiring and processing company is outdated and fragmented, hindering our ability to meet consumer expectations and keep up with competitors. With millions of financial transactions processed daily, our data team is struggling to reconcile data and establish a single source of truth for investor reporting and customer transaction analysis. The existing hybrid data architecture, comprising both cloud and on-premise components, is patched regularly but requires significant maintenance, resulting in limited agility for releasing new features and analyzing data. Furthermore, as we undergo rapid growth through mergers and acquisitions, we face challenges related to diverse data processing techniques, technologies, and representations of customer data across subsidiaries and divisions.

Proposed Solution: AWS Cloud-based Unified Data Analytics Architecture

To address the challenges and achieve the objectives outlined in the problem statement, we propose implementing a unified data analytics architecture using the AWS Cloud platform. This solution will provide scalability, flexibility, and enhanced security while reducing costs and improving agility in data analytics processes.

Data Architectural Strategy: Our Approach

We will migrate our data infrastructure to AWS Cloud, utilizing services like Amazon S3 for storage, Amazon Redshift for data warehousing, and Amazon EMR for big data processing. AWS Glue will automate ETL processes, while AWS Glue DataBrew ensures data cleansing and standardization.

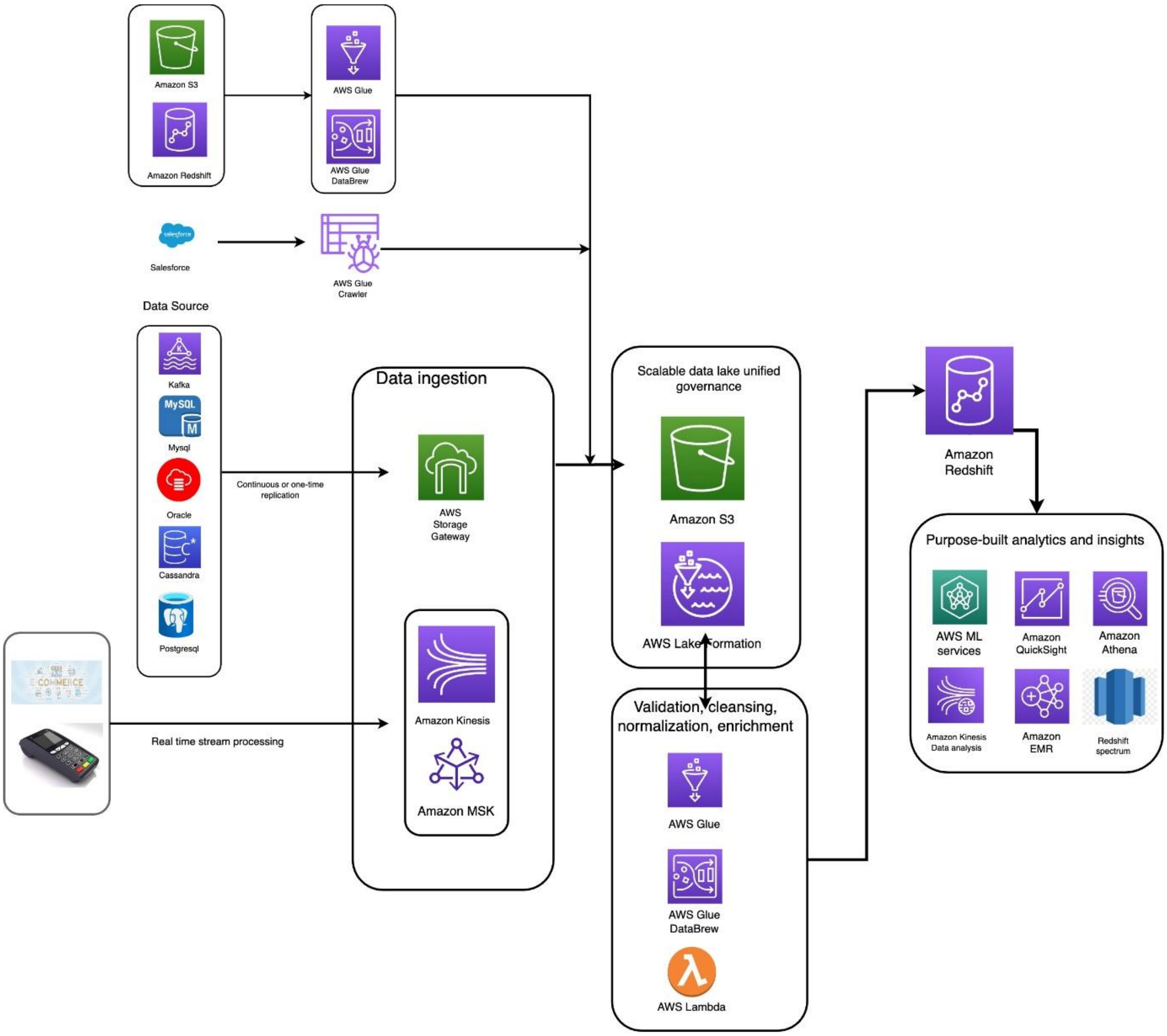
Data security will be maintained through encryption using AWS Key Management Service (KMS) and access controls with AWS Identity and Access Management (IAM). Compliance with regulations like PCI DSS, GDPR, and SOX will be met through AWS compliance programs.

Amazon Redshift will serve as our scalable data warehouse, and data analytics will be facilitated by Amazon QuickSight for interactive dashboards and visualizations. Ad-hoc querying capabilities will be provided by Amazon Athena.

Machine learning and AI capabilities from AWS, such as Amazon SageMaker and Amazon Comprehend, will enhance analytics with advanced features like anomaly detection and sentiment analysis.

Cost optimization will be achieved through AWS's pay-as-you-go model, allowing resource scaling and reducing infrastructure maintenance expenses. AWS Cost Explorer and AWS Budgets will assist in monitoring and managing costs effectively.

Data governance and metadata management will be ensured with AWS Glue for data cataloging, enabling data lineage, discovery, and governance.



Cloud Migration and Modernization: AWS S3

We are choosing Amazon S3 as the data lake storage platform for this case study for several reasons:

Scalability: Amazon S3 is highly scalable and can accommodate the massive volume of financial transaction data that our payment acquiring and processing company handles daily. As our company experiences hyper-growth and engages in mergers and acquisitions, the scalability of Amazon S3 ensures that we can easily store and manage the ever-increasing amount of data.

Cost-effectiveness: Amazon S3 follows a pay-as-you-go pricing model, allowing us to optimize costs by only paying for the storage and data transfer we use. This cost-effective approach is essential as we aim to reduce expenses associated with maintaining and upgrading our existing data warehouse system. By leveraging Amazon S3 as a data lake, we can store large amounts of data at a lower cost compared to traditional on-premise storage solutions.

Durability and Reliability: Amazon S3 offers exceptional durability and reliability, ensuring that our data is protected and available when needed. It automatically replicates data across multiple geographically diverse data centers, reducing the risk of data loss. With Amazon S3's 99.999999999% (11 nines) durability, we can trust that our financial transaction data will be securely stored and accessible at all times.

Integration and Compatibility: Amazon S3 integrates seamlessly with a wide range of AWS services and third-party tools. This compatibility enables us to leverage other AWS services, such as Amazon Redshift for data warehousing and Amazon Athena for ad-hoc querying, to perform analytics and gain insights from the data lake. Additionally, popular data analytics and ETL tools can easily interface with Amazon S3, allowing us to leverage existing data processing pipelines and workflows.

Security and Compliance: Amazon S3 provides robust security features to protect our sensitive financial transaction data. We can enforce encryption at rest and in transit using AWS Key Management Service (KMS) and secure data access using AWS Identity and Access Management (IAM). Amazon S3 is also compliant with various industry standards and regulations, such as PCI DSS, GDPR, and HIPAA, ensuring that our data lake meets stringent security and compliance requirements.

Data Lifecycle Management: Amazon S3 offers flexible data lifecycle management options, allowing us to define rules and policies for data retention, archiving, and deletion. This feature helps us optimize storage costs by automatically moving infrequently accessed data to lower-cost storage classes, such as Amazon S3 Glacier, while keeping frequently accessed data readily available in the appropriate storage tier.

Data Accessibility and Analytics: Amazon S3 provides fast and reliable access to data, enabling efficient data retrieval and analytics. With the capability to store a wide variety of data formats, including structured, semi-structured, and unstructured data, Amazon S3 allows us to perform diverse analytics tasks, such as data exploration, machine learning, and data lake analytics using AWS services like Amazon Athena and Amazon EMR.

Based on these reasons, choosing Amazon S3 as the data lake storage platform for our payment acquiring and processing company's unified data analytics architecture is a strategic decision that aligns with our objectives of scalability, cost-effectiveness, durability, security, compatibility, and accessibility. It will enable us to leverage the full potential of our data, unlock valuable insights, and drive informed decision-making across the organization.

Data Ingestion :

We used Amazon Kinesis Firehose and AWS Gateway for data ingestion in our unified data analytics architecture for the following reasons:

Amazon Kinesis Firehose for Streaming Data: Amazon Kinesis Firehose is a fully managed service that enables real-time streaming data delivery. It allows us to directly ingest and deliver streaming data to Amazon S3, which serves as our data lake storage platform. By utilizing Kinesis Firehose, we can efficiently handle high-velocity data streams, ensuring that real-time data is processed and stored in our data lake for further analysis and insights.

Transformation Capabilities: Kinesis Firehose provides transformation capabilities that allow us to modify and pre-process streaming data before storing it in Amazon S3. These transformations include compression, encryption, data batching, and the option to use AWS Lambda functions. These capabilities enable us to optimize data storage, enhance data security, and perform necessary data transformations to align with our analytics requirements.

Reduced Transaction Costs: By using Kinesis Firehose to deliver streaming data directly to Amazon S3, we can reduce transaction costs associated with storing data. Kinesis Firehose optimizes data delivery and batching, reducing the number of write operations and optimizing storage usage in Amazon S3. This helps us achieve cost efficiency while ensuring real-time data availability for analysis.

AWS Gateway for On-Premises Data Access: AWS Gateway provides a seamless integration between on-premises data sources and the AWS Cloud, allowing us to ingest data from on-premises systems into our unified data analytics architecture. By utilizing the AWS Gateway's NFS connection, we can establish a secure and scalable connection to on-premises data sources, enabling efficient and reliable data transfer to Amazon S3.

Scalable and Cost-Effective Storage: With Amazon S3 as the backend storage for the AWS Gateway, we can leverage its scalability and cost-effectiveness for storing data from on-premises systems. Amazon S3 provides virtually unlimited storage capacity, allowing us to handle large volumes of data from diverse sources. Its pay-as-you-go pricing model ensures cost optimization by only paying for the storage resources used.

Integration and Management: AWS Gateway integrates seamlessly with other AWS services, providing monitoring, access control, and encryption capabilities. It allows us to monitor data ingestion activities, apply access control policies to ensure data security, and leverage encryption mechanisms for data protection. Additionally, AWS Gateway offers intuitive interfaces for easy deployment and management, simplifying the configuration and maintenance of the data ingestion pipeline.

Backup and Disaster Recovery: AWS Gateway enables efficient backup and disaster recovery strategies by replicating on-premises data to Amazon S3. This ensures data availability and resilience in the event of a disaster or system failure. Native support for DistCP (Distributed

Copy) facilitates fast and reliable data transfer between on-premises systems and Amazon S3, enhancing backup and recovery processes.

By utilizing Amazon Kinesis Firehose and AWS Gateway in our unified data analytics architecture, we can effectively ingest and process both streaming and on-premises data. These services offer scalability, cost-effectiveness, transformation capabilities, integration with other AWS services, and efficient data transfer, enabling us to build a comprehensive and robust data ingestion pipeline that supports our diverse data sources and analytics requirements.

Data Integration and Standardization:

With AWS Glue, we can automate the extraction, transformation, and loading (ETL) processes, enabling seamless data integration from various subsidiaries and divisions. AWS Glue DataBrew can be utilized for data cleansing and standardization, ensuring consistency and quality across the entire dataset.

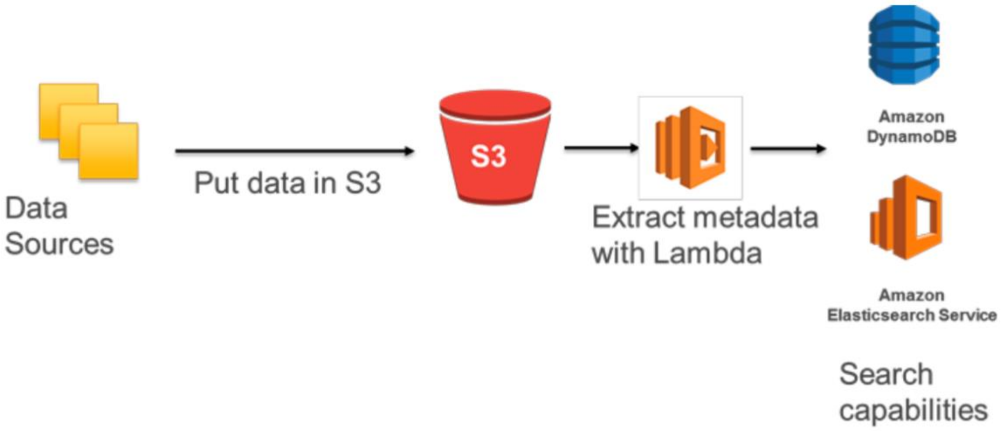
Role and Types of Data Catalogs in the Data Lake Environment

In the proposed unified data analytics architecture for our payment acquiring and processing company, a data catalog plays a crucial role in managing and organizing the data assets within the data lake environment. The data catalog serves as a centralized repository that keeps track of all the raw assets loaded into the data lake, as well as the new data assets and versions created through data transformation, processing, and analytics.

The data catalog ensures that there is a single source of truth about the contents of the data lake. It provides a comprehensive view of the available data assets, their metadata, and relationships, enabling efficient data discovery, exploration, and analysis. With the help of a data catalog, data users across the organization can easily find and understand the available data assets, reducing the time and effort required for data exploration and analysis.

There are two types of data catalogs commonly used in a data lake environment, both of which can be relevant to our case study:

Comprehensive Data Catalog: A comprehensive data catalog captures and catalogs all types of data assets within the data lake, including raw data, transformed data, and curated datasets. It maintains metadata about the data assets, such as data source, data quality, data lineage, schema information, and access permissions. This type of data catalog provides a holistic view of the entire data landscape, enabling users to understand the available data assets and make informed decisions during data analysis and reporting.



Hive Metastore Catalog (HCatalog): Hive Metastore is a metadata repository used in Apache Hive, a popular data warehousing and SQL query engine often used in data lake environments. HCatalog is a service built on top of Hive Metastore that provides a unified metadata management system. It enables storing and managing metadata about data tables, schemas, partitions, and views within the data lake. HCatalog simplifies data access and discovery by allowing users to query and analyze data using HiveQL (Hive Query Language).

In our case study, implementing a comprehensive data catalog would be beneficial as it would provide a centralized and unified view of the data assets across the data lake, including raw data, transformed data, and curated datasets. This would enable efficient data exploration, analysis, and reporting, supporting use cases such as risk and compliance monitoring, finance, merchant segmentation, sales/marketing insights, and customer 360.

Additionally, depending on the specific technology stack and requirements, integrating with a Hive Metastore Catalog (HCatalog) could be advantageous if Apache Hive or Hive-compatible tools are utilized for data processing and querying within the data lake environment. HCatalog would provide a standardized way to manage metadata and facilitate data access using HiveQL.

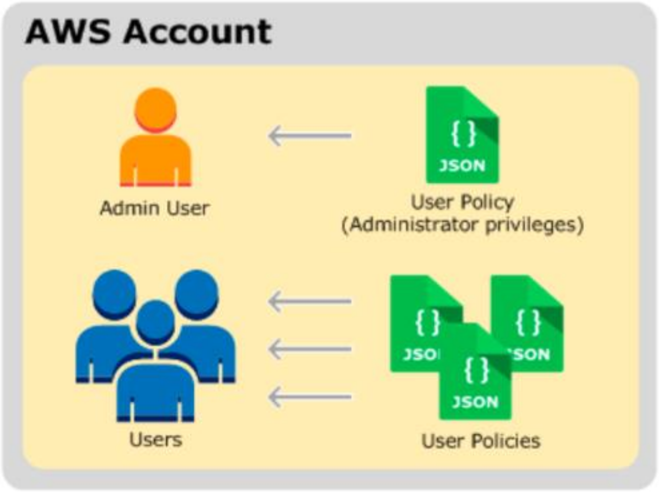
By leveraging the capabilities of a data catalog, we can establish data governance, enhance data discovery, promote data standardization, and ensure a single source of truth for data assets within our unified data analytics architecture, supporting the various use cases and objectives.

Securing, Protecting, and Managing Data in a Data Lake:

By consolidating data assets from various sources into a centralized data lake, it becomes crucial to implement robust security measures and access controls to safeguard sensitive information and ensure data integrity. Two key components for achieving this are access policy options and AWS Identity and Access Management (IAM).

Access Policy Options: Access policy options refer to the mechanisms that define who can access the data and what actions they can perform on it within the data lake. By implementing access policies, organizations can enforce fine-grained access controls, ensuring that only authorized users or groups can access specific data assets or perform certain operations on them. Access policy options can be configured at different levels, including bucket-level or object-level, to align with the required level of granularity in data access management.

AWS IAM (Identity and Access Management): AWS IAM is a service that enables the management of user identities and their permissions within the AWS environment. IAM allows organizations to define user policies, which specify what actions users are allowed or denied on AWS resources, including data assets in the data lake. User policies in IAM provide a flexible and centralized way to manage and enforce access controls based on the principle of least privilege. They allow the assignment of permissions to individual users, groups, or roles, ensuring that users have appropriate access rights based on their roles and responsibilities.



By utilizing resource-based policies and user policies with AWS IAM, organizations can establish a comprehensive security framework for the data lake:

Resource-based policies: Resource-based policies are attached to AWS S3 buckets and define the access permissions for the data stored within them. These policies specify which users or groups have access to the bucket and the level of access they have, such as read, write, or delete. Resource-based policies help control access to data assets at a bucket level and can be customized to suit specific security requirements.

User policies: User policies in AWS IAM are used to define fine-grained permissions for individual users, groups, or roles. These policies outline the actions users can perform on AWS resources, including data assets in the data lake. By configuring user policies, organizations can ensure that users have appropriate access to data based on their roles and responsibilities within the company. User policies can be tightly controlled and managed, providing a granular level of security and access control to the data assets in the data lake.

Implementing access policy options and leveraging AWS IAM in our unified data analytics architecture allows us to establish a robust security framework. This ensures that only authorized users have access to the data lake and that their permissions are defined based on the principle of least privilege. By implementing stringent security measures and fine-grained access controls, we can protect sensitive data, prevent unauthorized access, and maintain data integrity within the data lake environment.

Data Encryption with Amazon S3 and AWS KMS

By leveraging Amazon S3 and AWS Key Management Service (KMS), we can implement robust encryption measures to protect data at rest and in transit. Additionally, integrating with other AWS services such as AWS CloudTrail, Amazon API Gateway, Amazon Cognito, and IAM enhances data protection and provides auditing capabilities.

Data Encryption with Amazon S3: Amazon S3 offers various encryption options to protect data stored in the data lake. Server-Side Encryption (SSE) can be enabled, which automatically encrypts data at rest using either S3-managed keys (SSE-S3), AWS Key Management Service (KMS) managed keys (SSE-KMS), or customer-provided keys (SSE-C). By encrypting the data at rest, even if unauthorized individuals gain access to the data assets, they won't be able to view or use the data without the proper decryption keys.

AWS Key Management Service (KMS): AWS KMS is integrated with Amazon S3 to provide secure and centralized key management for data encryption. KMS allows us to create, manage, and control the encryption keys used to protect the data assets in the data lake. By utilizing KMS, we can ensure that the encryption keys are protected and managed effectively. AWS

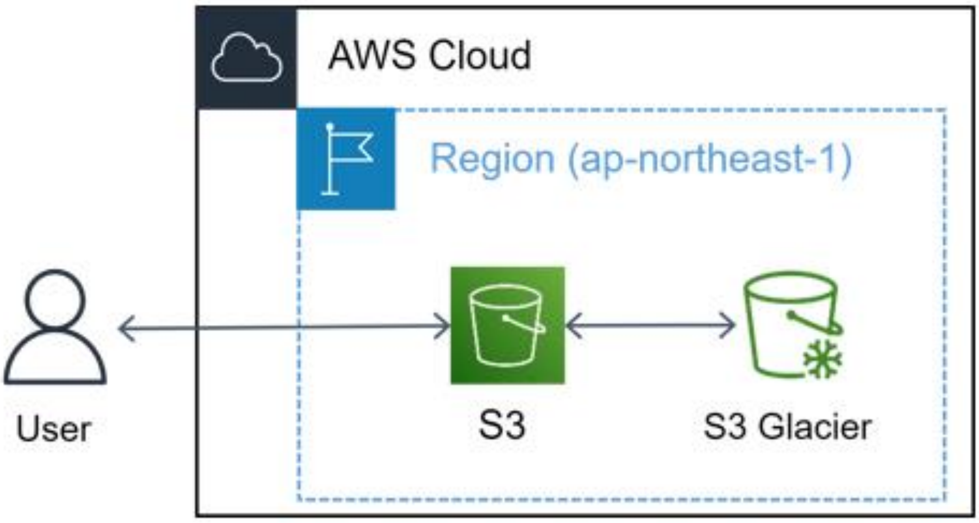
KMS also integrates with AWS CloudTrail, providing auditing capabilities to track key usage and monitor access to encrypted data.

Auditing with AWS CloudTrail: AWS CloudTrail is a service that logs and monitors API activity within the AWS environment. By integrating AWS KMS with CloudTrail, we can obtain detailed logs of key usage, including who used which keys, on which resources, and when. This auditing capability helps in identifying any unauthorized access attempts or suspicious activities related to key management and data encryption within the data lake.

Enhanced Data Protection: To further enhance data protection in the data lake, additional AWS services can be combined. For example, Amazon API Gateway can be used to create a secure and controlled access point for users to interact with the data lake. Amazon Cognito can provide user authentication and authorization, ensuring that only authorized individuals can access the data assets. IAM can be utilized to manage fine-grained access control to the data lake resources. These services, combined with data encryption, create a robust "shopping cart" model where users can securely check in and check out data lake data assets, reducing the risk of inadvertent or malicious access to sensitive data.

By implementing data encryption with Amazon S3 and AWS KMS, along with the integration of services like AWS CloudTrail, Amazon API Gateway, Amazon Cognito, and IAM, we can achieve a high level of data protection in the data lake environment. These measures ensure that even if unauthorized access occurs, the encrypted data remains secure and inaccessible. Additionally, auditing capabilities provide visibility into key usage and help in detecting and responding to any security incidents or policy violations within the data lake.

Protecting Data with Amazon S3



Data Integrity and Protection: Versioning in Amazon S3 acts as a safeguard against data corruption, loss, and accidental or malicious modifications. By enabling versioning, multiple versions of a data asset are retained, ensuring that each change to the asset is preserved. In the event of data corruption, such as due to software bugs, hardware failures, or network issues, previous versions of the data asset can be retrieved, restoring data integrity and mitigating the impact of corruption. This feature provides an extra layer of protection, as it ensures that valuable data assets are not permanently compromised by unforeseen issues.

Accidental or Malicious Data Modifications: Accidental or malicious overwrites, modifications, or deletions of data assets can have severe consequences. Enabling versioning in Amazon S3 allows us to retain previous versions of data assets, even after changes have been made. If a data asset is accidentally modified or deleted, the previous versions can be retrieved, effectively undoing the unwanted changes and preventing data loss. This feature provides a safety net, reducing the risk of irreversible damage to critical data and allowing for quick recovery from unintended data modifications.

Auditing and Compliance: Versioning in Amazon S3 provides an audit trail of changes made to data assets. Each version of a data asset is timestamped, creating a historical record of data modifications. This audit trail is valuable for compliance purposes, as it helps in demonstrating data integrity, tracking data lineage, and supporting regulatory requirements. By having a comprehensive record of data changes, organizations can prove that they have appropriate controls in place to protect and manage data assets effectively, ensuring compliance with industry regulations and data governance policies.

Data Recovery and Rollback: The versioning feature in Amazon S3 facilitates easy data recovery and rollback. If undesired changes are made to a data asset or if the data needs to be restored to a previous state, the appropriate version of the data asset can be retrieved and restored. This capability provides flexibility and agility in managing data assets, enabling organizations to revert to a known good state and recover from data incidents or errors promptly. Whether it's recovering from accidental data modifications or rolling back to a stable version, versioning simplifies the data recovery process and reduces the potential impact of data-related issues.

In summary, using Amazon S3's versioning feature enhances data asset protection in the data lake environment. It safeguards against data corruption, loss, accidental or malicious modifications, and deletions. By retaining multiple versions of data assets, organizations can ensure data integrity, quickly recover from unwanted changes, maintain compliance with regulations, and have the ability to roll back to previous states if needed. This feature provides peace of mind and strengthens the overall data management and protection strategy within the data lake.

Transform raw data assets in place into optimized usable formats.

Transforming Raw Data Assets: Raw data assets often come in various formats and structures, making them difficult to work with directly for analysis or processing. By transforming the raw data assets into optimized formats, such as columnar storage or compressed file formats, we can improve data query performance, reduce storage costs, and enable faster data processing. Transformations may involve data cleaning, normalization, aggregation, or other operations that enhance data usability and efficiency.

In-Place Data Transformation: In-place data transformation refers to performing data transformations directly within the data lake, without the need for extensive data movement or duplication. This approach minimizes data movement and reduces storage requirements, making it a more efficient and cost-effective solution. By leveraging tools and technologies available in the data lake environment, such as AWS Glue or Apache Spark, data assets can be transformed in place, optimizing their format and structure while retaining their original location within the data lake.

Object Tagging for Data Management: In a multi-tenant data lake environment, where multiple organizations, lines of businesses, users, and applications access and process data assets, it becomes crucial to associate data assets with relevant entities and set policies for coherent data management. Object tagging is a mechanism provided by Amazon S3 that allows you to assign custom metadata tags to individual data assets. These tags can represent various attributes such as organization, business unit, user, application, or data classification. By tagging data assets, you can categorize and organize them based on different criteria, enabling better data governance, access control, and policy enforcement.

Discussing in-place data transformation and object tagging is important because they address specific challenges related to data management and usability in a multi-tenant data lake environment. By transforming raw data assets in place, we can optimize their formats for efficient analysis and processing. Additionally, object tagging allows for better organization and management of data assets, ensuring that they can be associated with relevant entities and enabling the enforcement of coherent policies for data governance, access control, and data lifecycle management.

Monitoring and Optimizing the Data Lake Environment

data lake monitoring and optimization is essential to ensure the efficient operation and performance of the data lake environment.

Holistic Monitoring with Amazon CloudWatch: As the administrator of the data lake environment, it is crucial to have comprehensive visibility into the various components and resources that make up the data lake. Amazon CloudWatch is a monitoring and observability service provided by AWS that enables you to monitor and collect metrics, logs, and events from different AWS services. By utilizing CloudWatch, you can gain insights into the performance, utilization, and health of the data lake infrastructure, including storage, compute, networking, and other relevant metrics. This holistic monitoring approach allows you to proactively identify and address any potential issues or bottlenecks in the data lake environment, ensuring optimal performance and reliability.

Monitoring API Calls with AWS CloudTrail: AWS CloudTrail is a service that provides continuous monitoring and auditing of API calls made within an AWS environment, including the services that constitute a data lake. By enabling CloudTrail, you can capture detailed information about the API activities performed by various users, applications, and services interacting with the data lake. This includes tracking who made the API calls, which resources were accessed, and when the actions occurred. By leveraging CloudTrail, you can maintain a comprehensive audit trail of data lake activities, enhancing security, compliance, and governance capabilities. It helps in detecting and investigating any unauthorized or suspicious activities, ensuring the integrity and protection of data assets in the data lake.

Discussing data lake monitoring with Amazon CloudWatch and AWS CloudTrail is vital for several reasons. It allows the administrator to have a holistic view of the data lake environment, enabling proactive monitoring, troubleshooting, and optimization of the infrastructure. By leveraging CloudWatch, potential performance bottlenecks or resource utilization issues can be identified and addressed promptly, ensuring smooth operation and meeting the needs of internal teams and external stakeholders. Additionally, the use of AWS CloudTrail ensures robust security and compliance by capturing and retaining a detailed audit trail of API activities, providing visibility into who accesses the data lake and what actions they perform. This strengthens the overall data governance and security posture of the data lake environment, aligning with the company's objectives of protecting customer data and meeting regulatory requirements.

Data Lake Optimization

Amazon S3 Lifecycle Management: As the data lake grows, it becomes essential to optimize storage costs and performance. Amazon S3 Lifecycle Management allows you to automate the transition of data between different storage tiers within S3 based on predefined rules. This feature helps optimize costs by automatically moving less frequently accessed data to lower-cost storage tiers, such as Amazon S3 Glacier or Amazon S3 Glacier Deep Archive. By moving infrequently accessed data to more cost-effective storage options, you can significantly reduce storage costs while maintaining the necessary data accessibility for analytics and compliance purposes.

Amazon S3 Storage Class Analysis: Understanding data access patterns is crucial for optimizing storage costs in the data lake. Amazon S3 Storage Class Analysis provides insights into data access patterns by analyzing historical data access patterns and generating reports. These insights help you make informed decisions about data storage optimization, such as selecting the appropriate storage class based on data usage frequency. By leveraging the information provided by Storage Class Analysis, you can identify data that can be moved to lower-cost storage tiers or optimized for improved performance, ultimately optimizing storage costs and overall data lake efficiency.

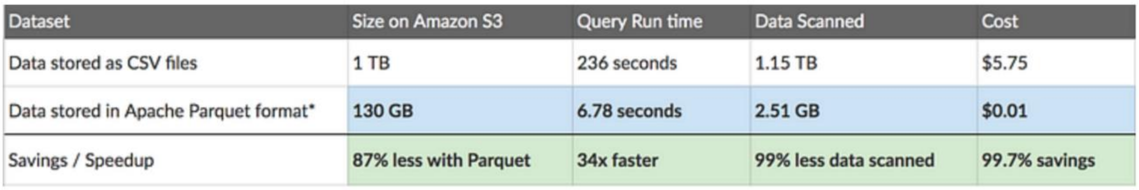
Amazon Glacier: Amazon Glacier is a secure and durable archival storage service designed for long-term data retention. In the data lake environment, there may be data that is accessed infrequently but still needs to be retained for compliance or historical purposes. Amazon Glacier provides a cost-effective storage option for such data, offering significantly lower storage costs compared to standard storage classes. By leveraging Amazon Glacier, you can store infrequently accessed data in an economical manner while ensuring its durability and availability when needed.

Cost and Performance Optimization

Reduced Storage Costs: Apache Parquet is a columnar storage format that offers efficient compression and encoding techniques. By using Parquet, you can significantly reduce the storage footprint of your data lake. The columnar storage format allows for better compression rates as similar data types are stored together, resulting in smaller file sizes. With smaller file sizes, you can save on storage costs as less physical storage space is required to store the data. This is particularly important in a data lake environment where large volumes of data are processed and stored.

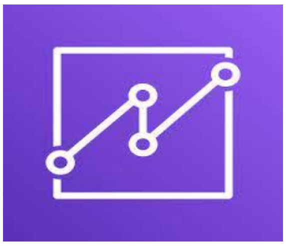
Improved Analytics Querying Performance: The columnar nature of Apache Parquet makes it highly optimized for analytics querying. Since each column is stored separately, query engines can selectively read only the columns relevant to the query, minimizing disk I/O and improving query performance. This allows for faster data retrieval and processing, enabling more efficient and responsive analytics workflows. Improved query performance translates to faster insights, reduced query execution times, and better overall data lake performance.

Compatibility and Interoperability: Apache Parquet is widely supported by various analytics and processing frameworks, including Apache Spark, Amazon Athena, and AWS Glue. This compatibility ensures seamless integration and interoperability across different components of the data lake ecosystem. You can leverage the benefits of Apache Parquet across the entire analytics pipeline, from data ingestion to transformation and analysis, without the need for format conversions or data movement. This reduces complexity, improves data processing efficiency, and enhances overall data lake performance.



Dataset	Size on Amazon S3	Query Run time	Data Scanned	Cost
Data stored as CSV files	1 TB	236 seconds	1.15 TB	$5.75
Data stored in Apache Parquet format*	130 GB	6.78 seconds	2.51 GB	$0.01
Savings / Speedup	87% less with Parquet	34x faster	99% less data scanned	99.7% savings

Reporting and Analytics



Amazon EMR, Amazon Machine Learning, Amazon QuickSight, and Amazon Data Analysis, because they play a crucial role in enabling effective data analysis, visualization, and reporting within the data lake environment. Here's why we are talking about these tools:

Amazon EMR (Elastic MapReduce): Amazon EMR is a managed cluster platform that simplifies big data processing and analysis. It allows you to easily process and analyze large datasets using popular distributed processing frameworks like Apache Spark, Apache Hadoop, and Apache Hive. EMR provides a scalable and cost-effective solution for running data-intensive workloads, including data transformations, aggregations, and complex analytics. By leveraging EMR, you can perform advanced data processing tasks within the data lake, facilitating the generation of insights and supporting reporting requirements.

Amazon Machine Learning: Amazon Machine Learning (Amazon ML) is a cloud-based service that provides a robust platform for building and deploying machine learning models. With Amazon ML, you can utilize machine learning algorithms to gain valuable insights from your data lake. This includes tasks such as predictive analytics, anomaly detection, and customer segmentation. By leveraging the capabilities of Amazon ML, you can enhance your reporting capabilities by incorporating machine learning-based insights into your data analysis and reporting workflows.

Amazon QuickSight: Amazon QuickSight is a business intelligence (BI) tool that enables data visualization, exploration, and reporting. It provides an intuitive interface for creating interactive dashboards, charts, and reports using data from various sources, including data lakes. QuickSight allows users to analyze and visualize data in real-time, providing actionable insights to support decision-making. By utilizing QuickSight, you can empower stakeholders to explore and interpret data from the data lake, facilitating self-service reporting and enabling a data-driven approach to decision-making.

Amazon Data Analysis: While not specified in detail, Amazon Data Analysis likely refers to the broader suite of AWS analytics services, including AWS Glue, AWS Athena, and AWS Redshift. These services provide comprehensive capabilities for data integration, data warehousing, and interactive querying. By leveraging these services, you can perform sophisticated data analysis tasks, build data pipelines, and create data models to support reporting and analytics within the data lake environment.

In summary, we discuss Amazon EMR, Amazon Machine Learning, Amazon QuickSight, and Amazon Data Analysis because they are key reporting tools that enable data processing, analysis, visualization, and reporting within the data lake environment. These tools provide the necessary capabilities to derive insights, generate reports, and support decision-making processes based on the data stored in the data lake. By utilizing these tools effectively, organizations can leverage the power of data to drive business growth, improve operational efficiency, and gain a competitive edge in the payment acquiring and processing industry.

Conclusion

In conclusion, adopting AWS cloud in a hybrid mode offers organizations the benefits of scalability, flexibility, and cost efficiency while maintaining control over sensitive data and on-premise systems. It enables innovation, operational efficiency, and accelerates digital transformation.

References

http://highscalability.com/blog/2017/8/14/why-morningstar-moved-to-the-cloud-97-cost-reduction.html

https://aws.amazon.com/solutions/case-studies/paytm/

https://aws.amazon.com/blogs/big-data/how-paytm-modernized-their-data-pipeline-using-amazon-emr/

https://d0.awsstatic.com/whitepapers/Storage/data-lake-on-aws.pdf