ORGANIZING FOR BUSINESS ANALYTICS PLATFORMS

INTRODUCTION

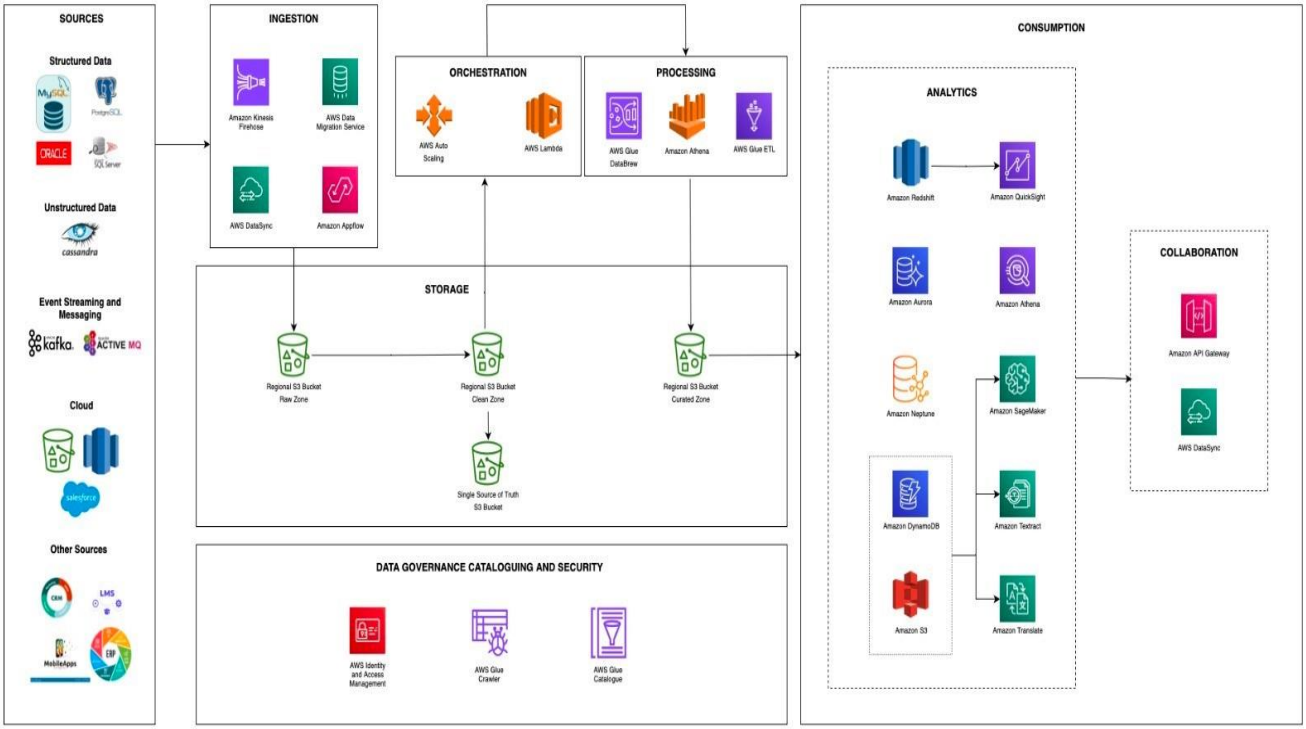
In navigating the dynamic higher education landscape, our university's data team grapples with the complexities of managing diverse datasets from both on-campus and online courses. Though attempting to adapt, the current hybrid infrastructure faces efficiency challenges due to outdated sources. With a focus on growth and global online programs, the Chancellor's office acknowledges the potential of predictive analytics and AI but recognizes the need to streamline reporting, standardize data, and enhance security. This sets the stage for a crucial, assertive push towards a modernized and strategic data management approach.

OBJECTIVE

The project aims to design and implement a Unified Data Platform to elevate our esteemed university's operational efficiency and decision-making capabilities.
Leveraging the advanced cloud services offered by Amazon Web Services (AWS), this initiative aims to seamlessly integrate and consolidate diverse data sources, providing a unified and scalable foundation for analytics, insights, and informed decision-making. By harnessing the power of AWS, we aspire to empower our university with a cutting-edge data infrastructure that not only enhances operational workflows but also opens new avenues for innovation, collaboration, and excellence in academics.

OUR APPROACH

DATA ARCHITECTURE



DATA SOURCES

Our university's data landscape comprises three primary categories of sources: cloud-based Amazon S3, Redshift, and Salesforce. On-premise sources such as MySQL, ORACLE, PostgreSQL, etc. and pre-existing systems like Microsoft (MS) applications. These sources include Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), and Learning Management System (LMS). Recognizing the inefficiency of upgrading or maintaining the existing on-premises data warehouse from both a cost and performance perspective, our strategy involves a decisive shift to a fully cloud-based architecture. To seamlessly transition our data ecosystem, we undertake a one-time migration using AWS Data Migration Service to transfer all data from the on-premises data warehouse to the cloud. This migration marks a pivotal moment, ensuring that our cloud infrastructure becomes the central repository for all our data. Subsequently, we redirect data flows from our applications to the cloud-based data lake. This strategic move not only enhances performance and scalability but also aligns with cost-effectiveness, allowing us to leverage the flexibility and efficiency of cloud resources. The consolidation of our architecture in the cloud signifies a forward-looking approach, providing the foundation for streamlined data management, advanced analytics, and responsiveness to evolving business needs.

INGESTION

Amazon Kinesis is a suite of services within Amazon Web Services (AWS) designed for real-time data streaming and analytics. Leveraging this AWS service, we aim to ingest streaming and messaging data.

AWS Data Sync

AWS Data Sync is a service provided by Amazon Web Services (AWS) that simplifies and accelerates the online transfer of large volumes of data between on-premises storage systems and AWS storage services. Leveraging this AWS service, we aim to integrate the cloud-native data.

AWS App Flow

AWS App Flow is a fully managed integration service offered by Amazon Web Services (AWS) that enables secure and seamless data exchange between AWS services and SoXware as a Service (SaaS) applications. We leverage this service to integrate data from mobile and other applications like Salesforce.

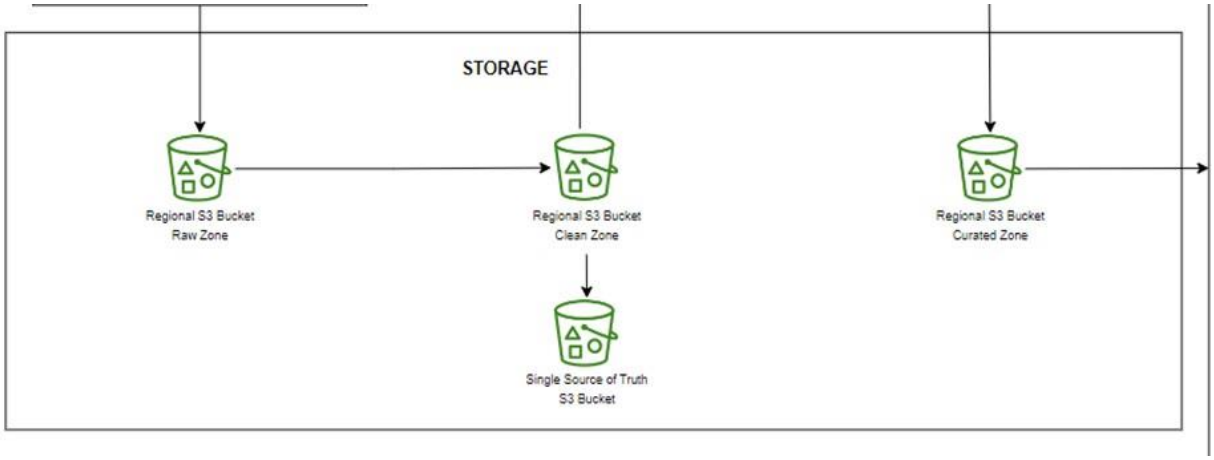
AWS Data Migra4on Service

AWS Data Migration Service is a fully managed integration service offered by Amazon Web Services (AWS) that enables secure and seamless data exchange between AWS services and SoXware as a Service (SaaS) applications leveraging this AWS service; we aim to integrate structured and unstructured data from other sources like ERP, CRM, and LMS.

STORAGE

Amazon Simple Storage Service (Amazon S3) is a scalable and secure object storage service that Amazon Web Services (AWS) provides. Here's a brief overview of its key features and functionalities: Object storage, scalability, data lifecycle management, versioning, etc.

Through the ingestion process, data arriving in the S3 bucket undergoes categorization into three distinct zones: raw, clean, and curated. Initially residing in its unprocessed state within the Raw Zone, the data transitions into the Clean Zone, where crawlers and cataloging techniques facilitate the extraction of schemas and metadata. From the Clean Zone, a replicated subset of the data is securely stored in the Single Source of Truth Bucket, serving as a global repository for consistency. Simultaneously, some data progresses into the Processing and Orchestration Zone, ultimately finding its place in the Curated Zone. This comprehensive data management strategy ensures a structured and refined storage approach, optimizing accessibility and analysis.



Raw Zone- In the Raw Zone, the data is initially stored in its original, unaltered form. This zone is a secure and untouched storage area where the raw data is preserved for auditability and traceability.

Clean Zone- From the Raw Zone, data moves to the Clean Zone, where preliminary cleansing and formatting occur. This zone ensures the data is consistent and adheres to basic quality standards, preparing it for downstream processing.

S3 Curated Zone- Post-processing, the curated data is stored in the S3 Curated Zone, representing the refined and enhanced dataset. This zone acts as an intermediate stage where data, now enriched and processed, awaits further consumption while still readily accessible.

Single Source of Truth S3 Bucket- A dedicated S3 bucket is the Single Source of Truth (SSOT), containing the data's most accurate and validated version. This SSOT S3 bucket ensures that all downstream processes and applications access a consistent and reliable dataset, reducing discrepancies and ensuring uniformity.

ORCHESTRATION

A hybrid approach combining AWS Auto Scaling and AWS Lambda offers the most flexibility and scalability. Allowing one to leverage the strengths of both services and use both depending on the situation to optimize performance, cost, and management complexity for diverse needs.

AWS AUTO SCALING AND AWS LAMBDA

1. **Dynamic Scaling for Online Courses**: AWS Lambda's horizontal scaling is invaluable in the context of online courses, where enrollment numbers can fluctuate. It ensures that the infrastructure scales dynamically to handle varying workloads, accommodating peak times during course registrations or assessments without manual intervention.
2. **Real-4me student analy4cs** - AWS Lambda can process real-time data streams related to student activities, such as course enrollments, assignment submissions, and attendance records. This enables the university to gain instant insights into student engagement and academic performance, facilitating timely interventions and support.
3. **Efficient Data Validation and Integration** - AWS Lambda can be crucial in validating and integrating data from diverse sources like CRM, ERP, and LMS systems. By automatically validating and transforming data as it flows through the pipeline, Lambda ensures data consistency and reliability across different university systems.

PROCESSING

The combination of Athena and Glue gives the right balance between different university's use cases. Athena remains suitable for simple ad-hoc analysis, while Glue provides a comprehensive suite of tools for data processing and preparation.

AWS GLUE

- Complex data processing: Transforming data from diverse sources like ERP, CRM, LMS, and mobile apps requires Glue's ETL capabilities for data extraction, transformation, and loading.
- Dynamic segmentation: Glue's flexibility allows for creating various student segments based on specific criteria, including dropouts, failures, financial hardship, and academic performance.
- Student/staff 360 view: Combining data from various sources into a unified view necessitates Glue's data integration and processing capabilities.
- Sales/marketing insights and alerts: Glue facilitates building data pipelines that extract and analyze student data for targeted marketing campaigns and personalized outreach.

AWS GLUE DATA BREW

- Provides a visual interface for cleaning and preparing data before it is used in Glue ETL jobs. Enables users to clean and normalize data without writing any code. Using Data Brew helps reduce the time it takes to prepare data for analytics and machine learning (ML) by up to 80 percent.

AMAZON ATHENA

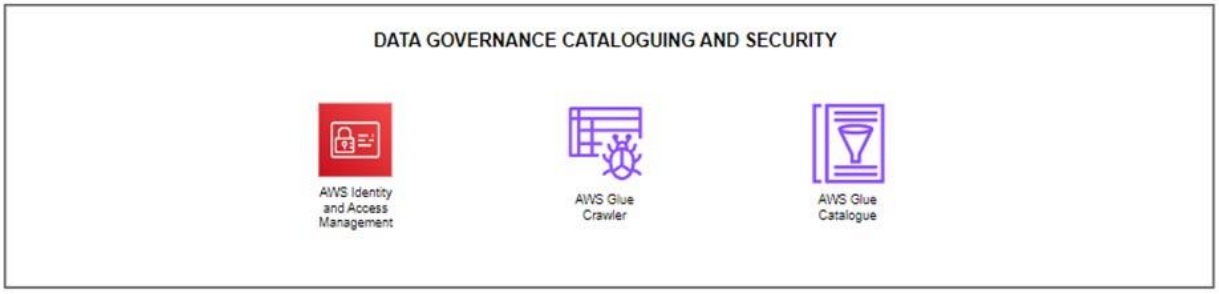
- For basic transformation & analyzing student data, gauging staff effectiveness, and identifying sales/marketing opportunities and analyzing datasets stored in Amazon S3 using standard SQL queries.

DATA GOVERNANCE, SECURITY AND CATALOGUING

AWS IAM

AWS Identity and Access Management (IAM) is a crucial service provided by Amazon Web Services (AWS) that allows our university to securely manage access to AWS resources. IAM enables the control and configuration of user permissions and access to AWS services and resources.

AWS IAM controls access to the Glue Catalogue, allowing the University to implement granular access controls. This ensures that only authorized personnel, such as administrators and data stewards, can access sensitive metadata, supporting data security and governance.



AWS GLUE CRAWLER

AWS Glue Crawler is a service provided by Amazon Web Services as part of the AWS Glue data integration service. It is designed to discover automatically and catalog metadata from various data sources, making it easier to analyze, query, and transform the data.

AWS Glue Crawler systematically scans data sources, including on-premises databases and cloud storage, to create a complete data catalogue. This catalogue is crucial for maintaining a unified view of various data assets across the University.

Glue Crawler's automated discovery process enables efficient data exploration. This is particularly valuable when introducing new courses or programs, as it ensures that relevant data is cataloged promptly for analytics and reporting purposes.

DATA CATALOGUING

AWS provides a comprehensive data cataloging service through the AWS Glue Data Catalog. The AWS Glue Data Catalog is a centralized metadata repository that enables easy discovery, management, and understanding of your data assets.

The Glue Catalogue organizes metadata related to student information, courses, and curriculum. This centralized metadata repository streamlines data governance and ensures consistency in managing and accessing information.

The Glue Catalogue catalogs individual data assets and captures relationships and lineage information. This is vital for understanding the origin and impact of data and supporting data quality management and compliance efforts.

DATA CONSUMPTION

AWS ANALYTICS

AWS extends powerful tools tailored for academic and research endeavors, enabling universities to unleash Data Insights. AWS services like Amazon Redshift and Amazon EMR provide robust analytics capabilities, empowering universities to delve into complex research datasets and derive valuable insights for academic advancement.
Machine Learning Applications: Leveraging tools like Sage Maker, universities can explore and deploy machine learning models, facilitating research breakthroughs and academic advancements through data-driven innovation.

Efficient Data Processing: With AWS services, universities gain access to streamlined data processing mechanisms, accelerating research workflows, and enabling swift data-driven conclusions.

These AWS tools empower universities to harness the potential of data analytics and machine learning, revolutionizing academic research and fostering innovation across various disciplines.

Amazon Quick Sight is a cloud-based business intelligence (BI) service provided by Amazon Web Services (AWS). It is designed to help organizations easily create interactive and shareable dashboards, perform ad-hoc analysis, and derive meaningful insights from their data.

With automatic data refresh capabilities, Quick Sight ensures that decision-makers access realtime information crucial for monitoring and responding to changes in student demographics, academic performance, and institutional effectiveness. The scalability of Quick Sight accommodates the varying analytical demands of different university departments, from academic affairs to student services.

COLLABORATION

API Gateway

Amazon API Gateway follows the principle of providing a secure and scalable gateway for applications to interact with university data and services.
API Gateway uses endpoints to define the URLs through which applications can access University data.

Web Application Integration: API Gateway facilitates the integration of university services with web applications, ensuring a seamless user experience.

Mobile App Integration: API Gateway is commonly used to integrate University data and services with mobile applications, providing students and faculty real-time access to information.

APIs for Student Portals: Amazon API Gateway is leveraged to create APIs that power student portals, allowing seamless access to personalized information such as grades, schedules, and announcements.
Integration with Learning Management Systems: API Gateway enables the integration of university services with Learning Management Systems, ensuring a unified experience for students and faculty.

Secure Data Access for Alumni Platform: API Gateway creates secure APIs for an alum platform, allowing graduates to access and update their information securely.

AWS Data Sync

AWS Data Sync employs data transfer agents to move data between on-premises and AWS storage services efficiently and securely.
Efficient Data Transfer: AWS Data Sync adheres to principles of efficiency and reliability in transferring data between different storage environments.

Hybrid Cloud Architectures: Data Sync is utilized in hybrid cloud architectures, efficiently transferring data between on-premises data centers and AWS.

Backup and Restore: Organizations use Data Sync for backup and restore processes, ensuring data durability and availability.

Data Transfer for Research Collaboration: AWS Data Sync transfers large datasets between onpremises research storage and Amazon S3, facilitating collaborative research efforts.

Effective Restoration of Research Data: AWS Data Sync guarantees a dependable and effective migration of research information from locally hosted storage to Amazon S3 for backup and archiving needs.

CONCLUSION

In conclusion, the integration of these enabling strategies serves as a cornerstone of our university's commitment to fostering a resilient, flexible, and compliant data infrastructure. Through the embrace of these strategic initiatives, we are dedicated to establishing an efficient, well-governed, and agile data ecosystem that not only meets the demands of our current data landscape but also lays the foundation for continuous innovation and informed decisionmaking. By prioritizing adaptability, robust governance, and compliance with industry standards, we are equipping our institution with the tools necessary to navigate the complexities of a rapidly evolving educational environment. This forward-looking approach not only ensures the integrity and security of our data but also positions us to leverage data-driven insights effectively, empowering our university community to excel in their respective roles and contribute meaningfully to our overarching mission of academic excellence.

REFERENCES

- https://aws.amazon.com/blogs/architecture/optimizing-data-withautomated-intelligent-document-processing-solutions/
- https://aws.amazon.com/blogs/big-data/unstructured-datamanagement-and-governance-using-aws-ai-ml-and-analytics-services/
- https://aws.amazon.com/neptune/
- https://thenewstack.io/amazon-aurora-vs-redshift-what-you-need-toknow/
- https://toluorina.medium.com/etl-on-aws-part-3-ingesting-data-from-amysql-rds-database-into-s3-based-data-lake-using-aws-52ab15a7623e