Group 9 (BUAN-6335-SEC-501) Group Project
Members :- Sahil Yadav, Siddhesh Daphane, Piyush Yerpude, Siddhant Bhujade, Saad Ur Rehman

INTRODUCTION

The Chancellor's Office at Major University has embarked on a strategic initiative to modernize its data and analytics capabilities. As enrollments spike in new online programs, the limitations of current infrastructure are being exposed. Legacy systems, siloed data, and rising costs threaten the University's ability to provide robust services that drive student success.

A cross-functional team has been tasked with designing a unified, cloud-based platform to consolidate the existing hybrid environment. This new architecture aims to enable advanced analytics, improve data governance, reduce expenses, and scale to accommodate rapid growth.

Problem Background

Over the last decade, Major University has added a multitude of new systems as on-campus and online courses expanded. With thousands of additional students and hundreds of cohorts, managing the accompanying data has become an pressing challenge.

The current environment relies on dated solutions that create data silos and drive complexity. A mix of on-premise and cloud databases power vendor and custom applications, resulting in fractured views of information. As duplicative systems proliferate, costs and data discrepancies balloon.

The outdated infrastructure hamstrings administrators' ability to quickly deploy new capabilities or generate insights through modern techniques like machine learning. Lacking standardization and a "single source of truth" puts institutional analytics and reporting at risk.

OBJECTIVE: -

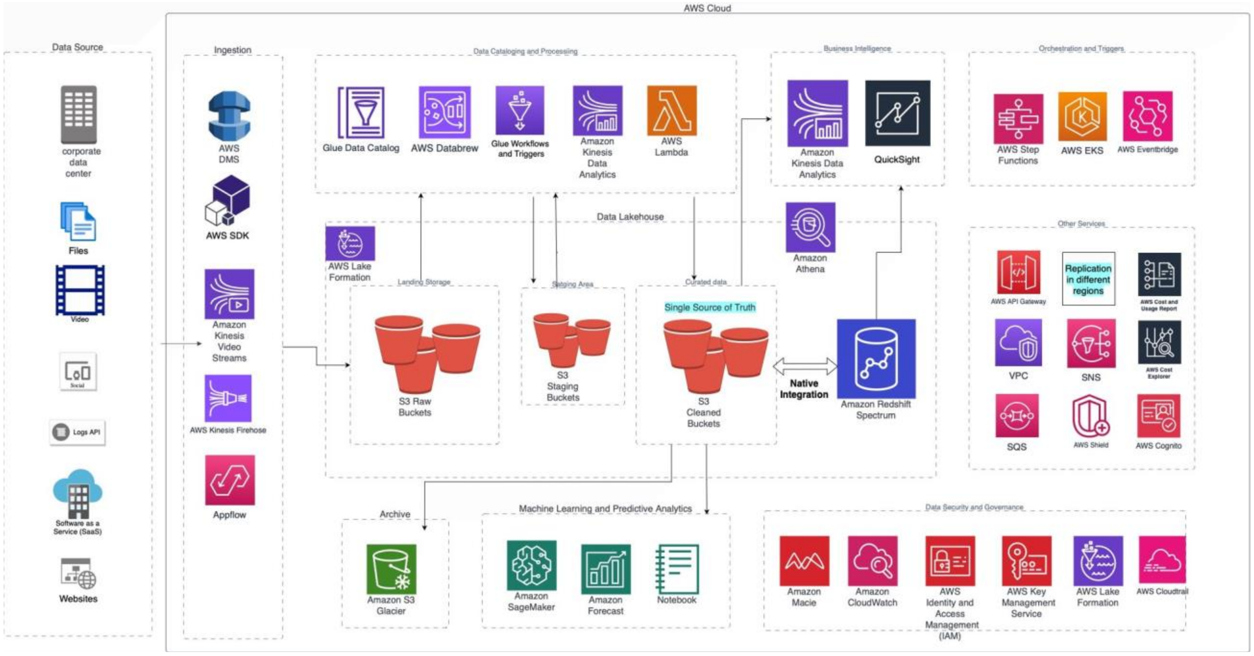
Our objective is to design and implement a robust, scalable data analytics platform on AWS cloud that serves as the centralized data repository for the university. The key goal is to unify disparate siloed data sources scattered across legacy on-prem systems, SaaS applications and existing cloud deployments into an integrated data lake powered by capabilities for seamless batch and real-time data ingestion.

Sophisticated data processing will structure unstructured content while master data management will maintain data accuracy and single source of truth for key entities like students, employees and courses. Easy self-service access coupled with customizable dashboards, strong governance standards encompassing security, privacy and access controls will enable democratized analytical insights for both technical and business users.

The platform aims to unlock advanced analytics through machine learning to answer critical questions around student outcomes, optimal curriculum design, compliance monitoring etc. Continual enhancement of predictive models and AI capabilities will establish thought leadership for the university. Through reliable, scalable, cost-optimized design leveraging AWS cloud, we target driving automation into data-intensive processes to inform strategic decisions that significantly improve educational experiences and business efficiency.

Proposed Solution: -

The proposed cloud-native solution leverages AWS Lake Formation to establish a scalable data lake ingesting batch and real-time streams. AWS Glue data catalog classifies datasets while Spark processes transformations for analytical readiness. Granular access controls, encryption enforced through Lake Formation coupled with CloudTrail, Macie enable trusted governance. Kinesis Firehose, AppFlow and EventBridge drive continuous data integration. Redshift Spectrum and SageMaker power ad-hoc querying over vast data scale and runway for advanced machine learning in a serverless fashion - collectively realizing a high-performance, cost-efficient analytics environment.

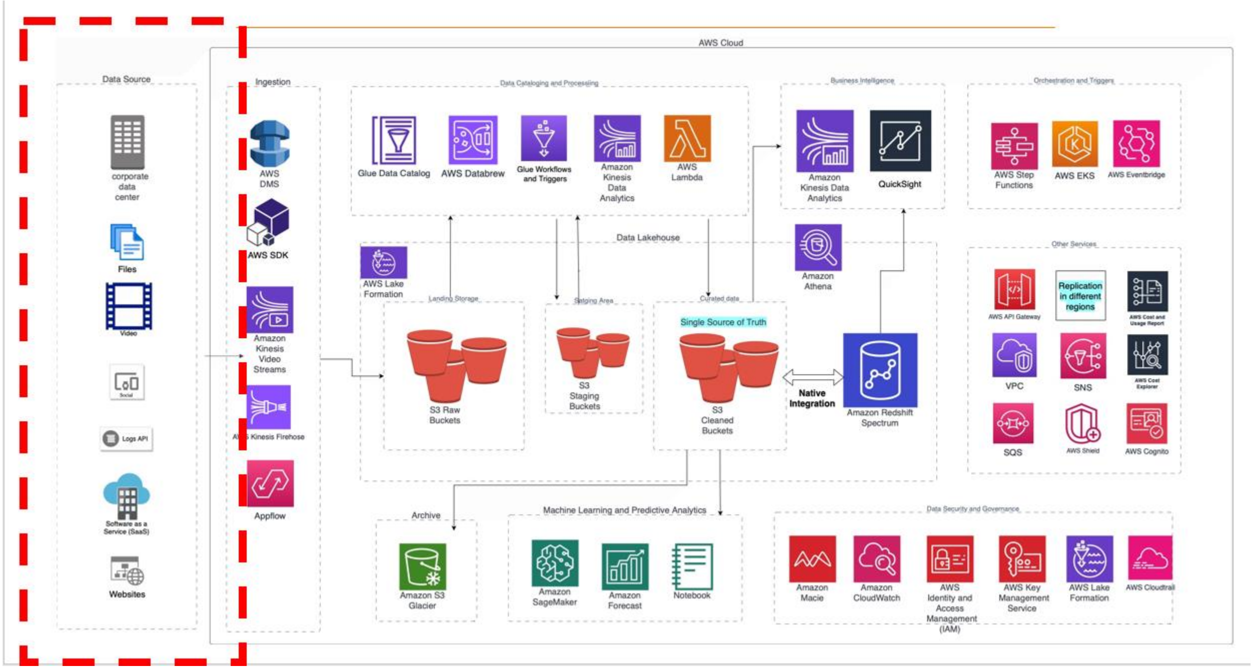


DATA SOURCES

University Data System may collect upstream data from various sources, including:

1. Corporate Data Center: These encompass internal data reservoirs, including structured data from various departments such as sports management, facilities, library, finance, and more.
2. Files: The compilation comprises a variety of file types, including CSV (Comma-Separated Values) and Parquet, which offer diverse data structures and formats to accommodate and streamline the storage, accessibility, and utilization of information across the platform.
3. Video: Within the realm of multimedia content, the dataset extends to encompass video-based sources, specifically targeting online data classes. This facet of our data compilation incorporates diverse video materials, contributing to a comprehensive reservoir that facilitates a nuanced exploration of educational and informational content available through online courses and classes.
4. Social: This facet of our dataset encapsulates a broad spectrum of information gleaned from various social media platforms, encompassing data stemming from advertising initiatives, promotional campaigns, and engagement activities across these diverse online channels.
5. Logs API: These logs serve as a rich repository of information, documenting and cataloging a myriad of events, transactions, and activities.
6. Software as a Service (SaaS): This is derived from various key sources, including but not limited to Customer Relationship Management (CRM) data, Salesforce data, and Google Analytics. This intricate compilation brings together insights from customer interactions, sales and lead management through CRM, the Salesforce platform's dynamic repository, and the comprehensive analytics provided by Google Analytics.
7. Websites: This consists of unstructured data from third party websites like e-learning, canvas, and blackboard.

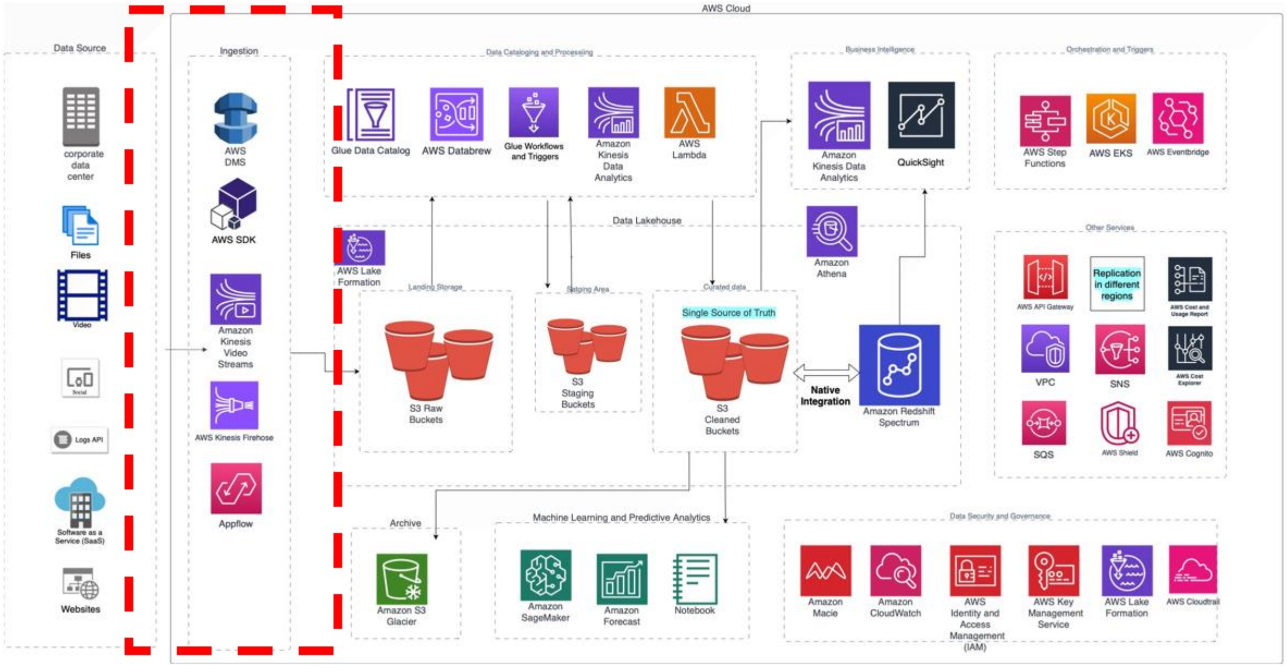
In our revamped data analytics architecture, the paramount objectives include prioritizing data security, scalability, and cost optimization. Our goal is to eradicate redundant data reporting, streamline data integration, and facilitate effortless access to a unified data structure. We're enhancing downstream reporting and business systems through microservices, catering to the diverse requirements of developers, finance teams, analysts, and data scientists. Emphasizing efficient storage utilization, we support historical and incremental data models for seamless data preparation, ensuring consistency and eliminating irregularities. This strategic approach ensures a robust foundation for advanced analytics and data-driven decision-making.



INGESTION

Data will be ingested through the following services:

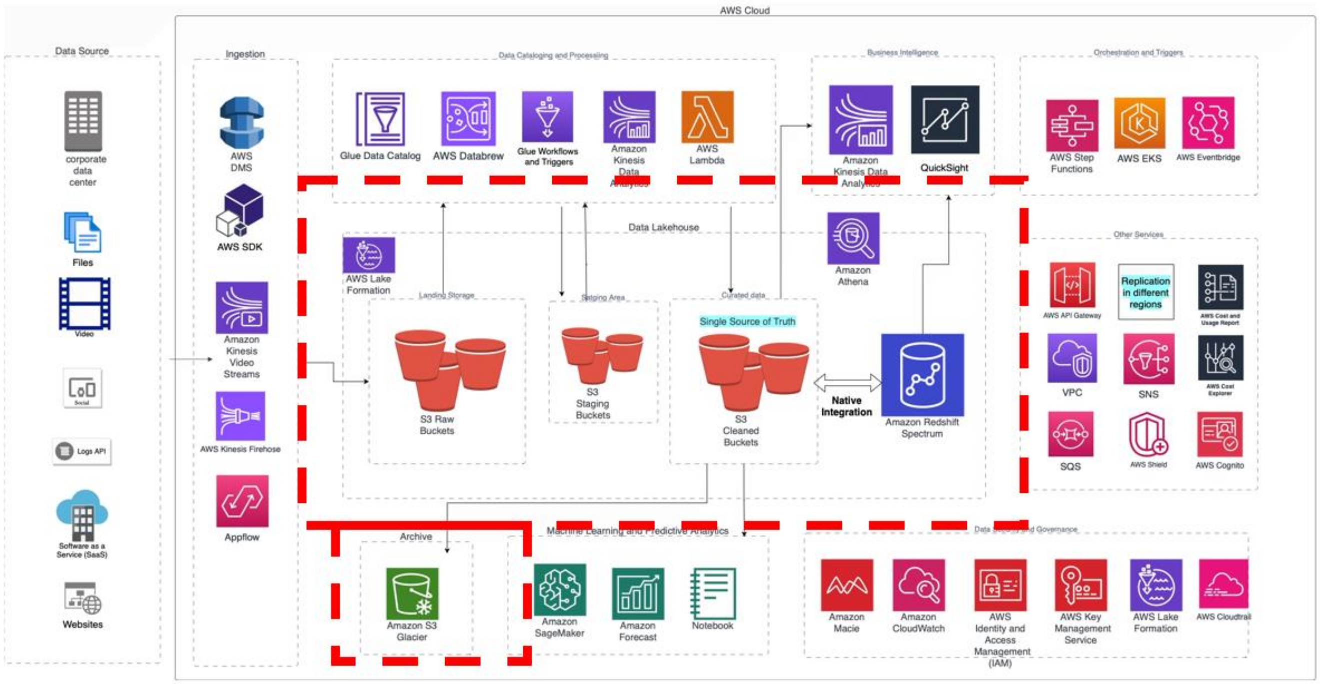
1. DMS: Facilitating the migration of on-premise data to the cloud, our solution offers comprehensive services encompassing data discovery, schema conversion, and seamless data migration. This streamlined process ensures a smooth transition with enhanced accessibility and efficiency.
2. SDK: Capable of ingesting both Parquet and CSV files, our solution is also a versatile platform utilized for writing Python code, providing a dynamic environment for diverse data processing tasks.
3. Kinesis Video Streams: Designed to efficiently ingest streaming data, our platform excels in handling dynamic information sources, such as data generated from online classes. This capability ensures real-time processing and analysis for timely insights.
4. Kinesis Firehose and Data Streams: This platform incorporates a serverless streaming data service, simplifying the seamless capture, processing, and storage of data streams at any scale. This ensures effortless scalability and efficient management of dynamic data sources.
5. Appflow: Ensuring secure data transfer, our solution seamlessly moves data from SaaS services to AWS destinations such as S3 and Redshift. This robust process guarantees a reliable and protected transition of information.



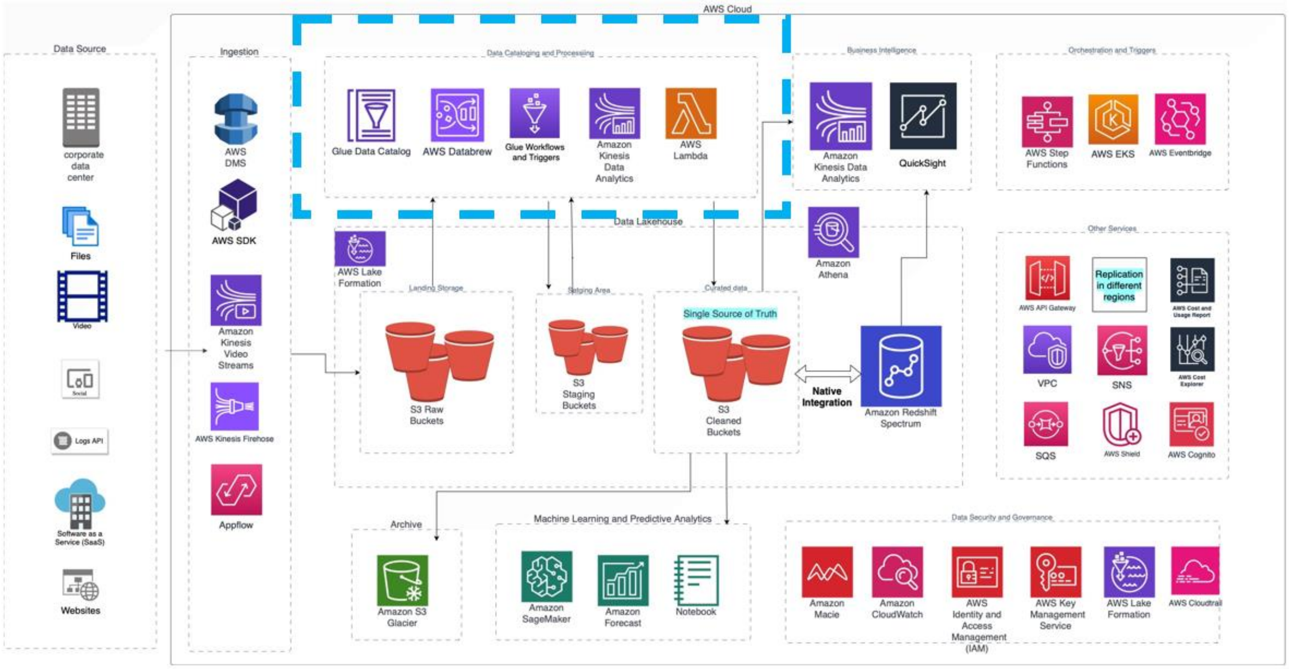
DATA STORE

AWS Lake Formation is explained through:

1. S3 Buckets: Upon ingestion, the raw data is stored in distinct S3 buckets, and subsequently undergoes processing via a staging method. This organized approach ensures efficient data management and processing workflows.
2. **Single Source of Truth:** The conclusive, cleaned data is housed in various S3 buckets within this repository, providing an organized and accessible storage solution. This arrangement facilitates easy retrieval and utilization of the refined dataset.
3. AWS Redshift Spectrum: Integration of data from the cleaned S3 buckets to this repository takes place, with Redshift Spectrum serving as the data warehouse for storing information intended for subsequent data analysis. This setup ensures a centralized hub for analytical insights and streamlined data processing.
4. AWS Athena: It is intended for conducting local queries, providing a user-friendly interface for extracting specific information and insights from the dataset. This feature enhances the efficiency of localized data exploration and analysis.
5. AWS S3 Glacier: It serves as a repository for archival data with a retention policy extending over 25 years, accommodating infrequently accessed information. This ensures a dedicated storage solution for long-term data preservation.



DATA CATALOGING AND PROCESSING



AWS Glue:
- Managed serverless Spark service for ETL data transformation and integration
- Flexibly handles both batch and streaming workloads
- Contributes jobs, crawlers, triggers and workflows to the unified Glue environment

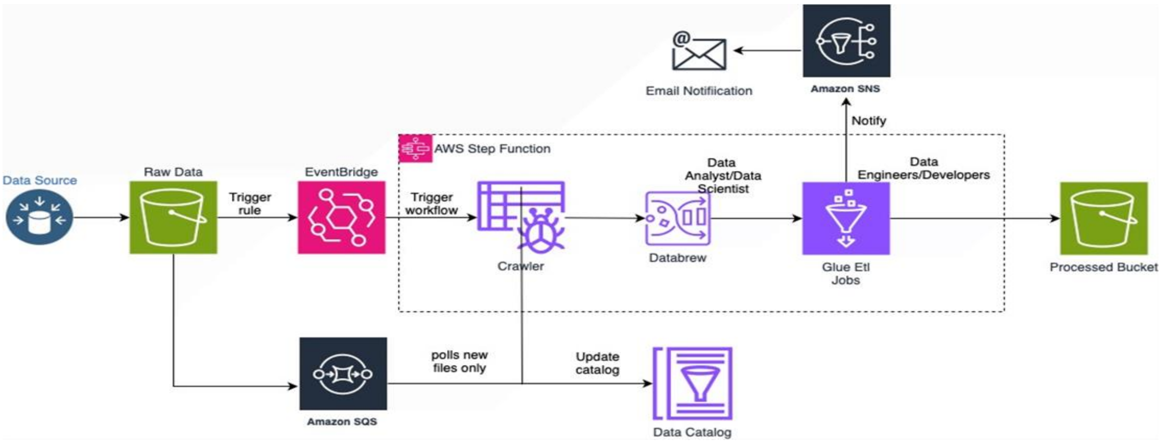
Glue Data Catalog:
- Centralized metastore to register, govern and discover data assets on AWS
- Provides rich integration across Athena, Redshift, Glue and related services
- Track data provenance and apply policy tags for compliance

DataBrew:
- No-code data preparation tool for Spark-based cleaning and shaping
- Visual workflows enable in-application transformation against data sources
- Lowers barrier for business users and data scientists preparing their own local datasets.
- Easy to use for Data Analysts and Data Scientist.

Kinesis Data Analytics:
- Enables real-time analytics on streaming data
- SQL or Apache Flink code analyzes records in motion vs at rest
- Autoscale without capacity planning
- Piped results output to variety of sinks

AWS Lambda:
- Serverless compute service to execute application logic as functions
- Helper scripts encapsulated to trigger on variety of workloads
- Handles ephemeral event-driven processing without provisioning infrastructure

BATCH PROCESSING PIPELINE



The workflow begins with new data files arriving in an S3 Raw landing zone from campus systems like the SIS (Student Information System).

Event Bridge allows chaining or kicking off dependent processes. Once new data comes into S3 raw buckets. It can trigger the Steps Functions workflow.

Step Functions are used to orchestrate cross-system workflows.

AWS Glue crawlers would classify the new data, extracting schema and technical metadata to catalog datasets for discovery in the AWS Glue Data Catalog. Crawlers can be triggered via events or scheduled.

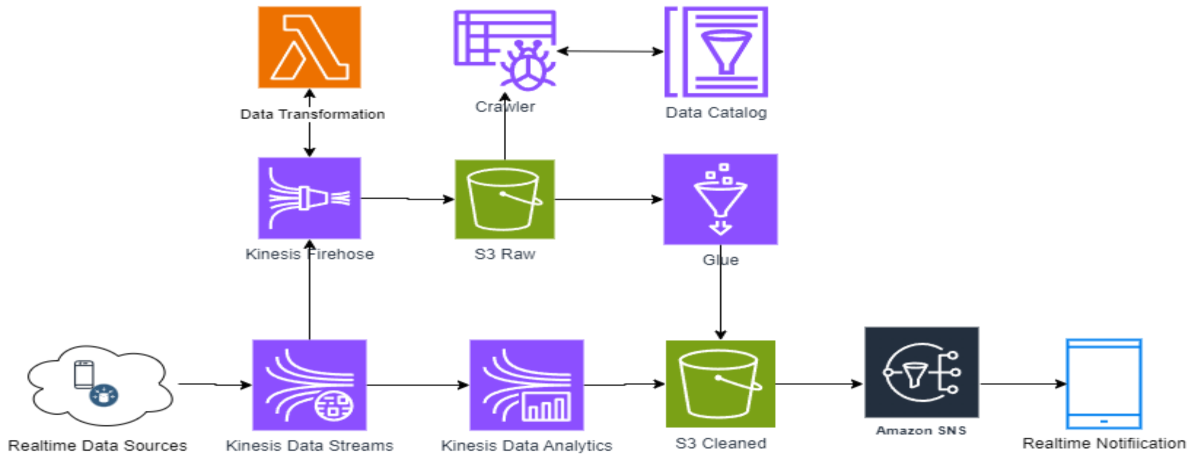
Glue job triggers would initiate the ETL processing on the batches of the data. Spark jobs transformation logic handles validation checks, enforcing data quality standards, integrating reference data, generating aggregates and analysis-ready datasets for consumption.

SNS is used for email alerts to data consumers on new dataset availability.
SQS service is implemented to have the recent data in the queue to be picked by crawler for change data capture and metadata discovery.

Processed student data zones in S3 enable direct access for BI tools like AWS QuickSight for reporting or analytics applications leveraging the polish datasets.by loading tables in Redshift.

The workflow delivers batch data to users with appropriate lineage, description, and quality enforcement - enabling analytics trust.

REALTIME STREAMING PROCESSING PIPELINE



At the ingestion layer, real-time data sources like application logs, website clicks or IoT sensor streams are captured using Amazon Kinesis Data Streams and Data Firehose. These provide managed, scalable transport streams to land raw events durably into S3 Raw storage layers.

For example campus badge swipe door access logs could flow to the raw area.

As the streams arrive, they can optionally be processed by Kinesis Data Analytics for initial querying or aggregations before landing in storage. Simple counts could be done on entry/exit timestamps for example.

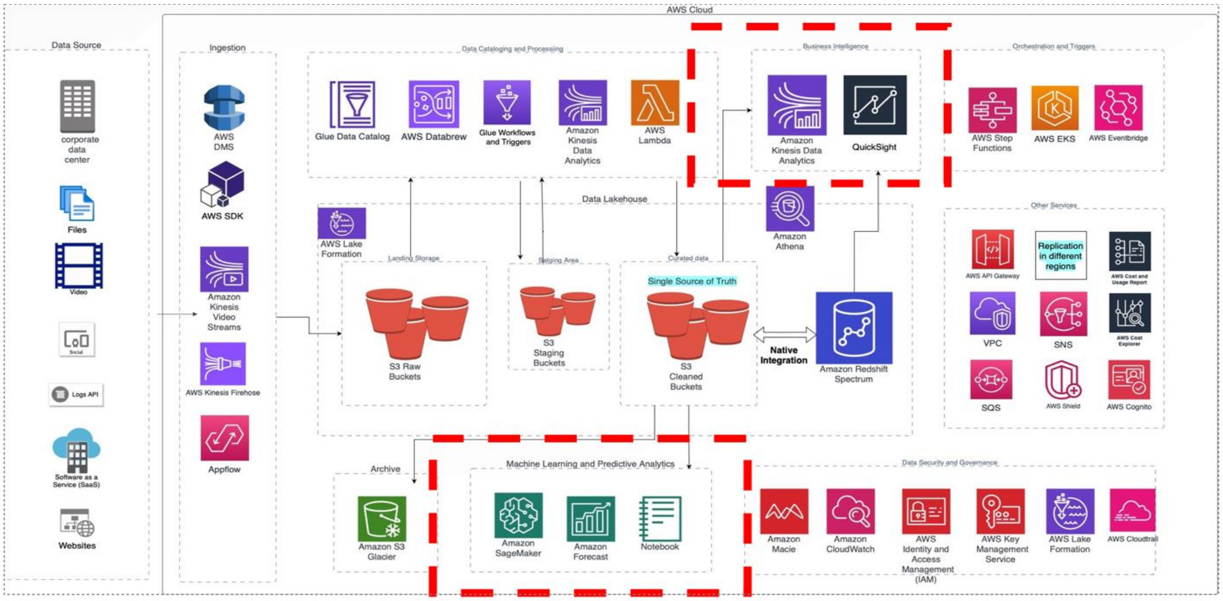
Next, If the data does not require decisions to be made in real time, it will be stored in S3 and S3-stored raw stream objects can trigger AWS Glue crawlers to classify and catalog the new partitions, essential for data discovery.

Serverless Spark jobs are also shown to pull the arriving streams from S3 Raw, execute any data quality rules, transformations and output cleaned, compliant analytics streams to S3 Curated zones for downstream analysis. Real-time parsing and standardization of badge access events could occur here.

Finally, to enable timely reaction where fits, the architecture shows leveraging Amazon SNS topics to publish alerts on attributes of interest in the streams like security policy violations for immediate notifications.

The components work together to intake, refine and reveal actionable signals from real-time system and event sources in a governed way - key to enabling real-time decision making.

ADVANCE DATA ANLYTICS



Machine Learning Use Cases

1. Predictive analytics for at-risk students:
- Build classification models leveraging random forest, XGBoost on SageMaker to predict students likely to fail or drop out based on academic, engagement, and psychosocial factors. Enables proactive student support.

2. Personalized course recommendations:
- Implement collaborative filtering algorithms like matrix factorization on student attributes, grades, career goals etc to provide tailored course suggestions maximizing progress or completion rates.

3. Curriculum and instruction analysis:
- Employ natural language processing techniques on open-ended course surveys to gather topic modelling-based insights and sentiment analysis on teaching methods, content quality etc to inform evidence-based curriculum redesign at scale.

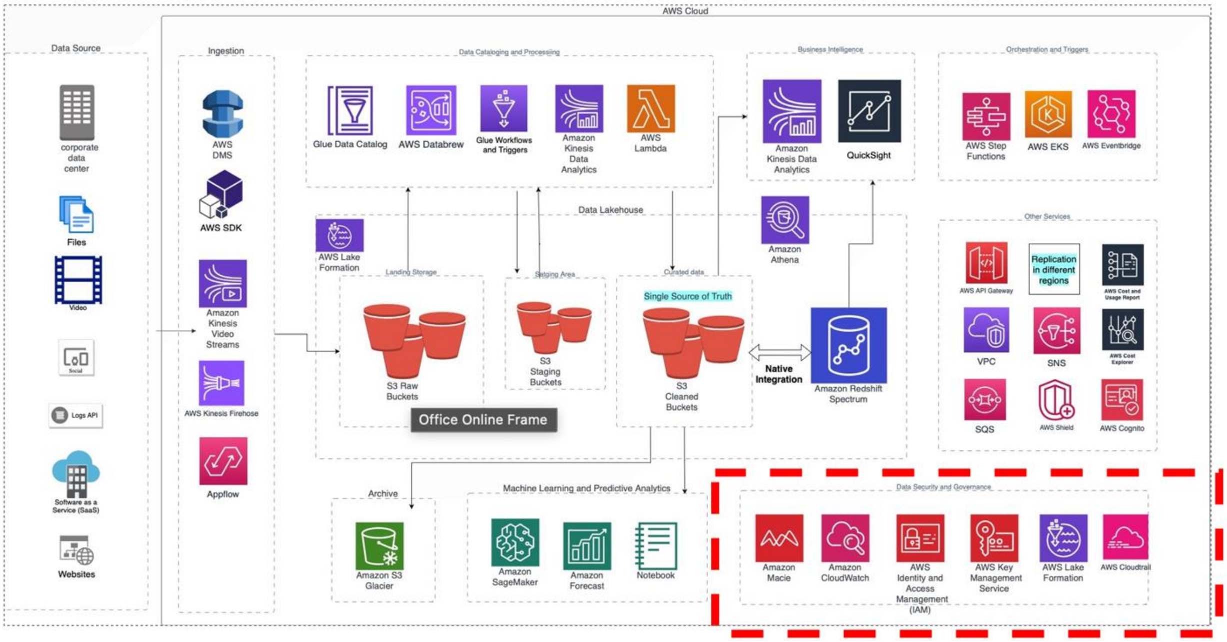
4. Automated document processing:
- Use optical character recognition and extraction techniques on application forms, test papers etc to structure unstructured student data, transcripts into searchable indices, minimizing manual effort.

Business Intelligence

5. Central data visualizations for tracking KPIs across student funnel from lead to alumnus leveraging QuickSight dashboards providing senior leadership instant visibility into enrolment trends, revenue forecasts etc.
6. Real-time dashboards for monitoring student and staff operational metrics using Amazon Kinesis Data Analytics SQL capabilities for streaming analysis as data arrives, informing load balancing of shared resources.
7. Interactive geospatial view generated in QuickSight analyzing application rates sliced by locational dimensions enabling strategically targeted marketing campaigns, optimal siting of new campuses

Data Governance and Security: -

A robust data governance framework underpins the reliability of analytics empowering strategic decisions for the university. Leveraging capabilities like Lake Formation, Macie and CloudTrail, we implement access controls to grant selective data access ensuring privacy compliance, scan storage for sensitive PII data exposure risks, enable encryption to secure data assets, maintain immutable logs of user activity and platform audits to ensure accountability and traceability of data pipeline executions - collectively fostering analytics in a compliant, controlled environment. Tight governance integration will provide the foundations to build trust and transparency on platform operations.

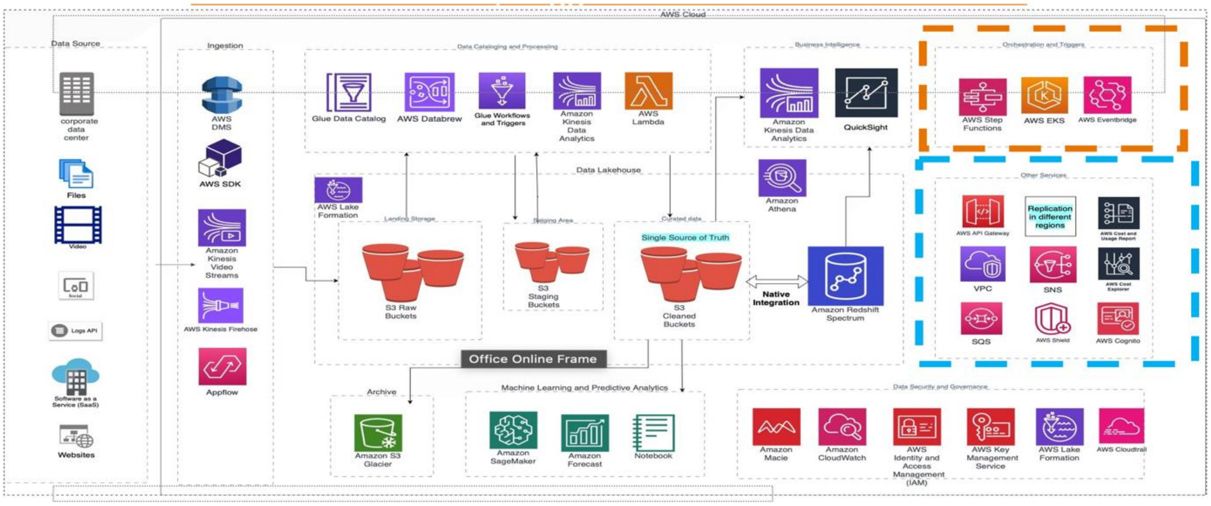


1. **AWS Macie**: Uses ML and pattern matching to discover and classify sensitive student PII/financial data stored in S3. Helps identify and protect private data while ensuring compliance with regulations.
2. **AWS CloudWatch**: Enables configuring alerts and alarms based on metrics thresholds crossed for usage, spend anomalies providing tight monitoring on infrastructure. Critical for cost management.
3. **IAM Policies & Roles**: Grants selective authentication and granular access controls for platform users and groups based on persona like data scientists, analysts, business users etc. Aligned to principle of least privilege.
4. **AWS KMS**: Encrypts data assets at rest and in transit across various services. Easy key management with ability to audit. Critical to enforce security best practices.
5. **AWS Lake Formation**: Centralizes access controls, auditing and metadata management over user access and actions on stored analytics data. Provides visibility into data lineage tracking.
6. **AWS CloudTrail**: Logs platform API calls and user activity supporting auditing requirements. Coupled with CloudWatch provides audit trail for security incident response.

This combination of preparatory data security, identity-based fine grained access controls, auditing trail for incident investigation and encryption meets complex governance needs for the analytical workloads while being cost-effective.

Orchestration and Triggers: -

A layered approach allows seamlessly orchestrating and responding to data events enabling real-time analytics. Step Functions coordinate batch ETL sequences and integration tasks providing visual workflows. EventBridge triggers instant reactions to data changes propagating updates while EKS Kubernetes clusters scale resources dynamically matching processing needs, unifying streaming and REST interfaces between systems.



1. AWS Step Functions: Orchestrates Lambda functions, Glue jobs into data workflows with sequence, branching, parallelization constructs. Simplifies building robust data pipelines.

2. Amazon EventBridge: Has event buses catching event streams like S3 data arrival, RDS table updates etc and routes to triggers activating data propagation, processing updates enabling real-time data pipelines.

3. Amazon EKS: Manages Kubernetes cluster for scalable spark jobs and microservices providing autoscaling of containers matching workload needs. Handles bursts maintaining SLAs. Integrates queueing and streaming data services.

The combination provides tools to reliably coordinate disparate systems, respond instantly to data changes and allocate just enough resources to workloads for high performance at optimized infrastructure costs.

Other Services: -

An integrated collection of services fortifies analytics operations, cost efficiency and continuity against disruptions. VPCs establish isolated secure analysis environments, SNS enables push communication of monitoring alerts while Cognito, SQS help coordinate user identity and task queues. Robust visibility into spends and usage is powered by Cost Explorer and Reports for ongoing optimizations. AWS Shield protects against DDoS attacks safeguarding availability.

1. Amazon VPC: Creates virtual private cloud to host analytics workloads securely isolating environments needed for sensitive student data analysis separate from public facing tools.

2. Amazon SNS: Facilitates push notifications via SMS, Email on service health events and alarms allowing teams to respond faster to incidents that could impact analysis.

3. AWS Cost Explorer: Interactive dashboard providing granular visibility into usage, forecasts and budgets helping identify opportunities to optimize cloud spend.

4. AWS CUR: Monthly report detailing AWS usage and expenditures at day level granularity feeding into chargeback, accounting and reconciliation needs.

5. Amazon Cognito: Manages user identities including authentication, account recovery and synchronizing user profile data across devices enabling tailored, unified experiences.

6. Amazon SQS: Manages messaging queue workload buffering requests, batching write operations helping smooth traffic spikes improving reliability for upstream APIs.

7. AWS Shield: Protects platform resources against Volumetric DDoS attacks ensuring system availability critical for round the clock analytical needs.

1 YEAR MILESTONES:

Quarter 1:

Implement core data ingestion pipelines for 2-3 critical on-premises sources into S3/Glue catalog. Performing quick PoC analysis and viz in QuickSight on a small sample dataset.

Quarter 2:

Expand the raw data pipelines by integrating 5 high-value additional sources, including streaming and SaaS data, enhancing foundational data processing ETL of core domains, address gaps in data integrity and analyze hardened datasets reported by analysts.

Quarter 3:

Launch centralized Lake Formation governance of hardened zones including access controls and column level security enabled. Furthermore, formalize processes for data request triage, privileged access audits, and enforcement of standards for extended community.

Quarter 4:

For the last quarter of the year, the milestones would be to support launch the ML application into production by grading process owners with measured impact on student complaints. Moreover, it will be important to continue model optimization. Finally, begin shift towards real-time and predictive use cases, followed by pulling in streaming analytics for student attendance monitoring.

CONSIDERATIONS:

While moving to a fully cloud-based AWS model for data processing offers many benefits, there are also several risks that should be considered:

1) Data security: Storing sensitive data in the cloud can pose security risks. Companies must ensure they have robust security measures in place to protect against data breaches and other security threats.
2) Dependence on third-party vendors: A fully cloud-based model means that companies will rely heavily on AWS services and their performance. Any outage or service disruption on AWS's part can impact business operations.
3) Latency and network issues: Cloud-based processing can introduce latency and network related issues, affecting data processing and response times. This can be especially problematic for real-time processing and applications with strict latency requirements.
4) Vendor lock-in: Migrating data and applications from one cloud provider to another can be complex and costly. Therefore, companies must consider the risk of vendor lock-in when moving to a fully cloud-based model.
5) Cost management: While AWS offers cost-efficient pricing models, cloud-based data processing can still result in unexpected expenses if not managed effectively. Companies must have robust cost management strategies to avoid overpaying for cloud services.

CONCLUSION

Implementing a cloud-based data lake architecture on AWS will provide our University with an integrated, scalable, and high performance platform to drive advanced analytics. By consolidating data siloes, establishing a single source of truth, and enabling new methods like machine learning, we can greatly enhance the insights available to better serve our students. With embedded security schemes, we will also maintain and possibly improve governance standards over sensitive data. And using Lakehouse design and autoscaling resources, we can reduce continued infrastructure expenses while future-proofing our capacity for growth. On the people side, the rollout of a transformative program like this does require planning for organizational change. We will need to ensure current teams are brought along on the journey, provide role-based training, and embrace new ways of developing and sharing analytics to see success.

With the right vision and commitment, this modernization initiative presents a massive value creation opportunity for the University. It can keep us competitive in attracting top students around the globe while improving learning outcomes through data-driven services.