DATA LAKEHOUSE ON AWS

FINAL PROJECT
BUAN6335.501.23F

Kamna Kumari (kxk220100)
Rahulchandra Marampudi (rxm220075)
Harshavardhan Kumar Jetty (HKJ220001)
Christena Darsi (cxd230013)

Agenda

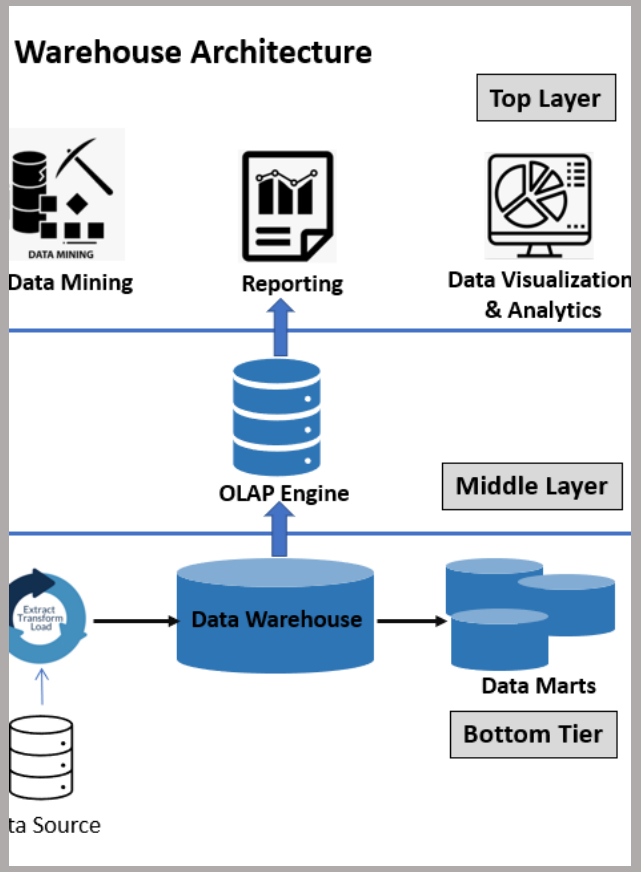
- Objective
- Traditional Architecture
- Data Lakehouse Architecture
- Data Governance
- Sharding Strategies
- Scaling
- Data Caching
- Use Cases
- Milestone
- Conclusion
- References

Objective

Propose a data Lakehouse architecture on AWS to resolve data management problem of a university

Current Architecture

- In-house Data Warehouse system: Archaic infrastructure
- Enormous Data Volume:
  Current data is approximately 255 TB
- Huge Cost
- Lack of single source of truth
- Huge in-house data warehouse
- Absence of reporting tools
- Hectic Data analysis



AWS Strengths Compared to GCP & Azure

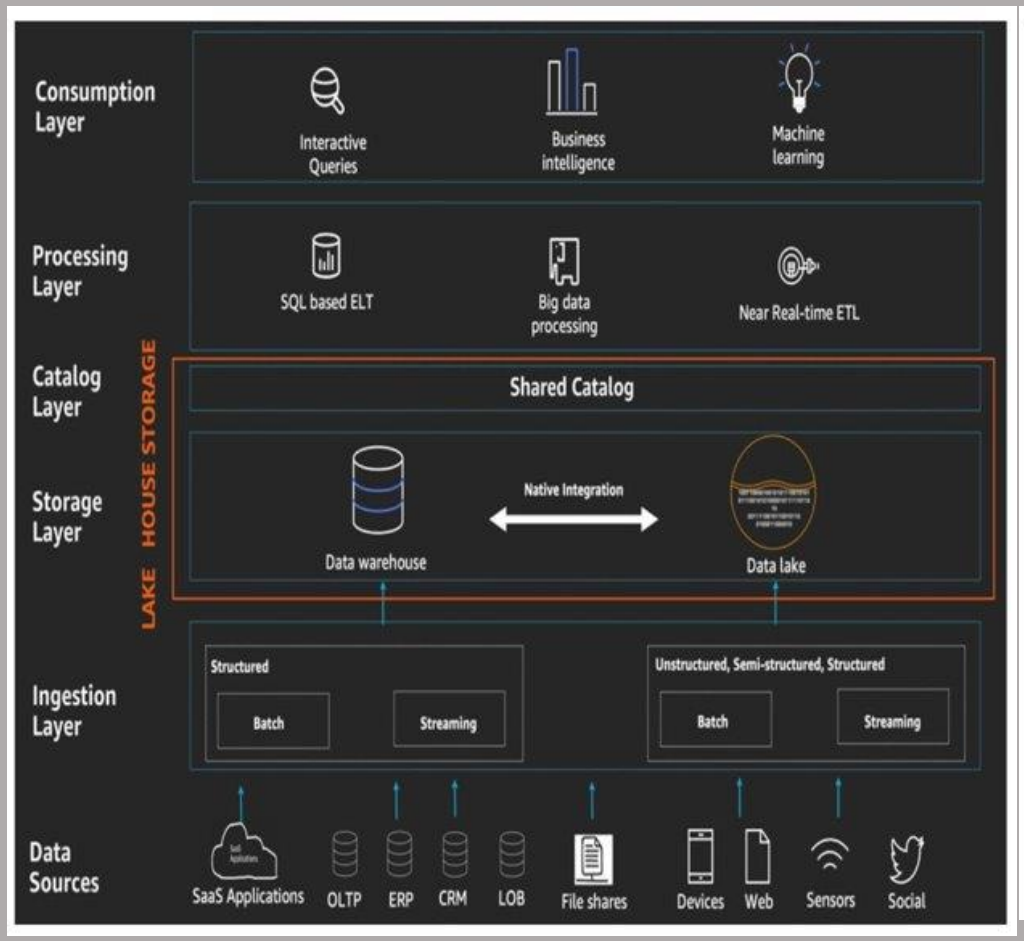
Market Dominance: AWS leads in market share, boasting maturity and a robust set of features compared to GCP and Azure.

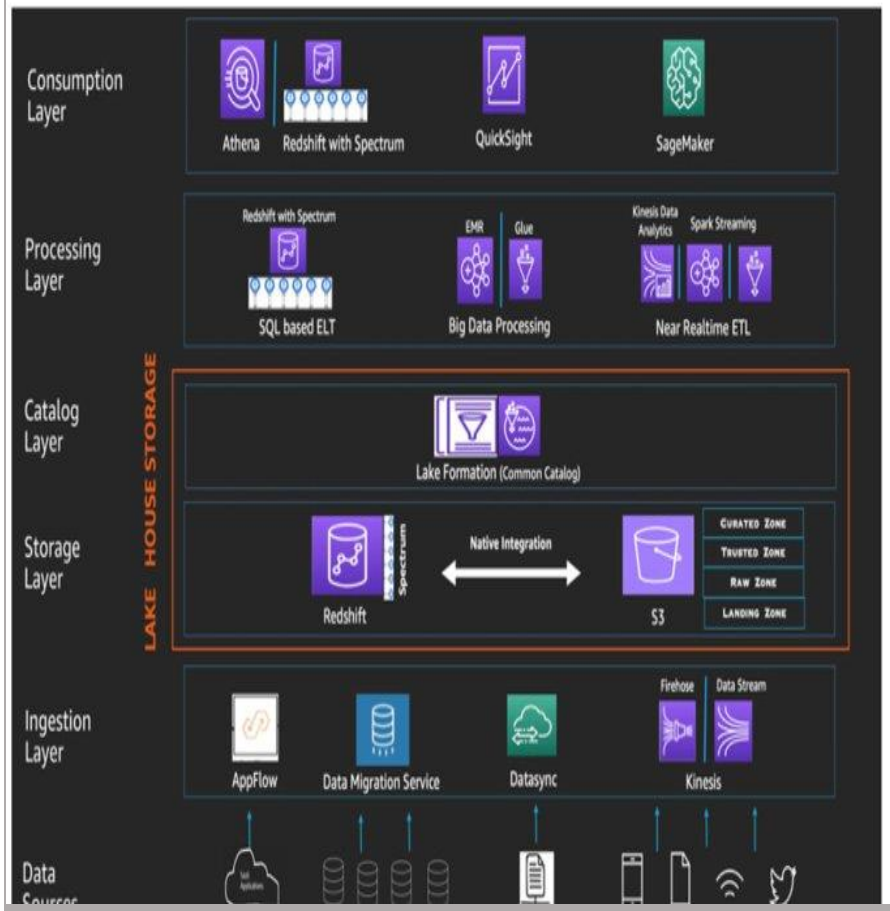
Geo-redundancy: AWS excels in automatic geo-redundant storage, providing a reliable solution, whereas Azure has caveats, and GCP has limitations.

Ease of Integration: AWS stands out with a wide range of APIs and connectors, while Azure and GCP showcase robust integration capabilities with Microsoft and Google products.

Compliance & Security: AWS has a wider range of compliance certifications compared to Azure and GCP.

Data Lakehouse





Data Sources

Structured Data:

Ex: Student information system, HRM system, Finance & Accounting, E-Learning, ERP, CRM, LMS systems

Semi structured Data:

Ex: Course syllabus, survey responses, event calendars, email communication

Unstructured Data:

Ex: Research data & Publications, Assignments, Video lectures, Social media content, web content

Unified Data Ingestion

The data ingestion layer within the showcased serverless architecture is crafted with purpose-built AWS services, facilitating seamless data ingestion from diverse sources.

Use Case:

- Streaming student enrolment data
- Batch loading historical data
- Transferring large files

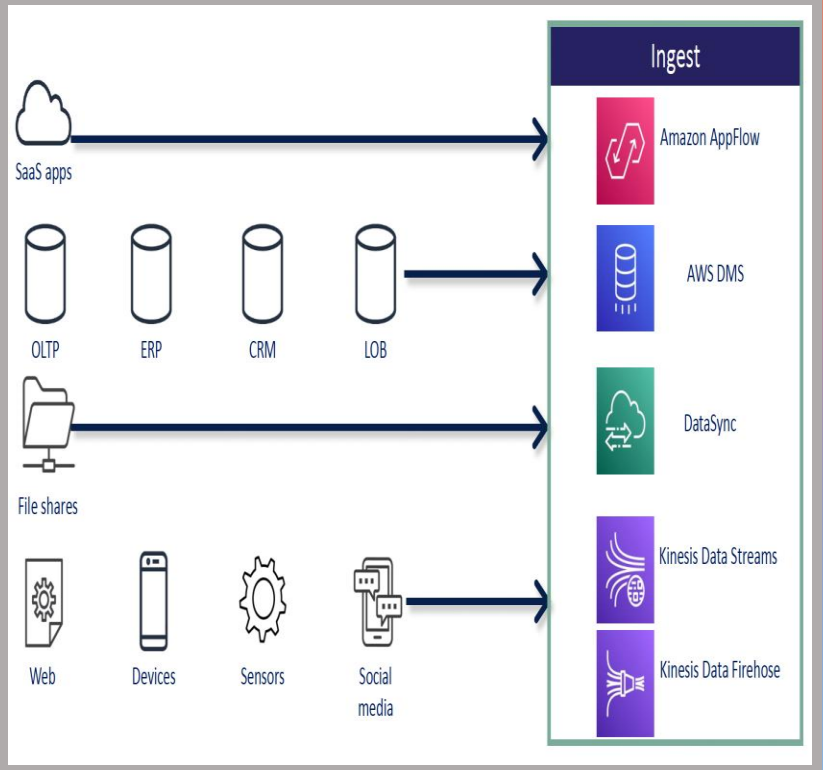
Benefits:

- **Precision:**

Ensure the accuracy and reliability of all the information you work with within the Data Lakehouse.

- **Versatility:**

After data ingestion, enjoy enhanced accessibility, manipulation, and analysis capabilities, surpassing the utility of raw data forms.



Lakehouse Storage

Key Components:

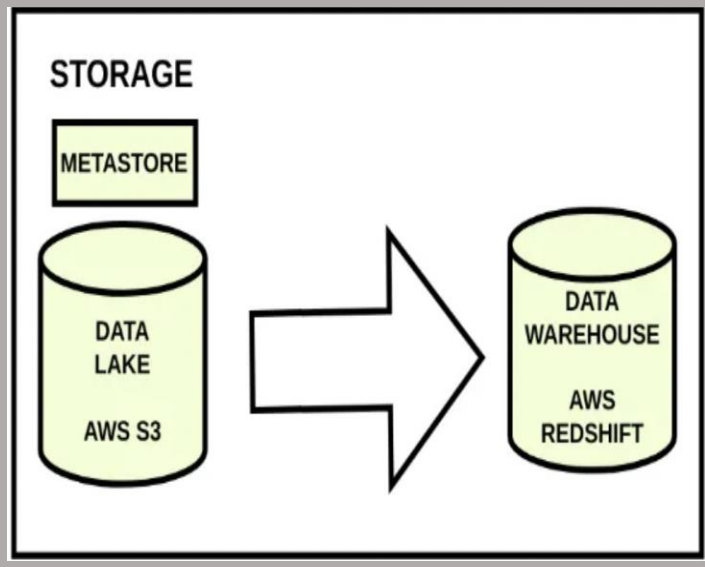
- Amazon Redshift: Stores reliable, consistent, and highly managed structured data in standard dimensional schemas.
- Amazon S3: Provides exabyte-scale data lake storage for structured, semi-structured, and unstructured data.

Use cases:

- Store & analyse student data from various sources
- Gain insights into student performance & engagement
- Improve administrative efficiency

Benefits:

- Scalability and Performance: Amazon S3 offers industry-leading scalability, data availability, security, and performance for open file formats.
- Unified SQL Interface: Redshift Spectrum empowers Amazon Redshift to process SQL statements referencing data in both data lake and warehouse



Understanding Data Catalog

Key Components:

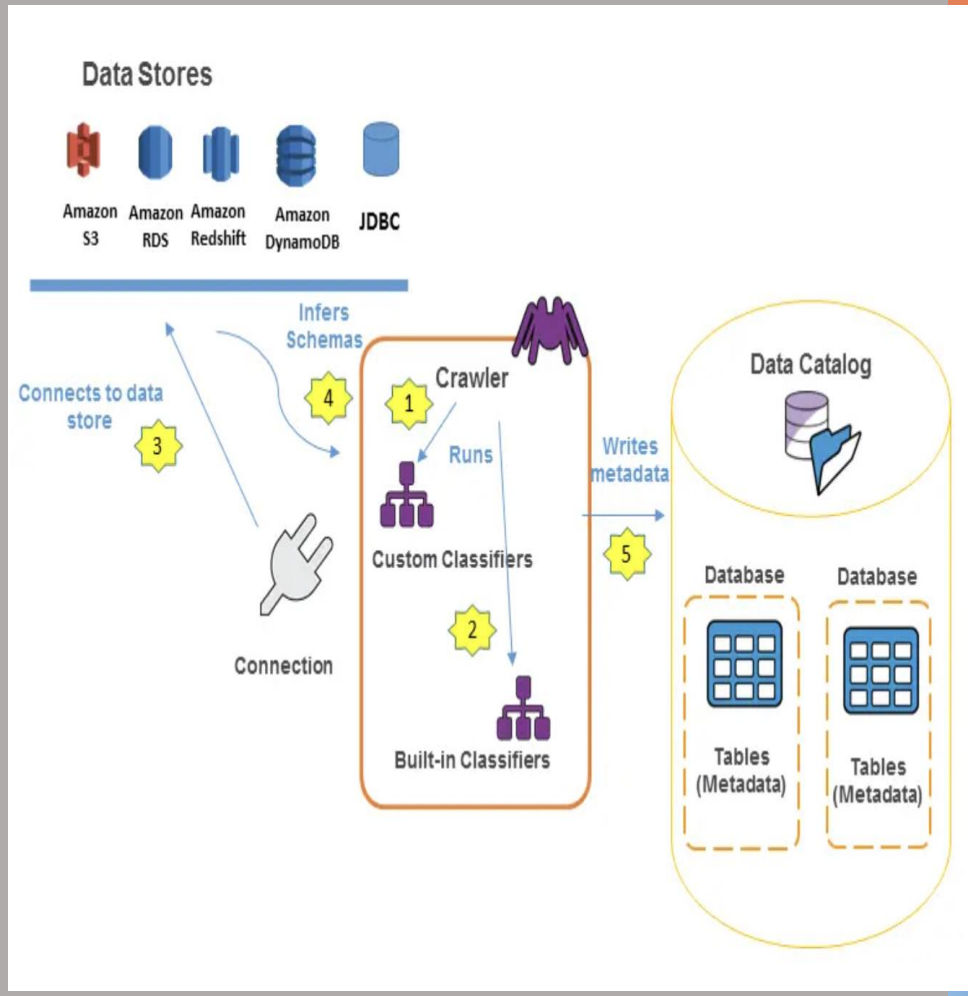
- Metadata Storage
- Data Classification
- Data Search and Discovery
- Data Governance Rules

Use Cases:

- Organize student records
- Enhance research collaboration
- Improve operational efficiency
- Support compliance with regulations

Benefits:

- Improved Data Visibility
- Enhanced Data Governance
- Simplified Data Access



Data Processing Layer

Key Components:

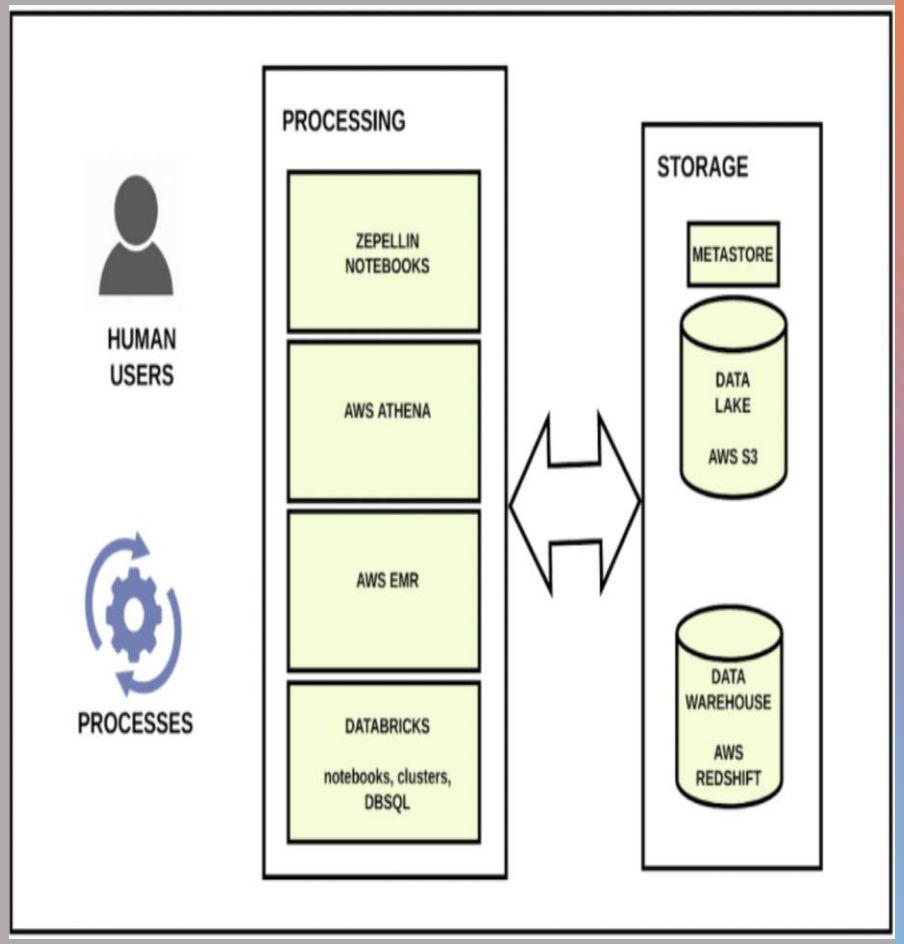
- Data Validation
- Data Cleanup
- Normalization
- Transformation
- Enrichment

Use Cases:

- Real-time ETL for Student Engagement
- Detect unusual spikes in enrollment numbers
- identifying and addressing missing or inconsistent data(student records, enrollment data & financial transactions)

Benefits:

- Eliminates Data Redundancies
- Scalability
- Data Flexibility
- Unified Data Access



Data Consumption Layer

Key Components:

- SQL Query Engines
- Business Intelligence (BI) Dashboards
- Machine Learning (ML) Integration
- Data Lake Data Access

Use Cases:

- query student enrollment data to identify enrollment trends
- Visualize data and track progress towards university goals
- Identify at-risk students

Benefits:

- Cost-Efficient Analysis
- Rapid Insights
- Support for Multiple User Personas

Data Governance

Key Components:

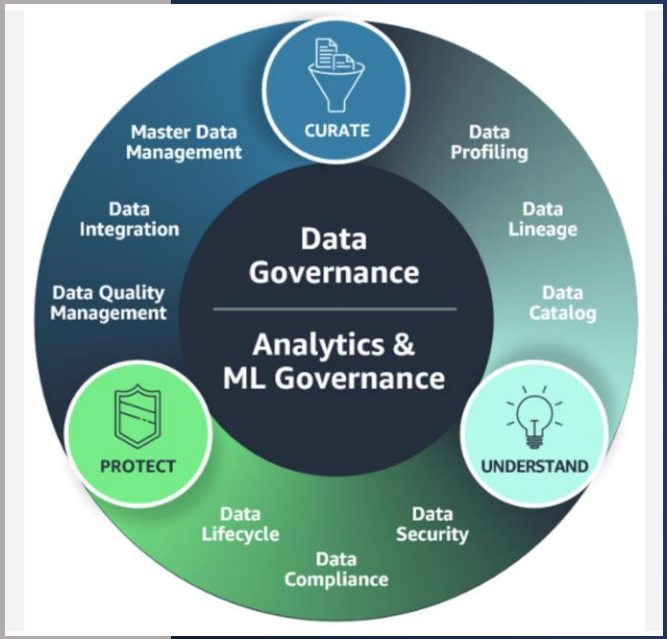
- Data Governance Team
- Data Quality Management
- Data Access Control
- Data Compliance

Use Case :

- Security and Audit
- Data Lineage and Transparency
- Compliance Management
- Identify and address data quality issues

Benefits :

- Data Trustworthiness
- Improved Decision-Making
- Data Transparency & Security
- Reduced Data Risks



Sharding/Partitioning Strategies

Key Components:

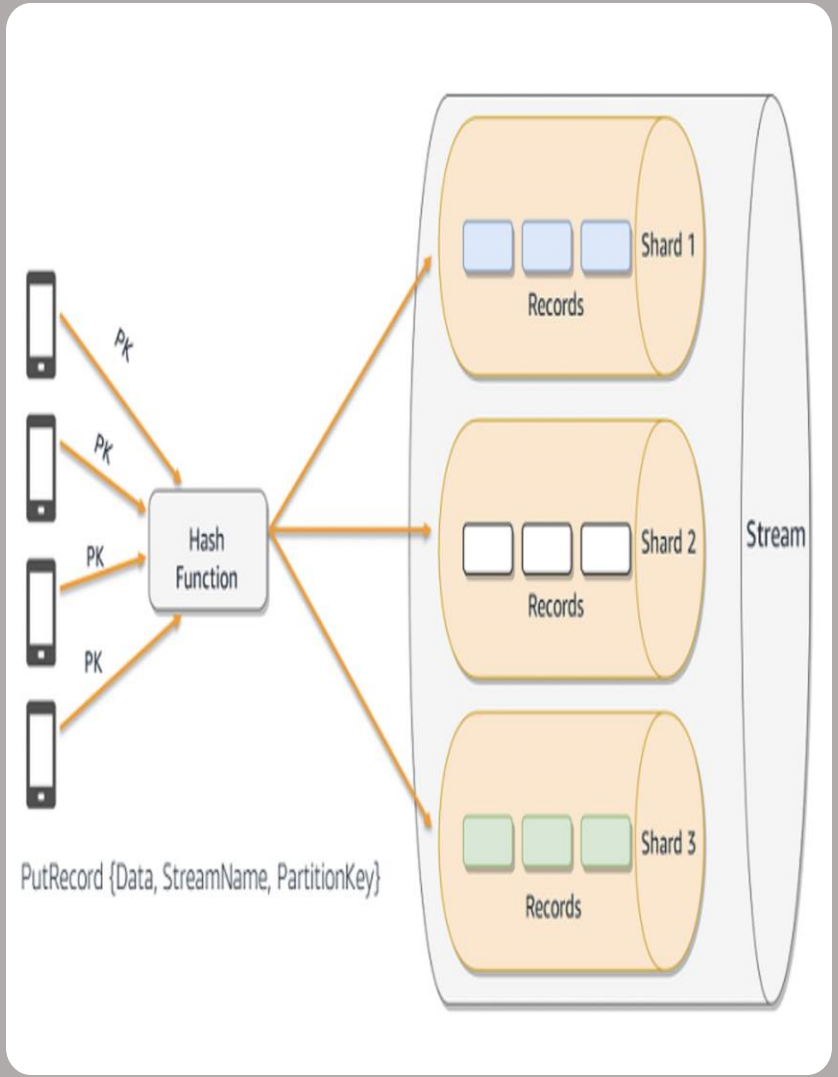
- Data Partitioning by Date
- Geographical Partitioning
- Content-Based Partitioning

Use Cases:

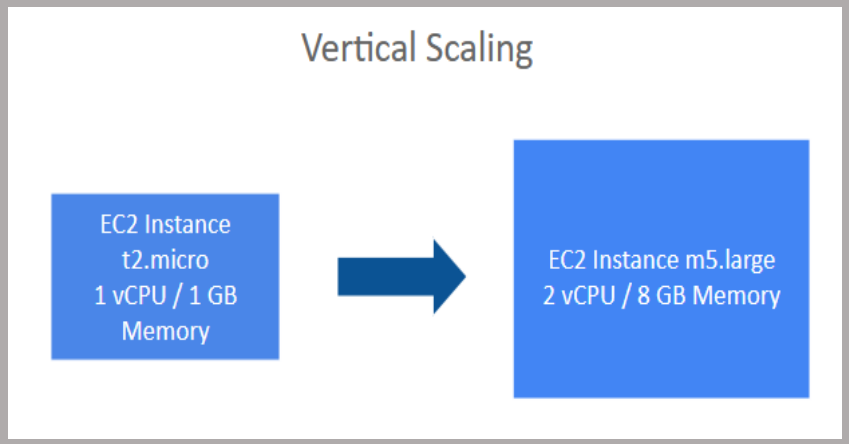
- Efficiently manage university data by segmenting data
- Quick data retrieval and efficient data analysis.

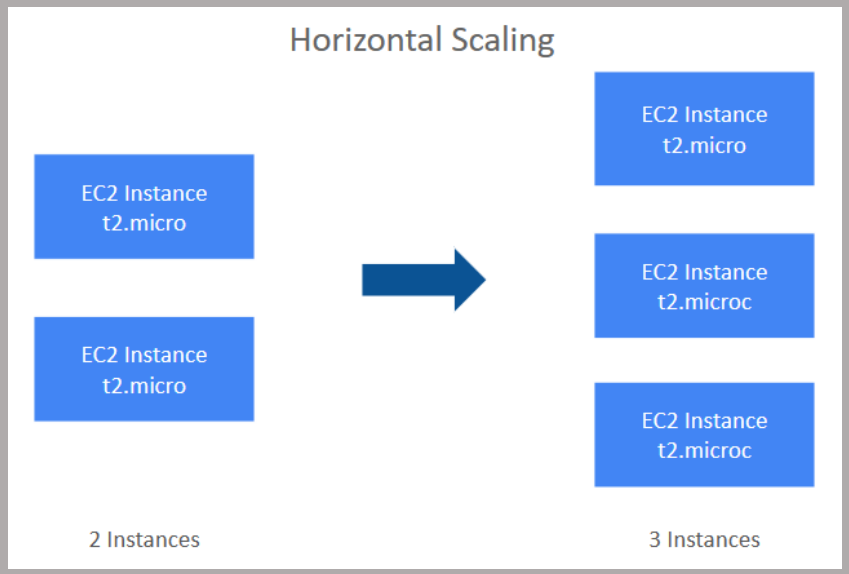
Benefits of Sharding/Partitioning:

- Improved Query Performance.
- Efficient Data Management.
- Better Scalability and Cost Management.



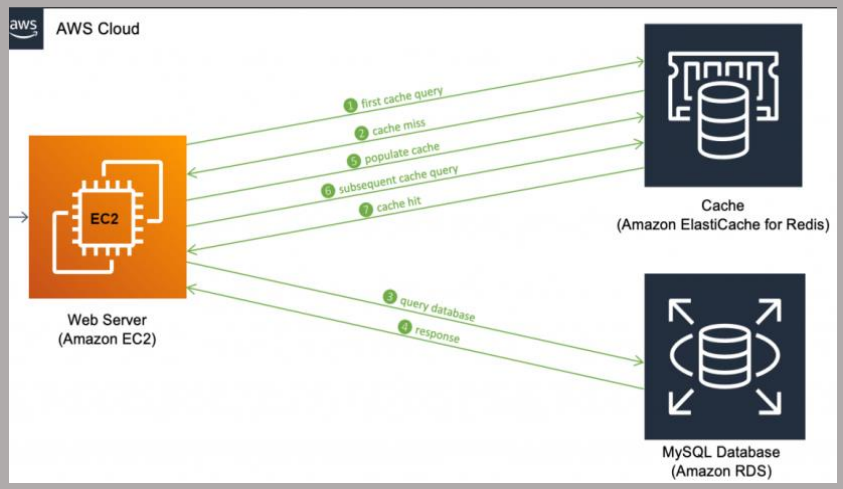
Vertical vs. Horizontal





- Horizontal scaling is preferrable for the current project

Data caching



Features of Elasti-cache service:

- Performance Gains
- Scalability
- Cost reduction
- Compliance and security
- Data durability
- Fully Managed

Use Cases

- Data Analytics and Reporting
- Personalized learning & student report
- Curriculum Effectiveness
- Security and Compliance
- 360-Degree Student and Staff View
- Financial Management
- Resource Allocation and Planning
- Sales and Marketing Insights
- Alumni Engagement and Fundraising
- Research and Grants Management
- Staff Performance and Development
- Mobile and Online Learning Insights

Milestones - Sample Project Plan (Three Year)

Assess and Prepare

Year 1: Assess and Prepare (Migration Phase 1)

Quarter 1: Project Initiation and Objective Definition

Define Migration Objectives

Assemble Migration Team

Initial Assessment

Quarter 2: Current State Assessment

Comprehensive Assessment

AWS Service Selection

Quarter 3: AWS Environment Setup and Security Configuration

AWS Environment Setup

Data Lake Design Initiation

Quarter 4: Data Lake Design and Planning

Data Lake Finalization

Data Migration Strategy

Build and Optimize

Year 2: Build and Optimize (Migration Phase 2)

Quarter 1: Data Ingestion and ETL Development

Data Ingestion

ETL Development

Quarter 2: Real-time Data Streams and Data Warehouse Deployment

Real-time Data Streams

Data Warehouse Deployment

Quarter 3: Predictive Analytics and Advanced Features

Predictive Analytics

Optimization

Quarter 4: Data Governance and Security Implementation

Data Security

Governance Initiation

Operate and Scale

Year 3: Operate and Scale (Migration Phase 3)

Quarter 1: Data Governance Expansion and Data Migration

Data Governance Enhancement

Data Migration Commencement

Quarter 2: Testing, Optimization, and Backup Strategies

Comprehensive Testing

Optimization and Protection

Quarter 3: User Training and Documentation

User Training

Documentation

Quarter 4: Ongoing Operations and Scaling

Ongoing Operations

Scalability Assessment

Documentation Updates

Conclusion

- Architectural Transformation
- Alignment with technical goals & objectives
- Data-Driven Excellence

References

- https://www.talend.com/resources/cloud-data-warehouse-architecture/
- https://vticloud.io/en/build-data-lakehouse-on-aws-part-2/
- https://atlan.com/aws-glue-data-catalog-explained/
- https://medium.com/@gu.martinm/how-to-build-your-own-data-platform-episode-2-authorization-layer-data-warehouse-implementation-ab1cbca04dfd
- https://www.analytics8.com/blog/why-it-is-time-to-consider-a-data-lakehouse-as-part-of-your-modern-data-architecture/
- https://mattturck.com/data2021/
- https://www.enablegeek.com/blog/aws-vertical-scaling-vs-horizontal-scaling/
- https://operisoft.com/aws-cloud/amazon-elasticach/
- Class Notes

Thank you