Data Lake

Cristian Bassotto

Introduction

What is a Data Lake?

A data lake is a centralized repository designed to store, process, and secure large amounts of data of various type. It is possible to store the data as it is, without having to first structure the data, and to run different types of analytics (from dashboards and visualizations to big data processing, real-time analytics, and machine learning).

Such as databases and spreadsheets, a data lake can store structure data, but also semi-structure data (like CSV files and XML documents) or unstructured data such as emails, social media and multimedia files. This simplifies the process for organizations to store various types of data without requiring predefined schemas or models, making it easier to adapt and scale as needed.

In a data lake, the structure of the data or schema is not defined when data is captured. Conversely, a data warehouse is a database optimized to analyze relational data, with its schema defined in advance to optimize fast SQL queries. However, data warehouses are limited in their ability to handle a wide variety of data types.

Data Lake for online banking

I am planning to launch an online banking company that will need to store various types of data, including customer information, transaction data, and authentication records. My aim is to offer customers both a mobile application and an online web application for easy account creation, access to banking services, and transaction management.

To handle the diverse data generated by my company, such as documents, images, videos, and financial data, I intend to set up a data lake. This data lake will serve as a centralized storage solution, enabling efficient management and analysis of my data assets.

To ensure compliance with regulations and protect sensitive data, I'll prioritize data governance and security. This involves implementing access controls, encryption, data masking, and monitoring mechanisms to safeguard data integrity and privacy.

Methods

Online banking data

The data collected has a diverse structure to accommodate various types of informations essential for online banking operations:

1. **Customer Information**:
   - Structured data as personal details such as name, address, contact information, date of birth.
   - Unstructured data for documents like images of identification documents and recordings of agreement confirmations.
2. **Account Data**:
   - Structured data as account numbers, types of accounts and balances.
   - Unstructured data for account's contract, agreement and privacy details.
3. **Transaction Data**:
   - Structured data for transaction details including amounts, dates, times, types, and parties involved.
4. **Authentication and Authorization Data**:
   - Structured data for login records, authentication methods used, and access permissions.
   - Unstructured data to store failed logins throw face recognition or fingerprint.
5. **Risk and Compliance Data**:
   - Unstructured data for regulatory compliance information imposed by financial authorities and fraud detection indicators or signals.
6. **Customer Interactions**:
   - Unstructured data formats for chat logs, call recordings, emails, and feedback forms to better conduct AI approaches to search for improvements.
7. **Market Data**:
   - External data sources that could be structured or unstructured and contains feeds for market trends, interest rates, currency exchange rates, etc.
8. **Internal Logs**:

- Structured logging formats for internal IT infrastructure components like servers, applications and network devices.

9. **Complaints and Dispute Resolution**:
    - Unstructured data fields for recording customer complaints, disputes and resolutions to recognise satisfaction also by voice detection or face recognition.

Users and permissions

Plenty of professional figures will need to have access to some specific part of the data lake in order to perform different types of analysis and measures to higher the performances of the company. Additionally, it is necessary to manage permissions following some strict security measures to protect sensitive data. In particolar, access to sensitive data should be restricted based on roles and responsibilities. Users should only have access to the data required for their specific job tasks. Role-based Access Control (RBAC) mechanisms are crucial in this regard. Robust controls need to be established to ensure that only authorized users can access specific datasets. This involves setting permissions at the dataset level, ensuring that individuals or groups are granted access only to the data they require. Also it is better to require Attribute-Based Access Control (ABAC). This method is a dynamic access control model that grants or denies access to resources based on attributes associated with users, resources, and the environment. It enables fine-grained access control by considering various attributes such as user roles, resource classifications, and environmental factors. Lastly, I will include also an Access Control List (ACL) e.g. a list of permissions attached to specific resources that defines which users or system processes are granted access to those resources and what operations they are allowed to perform. These three methods are combined into a single permission evaluation algorithm, as described in Figure 1.

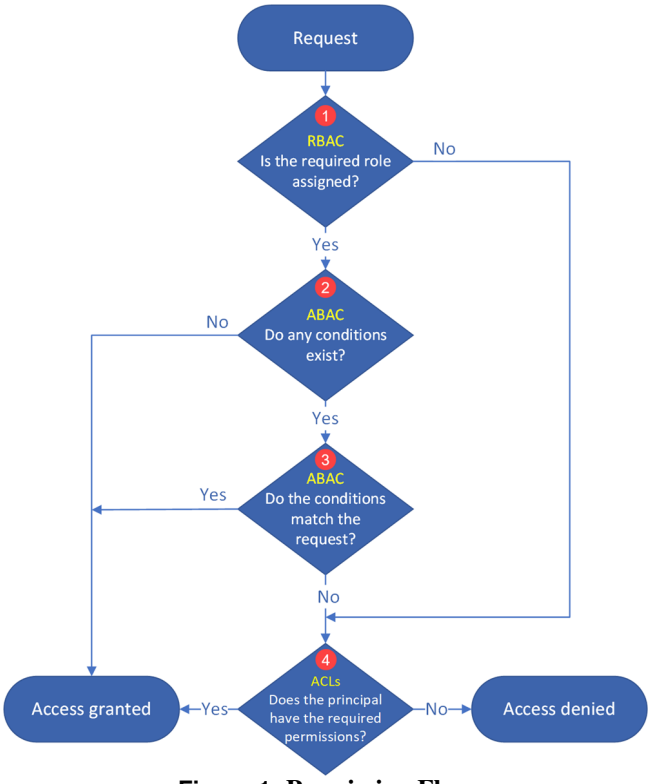

Figure 1. Permission Flow

- **Data Scientists**: For advanced analytics, machine learning and predictive modeling. They would require access to a wide range of data types, including structured customer information for segmentation analysis and unstructured customer interaction data for sentiment analysis. They would also need access to transaction data for pattern recognition and risk assessment.

- **Data Analysts**: For generating reports, performing ad-hoc analysis and extracting insights from the data. Analysts would primarily work with structured data such as transaction data and account information to generate reports and extract insights for business decision-making. They may also require access to authentication and authorization data to analyze login patterns and identify potential security breaches.

- **Business Analysts**: For conducting market research and performance analysis. They would focus on market data and customer interactions to identify trends, assess customer satisfaction, and evaluate the effectiveness of marketing campaigns. They may also need access to complaints and dispute resolution data to understand customer feedback and improve service quality.

- **Engineers Team**: For monitoring the system, managing data infrastructure and ensuring data security. They require access to most of the data to monitor system activities, troubleshoot issues, and ensure the overall security and integrity of the data infrastructure.

Sharing data

I'm going to share the data both internally within the organization and externally with regulatory authorities, auditors, and select third-party vendors as necessary for compliance and business operations.

Internally, authorized users will access data to facilitate collaboration and support various business functions. *Externally*, data sharing will occur with regulatory bodies, auditors, and vendors to ensure compliance with regulations and to enhance operational efficiency. I will establish data sharing agreements to govern the terms and conditions of data sharing, ensuring transparency, security, and compliance throughout the process.

Data Quality, Governance, and Security

To improve data governance and quality I will provide a dedicated team of engineers that will be responsible for building

and maintaining a data catalog along with the specific data analysts that will use that type of data, to provide a common practice to store and maintain data inside the data lake. The data catalog is an organized, comprehensive store of table metadata. It's important for this team to properly catalog new data as it enters in the data lake, and continually curate it to ensure that it remains updated. ACID properties (atomicity, consistency, isolation and durability) are also important data features that my team will need to take care about. To do so, they will have admin access to the data lake in order to solve the main associated problems.

To solve the security issues of data lakes, I will also need to encrypt data at rest and in transit to prevent unauthorized access and mitigate the risk of data breaches. Additionally, employing data masking and anonymization will helps protect sensitive information by obfuscating identifiable details, thereby enhancing privacy and confidentiality.

Success criteria

Success criterias are needed to evaluate my data lake and include measuring the quality of the data and the performances. In particolar:

1. **Data Accessibility and Availability**: Success can be measured by the easy access to data, essential for online banking operations. High availability and timely access to data indicate that the data lake infrastructure is functioning effectively.

2. **Data Quality and Consistency:** Data quality is also an essential criteria (particularly important for sensitive financial data) and can be measured using accuracy, completeness, and consistency. This criteria is the most important one when refering to an online banking, since inconsistent data or low quality can affect the future of the all company.

3. **Performance and Scalability:** performance and scalability of the data lake infrastructure is another important criteria, critical for handling diverse data types and growing data volumes. Some metrics could be query response times and data processing speeds.

4. **Usage analysis:** metrics related to the usage of the data lake by data scientists, analysts, and IT operations. Tracking the number of users accessing the data lake and frequency of data queries provides insights into the platform's effectiveness in supporting various business functions.

5. **Business Impact:** Ultimately, we should measure its impact on the company's business objectives, such as revenue growth, cost reduction, operational efficiency, and customer satisfaction. Although it's not easy to actually measure this impact, it's an important criteria to see how this improvement help the economy of the company.

6. **Return on Investment (ROI):** measures of ROI to evaluate the cost-effectiveness of the data lake investment. Quantifying the benefits derived from improved efficience against the costs associated with implementing and maintaining the data lake infrastructure.

The use of a data lake will improve the revenue of my company and the actual impact will be measures throw the success criteria. My company will benefit from personalized marketing campains, by analysing customer data, that better fit the needs of each user.

Analysing the trends of service selling and customers complains, my company can focuse more in specific improvements in specific services to higher customers satisfaction. Also risk management and fraud detection will increase the security in the company. Optimizing the efficiency of the operation will reduce non-profit times and leaad to cost savings and resource optimization.

Online providers

Amazon Web Services (AWS), Microsoft Azure, and Google Cloud are leading cloud service providers:

Amazon Web Services (AWS)

Amazon Web Services is the most standard storage service for object archiving, and it holds the biggest chunk of the current cloud market. It consists of a rich pair of tools like IoT, security, databases, management, analytics, and enterprise applications. On the other hand, AWS Service S3 offers reliability, scalability, accessibility of information, and efficiency. It provides a global network of data centers, allowing businesses to scale and deploy applications globally with ease.

Google Cloud Platform (GCP)

Google Cloud Platform is Google's suite of cloud computing services, offering infrastructure as a service (IaaS), platform as a service (PaaS), and serverless computing environments. GCP provides services such as Google BigQuery for analytics, Google Cloud Storage for scalable storage, and Google Cloud Dataflow for real-time data processing. It emphasizes machine learning and AI capabilities, leveraging Google's expertise in these areas.

Azure

Azure is Microsoft's cloud computing platform offering a wide range of services including computing, analytics, storage, and networking. Azure provides services such as Azure Blob Storage for object storage, Azure Data Lake Storage for big data analytics, and Azure SQL Database for relational database management. It integrates well with Microsoft's suite of products and offers hybrid cloud solutions for businesses with on-premises infrastructure. Data Lake Storage and Queue Storage are some of the best Azure options for big companies that need high data storage requirements. Bulk storage is ideal for companies opting for a large amount of

unstructured data, while File storage is reliable and designed for most business requirements.

Following a list of services for specific use cases offered by each provider.

Data management

Each of the leading cloud service providers offers a suite of tools and services for effective data management.

Storage

For storing any amount of data and retrieving it as often as needed, AWS, Google Cloud, and Azure each provide robust solutions:

- AWS Simple Storage Service (S3) : It offers various storage classes to optimize costs based on data access patterns and durability requirements.
- Cloud Storage : Its multi-regional and regional storage options allow users to choose the location that best suits their latency and compliance requirements.
- Azure Blob Storage: It provides features such as hot, cool, and archive tiers to optimize storage costs based on data access patterns.

Management Tools

In terms of management tools for monitoring, controlling, and optimizing cloud costs, each provider offers solutions tailored to their platform:

- AWS Cost Explorer and AWS Budgets: AWS provides Cost Explorer for analyzing AWS usage and costs, along with Budgets for setting custom cost and usage budgets to track spending.
- Cost Management: a suite of tools for cost management, including Cloud Billing reports, Budgets, and Quotas.
- Azure Cost Management: It provides insights into resource utilization, cost trends, and recommendations for cost optimization.

Additionally, each provider offers tools for configuring restrictions on resource usage:

- AWS Organizations policies: leverage Service Control Policies (SCPs) to restrict access to specific AWS services or resources at the organizational level. SCPs act as guardrails, allowing organizations to control what actions and services can be accessed by member accounts.
- Organization Policy Service: define constraints on resource configurations using policy constraints. These constraints can be applied at the project, folder, or organization level, ensuring compliance with organizational policies.
- Azure Policy: Policies can be scoped to specific resource types, locations, or tags, allowing for fine-grained control over compliance requirements.

Analytics

Analytics is another important field to take into consideration.

Data discovery and metadata management

Discover, understand, and manage data at scale with powerful search and seamless integration to the storage, secured using IAM.

- AWS Glue Data Catalog : allows users to discover, understand, and manage data at scale.
- Dataplex : offers advanced data discovery and metadata management capabilities.
- Azure Purview and Azure Data Explorer: provides data governance and metadata management capabilities, allowing users to discover, catalog, and govern data assets across the organization. Azure Data Explorer offers real-time data analytics and exploration capabilities for large-scale datasets.

Data processing

Deploy open-source data and analytics processing services (Apache Hadoop, Apache Spark, etc.) with improved efficiency and security.

- Amazon Elastic MapReduce (EMR), AWS Batch and AWS Glue : AWS offers a suite of data processing services including EMR for processing large datasets using Apache Hadoop and Apache Spark, AWS Batch for batch processing workloads, and AWS Glue for ETL (Extract, Transform, Load) tasks.
- Dataproc : managed Apache Spark and Apache Hadoop clusters for data processing and analytics tasks.
- Azure Data Lake Analytics and HDInsight: Azure offers services like Data Lake Analytics for big data processing and analytics, and HDInsight for managed Apache Hadoop, Spark, and other big data frameworks.

Core compute

To train and run machine learning models faster (GPUs and TPUs time computation).

- Amazon Elastic Compute Cloud (EC2) P3 and AWS UltraClusters: AWS EC2 offers a wide range of compute instances including P3 instances optimized for machine learning workloads, and UltraClusters for high-performance computing tasks.
- Cloud GPUs and Cloud TPU : for accelerating machine learning and AI workloads with high performance and scalability.
- GPU Optimized VMs and Azure Virtual Machines : Azure offers GPU Optimized VMs and Virtual Machines for running compute-intensive workloads such as deep learning, rendering, and simulation.

Document understanding

Automate data capture at scale to reduce document processing costs.

- Amazon Textract : is a machine learning service that automatically extracts text and data from scanned documents.
- Document AI : offers advanced document understanding capabilities, including document parsing, entity recognition, and content classification.
- Azure Form Recognizer : is a cognitive service that extracts information from forms and documents.

Image recognition

Derive insights from images in the cloud or at the edge, or use pre-trained Vision API models to detect emotion, understand text, and more.

- Amazon Rekognition Image : deep learning-based image recognition service that enables users to analyze images for objects, scenes, and facial analysis.
- Vision AI : provides pre-trained models for image recognition tasks such as object detection, facial recognition, and image labeling.
- Azure Computer Vision : cognitive service that analyzes images and extracts information such as objects, text, and facial attributes, allowing organizations to derive insights from visual data.

Security

Identity and Access Management (IAM)

Services that provide fine-grained access control and visibility for centrally managing resources:

- Amazon Identity and Access Management : allows to create and manage user identities and permissions.
- Identity and Access Management : offers similar capabilities for managing access to cloud resources, enabling organizations to control who can access specific resources and what actions they can perform.
- Azure Identity Management : offers the same capabilities of AWS

Encryption

All the three online providers supports server-side encryption, and have services that allow to manage encryption keys securely:

- AWS Key Management Service (KMS): offers secure and scalable key management services, compliant with FIPS 140-2 Level 2 standards. It allows to create and control encryption keys used to encrypt data stored in AWS services and applications.
- Cloud KMS and Cloud HSM: Host encryption keys and perform cryptographic operations in a cluster of FIPS 140-2 Level 3 certified hardware security modules (HSMs).
- Azure Key Vault: similar to AWS KMS.

Network Security

- **AWS Virtual Private Cloud (VPC)** : for network isolation, security groups for instance-level firewall rules, and AWS Web Application Firewall for web application firewall protection.
- **Google Cloud Virtual Private Cloud (VPC)** : for network isolation, Cloud Identity-Aware Proxy for controlling access to my web applications, and Cloud Armor for protecting against DDoS attacks.
- **Azure Virtual Network (VNet)** : enables network isolation, Network Security Groups (NSGs) provide firewall rules, and Azure DDoS Protection defends against DDoS attacks.

Security and risk management

Services for helping to detect, investigate, and respond to security threats:

- Amazon Guard Duty, AWS Security Hub, AWS Audit Manager and AWS Config: WS offers a suite of security services including Guard Duty for threat detection, Security Hub for centralized security management, Audit Manager for automating compliance audits, and Config for monitoring resource configurations.
- Security Command Center : provides visibility into security threats and vulnerabilities across Google Cloud services.
- Microsoft Defender for Cloud : unified security management and advanced threat protection capabilities for Azure resources.

Resource monitoring

Services for hierarchically manage resources by project, folder, and organization:

- AWS Resource Access Manager and AWS Organizations.
- Resource Manager.
- Azure Resource Manager.

Abuse prevention

Services that help protect websites from fraudulent activity, spam, and abuse without creating friction:

- AWS WAF CAPTCHA and AWS Fraud for fraud detection.
- reCAPTCHA Enterprise.
- Microsoft Dynamics Fraud.

Pricing Bill

AWS dominates the global market with a revenue share of 32 percent, while Azure has a 21 percent market share and GCP has an 8 percent market share. Microsoft and Google compete with AWS by reducing their prices.

AWS offers a pay-as-you-go model, allowing users to pay for the services they consume without any additional termination fees. The pricing for basic data storage of 200GB per month typically ranges between EUR 4-5 per month on average, depending on the location.

Azure also provides a pay-as-you-go pricing model similar to AWS, along with a free tier option and discounts through Reserved Instances, offering up to a 72 percent discount for committed usage. Additionally, Azure offers spot instances for purchasing VMs at low prices from spare capacity.

On the other hand, GCP offers multiple pricing models, including free tier options and long-term reservations. It provides 300 dollars credit for free, along with sustained use discounts based on usage percentages throughout the month. Moreover, GCP offers substantial discounts for products committed to certain usage levels for one or three years in advance, known as ”committed use.”

Discussion

In conclusion, after considering the strengths and features of Amazon Web Services (AWS), Google Cloud Platform (GCP) and Microsoft Azure, I have decided to choose Google Cloud Platform (GCP) for my online banking company.

GCP emerges as the best choice primarily due to its competitive pricing options, which include sustained use discounts, custom VMs, and generous free tiers and credits for new customers. This flexible pricing model aligns well with my company’s budget requirements and ensures cost-effectiveness in the long run.

Moreover, GCP offers superior encryption services, providing robust security measures to safeguard sensitive customer data. This is crucial for maintaining compliance with regulations and ensuring the privacy and integrity of financial information.

Furthermore, GCP excels in the realm of artificial intelligence (AI) and machine learning (ML) applications, with its focused approach to ML processes and the comprehensive Vertex AI platform. These capabilities will be instrumental in implementing document recognition and analysis during user registration and contract signing, enhancing the efficiency and accuracy of our operations.

While Azure offers strong hybrid cloud capabilities and AWS boasts an expansive catalog of services, GCP’s combination of affordability, advanced encryption features, and powerful AI and ML tools makes it the optimal choice for my online banking company’s cloud infrastructure needs.