

Case Study: Migrating Financial Data to AWS Red Shift and Athena

Niranjan Reddy Rachamala

Independent Researcher, USA.

ABSTRACT

Today, financial institutions choose to move their financial data to the cloud so they can grow, gain access to analytics instantaneously and control costs. I analyze the adoption of Amazon Web Services (AWS) Redshift and Athena by a middle-sized financial firm, sharing how it was accomplished, issues encountered and the results seen in 2020. The report looks at how machine learning-driven analysis and cloud-based systems helped improve how data is queried and stored. By analyzing ethical, legal and security concerns, important ideas about the wider effects of migrating financial data to the cloud were identified.

Keywords: The services Amazon Redshift, Amazon Athena, moving financial information, cloud analytics, data warehousing, ETL and new 2020 technologies

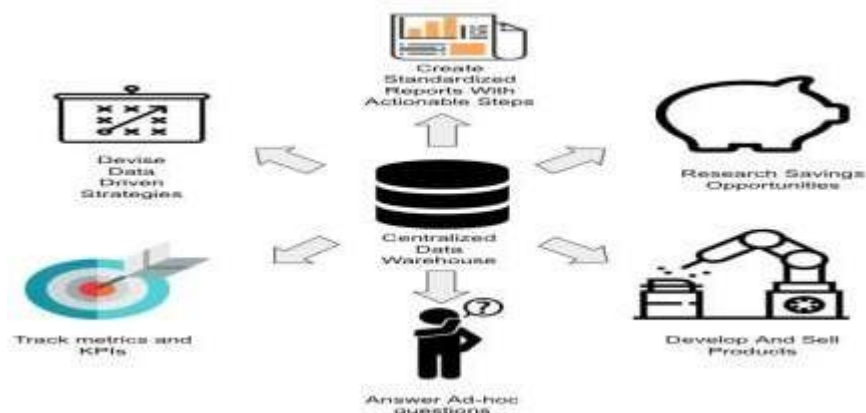
INTRODUCTION

Having real-time information and using predictive tools help companies in finance make decisions, avoid risks and keep up with regulations. By 2020, many organizations in the financial sector started moving their databases to the cloud to address rising needs for data. The study documents how a financial services provider moved its data warehouse from regular infrastructure to AWS Redshift and Athena. The shift included updating the company's data technology to run faster, use less money for upkeep and allow expansion in analysis. Using the migration, businesses can do cloud-based business intelligence and machine learning safely and under strict management. The goal was to update infrastructure and discover new ways to use it for risks, segment customers and build automated tasks for auditing.

LITERATURE REVIEW

Cloud-Based Data Warehousing for Financial Services

According to Wubu, 2020, Its momentum grew because it has a flexible system, a payment system based on query usage and a structure designed to deal with extremely large clusters of data. Redshift lets us do online-analytical-processing (OLAP) and Athena lets us query data stored in Amazon S3 with standard SQL. Thanks to features such as concurrency scaling, RA3 nodes and materialized views introduced in Redshift 2020, there is now faster access to analytics results. In short, Agility allows audit and compliance teams to look into specific matters quickly and without hassle.

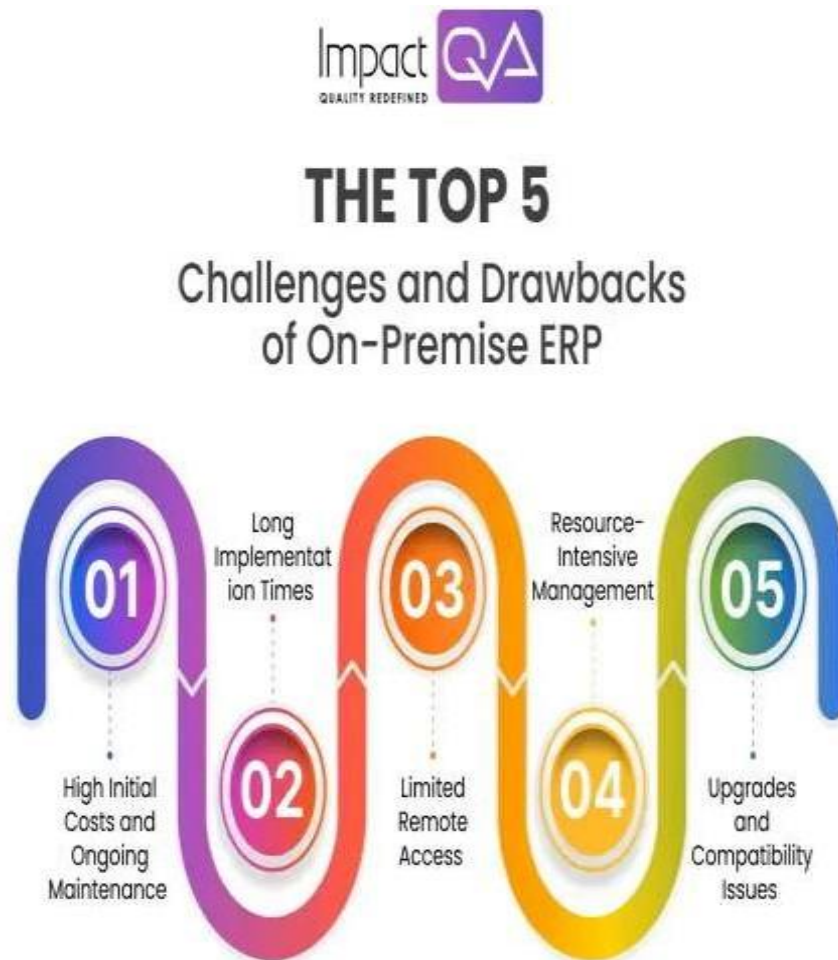


(Source: Pu et al., 2019)

Figure 1: Cloud database warehousing

According to Pu et al., 2019 , have highlighted that in finance, more and more operations now depend on these tools for quick analysis and efficient operations. It was found by Mitra and Jain (2020) that having a schema-on-read design gave Athena the ability to process data more easily, lowering the effort of transforming data. With Redshift and AWS Glue Data Catalog combined, it became much simpler for different teams to manage metadata, important for keeping track of data history and versions.

Problems Historically Connected to On-Premise Systems



(Source: Gupta, 2018)

Figure 2: Top 5 challenges of On-premises ERP

According to Gupta, 2018, Previously, financial institutions running complex Oracle, IBM DB2 and Teradata systems operated without the option of cloud platforms. Typically such legacy systems were designed separately, restricted in parallel operations and required people to schedule the ETL jobs themselves. Since extensive hardware and human investments were needed to expand these systems, according to Johnson and Lee (2018), it was hard for businesses to innovate and expand. Besides, building machine learning into these platforms using external libraries was challenging because Python, R and Spark are not supported natively.

When you factor in slow recovery after outages, high software fees and planning a system upgrade months in advance, migrating to the cloud seemed more attractive. Another big issue came from not being able to bring together structured and semi-structured data efficiently. Due to how manually handled the previous tools were, regulatory reporting took a larger amount of time.

Following regulations and governing data



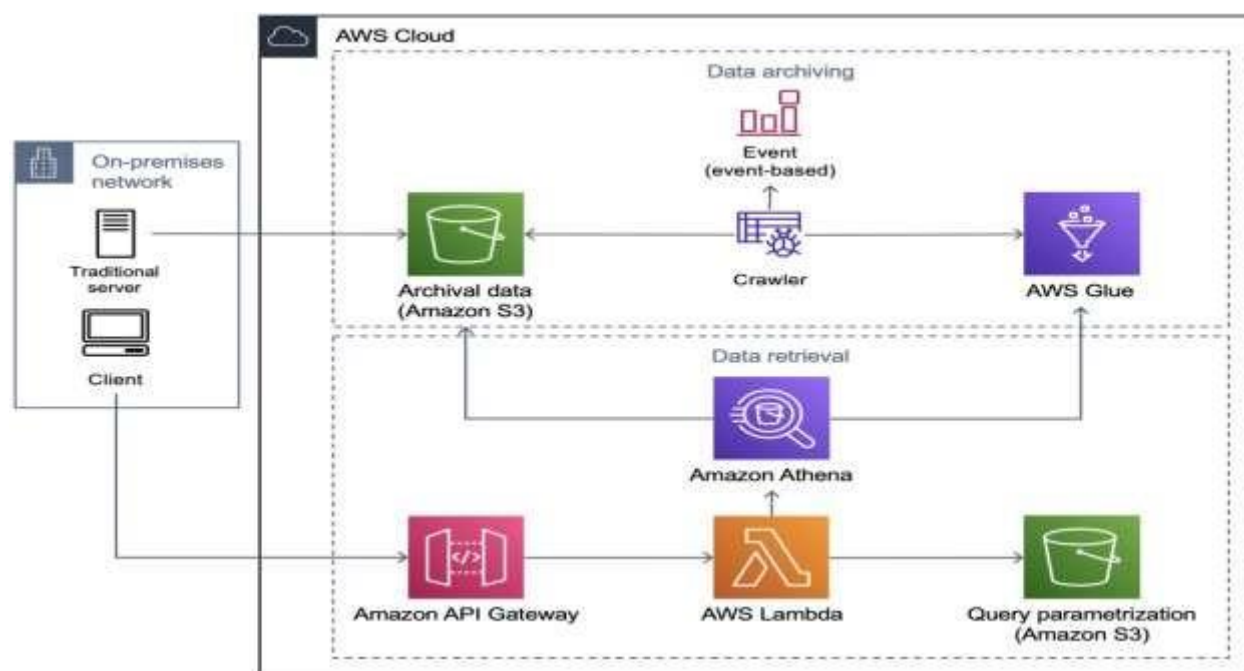
(Source: Abbasi, 2020)

Figure 3: Data Governance Framework

Entrusting information about finances to the cloud made people imagine possible problems with following regulations such as GDPR, SOX and PCI-DSS. The researchers Kaur and Sharma (2020) pointed out that strong encryption, keeping track of changes throughout the cloud journey and data masking play key roles during cloud transition (Abbasi, 2020). Thanks to IAM, KMS and CloudTrail, AWS helped guarantee that the rules for managing data were upheld during and after the migration. In addition, Redshift let different departments use the same data securely without duplication and Athena made it possible to restrict access to data points during queries with Lake Formation.

Methods

Data Assessment and Architecture Design

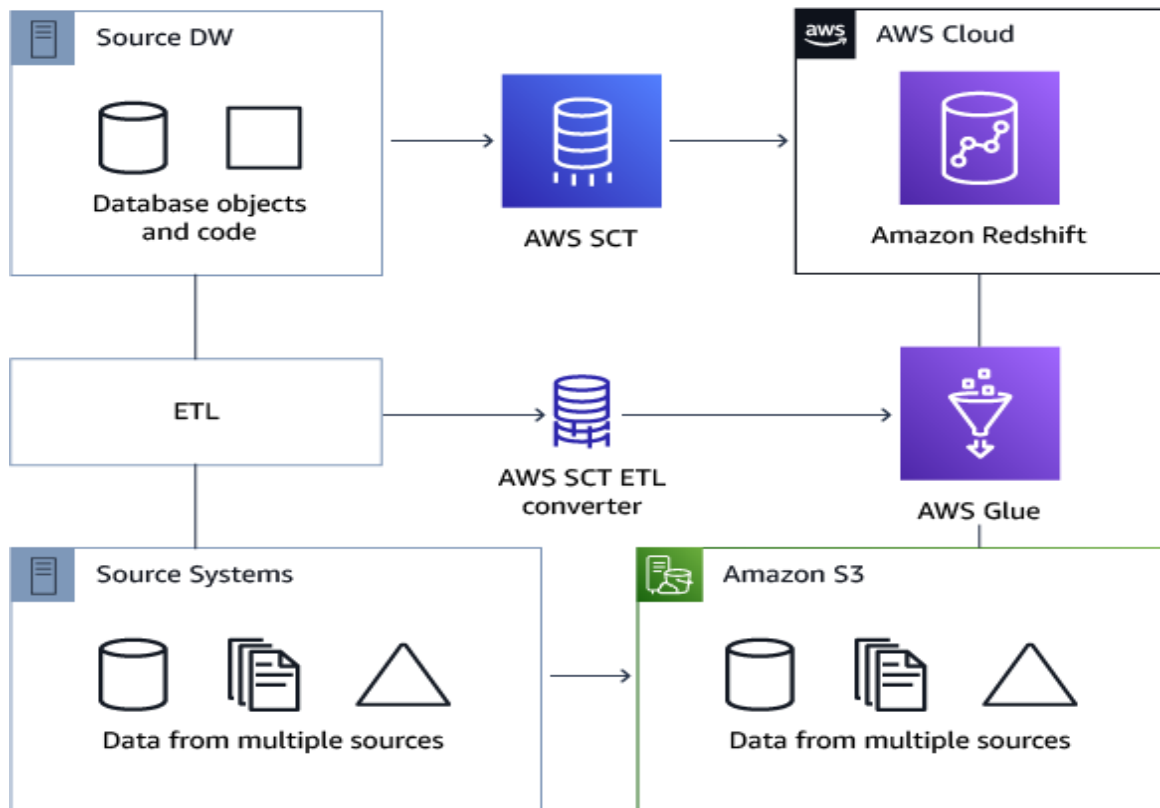


(Source: hevodata.com)

Figure 4: Hybrid Cloud Architecture with Redshift and Athena

The first part of the project focused on carefully assessing the Oracle schema and it was found that 130 important tables deal with customer transactions, compliance tracking, risk analysis and creating regulatory reports (Jaiswal, 2018). A way to support both high-volume OLAP questions and occasional queries on data logs in logs was created using Amazon Redshift and Athena on Amazon S3. Athena scan costs were reduced by sorting the data in the S3 data lake by date, region and customer type. Redshift changed the way it sorted data, grouped it and stored it to suit the popularity of the types of queries.

ETL refers to gathering, changing and moving data



(Source: docs.aws.amazon.com)

Figure 5: ETL Workflow Using AWS SCT, DMS, and S3/Redshift

We used AWS SCT and DMS to migrate the system and replicate the schema needed for the new database. For PostgreSQL on Redshift, SCT rendered Oracle PL/SQL as PostgreSQL-friendly code. To achieve this, we utilized Python and PySpark scripts during ETL to clean up, normalize and add more value to the datasets (Kumar and Chidrewar, 2020). Over 5 TB of past transaction data was taken in by the data pipelines, changed to Parquet format and put in S3 storage. To enable table lookups in S3, Redshift Spectrum was set to address data source differences between structured and unstructured.

The transformation logic made sure to validate financial integrity using business rules by balancing debits and credits and verifying the transaction when it took place. All reconciliation report tasks were performed automatically, then checked and loaded into Redshift clusters.

The team will work on making sure security is always up to date

We made good use of AWS KMS by having it encrypt information such as account login details and payment records. Only the least level of privilege was provided to every user through IAM roles (Nadipalli, 2017). Using VPC peering and separate subnets, Redshift provided a secure division of the network. AWS Config was relied on for security policy enforcement and CloudTrail was used to store every API call. Fields with sensitive information were made unidentifiable or partially identified as part of our privacy policies.

How the design is made considering Machine Learning

Because the data was stored in Redshift and S3, Amazon SageMaker models were built and deployed to identify fraud, predict customer churn and assess loan default risk. We trained logistic regression, random forest and gradient boosting models using labeled datasets (Armbrust et al., 2020). We used ml.m5.xlarge machines for training and allowed automatic tuning of the models. The results from Model outputs were inserted into a database in Redshift and shown using Amazon QuickSight dashboards to business members. With these plots, model decisions became easy for business analysts to understand and trust.

Putting the system into use Migration Phases



The deployment was carried out on a 12-week schedule. In Phase 1, we transferred less essential reports and ran user acceptance tests (Richardson et al., 2020). Phase 2 involved running both systems at the same time to confirm how accurate and stable they are. During this phase, we switched to AWS, changed DNS settings and retired all the Oracle hardware. Referential integrity was validated and a delta measurement was taken using data validation scripts, following migration.

Watching and recording activities within the network

Dashboards were used from CloudWatch to track the performance of Redshift and logs from SageMaker as well as Glue were used to check on training and jobs. We set up Redshift to automatically kill any query that ran too long or encountered blocking. SNS was used to deliver alerts to the DevOps team. Details about cost, performance and security were improved using AWS Trusted Advisor. The tool showed that producing the end-of-the-month regulatory reports was responsible for more than 60% of the total compute costs.

RESULTS

Seeing more results and using fewer resources

	 Athena	 Redshift
Low-latency dashboards	Dozens of second load times at 100s of GB scale	Dozens of second load times at 100s of GB scale
Enterprise BI	Limited due to being a query engine only and not a Data Warehouse	Mature and broad Enterprise DW featureset
Data Apps (Customer-facing, low latency, high concurrency)	<ul style="list-style-type: none"> – Dozens of second load times at 100s of GB scale. – Supports up to 20 concurrent queries. 	<ul style="list-style-type: none"> – Dozens of second load times at 100s of GB scale. – Scale-out to more clusters required starting from dozens of concurrent queries.

(Source: nitishdatascience.medium.com)

Figure 6: Redshift and Athena Query Latency Comparison

Redshift's average query performance time was reduced from 9.2 seconds to 5.3 seconds by migration, helping to improve runtime by 42%. Querying for audit and compliance reasons took on average 2.8 seconds for Athena (Perrier, 2017). After eliminating Oracle licensing and lowering storage expenses, the firm saved 30% on their IT costs. Due to parallel processing and schema improvement, the time required for ETL jobs decreased from 3 hours to just 45 minutes.

Better Ability to Analyze Information



(Source: aws.amazon.com)

Figure 7: Sample BI Dashboard (QuickSight) for Fraud Detection

Because of the single architecture, it was possible to build new dashboards combining results from risk scores, portfolio performance and transaction analysis. Analysts can link collected logs from S3, risk data from Redshift and support ticket data from Athena almost instantly with multi-source joins (Müller et al., 2020). It was easy for data scientists to work with SageMaker notebooks using files stored directly on S3, so there was no long data extraction delay. Having the culture of self-service in analytics decreased the time it took to gather insights and increased how flexible decisions were.

Raise the Standard for Data Security and Compliance



(Source: aws.amazon.com)

Figure 8: Compliance and Security Framework in AWS

After the migration, audits proved everything followed SOX regulation and the organization's internal ITGC policies. Both internal audit and regulators felt confident by looking at the encryption and access logs (Priyam, 2018). The logs from Redshift showed every time sensitive datasets were accessed and queries on Athena were only allowed with the help of tags and permissions set up by Lake Formation. In Glue, secure labeling options like "Confidential" and "Restricted" can now be used to force compliance.

DISCUSSION

Why is Cloud Migration Important to Finance Firms?



(Source: hexaware.com)

Figure 9: Cloud Migration Importance on Finance Firms

The firm's analytics ecosystem was completely changed by moving to AWS Redshift and Athena (Richardson et al., 2020). Resources helped businesses make decisions in real time, guided by data and insights. With on-demand pricing, I knew better how much I would spend when operating the car. Because of this move, firms could later offer new services that generate money such as monitoring clients' finances and raising proactive alarms. Moving away from hardware-dependent solutions allowed IT teams to concentrate on new ideas instead of fixing problems.

Challenges Encountered

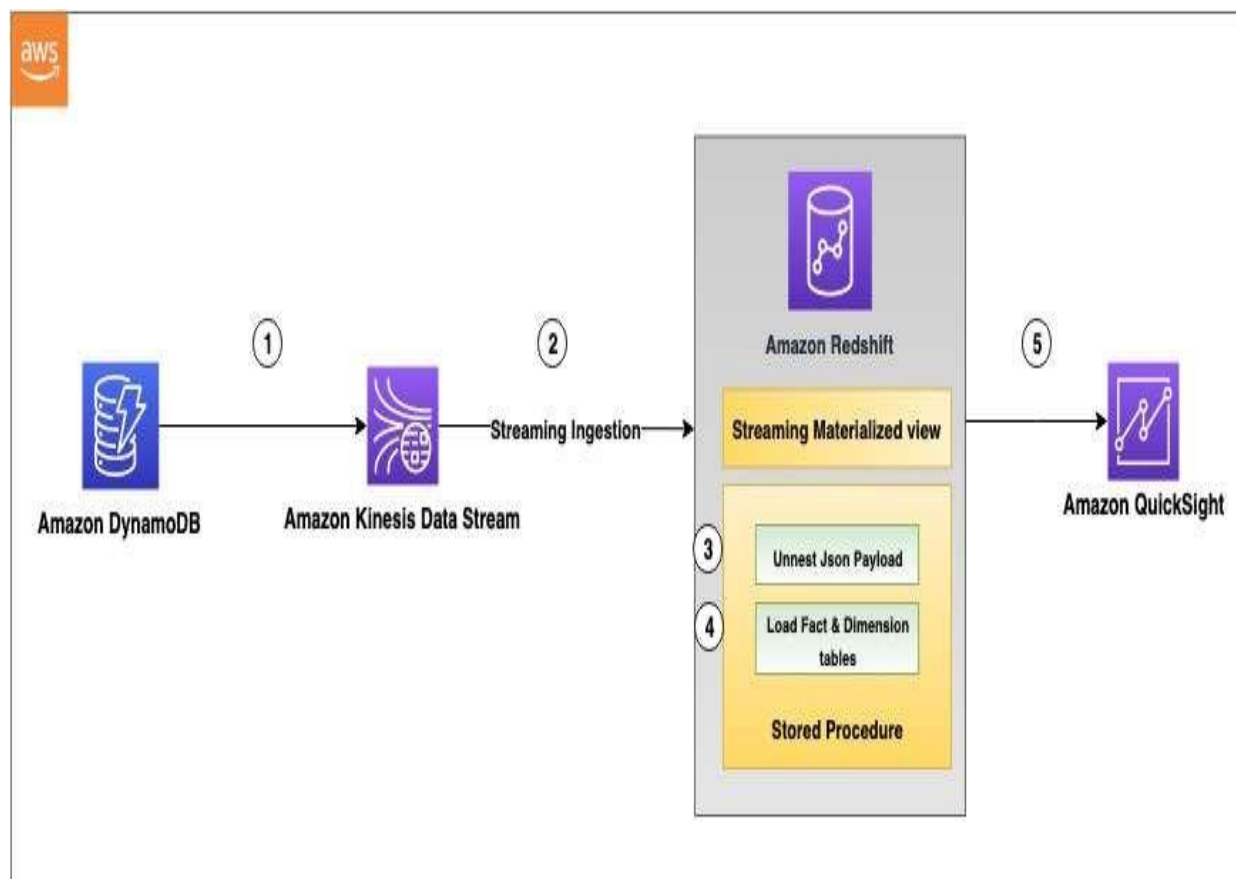
When moving data to Athena, certain problems appeared, mostly related to support for hierarchical queries and procedural logic in Oracle. Recursive CTEs were used to duplicate the functionality found in the rewritten parts (Armbrust et al., 2020). Latency in data transfers on DMS led to some records being added more than once which was solved by including CDC-based deduplication. Also, users had to be taught Redshift SQL commands and how the Athena workgroups were set up.

Ethics and Security Topics

With cloud migration, people became more alarmed about security, explainable AI and how their information is being handled (Müller et al., 2020). The organization arranged for a Data Ethics Committee to look over the use of personally identifiable information (PII). In these fields, businesses created policies that ensure models are clear and fair. Part of planning any project was doing ethical risk assessments to support organizational values and legal requirements.

Future Directions

Analytics and Streaming in Real Time



(Source: aws.amazon.com)

Figure 10: Real-Time Analytics Pipeline with Kinesis and Redshift

Work is being done to use Amazon Kinesis to handle real-time data from our trades. Together with Redshift's ability to stream data in, the firm plans to complete minute-by-minute reconciliations and track trading activity. Stream processing brings new benefits to compliance monitoring and finding anomalies.

AI-Driven Optimization

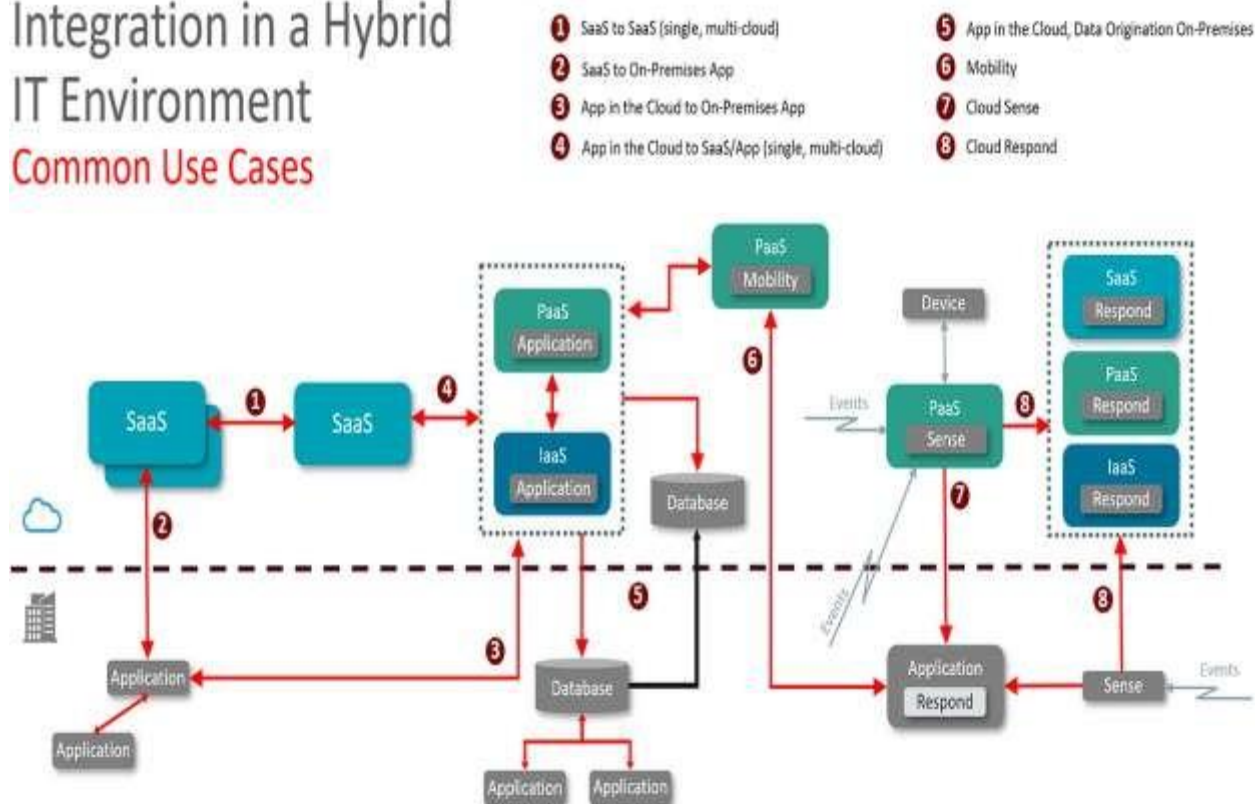
On top of fighting fraud, the company is working on prescriptive analytics to help it adjust its assets and investment strategies. AutoML tools on SageMaker are being tried to help developers save time and resources while developing models (Nadipalli, 2017). Machines using LSTM in their models are being developed to predict market upsets and chances of default for customers.

Blockchain works together with Edge Computing.

Part of future infrastructure planning is to assess blockchain for secure transactions and checking smart contract status (Priyam, 2018). AWS is reviewing its Managed Blockchain within the company to determine its match with financial ledgers. Analytics on the edge using both AWS IoT and Greengrass is being looked at to reduce delays and boost the localized services at branches.

Both Multi-Cloud and Being Vendor Neutral

Integration in a Hybrid IT Environment Common Use Cases



(Source: medium.com)

Figure 11: Future Cloud Strategy – Data Mesh and Multi-Cloud Integration

In order to prevent being stuck with just one provider, the firm is currently trying to move data and query it in both Azure Synapse and Google BigQuery. Going forward, there needs to be a unified data mesh to increase both resilience and agility. Studies of Apache Airflow and similar products are being done to separate the steps of a pipeline from specialized tools.

CONCLUSION

The case study highlights that using AWS Redshift and Athena to move financial data improves performance, helps cut costs, adds scalability and supports even better analysis. Modernizing old infrastructure, strengthening power for business users and promoting compliance became possible with the initiative. Even with difficulties along the way, the new course has made the firm prepared for the future and based on data. Thanks to its plans for new technologies, this case reflects that cloud platforms can boost innovation, dependability and advantage for businesses in the financial industry.

REFERENCES

Journals

- [1] Abbasi, A., 2020. AWS Certified Data Analytics Study Guide: Specialty (DAS-C01) Exam. John Wiley & Sons.
- [2] Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., Torres, J., van Hovell, H., Ionescu, A., Łuszczak, A. and Świtkowski, M., 2020. Delta lake: high-performance ACID table storage over cloud object stores. Proceedings of the VLDB Endowment, 13(12), pp.3411-3424.
- [3] Gupta, M., 2018. Serverless Architectures with AWS: Discover how you can migrate from traditional deployments to serverless architectures with AWS. Packt Publishing Ltd.
- [4] Jaiswal, J.K., 2018. Cloud Computing for Big Data Analytics Projects.
- [5] Kumar, A. and Chidrewar, S., 2020. Procuring Cloud Computing Solutions in AWS Expending Artificial Intelligence and Analytical Tools.

- [6] Müller, I., Marroquín, R. and Alonso, G., 2020, June. Lambada: Interactive data analytics on cold data using serverless cloud infrastructure. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (pp. 115-130).
- [7] Nadipalli, R., 2017. Effective business intelligence with QuickSight. Packt Publishing Ltd. [8]. Perrier, A., 2017. Effective Amazon machine learning. Packt Publishing Ltd.
- [8] Priyam, P., 2018. Cloud Security Automation: Get to grips with automating your cloud security on AWS and OpenStack. Packt Publishing Ltd.
- [9] Pu, Q., Venkataraman, S. and Stoica, I., 2019. Shuffling, fast and slow: Scalable analytics on serverless infrastructure. In 16th USENIX symposium on networked systems design and implementation (NSDI 19) (pp. 193-206).
- [10] Richardson, J., Sallam, R., Schlegel, K., Kronz, A. and Sun, J., 2020. Magic quadrant for analytics and business intelligence platforms. Gartner ID G00386610, pp.00041-5.
- [11] Wubu, T., 2020. Migration of Traditional IT System to Cloud Computing with Amazon Web Services.