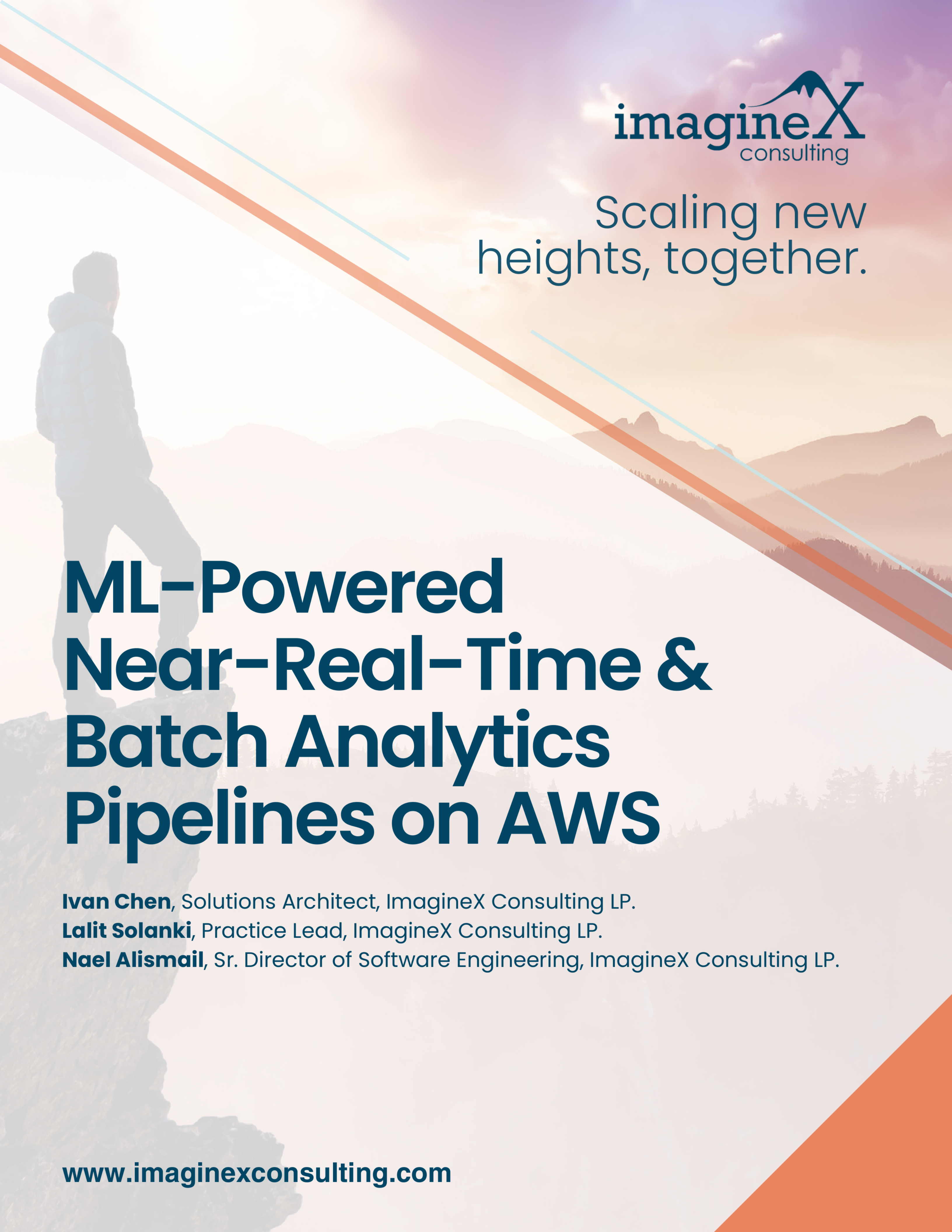


imagineX consulting

Scaling new heights, together.

ML-Powered Near-Real-Time & Batch Analytics Pipelines on AWS

Ivan Chen, Solutions Architect, ImagineX Consulting LP.
Lalit Solanki, Practice Lead, ImagineX Consulting LP.
Nael Alismail, Sr. Director of Software Engineering, ImagineX Consulting LP.

www.imaginexconsulting.com

About us

At ImagineX, we understand that every organization is unique and that there is no one-size-fits-all solution.

That's why we are committed to the #bebetter mission of developing and delivering scalable, secure, and user-centric solutions that optimize your operations, increase your competitive advantage, and fortify your defenses.

Table of contents

Executive Summary & Business Use Cases 3

Challenges 4-5

The Proposed Solution 6

Architecture Decisions and Recommendations..... 7

Comparison Between AWS Glue and Amazon EMR 8

Business Values and Benefits 9

AWS Cost Estimates 10

Conclusion & About the Authors 11

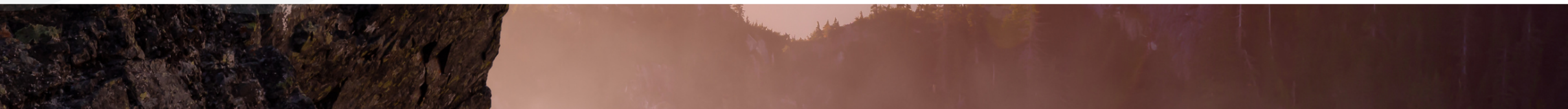
Executive Summary

Machine Learning (ML) is gradually gaining popularity. It allows organizations to improve their business operation efficiency, enable new products or services, or make more informed decisions. Many organizations are currently seeking to enable machine learning models as a part of their data processing pipelines for ML-powered analytics processes. In this white paper, we demonstrate solutions of how to leverage Glue and SageMaker services together to build ML-powered analytics pipelines for near real-time & batch data processing.

Business Use Cases

There are several popular business use cases that organizations can use to leverage machine learning (ML) models to achieve better outputs:

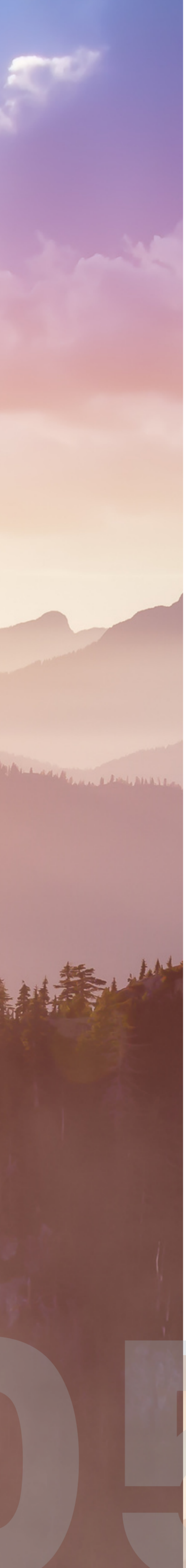
- Customer Churn Prediction: By analyzing customer behavior and historical data, ML models can predict the likelihood of customers churning. This enables businesses to take proactive measures to retain customers and improve customer satisfaction.
- Fraud Detection: ML models can be used to identify fraudulent activities in near real-time or batch data, helping businesses prevent financial losses and maintain the integrity of their systems.
- Predictive Maintenance: ML models can analyze sensor data in real-time or in batches to identify patterns and predict equipment failures. This helps businesses schedule maintenance activities, reduce downtime, and optimize resource allocation.
- Personalized Recommendations: By processing user behavior and preferences, ML models can provide personalized recommendations for products, services, or content. This enhances the user experience and increases customer engagement.
- Demand Forecasting: ML models can analyze historical sales data and external factors to predict future demand. This helps businesses optimize inventory levels, plan production, and improve supply chain efficiency.



Challenges

One of the challenges that organizations have to face is how to handle such a large amount of data effectively and efficiently with machine learning models in complex data processing pipelines. Generally speaking, there are three major approaches to handle data processing tasks for business in different time-sensitive scenarios. These include real-time, near real-time (i.e., micro-batch), and batch.

- **Real-time data processing**: Real-time data processing refers to the immediate processing and analysis of data as it is generated or received. This approach aims to provide instant insights and responses based on the most up-to-date information available. Real-time processing is typically used in applications where time-sensitive decisions or actions need to be taken, such as financial trading, fraud detection, or real-time monitoring systems. It requires low-latency systems capable of processing and analyzing data in near-instantaneous time frames (e.g., milliseconds).
- **Near real-time data processing**: Near real-time data processing involves processing and analyzing data with minimal delay, usually within seconds or minutes of its generation or receipt. While not as immediate as real-time processing, near-real-time processing still provides timely insights and enables near-instantaneous decision-making. Near real-time processing is commonly used in applications such as online customer support, inventory management, or social media analytics. It requires systems that can handle data with low to moderate latency, balancing timeliness with processing capacity.
- **Batch data processing**: Batch data processing involves processing and analyzing data in large volumes, typically collected over a period of time. In this approach, data is collected and stored, and then processed in batches at a later time. Batch processing is characterized by high throughput and is often used for tasks that don't require immediate responses, such as generating reports, running analytics on historical data, or performing large-scale data transformations. Batch processing can handle massive datasets efficiently but lacks the immediacy of real-time or near real-time processing.



Challenges...continued

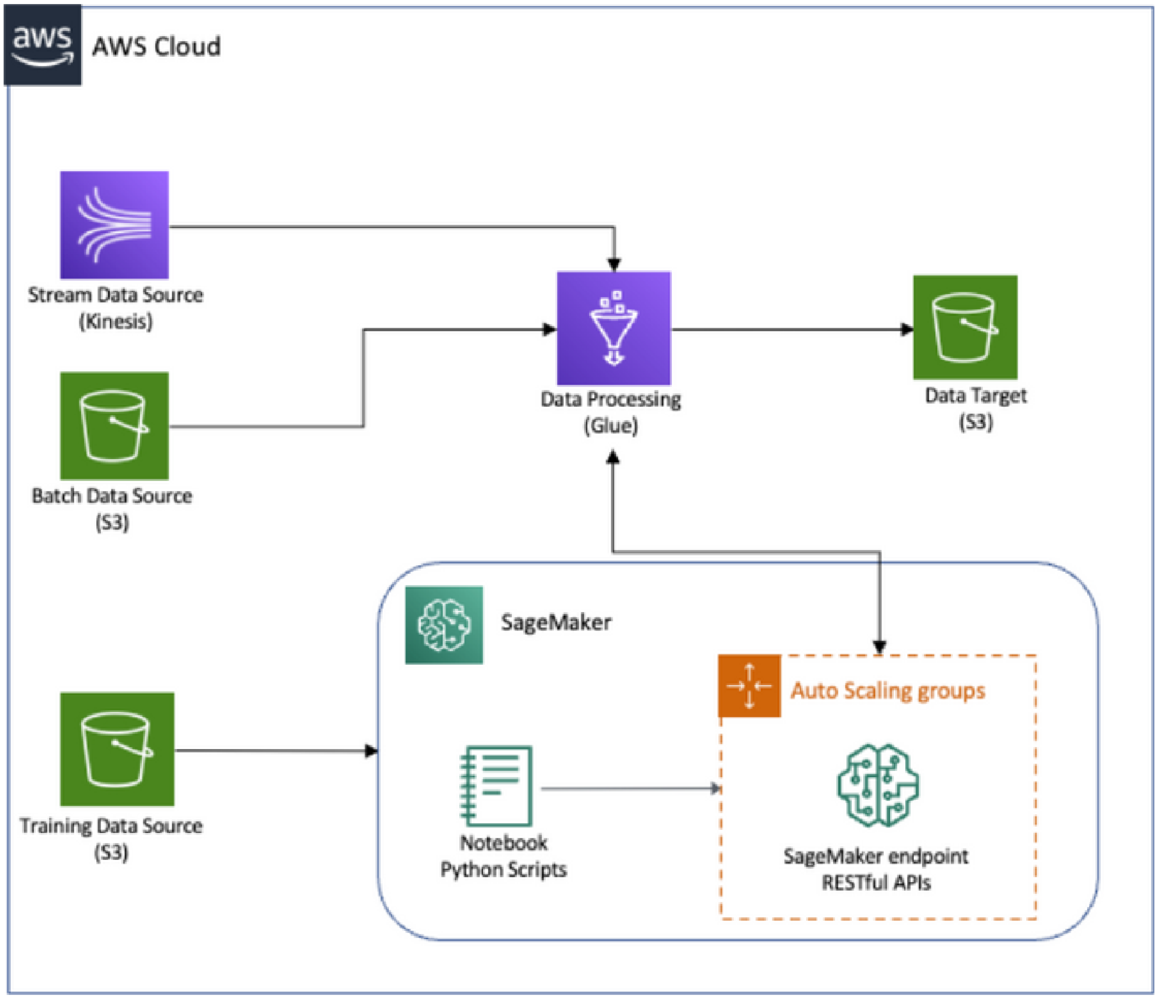
While all three approaches can handle a large amount of data with today's technology, real-time data processing can be more complex, costly and resource-intensive compared to near real-time and batch processing. It usually requires specialized infrastructure with high-performance hardware, such as powerful servers, high-speed networks, and real-time data streaming platforms. Building and maintaining such infrastructure can be very complicated and costly. Moreover, integrating large and complex machine learning (ML) models into real-time data processing systems can indeed impose additional burdens on an organization's limited budget and IT infrastructure and development teams, which can be seen in this AWS blog. Thus, many organizations would prefer to use near real-time or batch processing when considering leveraging ML models in their data analytics pipelines unless there are real business needs for real-time.

Fortunately, Amazon Web Services (AWS) offers a wide range of services that can help organizations to implement the solution. One particular approach is to host machine learning models on SageMaker endpoints, then create batch and near real-time (micro-batch) Glue jobs to invoke the machine learning model to generate prediction data along with the original input data, and finally save all the data into a data lake environment, like an S3 bucket.

In the rest of this white paper, we will discuss in detail about this proposed AWS solution, its business values and benefits, and cost estimates.

The Proposed Solution

With AWS, many organizations are trying to leverage ML model creation on Amazon SageMaker to improve their ML development processes. The AWS Glue service is a popular serverless tool for data engineers and developers to process a large amount of data, which is usually scaled from GB size to TB size daily. One can consider to invoke the ML models hosted on SageMaker from a Glue job for batch and near real-time processes. The following diagram illustrates the architecture of the solution.



This Github repo will walk through the detailed steps to setup and configure AWS services for this solution, which was coded for the use case of Customer Churn Prediction. This requires some basic machine learning (ML) knowledge and the knowledge of AWS services, such as SageMaker, IAM, Glue and Kinesis. The high-level steps include:

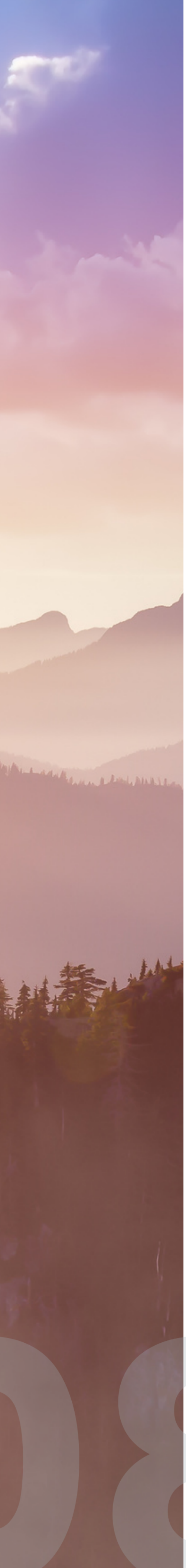
- Create proper IAM roles.
- Create and deploy a Machine Learning (ML) model.
- Create a Glue Spark batch job to invoke the ML model for predictions
- Create a Glue (real-time) streaming job to invoke the ML model for real-time predictions
- Clean up

All the code for this solution can be found in the GitHub repo as well.

Architecture Decisions and Recommendations

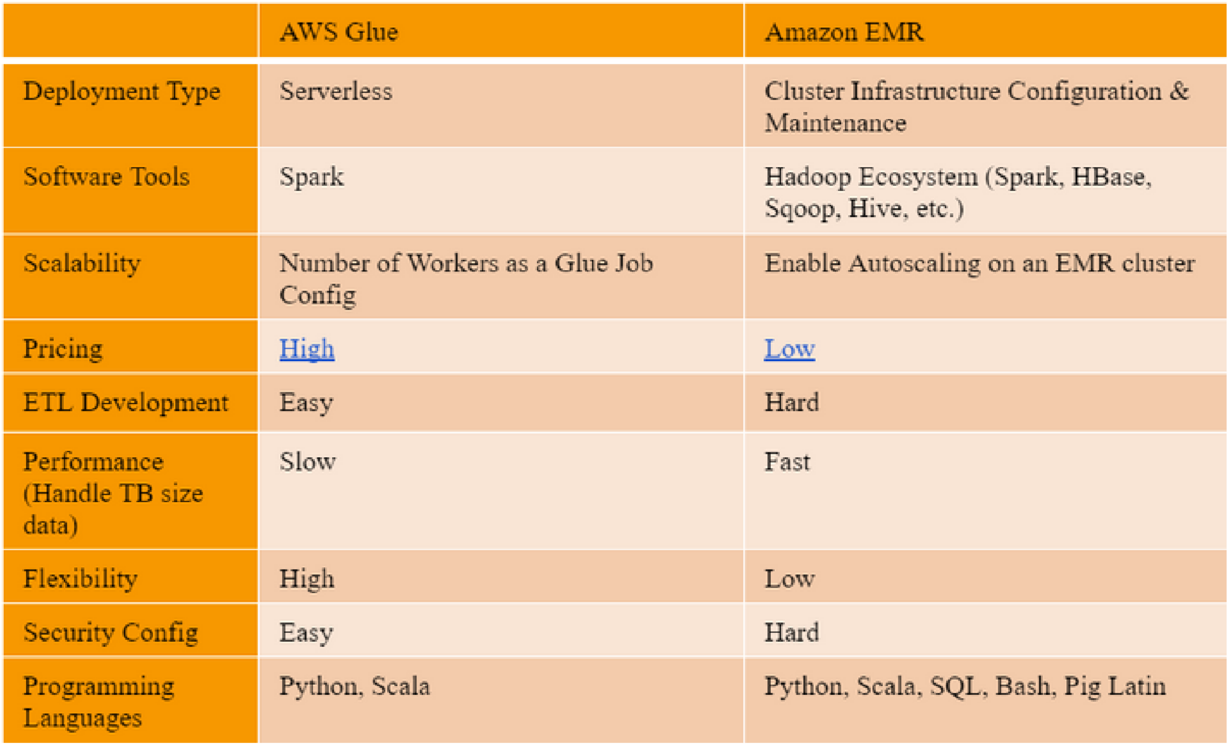
- **AWS Glue** – is a fully managed extract, transform, and load (ETL) service that helps with data integration and transformation. It is easy to discover, prepare and combine data for analytics, machine learning and application development. It provides all of the capabilities needed for data integration.
- **Amazon Kinesis** – is a fully managed streaming data service that makes it easy to collect, process and analyze real-time streaming data. With Amazon Kinesis, people can ingest real-time data such as key business product and price data, application logs, website clickstreams and IoT data for data lake storage, machine learning, analytics, and other applications.
- **Amazon Simple Storage Service (S3)** - is an object storage service that offers industry-leading scalability, data availability, security, and performance. It provides scalable and secure storage for storing and accessing large volumes of data. Amazon S3 is used for storing and protecting data for a range of use cases, such as data lakes, websites, big data analytics and many more.
- **Amazon SageMaker** - is a fully managed service that simplifies the end-to-end process and helps data scientists and developers to prepare, build, train, and deploy high-quality machine learning (ML) models quickly by bringing together a broad set of capabilities purpose-built for ML.

An alternative option to the Glue service is the Amazon EMR service. Creating Amazon EMR processes to invoke SageMaker ML endpoints can be very similar to the work demonstrated in this white paper.



Comparison Between AWS Glue and Amazon EMR

Both AWS Glue and Amazon EMR services can be used for data processing. We generally recommend AWS Glue to be the primary service in this proposed solution due to the flexibility of the serverless architecture. The following table summarizes the major differences between AWS Glue and Amazon EMR for ETL processes.



	AWS Glue	Amazon EMR
Deployment Type	Serverless	Cluster Infrastructure Configuration & Maintenance
Software Tools	Spark	Hadoop Ecosystem (Spark, HBase, Sqoop, Hive, etc.)
Scalability	Number of Workers as a Glue Job Config	Enable Autoscaling on an EMR cluster
Pricing	High	Low
ETL Development	Easy	Hard
Performance (Handle TB size data)	Slow	Fast
Flexibility	High	Low
Security Config	Easy	Hard
Programming Languages	Python, Scala	Python, Scala, SQL, Bash, Pig Latin

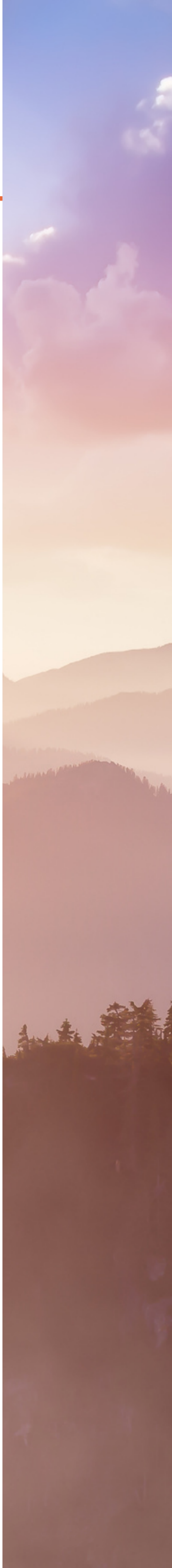
When To Choose The AWS Glue Service

- Data size is huge but structured (i.e., table structure and known format
- (CSV, parquet, orc, json)
- Data lineage is required
- Vertical scaling is not required. Due to the serverless architecture of Glue, horizontal scaling is preferred
- You don't want the overhead of managing a large cluster and pay only for what you use

When To Use Amazon EMR Service

- Data is huge but semi-structured or unstructured where you can't take any benefit from Glue catalog
- Lineage is not required
- Vertical & horizontal scaling is needed
- Run multiple jobs on a single EMR cluster reducing job maintenance and potentially saving costs
- In case of structured data, you should use EMR when you want more Hadoop capabilities like hive, presto for further analytics

Business Values and Benefits



Using AWS Glue and SageMaker can provide several business values and benefits. Here are some of them:

- **Data integration and ETL automation**: AWS Glue allows businesses to easily extract, transform, and load (ETL) data from various sources into data lakes or data warehouses. It automates the process, reducing manual effort and improving efficiency.
- **Data discovery and cataloging**: AWS Glue provides a data catalog that automatically discovers, catalogs, and organizes metadata from various data sources. This makes it easier for businesses to find and understand their data assets, leading to better decision-making.
- **Scalability, serverless and cost-effectiveness**: Both AWS Glue and SageMaker are fully managed services, meaning businesses don't have to worry about infrastructure management. They can easily scale up or down based on demand, paying only for the resources they use. This allows businesses to save costs and focus on their core activities.
- **Machine learning capabilities**: AWS SageMaker is a powerful machine learning platform that provides businesses with the tools and infrastructure to build, train, and deploy machine learning models at scale. This enables businesses to leverage their data for predictive analytics, personalization, fraud detection, and more, leading to improved customer experiences and operational efficiency.
- **Integration with other AWS services**: Both AWS Glue and SageMaker seamlessly integrate with other AWS services, such as AWS Lambda, Amazon S3, Amazon Redshift, and Amazon EMR. This allows businesses to build end-to-end data pipelines and leverage a wide range of services for analytics, storage, and processing.
- **Security and compliance**: AWS Glue and SageMaker come with built-in security features and compliance certifications, ensuring the protection of data and adherence to industry regulations. This gives businesses peace of mind when working with sensitive or regulated data.
- **Developer productivity**: AWS Glue and SageMaker provide easy-to-use interfaces, SDKs, and APIs, making it simpler for developers to work with data and build machine learning models. This improves developer productivity and reduces time to market for new applications and insights.

AWS Cost Estimates

In this section, we will provide you with AWS cost estimates on solution implementation for the use case of Customer Churn Prediction. There are two types of costs for this solution: cost for one-time ML model training and ongoing cost or total cost of ownership (TCO) for operation.

Cost for One-time ML Model Training

The cost for one-time machine learning model training on AWS SageMaker can vary depending on several factors, including the complexity of the model, the size of the dataset, and the duration of the training process. Here we assume:

- The training dataset is less than 10,000 records.
- The machine learning algorithm is xgboost.

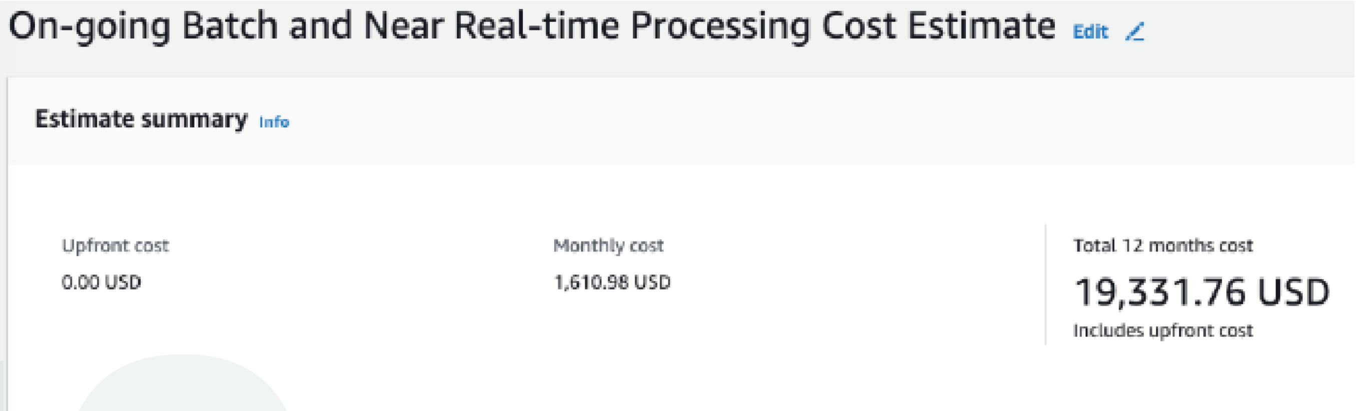
This leads basically to the fact that the training time is less than 1 hour with the AWS EC2 instance of ml.m4.xlarge (4 vCPU, 16 GiB memory, 10 GB storage). The one-time cost for this training process is about $3.11 USD. The details of this cost estimate on AWS pricing calculator can be found at calculator.aws/#/estimate. You can revise it based on your preferences or specific business requirements.

Cost for ML-powered Analytics Pipelines

The cost for ML-powered analytics pipelines can heavily depend on the size of the dataset as well as the complexity of the model. Here we assume:

- The data volume for batch and near real-time processing is 100 GB per day and about 3 TB per month.
- The duration of data processing is 2 hours batch and 12 hours near real-time processing per day.
- The Kinesis data stream transfers 50 GB data to Glue service per day.
- The trained machine learning model is based on xgboost.

This leads to the fact that we need 5 Data Processing Unit (DPU) for batch processing and 5 Data Processing Unit for near real-time processing (4 vCPU and 16 GB of memory per each). The total cost for this ML-powered analytics pipeline is about $1610.98 USD monthly or $19,331.76 USD annually, as shown below.



On-going Batch and Near Real-time Processing Cost Estimate

Estimate summary

Upfront cost	Monthly cost	Total 12 months cost
0.00 USD	1,610.98 USD	19,331.76 USD Includes upfront cost

The details of this cost estimate on AWS pricing calculator can be found here: calculator.aws/#/estimate. You can revise it based on your preferences or specific business requirements.

Conclusion

In this white paper, you learned how to prepare your own ML endpoint running on SageMaker and create Glue batch and near real-time streaming jobs to invoke the ML endpoint to add prediction results along with the original input data set. By doing this, you enable a capability to build an analytics pipeline with the power of the machine learning (ML) by leveraging AWS Glue jobs and SageMaker ML endpoints.

About the Authors

Ivan Chen is a Solution Architect at ImagineX Consulting LP. He helps customers create AI/ML solutions as well as data lake, data warehouse and database solutions. He holds a Ph.D. degree in Computer Science. In his spare time, Ivan enjoys travel and soccer.

Nael Alismail is a forward-thinking technology executive with a track record of maximizing business value for growth-stage tech organizations and startups. He collaborates closely with business visionaries to harness cutting-edge technology, creating innovative technology solutions that drive profitability. In his spare time, Nael spends his time watching football, camping, and playing with his children.

Lalit Solanki is tech evangelist, builder, mentor, and a business professional. In his spare time, he's actively shaping the future while helping others achieve their dreams. Lalit's mission is to guide and inspire people in the tech world, drive innovation, and support individuals and businesses on their journeys to success.