AGILISIUM
BIG ON CLOUD. BIG ON DATA.

aws PARTNER Advanced Tier Services

How to Optimize Big Data Analytics with Amazon EMR?

Migration Guide: Apache Hadoop to Amazon EMR

Contents

The Need For Amazon EMR Migration 3
- Challenges with Hadoop On-Premises 3
- The Need of the Hour for Organizations 3
- Why AWS EMR Migration? 4

How Does It Work? 5

Use Cases of Amazon EMR 6

Preferred Migration Strategy 7

General Best Practices for Migration 8

Setting Migration Goals: 9
- Key Challenges 9
- Security Model 9
- Data Catalog / Metastore 9
- High-Velocity Data 10
- Hive /Impala SQL to Redshift 10

Inventory of Current Platform 11
- Ingestion 11
- Transformations 11
- Analytics 12
- Security and Governance 12

Mapping 13
- Implementation 14
- Ongoing Managed Services 15

Case Study 16

AWS Funded Exclusive Agilisium Consulting Offers 18

Meet Agilisium 19

The Need For Amazon EMR Migration

In this technology-driven world, organizations are quite aware of the potential of data analytics and processing frameworks such as Apache Hadoop. While this awareness is appreciable, still, implementing the same effectively remains to be a challenge for several organizations. AWS EMR migration helps organizations shift their Hadoop deployments and big data workloads within budget and timeline estimates.

AWS EMR is recognized by Forrester as the best solution for migrating Hadoop platforms to the cloud. Upgrading and scaling hardware to accommodate growing workloads on-premises involves significant downtimes and is not economically feasible. This has further prompted organizations to re-architect using AWS EMR to build a modern system that is future-ready, high-performing, and cost-effective.

Challenges with Hadoop On-Premises

Scalability is challenging for organizations with Hadoop deployed on-premises as it involves the purchase of extra hardware. Additionally, it resists achieving elasticity and utilizing clusters for longer durations. The costs associated with workloads keep on increasing due to the 'always on' infrastructure while the data recovery and high availability must be managed manually.

The Need of the Hour for Organizations

- Easily scalable and flexible infrastructure that can be quickly provisioned as per requirements.
- Reduced admin dependency with a completely managed service.
- Cost optimization with the ability to switch the infrastructure on and off based on workload requirements.
- Innovative schemes for improved return on investment (ROI) in the long term.
- Exploring new open-source technologies by spinning up sandboxes in real-time.
- Integrating cloud security with the Hadoop ecosystem.

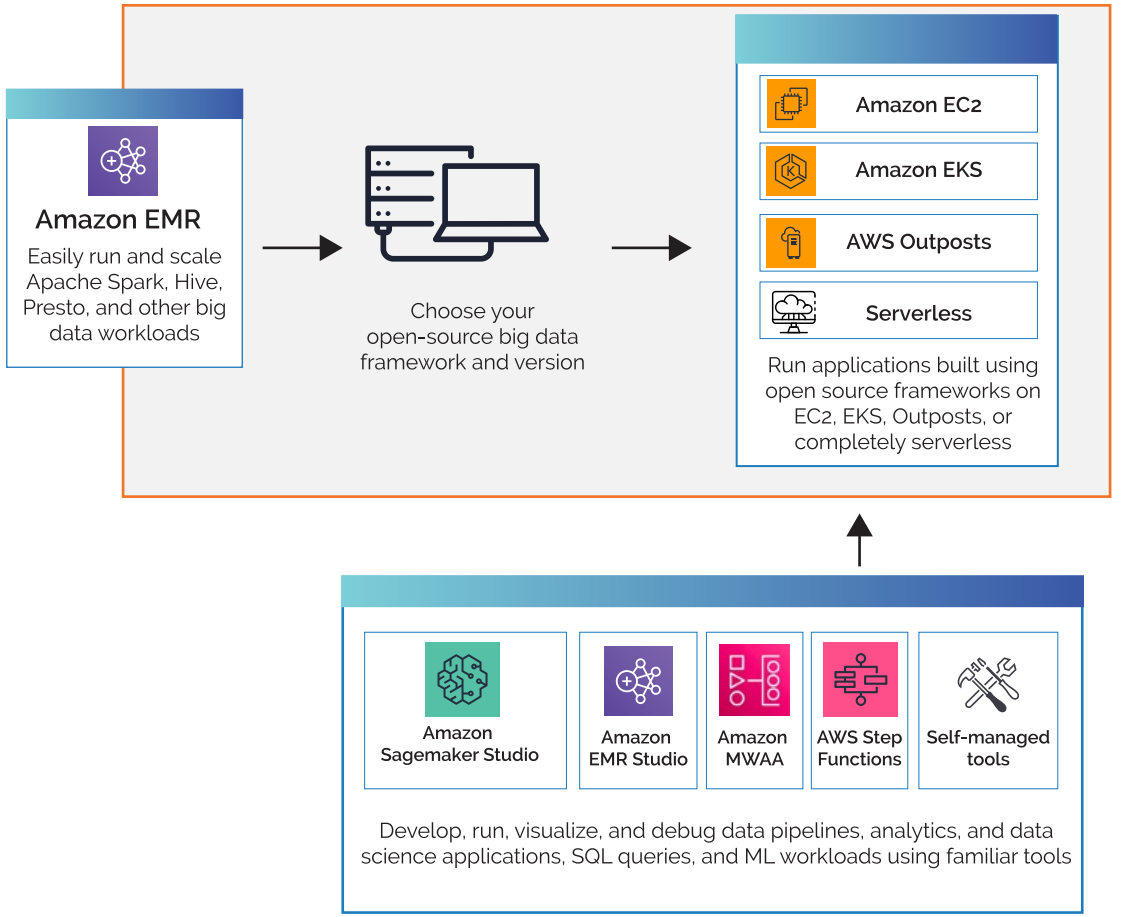
Why AWS EMR Migration?

Data-driven insights and cost optimization are primary considerations of organizations to achieve near-zero downtime of workloads with faster business value. Following are some key USPs of AWS EMR migration design that help organizations achieve the aforementioned.

- Decoupling of the storage and compute systems.
- A seamless data lake environment with Amazon S3.
- A stateless compute infrastructure.
- Cluster capabilities that are consistent and transient.
- Cluster fragmentation based on business units for improved isolation, customization, and cost allocation.

How Does It Work?

Amazon EMR is a cloud big data platform for running large-scale distributed data processing jobs, interactive SQL queries, and machine learning (ML) applications using open-source analytics frameworks such as Apache Spark, Apache Hive, and Presto.



Use Cases of Amazon EMR

Amazon EMR can be used to process applications with data intensive workloads.

Some of the common use case examples for Amazon EMR are:

Data Mining

Log file analysis

Web indexing

Machine learning

Financial analysis

Scientific simulations

Data warehousing

Bioinformatics research

Apart from these there could be several specific use cases in your organization that might require large scale data computation, for all such use cases you can use Amazon EMR.

Preferred Migration Strategy

Lift and Shift

This strategy helps companies achieve Hadoop to EMR migration faster to accelerate decommissioning of their on-premises data center. This enables organizations to eliminate cost-intensive hardware upgrades. The lift and shift strategy guides organizations to keep their existing Hadoop segregated and classified by utilizing AWS S3. Additionally, it helps them in decoupling resources, limiting code transformations to bare minimum. The simple lift and shift Hadoop to EMR migration approach moves the code as is to the cloud environment.

Replatform

The replatform strategy for Hadoop to EMR migration enables organizations to maximize their cloud migration advantages. This is basically done by utilizing the entire set of features provided by AWS EMR. With this strategy, organizations can fine tune their workloads and infrastructure for cost-effectiveness, scalability, and performance. Additionally, this strategy allows organizations to integrate their Hadoop ecosystem with cloud monitoring and security. Although replatform is similar to the lift and shift approach, it offers relatively lesser optimizations when it comes to cloud features and offerings.

Re-Architect

The strategy of re-architecting Hadoop on AWS EMR helps organizations to re-imagine their ecosystem of insights in the cloud. It helps them democratize their data to a larger customer pool while reducing the time-to-insight. This can be primarily attributed to the capabilities of streaming analytics, which provides organizations to self-service their requirements while building greater capabilities.

The re-architect strategy covers resolving all challenges of organizations, ranging from the analysis of business priorities to building a cloud-based data platform. The strategy basically involves changing the architecture with the help of cloud-native services to enhance performance, provision scalable solutions, and improve cost effectiveness of the infrastructure.

General Best Practices for Migration

Migrating big data and analytics workloads from on-premises to the cloud involves careful decision making. The following are general best practices to consider when migrating these workloads to Amazon EMR:

- Consider using AWS Glue, Amazon Redshift, or Amazon Athena. Although Amazon EMR is flexible and provides the greatest amount of customization and control, there is an associated cost of managing Amazon EMR clusters, upgrades, and so on.
- Consider using other managed AWS services that fulfill your requirements as they may have lower operational burden, and in some cases, lower costs. If one of these services does not meet your use case requirements, then use Amazon EMR.
- Use Amazon S3 for your storage (data lake infrastructure). A data lake is a centralized repository that allows you to store all of your structured and unstructured data at any scale.
- Data can be stored as-is, without having to first structure the data. You can execute different types of analytics on the data, from dashboards and visualizations to big data processing, real-time analytics, and machine learning.
- Amazon S3 is architected for high durability and high availability and supports lifecycle policies for tiered storage.
- Using Amazon S3 as the Central Data Repository.

Decouple compute from storage so that you can scale them independently. With your data in Amazon S3, you can launch as much or as little compute capacity as you need using Amazon Elastic Compute Cloud (EC2).

Setting Migration Goals:

Before beginning any migration, it is extremely important to define and document the goals of the migration upfront. Without clear, specific goals, the end-state solution will not meet the intended business objectives — especially since some goals may be in direct competition with others.

For example, migrating to the cloud as fast as possible, with as few changes as possible, would not result in the lowest cost solution.

Example goals:

- Data center exit project
- Enablement of new capabilities like GraphDB, Machine Learning Services, etc
- Better alignment of resource usage with costs
- Faster time-to-value for new data products

Key Challenges

Security Model

The number-one challenge our customers face when migrating is developing a proper security model on AWS. Configuring AWS IAM policy is significantly different compared to using traditional Active Directory for authorization and authentication. Companies often leverage 3rd-party tools, or have in-house tools that provide rights management to systems and data, and those systems leverage AD authentication and group membership for authorization. While some AWS services provide integration to AD, they do not provide consistent authorization to all services and data.

Data Catalog / Metastore

The number-one challenge our customers face when migrating is developing a proper security model on AWS. Configuring AWS IAM policy is significantly different compared to using traditional Active Directory for authorization and authentication. Companies often leverage 3rd-party tools, or have in-house tools that provide rights management to systems and data, and those systems leverage AD authentication and group membership for authorization. While some AWS services provide integration to AD, they do not provide consistent authorization to all services and data.

High-Velocity Data

Kudu and HBase are unique within Hadoop, providing a data store for high-velocity data with huge update rates. There are similar services within AWS, but they would require significant application changes along with a new architectural approach. Agilisium can identify those use cases quickly before the project gets too far down the wrong path. Catching these patterns early can help the business adjust expectations or ensure the proper amount of resources are applied to these use cases to deliver the desired business outcome.

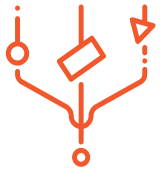
Hive /Impala SQL to Redshift

Converting and validating SQL between SQL engines can be time-consuming and error-prone. Agilisium has developed working models for identifying and transforming SQL from Hive/Impala to Redshift. This helps accelerate the migration activities and reduces the burden of manually rewriting these queries.

Inventory of Current Platform

The first step of any migration is to conduct an accurate inventory of the current platform, which then provides the basis for scoping the effort, timeline, technical capabilities, and strategy for the entire migration.

The inventory is broken into four broad capability categories: ingestion, transformation, analytics, and security/ governance. Within each category, the scope is further broken down into patterns, technologies, workloads, and complexity.



Ingestion

- Patterns — Streaming, File, CDC, RDBMS, etc.
- Technologies — Sqoop, Informatica, Talend, Attunity, etc.
- Workloads — Number of source systems, number of pipelines, number of databases, number of tables, frequency, etc.
- Complexity — Low (File, RDBMS), Medium (Streaming), High (Real-time dashboards from streams)

Transformations

- Patterns — Batch, data-triggered, time-triggered, OCR, data cleansing, data masking, etc.
- Technologies — Hive, Impala, Pig, Spark
- Workloads — Number of source databases, Hive jobs, Impala queries, Pig jobs, Spark jobs
- Complexity — Low (data filtering, data validation), Medium (creating new information models), High (OCR, real-time, translations)

Analytics

- Patterns — BI tool integration, report creation, dashboards, data exports
- Technologies — Hive, Impala, Pig, Spark
- Workloads — Number of users, number of queries, number of source databases, Hive jobs, Impala queries, Pig jobs, Spark jobs
- Complexity — Low (Reports), Medium (10+ Analysts), High (1,000+ Analysts)

Security and Governance

- Patterns — Sentry, Ranger, HDFS ACLs, Kerberos, AD, encryption-at-rest, encryption in-transit
- Technologies — Sentry, Ranger, AD, Kerberos, SSL
- Workloads — Number of AD groups, number of Sentry workspaces, databases.
- Complexity — Low (10 workspaces , Medium (100 workspaces), Hard (1,000+ workspaces)

Mapping

Dependency Mapping

Data products often have complex dependencies; that makes coordinating the inventory extremely important to a successful migration. Core data sets need to be identified, along with subsequent data model dependencies. Without an accurate dependency map, coordinating technology and migration timelines is impossible.

To help understand and codify these dependencies, customers should look to a well-established data catalog — one A that has lineage, discoverability, security, and governance as core capabilities. With this information, users will have self-service onboarding capabilities and the ability to request access to appropriate data sets; but without a proper catalog, migration to AWS can be extremely complex, and a massive burden to plan.

Technology and Pattern Mapping

The key to a successful migration is to develop templated patterns of the existing data applications to applicable AWS patterns that meet the stated goals of the migration. The patterns help accelerate the migration process by applying strict governance around the definition of the end-state data platform. Without patterns and governance, teams will be reinventing the wheel for each data application, and it will be hard to provide security and auditing in an automated, repeatable way.

Patterns

In AWS, a pattern is templated infrastructure-as-code — information architecture, automation, security, and governance. At Agilisium, we leverage cloud APIs and infrastructure-as-code services such as CloudFormation to define patterns as stacks that represent a data product. For example, an ingestion pattern that uses AWS DB Migration Service would be a CloudFormation template with appropriate parameters. The CloudFormation would templatize the creation of the following resources: RDBMS to S3 ingestion pattern, including the S3 bucket, S3 policy, IAM policy, DB Migration service, Cloudwatch configuration, etc.

After patterns are established, the existing inventory is mapped to the patterns so that it's clear what the endstate solution will look like. An accurate cost projection can be developed and measured against the stated goals of the migration. If these fail to meet the goals, patterns can be reworked and reevaluated until there is agreement with the overall architecture and implementation plan. Agilisium can help accelerate pattern creation and design, drawing from our experience with previous migrations.

Implementation

After taking a thorough inventory — along with dependency and technology pattern mapping — the project moves to implementation. Depending on the business objectives, inventory will be mapped to migration timelines, and the execution phase will begin, starting with the most critical data sets, as determined by the dependency mapping, and proceeding through the inventory. Implementations must have clear outcomes and validations. A framework is needed to audit data quality, timing, and accuracy of reports. The validations should be referenced against the current systems to ensure the migration is delivering clear and verifiable results.

Implementation Best Practices

Agilisium is committed to software engineering principles and best practices across all phases of the implementation, including:

- Proper source code, builds, dependency management, and artifact management
- Commitment to CI/CD, with automated deployments, releases, and configuration management
- Thorough unit and integration testing
- Centralized logging and metrics instrumentation to identify and resolve issues faster
- Dashboards with overall operational health

Without committing to these sound principles, migration projects often lose focus — resulting in unmanageable pipelines that lack the rigor required to scale properly.

Ongoing Managed Services

Many end-state AWS services reduce the complexity and operational ownership required to manage the solution; however, it's still important to have a dedicated, well-trained operational support team to look after the health and well being of your data pipelines — ideally one that can provide 24x7 monitoring and incident response to troubleshoot problems, respond to unforeseen business needs, and triage data quality issues.

Key Operational Responsibilities Include

- Deployment of patterns to support new use cases
- Alerting and monitoring of current pipelines
- Helpdesk for BI and Transformation users to troubleshoot data quality and warehouse issues
- Data quality monitoring
- Data pipeline monitoring and deployment
- Service upgrades (e.g. EMR, Redshift, RDS)
- Security audit and provisioning
- Data catalog support and integrations
- New user requests
- 3rd-party tool support

Case Study

Hortonworks to EMR: A migration enabling Big Data analytics at pace

Overview

The Media and Entertainment domain (M&E) is crowded with more players than ever before. Increased access to high speed internet has seen OTT streaming services emerging as a strong contender to traditional media studios. In this war for eyeballs, actionable business insights derived from processing big data is the most crucial weapon of all. Hence, big data processing technologies like Hadoop play an indispensable part in the technology stacks of M&E enterprises like our client – an entertainment conglomerate founded in 1923 and headquartered in Burbank, California, USA.

A customer insights application powered by Hortonworks was built for the client's Data Science team. For over a year, the application was managed by Agilisium, while the client's Data Engineering team supported and served requests to develop new queries for analysis.

This application was starting to show latency. Agilisium's ongoing relationship with the client and data analytics expertise enabled it to proactively recommend one significant change that accelerated client's data-to-insight journey time from days (on Hortonworks) to hours.

The Challenge

The primary change recommended by Agilisium's experts was that the client migrate from their existing Hortonworks data platform to a cloud service/product that better suited their current needs. Hortonworks was failing to meet the client's needs due to three main reasons,

- The average utilization peaked around 30% in spite of Data Science and Data engineering teams running data processing jobs on Hortonworks. However, it peaked at 80% – 90% during month-end activities. This meant that the client invariably paid peak utilization fees.
- There was a high volume of requests from both Data science and Engineering teams to the DevOps team. The service requests were operational in nature, hence the DevOps TurnAround Time (TAT) was at minimum a few days and could even go up to a few weeks.
- This led to increased overhead costs as two full-time resources were required to spin up clusters, monitor jobs and ensure that all teams followed DevOps processes while pushing jobs onto Hortonworks.

Consequently, it was a challenge for the client to realize their ROI from Hortonworks. In addition, Agilisium predicted that the client's data processing needs would only increase in number and complexity over the coming quarters. When the findings were presented, the client appreciated Agilisium's proactive recommendations and requested that we offer them a best fit solution.

Our Solution

The team of experts from Agilisium tackled the challenge presented to them systematically. Firstly, they built comparative POCs for the client's use case using three hand-picked technologies - Databricks, AWS EMR & Qubole – all in just under 12 weeks. Secondly, the POCs were presented to the client and they chose EMR for its - elastic auto-scaling compute power, pay as you go pricing, lack of licensing fee and an easy to use management console which also addressed the issue of the high overhead.

Thirdly, although on paper the migration was simple as Hortonworks and EMR are built on the same two open-source technologies - Hadoop & Spark – in reality, the two platforms are significantly different. Therefore, each data processing job on Hortonworks had to be carefully refactored for EMR. The client had stipulated that the migration would be executed with their inhouse resources working closely with Agilisium's team. The client's team worked closely with Agilisium and leveraged their expertise in both products to the hilt, easing their migration process.

Finally, while the client's team handled the migration of data processing jobs, Agilisium also worked on ensuring that EMR's connection to the rest of the architecture was stable and that it was user friendly. This involved,

- Simplifying the complex DevOps & CI/CD processes developed originally for Hortonworks to match EMR's features.
- Moving data storage from HDFS to S3, dramatically bringing down costs.
- Usage of tools like Jenkins, Ansible, Terraform and Airflow to automate the jobs flowing through the EMR platform, instead of complicating the architecture with point-to-point integration.

At the end of a 12-week migration effort, all the client's data processing jobs were migrated to AWS EMR from Hortonworks leading to the client gaining big data processing capabilities at pace.

Results and Benefits

- The Total Cost of Ownership (TCO) significantly decreased due to usage of open source technology and EMR's lack of licensing fee and pay-as-you-go pricing.
- On average, DevOps Team TAT reduced between 70 – 80% going from days to hours. Subsequently, Data Scientists could do continuous analysis on business data to obtain customer insights.
- Time-to-Market for application deployment reduced significantly by about 90% without loss in performance.

AWS Funded Exclusive Agilisium Consulting Offers

2 Days

Hadoop Migration Workshop

Understand EMR features and how to migrate Hadoop to AWS through a use case

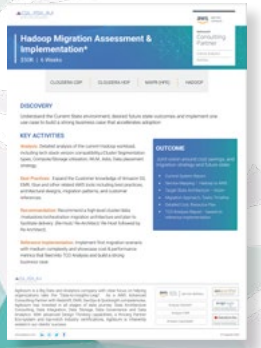
Download

3 Weeks

Detailed Assessment

Understand the State of your current data platform environment and provide a detailed approach for the desired future state

Download



6 Weeks

Assessment + Reference Implementation

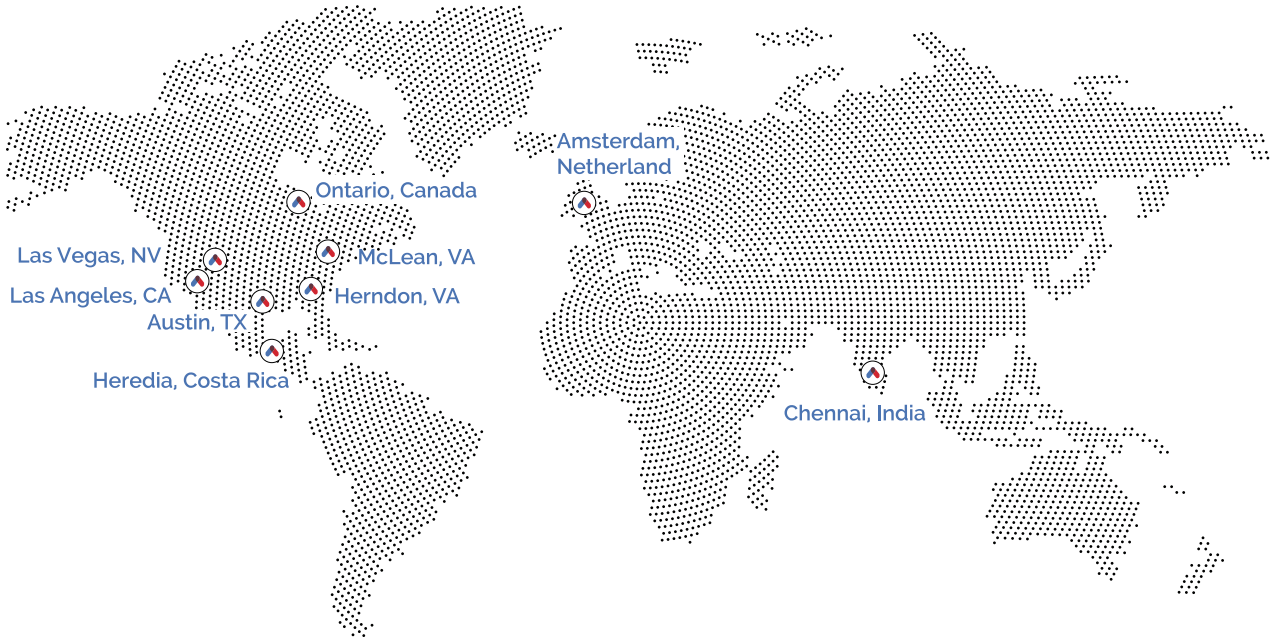
Detailed Assessment and implement one use case to build a strong business case that accelerates adoption

Download

Meet Agilisium

Agilisium is the fastest-growing Cloud Transformation & Data Analytics company with strong expertise in Data lake solutions, Data Warehouse Engineering, Data Migration & Modernization, Data Visualization, and Cloud Optimization services. Agilisium is an AWS Advanced Consulting Partner who helps companies architect, build, migrate, and manage their application workloads to accelerate their journey to the agile cloud, achieve desired business outcomes, and reach new emerging global markets.

The Cloud, Data Lake and Analytics Company with clear focus on helping organizations accelerate their "Data-to-Insights-Leap"



Global Captive Speciality Services firm offering services around Strategy, Implementation and Managed Services in AWS, Cloud, Data, Analytics, DevOps, Full stack development.

Competencies

aws PARTNER Advanced Tier Services

- Data & Analytics Services Competncy
- Migration & Modernization Services Competency
- DevOps Services Competency
- Well-Architected Partner Program
- Immersion Day

SDPs

aws PARTNER Advanced Tier Services

- Amazon EMR Delivery
- Amazon RDS Delivery
- AWS Lambda Delivery
- Amazon Kinesis Delivery
- Amazon DynamoDB Delivery
- Amazon Redshift Delivery
- Amazon Quicksight Delivery
- Amazon EC2 for Windows Server Delivery

AGILISIUM

BIG ON CLOUD. BIG ON DATA.

Contact

sales@agilisium.com

+(818) 241-4053

Corporate Head Quarters

United States | California

2629 Townsgate Road Suite 235
Westlake Village, CA 91361
Phone: +(818) 241-4053
Fax: +(818) 649-3321

Offshore Delivery

India | Chennai

World Trade Center,, 1st floor of Tower B ,Unit 102,,
Perungudi Real Estates Private Limited, 5/142,Rajiv
Gandhi Salai, Old Mahabalipuram Road,
Perungudi, Chennai-600096

USA | Canada | India | Netherlands | Costa Rica