White paper

Unlock Actionable Insights with a Modern Data Strategy on AWS

rackspace technology

Today, big data is bigger than many organizations realize — doubling in size every two years. Its sheer volume requires systems capable of both handling it and extracting actionable meaning from it. The more data an organization collects and accurately analyzes, the better it can understand its business, operations and customers. Used well, big data can reveal patterns, trends and associations — and fuel smarter and more profitable innovation across the enterprise.

But there is a downside. Many companies have collected vast quantities of data, but cannot make the most of it by extracting actionable business insights. The problem is a range of challenges, including data's sheer volume, disparate sources, unreliable pipelines and a lack of knowledge about the data.

How can organizations gain control of their big data and use it to fuel innovation? To start, they can deploy a cloud-based data platform equipped with modern data storage, components and services.

This white paper provides a guide to design patterns for a modern data platform and cloud technologies available from Amazon Web Services (AWS) that allow organizations to extract the most from their data. It overviews the anatomy of a platform that can handle big data — including the three critical functions: storage, processing and visualization. It also describes a next-step strategy to launch a data platform and leverage the power of big data for enterprise-wide innovation.

rackspace technology.

Part 1 — What is the state of big data today?

With an ever-increasing amount of business transaction data being stored and data-intensive mobile and Internet of Things (IoT) devices being deployed, big data is getting bigger than ever. Data is being created and collected at an astounding rate. The growth rate for big data was already increasing before the advent of mobile and IoT devices. Today, we are seeing customers overwhelmed by the amount of data they are collecting and unable to use it to advance their business.

Here are just a few stats about the explosion of big data:

- In 2020, 175 zettabytes of data is projected to be created (IDC)
- IoT devices are expected to generate 79.4ZB of data in 2025 (IDC)
- 53% of companies are deploying big data analytics (Dresner Advisory Services)
- 95% of businesses face a need to manage unstructured data (SharesPost)

How the "six Vs" are driving big data growth

Big data is the term used to describe large and diverse sets of information that continue to grow at ever-increasing rates. Big data management is a discipline that systematically extracts value from data sets that are too large or too complex to be managed by traditional data processing software.

The complexity of big data is driven by these six factors:

- **Volume:** the size of the data pool
- **Velocity:** the speed at which data is created and collected
- **Variety:** disparate sources of data
- **Value:** data's ability to deliver actionable business insights
- **Veracity:** data's truthfulness, taking into account biases like noise or inconsistencies
- **Volatility:** the relevance of data over time

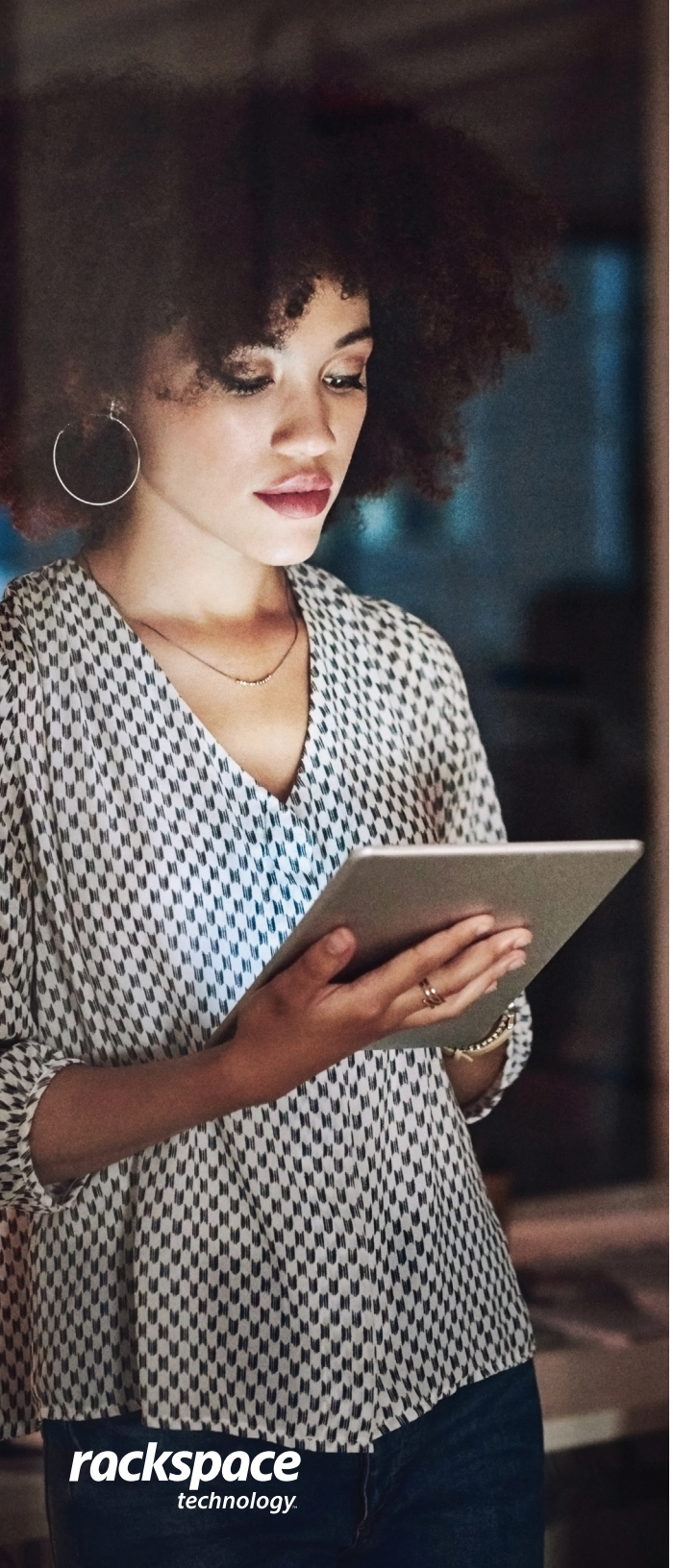
These factors must all be considered when enterprises embark on projects to build platforms to manage their big data.

Are you hamstrung by big data challenges?

Generating and collecting data is not the issue at most organizations. We've mastered that. However, there's a catch: the wealth of data only offers the potential to be useful. Simply collecting data is not enough. It must be stored, accessed, managed, analyzed and utilized. Otherwise, it delivers zero value.

While big data offers the potential to gain business-changing insights and support innovation, extracting this value means overcoming several challenges, including:

- **Overly complicated system design:** As complexity increases, systems must become more flexible, secure and able to deliver value in real time.
- **Users demand more:** More entities are making more demands on data access, including data analysts, scientists, business users and artificial intelligence (AI) systems.



- **Systems can't scale:** On average, data platforms need a life expectancy of approximately 15 years, which means they will need to scale approximately 1,000 times to meet future demands.
- **Systems aren't secure:** As the volume of data expands, so do data security challenges.
- **Data is siloed:** As companies have grown, each department usually has it's own data infrastructure, but data scientists and analysts want access to all of the data.

With a right big data strategy in place — using components that are up to the task — organizations can unlock the barriers to data-driven insights that inform meaningful innovation.

Are you leveraging your big data for big opportunities?

Organizations are leveraging big data for multiple purposes — all of which impact their bottom lines. On average, big data initiatives are increasing businesses profits by approximately 10% and reducing costs by approximately 10%. When organizations fail to properly leverage big data, they miss opportunities to improve their businesses.

Organizations that make the most of big data are those who rely on it in these key areas of business:

- **Product development:** Big data delivers a boost to product development by classifying key attributes of past and current products so systems can model relationships and use the source data to fuel product recommendation engines.
- **Predictive maintenance:** Big data can enable organizations to anticipate events, such as heavy equipment breakdowns, by analyzing both structured data (equipment year, make, model, etc.) and multi-structured data (log entries, sensor data, error messages, engine temperature, etc.), so they can take action before failures occurs.
- **Fraud and compliance:** Big data allows firms to spot anomalous usage and suspicious transaction patterns in real time, so they can take action and curtail long-term abuses.
- **Operational efficiencies:** Big data enables live operations monitoring, which provides real-time insight for taking actions that can impact everything from retail sales to medical outcomes.
- **Machine learning and predictive modeling:** Big data gives organizations the ability to conduct a wide range of analysis, such as revenue forecasting based on advertising or predicting the impact of ad campaigns based on past performance.

Part 2 — Modern big data processing begins with an old-school problem: storage

Data storage is an aspect of IT that we thought we had under control. In fact, many consider it a problem of the past, back in the older, non-innovative days. But data storage is a critical piece of the big data challenge. Without the right storage environment, gaining the maximum value from big data is impossible.

The fact is that big data is heavy. As a result, moving it between environments is a challenge, such as transferring data from an on-premises data center to a cloud environment. Enter cloud-based storage options — data lakes, data warehouses and data marts, which deliver innovative approaches to storing big data for optimal usage. Here's how they work.

Data lakes — a centralized, curated and secure repository. Data lakes store all data, both in its original form and after it's been prepared for analysis. A data lake enables organizations to break down data silos and combine different types of analytics to gain business insights and guide innovation. Amazon S3 is typically chosen to support data lake implementations due to its low cost, high availability, flexibility for storing many different types of data and integration with many supporting AWS services. This works well for talking to different data warehouses, such as Amazon Redshift, AWS's native data warehouse service. Key benefits of data lakes include robustness to work with data without moving it, flexibility for both relational and non-relational data, scalability to exabytes of data and cost effectiveness.

Data warehouses — collect and analyze volumes of data. Data warehouses contain large quantities of data from many different sources and enable fast and complex queries for data analysis and reporting. Amazon Redshift is an example of a data warehouse. Key benefits of Amazon Redshift include efficient storage in a columnar storage paradigm versus row-based storage, complete management and maintenance by AWS, ability to handle exabyte-level quantities of data at scale, fast and complex queries of any data, integration with Amazon S3 Data Lake, and added support for AWS IAM for access control and AWS KMS for encryption.

Data marts — a subset of data warehouses. Data marts are typically smaller than data warehouses and focus on a single domain, for example, financial reporting. This simplicity provides nimble and rapid access to specific datasets.

Data lake vs. data swamp — how to avoid the risks

Failure is an option when it comes to big data storage. If data storage is mismanaged, it can quickly create a "data swamp." The result can be a hodgepodge of data that is not well organized or adequately protected. To avoid this outcome, organizations need to answer these questions on the front end:

1. **Know your data:** What types of data is being collected? How fast is it being generated? How do the data systems relate to each other?
2. **Know your users:** How are they going to access the data? Are they going to try and make a bunch of temporary tables?
3. **Know what analytics will be used with the data:** How does the data need to be partitioned to efficiently support the use cases?

Part 3 — Working with the power tools designed to build modern data platforms

Building a robust modern data platform requires the support of a wide range of power tools. Amazon Web Services provides everything organizations need, including AWS Glue, Amazon EMR, Amazon Redshift and AWS Lake Formation. Each component brings advantages to the creation and deployment of a modern data platform.

AWS Glue and Amazon EMR

AWS Glue is a fully managed extract, transform and load (ETL) service for big data processing. It makes it easy for organizations to prepare and load their data for analytic functions. Once activated, AWS Glue discovers the data and stores the associated metadata (table definition, schema, etc.) in the AWS Glue Data Catalog. Once cataloged, data is immediately searchable, queryable and available for ETL.

Amazon Elastic MapReduce (EMR) is a processing framework that allows developers to write programs that process massive amounts of structured and unstructured data in parallel across a distributed cluster of processors or stand-alone computers using open-source tools. It delivers these advantages:

- **Elasticity:** Unlike the rigid infrastructure of on-premises clusters, Amazon EMR decouples compute and storage to provide the ability to scale each independently and take advantage of tiered storage of Amazon S3. Organizations can provision one, hundreds or thousands of compute instances to process data at any scale. The number of instances can be increased or decreased automatically using Auto Scaling (which manages cluster sizes based on utilization).
- **Cost savings:** Organizations can run petabyte-scale analysis at less than half of the cost of traditional on-premises solutions.
- **Speed:** Systems can run analysis much faster than standard processing.
- **Security:** Amazon EMR automatically configures Amazon EC2 firewall settings to control network access to instances and launches clusters in an Amazon Virtual Private Cloud (VPC). This supports encryption at rest and in transit using the AWS Key Management Service. Also, users can employ AWS Lake Formation to apply fine-grained data access control for data storage locations, databases, tables and columns.

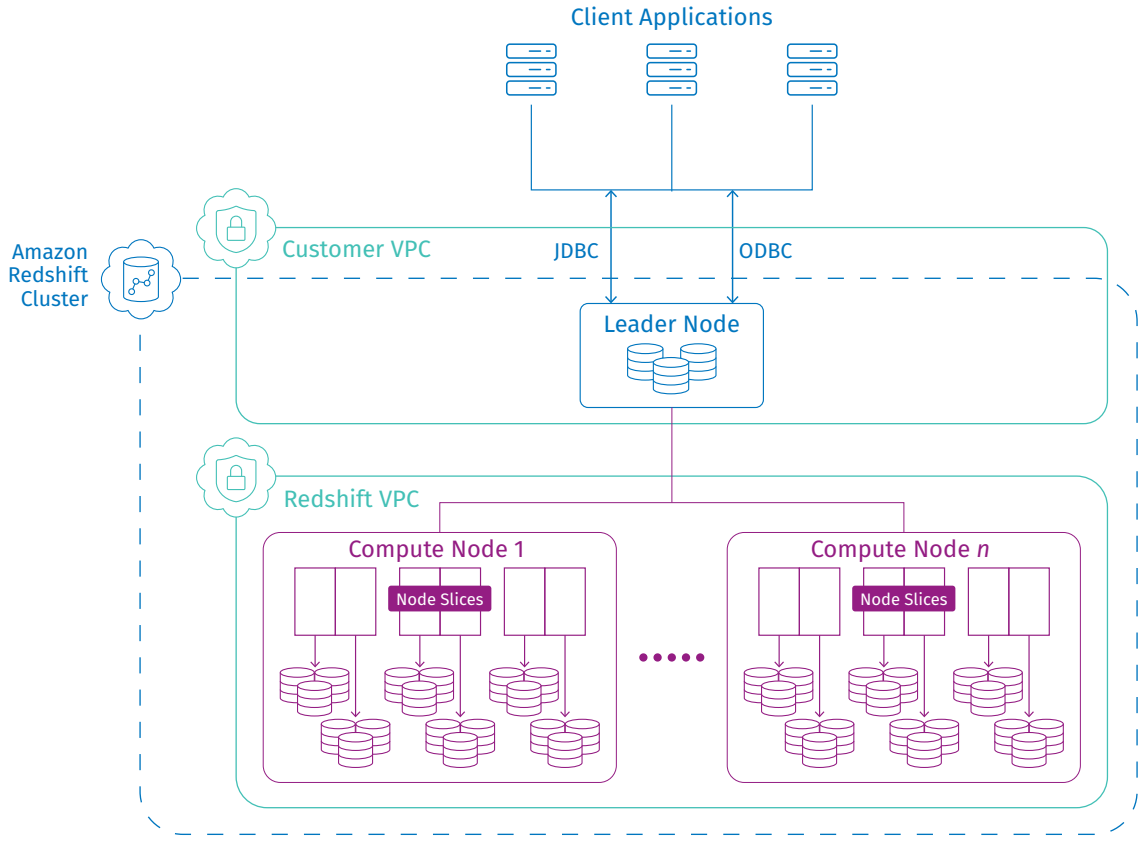
Amazon Redshift — a cloud-native data warehouse

Amazon Redshift is a column-oriented, fully managed, petabyte-scale, cloud-native data warehouse that makes analyzing data using existing business intelligence tools simple and cost-effective. Amazon Redshift achieves efficient storage and optimum query performance through a combination of massive parallel processing, columnar data storage, and efficient and targeted data compression encoding schemes.

With Amazon Redshift, an organization can query petabytes of structured and semi-structured data across its data lake, data warehouse and operational database using standard SQL. Results of queries can be saved in an S3 Data Lake using open formats, like Apache Parquet, to gain additional analysis using other analytics services, like Amazon EMR. Also, the custom database engine in Redshift is based on PostgreSQL, supports OLAP operations, and is ACID and ANSI SQL compliant.

Amazon Redshift cluster architecture. Amazon Redshift can be deployed as a single node or as a multi-node cluster that can scale as organizations build out their data warehouses. The core infrastructure component of an Amazon Redshift data warehouse is a node. A cluster is composed of one or more compute nodes. If a cluster is provisioned with two or more compute nodes, an additional leader node coordinates the compute nodes and handles external communications. Advanced filtering and join logic reside within the cluster to maximize aggregation and join processing performance.

There are several design considerations that went into Amazon Redshift for maximum performance and scalability. For example, unlike many scalable AWS services, Amazon Redshift clusters only reside in a single Availability Zone (AZ) for maximum network throughput. Multi-node Amazon Redshift clusters span two separate VPCs. The leader node is in a customer-managed VPC, and exposes endpoints for ODBC and JDBC endpoints to integrate with analytics tools. Compute node resources are created in an AWS-managed Redshift VPC, and are responsible with computation to support joins and data aggregations.



Amazon Redshift columnar storage. Amazon Redshift uses a columnar storage format for fast data retrieval, typically in analytical applications, unlike a relational database, which is optimized for storing rows of data, typically for transactional applications. Column-oriented storage for database tables is an important factor in analytic query performance, because it reduces the overall disk I/O requirements and the amount of data that organizations need to load from disks.

Amazon Redshift Spectrum. This is a feature within Amazon Redshift that lets data analysts conduct fast and complex analysis on objects stored external to the data warehouse on Amazon S3. Amazon Redshift Spectrum's key functionality offers a "lake house" design pattern in which the data warehouse can be brought to the data lake. This allows querying of potentially exabytes of data sitting on Amazon S3, and query and retrieve data from Amazon S3 without explicitly loading the data into the cluster. Powered by a separate fleet of Redshift Spectrum nodes, its highly parallelized execution can query hundreds of petabytes in minutes. It delivers high performance through multiple compression types, bloom filters for collocated and broadcast join queries, and cache-optimized aggregation and join processing. Beginning at just $5 per TB of data scanned, it delivers cost-effective access to an extremely large amount of data.

Amazon Redshift security and compliance

AWS is dedicated to the highest standards of security and compliance for its products and services. Amazon Redshift security and compliance functionality includes:

- End-to-end data encryption using AWS KMS
- AWS IAM integration
- SAML IdP for access federation with AD, Okta or Ping Identity
- Network isolation with Amazon VPCs
- Database security model
- Audit logging and notification
- Column-level security for internal and external tables
- SOC 1/2/3, PCI DSS, FedRAMP and HIPAA certifications

Amazon Redshift continuous evolution

AWS is dedicated to the continued evolution of its products and services. For Amazon Redshift this means:

- New RA3 instance type allowing independent scaling of compute resources and storage
- Managed storage feature
- Maximum cluster storage of 8.2PB
- Advanced Query Accelerator (AQUA)
- AWS-designed analytics processors optimized for compression, encryption, filtering and aggregations
- Federated queries
- Spatial data support
- Distribution and sort key advisor
- Elastic resize

Amazon Redshift Lake House architecture

Amazon Redshift powers the lake house architecture, enabling analysts to query data across data warehouses, data lakes and operational databases to generate faster and deeper insights, which would otherwise not be possible. With a lake house architecture, organizations can store data in open file formats in their Amazon S3 data lake, allowing them to make the data available to other analytics and machine-learning tools, versus locking it into a new silo.

Lake house architecture functionality streamlines data analysis in multiple ways. For example, it conducts queries in the data lake and writes back to the data lake in open formats. It uses familiar SQL statements to combine and process data across all data stores. It also executes queries on live data in operational databases without any data loading and ETL pipelines.

Anatomy of lake house architecture

Here's how all of the parts of a highly efficient data lake house work together.

- **Data sources:** Includes applications, databases, streaming services, third parties and IoT devices.
- **Ingestion layer:** Acts as a buffer to take data from a raw stage to a data lake. These are typically handled by services like Amazon Kinesis, Amazon EMR or AWS Lambda.
- **Monitoring and logging:** Helps provide a smooth ongoing flow of data within the system through monitoring, logging, altering, sending notifications and creating audit logs.

- **Data lake:** Resides at the heart of the architecture. Data lakes support data movement throughout all of the stages of a data platform by providing zones for managing the logical or physical separation of data types. The five basic zones include:
  - **Raw zone:** Where data is stored in a raw state, providing flexibility for future changes, and backups for reference. Essentially, this elevates data processing from an ETL model to an ELT model. ETL allows one opportunity to transform data, whereas ELT provides ongoing access to the raw data.
  - **Structured zone:** Applies processing to make data more uniform or create additional metadata, for example, transcribed video.
  - **Curated zone:** Combines data from a variety of sources, such as calculations, quality checks (NULLS, formats, etc.) and referential integrity.
  - **Consumer zone:** Provides a friendly format for consumers, such as a data mart zone.
  - **Analytics zone:** Populated from all of other zones, this is the domain of the data scientists, where they analyze data.
- **Security functionality:** Implemented using policies and roles, and includes AWS KMS keys to encrypt the data through AWS services.
- **Data warehouse:** Once data moves into the data lake, it will be loaded into the data warehouse, such as (typically) Amazon Redshift.
- **Data catalog:** The catalog in an index to the locations, schema and runtime metrics of the data. AWS Glue Data Catalog is the AWS-native solution to cataloging data. Once data is indexed in the catalog it may be processed using Amazon EMR or AWS Glue jobs, loaded into a data warehouse, like Amazon Redshift, or queried using Amazon Athena. Typically, the data catalog is populated by crawling the data to take inventory of a particular Amazon S3 location.
- **Data consumers:** These are application consumers, which may be data scientists, analysts and reporting personnel within an organization.

Build a secure data lake with AWS Lake Formation

Building data lakes can easily take months when using manual, complicated and time-consuming processes. AWS has introduced a new service called AWS Lake Formation to streamline the nuts and bolts of merging processes and security requirements together. It allows an organization to build a secure data lake in days or weeks with AWS Lake Formation depending upon security, monitoring and other customer requirements.

AWS Lake Formation functionality builds on the capabilities available in AWS Glue, making each step much easier to execute. Among them, it registers Amazon S3 buckets and paths where data resides, and orchestrates data flows that ingest, cleanse, transform and organize the raw data. It also creates and manages a data catalog containing metadata that resides in the data lake and data sources. Further, it defines granular data access policies through a grant-and-revoke permissions model.

Building a data lake manually versus with AWS Lake Formation

Creating data lakes manually — without AWS Lake Formation — requires organizations to execute these tasks on their own:

- Identify and load data from diverse sources
- Monitor data flows
- Set up partitions
- Turn on encryption and managing keys
- Define transformation jobs
- Monitor job operations
- Re-organize data into a columnar format
- Configure access control settings
- Deduplicate redundant data
- Match linked records
- Grant access to data sets
- Audit access over time

Creating a data lake with AWS Lake Formation delivers these automated advantages:

- Collects and catalogs data from databases and object storage
- Moves the data into new Amazon S3 data lake
- Cleans and classifies data using machine-learning algorithms
- Secures access to sensitive data
- Allows users to access a centralized data catalog
- Allows users to leverage data sets with their choice of analytics and machine-learning services

AWS Lake Formation blueprints

AWS Lake Formation accomplishes its functionality through blueprints built on AWS Glue. A blueprint is a data management template that enables easy ingestion of data into a data lake. AWS Lake Formation provides several blueprints, each for a predefined source type, such as a relational database or AWS CloudTrail logs.

AWS Lake Formation blueprints set projects up for success with its robust functionality. It combines catalog tables, jobs and AWS Glue crawlers to orchestrate loading and updating data. It handles the heavy lifting and nuts and bolts of configuring glue jobs to examine and process data. Blueprints also allow data into the data lake without re-inventing the wheel. While handling bulk or incremental loads from existing relational database management systems (RDBMS) and log ingest, it configures source and target parameters, frequency and more. It delivers unparalleled security by monitoring for errors or issues within the data pipelines, and securing once and allowing access in multiple ways.

Part 4 — From a secure edge to a robust core, AWS services think of everything

AWS has developed a wide range of services to meet any level of big data development and engineering — from customized support to fully managed solutions with robust machine-learning capabilities. They are elastic, pay-as-you-go, managed services. AWS service categories include:

- Analytics services, such as Amazon EMR and Amazon QuickSight
- Compute resources for load or reporting operations
- Compute and storage services that can scale independently
- Advisory services to achieve performance optimization

A hallmark of AWS services is their high level of security and compliance. When it comes to security, AWS thought of everything — from a highly secure edge to data encryption using AWS KMS. Columnar-level security ensures that data can easily be restricted to specific user groups.

IAM integration adds data protection and isolation between different user groups. Integration with SAML-based identity provides for Access Federation to manage users with Active Directory, Okta or Ping Identity. Network isolation with AWS using a VPC technology ensures clusters are secure from access by outside parties.

Additional security functionality includes:

- Database security to manage users in a traditional database fashion
- Audit logging and notification for monitoring
- Certifications, including SOC- 1/2/3, PCI DSS, FedRAMP

Data engineering provides a foundation for value and insight

Data engineering is the foundation that supports data science, collection and analytics. Before organizations can deploy data analytics or machine-learning operations, their data must be ingested, stored, transformed, cataloged and managed. Engineering is a combination of design, development and management to enable users to perform these functions and gain valuable and actionable insights from their data.

Part 5 — Best practices and next steps for launching a winning modern data strategy

Many organizations, large and small, around the world are realizing the benefits of AWS to build their big data platforms — and support their innovation. Here are just three examples.

A global product company needed to build a data lake that could combine data from multiple sources, including third parties, mobile analytics platforms and internal binary data sources. The company wanted a single access point for all of its users — marketing, advertising, data science and DSP teams — so they could access all of the mobile analytics and binary data to develop and promote new products and improve existing products. This objective was accomplished using Athena data connectors that allow Athena to query metadata information that may be stored in Amazon DynamoDB and Amazon S3.

A financial services firm wanted to improve the performance of its fraud analytics team by transforming a process that took days or weeks to a real-time analytics capability. Employing DMS and AWS Glue to transition data from the existing on-premises data source into Amazon S3, the firm achieved a high-performance cost-effective query capability using Amazon Redshift Spectrum.

A global leader of wearable safety technology, gas monitoring and cloud-connected software and data analytics needed a high-volume data streaming and processing infrastructure to support an expanding customer base and growing solutions portfolio. It also needed to support the requirements for a contact tracing reporting solution designed to keep workers safe amid the COVID-19 pandemic. The company successfully migrated from its conventional data ingestion and database storage to a contemporary AWS infrastructure. The result is a workflow designed to ingest the higher data-rate raw messages, and enrich them to deliver high-value data in Amazon Redshift for reporting.

Leverage big data strategy best practices

Optimizing the development of a data platform on AWS begins with creating a solid strategy. These best practices are the foundation of an ideal big data strategy:

- **Assess the data strategy:** Analyze your current data platform and define a roadmap to a modern data platform on AWS.
- **Migrate data to AWS:** Determine the best set of services and practices to meet future production-grade data platform requirements on AWS.
- **Build an MVP data platform:** Create a minimal viable product for analytics and visualization that supports processing business intelligence insights.
- **Optimize with Amazon Redshift:** Assess and tune Redshift cluster performance for optimal performance.

Take six next steps to a rapid launch on AWS

Here is the basic six-step path to launching a modern cloud-based big data platform on AWS:

1. **Set up storage:** Provision services and establish access policies.
2. **Configure security policies:** Refine and enforce security and compliance policies.
3. **Move data:** Identify data sources and move data from sources to data lake.
4. **Prep data:** Clean, prep and catalog all data.
5. **Deploy data:** Embed analytics functionality.
6. **Repeat:** Continue following these steps for all datasets, users and end services.

After following these steps your organization will be well positioned to effectively leverage your big data to extract the maximum actionable insight — so you can accelerate innovation across your enterprise.

Contact Rackspace Technology

Whether your organization is just getting started with your modern data platform, is ready to conduct a big data strategy assessment or aren't sure of your next steps, Rackspace Technology experts can analyze your needs and create a roadmap for deploying a modern data platform on AWS across your enterprise.

Learn more at **www.rackspace.com/en-gb** or call **+44 203 553 6268**.

rackspace technology.

About Rackspace Technology

Rackspace Technology is the multicloud solutions expert. We combine our expertise with the world's leading technologies — across applications, data and security — to deliver end-to-end solutions. We have a proven record of advising customers based on their business challenges, designing solutions that scale, building and managing those solutions, and optimising returns into the future.

As a global, multicloud technology services pioneer, we deliver innovative cloud capabilities to help customers build new revenue streams, increase efficiency and create incredible experiences. Recognised as a best place to work, year after year, by Fortune, Forbes, Great Places to Work and Glassdoor, we attract and develop world-class talent to deliver the best expertise to our customers. Everything we do is underpinned by an obsession with our customers' success — our Fanatical Experience™ — so they can work faster, smarter and stay ahead of what's next.

Learn more at **www.rackspace.com/en-gb** or call **+44 203 553 6268**.

TSK_2676_WP_Data_Unlocking_Actionable_Insights :: October 19, 2020