

Modern Data Warehousing on

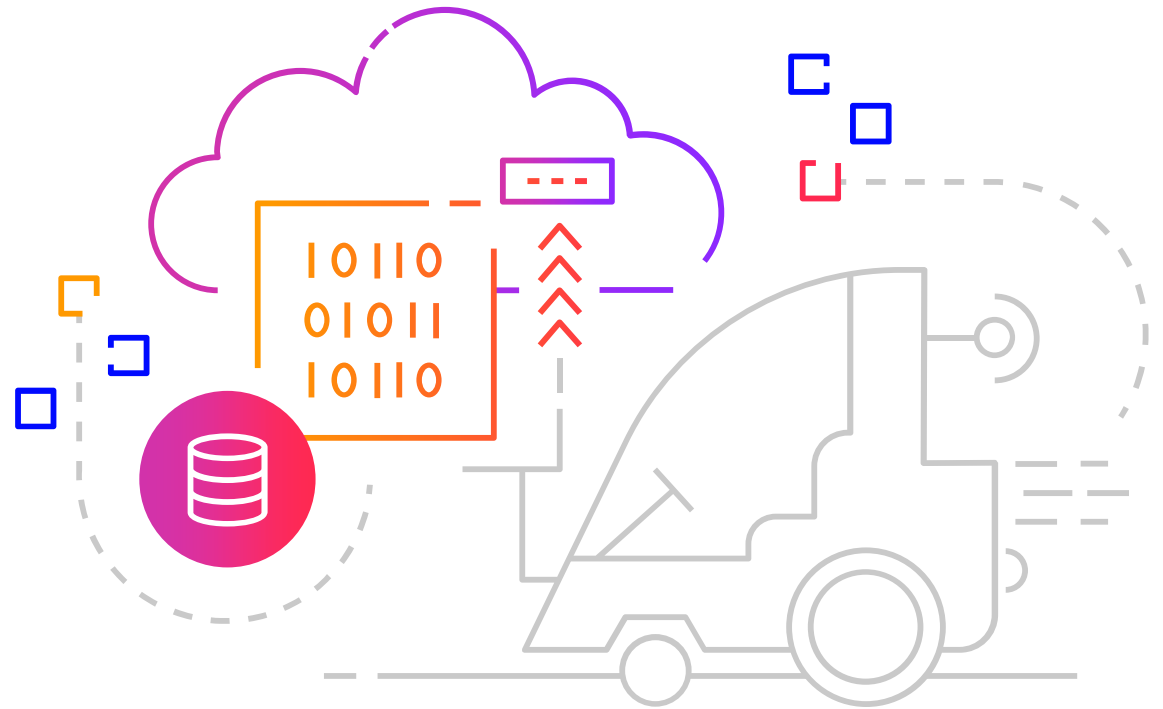


Table of Contents

1

Square Peg in
a Round Hole:

The Problem with
Traditional Data
Warehouses

2

What is a
Modern Data
Warehouse

3

How is a
Modern Data
Warehouse
Different

4

Introducing
Amazon
Redshift

5

The benefits
of data
warehousing
on AWS over
traditional
Data
warehouses

6

Amazon
Redshift
Use Cases

7

Migrating
to Amazon
Redshift from
On-Premises
Data Warehouses

Self-service Migration
Using Database
Migration Service

Migrating with AWS
Partner Network
Systems Integrator
Partners

8

Getting
Started with
Amazon
Redshift

9

Case Study:

Equinox Fitness
Clubs Realizes Faster
Time-To-Benefit
and Increased
Productivity with AWS

1

Square Peg in a Round Hole:

The Problem with Traditional Data Warehouses

Square Peg in a Round Hole:

The Problem with Traditional Data Warehouses

Most large enterprises today use data warehouses for reporting and analytics purposes using the data from a variety of sources, including their own transaction-processing systems and other databases. Data and analytics have become an indispensable part of gaining and keeping a competitive edge. But many legacy data warehouses underserve their users, are unable to scale elastically, don't support all data, require costly maintenance, and don't support evolving business needs. These challenges often force organizations to only use a fraction of their data for analysis. We call this the "dark data" problem: companies know there is value in the data they have collected, but their existing data warehouse is too complex, too slow, and just too expensive to analyze all their data. The rigid traditional data warehouse architectures are further buckling under the strain of growing needs to have insights available at the fingertips of every employee even in the largest organizations, new use cases to enable machine learning applications and analyze data almost as soon as it is generated. The following factors cause traditional data warehouses to break down for modern analytics:

Expensive to set up and operate



Most data warehousing systems are complex to set up, cost millions of dollars in upfront software and hardware expenses, and can take months in planning, procurement, implementation, and deployment processes. Organizations need to make large investments to setup the data warehouse and hire a team of expensive database administrators to keep the data warehouse running queries fast and protect against data loss or system outage.

Difficult to scale



When data volumes grow or organizations need to make analytics and reports available to more users, they choose between accepting slow query performance or investing time and effort for an expensive upgrade process. In fact, some IT teams discourage augmenting data, adding users, or adding queries to protect existing service-level agreements. Finally, restrictive licensing models are not conducive to quickly scale the data warehouse up and down, locking organizations down into expensive long-term contracts with long periods of low utilization.

The "Dark Data" problem:

companies know there is value in the data they have collected, but their existing data warehouse is too complex, too slow, and just too expensive to analyze all their data.

Square Peg in a Round Hole:

The Problem with Traditional Data Warehouses

Unable to handle data variety



Increasing diversity in information sources such as devices, videos, IoT sensors, and more results in increasing volume & variety of unstructured data, which creates new challenges to derive useful insights. Traditional data warehouses, implemented using relational database systems, require this data to always be cleansed before it is loaded, and to comply with pre-defined schemas. This poses a huge hurdle for analysts and data scientists, who are otherwise skilled to analyze data directly in open formats. Organizations try to handle the volume and variety of these data sources and bolt-on data lake architecture to their data warehouse to store and process all data. But, they often end up with data silos and run into error prone and slow data sync between their data warehouses and data lake.

Insufficient for modern use cases



Traditional data warehouses are designed for batch processing Extract, Transform and Load (ETL) jobs. This prevents data to be analyzed as it arrives, and in the formats (like Parquet, ORC, and JSON) in which it is stored. Traditional data warehouses also don't support sophisticated machine learning or predictive workloads or support them in a limited fashion. Therefore, they're unable to support modern use cases such as real-time or predictive analytics and applications that need advance machine learning and personalized experiences.

Need additional investment for compliance and security

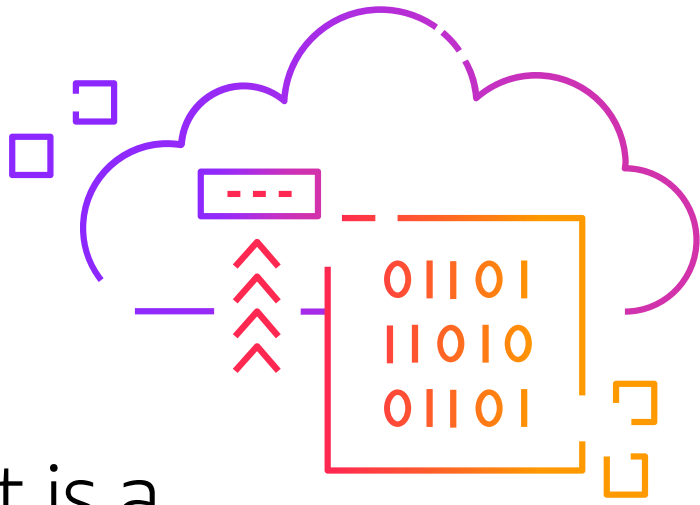


Industries such as healthcare and financial services that work with highly sensitive data require the data warehouse to be compliant with ISO, HIPAA, FedRAMP, and more. General Data Protection Rules (GDPR) further add to the burden on IT to ensure that sensitive customer data is encrypted in all states – in rest and in motion. Some of these regulations also require organizations to react quickly to retrieve and update or delete a record at short notice. Traditional data warehouses often require organizations to implement expensive work-around solutions, which often leave the sensitive data out of the analysts' reach.

Traditional data warehouses cannot query data directly from the data lake and from open formats such as Parquet, ORC and JSON

2

What is a Modern Data Warehouse



What is a Modern Cloud Data Warehouse

A modern cloud data warehouse is designed to support rapid data growth and interactive analytics over a variety of relational, non-relational, and streaming data types leveraging a single, easy-to-use interface with all the elasticity and cost efficiency cloud provides.



A modern data warehouse liberates you from the limitations of loading data into local disks before it can be analyzed. **This gives you virtually unlimited flexibility to analyze the data you need, in the format you need it, and using familiar BI tools and SQL**

3

How is a Modern
Data Warehouse Different

How is a Modern Data Warehouse Different

Capability	Modern Data Warehouse	Traditional Data Warehouse
Elasticity	Scale up for increasing analytical demand and scale down to save cost during lean periods – on-demand and automatically.	Hard limitations on growing or shrinking the storage and compute, slow to adopt, over provisioning for future demands, low capacity utilization.
Performance	Leverage easily scalable cloud resources for compute and storage, caching, columnar storage, scale-out MPP processing, and other techniques to reduce CPU, disk IOs, and network traffic.	Depend on fixed compute and storage resources, leading to performance ceiling and contention during data integration and loading activities.
Cost	Eliminate fixed costs of infrastructure and allow organizations to grow their data warehouse and evaluate new data sources with minimal impact on cost. High availability and built-in security eliminates cost of failure.	Need millions of dollars of up-front investment in hardware, software, and other infrastructure, and significant cost of maintaining, updating, and securing the data warehouse. Scaling may not result in economies of scale, as more fixed investment is required every time capacity maxes out.
Use cases	Work across a variety of modern analytical use cases because of versatility and seamless integration with other purpose built engines	Insufficient for predictive analytics, real-time streaming analytics and machine learning use cases.
Simplicity	Easy to deploy and easy to administer; avoid knobs and use smart defaults to offer better out-of-the-box performance than traditional data warehouses. Hardware upgrades happen seamlessly.	Rolling out a data warehouse is often slow and painful. On an ongoing basis, a lot of undifferentiated heavy lifting is involved in database administration, including in provisioning, patching, monitoring, repair, backup and restore.
Scale	Scale in and out automatically to absorb variability in analytical workloads and support thousands of concurrent users. Quickly scale up to accommodate high analytical demand and data volume and scale down when demand subsides to save cost. Scale out to exabytes of data without needing to load it into local disks.	Scaling up requires considerable time and investment. Additional hardware may need to be deployed. Under-utilized capacity leads to waste during periods of analytical demand. Scale out to large volumes of data also doesn't work as analysis is limited to data loaded into local disks only.
Rich ecosystem	Work with a wide variety of ETL, querying, reporting and BI, and advanced analytics tools.	May only work with proprietary tools or specific third-party vendors. Integrations may be difficult to implement.
Data ingestion	Take advantage of relational, non-relational, and streaming data sources.	Optimized for relational data sources.
Open data formats	Load and query data from, and unload data to, open formats, such as Parquet, ORC, JSON, and Avro. Run a query across heterogeneous sources of data.	May not support open data formats.

4

Introducing Amazon Redshift



Introducing Amazon Redshift

Amazon Redshift is a fast, scalable data warehouse that makes it simple and cost-effective to analyze all your data across your data warehouse and data lake. You can use the Business Intelligence (BI) tools you love and use familiar ANSI compliant SQL for even the most complex analytical workloads.

Redshift delivers ten times faster performance than other data warehouses by using machine learning, massively parallel query execution, and columnar storage on high-performance disk. You can setup and deploy a new data warehouse in minutes and run queries across petabytes of data in your Redshift data warehouse and exabytes of data in your data lake built on Amazon S3 with Redshift Spectrum. You can start small for just \$0.25 per hour and scale to \$250 per terabyte per year, less than one-tenth the cost of other solutions.

Redshift is
10 times faster than
other data warehouses.
You can setup and
deploy a new data
warehouse in minutes.

5

The Benefits of Data Warehousing on AWS over Traditional Data Warehouses

The Benefits of Data Warehousing on AWS over Traditional Data Warehouses

Amazon Redshift is the first MPP Data Warehouse solution architected for the cloud. Since its launch in 2013, Redshift has been one of fastest growing AWS services, with thousands of customers across industries and company sizes. Forrester ranked AWS, with the largest cloud DW deployments, as a leader in the [Forrester Wave: Cloud Data Warehouse, Q4 2018](#). Enterprises such as NTT DOCOMO, FINRA, Johnson & Johnson, Hearst, Amgen, and NASDAQ have migrated to Amazon Redshift and have leveraged a variety of AWS services such as S3, Kinesis, EMR, Glue, and QuickSight to build a modern data warehousing solution. These organizations have benefited from the following:

The perfect foundation for a modern analytics pipeline



As a fully managed data warehouse, Amazon Redshift allows you to automate most administrative tasks so you can focus on your data and your business. It delivers fast query performance, improves I/O efficiency, and scales up or down as your performance and capacity needs change. You can setup and deploy a new data warehouse in minutes. You can run queries across petabytes of data in your Redshift data warehouse and exabytes of data in your data lake built on Amazon S3. Most results come back in seconds. With Amazon Redshift, you can start small for just \$0.25 per hour with no commitments. You can scale out to petabytes of data for \$1,000 per terabyte per year, less than a tenth the cost of traditional solutions with Amazon.

Redshift has been one of fastest growing AWS services with the largest cloud DW deployments

Flexibility to query seamlessly across the data warehouse and data lake



Most organizations already have more data than they can load into a data warehouse, while data continues to grow at rates faster than it is analyzed. With your data warehouse on AWS, you can go beyond data stored on local disks in your data warehouse to query vast amounts of unstructured data in your Amazon S3 "data lake" - without having to load or transform any data. Amazon Redshift includes Spectrum, a feature that gives you the freedom to store your data where you want, in the format you want, and have it available for processing when you need it. Amazon Redshift Spectrum queries are just as easy to run as the queries you would perform against the data stored on local disks in Amazon Redshift, but without the need to ingest data into an Amazon Redshift cluster. No loading or transformation is required, and you can use open data formats, including CSV, TSV, Parquet, Sequence, and RCFile. Redshift Spectrum leverages the powerful Redshift Query Optimizer and automatically scales to thousands of nodes. Having your data available in open formats in your data lake ultimately gives you the flexibility to use the right engine for different analytical needs, such as Amazon Redshift for business analytics, Athena for Serverless Queries, and SageMaker for predictive analytics and machine learning.

The Benefits of Data Warehousing on AWS over Traditional Data Warehouses

A rich ecosystem for big data

AWS gives you fast access to flexible and cost-effective IT resources so you can rapidly build and scale virtually any big data solution, including data warehousing, clickstream analytics, fraud detection, recommendation engines, event-driven ETL, serverless computing, and Internet of Things (IoT). You can use Amazon QuickSight or other BI tools, such as Tableau, Looker, and Periscope to conduct exploration and visual analysis of any type of data and quickly arrive at impactful business insights.



Easy, fast, and flexible loading

You can load virtually any type of data into Amazon Redshift from a range of data sources including Amazon Simple Storage Service (Amazon S3), Amazon DynamoDB, Amazon EMR, Amazon Kinesis, and/or any SSH-enabled host on Amazon Elastic Compute Cloud (Amazon EC2) or on-premises instances. Additionally, you can leverage third party ETL tools, such as Matillion, Talend, and others to simplify and speed up data loading and transformation. Amazon Redshift loads your data in parallel into each compute node to maximize the data ingestion rate.

If you need to transfer petabytes or even exabytes of data to the cloud, you can use AWS Snowball or AWS Snowmobile respectively to move the data into Amazon S3. Once the data lands in S3 you can query it with Amazon Redshift Spectrum. You can load your most frequently accessed data into Redshift for faster query processing. This allows you to run queries against all your data, including unstructured data in Amazon S3 that is not loaded into Redshift, using Amazon Redshift and Redshift Spectrum.



Security and compliance

When you host your data warehouse on AWS, you benefit from the infrastructure built to meet the requirements of the most security-sensitive organizations. AWS actively manages dozens of compliance programs in its infrastructure, helping organizations like yours meet compliance standards such as PCI DSS, HIPAA, FedRAMP, and more. Amazon Redshift further keeps you secure and compliant - you can encrypt data at rest and in transit and be compliant with GDPR or industry-specific regulations. In addition, you can isolate your Amazon Redshift clusters with Amazon VPC, and manage your keys using AWS Key Management Service and Hardware Security Modules (HSMs).



Increased agility and reduced costs

Data warehousing on AWS offers the best price to performance ratio for your workload. Amazon Redshift is inexpensive – you can start at \$0.25 per hour and run your data warehouse for as low as \$1000 per TB per year. With up to 4x compression and the ability to analyze data without needing expensive loading processes, even Exabyte-scale data sizes are cost-effective to analyze so you'll never have to throw away valuable data. Amazon Redshift further brings down the cost of management- it alleviates capacity, provisioning, patching, monitoring, backups, and a host of other DBA headaches. On an ongoing basis, Amazon Redshift uses machine learning to automatically learn, provide recommendations for optimizing your cluster and performing automated maintenance tasks under the hood.

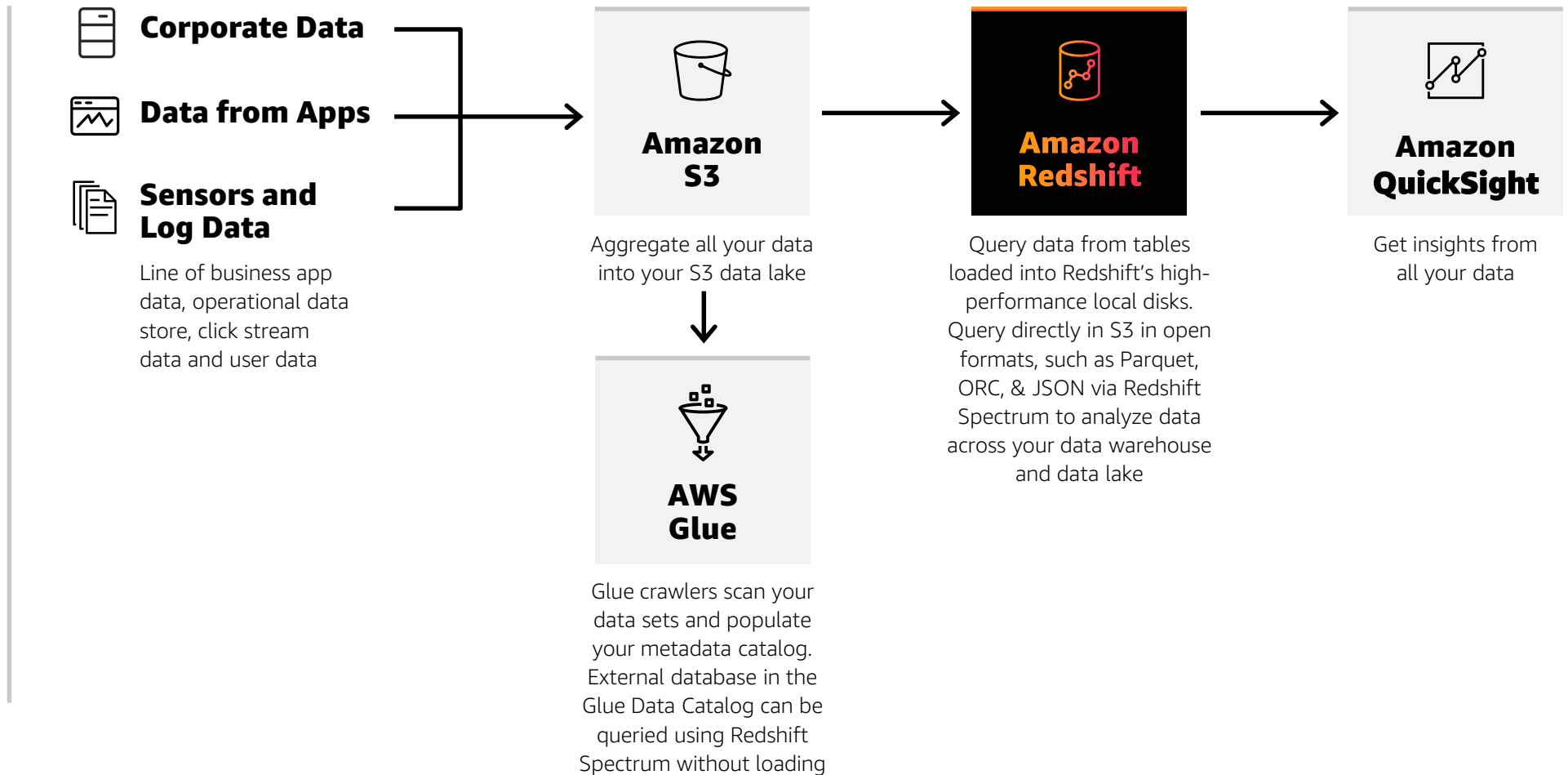


6

Amazon Redshift Use Cases

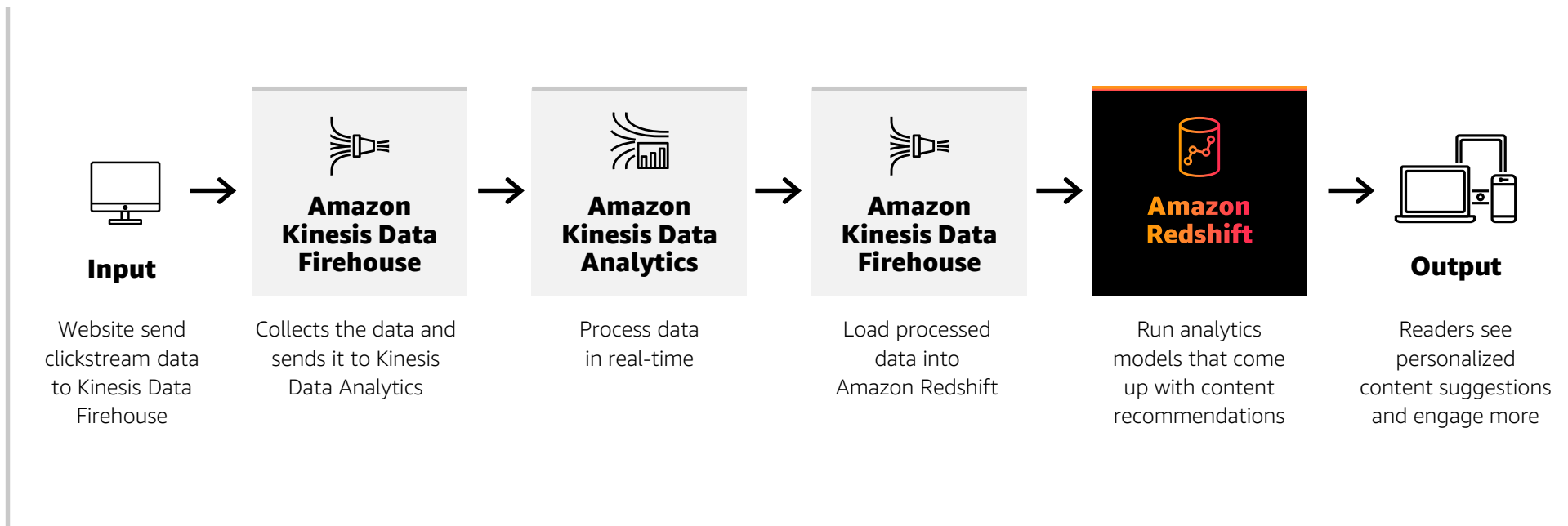
Amazon Redshift Use Cases

① Business Intelligence Using Modern Data Warehouse and Data Lake



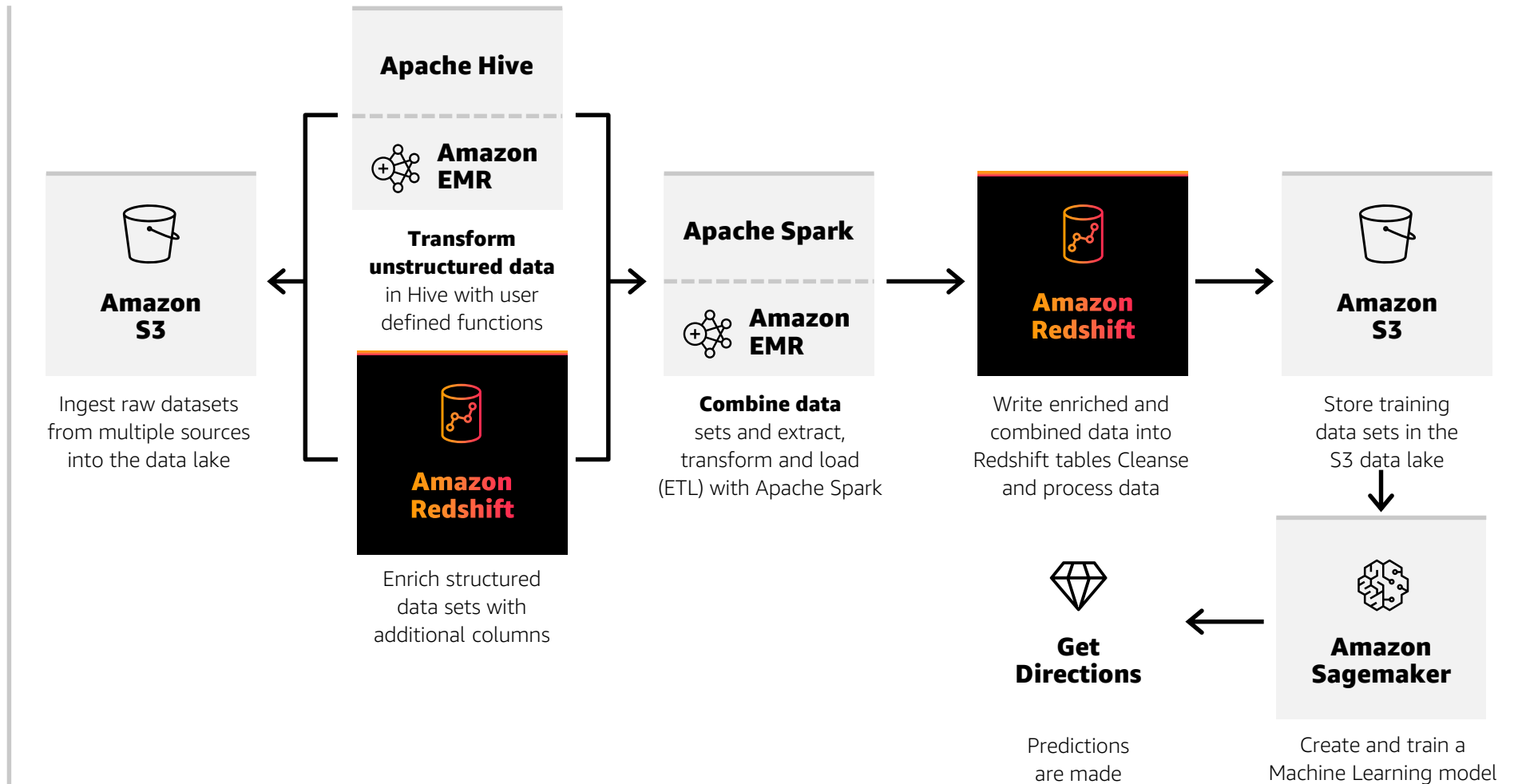
Amazon Redshift Use Cases

② Use Case – Real-time Streaming Analytics



Amazon Redshift Use Cases

③ Integrate with Other Analytics for Data Science and Machine Learning



7

Migrating to Amazon Redshift from On-Premises Data Warehouses

Self-service Migration using Database Migration Service

Migrating with AWS Partner Network Systems Integrator partners

Migrating to Amazon Redshift

If you decide to migrate from an existing data warehouse to Amazon Redshift, which migration strategy you should choose depends on several factors:

Network bandwidth between the source server and AWS

Whether the migration and switchover to AWS will be done in one step or as a sequence of steps over time

The rate of data change in the source system

Transformations needed during migration

The partner tool that you plan to use for migration and ETL

One-step migration



One-step migration is a good option for small databases that don't require continuous operation. Customers can extract existing databases as comma-separated value (CSV) files, then use services such as AWS Import/Export Snowball to deliver datasets to Amazon S3 for loading into Amazon Redshift. Customers then test the destination Amazon Redshift database for data consistency with the source. Once all validations have passed, the database is switched over to AWS.

Two-step migration



Two-step migration is commonly used for databases of larger size:

- 1. Initial data migration:** The data is extracted from the source database, preferably during nonpeak usage to minimize the impact. The data is then migrated to Amazon Redshift by following the one-step migration approach described previously.
- 2. Changed data migration:** Data that changed in the source database after the initial data migration is propagated to the destination before switchover. This step synchronizes the source and destination databases.

Once all the changed data is migrated, you can validate the data in the destination database, perform necessary tests, and if all tests are passed, switch over to Amazon Redshift.

Tools for self-service database migration



Several tools and technologies for data migration are available. You can use some of these tools interchangeably or you can also use other third-party or open-source tools available in the market.

AWS Database Migration Service supports both the one-step and the two-step migration processes described preceding. To follow the two-step migration process, you enable supplemental logging to capture changes to the source system. You can enable supplemental logging at the table or database level.

More than 100,000 databases have been migrated using the AWS Database Migration Service

Migrating to Amazon Redshift

Migrating with a Systems Integrator



Not all migrations can be done self-service. Organizations may have built custom schemas over time, may have data spread out in a variety of online and offline sources, and may have mission critical applications dependent on the existing data warehouse. In such scenarios, an AWS Partner Network System Integration and Consulting partner can provide specialized resources to ensure that the migration goes smoothly.



[Cloudreach](#) is a leading provider of software-enabled cloud services. They have empowered some of the largest and best known enterprises in the world to realize the benefits of cloud. Through intelligent and innovative adoption, their customers always adopt cloud rapidly, efficiently and at scale.



[Cloudwick](#) is the leading provider of enterprise business and technology modernization services and solutions to the Global 1000 with years of experience in architecting, scaling, and managing production enterprise big data services, including data warehouses and data lakes.



[TensorIoT](#) was founded on the instinct that all 'things' are becoming smarter. Their founders helped build world-class IoT, Analytics and AI platforms at AWS & other companies and are now creating solutions to simplify the way enterprises incorporate edge devices and their data into their day-to-day operations. Their Redshift practice helps customers prove-out, pilot, and accelerate their cloud data warehouse initiatives with best-of-breed architecture, rapid data ingest, transformation and modeling for analytics.



[Agilisium](#) is a Los Angeles based Big Data and Analytics Company with clear focus on helping organizations accelerate their "Data-to-Insights-Leap". To this end, Agilisium has invested in all stages of data journey: Data Architecture Consulting, Data Integration, Data Storage, Data Governance and Data Analytics. With advanced Design Thinking capabilities, a thriving partner Eco-system and top-notch industry certifications, Agilisium is inherently vested in our clients' success.



[47Lining](#), now part of Hitachi Vantara, is an AWS Premier Consulting Partner with Big Data and Machine Learning Competency designations. They develop big data solutions and deliver big data managed services built from underlying AWS big data building blocks like Amazon Redshift, Kinesis, S3, DynamoDB, Machine Learning and Elastic MapReduce. They help customers build, operate and manage breathtaking "Data Machines" for their data-driven businesses.

For more information on data integration and consulting partners, see [Amazon Redshift Partners](#).

8

Getting Started with Amazon Redshift

Getting Started

For more information about data warehousing on AWS and how to get started, visit:

[What is a Data Warehouse?](#)

[What is Amazon Redshift](#)

[Migrate your Data Warehouse to Amazon Redshift](#)

[Getting started with Amazon Redshift](#)

[Deploy a data warehouse on AWS in 60 minutes](#)




For 10 years, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud platform.

AWS offers more than 90 fully featured services for compute, storage, databases, analytics, mobile, IoT and enterprise applications from 54 Availability Zones (AZs) across 18 geographic regions in the U.S., Australia, Brazil, Canada, China, Germany, India, Ireland, Japan, Korea, Singapore, and the UK. AWS services are trusted by millions of active customers around the world monthly -- including the fastest growing startups, largest enterprises, and leading government agencies -- to power their infrastructure, make them more agile, and lower costs. To learn more about AWS, visit aws.amazon.com.

9

Case Study:

Equinox Fitness Clubs Realizes Faster
Time-To-Benefit and Productivity with AWS

A photograph of a wooden floor with a herringbone pattern. The word "EQUINOX" is painted in large, white, sans-serif capital letters across the floor. To the left of the letters, there is a small red and black ball. In the background, a green wall and a ceiling fan are visible.

Equinox built an analytics pipeline with AWS that is extremely flexible, friendly on storage size, and very performant when queried. The new pipeline reduced time-to-benefit and increased end-user productivity.

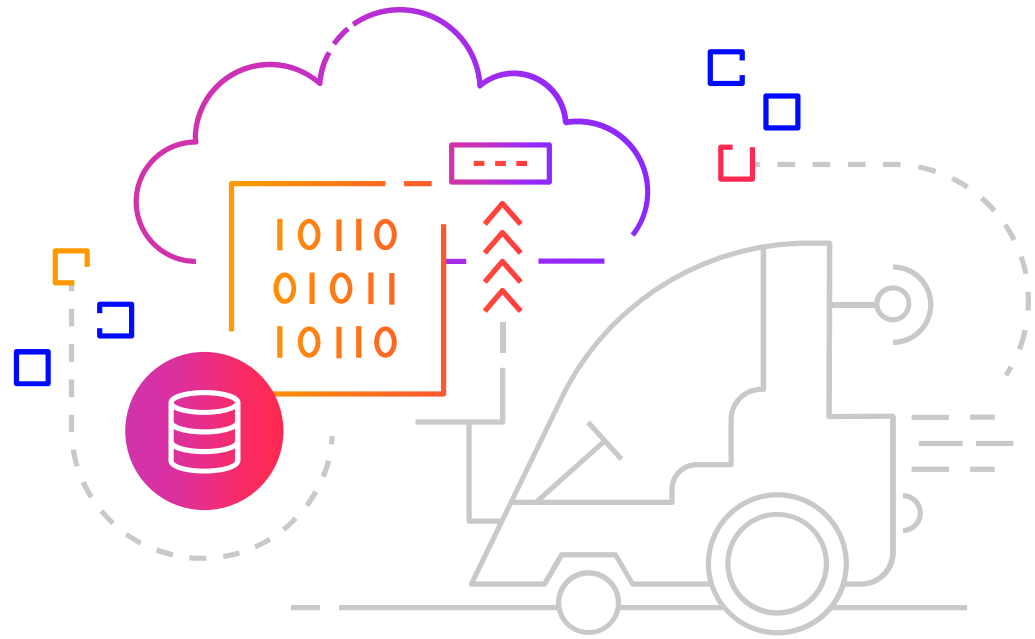
Case Study:

Equinox Fitness Clubs Realizes Faster Time-To-Benefit and Productivity with AWS

Equinox Fitness Clubs is a company with integrated luxury and lifestyle offerings centered on movement, nutrition and regeneration. Their brands include Blink, Pure Yoga, SoulCycle, Hotel-Equinox, etc. Equinox built connected experiences using applications that connect to Apple Health, and built data collection in their exercise equipment (gamified cycling machines, cardio machines, etc.).

Equinox moved their data warehousing infrastructure from Teradata to Amazon Redshift, combined with their Amazon S3 data lake where they store data from disparate sources (clickstream data, cycling logs data, club management software, data from software that enhances their services). Now they can perform queries from Amazon Redshift, blending structured Salesforce data with semi-structured, dynamic Adobe Analytics data to get a 360 degree view of each customer. Their analytics pipeline is extremely flexible, friendly on storage size, and very performant when queried. They are now leveraging Amazon Redshift, using Spectrum, for many new use cases such as data quality checks, analyzing machine data, historical data archiving, and empowering data analysts and scientists to more easily blend and onboard data.

As a result, Equinox has reduced time-to-benefit and increased end-user productivity. Their new analytics platform, powered by Amazon Redshift, is very dependable and resilient to schema changes in the source data. Finally, they achieved huge cost savings over Teradata by eliminating the maintenance and support contracts, as well as hardware and original licensing costs.



About the Authors

[Ayushman Jain](#) is Sr. Manager, Product Marketing at Amazon Web Services. He loves growing cloud services and helping customers get more value from their cloud deployments. He has several years of experience in Software Development, Product Management, and Product Marketing in developer and data services.

[Vinay Shukla](#) is a product manager on the Redshift team. Prior to Amazon, he has worked at Oracle, Hortonworks, and Teradata. When not at work he hopes to be on a Yoga mat or on a mountain. He still thinks he was not a bad programmer.