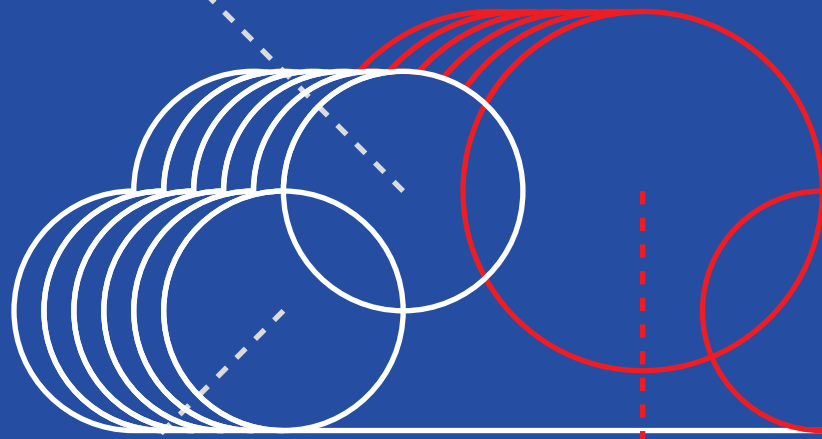
Informatica™

amazon web services

The Ultimate Workbook for Cloud Data Warehousing With Amazon Redshift.



Accelerating successful cloud analytics on Amazon Web Services.

Contents

Tip: Click on parts to jump to the particular section you want.

Contents

Tip: Click on parts to jump to the particular section you want.

Introduction.

Businesses Shift to Public Clouds for Agility.

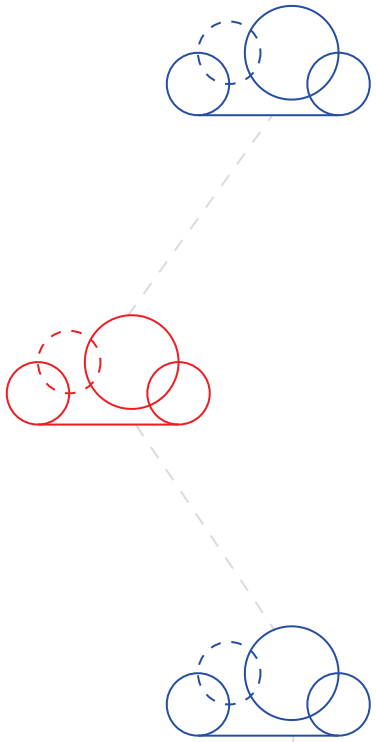
Data is a powerful force enabling businesses to disrupt competitive landscapes and to discover new revenue streams.

To make game-changing business decisions based on accurate insights, decision makers and analysts need fast access to trusted and timely data, regardless of whether the data is in the cloud, on-premises, or in a hybrid cloud environment.

Companies create and pull data from many different parts of the organization, as well as from external sources. And the types of data vary, ranging from structured, semi-structured, or unstructured.

To ensure performance, agility, and flexibility, and to benefit from cloud economies, IT organizations are moving on-premises workloads to public clouds, such as Amazon Web Services (AWS), at an increasing rate. In its "2017 State of the Cloud" report, InteropITX Research[1] found that since 2012, the rate of businesses shifting workloads to public clouds nearly doubled to 57 percent in 2016. The research also notes "substantial growth" in AWS usage, which increased from 39 percent to 52 percent in the past two years. The top three drivers to public clouds, according to the InteropITX report, are scalability, performance, and better/faster access to resources.

Clearly, cloud computing is the new normal.

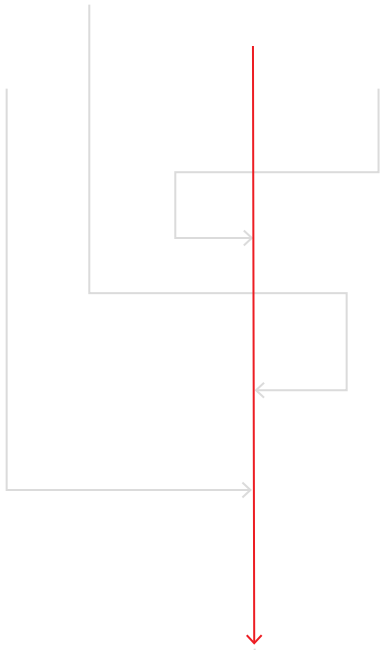


Better, Faster Decision-Making in the Cloud.

The agility of public clouds provides an ideal platform for businesses to use advanced analytics techniques, such as predictive and real-time analytics.

A cloud computing environment is ideal for processing the vast quantities of data necessary for advanced analytics, and the most successful businesses are helping to speed decision making by giving their users cloud-based self-service analytics tools. According to The State of Cloud Analytics 2016 report by EMA[2], 70.1 percent of respondents said cloud is a key part of their analytics strategy, with a further 21.6 percent citing that it is an important part of their analytics adoption.

With the focus shifting to analytics in the cloud, it makes sense that businesses are deploying or migrating data warehouses to public clouds, so that data and processing are close together for improved efficiency and cost savings. When TDWI researchers asked respondents to an emerging technologies survey, the following question: "What are you doing in the cloud for data management and analytics?" The top answer from respondents was, "Using a data warehouse in the cloud."[3]



Cloud Analytics Require Modern Data Management Techniques.

As cloud analytics and cloud data warehousing become strategic to many businesses, it's imperative that IT organizations support these initiatives with data management optimized for the cloud.

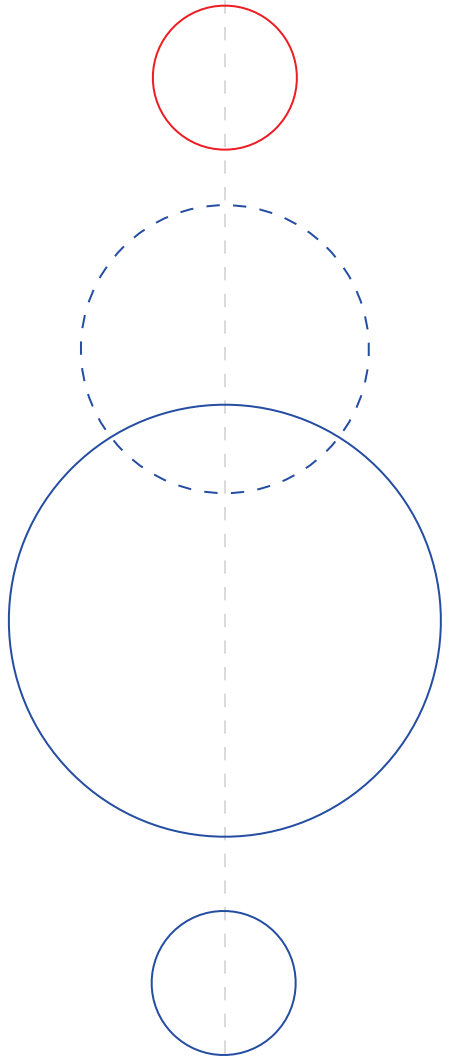
Traditional data management strategies for on-premises data warehouses were not designed to support cloud data warehousing or advanced analytics, and customers are finding a number of robust, cloud-friendly, or cloud-native tools to support these needs.

With AWS seeing "substantial growth" in usage overall, Amazon Redshift, AWS's cloud data warehouse platform, has gained traction among businesses. Whether you're already using Amazon Redshift or evaluating it, you should optimize your data management strategy for it.

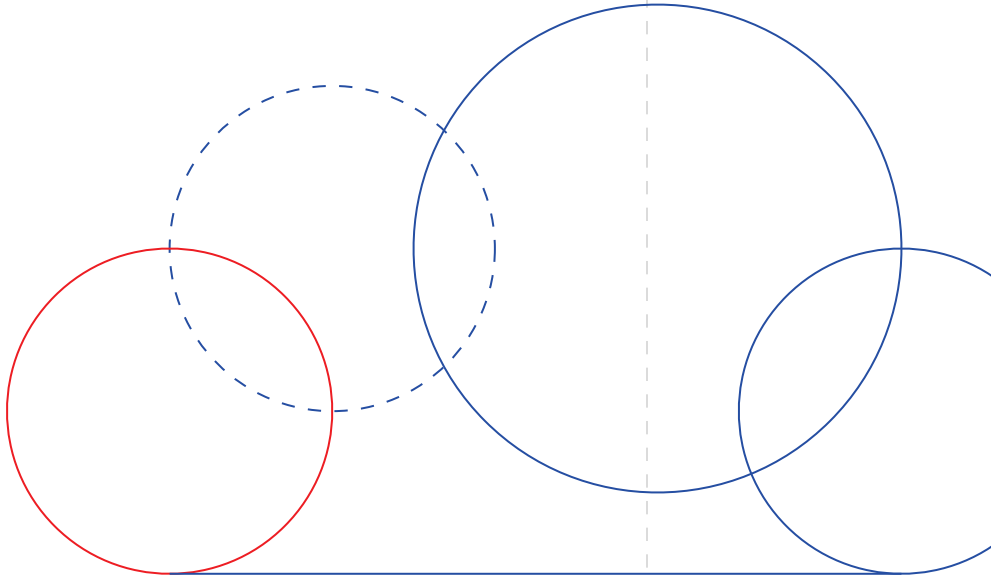
We've written this workbook to guide you through the steps to modernizing your data warehouse architecture with Amazon Redshift. We will show you how public cloud data management can enable the next generation of agile analytics initiatives. We'll describe the key cloud data management challenges and how to overcome them in order to support three common usage patterns for cloud data warehouses with Amazon Redshift.

By the end of the workbook, you'll be prepared to start your journey toward delivering trusted data to your business decision makers via Amazon Redshift cloud data warehouse.

Let's start by building the foundation.



A Bit About Data Warehousing in the Cloud.



Traditional on-premises data warehousing initiatives are often costly to acquire, need extensive upfront investment, and can take months to purchase the hardware and deploy the necessary software.

Public cloud data warehouses, such as Amazon Redshift, provide robust functionality, require virtually no upfront investment, and offer the added benefits of the cloud's elasticity, infinite scale, and agility. With cloud data warehouses, you pay for what you need, when you need it, eliminating the need for complex capacity planning and fixed capital investment.

A Bit About Data Warehousing in the Cloud.

Here are just some of the benefits of public cloud data warehousing with Amazon Redshift:

- Agility: Cloud data warehouses can be set up in days or hours, rather than months, and can quickly evolve to support changing data requirements of the business.
- Elasticity and Scalability: You can easily resize your Amazon Redshift cluster up and down as your performance and capacity needs change with just a few clicks in the console or a simple API call. No need to do complex capacity planning.
- Cloud Bursting: If your existing on-premises environment needs a temporary increase in processing or data capacity, you can burst into a public cloud environment to handle the increased load and then shrink the environment back to on-premises later.
- Offloading: Some enterprise data warehouse workloads can be selectively offloaded to Amazon Redshift injecting the cloud's speed and elasticity into the on-premises data warehousing environment.
- Fault Tolerant: Amazon Redshift has multiple features that enhance the reliability of your data warehouse cluster. All data written to a node in your cluster is automatically replicated to other nodes within the cluster and all data is continuously backed up to Amazon S3. Amazon Redshift continuously monitors the health of the cluster and automatically re-replicates data from failed drives and replaces nodes as necessary.
- Performance: Amazon Redshift delivers fast query performance by using columnar storage technology to improve I/O efficiency and parallelizing queries across multiple nodes.
- Security: Security is built-in. You can encrypt data at rest and in transit using hardware-accelerated AES-256 and SSL, isolate your clusters using Amazon VPC, and even manage your keys using AWS Key Management Service (KMS) and hardware security modules (HSMs).

Part 1:

Cloud Data Management Challenges.

New Data Types and Volumes Challenge Traditional Data Management Approaches.

Now that we've demonstrated how public cloud data warehousing enhances analytics, let's look at why traditional data management strategies aren't ideally suited for this new environment, and what an optimized approach looks like.

Data management requirements have changed since businesses began to introduce traditional on-premises data warehouses. Here are some reasons new paradigms are challenging traditional approaches:

- Increased data sources and volume: Businesses are consuming data from more sources than ever before: from big data, to Internet of Things devices, to business applications, and social media feeds. Data may reside in on-premises systems, in the cloud or often, in multiple locations. Along with new sources, the volume of data businesses process is increasing as well, necessitating a more complex data management strategy. It's also expensive and challenging for IT organizations to maintain the specialist skills needed to produce and maintain custom applications that smoothly integrate with continually changing APIs. Features and function can evolve rapidly when using SaaS data sources, especially when compared with traditional on-premises data sources.
- Increased data silos: Data silos continue to proliferate across enterprises through traditional IT, but SaaS data sources can create silos as well. A single business unit may use a number of SaaS solutions to meet its needs and when you multiply that across the entire company, the number of sources rapidly increases. Each solution may generate valuable data, but without an enterprise integration strategy, it's hard to get the insight your business needs to stay competitive. Silos slow down data access, and hamper your ability to get a holistic view of all your data.

New Data Types and Volumes Challenge Traditional Data Management Approaches.

- The need for immediacy: Today's "fail fast, fail often" business ethos and agile methodologies rely on fast prototyping, development, and speed to deployment of analytics initiatives. Cloud computing accelerates development and deployment of new analytic solutions, but puts a strain on traditional data management approaches. A more agile data management approach is needed to support agile analytics.

- Data volume and variety brings complexity: Monitoring and managing data in production is key to catching and correcting problems early before they escalate out of control. The increased variety of data sources brings a new set of complexities to this task.

- Data security continues to challenge: As organizations evolve toward a hybrid data management model, they need to pay extra attention to securing sensitive data at rest and in motion, whether they are moving data from on premise to cloud or from cloud to cloud.

- Data migration planning is critical: Migrating workloads to AWS can be a large and complex undertaking and often involve the migration of large volumes of data. This is particularly true if you're moving many different data types, from standard relational structures and predictable batches to huge volumes, massive variety, and real-time streaming data. Having the right plan can make a huge difference in this process.

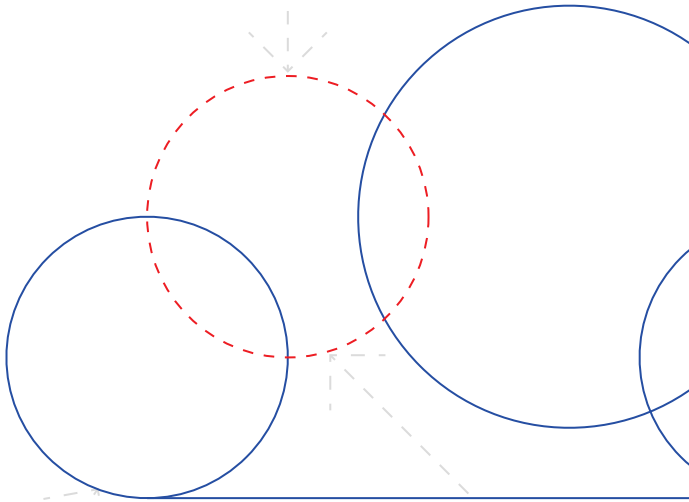
- IT must respond to business needs faster: Business decision makers have more new data types to analyze and their analysis needs can evolve rapidly. Often this means IT must develop new processes—quickly—to support this. When new data sources come online, IT must quickly find a way to integrate this data even if it is a new concept for the organization.

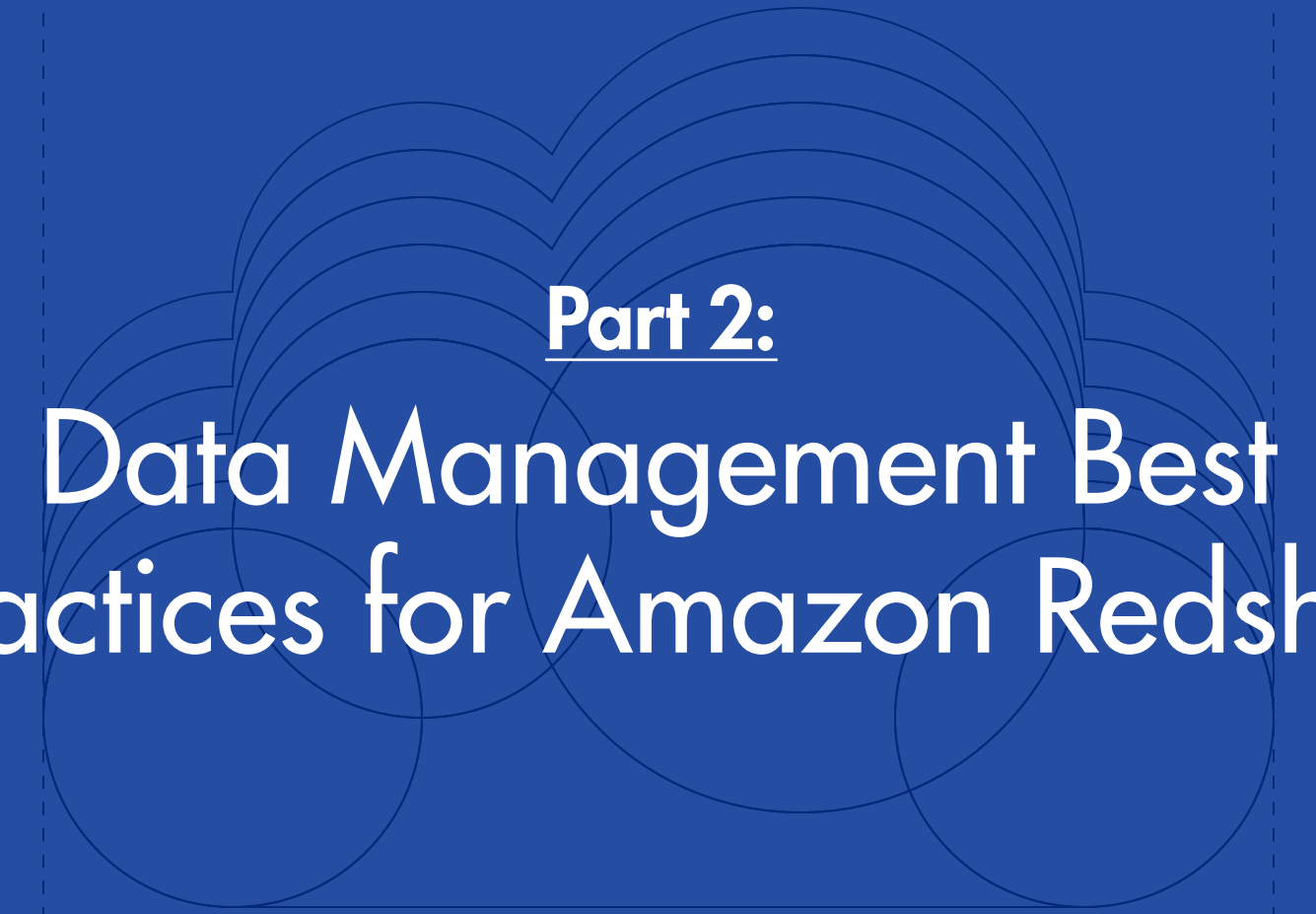
New Data Types and Volumes Challenge Traditional Data Management Approaches.

- Context is becoming more crucial: Business analysts can get valuable insights by knowing their data's "life story", enriching many analytics initiatives. Data visibility and lineage allows businesses to discover and understand the history of the data: where it came from, the path it took, who uses it, and so on. Data governance—the management and oversight of data by aligning people, process, policy, and technology—plays a critical role in helping you to deliver data that's complete, contextual, and trustworthy for analytics. But data governance is challenging when you're dealing with many new data sources and types across on-premises and cloud environments.

- The need for high quality data remains: "Garbage-in, garbage-out" has never been more true. Data assurance, quality, and integration are critical factors in IT and especially analytics. Ensuring data quality remain an intensive, complex requirement, magnified by the volume, velocity, and variety of the modern data landscape.

All this drives the need for a modern, agile approach to data warehousing that's based on cloud computing and supported by a data management strategy that's optimized for a cloud first or hybrid environment.





Part 2:

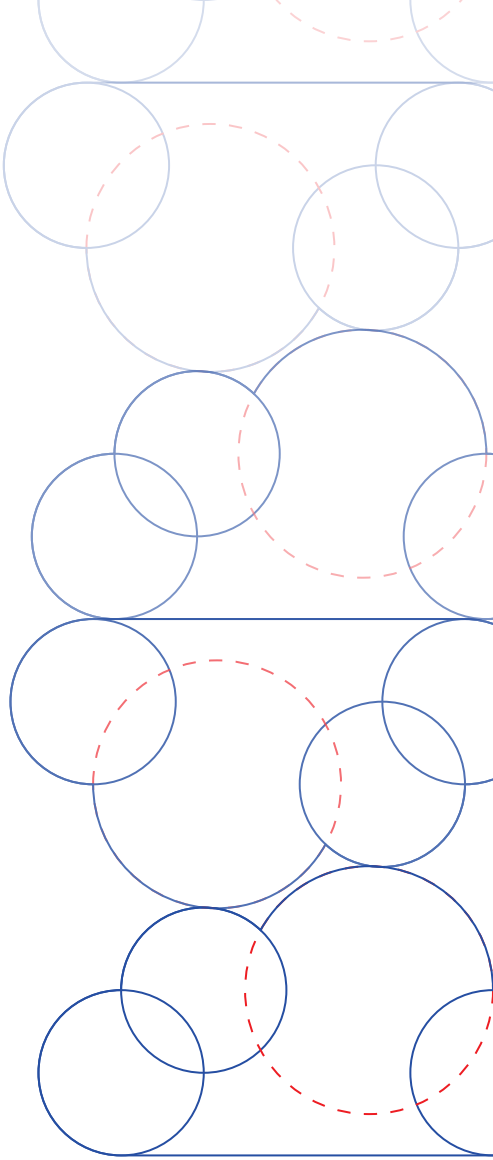
Data Management Best Practices for Amazon Redshift.

Stack the Odds in Your Favor with Automated Data Management.

Amazon Redshift is a powerful public cloud data warehouse that is well suited to support modern cloud analytics in the worlds of big data and IoT.

A number of features ensure Amazon Redshift provides the agility, speed, performance, and security that businesses need to deliver great data for business decision makers, while maintaining cost efficiencies. These include massive parallel processing (processing spread across multiple nodes), elastic scalable processing nodes for high performance, columnar storage for fast return of large warehouse-size data sets, built-in fault tolerance, and data encryption in transit for security, among others.

Automated data management and data integration solutions for Amazon Redshift go a long way to ensuring an agile environment for your business analytics, as well as the added capability to fully and easily harness the inherent benefits offered by the Amazon Redshift architecture. Successful cloud data warehouse projects share common characteristics.



Stack the Odds in Your Favor with Automated Data Management.

Here's a list of imperatives and best practices to guide your success:

- Robust connectivity architecture: A typical Amazon Redshift data warehouse requires data integration from any number of cloud and on-premises data sources. A data integration architecture that separates data integration logic from the underlying data source APIs, delivers a higher degree of code reuse and investment protection. You can simplify and accelerate your data integration development (separation of concerns), and can swap sources and targets in and out. Best of all, you don't need a lot of domain knowledge to understand the underlying APIs. This is especially important in the cloud, where SaaS applications and data services rapidly change and evolve their APIs.

- Data visibility via metadata: The creation of technical metadata to support change management is critical. A data integration solution that automatically captures the metadata of your integration logic, as you develop, helps you reduce both the risk of change and the time lost by managing the change. Technical metadata enables you to better plan for change by conducting impact analysis of changes in data and data integration code on upstream and downstream systems.

- Operational confidence: You can catch production problems in your data early by monitoring your data integration processes to ensure they are delivering timely data as expected. Having complete visibility into your entire enterprise data production environment is a powerful asset to ensure the right data is delivered in a timely fashion to your data consumers.

- Security: It's important to build the right level of security into the architecture at every level, both within technology and at the process level. Sensitive data should be protected at both rest and in motion via encryption and access controls. Regular patch maintenance and penetration testing keep you ahead of your adversaries. Secure network access and a layered security model goes a long way toward protecting your data. Security is not bolted-on after the fact; security must be integrated throughout the environment. Special attention must be given to instances whereby on-premises data is being accessed by cloud systems or is being moved between on-premises and cloud systems.

Stack the Odds in Your Favor with Automated Data Management.

- Agility: The definition of an agile IT environment is the ability to meet changing requirements quickly and without excessive retooling or cost. By leveraging platforms and architectures that enhance the productivity of all your analytics stakeholders—both technical and citizen integrators—you can ensure that connected, trusted data can be rapidly delivered to support business initiatives. Having the right data management systems in place allow you to easily add or remove new capabilities and onboard new data systems quickly.

- Scalability: One of the great benefits of cloud environments is the ability to scale up and down to match the workload needs of the business. You can leverage this elastic capability offered by both cloud data warehousing and cloud data integration solutions to support variable data volumes and almost infinitely scale your solution as your data grows. Pipeline optimization, parallelism, clustering technology.

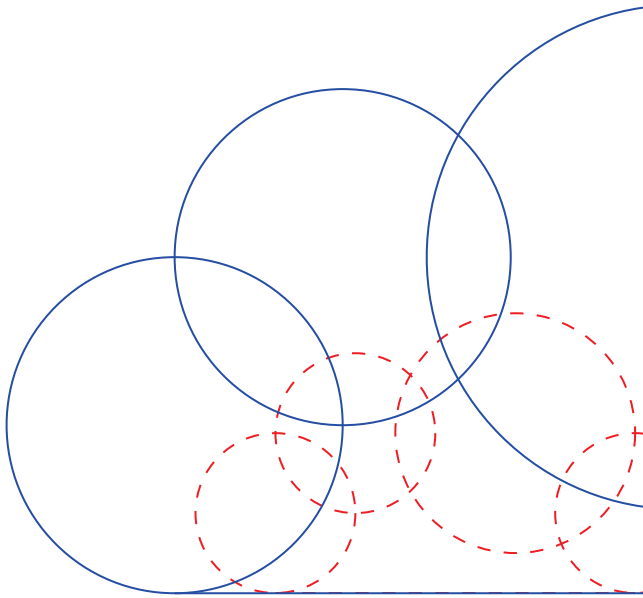
  Push Down Optimization (PDO) and partitioning are components of the entire stack, including the data warehouse and the data integration solution, all of which allows you to scale to meet workload requirements.

Using automated data integration solutions can help you automatically leverage the inherent scalability and performance benefits offered by Amazon Redshift, while eliminating manual, multi-step programming methods. During data load and transformation operations and in support of compute-intensive analytical operations, the ability to scale upward ensures processing completes accurately and on time.

Stack the Odds in Your Favor with Automated Data Management.

- Performance: Easy-to-scale performance is an important benefit of cloud data warehouses. To get the best performance, leverage database software and schema architecture optimized for data warehousing and analytic processing. Ensure underlying cloud infrastructure (compute nodes, cluster capabilities, storage, and networking) are modern and appropriate to the level of performance you require. Employ parallel processing, clustering, and PDO on modern multi-node columnar architectures and ensure that your data management solution automates and fully supports these types of data warehousing use cases.

- Leverage the inherent capabilities of your data warehouse: Learn the characteristics of your data warehouse software, design, and the supporting cloud infrastructure; understand what it does well and where its limitations exist. For example, if you know your data warehouse supports massive parallel processing and native pushdown capabilities across multiple compute nodes, then executing the data transformation processing power in the target warehouse environment makes sense, especially if your solutions allow you to perform this capability and automate the process



Stack the Odds in Your Favor with Automated Data Management.

- Standardization: Within cloud data warehousing, the practice of standardization applies to multiple areas: data integration, governance procedures, application processes, code, and skills promoting reuse, and data preparation. Standardization is a prerequisite step to repeatability and automation, which enhances agility and time to market of new solutions.

- Repeatability: Processes and procedures should be well-defined, documented, and repeatable so they can be automated. Ad-hoc and one-off processes requiring manual intervention or monitoring return less value to the business. Implement solutions supporting the automation of repeatable processes so your resources are freed up to work on other, higher impact projects, such as business analytics.

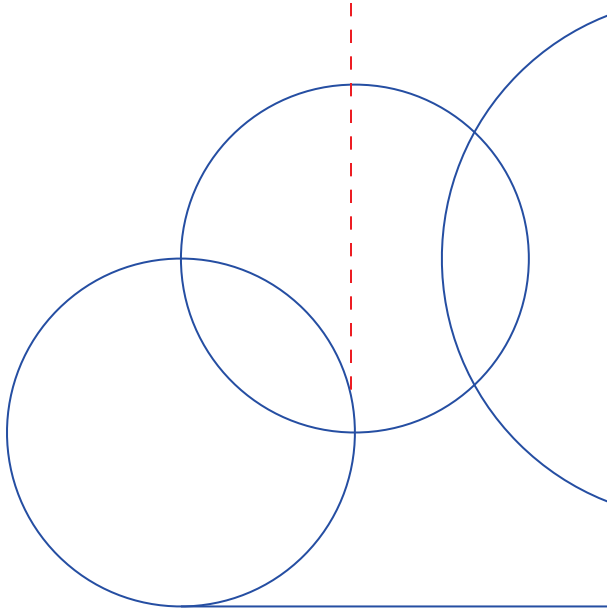
- IT and business collaboration: Teamwork, communication, and sound project management principles are often the difference between success and failure. If the collaboration between IT and business is insufficient, communication is weak, or expectations and goals are not managed, then even a great technical solution will struggle to be successful. Close collaboration between stakeholders is the key to success for any IT project, regardless of technology. Look for solutions that promote business and IT collaboration. For example, role-based tools allow different stakeholders to leverage their unique expertise (e.g. data knowledge vs. technology knowledge) and transparently "swap" work products back and forth as part of the data management work stream.

- Flexible Integration: Data integration software should connect all the data sources and applications you need regardless of their location on-premises or in the cloud. High quality, mature data integration solutions with robust capabilities in cloud as well as on premise promote this flexibility in hybrid environment leading to increased business agility.

Stack the Odds in Your Favor with Automated Data Management.

- Data governance: Implementing strong data governance procedures becomes even more important when you move data off-premises into a cloud environment. While cloud computing brings a new set of concerns for data governance, they are entirely manageable with careful planning, forethought, and with the right technology stack. AWS provides robust security and access controls in the cloud. Understanding your data and the impact of changes to source and target systems is just as important as building out your data integration architecture.

- Data management solutions: User-friendly, mature data management and integration solutions, with ETL and ELT capabilities, is a proven success factor for agile analytics initiatives in a cloud and hybrid environment. Furthermore, leveraging the right data management platform means it can be reused for future use cases and projects, will future-proof your analytics environment against changes and ease on-going maintenance operations while reducing manual workload. Supportive change management via toolsets is a key asset to this approach. The intent is to use integrated end-to-end solutions which work with you, not against you.



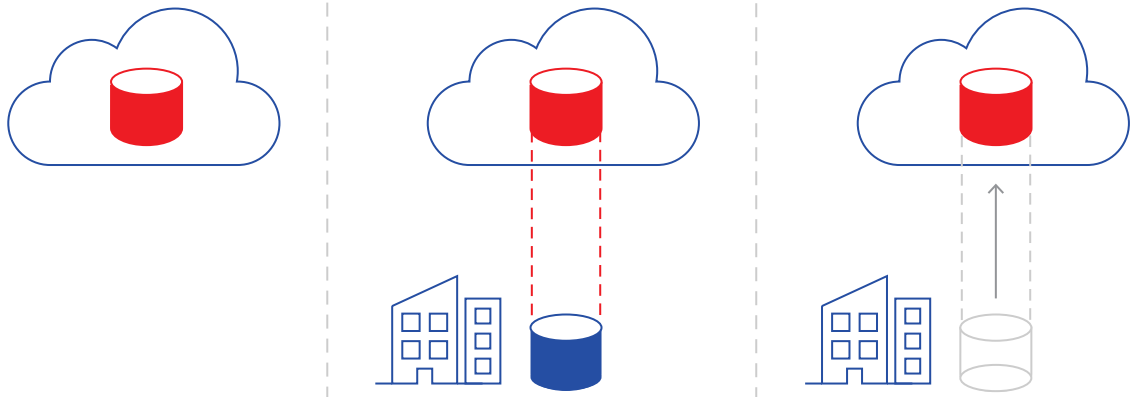
Part 3:

Three Ways to Leverage Cloud Data Warehousing With Amazon Redshift.

Exploring Cloud Data Warehouse Use Cases.

Now that we know more about data management best practices for cloud data warehousing, how do we start using it for Amazon Redshift? Organizations typically fall into these three common scenarios for cloud data warehousing with Amazon Redshift as can be seen in figure 3.1.

Figure 3.1 — Paths to Cloud Data Warehousing and Analytics

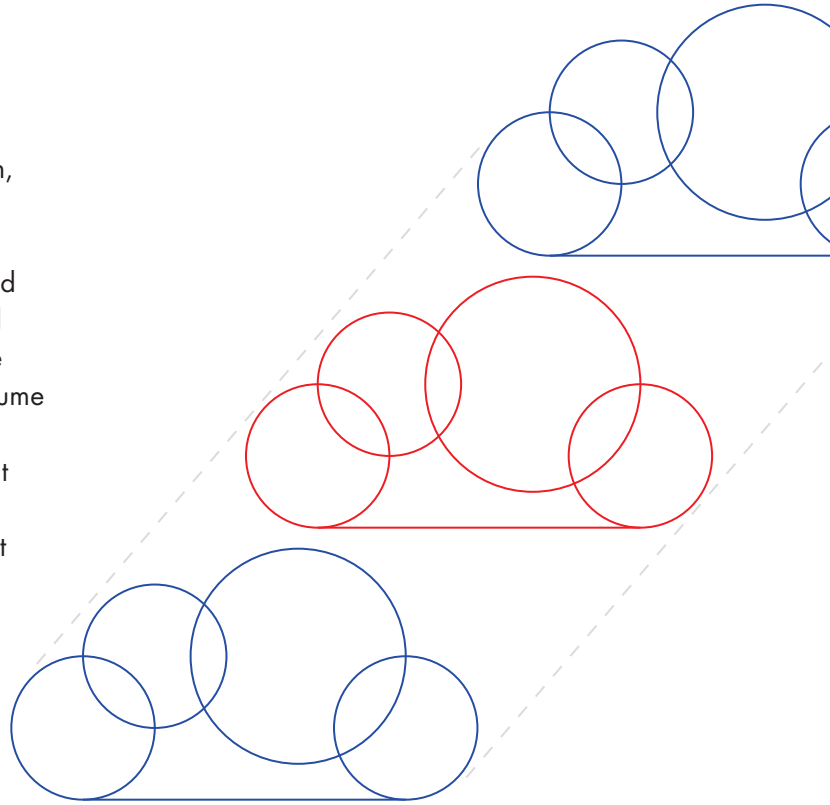


Exploring Cloud Data Warehouse Use Cases.

1. New Cloud Data Warehouse on Amazon Redshift

Rather than building your data warehouse on-premises, you start with a net-new data warehouse created in the cloud; your data warehouse is "born in the cloud" using Amazon Redshift. This often comes about when you are planning a new analytics initiative for a project, a business unit, or for an area that previously had not benefited from a strong analytics effort. A new Amazon Redshift data warehouse is highly complementary to an agile development effort and "fail fast" philosophy, which is often part and parcel with new business ventures or initiatives.

Amazon Redshift, when supported by a robust cloud data management platform, allows you to rapidly experiment, test, and evolve new ideas as they come in, without long development lead times and heavy upfront investment in systems and resources. As your new data warehouse project grows and data increases in volume and complexity, you can rapidly scale and enhance your analytics environment with the elasticity and performance capabilities inherent to Amazon Redshift and a robust data integration solution.



Exploring Cloud Data Warehouse Use Cases.

2. Extension of Enterprise Data Warehouse with Amazon Redshift

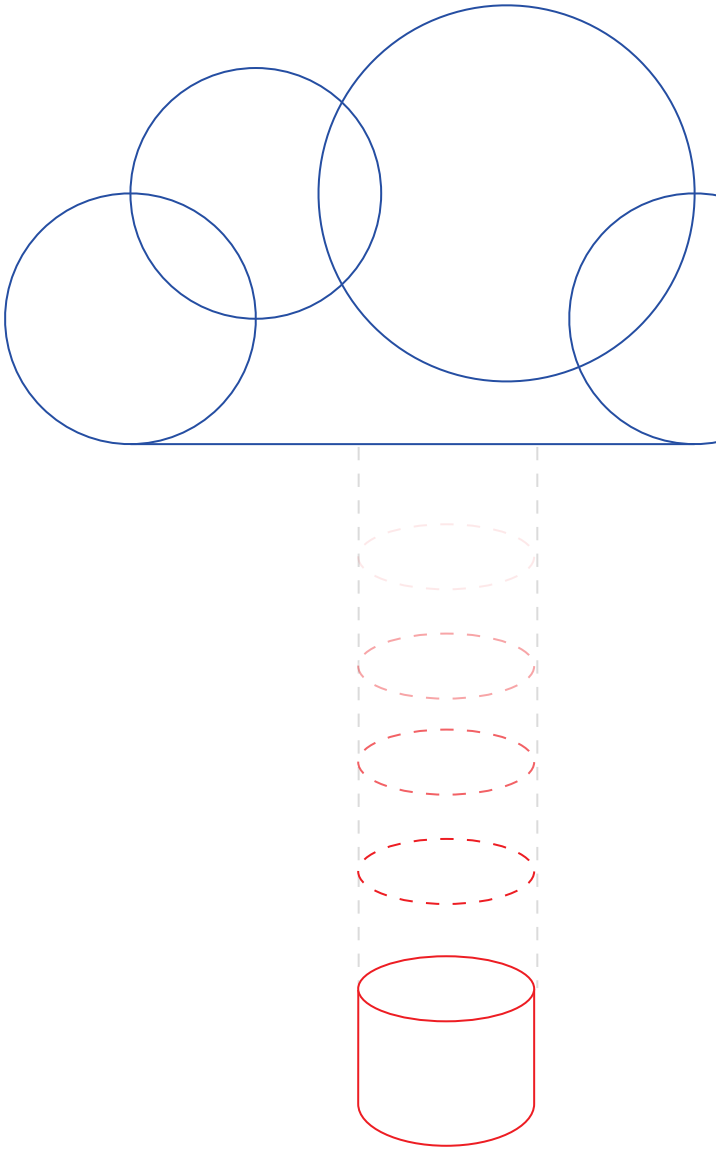
There are many reasons why you may need to keep assets (data and applications) on premises, such as for regulatory compliance. But for agility, scalability, and flexibility, it makes sense to create a hybrid cloud data warehouse environment by extending your existing enterprise data warehouse into the cloud with Amazon Redshift.

Hybrid cloud environments integrate on-premises resources with cloud based computing environments, and this is an easy way to start using the cloud without a major change to your current IT methodology.

To create a hybrid cloud data warehousing environment, you create a new cloud data warehouse with Amazon Redshift and extend your existing on-premises enterprise data warehouse architecture into it. You can leverage cloud-based data integration solutions to move your newest, most frequently used data to the cloud and leave less frequently used data on-premises (or data that must remain on-premises). This approach solves many challenges for organizations that want the benefits of cloud computing but whatever reasons may need to move at a slower or more controlled pace.

Hybrid cloud data warehousing with Amazon Redshift lends itself to cloud bursting situations as well. As workload pressures increase within the on-premises data center, processing can burst into AWS where near-infinite capacity and performance absorbs the increase in workload. Cloud bursting provides unlimited capacity and performance during critical spikes in processing when it is needed, and provides some elasticity and cost benefits.

Exploring Cloud Data Warehouse Use Cases.



3. Migration (a.k.a. "Lift and Shift") of Enterprise Data Warehouse to Amazon Redshift

Some organizations modernize their traditional on-premises enterprise data warehouse by moving it into Amazon Redshift. This may require significant effort for some businesses, but it can bring substantial value to the organization. In this scenario, businesses migrate (or "lift and shift") their on-premises data warehouse into a cloud-only environment.

This approach results in an entirely cloud-based data warehouse where your on-premises enterprise data warehouse is ultimately decommissioned. Your data is in the cloud, so it readily integrates with cloud based applications and services.

Benefits of this approach include lower upfront costs and the "pay-as-you-go" model of cloud computing. Just as important, this architecture offers rapidly expandable computing resources (storage, memory, CPU) so you can scale your analytics environment as your data sources increase and data volumes expand. With the right set of solutions, you can seamlessly integrate data from applications and databases, in cloud and on-premises, to your new cloud data warehouse.

Why It Makes Sense to Build Your New Data Warehouse in the Cloud.

Traditional approaches for building and evolving on-premises data warehouses are robust, but it can take months and cost millions of dollars just to get started.

Once you add in the maintenance and growth requirements of an on-premises system throughout its life cycle, the cost of this type of solution quickly adds up. There is a better way: build a new Amazon Redshift data warehouse in the cloud and bypass the traditional on-premises approach.

Just as important, new cloud data warehouses can be business-opportunity driven to provide quick agility and ease of experimentation to give your organization a competitive differentiation. These environments are great to test proof of concepts, and to quickly kick off new analytics projects and then rapidly scale and evolve as your needs grow.

You can experiment to try different reporting, as well as to test elements of your business to analyze data in different ways. You can do all this without incurring the high costs of both time and money inherent to a traditional, on-premises analytics approach.

Agility is a key driver for new analytics projects, and using Amazon Redshift with a complementary data management solution is a great way to start. With Amazon Redshift, you can spin up a new analytics project within hours or days—much faster than with traditional on-premises approaches.

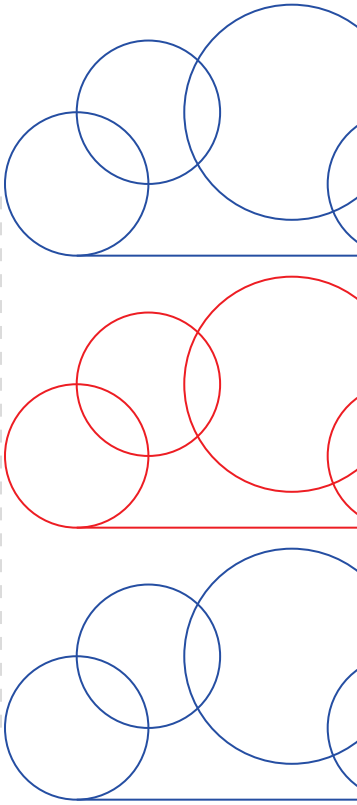
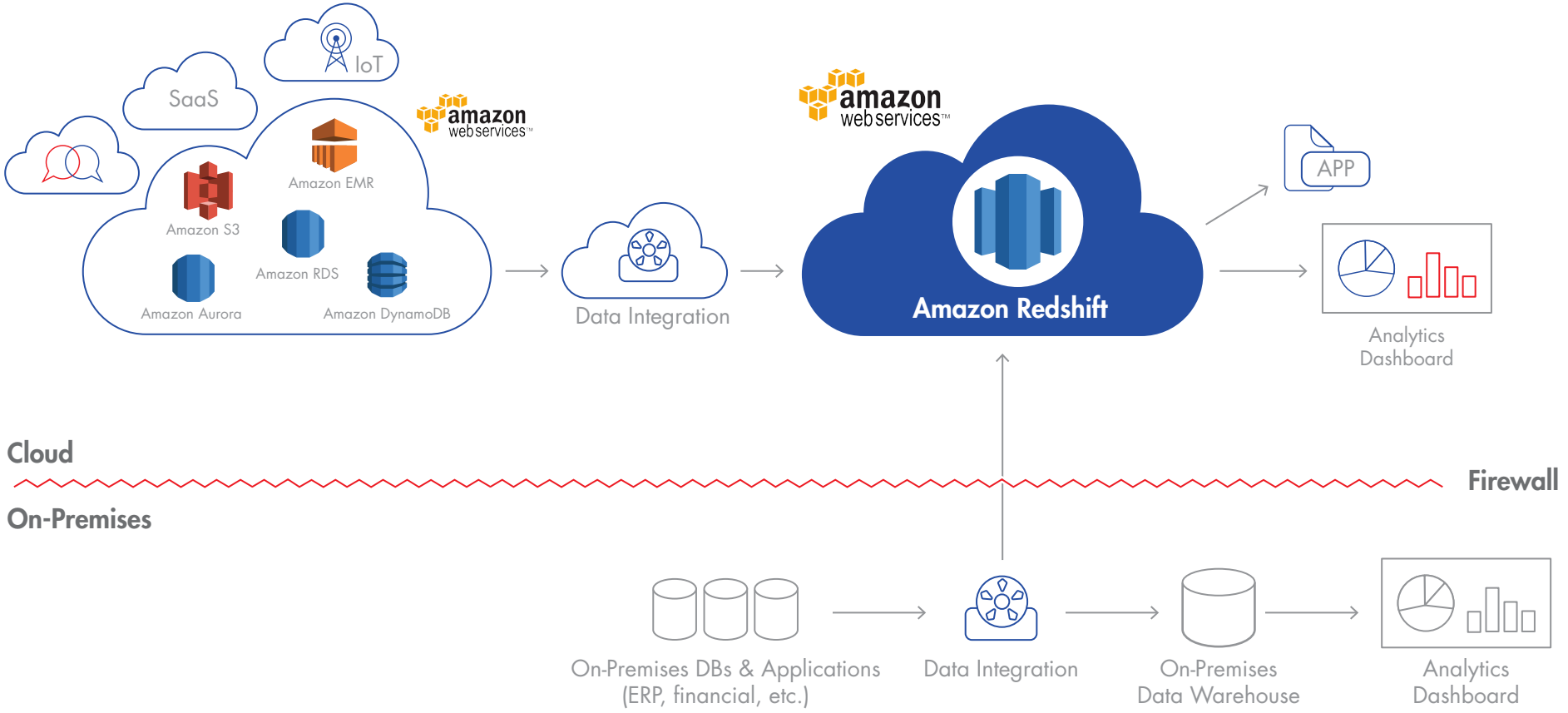


Figure 3.2 — New Cloud Data Warehouse (Born in the Cloud)



Why It Makes Sense to Build Your New Data Warehouse in the Cloud.

About This Architecture.
The new cloud data warehouse is primarily focused on increasing agility while reducing costs in terms of time and money to promote low risk experimentation and analytics projects.

Leveraging the inherent strengths of the AWS cloud coupled with Amazon Redshift, you have a highly capable and empowering architecture that lets you to support many different initiatives and needs, including:

- New analytics initiatives requiring speed and agility.
- Experimental efforts that don't require a lot of upfront commitment to set up so you can "fail fast" and try again until you get the results you want.
- A modular architecture that provides growth and new capabilities as you need it.
- Paying only for what you use by using elasticity to scale up or down your data architecture, and increase or decrease performance as required.

Note the role of data management and integration as you move to the cloud. No longer is your data management and integration platforms limited only to on-premises locations, but now they also exist in the cloud. Your data management and data integration platforms should complement and simplify both on-premises and cloud computing environments, and integration between those environments.

Integration is Everywhere.
The move to cloud is a journey, and even though you have created a new cloud data warehouse, you still will likely have on-premises assets in the form of databases, applications, and infrastructure; this is your hybrid cloud computing environment. Data integration between your on-premises data sources and your cloud data warehouse must still occur; as must data integration between your Amazon Redshift data warehouse and other cloud based data sources. Integration between your cloud data warehouse and your on-premises data warehouse may also be required. A comprehensive data management platform supporting on-premises and cloud-based data integration play a critical role in supporting each use case.

Extending Your On-Premises Enterprise Data Warehouse to the Cloud with Amazon Redshift.

One approach to data warehouse modernization is for your existing on-premises computing assets to leverage new cloud based applications and databases.

This scenario is conceptually simple yet powerful—you extend your existing on-premises enterprise data warehouse into a new cloud data warehouse with Amazon Redshift. By segregating data, you pick which data remains on-premises and you move the rest into the new cloud data warehouse.

Hybrid data warehousing environments can solve certain business requirements; businesses keep the data and applications they want on-premises and move the rest to the cloud for improved agility and lower cost.

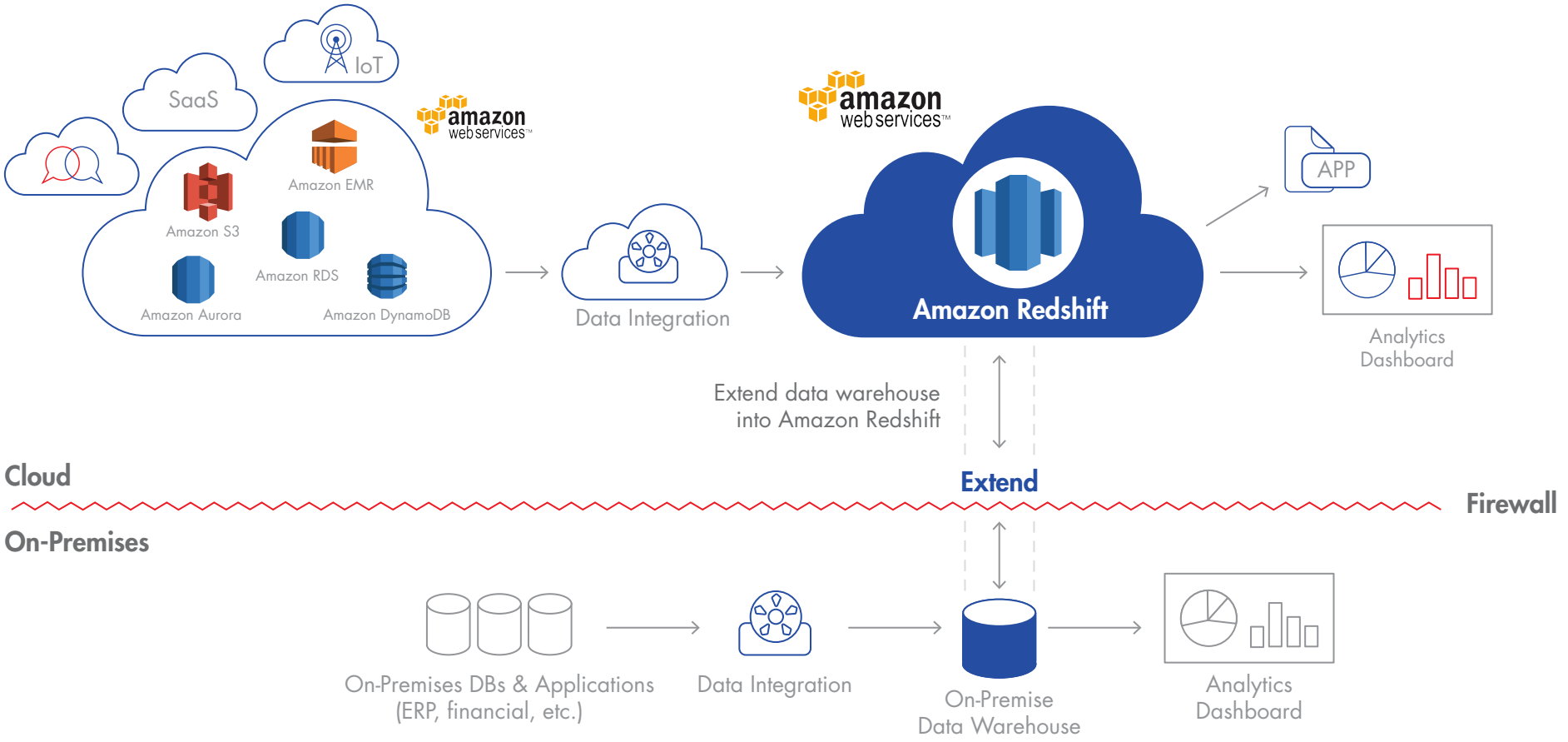
Some of the most common yet powerful drivers to this approach include:

- New technologies such as SaaS applications that must be adopted quickly to capitalize on business opportunities. Businesses can rapidly integrate new SaaS applications by leveraging an agile cloud data warehousing solution.
- Traditional data warehouses that are running out of capacity and are cost prohibitive to support. This issue disappears when businesses use scalable and lower cost cloud data warehousing platforms for tasks such as bursting workloads to the cloud during peak times for data processing.
- The high cost and effort to scale existing enterprise data warehouses. Elastic cloud computing allows you to quickly increase performance and scale storage capacity up and down as needed.

To get the most benefit from a hybrid data warehouse approach, leverage automated data integration solutions to migrate and load the data into your new Amazon Redshift data warehouse. Then, deploy a unified data integration platform to manage your data through subsequent updates. You can conduct analysis and testing of cloud bursting capabilities to prepare for workload spikes. Using a single data integration platform to support all data warehousing workloads — on-premises and in the cloud — enables a more consistent data-centric approach, promoting increased productivity and reuse across these platforms.

Integrating cloud analytics capabilities on Amazon Redshift with your on-premises enterprise data warehouse and assets is a measured, low risk approach many organizations use to achieve agility and lower costs.

Figure 3.3 — Extend into Cloud Data Warehouse



Extending Your On-Premises Enterprise Data Warehouse to the Cloud with Amazon Redshift.

About This Architecture.
This scenario is primarily focused on increasing agility and lowering the costs of your data warehousing initiative, while maintaining an on-premises data warehouse footprint.

Amazon Redshift running on a scalable and secure AWS infrastructure is a natural fit for this architecture as you can:

- Store your newest and most frequently used data in the cloud where storage is lower cost and efficient, leaving your less accessed legacy data on-premises.
- Offload your undifferentiated heavy lifting to a managed service like amazon Redshift.
- Move your newer workloads or those experiencing performance or reliability problems from on-premises to the cloud allowing you to take advantage of cloud based applications and more easily integrate with new data sources.
- Provide additional processing power and storage capacity to make your IT operations faster and cheaper via AWS than you could achieve by upgrading on-premises hardware, networking, and storage.
- Expand capability and test new technologies quickly without investing in on-premises infrastructure.
- Leverage cloud bursting to provide additional capacity (storage and compute power) during times of increased demand (think Black Friday shopping or year-end reporting).
- "Test the waters" of cloud data warehousing prior to migrating your entire on-premises enterprise data warehouse to the cloud.

Not every organization is ready to jump entirely into the cloud, and that's okay as hybrid cloud data warehousing solutions, leveraging Amazon Redshift enable you to begin to enjoy the benefits of cloud agility and lower cost, while still maintaining the necessary on-premises footprint for your business requirements.

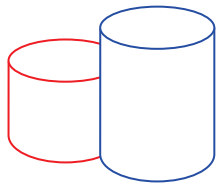
Take the Bold Leap—Move Your On-Premises Data Warehouse to the Cloud.

Migration (a.k.a. "Lift and Shift") of your on-premises enterprise data warehouse to the cloud is the logical desired outcome for many organizations.

Facing increased costs, complexity, and the crushing overhead of maintaining on-premises infrastructures, many organizations are asking "why are we still doing this to ourselves?" This is coupled with the frustration of trying to meet increasingly aggressive business requirements with new technology while maintaining a legacy infrastructure. For those organizations, cloud data warehousing is the solution to many of their most pressing challenges.

While the core drivers for migration are very similar to those for new and extending cloud data warehousing, specific drivers include:

- Eliminating data center maintenance and support costs (both technology and people).
- Enhancing security by eliminating the patchwork of on-premises safeguards and moving to a unified, built-in public cloud-based security model. There is a growing industry sentiment that inherent security in public cloud is higher than can be attained on-premises. This is because public cloud offerings like Amazon Redshift build in security that meets many of the compliance requirements of businesses and organizations across industry and geography.
- Addressing the need for elasticity by providing scale-up and scale-down capacity as needed.
- Consistency, modernization, and simplification of software and infrastructure. Cloud service providers, such as AWS, ensure all components are on the most recent release, and upgrades are simple, and can be automated. Take advantage of new features immediately, without waiting for a manual refresh, like in the on-premises world.

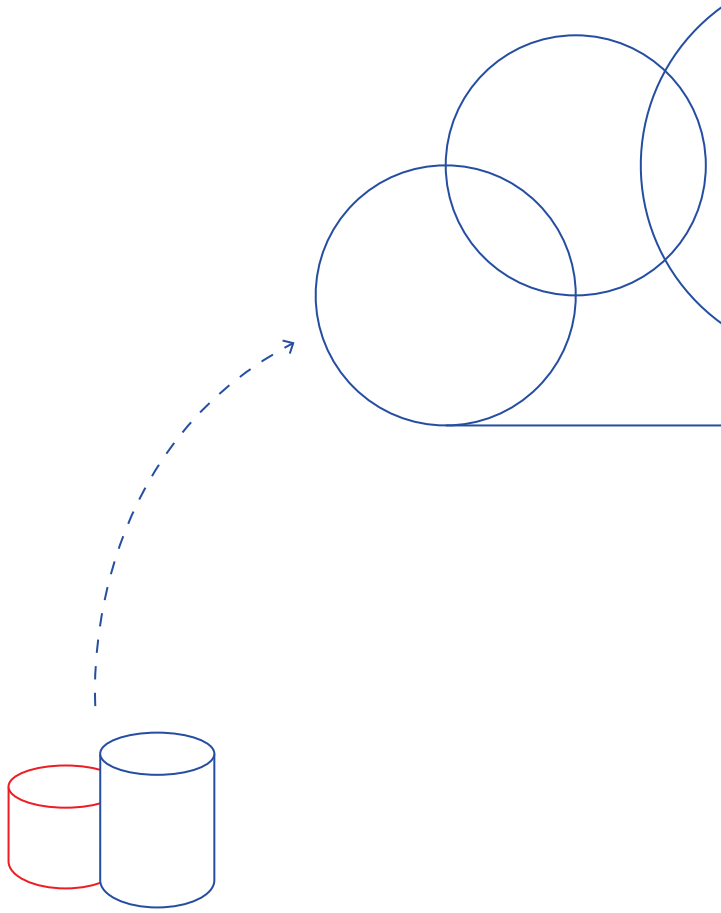


Take the Bold Leap—Move Your On-Premises Data Warehouse to the Cloud.

Data warehouse migration to cloud may occur as part of a broader corporate initiative to migrate to cloud, or as a dedicated data warehouse modernization initiative. At times, this effort can result from a confluence of the two trends. With migration, businesses may decide not to move all their data to the new environment. Some data may be left behind or archived if no longer needed. Similarly, when dealing with legacy applications or antiquated processes, savvy businesses find this an opportunity to migrate only what is needed and discard the rest. Additionally, this can be an opportunity to "clean house" and improve data quality as part of the data migration exercise, which is also where solid data quality tools can come into play.

A robust data migration and data integration strategy is critical especially if you are moving a lot of data, so select your data migration and integration solutions carefully. Diligently plan the initial data migration, utilizing data management tools to profile and understand all of your on-premises data, but also consider on-going data integration requirements, maintenance, and accessibility.

Despite moving your data warehouse, in most situations you will still have a hybrid cloud environment with on-premises and cloud based data sources. This is where data integration solutions yield the greatest benefits; consider both on-premises data integration and cloud-based iPaaS (Integration Platform as a Service) solutions for your hybrid environment.



Take the Bold Leap—Move Your On-Premises Data Warehouse to the Cloud.

Organizations with many years of investment in on-premises data integration solutions may choose to migrate their on-premises data integration solution to AWS, along with migration of the on-premises data warehouse to Amazon Redshift.

Also, consider performance—how long will it take to migrate the data and once there, how long will it take your ETL tools to load data into the tables? Does the target software stack support parallel processing and deliver high performance? Make sure your data integration solution can leverage the strengths and features of Amazon Redshift.

What is iPaaS?

Integration Platform as a Service (iPaaS) is a cloud service that gives IT a single platform from which to manage application, data, and process integration. It allows enterprises to rapidly execute any integration pattern, logically manage any data, and easily serve any user in need of integration.

Read our eBook "The Cloud Architect's Guide to iPaaS" for a complete definition, as well as common enterprise use cases for iPaaS.

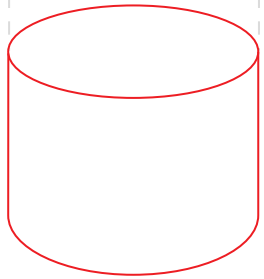
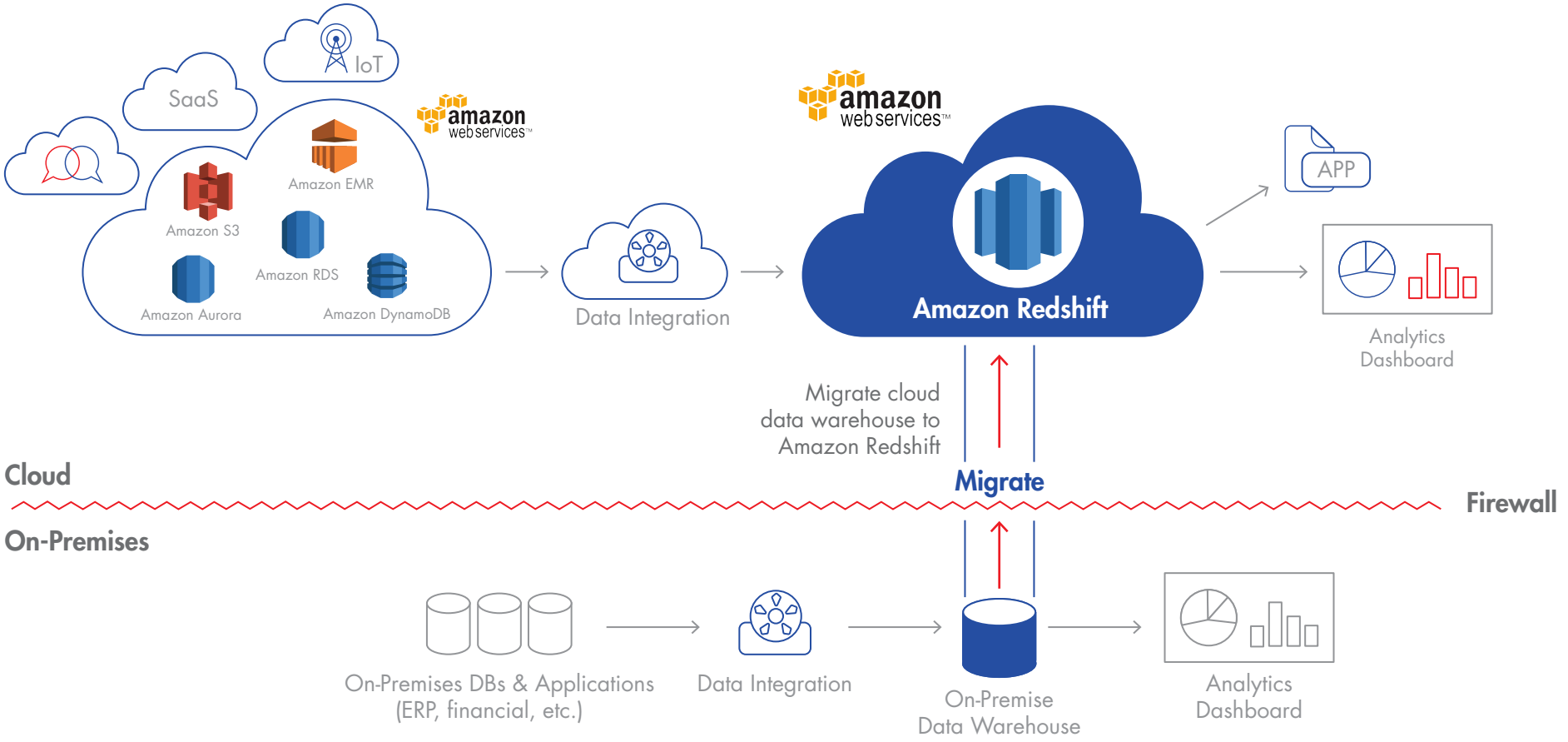
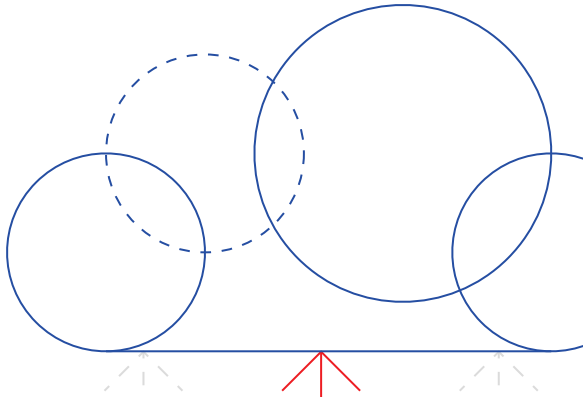


Figure 3.4 — Migrate (Lift & Shift) into Cloud Data Warehouse



Take the Bold Leap—Move Your On-Premises Data Warehouse to the Cloud.



About This Architecture.
This migration scenario embraces the full benefits and capabilities of the cloud. Once migrated, the cloud data warehouse is still accessible to on-premises assets, but it is poised for maximum integration with modern cloud applications, data stores, and analytics.

Using Amazon Redshift, a migrated data warehouse represents a powerful asset as it allows you to:

- Support cloud analytics efforts with new technologies very quickly.
- Provide flexibility and easy integration with other cloud components (SaaS applications, data, analytics).
- Scale up and down as needed, providing elasticity not possible with on-premises solutions.
- Reduce complexity and increase security and modernization due to integrated security and upgrade policies.
- Lower initial stand-up and on-going sustainment costs compared to supporting on-premises solutions. The cloud model of "pay only for what you use" is compelling to new and established companies alike.

Organizations may consider cloud data warehouse migration after testing out hybrid cloud computing and/or smaller new cloud data warehouses. Once an organization has cloud experience and encounters first-hand the lower costs and increased agility, they're much more willing to embark on a cloud data warehouse migration where the benefits are magnified.

Hybrid Computing for the Foreseeable Future.

Pure, "cloud only" computing environments without any on-premises footprint do exist, but they are less common in existing enterprises and are most frequently used by new companies that have never established an on-premises environment.

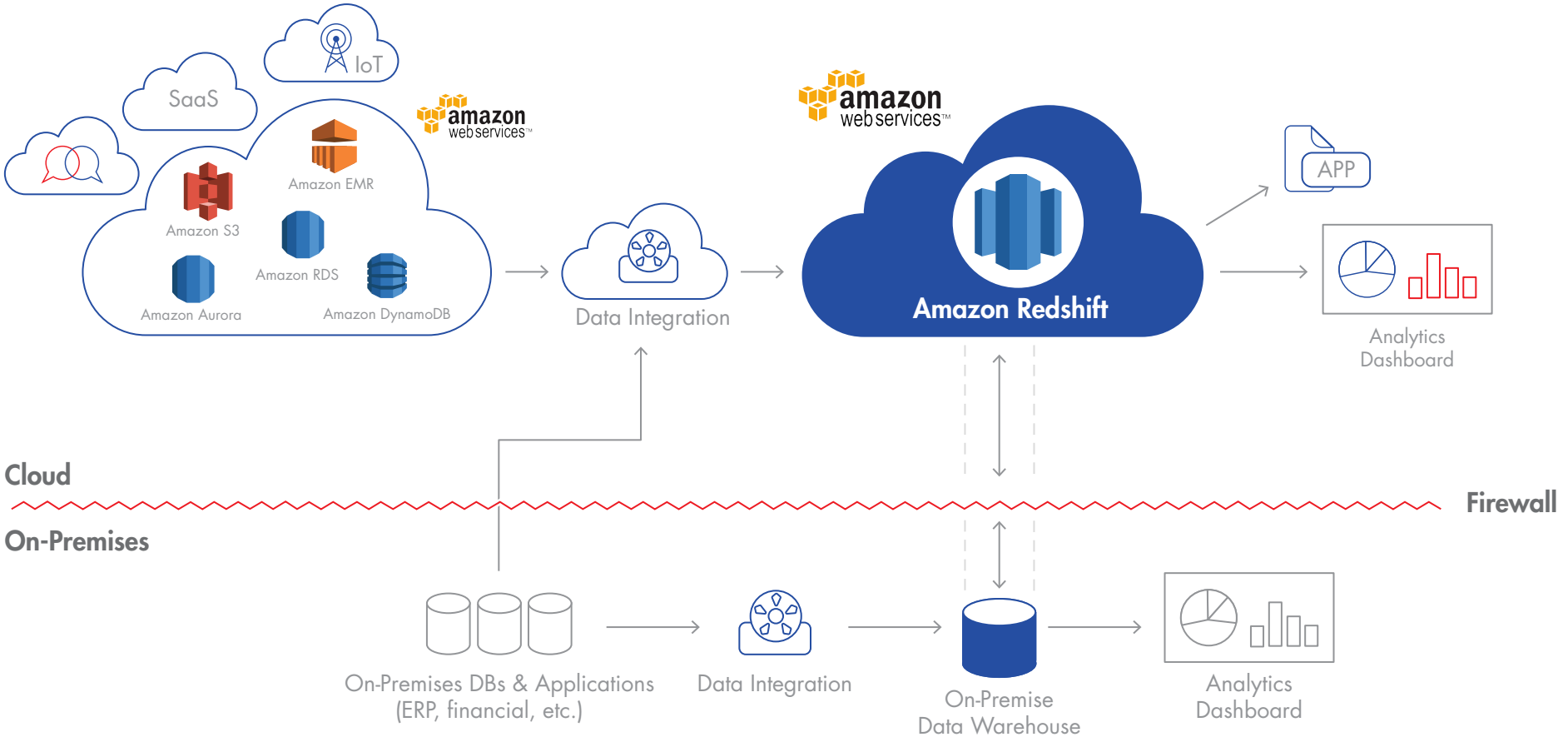
Most organizations, particularly medium to large size, have a substantial, existing investment in on-premises computing infrastructure, processes, and support, and these will not go away overnight. Furthermore, there are some special cases where the cloud is not the best fit. This is true for highly complex, highly custom, or specialized systems that don't fit in the standard, commodity-based model that powers cloud computing.

Despite compelling use cases and drivers for cloud computing, on-premises environments will endure for some time at certain companies, and so too will hybrid computing. It may take decades before the last on-premises workloads are migrated to the cloud or decommissioned; just look at the number of mainframe COBOL systems still in existence today.

The result is hybrid computing—on-premises integrated with cloud assets—which must be effectively supported for years to come. On-premises data integration and cloud-based iPaaS solutions play a critical role in hybrid environments. When you leverage a comprehensive data management platform that includes on-premises data management as well as iPaaS capabilities, you will enhance and simplify your data management activities.

In Figure 3.5 we see the hybrid cloud computing environment that will be common for some for years to come. These environments will typically include SaaS and on-premises applications and data sources on-premises and in the cloud. Expect to see cloud data warehouses such as Amazon Redshift as well as on-premises data warehouses integrating with these data sources and also with each other.

Figure 3.5 — Hybrid Data Warehousing



Part 4:

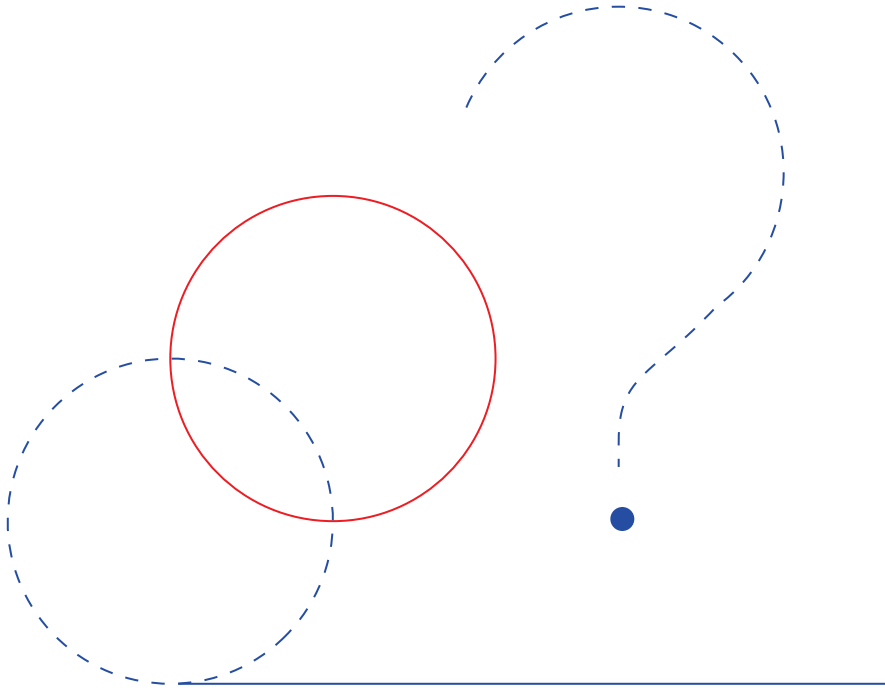
Your Path Toward Cloud Data Warehousing.

Questions for Your Cloud Data Warehousing Projects.

In this section, we're going to help you identify what you need to consider as you develop cloud data warehouses with Amazon Redshift via series of questions.

Most of the questions are applicable to any use case (new cloud data warehouse, data warehouse extension to cloud, and data warehouse migration) although some topics (such as bursting) are more applicable to a particular use case. Regardless of your specific use case, we invite you to work through all the questions to enhance your perspective and generate new ideas.

These questions aren't intended as a 1-2-3 stepwise install guide, but rather a way to identify and consider key points as you move forward. So grab a pencil and let's start planning.



Searching for Your Initial Opportunities.

Amazon Redshift is great for new cloud data warehouse projects where you need agility to start small, try out new analytic solutions, conduct some experimentation, or pursue a time-sensitive business opportunity—all at a lower start-up cost (time and money) and risk.

Tip: This list is free-form and open ended; we want to unleash your creativity. List out some initial ideas yourself, but also brainstorm with your colleagues to capture a wider scope of ideas. Set the baseline understanding for what we are looking for, but then dive in using whatever brainstorming techniques you prefer. Don't fall into the trap of trying to write full project charters or get bogged down into preconceived notions of "we can't"—we are just seeking some initial targets of opportunity.

List out ideas for your first Amazon Redshift projects based on these thoughts:

I've wanted to test this out for a while…

What is a minimum viable analytics prototype that we could kick-start this project with?

We think there may be some new opportunity for improved analytics with this idea, if only we could stand up a prototype environment…

We've been wanting to experiment with this but haven't due to hurdles of time, money, or infrastructure…

Identifying Your Stakeholders and Their Needs.

Any successful project needs to identify those people or groups impacted by the effort—these are you stakeholders. The impact may be good or bad (we hope good!) and not everyone is impacted to the same degree.

Creating a list of your stakeholders, and especially your key stakeholders, will ensure your impact analysis, scope, and communication plans are effective.

Once you've identified your stakeholders, you need to understand what data they need and define their expectations (you'll need to communicate with them—often). The cost of change only increase in terms of money, time, and complexity as a project progresses—good stakeholder communication can help manage (but not entirely eliminate) change.

In addition to good communication, empowering controlled self-service for stakeholders increases their effectiveness and reduces your workload and complexity. Where possible, self-service capabilities should be embraced.

> **Tip:** This isn't a project management guide per se, but it doesn't hurt to familiarize yourself with some project management fundamentals to increase your effectiveness.

Who are the people and groups that will most directly benefit from this work (key stakeholders)? Also, who will be indirectly impacted (positively or negatively)?

Identifying Your Stakeholders and Their Needs.

What is it that your stakeholders are looking for? What do they absolutely need (requirement) and what do they want (wish list)?

How will changes from your stakeholders be communicated, how often, and do they understand the balance of scope vs cost vs time? Use disciplined change management, well-defined expectations, and (if necessary) project sponsor involvement to prevent run-away scope creep that will torment any project manager (and their schedule!).

What can you do to provide self-service for your different stakeholders based on their skillsets? For example, can you add self-service analytics and self-service data integration for your citizen integrators? Where can you implement standardization and automation to enhance speed and repeatability (while also reducing cost and complexity)?

What about customers as stakeholders? How could this benefit your company's "stakeholders"? And ultimately your company?

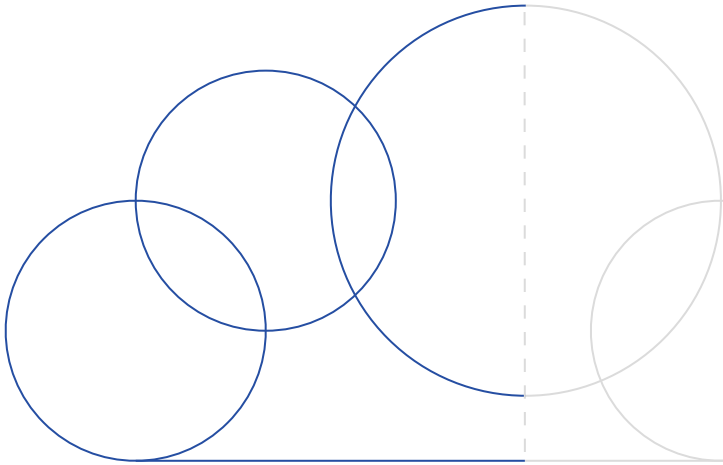
Migrating and Integrating Your Data.

Your analytics are only as good as the data that fuel them; having trusted and timely data available is key to your success. That success starts with first identifying what data you need and where it is, and then knowing how to get it.

Next, you must understand if you are migrating or integrating your data. Is there a difference between data migration and data integration? Yes! Below are some definitions in the context of cloud data warehousing.

- Data migration is a onetime event to move data from a data source to a target (Amazon Redshift in our case). Note that the data is actually moved, not merely copied. Once complete, you decommission the original data source.
- Data integration is an ongoing process of combining data, often from multiple data sources, into a target data warehouse, commonly also transforming the data in novel ways, in order to support analytical needs. During the integration process, records from the source are mapped to the records in the destination. The data remains in both its source, but now also in the target. Data integration does not result in decommissioning the original data source.

Visit the Informatica data integration web site for a deeper dive into this topic.



Migrating and Integrating Your Data.

Data Migration

Consideration must be given to how your data will be loaded and how long that process will take. Moving away from manual data loads to automated methods makes your life easier so leverage data management solutions to simplify and automate these tasks.

Moving large amounts of data over the network can take a long time. One technique is to compress the data before you move it. The disadvantage to this approach is compressing and decompressing is itself CPU-intensive and takes time, but by leveraging the massively parallel processing (MPP) capabilities of Amazon Redshift, it is still faster overall to compress then decompress data, rather than trying to move it all uncompressed.

Another technique is to move and load data in smaller chunks or files in parallel rather than as one large data file serially. This approach further maximizes the MPP architecture of Amazon Redshift.

Some organizations see a data warehouse modernization as an opportunity to "clean house" and start a new slate. This involves extensive profiling and cleansing of the data as well as data purging and/or archiving prior to migration.

As organizations develop a plan to migrate data from multiple sources to AWS, they may need to utilize data cataloging tools, which enable teams to fully understand the data and associated metadata and determine what will be moved over. This is an exercise that you may wish to undertake jointly with your business stakeholders.

You should also conduct user acceptance testing once your data warehouse has been migrated to ensure that your business stakeholders are satisfied that all the necessary data and reports have moved over as expected.

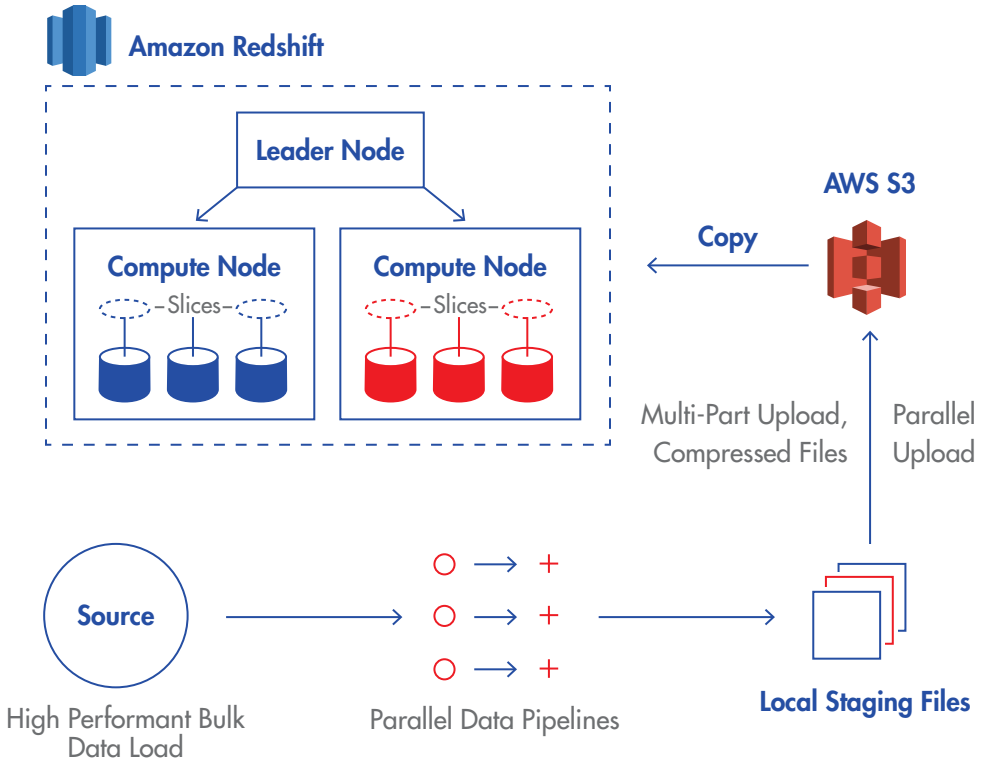
The diagram titled **"Fast Loading Data into Amazon Redshift"** (as can be seen on page 46) conceptualizes compressed parallel data loads from the source to a staging directory, then into Amazon S3 storage, and then that data is copied into the Amazon Redshift data warehouse. As data is loaded into Amazon Redshift using the COPY command, multiple files are loaded in parallel and Amazon Redshift automatically optimizes data distribution, indexing (encoding), and compression operations.

Migrating and Integrating Your Data.

Data Synchronization

Moving and keeping your data updated is a multi-phase process: first you have your initial data load (migration), but after that, ongoing data synchronizations occur to keep the target system current. The data synchronization tasks can upsert data from multiple sources into your target data warehouse to keep the data fresh and relevant. This is ongoing and these operations should be automated. Periodic changes in underlying data sources or appearance of new data sources will often occur, requiring maintenance upkeep of your dependent code and APIs. If you are leveraging quality solutions and automation will obviate the need for a lot of recurring maintenance work to do in response to these inevitable changes.

Figure 4.1 — Fast Loading Data into Amazon Redshift



Migrating and Integrating Your Data.

Data Integration

ETL vs ELT – what is the difference and why do you care?

ETL (Extract, Transform, Load) is used for high-speed transfer of huge volumes of data outside of RedShift (Teradata, Netezza, Oracle) into RedShift. Once you land the data into RedShift Staging Tables, further data processing needs to be done to transform the data into facts and dimensions for Analysis. The data transformation between staging, intermediate, and analytics RedShift tables can be accomplished in a highly performant way using SQL-based **pushdown optimization (PDO and also known as Extract, Load, and Transform - ELT)** where in the transformation is converted to SQL predicates and pushed down to SQL queries.

Extract, Load, and Transform (ELT) is an alternative approach. With ELT, you Extract the data and then Load your data to a target staging table in its raw data form. At this point, the data is then Transformed within the staging table and then moved to the target data warehouse. With ELT, the process of transformation occurs on the target server environment, utilizing native performance capabilities within the target data system with a larger, raw data set.

ELT is an approach well suited for large volumes of data in the cloud as is common with all of our use cases with Amazon Redshift. Consider the expansive S3 storage size for raw data, MPP processing, and the scalability power of Amazon Redshift architecture; it is built to handle the heavy Transformation workload with ELT processing.

Migrating and Integrating Your Data.

Push Down Optimization (PDO) is another concept to understand for data integration. PDO allows transformation logic to be pushed down to either the source or target server to optimize performance. For cloud data warehousing, it makes sense to use PDO to the target server in support of ELT optimization to leverage the strong processing capabilities of the cloud infrastructure.

Generally, your ELT data load path will look like this:

1. Bulk loading data into Amazon Redshift staging tables.
2. Move that data into intermediate tables for transformation processing.
3. Load that data into final analytics tables (fact and dimension tables) to conduct your BI analysis.

PDO is split between steps 2 and 3 above, with the transformation occurring on the target environment.

Tip: Data migrations and data integration can be complex; especially if you rely on in-house, manual solutions. Leverage powerful, prebuilt solutions and methods, which enhance and automate tasks to reduce complexity while accelerating your project.

Migrating and Integrating Your Data.

Consider your data needs for production:

a. List your data sources required to integrate into your data warehouse

b. Will some of your data sources be external to your organization?

c. Will your data integration need to work across both cloud and on-premises data sources and targets?

d. Do you have access to all these data sources?

e. Do you have a good understanding of the data from each data source?

f. What expertise do you have in-house to connect to each source?

g. How long does it take to onboard each data source?

h. How will you maintain independent and reusable data integration code as the underlying data sources and analytics requirements change?

i. How will you evolve your data integration as business needs change?

j. What data warehouse changes will be managed by developers and what can you enable your citizen integrators to do themselves.

k. What is your expected quality of data? Will you need tools to manage and monitor data quality?

l. Are there any other data production issues for you to consider?

Migrating and Integrating Your Data.

Plan your data integration:

a. How are you addressing data integration?

b. Manual or automated and what is the time and complexity involved? Consider the resources you have today and in the future.

c. How does this impact your agility and what can you do to optimize your processes?

d. How are you scaling out your data integration projects as complexity of your projects scales?

What data are you migrating and what are you leaving behind and/or archiving? Consider both your data sources and your data warehouse data. Can you migrate data from multiple systems so you decommission multiple data sources (databases)?

Prior to migration, data preparation is critical. What data profiling and cleansing procedures and tools do you have in-house? Data cataloging tools enable you to understand your data and the associated metadata to determine what to migrate – do you have these tools and expertise? Have you considered that some data should be purged or archived rather than migrating it?

Migrating and Integrating Your Data.

Think about each of these data migration factors:

a. What data management solutions will you use for this data load?

b. Manual or automated and how long do you think it will take?

c. Will this time impact your overall project and if so, what can you do to address that?

d. How will you gain business user approvals post migration?

e. What are your data management Service Level Agreements (SLAs)? Do you have a need for real time data updates to support business requirements?

f. Do you plan to use an Operational Data Store (ODS) technology to gather the data prior to loading into the data warehouse?

g. How complex would be the process of joining and merging the data? What tools do you plan to use?

h. How often do you expect to have to add new data sources and types to the data warehouse?

i. Will you need periodic data synchronization between this data warehouse and other systems, cloud and on-premises?

j. Are there any other data production issues for you to consider?

Maximizing the Power of Amazon Redshift.

Amazon Redshift together with other data-oriented AWS services gives you the ability to make your data warehouse incredibly powerful. How can you leverage these technical capabilities to benefit your data warehousing projects?

Several characteristics of Amazon Redshift are particularly useful here. MPP promotes data loads across a wide highway of processors, much faster than a string of cars on a two-lane road. Architect your data integration and analytic processing to leverage parallel processing across multiple compute nodes.

Powerful processors, vast memory, and fast storage I/O enable analytical processing on a large scale. The elastic cloud platform provides more computing resources to scale up to meet processing requirements and scale down when not needed. These capabilities aren't available with traditional on-premises computing solutions.

Amazon S3 storage is virtually unlimited, highly durable and low cost, which are important considerations for data warehousing. Loading your raw data into S3 storage and then performing transformation steps on top of this cloud infrastructure maximizes the incredible compute power of the cloud while providing low cost storage for data.

Good data management and integration solutions complement the inherent power of Amazon Redshift and AWS. Mature data integration solutions simplify access to the many data sources accessible via AWS and on-premises data centers, and then automate their access. High performance data management solutions provide native automated connectivity to key AWS data services such as Amazon Redshift, Amazon S3 storage, Amazon EMR and Amazon Aurora, Amazon DynamoDB, and Amazon RDS databases. Your data management platform should have the capability to manage and integrate data at petabyte scale; thus scalability is a key feature to look for in a solution.

Maximizing the Power of Amazon Redshift.

Additionally, leverage metadata-driven, visual role-based tools to deliver data visibility and improved productivity, reduced risk, and improved security and governance. You will find increased operational confidence by having a single point of control for end-to-end data flows. Finally, ensure your data management platform is certified to run on AWS to ensure it maximizes full capabilities of Amazon Redshift.

Tip: Understanding key components of AWS and how Amazon Redshift actually works will only make you more effective as you pursue your data warehousing analytics projects. You don't need to be a technical expert, but a reasonable investment in time to learn core fundamentals will yield dividends as you move forward.

For more information about Amazon Redshift, visit http://aws.amazon.com/redshift/

How can MPP processing of Amazon Redshift aid your initial data load and reduce your load time? How can parallel processing of Amazon Redshift enhance your performance?

Analytic processing can be intense—how can you leverage elastic cloud computing resources and modern infrastructure to fully enable and compliment these activities?

Maximizing the Power of Amazon Redshift.

Cloud storage, particularly Amazon S3, is low cost, reliable, and expandable. Amazon provides storage tailored to three different use cases: S3 Amazon Standard storage for normal, online operations, Amazon S3 Infrequent Access for less frequent access operations, and Amazon Glacier for archive data. How can you leverage the different levels of storage for your Amazon Redshift warehousing needs?

You must connect to all your data sources, but this process does not have to be complex. Does your data management solution easily connect to all your data sources within AWS, the cloud, and on-premise? Are these connections out-of-the-box, easy to maintain, and high performance or are they manual and cumbersome?

Your data management solutions should work in harmony to compliment Amazon Redshift. Is your solution metadata-driven using role-based tools? Does it provide data visibility and a single point of control for end-to-end production data-flows? To fully leverage the power of Amazon Redshift, your data management platform must be highly capability and well-integrated into AWS.

Bursting to Meet Temporary Workload Surges.

Amazon Redshift is built on a robust infrastructure that allows for rapid addition of storage, memory, and processor compute power in times of need.

When no longer needed, these resources are released back into the cloud to be used by another customer. This is scaling up and down of resources is the definition of elasticity and is a major benefit of cloud architecture.

Cloud bursting is an innovative way hybrid cloud environments solve the challenge of surge workload, such as during holiday sales events. On-premises computing occurs until a surge event occurs. At that point, workload processing is extended to the cloud environment, i.e. bursting into the cloud.

A unified data integration platform simplifies your bursting of data into the cloud. Additional compute and storage resources are added to the cloud environment as needed to meet the processing demand. Once the demand subsides, resources are released and the cloud shrinks back to normal size. Of course, this also holds true if your data warehouse is fully housed in Amazon Redshift. You can still burst your workloads by temporarily increasing your Amazon Redshift cluster size and then scaling it down again.

Tip: Don't simply assume your workload will automatically take advantage of cloud bursting. Do your homework and test to ensure the bursting workload is processed as expected.

Reuse Your Integrations.
When you use an automated tool that's supported by a robust connectivity architecture to load your on-premises data warehouse, you can easily reuse the integrations to integrate the same data to an Amazon Redshift data warehouse. Because the integration logic is separated from the underlying sources or targets, there is minimal change needed to adapt the code (or data mappings) to a new data warehouse.

Bursting to Meet Temporary Workload Surges.

What workload surge events occur within your organization? Are they predictable and well planned or are they unpredictable and disruptive?

What compute (memory and processor) and storage aspects of Amazon Redshift will be most needed during your next workload surge?

What is the cost of not reacting to a surge workload in terms of lost opportunity or failure to meet a regulatory requirement? If your on-premise resources are sized for peak workload, how much capacity is unused during normal operations?

How will you manage the data integration aspect of cloud bursting? Can you seamlessly move the same data integration processes from your on-premise data warehouse to the cloud data warehouse with minimal effort?

Reducing Your Risks and Promoting Success.

Leveraging Amazon Redshift for cloud data warehousing to support analytics yields many benefits and, with the right data integration solutions, is fast and relatively easy.

That said, there are some tips that improve your odds of success. As the core of cloud data warehousing is the data (surprise!), consider these tips within that context.

This last series of questions is purposefully light on the technical education aspect of cloud data warehouse concepts. Rather, we want you to think through your processes and their impacts before you implement them.

Tip: It may sound like a cliché, but "failing to plan is planning to fail" resonates with anything you do in IT. Perform your own independent research and due diligence, talk to people with more experience, and follow sound project management principles.

How will you validate that all your data has been moved? What prepared validation checks and tools do you have in place?

Reducing Your Risks and Promoting Success.

Do your key stakeholders believe that all of the data has been moved? How do the stakeholders define success and have you managed their expectations?

What is the impact to the organization of moving its production data store? Think in terms of technical, but also non-technical governance issues. Cloud computing is secure, but are there regulatory or contract restrictions on the placement of some data elements?

Are you using the right tools and automation? Are the data management and integration solutions working together as a complete solution or are they fractured and stove-piped? Consider what features in a data management platform benefit you the most (e.g. connectivity, ease-of-use, rapid integration, complete solution, scalability etc).

Part 5:

How the Informatica Platform Empowers Amazon Web Services and Amazon Redshift.

Enabling Success with Informatica and Amazon Web Services.

Today's enterprises have a mandate to do more with less: accelerate innovation, while reducing IT expenditures.

With AWS and Informatica, your organization can increase IT agility and speed, pay as you go for resources when you need them, benefit from cloud economies and scale, and free up IT resources for strategic business initiatives.

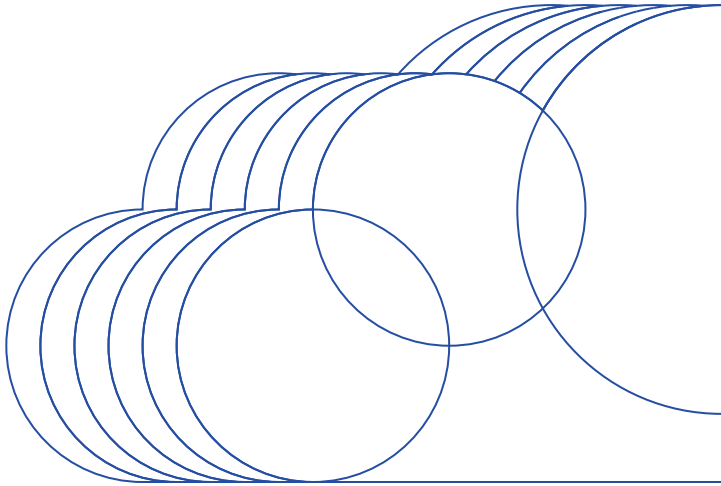
Informatica powers data management initiatives for a successful AWS cloud journey by delivering connected, trusted, meaningful data from cloud, on-premises, and big data sources.

Informatica's approach accelerates your Redshift data warehouse project deployment by automating your data integration development lifecycle, including connectivity, development, deployment and management.

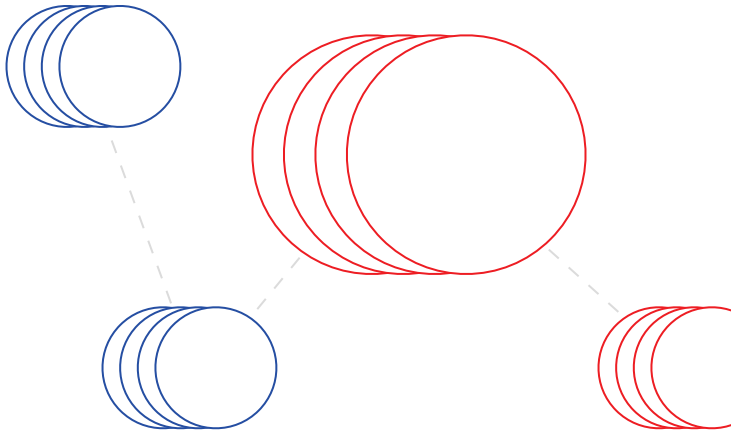
Data connectivity eliminates silos.

Optimize cloud data warehousing with Informatica's native Amazon Web Service Redshift connector that integrates data from both cloud and on-premises sources. Benefits include:

- Out-of-the box connectivity with data from hundreds of sources, including Amazon S3, RDS, Aurora, and DynamoDB; Salesforce, NetSuite, Workday, Marketo, big data, and social media sources.
- Reuse of integration code across sources and targets, effectively future-proofing your investment
- Connectivity as service, which relieves you from management of data sources' versioning and fast-changing SaaS releases
- Support for both ETL and ELT patterns for end-to-end high performance data pipelines.



Enabling Success with Informatica and Amazon Web Services.

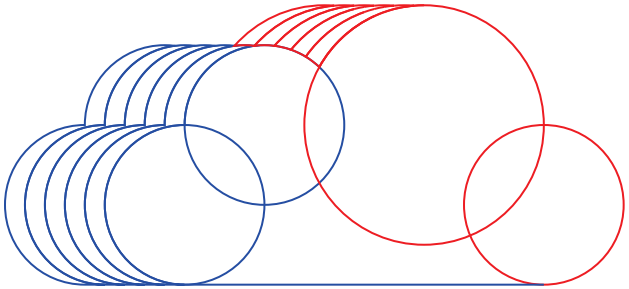


Development agility accelerates time to market.

Role-based SaaS development tools, optimized for both developers and business users are accessible anywhere, any time. Features include:

- A visual, easy-to-use user interface that automates tedious hand coding by techniques such as data 'upsert'
- Out-of-the-box templates and 300+ transformations, including aggregators, joiners, lookups and masking, which accelerate development
- Dynamic mappings, automatic target table creation, pre- and post-SQL processing, which simplifies the development lifecycle
- Hourly consumption software delivery via the AWS Marketplace, so you can rapidly kick start new Amazon Redshift project in minutes, not months
- No lengthy infrastructure and software
- No lengthy infrastructure and software procurement and installation
- A six-step Data Wizard that guides development by citizen integrators

Enabling Success with Informatica and Amazon Web Services.



Deployment scalability and operational confidence.

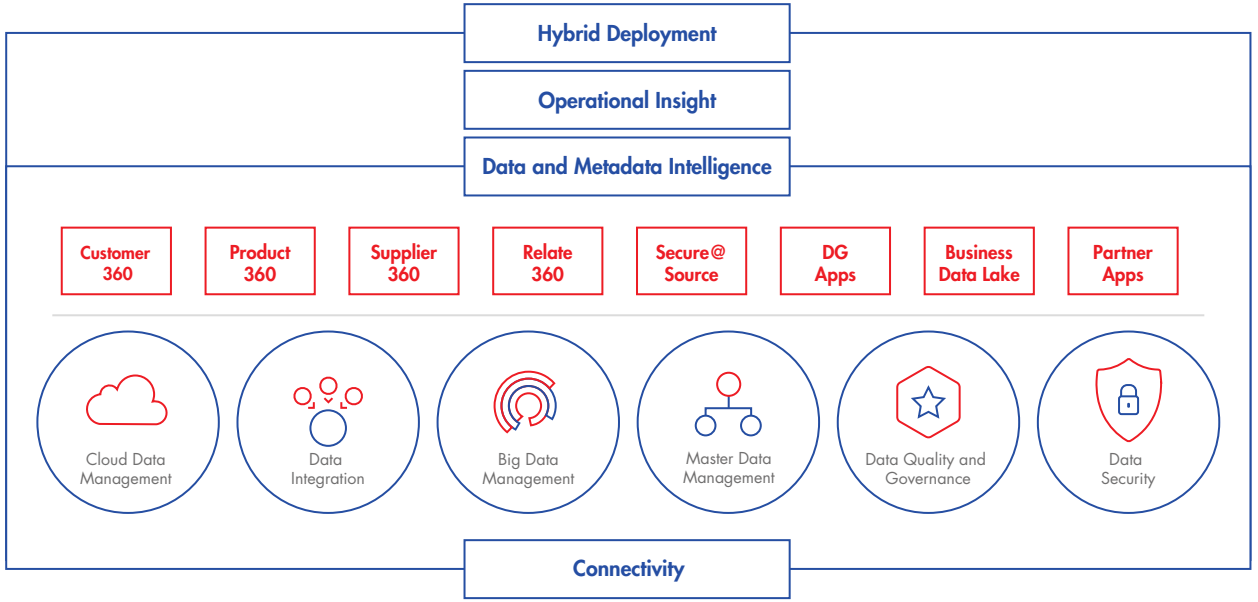
Pushdown optimization takes full advantage of Amazon Redshift's native processing, maximum scalability, and performance, including:

- Automatic partitioning for maximum throughput, with no developer pre-configuration required
- Automatic parallel loading of compressed files from Amazon S3 to Amazon RedShift cluster, which auto-scales to ingest massive data volumes and eliminates tedious programming
- Clustering of secure agents on a grid for maximum performance
- Data transformation optimizations such as cached lookups
- Comprehensive monitoring of production data workflows via user-friendly tools that deliver operational confidence to catch and correct production problems early, which improves data quality and timeliness

Enabling Success with Informatica and Amazon Web Services.

Informatica offers a comprehensive data management platform for AWS, including data integration, data lakes and big data management, MDM, data quality, and data security to accelerate AWS deployment and deliver trusted data in the cloud, on premise, and in hybrid environments.

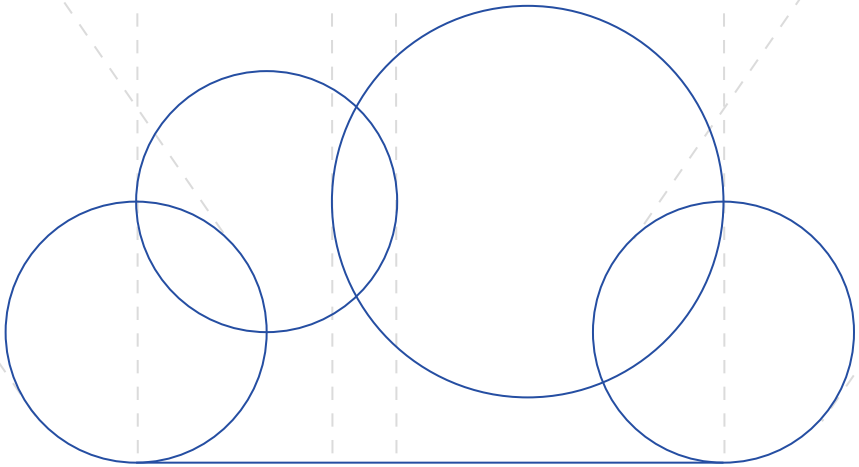
Figure 5.1 — The Informatica Intelligent Data Platform



Conclusion.

Conclusion.

Cloud data warehousing with Amazon Redshift powers analytics and agility, while reducing the support burden of on-premises computing. Informatica is your trusted data management partner as you deploy to Amazon Redshift and start your cloud journey.



Learn more

Take the Next Steps.

Now that you have a better idea on how to tackle your cloud data warehousing project with Amazon Redshift and the importance of a robust data management solution to support your efforts, here are concrete next steps you can take to advance your project.

Reference Architecture Guide

Cloud Analytics with Amazon Redshift and Informatica Cloud
Deep-dive on the technical implementation details of a cloud data warehousing solution with Amazon Redshift and Informatica Cloud.

Read more

Try & Buy Informatica Cloud for Amazon Redshift on AWS Marketplace

Get started today on a new data warehousing project or POC with Amazon Redshift and Informatica Cloud. Start a free trial or Pay as you Go with low hourly pricing on AWS Marketplace.

awsmarketplace

PowerCenter QuickStart on AWS

Benefit from 'single click deployment' of PowerCenter on AWS EC2, launched in your AWS environment and accompanied by detailed technical documentation and streamline PowerCenter deployment for Amazon Redshift projects.

Read more

Contact us to learn more.

About Informatica.

Digital transformation is changing our world. As the leader in Enterprise Cloud Data Management, we're prepared to provide you with the foresight to realize new growth opportunities or even invent new things. We invite you to explore all that Informatica has to offer—and unleash the power of data to drive your next intelligent disruption.

About AWS.

For 10 years, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud platform. AWS offers over 70 fully featured services for compute, storage, databases, analytics, mobile, Internet of Things (IoT) and enterprise applications from 33 Availability Zones (AZs) across 12 geographic regions in the U.S., Australia, Brazil, China, Germany, Ireland, Japan, Korea, and Singapore. AWS services are trusted by more than a million active customers around the world – 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 http://aws.amazon.com.

Worldwide Headquarters, 2100 Seaport Blvd, Redwood City, CA 94063, USA Phone: 650.385.5000 Fax: 650.385.5500 Toll-free in the US: 1.800.653.3871 informatica.com linkedin.com/company/informatica twitter.com/Informatica

Sources

[1]**InteropITX Research Reports,** 2017 State of the Cloud, January 2017

[2]**EMA, Informatica, Deloitte,** The State of Cloud Analytics 2016, February 2017

[3]**TDWI,** Best Practices Report: Emerging Technologies for Business Intelligence, Analytics, and Data Warehousing, October 2015