AWS Data & Analytics

Information and Data Management Solutions Generate Value

aws marketplace

Overview

Cloud Information and Data Management

In today's data-driven digital environment, companies are collecting, transforming and connecting data in innovative and meaningful ways. A robust information and data management solution enables you to leverage the power of your data, exploring and uncovering insights in documents, images and audio across billions of inputs and existing data assets.

Table of Contents

Cloud Information and Data Management

Our customers are implementing self-managing Data Lake and unified data management functions across services. In this e-book, we explore the possibilities of information and data management on AWS through two use cases: Cloud bursting and cloud data lab. For each use case, we outline the specific steps you can take to implement AWS and solutions available in AWS Marketplace when managing your organization's data and striving towards automation.

Generating Value from Data & Analytics

As businesses become increasingly digital in nature, the growth of applications and mobile apps has caused a rise in complex hybrid ecosystems of data. IT administrators are driven by business needs to combine cloud and on-premises deployment models to give their organizations the advantage they need to compete in a fast-moving digital economy. Improving the scale and flexibility of data integration across both cloud and on-premises environments to deliver a hybrid offering is essential to providing the right data to the right people at the right time.

According to Tolga Tarhan, CTO of Onica, a fast-growing Amazon Web Services premier consulting partner, data analytics automation can make a big difference for your business: *"You can answer some limited questions and make some limited predictions with little data. But, if you're not collecting and storing a larger stream of data, then you run the risk of not being able to answer tomorrow's questions."* Automating your data platform so it is always optimized means you can make timely decisions based on relevant data.

Whether you have an on-premises, cloud-native or hybrid infrastructure, integrating, augmenting and automating your big data infrastructure with Amazon Web Services and solutions available in AWS Marketplace can help you better collect, integrate, secure and manage data to enhance decision making. With AWS you can automate time-consuming administration tasks such as hardware provisioning, database setup, patching and scaling up and down.

Database and Data Migration

Data is the cornerstone of successful cloud application deployments. Your evaluation and planning process may highlight the physical limitations inherent to migrating data from on-premises locations into the cloud. AWS offers a suite of tools to help you move data via the network, through innovative technology partners and even by physically shipping a data storage device. For example, integrating your AWS instance with Snowball offers petabyte-scale data transport while addressing common challenges associated with large-scale data transfers.

Cloud data migration can be a daunting task. How do you safely and efficiently move data from your current location to your AWS cloud infrastructure with minimal disruption, cost and time? What is the smartest way to actually move your data, whether it's gigabytes or petabytes? IT professionals typically choose one of two ways to accomplish this task. Either they use very basic, unmanaged migration tools or they select one of AWS' suite of data migration services.

The suite of migration services from AWS includes different methods that help you efficiently manage the task. If you are moving large archives, data lakes in situations where bandwidth and data volumes are limited, there is a suite of AWS tools to make the process easier.

Data replication solutions available from vendors in AWS Marketplace give your infrastructure additional connections. Working with these solutions not only speeds up data replication, but it can also drive down your costs per transfer. The most basic solutions include network multiplexing. Intermediate solutions offer replicating and storing dev/test and production applications while minimizing your storage footprint with data deduplication and advanced compression. Other, more advanced replication solutions include those that specialize in backup and disaster recovery architectures like geo-distributed data consistency and “pilot light” environments. Whatever your needs may be, these solutions enable rapid recovery of your IT infrastructure and data.

Managing Data in a Data Lake

Businesses of all sizes are moving to a data lake infrastructure for their data. A data lake is a centralized repository that allows you to store all structured and unstructured data as is, which allows decision makers and analysts to use the tool and approach of their choice to answer business questions. When your data lake is built on AWS, you have an incredibly cost-efficient and durable solution, with access to native AWS services like Amazon EC2, EFS and S3 and AWS Marketplace solutions like Informatica PowerCenter and Teradata Database. When managing a hybrid environment of multiple ecosystems and applications, AWS is the ideal place to house your data and achieve data management and analytics power, while ensuring flexibility and cost-efficiency.

It's because of these benefits that many IT organizations choose to build their data lakes on AWS. A cloud-based data lake complements existing data warehouses and enables you to integrate multiple data sources, explore unknown datasets, discover new insights and analyze structured, semi-structured and unstructured data. The AWS data lake Quickstart automatically configures AWS services that are necessary to tag, search, share and govern specific subsets of data across a company or with other external users.

Tips from the Pros

Real-World Advice from Experts that Live Data Everyday

Ronald van Loon, Social Media Influencer

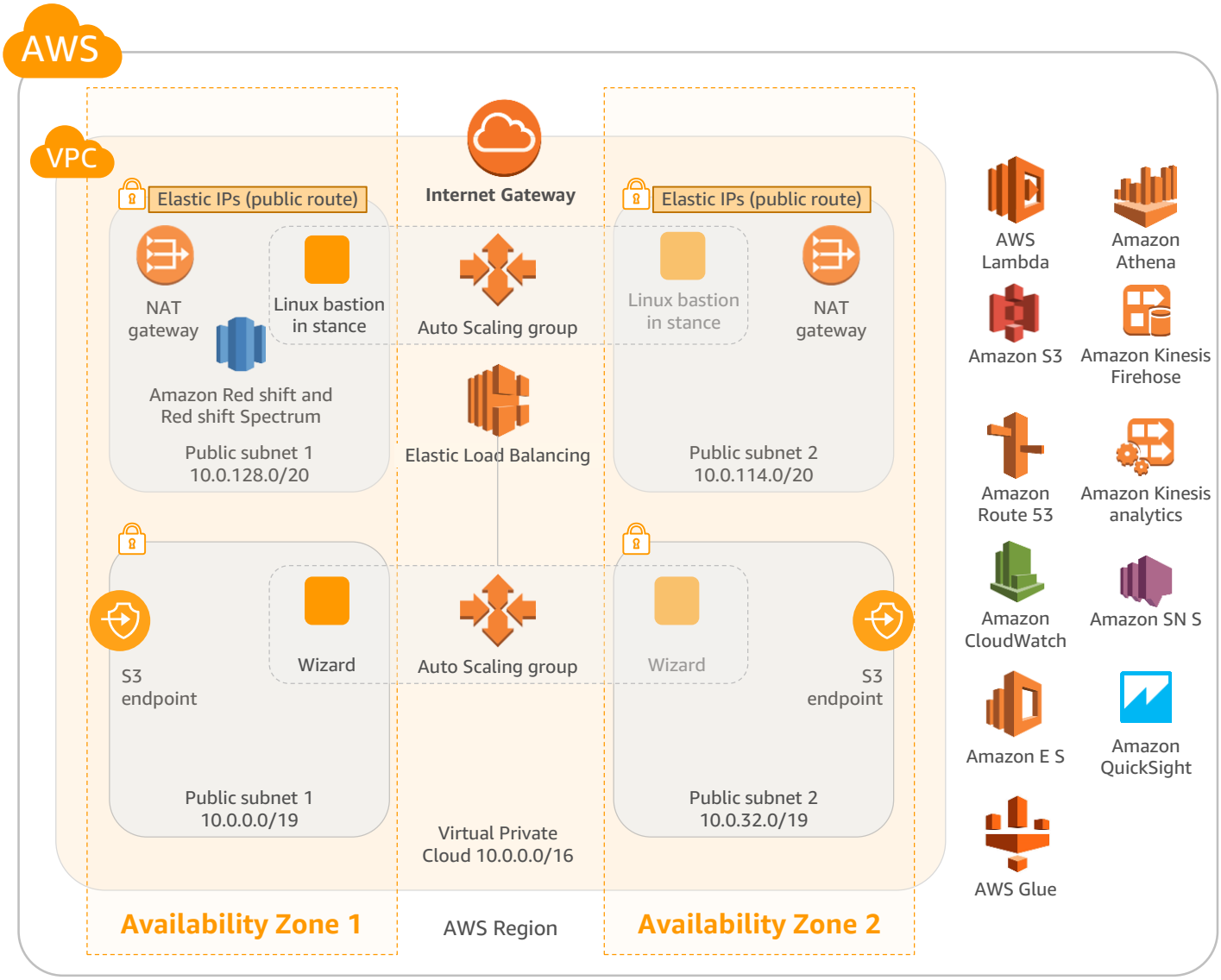
Adversitement delivers data management solutions and customer insights to optimize the customer experience.

It's simple: Being data driven is the difference between remaining competitive, as the entire world transitions to digital, or being left behind. Data analytics and big data are considered advantageous by nearly 91% of executives, and positively impacts revenues for at least ⅓ of Global 2000 businesses.

However, organizations struggle because of a lack of agility and the ability to make the right decisions based on all of their data. A data management platform provides a competitive edge, and should have five service layers: engagement, integration, development, data and modern core IT, which are all key components of every digital business.

Ronald van Loon, Director, Adversitement. @Ronald_vanLoon

Figure 1: Data Management on AWS



The above diagram is a standardized architecture that can be deployed to integrate AWS services such as Amazon Simple Storage Service (Amazon S3), Amazon Redshift, Amazon Kinesis, Amazon Athena, AWS Glue, Amazon Elasticsearch Service (Amazon ES) and Amazon QuickSight. The data lake foundation uses these AWS services to provide data submission, ingest processing, dataset management, data transformation, aggregation, analysis search, publishing and visualization.

The deployment includes an optional wizard and a sample dataset that is loaded into Amazon Redshift and Kinesis streams to demonstrate data lake capabilities. The reference architecture is automated by AWS CloudFormation: Runnable templates that you can customize to meet your specific requirements.

Amazon Redshift Case Study

A Siemens Power and Gas division was looking for ways to accelerate its data-analytics and customer-digitalization projects. After choosing Amazon Redshift, it wanted to build out the solution with best-in-class software. A search through AWS Marketplace led Siemens to run a free trial of Matillion ETL for Amazon Redshift, which took less than one hour to try, buy and implement within its existing AWS environments. For more about the story, view the customer success story.

Tips from the Pros

Real-World Advice from Experts that Live Data Everyday

Presidio

Presidio is an AWS advanced consulting partner paired with AWS-certified professionals and support services to help clients with digital infrastructure, cloud and security solutions.

You can't use traditional analytics and application platforms if you want to do anything real time. The reason why businesses are using cloud-native application and stateless application delivery is because you have to get away from the limitations of latency.

The best way to go from more traditional enterprise data warehouse and business intelligence to moving more toward managing data differently is to have the capabilities of doing more predictive and real-time analytics.

Robert Kim, Director – Digital Solutions, Presidio

Use Case #1: Implementing Cloud Bursting on AWS

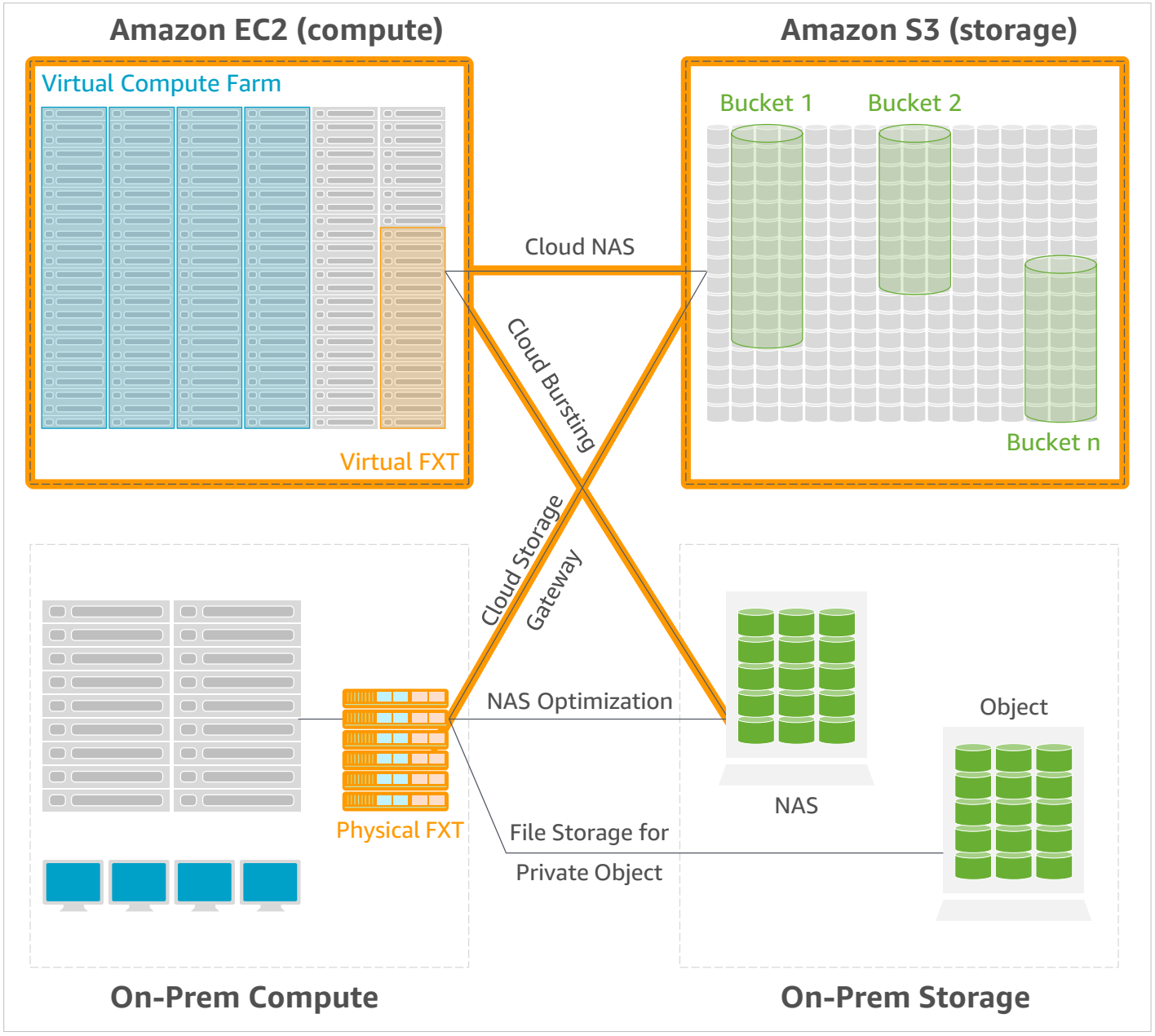
When evaluating IT infrastructure options for big data workloads, the AWS cloud and on-premises environments are not mutually exclusive. You can reap the benefits of cloud services including ease of deployment, reduced maintenance and lower costs while maintaining the use of on-premises resources based on your requirements.

Cloud bursting makes the most of cloud and on-premises environments. When your data center reaches capacity, directing the overflow demand to a public cloud is an efficient and economical approach. Processing can burst to AWS where near-infinite capacity and performance absorbs the increase in workload during processing spikes when needed. Services are uninterrupted and resources are paid for only in accordance with use, providing both elasticity and cost benefits.

A cloud bursting architecture addresses these common challenges associated with managing data:

- Pressure from the growing volume, variety and velocity of data.
- New data types that legacy systems aren't capable of capturing.
- Complexity and lack of flexibility from data warehouses.
- Lack of ease and trust in data backup and disaster recovery solutions.
- Data quality concerns without one version of the truth.
- Limited or no repeatable processes for information management.

Figure 2: Cloud Bursting with AWS



Implementing Cloud Bursting with Solutions in AWS Marketplace

Organizations are running data integration and data warehouse solutions on premises, though operating in a hybrid environment with AWS would offer cost benefits. There is an active need to satisfy business demand and service level consistency with dynamic capacity scaling by augmenting on-premises resources with a temporary or permanent cloud-based environment. This can happen, for example, with end-of-quarter financial reporting or when establishing a new subsidiary. AWS and seller solutions enable you to maximize these cost benefits for your cloud-bursting needs. Popular AWS Marketplace solutions include Informatica PowerCenter for data integration and Teradata Database Advanced or Enterprise in AWS Marketplace for Teradata customers.

When data is stored in disparate sources, users may not have access to the same information. Even if you frequently replicate your database, your users may still experience latency. Third-party sellers extend the capabilities of the Amazon Database Migration Service to additional sources and destinations and might leverage existing metadata used for data transformation.

For most organizations with on-premises technology investments, operating in a hybrid architecture is a necessary part of the journey to cloud-native adoption. Migrating legacy IT applications to new solutions takes time and effort. AWS can help implement a thoughtful hybrid strategy without requiring costly new investments in on-premises hardware and software, simplifying operations and more easily achieving business goals.

Implementing Informatica PowerCenter on AWS

Informatica PowerCenter Overview

Informatica PowerCenter is a metadata-driven, visual, scalable data integration platform that connects disparate data systems on-premises and in the cloud. The solution delivers data integration to support agility and deliver data to support decision making. You can collaborate, prototype, iterate, analyze, validate and deploy projects in days instead of months, with PowerCenter serving as the foundation of your data integration investments.

PowerCenter offers capabilities to integrate raw, fragmented data from disparate sources and transform it into complete, business-ready information. The solution works with your hybrid architecture, supporting your IT strategy and offering the agility the business needs. It leverages existing metadata models to new AWS data sources.

Informatica PowerCenter Features and Benefits

PowerCenter on AWS Marketplace allows you to integrate on-premises and cloud data sources. You can bring your existing PowerCenter metadata and jobs to AWS, leverage AWS infrastructure and spin up and down AWS infrastructure resources, paying only for what you need rather than over-provisioning and keeping resources idle. You can also reuse skills and efforts such as mappings, workflows and metadata across on-premises and cloud environments.

Deploying PowerCenter in the AWS cloud provides the following benefits to your hybrid cloud bursting environment:

- **Connections to existing data sources and quick onboarding of new data sources and types:** PowerCenter's connectors allow you to connect to on-premises data sources or AWS services like Amazon Redshift, Amazon Relational Database Service (Amazon RDS) and Amazon Simple Storage Service (S3).
- **Faster data loading:** By leveraging your existing on-premises PowerCenter mappings, metadata and workflows, you can quickly load data into AWS services.
- **Reduced costs for data integration:** On-demand compute bursting, in combination with an AWS pay-as-you-go model, can reduce costs by allowing you to increase or decrease the computing power of your data integration platform as needed.

With PowerCenter's metadata-driven AI and enterprise cataloging capabilities, you can ingest, cleanse, process, govern and secure multiple data types from multiple sources into a data lake on AWS. PowerCenter relies on Amazon S3 as the data lake storage system and leverages Amazon's native Hadoop processing power with Amazon Elastic MapReduce or other Hadoop distributions hosted on Amazon Elastic Compute Cloud.

How Informatica PowerCenter Works

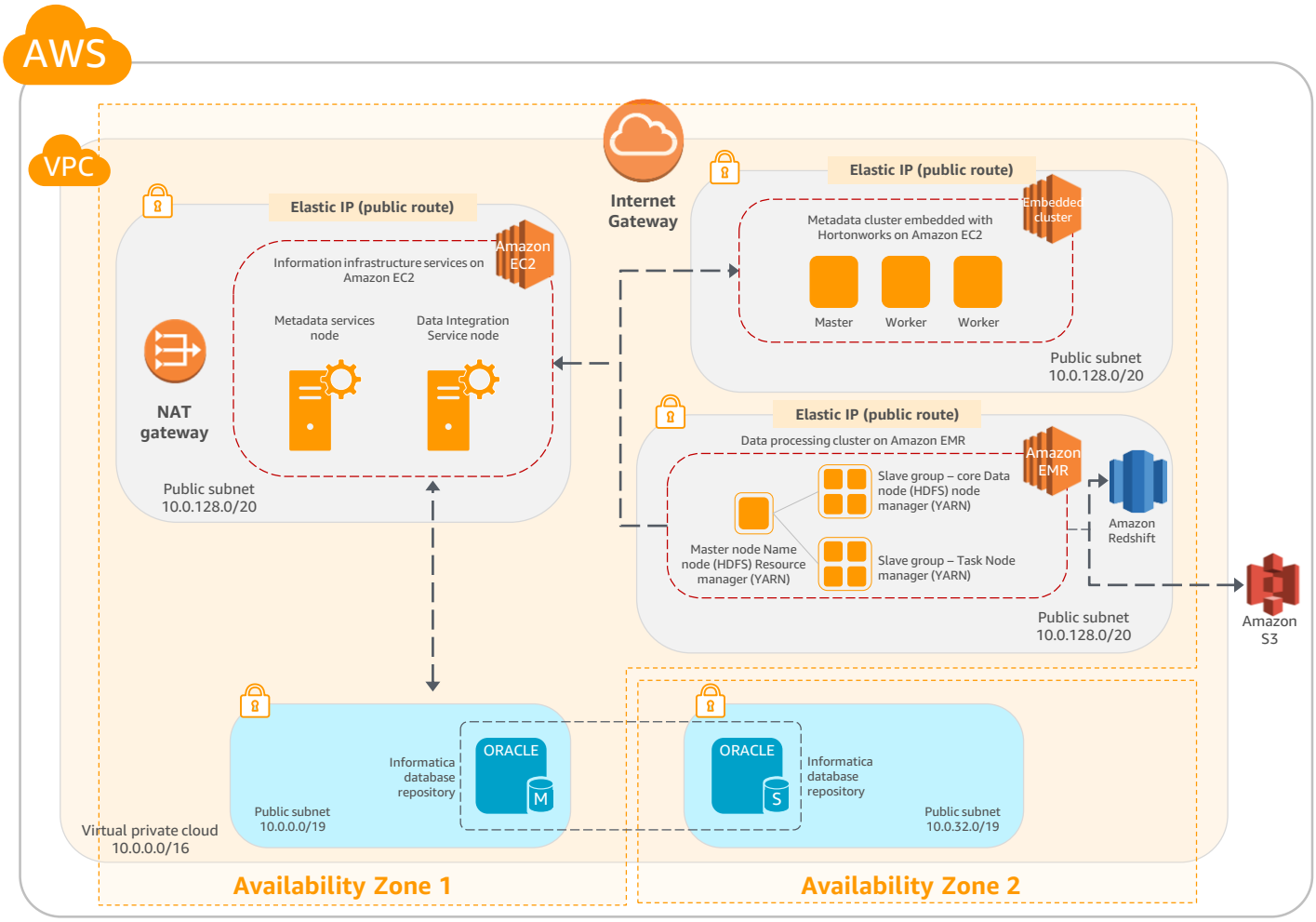
A PowerCenter environment on AWS includes:

- An AWS virtual private cloud (VPC) configured with public and private subnets across two Availability Zones. This provides the network infrastructure for your PowerCenter deployment.
- An Internet gateway to provide access to the Internet.
- The PowerCenter domain, which includes the PowerCenter Repository Service for object locking and access, and the PowerCenter Integration Service.
- Amazon RDS for Oracle repository and domain databases. The repository database holds all the metadata about objects and the domain database manages the service-oriented architecture (SOA) namespace.
- Amazon Elastic Block Store (EBS) for local persistent storage for PowerCenter. The PowerCenter services write cache, source and target files and store logs on Amazon EBS.
- Your choice of two operating systems—Microsoft Windows Server or Red Hat Enterprise Linux (RHEL)—for the PowerCenter deployment.
- On Linux, the choice to provision shared storage for your instances with Amazon Elastic File System (Amazon EFS).
- Your choice to create a new VPC or deploy into your existing VPC on AWS.

Informatica PowerCenter Architecture

The following PowerCenter environment is created in the AWS cloud when following the Quick Start Guide.

Figure 3: Informatica PowerCenter on AWS



The PowerCenter domain runs in the AWS cloud and includes the following application services:

- PowerCenter Integration Service runs PowerCenter sessions and workflows.
- PowerCenter Repository Service manages the PowerCenter repository. It retrieves, inserts and updates metadata in the repository database tables. Oracle serves as the PowerCenter repository database.

The PowerCenter services write files, such as cache, source and target files, to Amazon EBS. If the Create Elastic File System parameter is set to "yes" on a Linux system, the Quick Start uses Amazon EFS instead for shared storage if it's available in your region. While the PowerCenter services run on AWS, you can create, configure and run PowerCenter mappings and workflows from on-premises PowerCenter clients.

Informatica PowerCenter Technical Requirements

The PowerCenter domain can be deployed to an existing VPC or to a new VPC. Connectivity to the required sources and targets must be established from the VPC. All nodes are gateway nodes configured with optional Elastic IP addresses so that client tools running outside the VPC can connect to the domain.

You'll want to make sure that the required number of Elastic IP addresses are available in your AWS account. Also, you'll need to enter the IP addresses of all the nodes in the etc\hosts file on the client machine to establish connectivity to the domain. You can create additional nodes using a single Amazon Machine Image (AMI) to expand compute capacity of the domain as needed.

For more information, visit our Quick Start Guide on Informatica PowerCenter on AWS.

Tips from the Pros

Real-World Advice from Experts that Live Data Everyday

Informatica

Informatica is the world's leader in enterprise cloud data management. They deliver big data insights and allow businesses to be more agile and realize new growth opportunities.

We have a full set of different solutions, whether you want to do batch ETL, and you want to integrate directly from structured data from Oracle RDS onto Aurora, for instance, as a part of your migration. We can load it into any data source. Whether it be on-premise or on cloud, we have over 150 different data sources and native connectivity into a lot of these different data sources.

Many customers say, 'I want to use AWS in order to do all of my processing. I just want to move the data into it.' We recently came out with our mass ingestion capability in which we automatically do everything that I've described but load that right up into Amazon Redshift in bulk. So you don't have to do multi-part uploads. You can actually do a full mass ingestion type of process all the way up into Amazon Redshift as well as into S3.

As you're migrating mass amounts of data onto AWS, it probably makes sense to mass ingest that up into S3 or directly into an Amazon Redshift table, and then use Amazon Redshift analytic table or intermediate analytic table to do the processing as you go.

The same goes for some of the SparkBase use cases of unstructured and structured data. You really want to load that into S3 and then do that temporary processing just so you can leverage the power of Amazon Redshift and EMR for those particular types of use cases. Use Amazon Redshift push down, use EMR Spark processing. These are all some best practices that we talk about.

Andrew McIntyre, Director of Strategic ISV Alliances, Informatica

Implementing Teradata Database on AWS

Teradata Database Overview

With Teradata Database on AWS Marketplace, you get the advantages of Teradata's data software, as well as the agility and flexibility of the AWS cloud infrastructure. Running Teradata Database on AWS is similar to running it in your on-premises data center, giving you the ability to extend traditional multi-system Teradata solutions into the cloud.

Teradata Database Features and Benefits

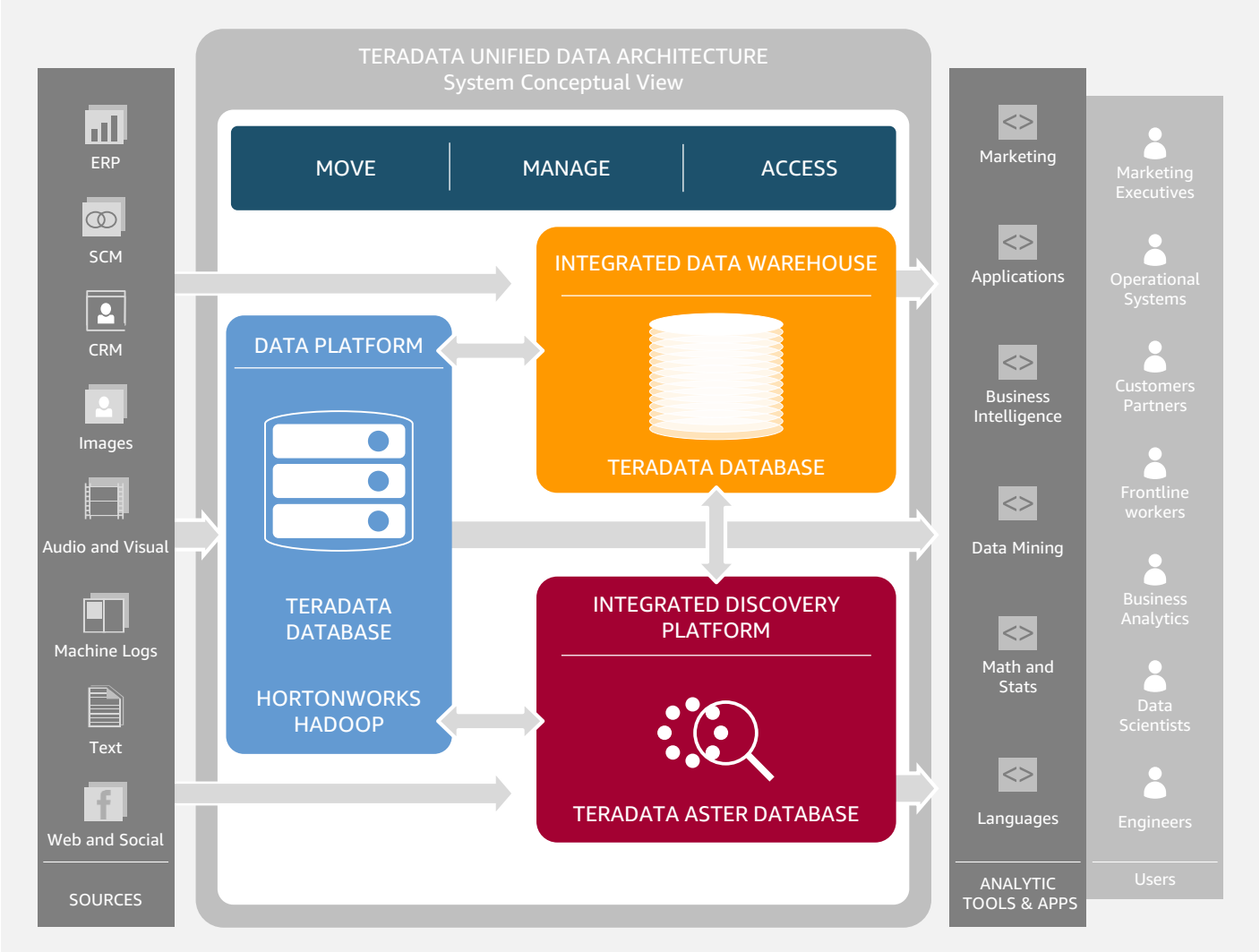
With Teradata Database on AWS, the Teradata Hybrid Cloud Manager helps to support three main hybrid use cases:

- Cloud Bursting: Satisfy business demand and service level consistency with dynamic capacity scaling by supplementing on-premises resources with a temporary cloud-based environment.
- Hybrid Cloud Disaster Recovery: Host a secondary system designed to stay in sync with and take over in case the primary production system is unavailable.

For cloud bursting during peak times or unexpected needs, Teradata temporarily instantiates a secondary cloud environment that you can keep up to date with a production system or a copy of the data necessary to provide the extra processing relief. Workloads can be directed to each system as service levels and performance demand.

Teradata has been working in this way with dual on-premises systems for some time, but with Teradata Database on AWS, you can lower the cost of your data management system while maintaining a high level of responsiveness.

Figure 5: Teradata Database Unified Data Architecture



How Teradata Database Works

Teradata Hybrid Cloud Manager provides automated orchestration of instantiation, configuration and population of Teradata Unified Data Architecture (UDA) multi-system environments. The initial release provides the following autonomous operations:

- Instantiation of Teradata Database system with configuration seeding.
- Addition of selected users and their permissions to the target Teradata Database system.
- Instantiation of Teradata Data Mover, enabling automated creation and execution of object and data migration.
- An existing Teradata QueryGrid for automated creation of “foreign views”, providing access to on-premises objects from the target Teradata Database system.
- Instantiation or leverage of an existing Teradata Viewpoint for automated system monitoring of the instantiated Teradata Database system.
- Stop, start and destroy operations for individual products or the entire environment.

Use Case #2: Implementing a Data Lab on AWS

Implementing a Data Lab with Solutions in AWS Marketplace

As big data continues to grow, data management and business analytics are becoming increasingly complicated. Business Analysts and Data scientists require more analytics capability with less interference. They want self-service access to data and the resources and computing power to analyze and ask questions of their data. To meet this need while helping to ensure agility, flexibility and security, you need coordination and cooperation across the analytic ecosystem.

The data lab is the command center of the data-driven enterprise, enabling the organization to reach the new level of innovation required for data-driven products. The data lab works to answer business needs by uncovering insights found in raw data from a number of structured and unstructured sources. Companies that create their own data labs face a number of obstacles as they ramp-up their data operations. These include ensuring timely data, driving collaboration, leveraging the right skillsets, limiting platform incompatibilities and preparing for growth.

In addition, to run a data lab, data scientists need robust compute power to use large datasets and run multiple experiments in parallel. Because these tasks are computationally intensive and require specialized, powerful CPUs and GPUs, on-premises solutions are expensive and inefficient. Demand is unpredictable and managing these requirements on-premises leads to idle resources and wasted costs. Ultimately, these solutions can't keep up with data deluge, aren't capable of capturing necessary data types and offer few repeatable processes for information management.

A cloud data lab solution resolves these challenges by automatically spinning machines up and down to meet demand. When the data lab trials are complete, users can terminate the cloud infrastructure and reduce the cost and overall environment processes with little impact to the production warehouse.

A data lab solution in the cloud enables data scientists to not only load and explore new data but also integrate the core production and reference data into their query as well. It allows access to large or specialized hardware when needed, then reverts to the on-premises solution when possible. This ultimately limits the challenges associated with idle resources in the data center, enabling the organization to build a data lab while avoiding unnecessary costs.

Domino Data Lab Overview

Domino Data Lab is a data science platform that unifies people, tools, artifacts and work products used across the data science lifecycle, from development to deployment. The platform enables you to create more maturity and discipline around data science as an organizational capability, instead of only a technical skill.

Domino Data Lab accelerates the development and delivery of models while offering infrastructure automation, seamless collaboration and automated reproducibility. This increases the productivity of data scientists and removes bottlenecks in the data science lifecycle.

Domino Data Lab Features and Benefits

Deploying Domino Data Lab on AWS enables you to:

- Accelerate research by streamlining collaboration, sharing and communication.
- Increase productivity by scaling resources and running more experiments per day.
- Enable reproducibility by saving experiment setups and re-establishing test conditions.
- Streamline deployment, choosing from flexible production options.
- Reduce costs by visualizing and intelligently managing your resources.

How Domino Data Lab Works

Domino provides a central system of record that tracks all data science activity across an organization and acts as an orchestration layer on the AWS compute and storage foundation. The result is a platform designed for all stakeholders to rely on. Data scientists get flexibility and scalable compute. IT gets visibility into and management over resource consumptions. DevOps gets a scalable deployment platform. And management gets a reliable, repeatable process for implementing model-driven business programs.

Domino replaces data science siloes and broken workflows with a process for the entire lifecycle. Both individual data scientists and teams can benefit from Domino's robust platform. Individual data scientists can accomplish more with Domino's scalable compute, containerized environments, automatic version control and publishing features. For teams, Domino centralizes work in one place, making it shareable, reproducible and reusable.

When implemented, Domino encourages nimble experimentation and innovation, collaboration and the ability to build a mature, reliable data science capability. Domino provides a central system of record that keeps track of all data science activity across an organization, so you can scale and reduce regulatory and operational risk. Domino helps data scientists orchestrate AWS software toolkits, increase flexibility and innovation and maintain required IT controls and standards. Together, Domino and AWS can provide a self-service cloud resource with the potential for scalability and reduced operational risk, yielding a platform for measurable data science success.

With Domino Data Lab you can:

- Build team knowledge through enhanced collaboration, sharing and communication.
- Work in parallel and run more experiments per day.
- Reduce overhead with automatic version control and reproducibility.
- Streamline deployment and reduce DevOps costs associated with model deployment.
- Reduce costs via central resource management and reporting.

Tips from the Pros

Real-World Advice from Experts that Live Data Everyday

Domino

Domino Data Lab is a data science platform that unifies people, tools, artifacts and work products used across the data science lifecycle, from development to deployment.

There are a couple common challenges where data science and IT struggle. The first of which is access to elastic compute. Data scientists will sometimes be working on their project and at some point they'll think, 'Now it's time for me to spin up a GPU.' It's not like they can plan that out three days in advance. That's not the way data science works. It's a research based process that is un-deterministic. They need to be able to click on a button and to spin it up when they want. That is one of things that Amazon provides a lot of value to but also from an IT perspective, IT doesn't want to have to be on call to release that particular piece of equipment or server to them. Nor do they want to have that data scientist who may be working on an incredibly important project wait for three days either.

The second part is that there's something called package availability. In data science, there are a lot of new open source packages. In fact, there are probably a couple of new ones every day and data scientists leverage those different packages. Those can come from different languages like Python or R or PetraFlow or MXData, lots of different ones. Those data scientists want to be able to get a machine spun up rapidly that has the latest package on it or a very specific package on which they've been working previously and again, in minutes not, 'Hey let me go ask someone to do this and wait for a couple days for me to get me that machine.'

That's where Amazon plus Domino help IT and data science work together to ensure that the project is done faster. And IT doesn't have to worry about managing packages all the time.

Southard Jones, VP of Marketing, Domino Data Lab

Learn more about Domino Data Lab on AWS.

Next Steps: AWS Marketplace

By collecting, managing, securing and analyzing large amounts of structured, semi-structured and unstructured data, you can inform business decisions, act on key insights and guide customer interactions and business processes. This can be done without growing your capacity on-premises to handle big data workloads. With AWS and solutions in AWS Marketplace, you can quickly find and deploy the big data solution you want, save money with pay-as-you-go pricing and scale globally across AWS regions.

Data management solutions on AWS Marketplace include:

Informatica PowerCenter

PowerCenter is a hybrid data integration tool that transforms fragmented, raw data from any source and at any latency into complete, actionable information. It enables you to integrate data from on-premises data sources, cloud data sources or AWS services. You can rapidly jumpstart your PowerCenter environment on AWS and quickly and seamlessly run the exact same data integration environment in the cloud or on-premises depending on your needs and get the same experience. You can then spin up or down AWS infrastructure resources based on need and reuse skills and existing investments such as mappings, workflows and metadata across on-premises and cloud environments.

https://aws.amazon.com/marketplace/pp/B0752DY9DV

Teradata Database

Teradata offers data warehouse software featuring Teradata Integrated Workload Management and no system concurrency limit. Customers can achieve quick time to value and low total cost of ownership. Applications are portable across cloud and on-premises hybrid deployments.

Teradata Database Advanced includes Teradata Database with Teradata Intelligent Memory, Teradata Columnar, Teradata Temporal, Row Level Security, Secure Zones and rights to use Teradata Integrated Workload Management, Teradata Studio and Teradata Tools and Utilities. These tools are available as a free download. See product documentation for details.

https://aws.amazon.com/marketplace/pp/B06Y4PRMKQ

Domino Data Lab

Domino's data science platform is the central hub and system of record for data science teams of all sizes. Domino replaces data science siloes and broken workflows with a seamless process for the entire lifecycle, from exploring new data sets to publishing and managing production models.

Individual data scientists work faster with Domino's easy access to scalable compute, containerized environments, automatic version control and publishing features. For teams, Domino centralizes work in one place, making it shareable, reproducible and reusable. Companies benefit from faster cycle times, collaboration and the ability to build a mature, reliable data science capability. Domino supports publishing models as REST APIs, dashboards, batch runs and as self-service reporting for non-technical users.

Domino is delivered as multi-tenant SaaS or an Amazon VPC. Amazon VPC deployments are available in the region of your choice, including GovCloud, and can be installed in an existing or new Amazon VPC.

https://aws.amazon.com/marketplace/pp/B075FWF96B