Modernizing the Amazon Analytics Infrastructure

Migrating from Oracle Data Warehouse to AWS

March 8, 2021

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided “as is” without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

Contents

Abstract

This whitepaper is intended for AWS customers interested in modernizing their existing on-premises Oracle data warehouses by building cloud-native, scalable and performant analytics platforms using AWS storage and analytics services. The whitepaper is based on Amazon's recent experience of modernizing its legacy Oracle RAC based data warehouse and building a next generation AWS based data lake and analytics services.

The whitepaper starts with an overview of the scale and complexity of Amazon's analytics requirements, challenges faced while operating its legacy data warehouse, and a brief description of the various AWS storage and analytics services used to construct its new analytics systems. The whitepaper goes on to describe the design principles that shaped the design of the new systems, its architecture, strategies used to migrate data. The whitepaper ends with the benefits Amazon experienced after the modernization.

This whitepaper is targeted at IT decision makers, data warehouse solution architects, data warehouse engineers, and analytics managers who are interested in undertaking a similar modernization at their organizations. It assumes that the reader is familiar with basic concepts of databases, data warehouses and cloud computing services offered on AWS.

Introduction

Amazon builds and operates thousands of services to serve its hundreds of millions of customers. These services enable customers to accomplish a range of tasks including browsing the Amazon website, placing orders, submitting payment information, subscribing to services, initiating returns, watching videos, listening to music, and interacting with Alexa devices. They also enable Amazon employees to perform a range of activities such as optimizing inventory in fulfillment centers, scheduling customer deliveries, reporting and managing expenses, performing financial accounting, launching new products, and running business analytics. Amazon's business analytics and data warehouse requirements are complex and have grown rapidly in recent years. The large and diverse analytics user base comprises over 1,800 data producers and 80,000 data consumers. An average data producer publishes ten tables and an average data consumer consumes fifty tables. The service is expected to execute over 900,000 extract-transform-loads (ETLs) every day. To meet these challenging requirements, Amazon operated Oracle RAC clusters in the past to meets its needs, but the growing set of challenges prompted a redesign. In 2017, Amazon decided to modernize its legacy data warehouse and invest in building a robust, scalable, and performant analytics system using AWS services.

Challenges with the Legacy Data Warehouse

This section describes the key challenges faced by users and the data warehouse engineers while operating the Oracle data warehouse.

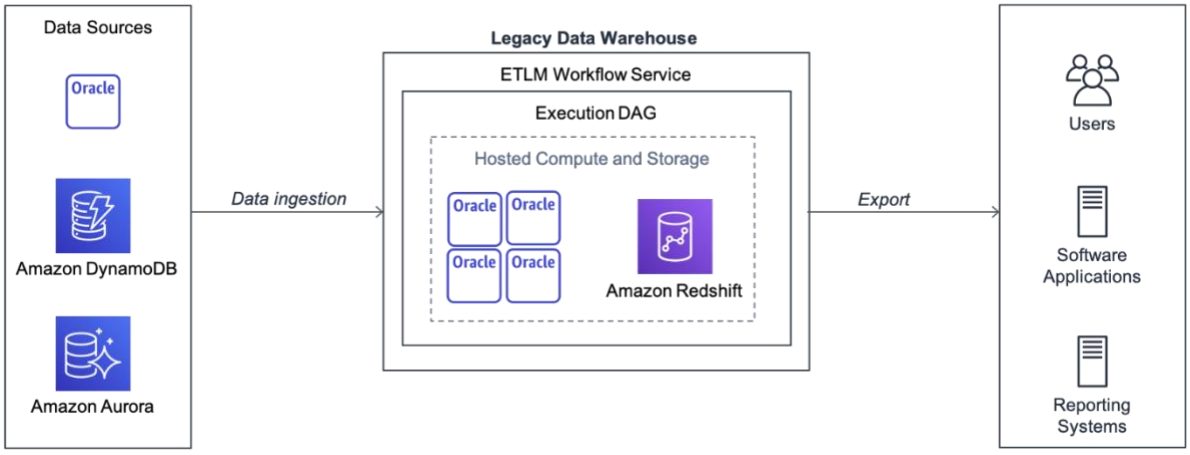

Figure 1: Architecture of the legacy Oracle data warehouse

Poor Performance at Scale

The data warehouse could not keep pace with the rapid growth in data volume and the increasing complexity of user queries. Database engineers used database partitions to optimize the amount of memory utilized by transforms; however, their design and implementation required extensive analysis of hundreds of terabyte-sized tables and thousands of user queries. Their implementation also grew increasingly complex each year due to growth in data volume. User queries processing tables containing over 100 million rows consistently failed restricting the ability of data consumers to use large tables in the preparation of complex reports, generation of deeper insights, or the deployment of machine learning models. These limitations left many data producers with no option but to make their datasets unavailable in the data warehouse.

Expensive and Suboptimal Usage of Infrastructure

The second issue with the legacy data warehouse was escalating costs. Amazon spent millions of dollars each year in licenses and specialized hardware. The data warehouse also had to be resourced for peak loads leading to significant underutilization during the rest of the year.

Tedious Database Management

Management of the legacy database software was tedious. Database engineers spent hundreds of hours each month performing routine database administration tasks such as software upgrades, database backups, operating system (OS) patching and performance monitoring. To fix certain issues, they had to work with Oracle support and obtain patches. The database engineers had to perform some of these tasks manually and these activities could trigger failures. This inherent unreliability prompted them to prepare contingency plans to recover and restore databases.

Cumbersome Hardware Provisioning

Hardware provisioning was cumbersome and complicated. Database engineers had to spend hundreds of hours each year forecasting user requirements, sizing hardware to meet demand, and then procuring it. After hardware procurement was complete, engineers spent hundreds of hours commissioning it. They also had to deal with the overhead of managing spares. Due to these challenges, Amazon decided to build a scalable, robust, secure and performant analytics system using AWS storage and analytics services.

Overview of AWS Services Used in New System

AWS offers an integrated suite of services required to quickly and easily build, and manage an analytics system. AWS powered data lakes can handle the scale, agility, and flexibility required

to combine different types of data used by Amazon in ways that the legacy data warehouse could not. AWS also offers the widest range of analytics and machine learning services for easy access to relevant data without compromising on security or governance. The following section provides a brief overview of the key AWS services used to build the new system.

Amazon Simple Storage Service (Amazon S3)

Amazon S3 is secure, highly scalable, durable object storage with millisecond latency for data access. Amazon S3 is built to store any type of data from anywhere—websites and mobile apps, corporate applications, and data from Internet of Things (IoT) sensors or devices. It is built to store and retrieve any amount of data, with high availability, and built from the ground up to deliver 99.999999999% (11 nines) of durability. Amazon S3 also offers tiered storage based on data retention requirements. Amazon S3 Select focuses data read and retrieval, reducing response times up to 400%.

Amazon EMR

For big data processing using the Apache Spark and Hadoop frameworks, Amazon EMR provides a managed service that makes it easy, fast, and cost-effective to process vast amounts data. Amazon EMR supports 19 different open-source projects including Hadoop, Spark, HBase, and Presto, with managed Amazon EMR Notebooks for data engineering, data science development, and collaboration.

Amazon Redshift

Amazon Redshift is a fast, scalable data warehouse that makes it simple and cost-effective to analyze data across your data warehouse and data lake. For data warehousing, Amazon Redshift provides the ability to run complex, analytics queries against petabytes of structured data, and includes Amazon Redshift Spectrum that runs SQL queries directly against exabytes of structured or unstructured data in Amazon S3 without the need for unnecessary data movement.

Amazon QuickSight

For dashboards and visualizations, Amazon QuickSight provides you a fast, cloud-powered analytics service, that that makes it easy to build stunning visualizations and rich dashboards that can be accessed from any browser or mobile device.

Design Goals

To design a scalable, robust, and secure analytics system, the solutions architects and engineers of the system adopted three key guiding principles. This section describes these principles and their influence on the architecture of the new system.

Unbounded Scalability

Amazon wanted to ensure the new system could scale well beyond current user requirements. To ensure that scale was unbounded, they decided to modularize the two core components of the system – storage and compute. With each of these components allowed to scale independently, the system could scale to handle workloads of higher orders of magnitude.

Open Systems Architecture

Amazon analytics needs are diverse and continue to grow in variety. As analytics and data warehouse technologies and frameworks continue to evolve, the data warehouse architects wanted to provide users options to pick compute services of their choice. To enable this, they designed the new system to store and retrieve data in native, raw formats thus allowing interoperability.

Enhanced Transparency

The legacy data warehouse struggled to meet the Amazon requirements of data transparency. The data warehouse architects analyzed this gap and discovered that most data warehouses and analytics services make data transparency a challenge creating an asymmetry of information between data producers and data consumers. The solutions architects wanted to eliminate this challenge by making it easier for users to search for datasets, trace their lineage, monitor their health, and actively track changes.

Architecture and Components of Andes

The solutions architects designed an analytics and data storage system that was influenced by the guiding principles named Andes. Andes contains four components – storage, data discovery services, ingestion services, and synchronization services. The following section describes the function of each of these components and the overall architecture of the system.

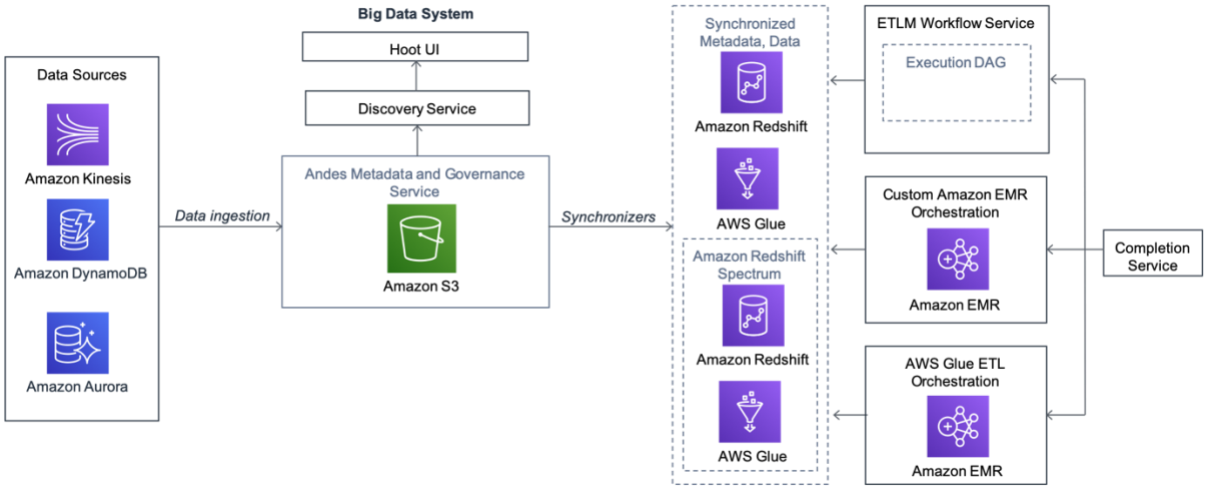

Figure 2: Architecture diagram of Andes

Storage Layer or Data Lake

The foundational layer of Andes was an Amazon S3 based data lake that ingested and stored vast amounts of raw data and made it available for data consumers while enforcing appropriate security and governance policies. Amazon used Amazon S3 to hold raw data in native format until required for analysis. Amazon S3 provides Amazon exceptional durability and a comprehensive set of security features to enforce access control and encryption. Beyond the natively supported features of Amazon S3, Amazon integrated its internal authentication, authorization, and data governance services with the data lake to deliver seamless and secure access to its users. This data lake can store petabytes of data and offer simultaneous access to thousands of users accessing it.

Data Discovery Services

After the datasets were made available in the data lake, data consumers needed a way to discover, access, and analyze them. Amazon developed a set of services to simplify the dataset discovery process through the data discovery services. These services allowed data consumers to easily search, sort, and identify datasets for analysis; explore schemas; and evaluate table sizes, partitions, indexes, keys, and other information. The data discovery services include the following key functions:

- Crawling the data sets on the data lake
- Extracting the metadata and preparing a data catalog
- Providing search, sort, and filter functions to users
- Delivering information about datasets to users through the browser interface

The discovery services also provided a way for data consumers to interact with the data producers through a single interface synchronized with Amazon's internal issue management system. For additional value to users, the data discovery services also displayed precomputed dataset health metrics to allow data consumers to assess data quality.

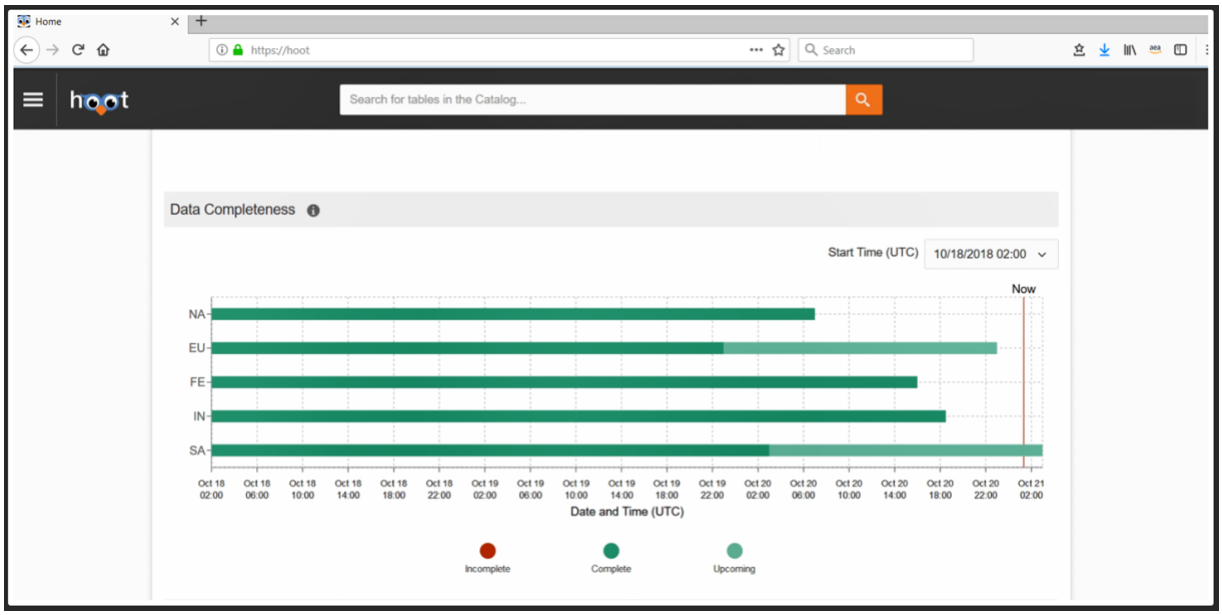

Figure 3: Data health metrics

Data Ingestion Services

After building the data lake and the data discovery services, the next challenge was to design an easy, efficient, and secure way for data producers to move data into it. This was accomplished through the Andes data ingestion services. These services enabled data producers to move data from a variety of sources such as Amazon Kinesis, Amazon DynamoDB, and Amazon Aurora into the data lake. It also enabled these data producers to create, update, and delete datasets. These services are like traditional ETLs services with tightly coupled integration with Amazon's internal access control and authorization mechanisms for security.

Data Synchronization Services

Data consumers require datasets from the data lake for analysis. To serve this need, the data warehouse architects designed the data synchronization services to create copies of datasets residing in the data lake on users' Amazon Redshift clusters. The synchronization services also enforce access control by integrating with internal access control and authentication systems. These services trigger initial bulk loads of datasets on users' Amazon Redshift clusters when initiated. After the initial load is complete, ongoing change data is propagated daily. The synchronization services also complete the following tasks:

- Update metadata changes, if applicable

- Propagate schema changes
- Synchronize change data with user clusters
- Warn users of dataset deletions
- Initiate new synchronization requests

The synchronization services also enable data consumer – data producer communication and issue resolution by integrating with Amazon's internal ticket management system.

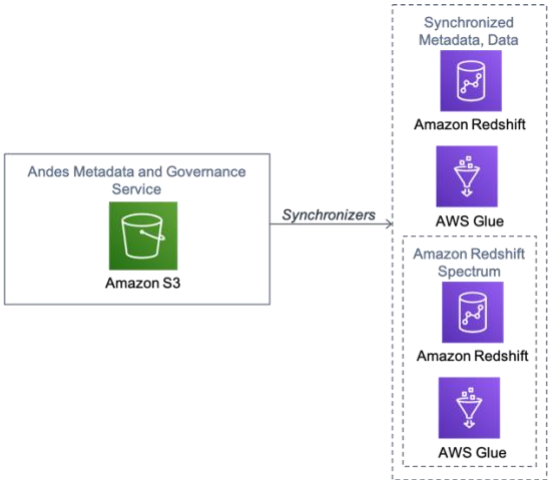

Figure 4: Andes data synchronization services and target platforms

Compute Layer

For the compute layer, Andes was designed in way that gave each team the freedom to pick any compute framework or service offered by AWS. Most data consumers chose Amazon Redshift due to the relative ease of transition from Oracle. Amazon Redshift was also popular because it enables business analysts, product managers, and financial analysts to run SQL queries and visualize the results in Java Database Connectivity (JDBC) compliant tools such as Amazon QuickSight. Another compute service used was Amazon Redshift Spectrum that allows users to directly query open data formats stored in Amazon S3. It enables users to analyze data across the warehouse and data lake in a single service using the familiar Amazon Redshift SQL syntax at only a fraction of the costs of Amazon Redshift.

Data scientists and advanced analytics users preferred using Amazon EMR over Amazon Redshift as Amazon EMR has the capability to compute without duplicating the data. This eliminates the need for data transfer leading to faster compute for heavy workloads. AWS Glue offers much needed ETL scheduling functionality for Amazon EMR. Amazon EMR and AWS Glue allowed users to run large, complex queries and deploy machine learning models that were previously not accessible on the Oracle data warehouse.

Migrating Data into Andes

The data warehouse architects, engineers, and program managers accelerated the migration through meticulous planning and tools. This migration was broken into two phases – seeding phase and transform phase. This section briefly describes each of the two phases.

Seeding Phase

In the seeding phase, the Program Management Office (PMO) moved select datasets from the Oracle data warehouse to the data lake. These datasets were selected based on frequency of usage and their importance to business decision making. These datasets seeded the data lake and encouraged the initial set of users to begin using it. It also created a demand for the rest of the datasets to follow. In parallel, the data warehouse engineers modified the loads jobs of these selected datasets to read from the source database and simultaneously write to two destinations—the legacy Oracle data warehouse and the Andes data lake. This approach ensured that these select data sets were available and updated in both locations, and it allowed data consumers to test their ETLs, transforms, reports and metrics on the new platform before eliminating their dependency on the Oracle data warehouse. After seeding the initial datasets, subsequent datasets were migrated to the new data lake in waves.

Transform Phase

The next phase of the migration involved migrating the end consumer compute. It primarily involved data consumers modifying their ETLs written in Oracle SQL to run on Amazon Redshift or Amazon EMR. As both Oracle and Amazon Redshift use similar SQL syntaxes, the modification of transforms running on Oracle to run on Amazon Redshift was easy to manage. To accelerate the translation of many transforms, the program managers advocated the use of AWS Schema Conversion Tool (AWS SCT). AWS SCT offered many features that accelerated the migration from Oracle to Redshift. In certain instances, where the database features could not be converted directly, the AWS SCT extension pack wizard installed AWS Lambda functions and Python libraries to emulate these features. AWS SCT also optimized Amazon Redshift databases by recommending sort and distribution keys. The data warehouse engineers worked closely with the AWS SCT product team and improved the conversion accuracy of the tool to over 95% saving Amazon hundreds of engineering hours. Teams using Amazon EMR developed custom tooling to convert their transform jobs. After testing the accuracy and performance of each migrated ETLs, the teams shut down the equivalent ETL on the Oracle data warehouse.

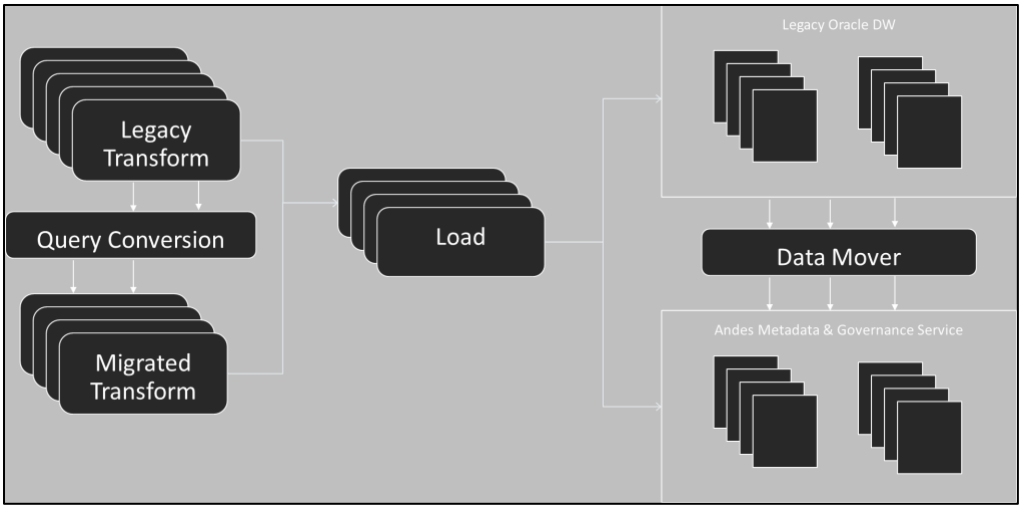

Figure 5: Step in migrating ETLs from Oracle to Amazon Redshift

Managing the Migration

This section describes the organization challenges encountered during the migration and the mechanisms used by the program managers to overcome them.
After Andes was released in beta, Amazon assembled a team of experienced program managers to build mechanisms that would ensure that all artifacts including datasets, ETLs, reports and applications built by a vast and globally distributed user base was migrated successfully. The migration was a cultural change as much as a technological one and the program managers implemented processes and mechanisms to ease the transition. The migration included the following key challenges:

- ensuring that all teams and users see the vision of the program and support it
- establishing uniform, consistent, and achievable goals for teams
- diligently tracking the progress of teams against these goals
- providing all users with enough training and support during the project

The program management team adopted the following three core tenets for the project:

- The migration will not impact business continuity.
- The older data warehouse will remain active until all teams complete their migration.
- The new system will operate in a decentralized manner.

To simplify project planning and execution, users were segmented into two groups—data producers and data consumers. Data producers own services that generate datasets and are responsible for publishing them to the data warehouse. Data consumers use these datasets to analyze data, prepare business reports, draw insights, and make business decisions. To ensure a

successful outcome, the program managers assigned realistic goals for teams within each group, tracked their progress, reported it to leadership, and extensively trained the user base. Each team was also asked to nominate a program manager to set goals, prepare project plans, and track progress.
The program team planned the migration in three waves. The first wave comprised teams that were the most extensive users of the data warehouse. During this first wave, the program managers refined their processes, mechanisms, and goals. After successfully migrating the first wave of users, the second and third phase of users were migrated.

Benefits

Amazon saw multiple benefits from the modernization of its analytics infrastructure. This section briefly describes a few of these benefits

Reduced Costs

The first benefit was a reduction in operating costs achieved through a combination of cheap storage and compute, decentralized capacity management, and the elimination of database administration overhead.

Agility

The second benefit was business agility. As business units were able to rapidly scale compute resources, teams could make business decisions quicker. As the modernization of the infrastructure also opened fresh analytical methods and access to new data sources, teams could run analytics at a higher magnitude of precision and scale.

Improved Security

The third major benefit was the elimination of physical security threats and real estate costs associated with datacenters. Andes allows data producers to control access to data sets. In addition, the system can restrict access to data consumers ensuring that authorized consumers can read these restricted datasets.

Decentralized Infrastructure

In the new analytics system, business units manage their compute instances locally, decentralizing the infrastructure management. In the legacy architecture, a centralized team had to manage the hardware and charge each business unit for their usage. This cost allocation model was complicated. Now, each team can optimize the ratio of reserved to on-demand instances based on usage patterns and data growth projections. Business units with cyclical loads have the flexibility to allocate a higher percentage of on-demand instances compared to

teams with steadier loads. The solutions architects and program managers developed heuristics to suggest an optimal mix of instance types based on projected service usage growth, load cyclicality, and discounting. They also built automated solutions to monitor the usage of fleets of AWS accounts to optimize usage.

Consistent Deployment of Services

To ensure that AWS service configurations were deployed uniformly and consistently across all teams with one-click, the program managers distributed preconfigured AWS CloudFormation templates. As the analytics stack is extensively used by technical and business users, the engineers developed additional interfaces to abstract the technical details of the deployment thus making adoption easier for business users. These templates ensured consistent usage of these service compliant with IT policy.

Enhanced Transparency

Andes improved data quality due to the enhanced transparency introduced by the metadata and governance services. This enhanced transparency translated to improved accountability from data producers, easy identification of error, and faster turnaround cycle for error resolution.

Contributors

Principal contributors to this document are

- Venkata Akella, Senior Product Manager, Database Freedom
- Naveen Yajaman, Principal Technical Program Manager, Business Data Technologies
- Craig Woods, Senior Solutions Architect, Business Data Technologies

Document Revisions

Date	Description
March 8, 2021	Reviewed for technical accuracy
November 2019	First publication