Amazon Connect Data Lake Best Practices

AWS Whitepaper

Amazon Connect Data Lake Best Practices : AWS Whitepaper

Table of Contents

Amazon Connect Data Lake Best Practices

Publication date: **May 13, 2021**

Abstract

Customer service is a crucial element of brand reputation and business success. Contact centers are vital to enabling a two-way agent-customer interaction and essential to delivering a superior customer service experience. Conversely, a poor experience can lead to customer churn. Organizations invest in omnichannel contact centers for a competitive edge in enhancing customer experience.

According to an Aberdeen survey, organizations manage an average of 33 unique data sources for analytics and experience 50% year-over-year data volume growth. Rapid data volume growth creates challenges in data management and storage capacity. Today, organizations are developing data lake strategies to harness intelligence from the diverse and ever-growing data. The survey indicates a 9% increase in organic revenue growth for organizations that implemented a data lake.

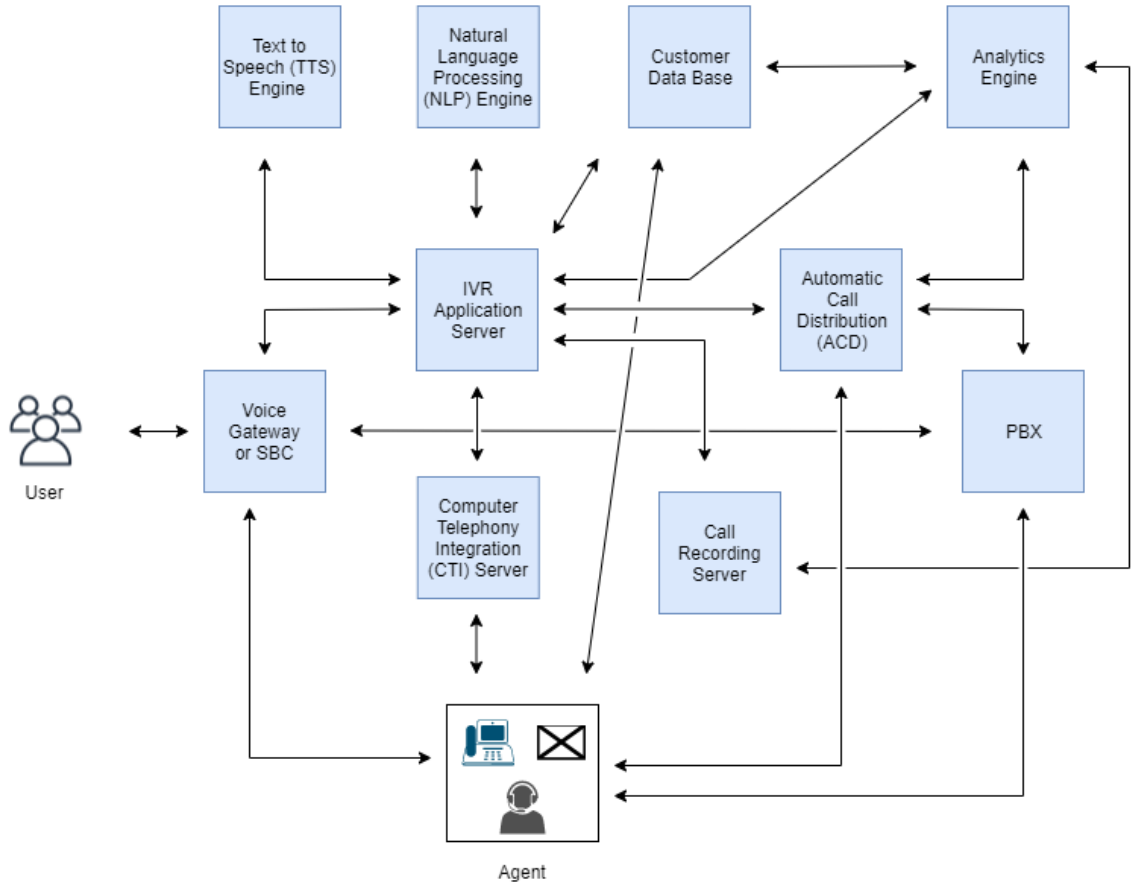
To get the most advanced analytics benefits, organizations need a robust platform and cost-effective solution to run a thriving contact center. Amazon Web Services (AWS) provides customers with a comprehensive set of services and a scalable platform to ensure high availability, security, and resiliency of a data lake in the cloud.

This whitepaper outlines the best practices for architecting a contact center data lake with Amazon Connect.

Introduction

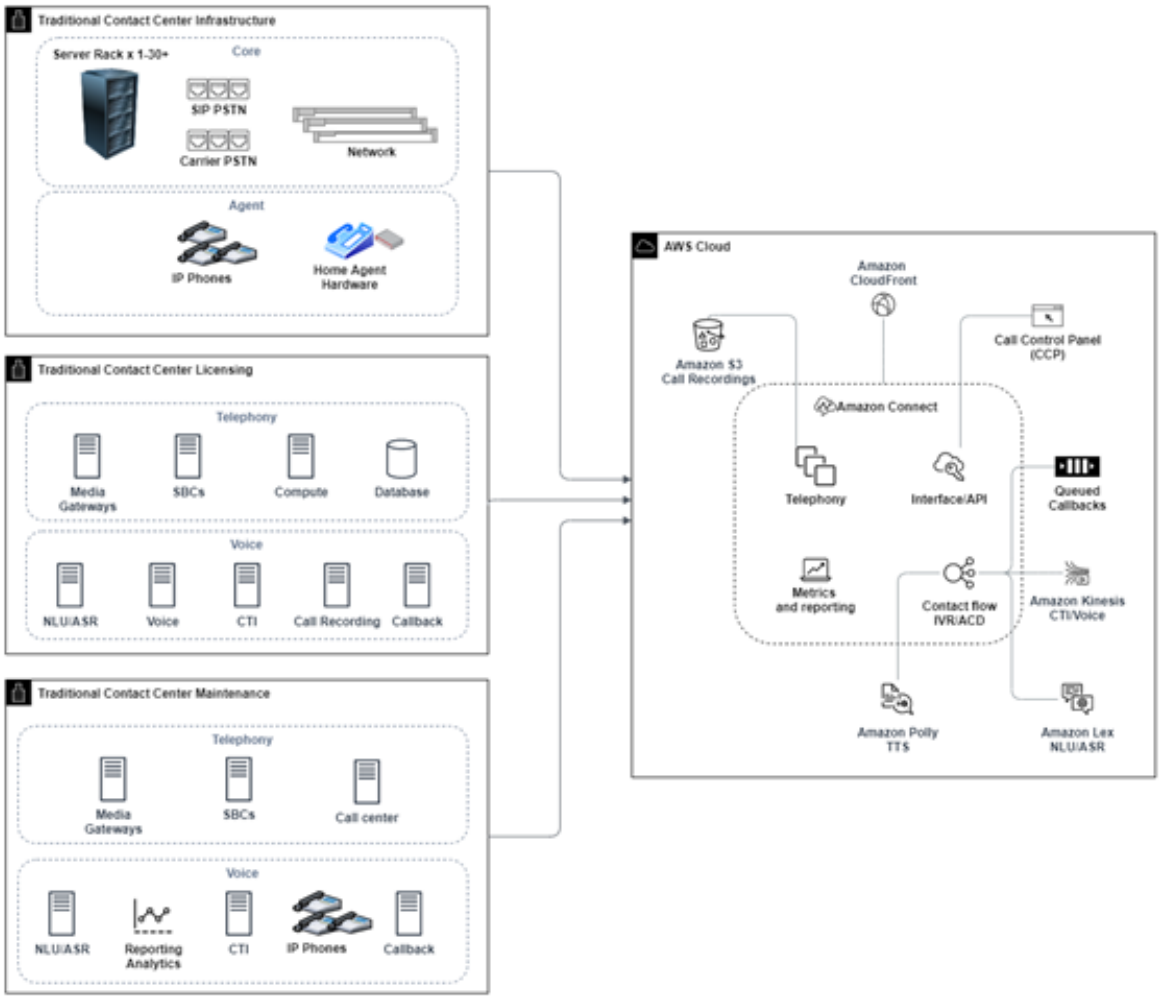
Traditional on-premises contact centers often involve multiple proprietary systems, resulting in disparate data sources containing data in various formats. Challenges in standardizing and consolidating information slow down the discovery of new business insights or possible operational issues.

The following figure shows the architecture of a traditional on-premises contact center.



On-premises contact center architecture

Data plays a crucial role in driving the success of a contact center. Managing data silos is inefficient and may lead to a poor customer experience. Customers expect a frictionless experience when connecting with a contact center. Having a streamlined data lake solution helps agents deliver a personalized experience to customers. The following figure shows a strategic approach to simplifying complex traditional contact center data spans across infrastructure, licensing, and maintenance environments into Amazon Connect.



A strategic approach to simplifying complex traditional contact center data into Amazon Connect

A data lake is a centralized, curated, and secured repository that stores and governs all your structured and unstructured data in its native or transformed formats for analysis. AWS delivers the breadth and depth of services to build a secure, scalable, comprehensive, and cost-effective data lake solution. You can use the AWS services to ingest, store, find, process, and analyze data from a wide variety of sources.

This whitepaper provides architectural best practices to technology roles, such as chief technology officers (CTOs), architects, developers, and operations professionals when building a contact center data lake with Amazon Connect.

Amazon Connect

Amazon Connect is an easy-to-use and cost-effective omnichannel cloud contact center. You can get started with a fully managed cloud-based and artificial intelligence (AI) enabled contact center within minutes. With the pay-as-you-go model, you pay only when the service is in use. There is no infrastructure to manage or upfront costs.

Forrester Research Consulting conducted a Total Economic Impact (TEI) study on Amazon Connect and concluded a three-year financial impact on how Amazon Connect helps customers with significant cost savings, increased revenue, and improved agent productivity. Key findings include:

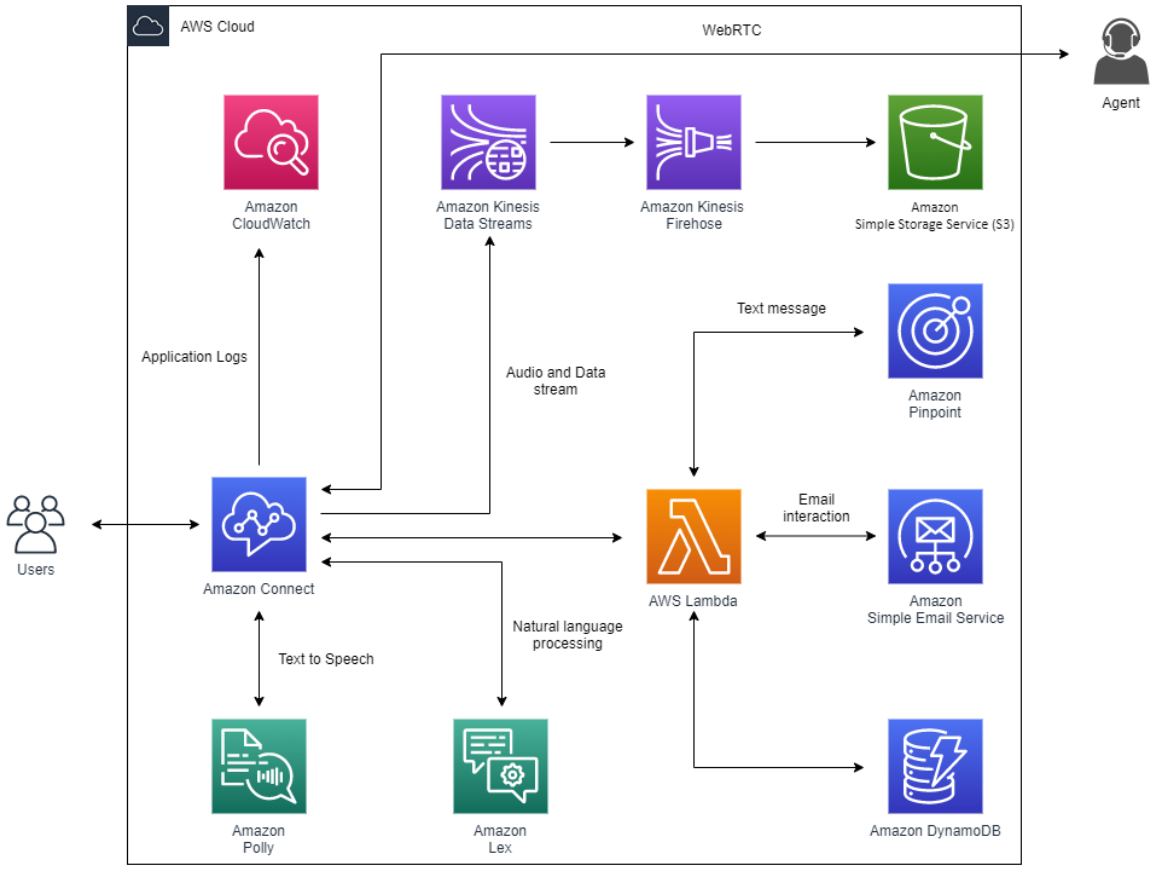
- Reduction in cloud technology costs of $4.3 million
- Subscription cost savings of 31%
- Agent labor savings from reduced call volume of $4.6 million
- Increased operating income by $2.6 million with enhanced customer experience
- Return on investment (ROI) of 241%

Amazon Connect provides skills-based routing, task management, powerful real-time and historical analytics, and intuitive management tools. You can focus on improving customer service experience and measuring contact center performance with ease using Amazon Connect. Agents can be productive quickly with a web-based softphone from any location.

With built-in analytics capabilities such as Contact Lens for Amazon Connect, contact center supervisory personnel can discover sentiment in contact interaction and operational efficiency.

Amazon Connect is an open platform. Using Amazon Connect's extensive set of published APIs, you can programmatically integrate with other AWS services and third-party systems, including customer relationship management (CRM) solutions and anti-fraud solutions.

The following figure shows a high-level Amazon Connect contact center architecture. Amazon Connect provides a unified and seamless customer experience across multiple channels. Along with voice and webchat, Amazon Connect integrates with Amazon Pinpoint and Amazon Simple Email Service (Amazon SES) to expand the contact center's capability on text messages and email delivery. Amazon Connect integrates with Apple Business Chat for Apple device users.



Amazon Connect contact center architecture

Data lake design principles

Building a data lake can break down data silos and democratize data for value extraction. A central data repository empowers organizations to make data-driven decisions and innovate quickly.

Organizations want a cost-effective and elastic storage capacity to store disparate data sources that grow exponentially. They want to centrally govern and share vast amounts of data across different business units. Furthermore, they want to empower their employees and stakeholders to derive business insights with shorter time-to-value.

Considerations when designing a data lake:

- How do you collect, store, and analyze high-velocity data across various data types, including structured, unstructured, and semi-structured?
- How do you store and share petabytes of data on-demand globally and cost-effectively?
- How do you scale IT resources to support a high number of concurrent queries against your data and scale down automatically for cost savings?
- How do your users view, search, and run queries on multiple data repositories today?
- How do you derive future insights using historical data patterns and past scenarios?

Data types

Amazon Connect manages a variety of contact center data, including:

- Resources and configurations such as queues, contact flows, users, and routing profiles
- Contact metadata such as connection time, handle time, source number or automatic number identification (ANI), destination number or dialed number identification service (DNIS), and user-defined contact attributes
- Agent-related performance data such as login time, status changes, and contacts handled
- Phone call audio streams such as call recordings
- Chat transcripts
- Attachments
- Integration configuration with external applications
- Knowledge documents
- Voiceprints for authenticating customer's voice

This section gives an overview of various data types available in Amazon Connect.

Customer profiles

Amazon Connect Customer Profiles enables agents to deliver efficient and personalized customer service by importing customer information from various applications into a unified customer profile. You can ingest customer data from homegrown or third-party applications such as Salesforce, ServiceNow, Zendesk, and Marketo into your Amazon Simple Storage Service (Amazon S3) data lake using pre-built connectors.

Contact trace record

Contact trace records (CTR) captures transactional metrics such as hold time, wait time, and agent interaction time in JavaScript Object Notation (JSON) format. Amazon Connect aggregates CTR data to create metrics reporting. Data retention for CTR is 24 months upon contact initiation. You can stream CTRs to Amazon Kinesis for extended retention and advanced analysis. The CTR data model describes various event types available in CTRs.

Contact flow logs

Amazon Connect contact flow logs capture real-time events and metrics about how your customers interact with contact flows. Amazon CloudWatch creates a log group for each Amazon Connect instance when you enable contact flow logging and include a set logging behavior block for the contact flows.

Contact flow logs contain the contact flow ID, the customer's contact ID, and the block's actions. Using contact flow logs, you can compare customer's interactions with different contact flow versions or trace their interactions through each contact flow. Contact flow logs help you debug and roll back contact flows to previous versions should any issues arise.

Contact Lens output files

Using natural language processing (NLP) and speech-to-text analytics, Contact Lens for Amazon Connect provides insights to analyze customer sentiment, identify conversations trends for product feedback, and compliance audits for standard greetings and sign-offs.

With advanced conversational search, you can perform a fast full-text search for relevant calls by sentiment scores and non-talk time to identify common utterances that result in positive or negative customer sentiment. Contact Lens automatically redacts sensitive personally identifiable information (PII) for data privacy.

You can intercept potential poor customer experience by creating rules to send alerts on specific keywords or phrases. Agents can escalate the issue and transfer calls while passing real-time transcripts to ensure proper handoff.

Contact Lens stores metadata for call transcript, sentiment analysis, non-talk time, talk speed, interruptions, and categorization labels in Amazon S3. You can create custom visualization or machine learning (ML) models using data from Contact Lens and CTR stored in S3.

Agent events streams

Amazon Connect agent event streams capture and store agent activity in S3 via Amazon Kinesis Data Streams. You can create dashboards for near real-time agent reporting such as agent login, agent logout, agent connects with a contact and agent status change.

You can integrate agent event streams into workforce management (WFM) solutions for agent staffing management or configure alerts on specific agent activity.

Voice and chat recordings

Amazon Connect records a conversation only when a customer connects to an agent. When the contact disconnects, the call recordings are available in your S3 bucket, or accessible in the customer's contact trace record (CTR).

As an omnichannel contact center, Amazon Connect Chat enables customers to chat with agents across your business applications, web, or mobile. Customers can resume conversations and switch devices during the chat.

Amazon Connect redacts, encrypts, and stores voice and chat conversations between the agent and the contact in your S3 bucket for advanced analytics.

Third-party integration

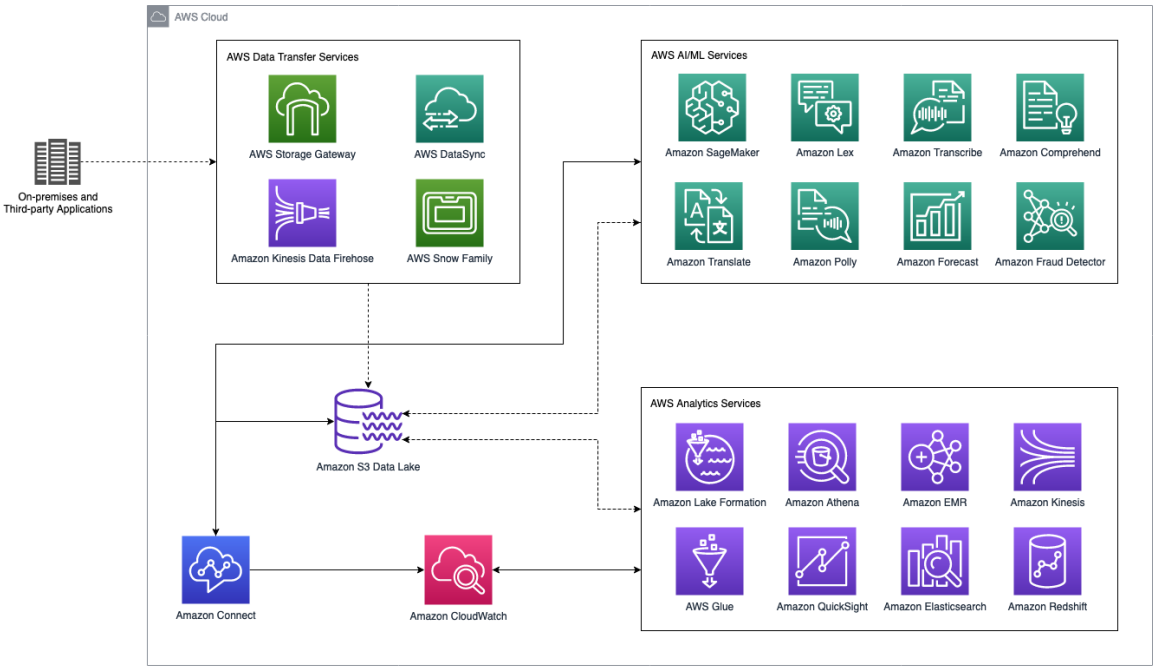
When using AWS Partners or other third-party solutions with Amazon Connect, you can consolidate logs and external data sources in Amazon S3.

Data lake lifecycle

Building a data lake typically involves five stages:

- Setting up storage
- Moving data
- Preparing and cataloging data
- Configuring security policies
- Making data available for consumption

The following figure is a high-level architecture diagram of an Amazon Connect contact center data lake that integrates with AWS analytics and artificial intelligence / machine learning (AI / ML) services. The following section covers the scenarios and AWS services shown in this figure.



Amazon Connect contact center data lake with AWS analytics and AI / ML services

Storage

Amazon S3 is an object storage service that offers industry-leading scalability, data availability, security, and performance. S3 provides 99.999999999% durability and 99.99% availability with strong consistency and unlimited data storage globally. You can use Cross-Region Replication (CRR) to copy data across S3 buckets in multiple Regions for regulatory compliance and low-latency requirements. S3 scales throughput automatically for performance and operational efficiency.

S3 buckets and objects are private with S3 Block Public Access enabled by default to all Regions globally. You can set up centralized access controls on S3 resources using bucket policies, AWS Identity and Access Management (IAM) policies, and access control lists (ACLs). You can evaluate and identify any buckets

with public access using Access Analyzer for S3. With object prefixes and tagging, you can manage access controls, storage tiering, and replication rules at the object-level granularity.

AWS CloudTrail logs every API call to S3 server access logging. S3 inventory audits and reports replication and encryption status for your data.

S3 Intelligent-Tiering provides automatic cost savings by moving data between frequent and infrequent access tiers when the access patterns change, without performance impact or operational overhead. S3 Glacier Deep Archive saves up to 95% on storage costs for rarely accessed objects that require long-term retention.

Storing data in columnar formats such as Apache Parquet and Optimized Row Columnar (ORC) enables faster queries and reduces processing costs with Amazon Athena. Compression options such as Snappy with Parquet reduce capacity requirement and storage cost.

With S3 Select and S3 Glacier Select, you can query object metadata using structured query language (SQL) expression without moving the objects to another data store.

S3 Batch Operations automate bulk operations on S3 objects, such as updating object metadata and properties, performing storage management tasks, modifying access controls, and restoring archived objects from S3 Glacier.

S3 Access Points simplify and aggregate access for shared data on S3 by different teams and applications. Each access point is associated with a unique DNS name for a single bucket. You can create service control policies (SCPs) to restrict access points to an Amazon Virtual Private Cloud (Amazon VPC)and isolate data within your private networks.

S3 Transfer Acceleration enables file transfer over long distances between your client environment and S3 buckets.

As your data lake grows, S3 Storage Lens provides organization-wide visibility into object storage usage and activity trends with actionable recommendations to reduce cost and operational overhead.

Ingestion

AWS provides a comprehensive data transfer services portfolio to move your existing data into a centralized data lake. Amazon Storage Gateway and AWS Direct Connect can address hybrid cloud storage needs. For online data transfer, consider using AWS DataSync and Amazon Kinesis. Use the AWS Snow Family for offline data transfer.

- **AWS Storage Gateway** extends your on-premises environments to AWS storage by replacing tape libraries with cloud storage, providing cloud storage-backed file shares, or creating a low-latency cache to access your data in AWS from on-premises environments.
- **AWS Direct Connect** establishes private connectivity between your on-premises environments and AWS to reduce network costs, increase throughput, and provide a consistent network experience.
- **AWS DataSync** can transfer millions of files into S3, Amazon Elastic File System (Amazon EFS), or Amazon FSx for Windows File Server while optimizing network utilization.
- **Amazon Kinesis** provides a secure way to capture and load streaming data into S3. Amazon Kinesis Data Firehose is a fully managed service for delivering real-time streaming data directly to S3. Kinesis Data Firehose automatically scales to match the volume and throughput of streaming data and requires no ongoing administration. You can transform streaming data using compression, encryption, data batching, or AWS Lambda functions within Kinesis Data Firehose before storing data in S3. Kinesis Data Firehose encryption supports S3 server-side encryption with AWS Key Management Service (AWS KMS). Alternatively, you can encrypt the data with your custom key. Kinesis Data Firehose can concatenate and deliver multiple incoming records as a single S3 object to reduce costs and optimize throughput.

AWS Snow Family provides an offline data transfer mechanism. AWS Snowball delivers a portable and ruggedized edge computing device for data collection, processing, and migration. For exabyte-scale data transfer, you can use AWS Snowmobile to move massive data volumes to the cloud.

DistCp provides a distributed copy capability to move data in the Hadoop ecosystem. S3DisctCp is an extension to DistCp optimized for moving data between Hadoop Distributed File System (HDFS) and S3. This blog provides information on how to move data between HDFS and S3 using S3DistCp.

Cataloging

One common challenge with a data lake architecture is the lack of oversight on the contents of raw data stored in the data lake. Organizations need governance, semantic consistency, and access controls to avoid the pitfalls of creating a data swamp with no curation.

AWS Lake Formation can manage data ingestion via AWS Glue by automatically classifying data and storing definitions, schema, and metadata in a central data catalog. Lake Formation has built-in machine learning capabilities for deduplication and finding matching records to improve data quality. For faster analytics, Lake Formation converts data into Apache Parquet and ORC before storing it in your S3 data lake. You can define access policies, including table and column level access controls, or enforce data encryption at rest. With consistent security enforcement, your users can access and analyze a curated and centralized dataset using their choice of analytics and machine learning services.

AWS Glue DataBrew, a visual data preparation tool, allows data owners, subject matter experts, or users of all skill sets to participate in the data preparation process. Without having to write any code, your teams can choose from over 250 pre-built transformations to automate data preparation tasks, including filtering data anomalies, converting data to standard formats, and correcting invalid values. The transformed data is ready for advanced analytics and machine learning projects.

Security

Amazon Connect segregates data by AWS account ID and Amazon Connect instance ID to ensure authorized data access at the Amazon Connect instance level.

Amazon Connect encrypts personally identifiable information (PII) contact data and customer profiles at rest using a time-limited key specific to your Amazon Connect instance. S3 server-side encryption secures both voice and chat recordings at rest using a KMS data key unique per AWS account. You maintain complete security control to configure user access to call recordings in your S3 bucket, including tracking who listens or deletes call recordings. Amazon Connect encrypts the customer voiceprints with a service-owned KMS key to protect customer identity. All data exchanged between Amazon Connect and other AWS services, or external applications is always encrypted in transit using industry-standard transport layer security (TLS) encryption.

Securing a data lake requires fine-grained controls to ensure authorized data access and use. S3 resources are private and only accessible only by their resource owner by default. The resource owner can create a combination of resource-based or identity-based IAM policies to grant and manage permissions to S3 buckets and objects. Resource-based policies such as bucket policies and ACLs are attached to resources. In contrast, identity-based policies are attached to the IAM users, groups, or roles in your AWS account.

We recommend identity-based policies for most data lake environments to simplify resource access management and service permission for your data lake users. You can create IAM users, groups, and roles in AWS accounts and associate them with identity-based policies that grant access to S3 resources.

The AWS Lake Formation permission model works in conjunction with IAM permissions to govern data lake access. The Lake Formation permission model uses a database management system (DBMS)-style

GRANT or REVOKE mechanism. IAM permissions contain identity-based policies. For example, a user must pass permission checks by both IAM and Lake Formation permissions before accessing a data lake resource.

AWS CloudTrail tracks Amazon Connect API calls, including the requester's IP address and identity and the request's date and time in CloudTrail Event History. Creating an AWS CloudTrail trail enables continuous delivery of AWS CloudTrail logs to your S3 bucket.

Amazon Athena Workgroups can segregate query execution and control access by users, teams, or applications using resource-based policies. You can enforce cost control by limiting data usage on the Workgroups.

Monitoring

Observability is essential to ensure the availability, reliability, and performance of a contact center and data lake. Amazon CloudWatch provides system-wide visibility for resource utilization, application performance, and operational health. Log relevant information from Amazon Connect contact flows to Amazon CloudWatch and create real-time notifications when operational performance falls below predefined thresholds.

Amazon Connect sends the instance's usage data as Amazon CloudWatch metrics at a one-minute interval. Data retention for Amazon CloudWatch metrics is two weeks. Define log retention requirements and lifecycle policies early on ensure regulatory compliance and cost savings for long-term data archival.

Amazon CloudWatch Logs provides a simple way to filter log data and identify non-compliance events for incident investigations and expedite resolutions. You can customize contact flows to detect high-risk callers or potentially fraudulent activities. For example, you can disconnect any incoming contacts that are on your predefined Deny list.

Analytics

A contact center data lake built on a descriptive, predictive, and real-time analytics portfolio helps you extract meaningful insights and respond to critical business questions.

Once your data lands in the S3 data lake, you can use any purpose-built analytics services such as Amazon Athena and Amazon QuickSight for a wide range of use cases without labor-intensive extract, transform, and load (ETL) jobs. Alternatively, you can bring your preferred analytics platforms to your S3 data lake. Refer to this blog for a walkthrough on analyzing Amazon Connect data with Amazon Athena, AWS Glue, and Amazon QuickSight.

The overall contact center service quality can make a significant and lasting impact on the customer's impression of your organization. Measuring call quality is essential to ensure a consistent customer experience. This blog describes capturing real-time call metrics using AWS Lambda and Amazon API Gateway, indexing data into an Amazon Elasticsearch Service cluster, and visualizing audio quality metrics such as increased latency or packet loss using Kibana.

For a highly scalable data warehousing solution, you can enable data streaming in Amazon Connect to stream CTRs into Amazon Redshift via Amazon Kinesis.

Machine learning

Building a data lake brings a new paradigm to contact center architecture, empowering your business to deliver enhanced and personalized customer service using machine learning (ML) capabilities.

Traditional ML development is a complex and expensive process. AWS provides the depth and breadth of high-performance, cost-effective, scalable infrastructure, and flexible ML services for any ML project or workload.

Amazon SageMaker is a fully managed service that enables your data scientists and developers to build, train, and deploy ML models for contact center use cases at scale. Data preparation contributes up to 80% of data scientists' time. Amazon SageMaker Data Wrangler simplifies and accelerates the data preparation and feature engineering from various data sources using over 300 built-in data transformations without writing any code. You can store standardized features in the Amazon SageMaker Feature Store to enable reuse and share with the rest of your organization.

Reducing friction in a customer journey is essential to avoid customer churn. To add intelligence to your contact center, you can build AI-powered conversational chatbots using Amazon Lex automatic speech recognition (ASR) and natural language understanding (NLU) capabilities. Customers can perform self-service tasks such as password reset, account balance check, and appointment scheduling via chatbots without speaking to the human agents. To automate the contact center's frequently asked questions (FAQs), you can build a question and answer (Q&A) chatbot with Amazon Lex and Amazon Kendra. Enabling text logging in Amazon CloudWatch Logs and saving audio inputs in S3 enables you to analyze conversation flow, improve conversational design, and increase user engagement.

Understanding caller-agent dynamics is essential to improve the overall service quality. See this blog on how to stream voice recordings to Amazon Transcribe via Kinesis Video Stream for speech recognition, and transform audio to text and run sentiment analysis on the transcripts using Amazon Comprehend.

For organizations with an international presence, you can build a multilingual voice experience in Amazon Connect using Amazon Polly or Amazon Translate for language translation.

Traditional financial planning software creates forecasts based on historical time-series data without correlating inconsistent trends and relevant variables. Amazon Forecast provides up to 50% higher accuracy using machine learning to discover the underlying relationship between time-series data and other variables such as product features and store locations. With no machine learning experience required, you can easily create an agent demand or inventory forecast by providing time-series and associated data in your S3 bucket to Amazon Forecast. You can encrypt confidential content using AWS KMS and control access to Amazon Forecast using IAM policy. Amazon Forecast trains and hosts a custom machine learning model in a highly available environment. You can generate highly accurate business forecasts quickly without managing any infrastructure or complex machine learning process.

Amazon Connect provides call attributes from telephony carriers, such as voice equipment's geographic location to show where the call originated, phone device types such as landline or mobile, number of network segments the call traversed, and other call origination information. Using the fully managed Amazon Fraud Detector, you can create a ML model to identify potentially fraudulent activity by combining your datasets with Amazon Connect call attributes. For example, you can customize the contact flow to intelligently route phone calls with potential fraud signals to a specialized agent.

Conclusion and further reading

Amazon Connect is a purpose-built omnichannel cloud contact center that provides a seamless and frictionless experience for your customers and agents. You can simplify operations, improve agent efficiency, and lower contact center cost with Amazon Connect.

Amazon S3 is a scalable, durable, and reliable service to build and manage a secure data lake at scale for contact centers. You can store all your contact center data as-is in the S3 data lake without restructuring the data, accelerating value extraction with shorter time-to-value. Your employees and stakeholders can run various analytics on the contact center data lake, including big data processing, real-time dashboards and visualizations, and ML to guide data-driven business decisions.

An efficient and streamlined contact center data lake can be a key driver for improving customer experience and developing market adoption. With a comprehensive portfolio of analytic services and scalable infrastructure on AWS, you can harness the power and unleash the intelligence of your contact center data lake to accelerate business growth.

Further reading

For additional information, see:

- Data Lake Storage on AWS
- Analytics on AWS

Document history and contributors

To be notified about updates to this whitepaper, subscribe to the RSS feed.

update-history-change	update-history-description	update-history-date
Initial publication (p. 15)	Whitepaper first published	May 13, 2021

Contributors

Contributors to this document include:

- Ankur Taunk, Senior Specialist Solution Architect: Amazon Connect, Amazon Web Services
- Cher Simon, Senior Partner Solutions Architect, Amazon Web Services

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.