Persistent NORC at the UNIVERSITY of CHICAGO

Case Study

Data Platform-as-a-Service utilizing DevOps for NORC

Contents

About the client 3

Problem statement 3

What we proposed 3

Outcome results and benefits 4

About the Client

National Opinion Research Center (NORC) is one of the largest independent social research organizations in the US established in 1941. Its corporate headquarters is located in downtown Chicago, with offices in several other locations throughout the US. They are an objective non-partisan research institution that delivers reliable data, statistically accurate surveys, and research analysis to help guide business and policy decisions.

Problem Statement

NORC launched an initiative for data and analytics as a new competency to release to their customers as “Data Platform-as-a-Service”. The platform was to be built using cloud technologies to provide the benefits of no upfront, low operating cost.The platform also needed to be able to scale to support multiple use cases in a multi-tenant environment. It should support variety of data (structured, sensor data/unstructured data) include flexible ETL & data pipeline mechanisms, batch & real-time data processing and be compliant with SOC 2, FISMA.

What we Proposed

Utilizing AWS DevOps tools and best practices, the Persistent team built a multi-tenant scalable data platform using IaaS, PaaS, and serverless services from AWS, implements use cases, and provides production support to end users. Some of the AWS services used were IAM, EC2, EBS, RDS, Redshift, S3, Lambda, SES, Kinesis Firehose, and Glue. Persistent implemented security and compliance on the AWS infrastructure using VPC, IAM, GuardDuty, Inspector, and KMS.

DevOps

Infrastructure as Code (IaC) and Automation: The infrastructure necessary for the platform is provisioned through infrastructure-as-code using CloudFormation to automate the creation and deployment. The automated process builds the infrastructure and applications in a repeatable and consistent manner and saves a significant amount of manual provisioning time.

Chef is used for configuration management on Linux systems. The Chef server is created using the AWS OpsWorks for Chef Automate server.The cookbook contains the recipes for configurations related to the CIS standards, NTO, SSH. SaltStack is used to automate the Linux patch management.It allows the scheduling of patches to be downloaded and installed on Linux systems. Python Scripting is used in multiple areas of the platform, mostly for automation. Various tasks such as infrastructure automation, test automation, and report generation are developed using Python.

CI/CD and Source Control: CodeCommit is used to store all the source code and binaries. The code release pipeline is created using the combination of CodePipeline, CodeBuild, and CodeDeploy. The code repository is created in the development environment and is the single source of truth for all code. A regular code commit process is used by the development team.

The pipeline is created using CodePipeline which pulls the code from the CodeCommit repository and stores into S3 bucket for the next stages. The CloudFormation templates are used in the pipeline to create the AWS resources.

Management

Persistent manages and operates the data and analytics workloads for onboarding new use cases. AWS VPC is used to setup the infrastructure for all customer use cases. Active Directory, file transfer tools, compliance tools for virus scanning, monitoring, and auditing are also created in the management area. End user desktops are run using Citrix XenDesktop and accessible through NetScaler Gateway deployed on AWS EC2 instances.

Each customer use case is implemented in a separate customer VPC to provide isolation. Statistical tools such as R, SAS, STATA, etc. are made available to end users. Each VPC can have a variety of storage options, such as file server, database, or a warehouse.

Storage

Appropriate storage options are used for different use cases. EBS volumes are used where the customer's data is in form of flat files. Periodic snapshots are created for the EBS volumes to create backup of the data in file system. Relational databases, such as MySQL, are created using RDS where the data is in structured/relational format. Redshift is used to create a fully managed and reliable data warehouse.

ETL

ETL jobs are run for retrieving data and loading into the corresponding storage. Tools such as Glue and Informatica are used for ETL depending on the use case requirements.

Security

Persistent is also responsible for the security and compliance on the AWS infrastructure using VPC, NACL, IAM, GuardDuty, Inspector, and KMS. End customer data is securely stored in its own VPC which is physically isolated from other VPCs. Active Directory and IAM policies and used to provision access to the desktops and data. The users have access to their own data. Multi-factor authentication is implemented for users in Active Directory using FreeRADIUS server. All privileged and administrator users in AWS IAM use MFA for authentication. Data on the EBS volumes and S3 buckets are encrypted using keys stored in KMS.

Logging and monitoring

VPC flow logs and CloudTrail events are monitored by a third-party vendor who has a SOC team and creates tickets for activities which need to be reviewed.

Disaster Recovery and Failover

The management and customer VPCs are created in two different AWS regions (active-passive deployment) which are physically isolated. Active Directory and database replication are implemented within each VPC. RTO of 72 hrs and RPO of 24 hrs is maintained in case of infrastructure failure. The failover testing has been performed in the production environment during the maintenance window.

Cost optimization

The resource configuration and counts are selected in a way to meet the business requirements without over provisioning. A daily and weekly billing report is reviewed by the customer to ensure the cost is within the budget.

Outcomes and Benefit

NORC now offers one of the lowest cost data platforms on AWS to guide business and policy decisions for their customers. The platform supports a pay as you go model and allows storage and compute to scale up and down depending on the data and analytics needs. It has created a new stream of data analytics business for NORC using nextgen technology and enables them to reduce the time to market while implementing use cases for various clients.

It is now possible to provision the infrastructure required for use cases in a few hours instead of days and also avoids repetitive compliance audits due to the DevOps framework deployed. The platform is SOC2 certified and multiple NORC use cases have been implemented on the platform.