BEYONDSOFT
beyond your expectations

CASE STUDY

LEADING DIGITAL ADVERTISING AGENCY (NEW YORK) ADVANCED ANALYTICS USING VERTICA ON AWS

CHALLENGE

A leading digital advertising agency wanted to use their business intelligence tool to query their historical data for 90 days under 2 secs with concurrent users of more than 80 users at a time. Their current system included data in Hadoop clusters and Infobright DB on-prem cluster, which is not able to handle above requirements.

TECHNOLOGIES USED:

Vertica on AWS, AWS SSM, AWS Cloudwatch logs, AWS S3, AWS ELB, AWS Fargate, AWS Parameter Store, AWS ECR, Python, Jenkins etc.

SOLUTION

Beyondsoft's BigData consulting team proposed a solution using Vertica, a columnar database on AWS. Bringing into AWS helped with added scalability, elasticity and performance that the customer was looking for. The project consisted of creating Vertica clusters in a repeatable manner, and a pipeline-based approach for Vertica DDL. It also included moving large amounts of data daily to Vertica cluster.

The project involved 3 phases as below:

- Creating a repeatable deployment process through infrastructure as code (Terraform) for Vertica cluster. The Vertica AMI is procured from AWS marketplace. Beyondsoft engineers added the ability to launch different Vertica nodes as part of a cluster through tags to AWS Elastic Load Balancer (ELB). The Vertica cluster consisted of 2 AZ's in active/active nodes, with 16 nodes total. The cluster consisted of 90 days of data which came to around approximately 40TB. AWS Systems Manager Service (SSM) and Cloudwatch Logs are used for administration of the cluster. This Vertica infrastructure as code also integrates with the customer self-servicing tool for their developers.
- Pipeline based approach for Vertica DDL Vertica DDL is pushed through pipeline using Liquibase, a java framework for database change and deployment. This ensured that production Vertica clusters are not touched manually for schema changes.
- ETL from Hadoop cluster to Vertica using a producer/

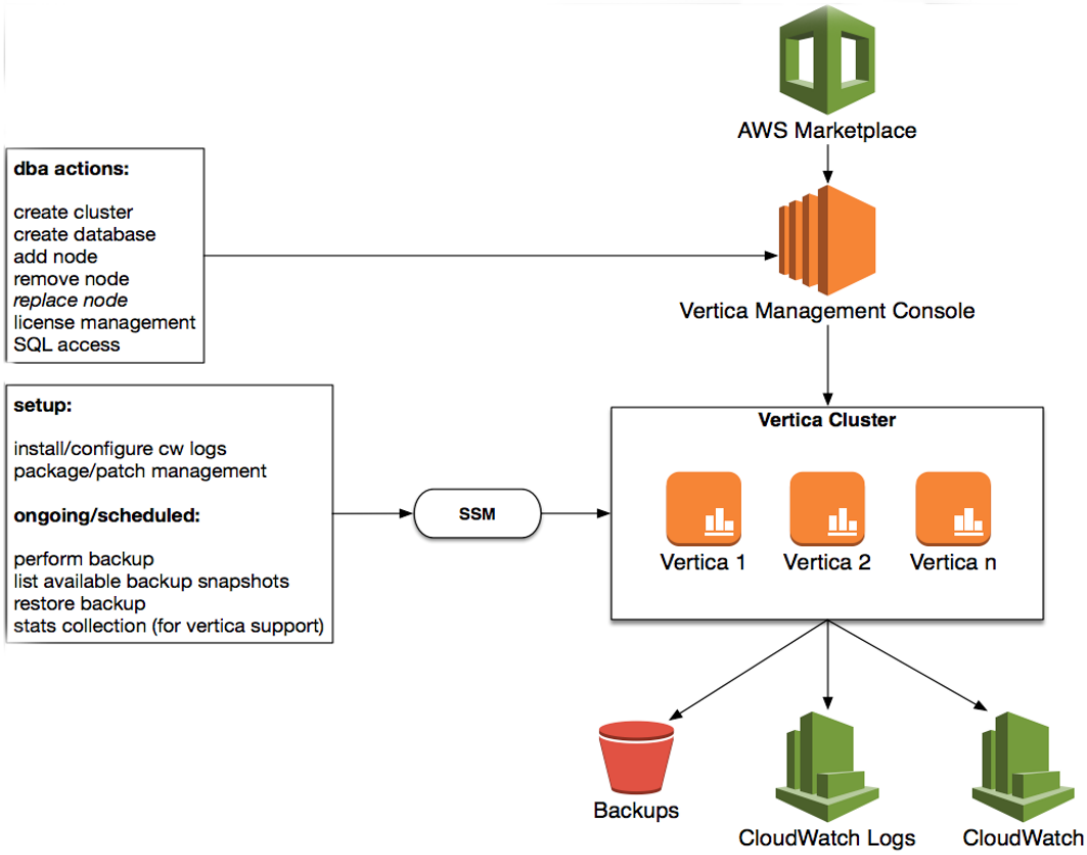
consumer pattern The ETL code is written in python code with on demand Fargate containers which extracts data from Hadoop and store it in zipped files in S3. From there, jobs are created to load data into Vertica from S3. The data is around 120Gb/day with around 570M rows loaded at its peak. The customer facing java application has several dashboards which is able to procure data from Vertica under 2 secs query time with concurrent usage.

KNOWLEDGE TRANSFER

Customer's data analytics team is educated about the new created solution, Terraform and provided with runbook so that they can not only manage but add to the solution in future. At Beyondsoft, we also educate customers in various AWS services and provide them customized training sessions on various topics.

BENEFITS

Moving from on-prem cluster to cloud increased the scalability, agility and performance of the whole solution. DevOps approach through data pipelines decreased the go the market time for the code changes. Infrastructure as Code provided the repeatable manner to create infrastructure which brought in consistency of the operations resulting in fewer bugs.



ABOUT BEYONDSOFT

Beyondsoft is a global full-service IT Solutions and Service Provider. It was founded in 1995 and headquartered in Beijing, China. Beyondsoft has over 14K employees distributed over 34 nationwide offices in United States, Japan, Indian, Spain, Canada, Singapore and Malaysia. Beyondsoft is trusted AWS Advanced Tier Partner.