CPG major increases business agility with cloud-based big data analytics platform

Client Overview

The client is a CPG major with a business scale of 1 million+ outlets, 35+ product categories, 65+ brands, and 1,100+ products. The project was carried for India/South Asia market.

The Challenge

The client needed to improve its on-shelf availability and product assortments in order to meet its ambitious growth targets. This required an understanding of changing buying patterns based on the market insights and analytics, also undertaking course corrections in the shortest possible time.

The client's existing Decision Support System (DSS) comprised an on-premise analytics platform based on machine learning and advanced predictive analytics. The platform churned nearly 500 million recommendations every month, computing around a billion records amounting to 6-8 TB of data. In order to drive competitive advantage, the client needed to significantly increase the performance of its DSS.

The high lead time required for generating Key Performance Indicators (KPIs) and building statistical models was yet another challenge was. As a result, the client was unable to leverage the latest market information for business insights. The existing platform was also not able to scale up

Executive summary

A leading CPG company needed to improve its on-shelf availability and product assortments in order to meet ambitious growth targets. Mindtree migrated the existing on premise Decision Support System to AWS Cloud platform. As a result, the company was able to reduce the lead time required for insight generation from a month to less than 10 hours for 6 TB of data, thereby significantly increasing business agility.

to accommodate the new age data sources for generating deeper insights.

The client needed a DSS based on a comprehensive data lake — with data from both traditional and new-age sources — to answer ad-hoc business questions, accurately. Not only this, the DSS would also have to perform significantly better to drive competitive advantage.

Mindtree's Solution

As a first step, Mindtree started the process of identifying opportunities to increase the performance of the DSS. We decided to move the workload to AWS to utilize the power of cloud and distributed computing.

Mindtree developed a cloud-based big data analytics platform leveraging on-demand scalability and massive parallel processing capabilities. This involved re-platforming and re-designing the entire process from data gathering to analytics and insights generation. To achieve this, Mindtree built an entirely new architecture using big data, AWS, AWS Redshift and Spark technologies with Machine learning capabilities.

The solution resulted in a comprehensive solution stack comprising a data lake which seamlessly stitched traditional data sources with new age data for newer insights. The solution leverages Mindtree's Big Data ETL framework & ABC (Audit Balance Control) framework for data integration and automation.

To provide the client with efficient performance and value for money, we opted for an AWS Trusted Advisor. This helped us stand by our promise of providing the best resource at optimised costs. Not only this, Mindtree could now focus on performance including providing fault-tolerant information.

On the security front, Mindtree was able to ensure security of resources using the access provided by the AWS Trusted Advisor. This included access to specific ports having unrestricted data, IAM User data, MFA on Root Account and service limits. Also, protection of root account credentials are considered as first priority.

We created fine grained permissions to resources under AWS account, applied them to groups and then assigned users to those groups using AWS IAM and ensured users have least privilege to accomplish tasks.

The entire security setup is inside a Virtual Private Network with limited and need-based ingress and egress to ensure high security. Firewall rules and more realistic server environment have also been setup. AWS Key Management Service (KMS) is used for encryption of data in EMR and Redshift. This makes creating, controlling and managing encrypted keys, easy. Data encryption has been done for data at rest as well.

The platform includes CloudWatch which helps collect metrics and log files. It also enables tracking of the collected metrics, setting alarms and making changes whenever required. The CloudTrail has helped administrate and control AWS account. As a result, Mindtree has been able to further log, monitor and retrospect the activities across the client's infrastructure.

Highly efficient Assortment optimization with Predictive Modeling
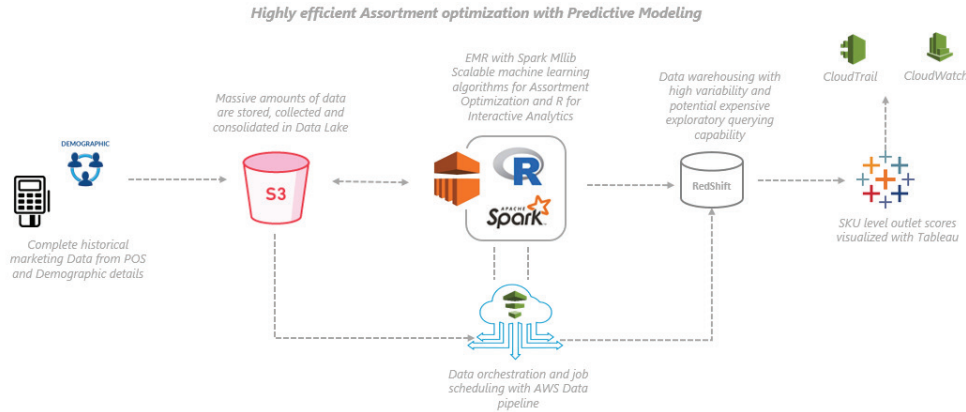
Automated Insights generation with DevOps & Automation:
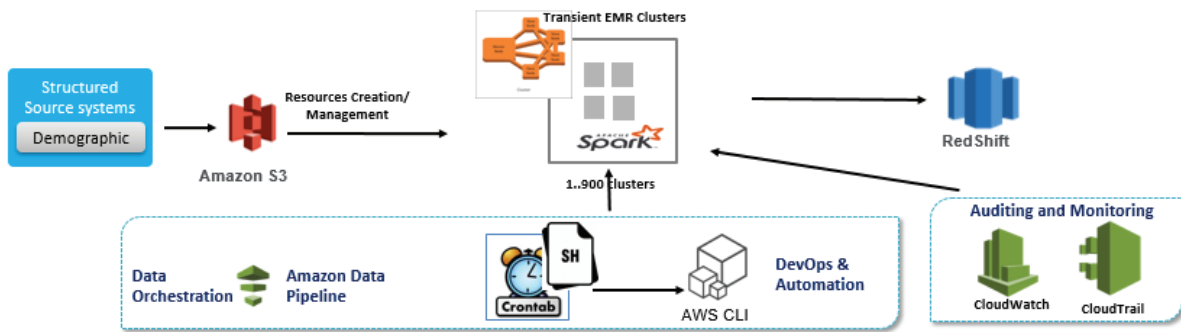
Automated insights generation is performed with shell scripts and AWS CLI automation. The analytics involves collaborative filtering distributed algorithm executed in Spark clusters where AWS DataPipeline shell command activity will be activated for creation of this transient EMR clusters and is monitored continuously with CloudWatch.

Bayesian network algorithm, which doesn't work in distributed environment is executed in parallel clusters with 70-80 segments assigned for each cluster on R images.

• Cron tab triggers shell scripts which in turn invokes AWS CLI for parallel cluster creation. The instances are based out of R images and the data is fetched using CLI automation from S3 for algorithm execution and once it is completed the data is copied to RedShift.

• Transient clusters are monitored using CloudWatch and configured to delete if it remains idle.

Effort is reduced to 8 hours from 25 days after CLI automation. Cost optimization is high where only 6000 instance hours for a span of 15 hours job with 900 clusters is used for insights generation using Bayesian network algorithm which gets automatically deleted once the job is completed.

Business Benefits

Increased business agility: The new platform offers increased frequency of insight generation. As a result, the client has been able to achieve greater business agility in terms of reacting quickly to changing buying patterns and competition scenario.

Reduction in lead time for generating insights: The solution has significantly reduced the lead time required from data processing and analysis to insight generation from a month, to less than 10 hours for 4-6 TB of data

Improved predictive capability: The new platform provides improved predictive capability as it seamlessly integrates traditional data sources with new age data sources for deeper insights.

Improved prediction accuracy: It ensures improved accuracy of prediction because of the platform's ability to use up-to-date market data for generating insights.

Better economies of scale: It has resulted in better economies of scale because of the platform's ability to perform on-demand analytics which was not possible earlier.

Reduced Churn Time: Churn time reduced from 20 days to 10 hours for approx. 5 TB of data with 30% cost savings on like to like churn compared to on premise.