CYBAGE
Delivering Value. Scientifically.

Big data analytics platform for independent fashion and style website

About the client
The client is a leading fashion and style website in the United States.

Technology problem
The client was facing the following technology issues:

- The legacy platform was facing performance challenges because report generation took up one whole day.
- High manual processing was required to consolidate data from multiple sources and generate reports. This resulted in loss of productivity.
- Multiple groups worked on the same data, resulting in duplication of efforts.
- The lack of a central reporting system (decentralized data), data inconsistency, and quality issues for WebPrints of one billion page views resulted in rework for reconciliation.

Technology solution
Cybage's solution comprised the following:

- Business Intelligence (BI) solution on Big Data platform, capitalizing on Hadoop technologies and open-source tools, using configurable ETL scheduled workflows.
- Data assimilation platform, which consolidated data from sources such as Google Analytics, Sailthru, Ooyala, Chartbeat, Disqus, and Pinterest.
- Business Intelligence (BI) solution on Big Data platform, capitalizing on Hadoop technologies and open-source tools, using configurable ETL scheduled workflows.
- Data assimilation platform, which consolidated data from sources such as Google Analytics, Sailthru, Ooyala, Chartbeat, Disqus, and Pinterest.
- Extractor for individual sources that were run on AWS-EC2, raw JSON files were stored on AWS-S3, ETL and analytics were done on AWS-EMR cluster using Hive and Impala, result TSVs were stored in AWS-S3, results were copied to AWS-RDS (MySQL) using copy activity in data pipeline, and the reports were run on this data. All the schedules were managed by AWS–Data pipeline.
- Rich reports (Self-service BI and Ad-hoc reporting) on aggregated data.
- Data mining capabilities for identification of specific patterns and trends.

Execution strategy
As part of its execution strategy, Cybage did the following:

- Capitalized on phased methodology, which enables targeted solution development (infrastructure, centralized DB, ETL (Extract, Transform and Load), reporting, Self-service BI, and data mining capability).
- Capitalized on SCRUM Agile methodology for active tracking, greater agility, and recurrent deployments.

Value realized
The solution provided the following benefits to the client:

- Ability to develop comprehensive reports with higher quality and in-depth data analysis, aiding informed decisions.
- Faster and efficient data processing on a centralized database.
- Encouragement to end-users to use Self-service BI for analyzing and visualizing data.
- 75% reduction in report generation efforts.

Tools and technologies
Cybage used the following tools and technologies:
Development Java, MapReduce, Pig, Hive, Impala
Testing Manual testing
Tools AWS-S3, AWS-EC2, AWS-EMR, AWS-RDS (MySQL)

Cybage services utilized
Architectural Services, Development, Testing, Big Data and BI CoE capabilities.
Technology