Big data.
Big opportunities.

Exploring the benefits of AWS Data Lake

January 2020
THE POWER OF ANALYTICS

“Data is the new oil”
It’s a phrase you hear everywhere, from social media to client meetings. So, it’s worth considering a layman’s definition of big data by a principal architect from one of the tech giants:

“Any data that cannot be processed optimally with the tools and technologies that are available at that period of time”

10 GB of data was big a few decades back, but a social media giant is now ingesting about 4 PB of data every day, and the number is ever-increasing.

Oil is of no value until it is extracted from the oil field. Similarly, data is of no value to enterprises until they gain insights out of it.

That’s why it’s called “the new oil”, because enterprises can now analyse any variety, at any velocity, irrespective of the volume, and gain insights that propel the business.

TRADITIONAL DATA ANALYTICS

At the high-level, the traditional approach has the following steps:

• Data from different OLTP sources like SQL Server, Oracle, MySQL etc., are imported into staging tables using ODBC connectors. Flat files are also imported. Some of the user data gathered through web apps is stored directly in the staging area.
• The transformation logic is run on the staging tables with any of BI tools/custom logic
• Transformed data is loaded into a data warehouse and the analytics on the environment is processed

The limitations of the traditional approach are:

• The architecture predominantly revolves around structured data, while we have to deal with many varieties of data (structured, unstructured, semi-structured, streaming, social media etc)
• The data warehouse was not optimised for faster performance. Not too long ago, we were working on nightly ETL jobs. Based on the volume of data and the complexity of the transformation we have seen SLAs of 24-72 hours.
• Scalability is a major issue at all layers

For more information, visit us at www.tatacommunications.com
BIG DATA PROCESSING WITH HADOOP

With the evolution of technology and devices, we can deal with large volumes of different varieties of data at very high velocity. Traditional architecture was not able to tackle these. When the world needed a distributed storage and distributed processing software platform, Hadoop was born.

Hadoop solved most of our problem statements:
• It can ingest any kind of data
• Hadoop cluster is built on the Hadoop Distributed File System (HDFS), distributed storage that spans across multiple nodes, making it highly available (multiple copies are stored across nodes)
• The developer’s familiarity factor is also taken into consideration. For example, a BI developer who is more comfortable in SQL can still use Hive SQL for their analytics purposes.

The limitations of Hadoop:
• Even though storage is distributed, it is tightly coupled with compute
• Even though storage is scalable, it is not auto-scalable
• A big data solution may not work for your existing IT team. You need specialised skills to deploy, manage, maintain and run analytics on a big data ecosystem.

Since Hadoop is open source, it led to the creation of lots of open source components that fit into its ecosystem.

Enterprises can now leverage the power of Hadoop ecosystem and forget about the pain of deploying, managing, scaling, and maintaining the ecosystem. They don’t need to maintain an analytics environment and can concentrate on the code that’s doing the analytical workload.

For more information, visit us at www.tatacommunications.com
Ingestion layer
AWS offers fully managed services for ingesting any kind of data:
- AWS DMS can ingest data from SQL and NoSQL sources to the S3 bucket
- AWS Kinesis can ingest streaming data
- Log files and flat files can be uploaded directly to S3 by AWS CLI, Console or SDK
- If the volume of data is very big, AWS Snowball can be used to transport data to S3

ETL layer
- AWS Glue offers a fully managed ETL service
- The ETL layer is also called ‘The Curation Zone’, where data is cleansed, deduplicated, transformed etc., which prepares it for analysis
- AWS Glue in the background, takes care of spinning up the Hadoop cluster and scaling it. It’s based on your transformation workload and discards the cluster after your workload is completed, charging you only for the duration of your job.
- All the transformed data is stored in an S3 bucket, with metadata stored in the AWS Glue data catalogue

For more information, visit us at www.tatacommunications.com
Analytics layer
AWS has a range of fully managed analytics services that cater to professionals including data analysts, data scientists, BI professionals and more.

• Amazon Athena is an interactive query service that makes it easy to analyse data in Amazon S3 using standard SQL
• AWS EMR (Elastic MapReduce) allows you to deploy your Hadoop ecosystem in AWS cloud with few clicks. While AWS takes care of the availability, scalability, manageability etc, Big data engineers can focus on writing the MapReduce jobs that extracts the insights from data in S3.
• AWS EMR gives analytical teams the engines and elasticity to run petabyte-scale analysis for a fraction of the cost of traditional on-premises clusters
• Amazon QuickSight is a fast, cloud-powered business intelligence service that makes it easy to deliver insights to everyone in your organisation. It’s also fully managed
• Amazon Rekognition makes it easy to add image and video analysis to your applications using proven, highly scalable, deep learning technology
• Amazon SageMaker is a fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning (ML) models quickly

This is not the complete list of Analytics services offered by AWS, there are many more please feel free to explore.

AWS Data Lake has the solution to all the limitations discussed in relation to Hadoop:

• Storage and compute are not tightly coupled. All data is in a centralised S3 bucket with the transformation and analytics workload taken care of by separate AWS services.
• It is fully managed, so, you don’t have to worry about deploying, managing, maintaining, HA etc.
• It offers different analytic services, offering developers the option to choose their language, for example, Python, Scala, R, Java etc.

READY TO TAKE THE NEXT STEP?

At Tata Communications we can help you design, deploy, and manage, the right solution for you, following the best practices for security, encryption, and high availability.

If you are new to AWS, want to set up Data Lake solutions for Analytics in AWS, or need to transform your existing analytics solution, we’d love to hear from you.

For more information, visit us at www.tatacommunications.com