February 2, 2024

Data Storage Technologies for Big Data and Analytics

DSMM - Maple Mapping

PREPARED BY:

Auradee Castro (c0866821)

Bhumika Rajendra Babu (c0867081)

Lakshmi Kumari (c0867090)

Maricris Resma (c0872252)

TABLE OF CONTENTS

I. Case Study Overview

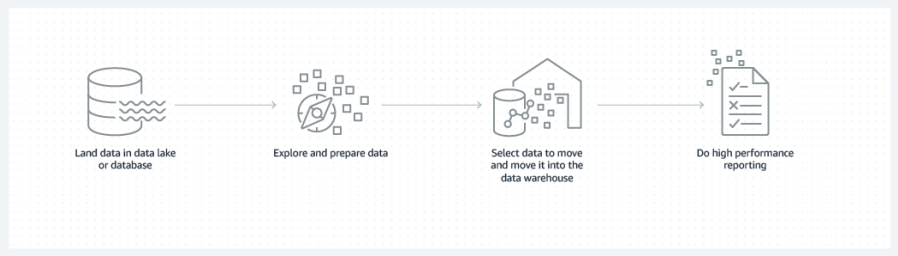
As a company that specializes in creating customized digital maps, Maple Mapping manages extensive geographical data in various formats like geospatial data, images, texts along with other customer data. Due to collecting different data from a variety of sources, the company requires modern storage solutions to manage them. Therefore, the goal of this project is to research different data storage technologies relevant to big data and analytics and create a podcast that would summarize key findings.

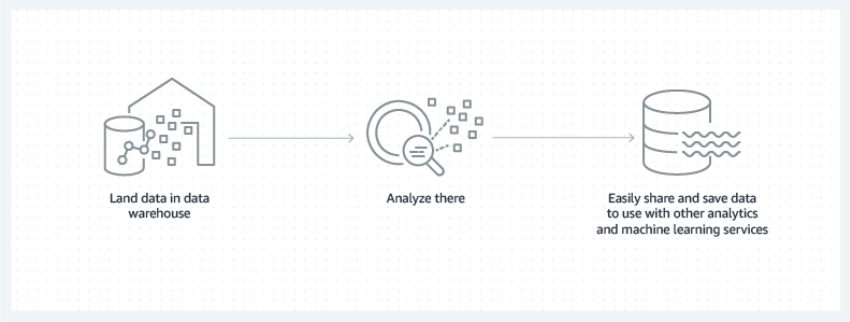
II. Data Storage Technologies

Introduction

In data analytics, a complete architecture spans from data collection, storage, processing, analysis and visualization but we will only focus on the storage part. Capacity, performance, scalability, and efficiency are just one of the many considerations that should be assessed when looking for the best one. We must ensure data movement would be quick and that the storage systems would be scalable, meaning able to expand without disrupting existing workflows. Lastly, we also want something that is cost efficient without sacrificing the quality of the storage solution itself.

Practically in terms of storing data we need both Data Lake and Data Warehouses. Data Lakes store raw data while Data warehouses store the relational structured data and/or treated/processed data. The layer for the warehouses is closer to and are specifically designed for analytics.

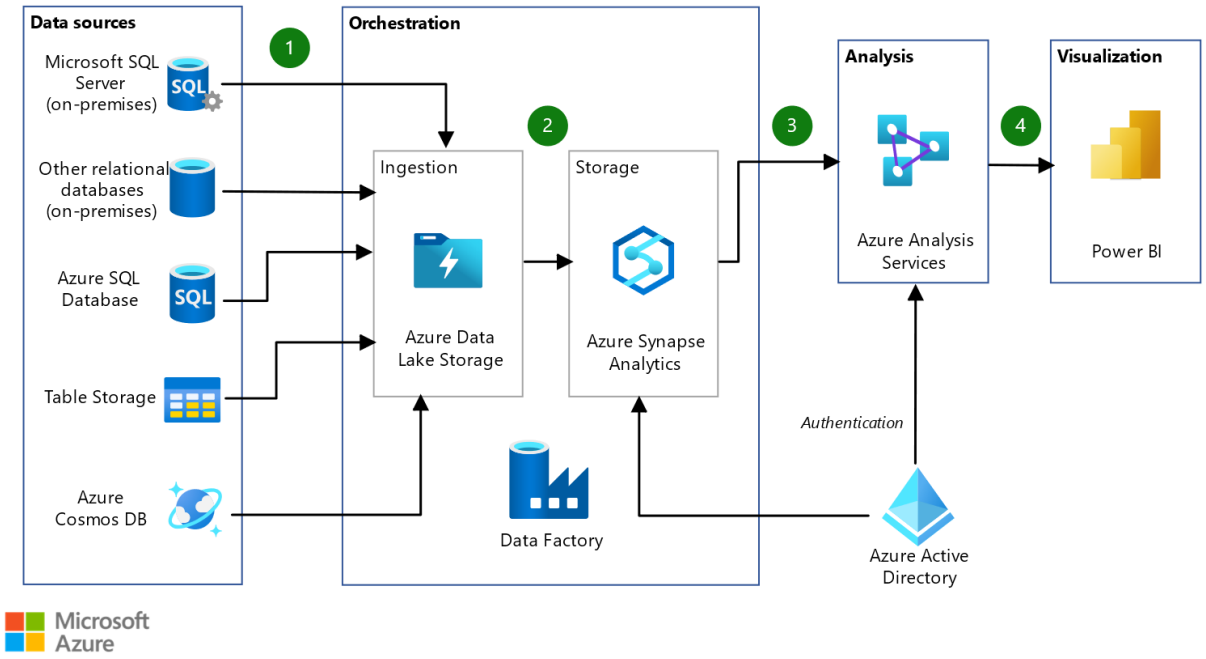




Our company utilizes the combination of the two in order to really complete the whole architecture from data ingestion, storage, processing and analysis, but for this specific study we shall focus on comparing only the various data lake storage in relation to their use cases for big data storage and integration into a whole IT solution architecture for data management and analytics. Here we list the 3 different data storage technologies and their key features.

1. Azure Data Lake Storage

Azure Data Lake is a cloud-based storage provided by Microsoft Azure. It is designed to handle large amounts of data in various formats, such as structured, semi-structured, and unstructured data.

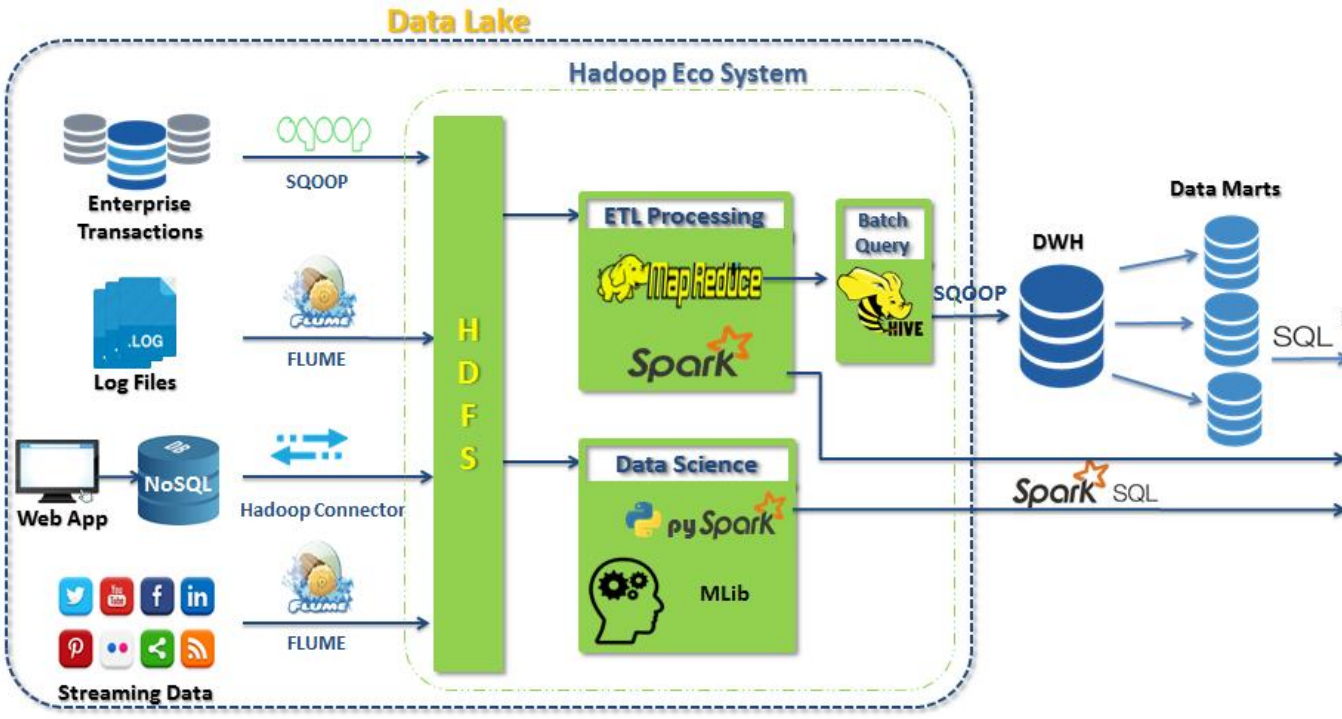


- Features: Azure Data Lake Storage Gen1 is a hyper-scale repository for big data analytics workloads. Offers features like a hierarchical namespace, security features, and integration with Azure services.
- Storage Capacity: Virtually unlimited, with individual file sizes up to the limits of the storage account.
- Performance: Designed for high performance and scalability, with support for parallel processing and optimized for big data analytics workloads.

- Performance Requirements: Suitable for demanding performance requirements, with options for optimizing performance based on access patterns.
- Cost: Pay-as-you-go pricing based on storage used, transactions, and data transfer. Costs can vary based on storage class and access patterns.
- Integration Needs: Integrates tightly with other Azure services and tools, providing a comprehensive data analytics platform.
- Advantages: Strong integration with the Azure ecosystem, hierarchical namespace (Gen2), and strong security features.
- Disadvantages: Cost management can be complex, and costs can increase as storage and access levels scale.

2. Hadoop HDFS

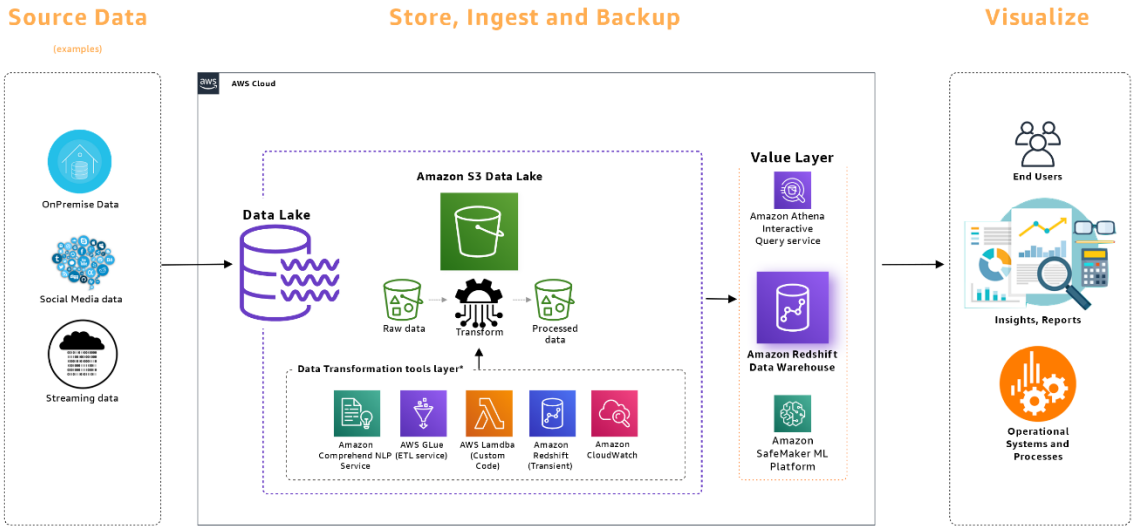
Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications. It is designed to store and manage large volumes of data across a distributed network of computers. HDFS is inspired by the Google File System (GFS) and is highly fault-tolerant, scalable, and reliable. Below is a sample architecture of how it is utilized as storage.



- Features: Distributed file system designed to store large volumes of data across a cluster of commodity hardware. Fault-tolerant and highly scalable. Supports replication and data locality for improved performance.
- Storage Capacity: Scales horizontally, so capacity is virtually unlimited and can be expanded by adding more nodes to the cluster.
- Performance: Optimized for handling large files and batch processing. Performance can vary based on cluster size, configuration, and workload.
- Performance Requirements: Requires a cluster of machines with sufficient storage and processing power to handle big data workloads.
- Cost: Open-source software, so no direct cost for the software itself, but requires hardware and maintenance costs for the cluster.
- Integration Needs: Integrates well with the Hadoop ecosystem and related tools for data processing and analytics.
- Advantages: Scalable, fault-tolerant, and designed for big data workloads.
- Disadvantages: Requires a complex setup and maintenance of a Hadoop cluster, which can be challenging and costly. Not as efficient for small file processing.

3. Amazon S3

A data lake built on Amazon Simple Storage Service (Amazon S3) provides the ideal target layer to store, process, and cycle data over time. As the central aspect of the architecture, Amazon S3 allows the data lake to hold multiple data formats and datasets. It can also be integrated with most if not all AWS services and third-party applications. Below is a sample architecture of how it is utilized as storage.



- Features: Scalable, durable, and secure object storage. Supports various storage classes for different access needs. Integrates with AWS services and third-party tools. Offers features like versioning, lifecycle management, and encryption.
- Storage Capacity: Virtually unlimited, with individual object sizes up to 5 TB.
- Performance: Designed for high durability and availability. Performance can vary based on storage class and configuration. Query performance can be enhanced using services like Amazon Athena or Redshift Spectrum.
- Performance Requirements: Suitable for a wide range of performance requirements, from low-latency access to infrequently accessed data.
- Cost: Pay-as-you-go pricing based on storage used, requests made, and data transfer. Costs can vary based on storage class and access patterns.
- Integration Needs: Integrates seamlessly with other AWS services and third-party tools through APIs and SDKs.
- Advantages: Easy to use, highly durable, scalable, and integrates well with the AWS ecosystem.
- Disadvantages: Direct query performance can be slower compared to purpose-built data lake solutions. Costs can increase as storage and access levels scale.

III. Comparison Matrix

Feature	Hadoop	Azure Data Lake Storage Gen2	Amazon S3
Storage Type	Distributed File System	Object Storage	Object Storage
Elasticity	No	Highly Elastic	Highly Elastic
Cost/TB/month	$206	$20	$23
Availability	Depends on cluster setup	99.9%	99.99%
Durability	Depends on replication	99.999999999%	99.999999999%
Scalability	Scalable, but manual	Highly Scalable	Highly Scalable
Performance	Depends on cluster setup	High	High
Integration	Hadoop ecosystem	Azure ecosystem	AWS ecosystem
Security	Kerberos, encryption	Azure AD, encryption	IAM, encryption
Real-time Data	Limited	Yes	Yes
Use Cases	Big data processing	Data lake, analytics	Data storage, analytics

III. RECOMMENDATION

To summarize, S3 and Azure cloud storage offer scalability, significantly higher availability and durability, and twice the performance compared to traditional HDFS clusters, all at a much lower cost.

Hadoop and HDFS revolutionized big data storage by making it affordable to store and distribute vast amounts of data. However, in a cloud-native setup, the advantages of HDFS are limited and not worth the added operational complexity. This is why many organizations choose not to use integrated cloud storage.

Cloud based storage is the clear winner so now choosing between Azure and AWS we recommend going for AWS S3 because Amazon S3 is one of the already widely adopted cloud storage services and has integration with various computational services such as Amazon Kinesis. The company also already specializes on the AWS ecosystem thus it’s better to be consistent.

IV. Podcast Episode: Data Storage Technologies

The Podcast created for this project highlights the key features of three data storage technologies with the comparison matrix. It is a podcast video presentation and at the end it concludes which storage technology the team has decided as the best one for Maple Mapping to utilize. The link for the video can be found: xxxxxxxxxxxxxxx

V. References

Amazon Redshift. (n.d.). Retrieved from https://aws.amazon.com/redshift/

Analytics end-to-end with Azure Synapse. (n.d.). Retrieved from https://learn.microsoft.com/en-us/azure/architecture/example-scenario/dataplate2e/data-platform-end-to-end?tabs=portal

Announcing General Availability of Qubole on Google Cloud. (n.d.). Retrieved from https://www.qubole.com/blog/announcing-general-availability-of-qubole-on-google-cloud

Automated enterprise BI. (n.d.). Retrieved from https://learn.microsoft.com/en-us/azure/architecture/reference-architectures/data/enterprise-bi-adf

AWS to Azure services comparison. (n.d.). Retrieved from https://learn.microsoft.com/en-us/azure/architecture/aws-professional/services#big-data-and-analytics

Benefits of Modernizing On-premises Analytics with an AWS Lake House. (n.d.). Retrieved from https://aws.amazon.com/blogs/architecture/benefits-of-modernizing-on-premise-analytics-with-an-aws-lake-house/

Google Cloud to Azure services comparison. (n.d.). Retrieved from https://learn.microsoft.com/en-us/azure/architecture/gcp-professional/services#big-data-and-analytics

Storage, A. D. (n.d.). Retrieved from https://azure.microsoft.com/en-ca/products/storage/data-lake-storage#features

The Best Place to Store Your Data: Amazon Redshift vs. Amazon Simple Storage Solutions (S3). (n.d.). Retrieved from https://www.zuar.com/blog/amazon-redshift-vs-amazon-simple-storage-solutions-s3/

What is Data Lake | Understand the Data Lake Architecture | Data Lake using Apache Spark. (n.d.). Retrieved from https://www.youtube.com/watch?v=B6RDjs7D-qY

What is Data Warehouse. (n.d.). Retrieved from https://aws.amazon.com/what-is/data-warehouse/

Feature	Amazon S3	Amazon Redshift	Google Cloud Storage (GCS)	Google BigQuery	Azure Data Lake	Azure Synapse
Type	Object storage	Data warehouse	Object storage	Serverless Data warehouse	Scalable data lake	Data warehouse
Scalability	Highly scalable	Scalable	Highly scalable	Highly scalable	Highly scalable	Scalable
Pricing Model	Pay-as-you-go	Pay-as-you-go	Pay-as-you-go	Pay-as-you-go	Pay-as-you-go	Pay-as-you-go
Data Formats	Supports various data formats	Structured data	Supports various data formats	Supports structured and semi-structured data	Supports various data formats	Supports various data formats
Query Performance	N/A	High performance queries	N/A	Optimized for fast query processing	Supports parallel processing for fast query performance	High performance queries
Integration	Integrates with other AWS services	Integrates with other AWS services	Integrates with other Google Cloud services	Integrates with other Google Cloud services	Integrates with other Azure services	Integrates with other Azure services
Security	Offers encryption and access control	Offers encryption and access control	Offers encryption and access control	Offers encryption and access control	Offers encryption and access control	Offers encryption and access control
Real-time Data	N/A	N/A	N/A	Supports real-time analytics	Supports real-time analytics	Supports real-time analytics
Use Cases	Ideal for storing and retrieving large objects	Ideal for data warehousing and analytics	Ideal for storing and retrieving large objects	Ideal for ad-hoc queries and interactive analysis	Ideal for storing and analyzing large volumes of data	Ideal for data warehousing and analytics

This matrix provides a high-level overview of the key features of Amazon S3, Amazon Redshift, Google Cloud Storage, Google BigQuery, Azure Blob Storage, Azure Data Lake,

and Azure Synapse, helping you understand their strengths and suitability for different big data and analytics use cases.

Storage Type
Elasticity
Cost/TB/month
Availability
Durability
Scalability
Performance
Integration
Security
Real-time Data
Use Cases