DATAMATICS

WHITEPAPER

BUILDING MODERN DATA PLATFORM WITH APACHE HADOOP (DATA LAKES)

TABLE OF **CONTENTS**

INTRODUCTION

Organizations are collecting and analyzing large amount of data from disparate data sources using enterprise data warehouses and analytics tools making it difficult to manage the storage, processing, and analytics at scale.

Apache Hadoop provides an ideal storage solution for **Enterprise Data Lakes** either on-premise or on-cloud that is scalable enough to manage petabytes of data. The data platform provides multitudes of integration capabilities with traditional database systems, analytics tools as well as in-house query engines with business reporting to facilitate extract, transform, and load processes.

APACHE HADOOP PROVIDES THESE BENEFITS THROUGH A TECHNOLOGY CORE COMPRISED OF:

01

Hadoop Distributed Filesystem (HDFS)

It is a highly scalable, fault-tolerant, distributed storage file system that works closely with a wide variety of concurrent data access applications, coordinated by YARN service.

02

Apache Hadoop YARN

It is an architectural center of Enterprise Hadoop platform. YARN allows multiple data processing engines such as interactive SQL, real-time streaming, data science, analytics workbench, and batch processing to handle data stored in a single platform and provide comprehensive and efficient analytics at scale.

IN ADDITION TO THE ABOVE, HADOOP ALSO EXTENDS HIGH LEVEL OF SUPPORT FOR:

01

Apache Spark

It is a lightning fast unified analytics engine for large-scale data processing, which runs numerous workloads (100x faster) than any modern Big Data Analytics system engine. It provides ease of use such that parallel applications can be written in Java, Scala, Python, R, and SQL. Spark provides great support to combine SQL, streaming, complex analytics and runs on a majority of Big Data Analytics systems.

BUSINESS CHALLENGE

Organizations collect and analyze large amounts of data from disparate data sources using Enterprise Data Warehouses and Analytics Tools making it difficult to manage storage, processing, and analytics at scale. Data Marts in the traditional Enterprise Data Warehouses are designed individually that doesn't allow organizations to perform comprehensive and efficient analytics. This limits organization's ability and agility to derive more value from its data, and capability to integrate with other analytics tools and processes.

Hadoop based Data Lake provides a centralized repository, which is scalable across numerous machines. It allows ingesting and storing structured and unstructured data, processing and transforming it, as required. It helps in conducting data discovery, reporting, advanced analytics, and visual reporting on the stored data regardless of its native form. It also helps the business savvy users to draw insights in real-time or near-real time and access it on-demand.

This paper discusses each of these options in detail and provides best practices to build Hadoop based Enterprise Data Lake. The following figure illustrates the high level architecture of Hadoop based Data Lake platform.

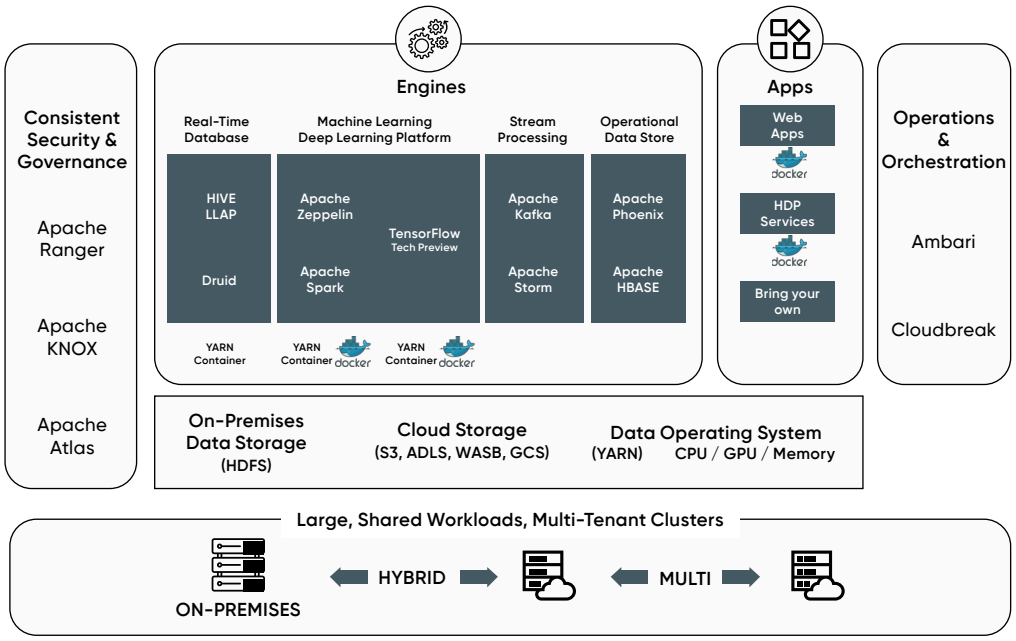


Figure: Hadoop based Data Lake platform (High-Level Architecture)

PROPOSED SOLUTION

ENTERPRISE HADOOP WITH DATA LAKE PLATFORM - DATA PLATFORM OPTIMIZATION & ANALYTICS

Organizations can leverage Data Lakes to re-architecture Enterprise Data Warehouse (EDW) or setup a new initiative to build Data Lake platform as per the organization needs. Data Lake can also be a source as well as sink to Data Warehouse for multiple types of data (structured, semi-structured and unstructured), and can be analyzed ad hoc to quickly explore unknown datasets and discover new insights.

As organizations extend their Data Warehouses with Data Lakes, it enables uniform querying capabilities, data science and workbench support.

Following are the common best practices for building Data Lake solution either as a new initiative or as a re-architecture of Data Warehouse systems:

1. Configure Data Lake to be flexible and scalable to aggregate and store all types of data following an organization's growth.
2. Include Big Data Analytics' ecosystem components, which support data encryption, with Apache Ranger & Knox, search using Apache Solr, complex analysis using Apache Spark, interactive analytics with Druid and querying with Apache Presto and Drill.
3. Implement access-control policies to authenticate users and data security mechanisms to protect stored data in Data Lake.
4. Provide mechanisms, which enable users to quickly and easily search and retrieve relevant data, and perform new types of data analysis.

KEY ELEMENTS OF A DATA LAKE WITH ANALYTICS SOLUTION

01

Data Movement

This step allows importing of any amount of data from disparate data sources and moving it to the unified storage platform in its native format. This process saves time by eluding data definitions, schema and transformations.

02

Securely store and catalog data

It allows storing real-time streaming data from mobile apps, IoT devices, and social media. It provides the options to understand, search and lookup where data resides in the lake through crawling, cataloging, and indexing. It enables security on the stored data across different users and departments to ensure security on data.

03

Analytics

In this step, data analytics is performed with a choice of analytics tools and frameworks. This includes popular Big Data frameworks, such as Apache Hadoop, Spark, Hive with PrestoDB, Druid, and commercial offerings from EDW and Business Intelligence vendors. Here analytics is performed comprehensively and efficiently without moving the data to a separate analytics system.

04

Machine Learning

It helps derive valuable insights including faster business reporting, perform self learning using models, which are built to predict possible outcomes and suggest actions to achieve optimal results.

05

Dashboards and Visualizations

For dashboards and visualizations, many third-party BI tools are available, which integrate well and provide faster business analytics service as well as make it easy to build dashboards that are accessible from any browser and mobile devices.

START YOUR DATA PLATFORM OPTIMIZATION WITH DATA LAKE AND ANALYTICS PROJECT

Building Modern Data Architecture with Data Lake:

Depending on the requirements, an organization requires Data Warehouse as well as Data Lake to serve the needs of modern data management and analytics.

Data Warehouses augment very well with Data Lakes ensuring business continuity and offer a broad range of Advanced Analytics and Data Science capabilities, which were previously unavailable in traditional systems. Disparity in data volume, variety in data formats, and the need to use emerging Analytics technologies, such as Machine Learning and Artificial Intelligence to perform high octane analytics, has forced organizations to re-architect their Data Warehouses and provide Data Lake management platform. This technology augmentation helps to overcome the below challenges:

Increasing cost

The cost of data warehousing is increasing significantly due to an expensive licensing model, which requires massive amount of data to be stored and maintained over the years.

Inability to scale-out

Scaling-out linearly using commodity hardware is expensive.

Unused data driving cost up

Research shows that 70% of data in Data Warehouse is unused, i.e. never queried in past one year.

Misuse of CPU capacity

Almost 60% of CPU capacity is used for extraction, transformation, and loading (ETL/ELT). 15% of CPU power is consumed by ETL to load unused data. This affects performance of queries.

Inability to support non-relational data

Data Warehouses are not suitable for semi-structured or unstructured data formats coming from IoT devices, mobile apps, and social media.

Inability to support modern analytics

Traditional Data Warehouses don't support modern analytics technologies, such as Machine Learning and stream processing.

OPTIMIZE YOUR DATA WAREHOUSE WITH DATA LAKE: PROPOSED SOLUTION

Step 1: Select offload data and the new data sources

Seek expertise to select the data sources and ETL workload for offloading from Data Warehouse. Keeping business continuity in place, retain the frequently used data in the warehouse and offload and archive the unused data to Data Lake repository. This ETL workload is taken over by Apache Spark fast in-memory processing capabilities. Also the organization can identify the data from new sources as per business KPIs to initiate data ingestion and storage for performing further analytics.

Step 2: Build the data pipelines

Migrate data in batches using Network File System (NFS) or Apache Sqoop or real-time methods using tools such as Kafka Connect. Many data warehouses provide connectors for Hadoop ecosystem that help simplifying the migration. Upon data migration, the data is stored in Hive tables, or Parquet or Avro files depending on requirements.

Step 3: Deliver the data to the stakeholders

Utilize unified SQL engines, such as Apache Drill, Hive, or Spark SQL to deliver data to Business Intelligence team. Leverage stored data in tables with popular BI tools, such as Tableau, Qlik, MicroStrategy, etc. The BI teams can continue querying the offloaded data using SQL. In addition, the newly sourced data, which is stored in the Data Lake, can be used to derive new insights by the Data Science team using Analytics workbench, such as Apache Zeppelin, and in-built Machine Learning libraries.

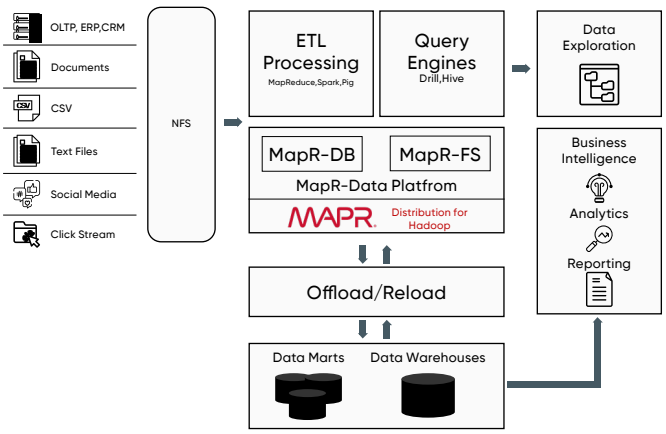


Fig: Sample MapR Data Lake Architecture platform

DATA LAKE PLATFORM
- KEY BUSINESS BENEFITS

Using a global data storage repository, the organization achieves MORE than simply storing the data in its native format. Data Lake provides many business benefits -

01

Reduce TCO of data analysis

Data Lake reduces the cost of data management and analytics. By deploying prepackaged applications, organizations are able to reduce the time to insight and deployment from days to minutes.

02

Unified Storage Platform

Storing data in unified storage repository, allows the organization to reduce the number of Data Marts, employ business centric rules, and access policies within the same storage platform.

03

Answer new questions

The augmented platform provides a wide range of analytics tools, which were unavailable in legacy Data Warehousing. It helps to react quickly to the new requirements with reduction in time for insight creation and complex analytics (Drill, Parquet, Spark, Machine Learning, others).

DATA ANALYSIS AS A SERVICE (DAaaS)

In addition to capabilities discussed above, Data Lake augments data storage, intelligent processing and complex analytics with Apache Hadoop and Spark ecosystem components on IaaS (Infrastructure as a Service) cloud platforms. This also explains why organizations consider Data Lake implementation on-cloud infrastructure:

Data Lake Analytics - analytics at scale to derive insights

This can be achieved by scaling the data storage and processing over Infrastructure as a Service (IaaS) platform offered by many competitive cloud vendors. Some popular ones are Amazon Web Services (AWS) and Microsoft Azure HDInsights. IaaS provides a platform, where the organization can easily develop and run massive parallel data transformations and processing programs without the overheads of managing a monolithic infrastructure, process and scale on demand, and only pay for the used time and or pay per job.

On Premise or Cloud Apache Spark and Hadoop services for the enterprise

AWS EMR and Azure HDInsights provide a fully managed cloud based Hadoop cluster with Analytics capabilities and extended support for Machine Learning libraries. It allows using popular Open Source frameworks including Apache Hadoop, Spark and Kafka. It helps to quickly spin up the cluster on demand and scale up and down based on organization requirements and needs.

Data Lake Store - Data Lake that powers Big Data Analytics

AWS s3 provides massively scalable secure storage. AWS s3 is highly available and designed to deliver 99.999999999% durability, and stores data for millions of applications used by many market leaders for data storage. It also provides 'query in place' functionality, which allows running a query on the data set at rest. AWS s3 is supported by the large community of third-party applications and AWS services.

CASE STUDIES

How a television broadcasting company leveraged Data Lake strategy for more efficient data exploration

CHALLENGE

The client was experiencing problems with infrastructure usage in their existing Data Warehouse system, where the computing resources were consumed up to ~99% resulting in job failures at a particular time. The identified bottlenecks found to be related with number of users accessing the data assets at the particular time via interactive querying.

Additionally, the client had the requirement to manage new data sources from their applications, where semi-structured data and unstructured data had to be stored in an appropriate storage system.

To meet growing demands for quick and responsive analytics, the client required a fast, searchable, and reliable data management solution.

APPROACH

Datamatics recommended Data Lake implementation as a pilot run with a compact cluster capacity as per the size of the data housed by the client. The pilot implementation of the configurable Data Lake architecture included features to secure and curate the new data. Upon successful deployment, the structured data was proposed to be managed in the Data Lake platform.

Also to ensure process continuity, the historical data was archived and stored.

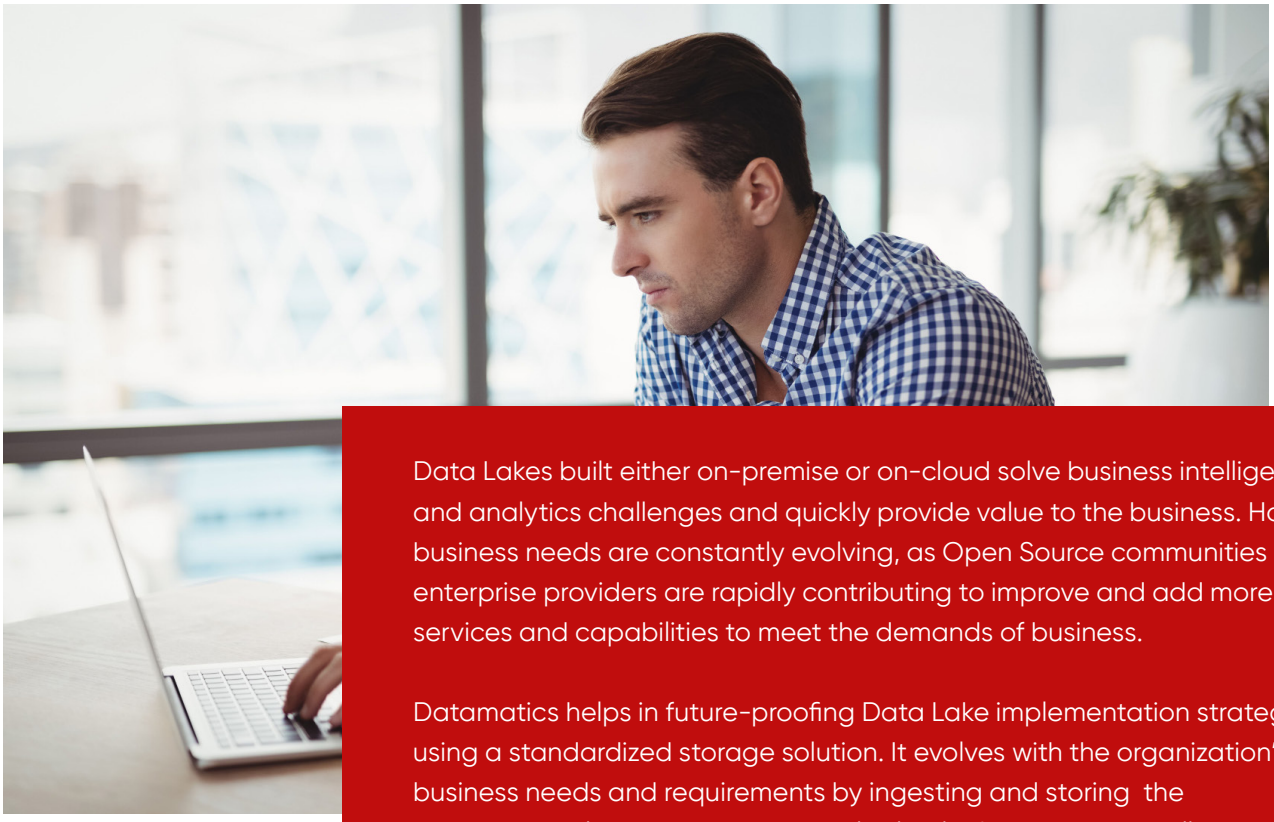
For getting the most out of the implementation, the platform was tailored in accordance with the growing business needs by using Open Source and commercial technology applications and their features.

RESULTS

At the end of the implementation, the client had recommendations for a fully integrated, reliable Data Lake that would allow business users and stakeholders to access data on demand. Users would be able to perform data exploration, interactive queries at scale, and also analyze variety of data.

With the successful implementation, the client was able to gain a flexible and reliable data management platform with reduced Total Cost of Ownership (TCO). The solution had the ability to integrate new data sources and continue delivering the maximum business values to the organization.

CONCLUSIONS



Data Lakes built either on-premise or on-cloud solve business intelligence and analytics challenges and quickly provide value to the business. However, business needs are constantly evolving, as Open Source communities and enterprise providers are rapidly contributing to improve and add more services and capabilities to meet the demands of business.

Datamatics helps in future-proofing Data Lake implementation strategy using a standardized storage solution. It evolves with the organization's business needs and requirements by ingesting and storing the organizational data assets on a scalable platform, which is well integrated with a variety of data processing tools. Additionally, the organization can also perform analytics using Data Lake assets to quickly explore new methods and tools and then scale the Data Lake in to production environment. Data Lake built on Hadoop platform helps to grow the business around existing and new data assets, and use them to quickly and easily derive business insights without limitations.

Resources : (Refer. Apache Hadoop Ecosystem components) https://hortonworks.com/ecosystems/

ABOUT
THE AUTHOR

GAURAV GANDHI
Technical Consultant

He has used his analytical, technical and problem solving skills to contribute in successful implementation of several Data Warehousing and Data Lake based projects.

Gaurav is an experienced Software Engineering professional passionate about contributing to the initiatives that leverages data and analytics in business - technology integration and innovation.

Gaurav is working in technology consulting for 5 years where he has used his analytical, technical and problem solving skills to contribute in successful implementation of several Data Warehousing and Data Lake based projects. He is skilled in Data Management, Analytics, NoSQL and Cloud technologies.

Gaurav holds a Bachelor's Degree in Electronics and Communication Engineering.

ABOUT **DATAMATICS**

Datamatics provides intelligent solutions for data-driven businesses to increase productivity and enhance the customer experience. With a complete digital approach, Datamatics portfolio spans across Information Technology Services, Business Process Management, Engineering Services and Big Data & Analytics all powered by Artificial Intelligence.

It has established products in Robotic Process Automation, Intelligent Document Processing, Business Intelligence and Automated Fare Collection.

Datamatics services global customers across Banking, Financial Services, Insurance, Healthcare, Manufacturing, International Organizations, and Media & Publishing.

The Company has presence across 4 continents with major delivery centers in the USA, India, and Philippines. To know more about Datamatics, visit www.datamatics.com

FOLLOW US ON

website: datamatics.com | email: business@datamatics.com

_012022

USA | UK | UAE | India | Philippines