International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Next Gen Data Lake on Cloud for Pharma and Healthcare Organizations

Rohit Malik

Chief Architect Multi Cloud – Tata Consultancy Services

Abstract: ***Data Lake concept was paved to solve a problem of managing a large set of datasets coming from new sources of data such as sensors, devices, medical instruments, social media, and website interactions in different type of formats and sizes. Traditional data management solutions were effective but were relevant for few types of structured datasets such as data warehouses and data marts. To enable storage and analysis of semi - structured and unstructured data sets a better system was needed to not only store data of such large sized, but to also to process it with speed and intelligence.***

Keywords: Cloud Data Lake, Data Warehouse, Health Lake, Analytics, AI / ML

1. Introduction

As storage requirement for ever - expanding data is challenge for every Pharma Researcher, Drug Manufacturer, Clinical Trial Scientists and Healthcare Organizations. Setting up data servers for storage and delivery not only requires huge capex but also a high level of expertise and floor area. Public cloud is the solution - Why procure and invest in the infrastructure which can be rented on pay as you go model in a very cost effective and efficient way? The cloud is made up of hundreds of physical data centers spread across the world, rife with cables connecting heavy - duty equipment.

While all the benefits of cloud are widely understood, migrating contents, files and database is not always smooth sailing and there can be challenges to overcome when transitioning. Below are the key challenges associated with data migration to the public cloud.

- Selecting the right cloud partner
- Regulatory policies, compliances and data security
- Storage selection for different types of data sets (Files, Objects & Database)
- Data preparation – Filtering unnecessary data and prioritizing
- Information validation – Before and after migration

Cloud Adaptation:
Most Life Sciences & Healthcare organizations have on - premises applications burdened with high capital expenses, complex management, scalability challenges, and hardware that needs to be replaced/upgraded every 3 - 5 years. With these on - premises challenges, Cloud Services Providers are increasingly offering cloud - based solutions for production and management while in delivery Life Sciences Organizations are turning to the large web - scale cloud players.

There have been a few high - profile use cases for cloud adoption where Pharma or Healthcare firm have shifted from reliance on traditional IT stacks in owned and operated data centers to the public cloud. Organizations havesetup some workloads to the cloud and pushing itself as a keen adopter of multi cloud technologies.

To enable Advanced Analytics the owners had choice either to turn a pharma or healthcare organization into a world - class data center operations company or move the services to the public cloud.

Data Lakelatest Trends enabled using Public Cloud Capabilities

- Despite concerns about the security & privacy, regulatory & compliance, data management, higher latency and lack of control over cloud storage, many businesses see that the cost savings, accessibility and disaster recovery are more valuable than the associated risks. Cloud storage is often the ideal solution for most business owners. It is inexpensive, easy to set up, and doesn't require any extra time or effort to maintain. Life Sciences and Healthcare companies today rely on the cloud for significant elements of their operations.
- Many Large and mid - sized companies already have implemented the cloud for long - term content storage, and now an increasing number of these companies are migrating production and manufacturing tasks - indeed their end - to - end workflow - to the cloud. There are several reasons why customers are migrating their application & data workloads to the cloud. Some are migrating to the cloud to increase the productivity of their workforce.
- Many Life Sciences companies with a data center consolidation or rationalization projects migrating to the cloud. Additionally, there are companies that are looking to completely re - imagine their business using modern cloud technology as part of a larger digital transformation program. Organizatons efforts to create their own digitally integrated workflows using traditional components in - house have struggled because of the huge costs and server requirements involved. Advances in technology and consumer behavior are driving a transformation in the way different type of content is delivered to researchers & consumers.
- The change involves a migration away from traditional data models and platforms towards digital content over the Internet to a widening array of connected devices. This fundamental shift is triggering three major disruptions for Life Sciences Organiztions, each calling for the scalability, cost flexibility and agility of cloud computing. Organizations should incorporate cloud technology to facilitate a cost - effective, collaborative

and futureproof creative environment to collect, store and conduct analytics on vast amounts of data, generating insights to drive personalization, service development, customer experience and one - to - one relationships, scalability to handle spikes in workload, including live events, and surges in the popularity of new services.

- Most migrations happen in phases to minimize risk and speed time to production. Moving contents and database could be organiztions first step to cloud migration. Cloud enables greater agility and efficiency in meeting increasing demands on time and capacity. Giving the full creative team immediate and flexible access to content and key production applications on scalable, costeffective cloud - based storage, media facilities position themselves to take advantage of any new opportunities that come their way.

Pharma and HealthcareData Lake Setup Uses Cases

- Cloud Health Data Lake helps Pharma and Healthcare organizations analyze industry trends, inferences and TCO to identify right solution for themselves.
- Cloud Health Data Lake supports the Digital Health care systems, Life and Health insurance organization and life sciences organizations to improve quality of life and health care system across the world.
- Cloud Data Lake also offers Pharma and Healthcare organizations to do advanced analytics using AI / ML techniques, provide better insight of patient data, hospitals usage and Pharma Manufacturing companies efficiency and automation.

Cloud Offerings to setup Health Data Lake for LifeSciences and Healthcare companies

- Cloud Services to maintain the data integrity when different applications or services are doing concurrent operations or in case of failures. This basic property of a data warehouse is inherited into a health data lake.
- Support for all raw data types, including audio, video, device formats, etc.
- Support streaming data and generating insights in real - time.
- Decouple Storage and compute to make it independently scalable as per LS & HC Use case needs.

Health Data Lake architecture is created for specific use cases in Life Sciences and Healthcare organizations. Ask is to ensure Compliance (HL7, HIPPA, etc…), Regulated (GDPR, etc…) and Secure Data (example encrypted, etc…) across organization or to third parties.

AWS Health Lake Solution

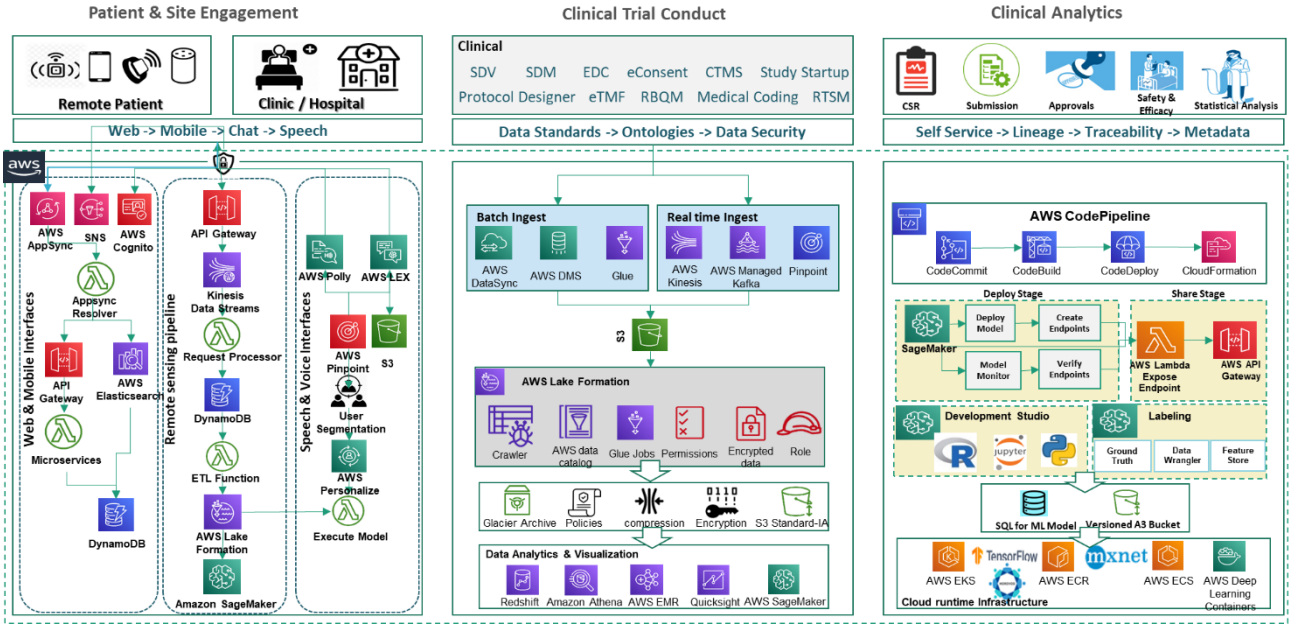
AWS offers a wide variety of services and partner tools to help us migrate our data sets, whether they are large content files, databases, machine images, block volumes or even tape backups. The most common approach is to lift - and - shift an application and its data with as few changes as possible providing the fastest time to production with follow - on projects to update application elements over time leveraging cloud services and optimizations that provide the most tangible benefits. AWS makes it fast and easy to lift - and - shift our applications to the cloud by providing storage services that are equivalent in features and functionality to what was being used on - premises with the added cloud benefits of flexibility, performance, security, and scalability. AWS Database Migration Service supports homogeneous migrations such as Oracle to Oracle, as well as heterogeneous migrations between different database platforms, such as Oracle or Microsoft SQL Server to Amazon Aurora, from one RDS database to another RDS database. For content migration we can opt for one - time offline migration using AWS SnowBall and further continues incremental movement to cloud through AWS DataSync. SnowBall (offline petabyte - scale data transport solution) offers secure appliances to transfer large amounts of data into and out of AWS. Using Snowball addresses common challenges with large - scale data transfers including limited network bandwidth, long transfer times, and security concerns. Transferring data with Snowball is simple, fast, and secure. AWS DataSync, makes it simple and fast to move large amounts of data online between onpremises storage and Amazon S3, Amazon Elastic File System (Amazon EFS), or Amazon FSx for Windows File Server. DataSync automatically handles many of the tasks related to data transfers that can slow down migrations or burden our IT operations, including running our own instances, handling encryption, managing scripts, network optimization, and data integrity validation. We can use DataSync to transfer data at speeds up to 10 times faster than open - source tools. We can use AWS DirectConnect to establish a dedicated network connection from our premises to AWS, which ensure increased bandwidth throughput and provide a more consistent network experience than Internet - based connections.

Key Benefits of using AWS native data migration services:

1) Simplify and automate data movement
2) Archiving of cold data
3) Reduce operational cost
4) Data movement for timely in - cloud analysis

Below Reference architecture recommended solution for using AWS Services to build a scalable, distributed advanced analytics workflow. The advanced analytics solution ingests metadata files and source content, processes the files & images for storage on a wide range cloud native data services, and delivers the advanced analytics using various analytics streaming and dashboard services.

AWS Data Lake Reference Architecture



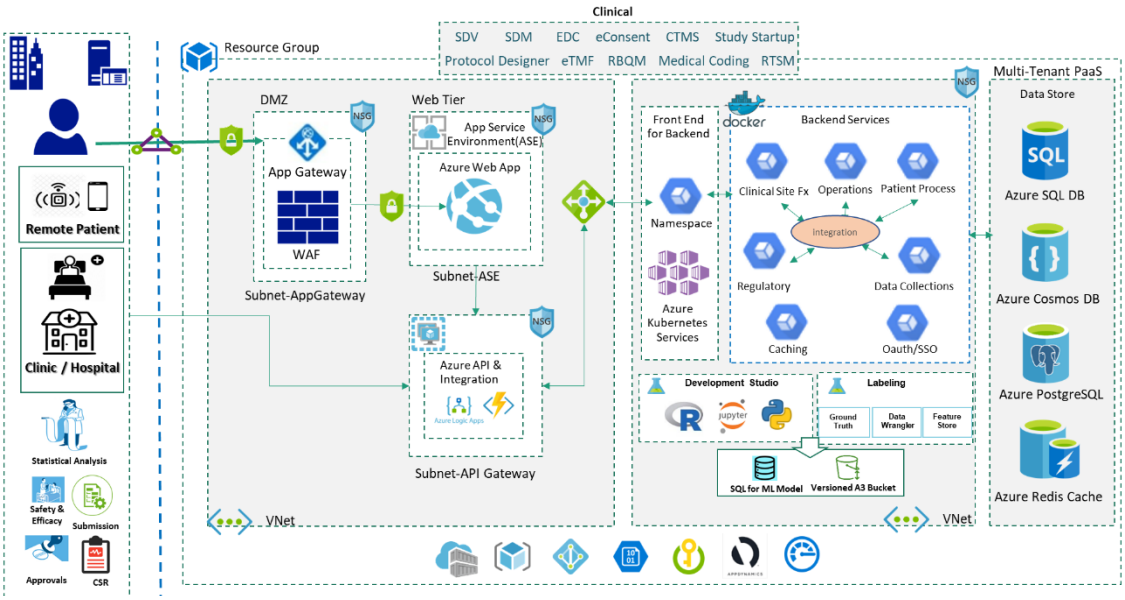
AWS Solution architecture ingests source data, processes the files & images and stores on a AWS Health Lake and other Data Storage Services for further intelligenc using AI / ML and Advanced Analytics. AWS has multiple Life Sciences and Healthcare industry specific services:

a) Amazon Healhtlake
b) Amazon Omics
c) Amazon Comprehend Medical

Azure Health Data Lake Solution

Azure Data Lake for Life Sciences Organizations has all the capabilities required to make it easy for data scientists and analysts to store data of any size, shape and velocity, and do data processing and analytics across platforms. It removes the complexities of ingesting and storing all of your data while making it faster to get up and running with batch, streaming and interactive analytics. Azure Data Lake works with existing IT investments for identity, management and security for simplified data management and governance. It also integrates seamlessly with operational stores and data warehouses so you can extend current data applications. We have drawn on the experience of working with enterprise customers and running some of the largest scale processing and analytics in the world for Microsoft businesses like Office 365, Xbox Live, Azure, Windows, Bing and Skype. Azure Data Lake solves many of the productivity and scalability challenges that prevent you from maximising the value of your data assets with a service that is ready to meet your current and future business needs.

Azure Data Lake Reference Architecture



Azure Data Factory which is a fully managed cloud native data integration service from MS Azure and can be used to populate the data from a rich set of on - premises and cloud - based data stores and save time when building your data lake and analytics solutions.

Azure Data Factory supports moving data from the following sources to Azure Data Lake Store:

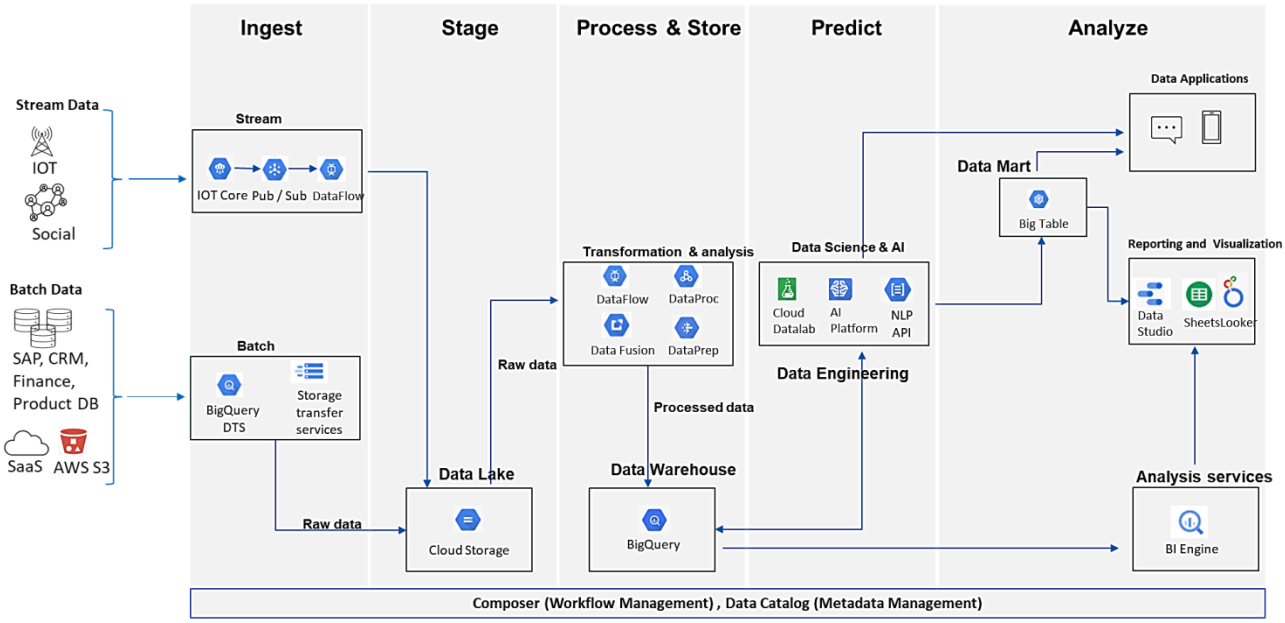
- Azure Blob
- Azure SQL Database
- Azure Table
- On - premises SQL Server Database
- Azure DocumentDB
- Azure SQL DW

On - premises File System

- On - premises Oracle Database
- On - premises MYSQL Database
- On - premises DB2 Database
- On - premises Teradata Database
- On - premises Sybase Database
- On - premises PostgreSQL Database
- On - premises HDFS

Google Cloud Solution for Health

GCP Data Lake Reference Architecture



Google Cloud healthcare data engine provides 360 - degree view solutions for Life Sciences and Healthcare customers using it's native services for combining medical records, clinical trials and research data, along with other informational sources

2. Conclusion

With the rapid advancement in technology including cloud, data lake, analytics and AI / MLHealth Lake is the future of Pharma and Healthcare industry. This white paper is written to provide details of current trends in Pharma and Healthcare industry which has made it very important to setup industry specific Health Data Lakes with enormous use cases which can be solved by 3 biggest public cloud providers AWS, Microsoft Cloud (Azure) and Google Cloud.

Acknowledgement
The Author would like to thank **Sanjeev Sachdeva** (Head - Life Sciences and Healthcare Domain Strategic Capability Group of TCS) and **Kapil Naudiyal** (Head Digital Transformation - Life Sciences and Healthcare Domain Strategic Capability Group of TCS) for giving the necessary support.

References

[1] Lock, M. Maximizing Your Data Lake with a Cloud or Hybrid Approach.2016. Available online: https: //technology - signals. com/wp - content/uploads/download - manager - files/maximizingyourdatalake. pdf
[2] Kumar, N. Cloud Data Warehouse Is the Future of Data Storage.2020. Available online: https: //www.sigmoid. com/blogs/cloud - data - warehouse - is - the - future - of - data - storage/
[3] Big Data and Analytics Services Global Market Report. Available online: https: //www.reportlinker. com/p06246484/Big - Data - and - Analytics - Services - Global - Market - Report. html
[4] Dixon, J. Pentaho, Hadoop, and Data Lakes.2010. Available online: https: //jamesdixon. wordpress. com/2010/10/14/pentaho - hadoop - and - data - lakes/

Author Profile

Rohit Malik, Cloud Chief Architect (Life Sciences and Healthcare Domain Strategic Capability Group of TCS),is an accomplished professional delivering over 23 years' managerial and functional career success in driving Futuristic IT Ecosystems, IT Solution Delivery, Innovation, Business Process Reengineering/Benchmarking using Digital Technologies. He has mastered the administration of establishing businesses, managing IT program, articulating technology market developments, invigorating businesses, and service delivery. He has a strong expertise in AWS Cloud Platform with a good knowledge of Azure and GCP Cloud Platforms, Application Migration &Modernization, Microservices, Big Data, Application and Hybrid Integration.