LTIMindtree

Whitepaper

# Serverless Data Warehouse and Analytics Platform

Authors : Himansu Sekhar Tripathy, Deep Sharma

# Contents

# 1. Background

Cloud computing has enabled ubiquitous access to system resources and higher-level services, which can be provisioned with minimal management effort. Most of digital disruption occurring today has some form of cloud computing at its core. According to Forrester's survey of data analytics professionals, 'public cloud is the technology priority for Big Data.' Depending on the regulatory framework enforced in their industry, enterprises have adopted private, public or hybrid clouds to ensure agility and innovation in their services. However, many organizations who have embarked on the cloud journey are grappling with the complexity of the hybrid architecture entailing—the integration of on-premise platforms with cloud-native applications –as well as data modelling on cloud, and infrastructure sizing. Hence, it is imperative to build a framework that combines the cloud's unlimited scale of resources with the best practices of a well-governed data platform.

# 2. Need for Serverless Architecture

Infrastructure sizing is considered more of an art than science, which often confounds seasoned IT architects. Anticipated and unanticipated business growth and data spikes during special events often expose infrastructure inadequacies. A typical compromise includes building additional capacity which leads to under-utilization, or building to capacity and under-performing in some situations. However, cloud adoption can prove to be an ideal solution since its capacity is unbounded. But for a data platformthe window of opportunity to make interventions is narrow, and cost of failure is high. Hence, enterprises require a platform that can scale without manual intervention, in order to ride the spikes in growth.

Furthermore, collaboration platforms for data scientists is becoming a norm, enabling them to access and understand analysis done by their colleagues. As these environments support multi users, individual users can develop models independently. Subsequently, the analytics platform would demand computing power that is elastic in nature.

Serverless architecture also plays a critical role in the success of use cases such as real-time analytics for IOT devices. Industrial devices produce several terabytes of data daily, making it difficult to store and analyze through conventional on-premise data centers. This creates demand for a platform capable of receiving unbounded messages, and applying flexible rules to data-in-motion. In essence, serverless architecture can be leveraged by any activity where estimating resource usage is difficult and actual usage shows a wide variance.

# 3. The Evolution of Serverless Data Platform

After establishing the need for building serverless architecture, data platforms such as Data Warehouses and Data Lakes need to be evaluated for suitability of migration to serverless architecture. Enterprises that have more than one data platform need to choose the best candidate for serverless architecture.

Enterprises usually employ data warehouses as central repositories of integrated data from multiple disparate sources, and derive business insights from them. Data is cleansed, transformed, catalogued, and presented to business executives and data scientists, for analysis, market research, decision support and data mining. Data warehouses employ matured data governance practices that drive superior analytics-based business strategies.

In its formative years, data warehouses were developed on relational database management systems (RDBMS) such as Oracle, Sybase. These databases are only vertically scalable and hence can achieve only limited scalability. As a result, modern data warehousesbuilt in last couple of decades - such as Teradata, Netezza, have moved to distributed computing architecture which enables horizontal scaling. However, this infrastructure is expensive and can store only structured data.

Emerging in the last decade,Hadoop and NoSQL databases were equipped to aggregate multi structured data from disparate sources and create'data lakes'. Data Lakes possess the required computing power and storage at low cost, and can store multi-structured data in raw format. A sandbox environment within the Data Lake permits analysts and data scientists to review and apply transformations to raw data for analytics. However, it has not done well in governance due to concerns regarding the veracity of quality and lineage of data. Moreover, these varied technologies have a high learning curve and are yet to mature.

In comparison, Data Warehouses are better placed as Data Governance is built with years of effort and is difficult to replicate to a new environment. Also, they can capitalize on cloud storage which has the capacity to store unstructured data. Hence, data warehouses are presumed to be the next generation data and analytics platform, which can be extended to store multi-structured data, and scale while maintaining low total cost of operations (TCO). Serverless data warehouses will allow organizations to focus on analytical processes that bring business insights and not worry about the underlying infrastructure.

# 4. Migration Approach to Serverless Data Platform

Enterprises can build serverless data warehouses and analytics platforms from scratch, or migrate their existing data warehouse to make it serverless. The scope of this paper does not include the pros and cons of these approaches, or the suitability of one approach over the other. Migration of existing data warehouses is a plausible solution as it has been built with years of effort and has matured models and data governance processes. However, a few concerns are likely to emerge with regard to existing processes, such as what to do with complex but effective ETL processes, layered data models with consumption from each layer, and existing BI reports. As with other contextual questions, these don't have definitive answers, but rely on best practices followed by enterprises. Successful migration from prior versions of data warehouses to a serverless platform involves the following distinctive steps:

a. Evaluate Enterprise Data Model

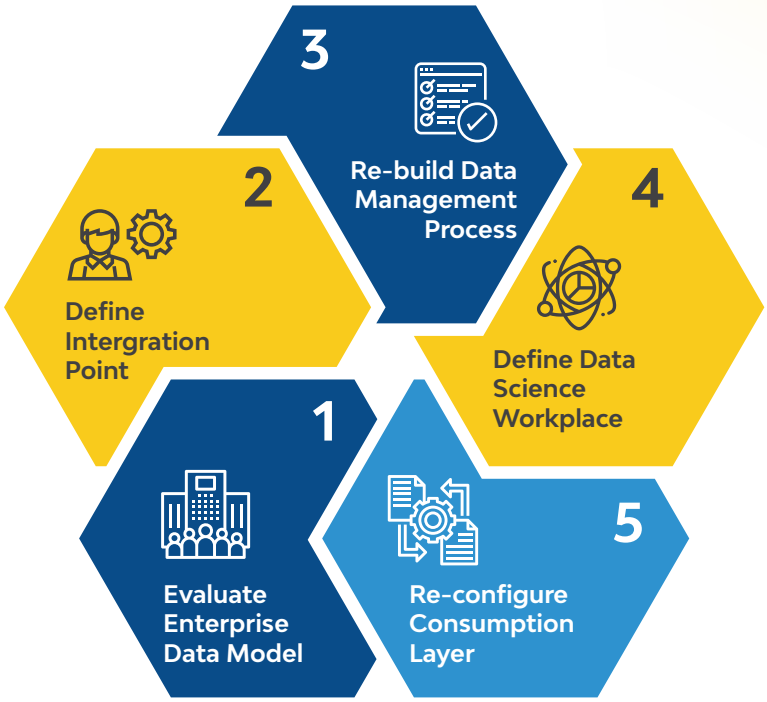
b. Define Integration Point

c. Rebuild Data Management Process

d. Define Data Science Workbench

e. Re-configure Consumption Layer

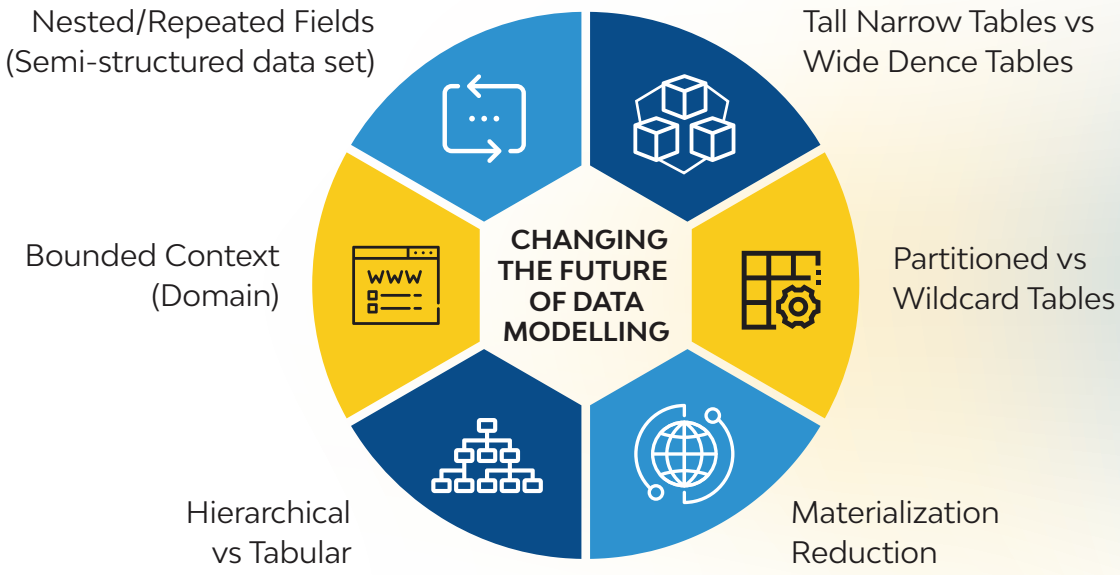
## RE-INVENT Serverless DWH & AI Platform STRATEGY



## Step 1 – Evaluate Enterprise Data Model

Data Models depend on the choice of platforms, and have been developed using various techniques including 3NF, dimensional (star / snowflake), object-oriented and document modelling. Migration to serverless platform requires thorough review of the current model, and optimizations on the number of tables, entities and collections. A large European retailer was able to reduce the number of tables by one-sixth (from about 500+ to 99), by adopting best practices of the chosen platform. Changes to existing data models even if too big, complex and multi-layered prove to be cost effective,as lean models out weigh the cost of building it. Enterprises that have large number of entities in their data model, can start re-modelling their outermost layer and move to inner layers subsequently.

Data modelling is a vast area of research and this paper doesn't prescribe any particular methods to build models for multi-structured data, but the following are a few considerations.




## Step 2 – Define Integration Points

Transferring data from on-premise applications to cloud based serverless data warehouses is done through cloud storage for batch processing and queue or topic for real time processing. As many data sources still remain within the data center of enterprises, well-defined integration points ensure that integration doesn't become the weak link in the overall process. Cloud storage provides a staging environment in the data integration pipeline and its underlying distributed architecture has the capacity to read data streams faster than any conventional staging area.

Moreover, integration points for streaming data will be a queue service provided by the cloud vendor. It is essential to use the cloud provider's queue service for auto-scaling rather than using external queuing services provided by third party vendors.

## Step 3 – Rebuild Data Management Process

Data Management processes include quality check, exception handling, applying business rules, and master or reference data management. In matured environments these are decoupled components, with clearly defined specifications. Moving to serverless platforms will involve migrating these components to cloud platforms. However, most cloud providers don't have specific services that ensure data quality and master data management. Rebuilding these components will be expensive and would discard years of effort spent on building the on-premise version. As a result, many enterprises have moved their on-premise solution to cloud using a lift-shift approach. In fact, a leading media conglomerate employed a third party cloud MDM solution that was compatible with their cloud data warehouse platform. This is possible, if product vendors used for on-premise solution has an easily portable, cloud version of the solution.

Most best practices developed in traditional data warehouse are applicable in cloud data management processes such as applying filters at the sources or close to sources, removing unwanted fields, checking data skew, applying map level aggregations and bucketing.

## Step 4 – Define Data Science Workbench

Given the evolution and emergence of next generation hybrid data and analytics technology components, Serverless analytical platforms play a key imperative for organizations to leapfrog in there analytics journey. The key is to employ a data science pipeline that enables businesses by providing:

a) Visual data preparation and aggregation with key data points, using scalable components of cloud computing for predictive analytics
b) Advanced API-based interface that adds intelligence to analytical models
c) Portable workflow components that train sample data locally and use serverless cloud platforms for testing and deployment at scale
d) Pre-built analytical frameworks in a marketplace which facilitates improved TAT for data scientists
e) Seamless integration of actionable insights with value realization components that expedite the outcomes

After building data management processes that load serverless data warehouses, they need to be reviewed by business analysts, power users and data scientists. The process also requires a collaborative environment where analysis work can be saved and shared with other users. Containers such as Jupyter workbook that can support multi-users can be used as a workbench, to enable an easy sharing of insights generated from multiple users.

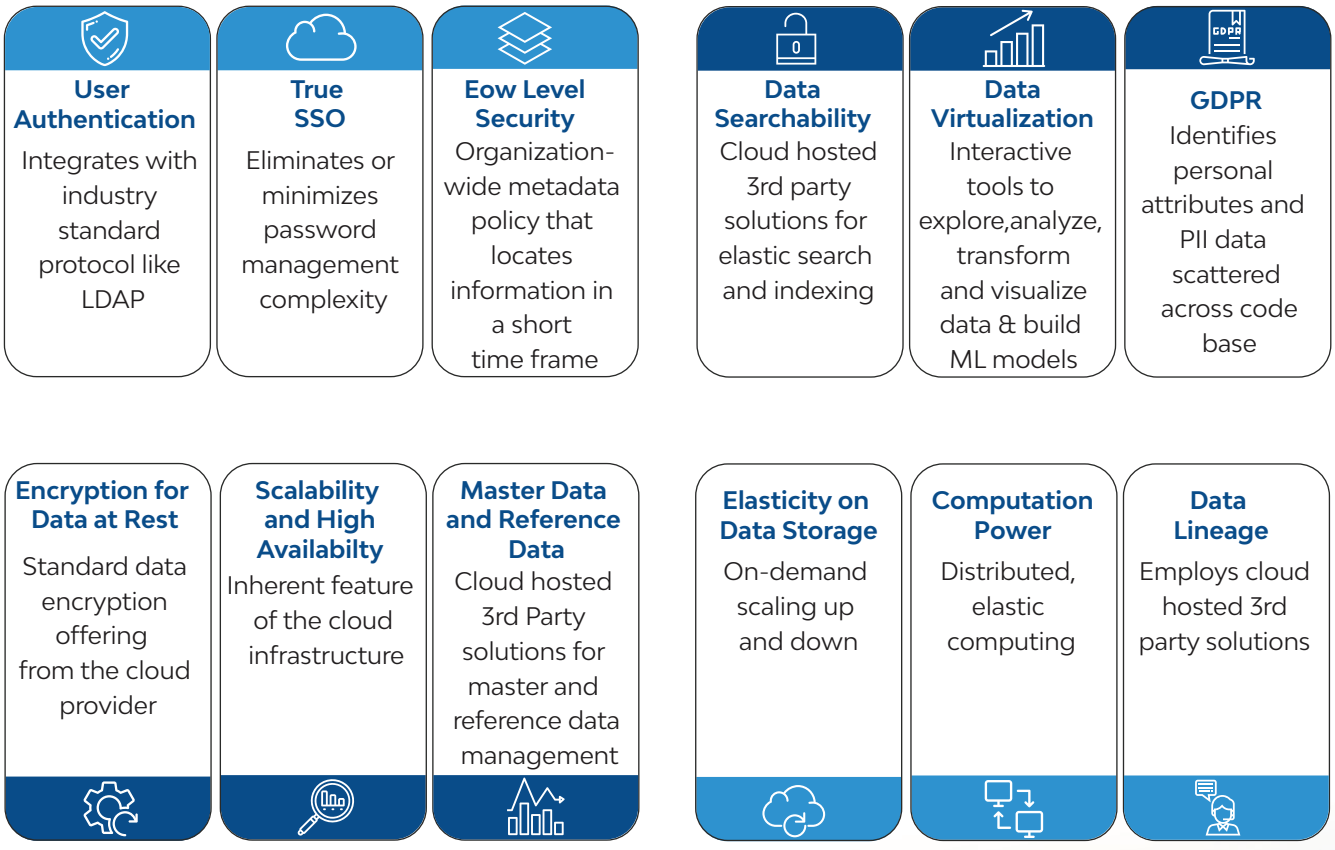
Data scientists can utilize data from staging i.e. cloud storage or from warehouses, and use supervised or unsupervised learning methods to prepare and train the model. They can also use APIs exposed by cloud providers for audio, image or video analytics. Furthermore cloud ML engines for deployment provide unlimited executing capacity and improved status on the health of the model.

## Step 5 – Re-Configure Consumption Layer

Last step of building serverless platform is the consumption of analytics via reports and dashboards. Most BI tools provide connectivity to cloud data warehouse platforms, enabling users to explore the data and derive insights. However, often many of these tools struggle to refresh their reports and dashboards at the same rate as the data platforms. Moreover though it sounds innocuous, evaluation of BI tools using any comparative matrix is essential while choosing a tool. Other than visualization, the consumption layer provides data to downstream systems using data management process with cloud storage and cloud APIs.

# 5. Data Governance

Data Governance enables organizations to ensure high quality of data in its complete life cycle, and focusses on areas like availability, usability, consistency, integrity and security. In serverless cloud data warehouse platforms, many facets of data governance such as data security, master data management, business taxonomy, lineage and high availability, needto be configured by vendor services or custom developed. An example of custom development is for "Searchability of Data" which needs to be built by developing a data catalog. In fact, a large Asia-pacific retail bank used third party software to develop a data catalog through regular expressions, building dictionaries and using AI from tribal knowledge.



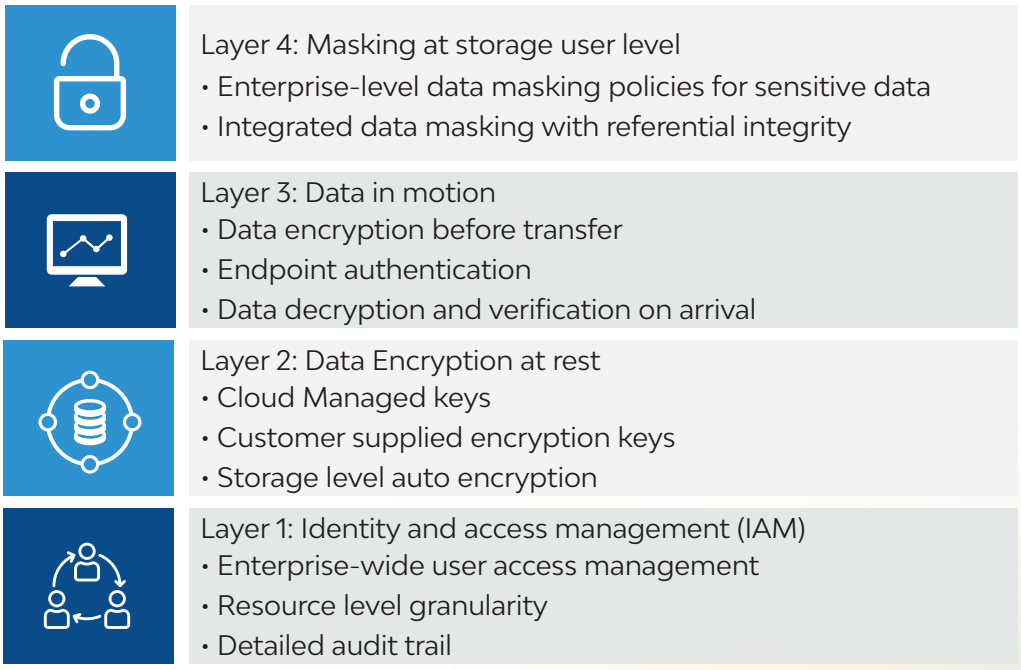
Before adoption, organizations need to recalibrate their approach to data security in a cloud environment. Cloud service providers offer access management and encryption services that safeguard data at rest or motion. However, since it is not common to hear breaches in data security, large organizations need to employ their own methods to ensure data safety in the warehouse. In particular it is essential that PII attributes are identified and security layers evaluated for these attributes. Many organizations implement a four layered architecture to safeguard information and insights from wrongful access.

a. Identity and Access Management (IAM)

b. Encryption by cloud-provider or own keys

c. Encryption of data-in-motion

d. Dynamic Masking in consumption layer



# Conclusion

Serverless computing is offered by cloud providers that dynamically manage the infrastructure allocation. Pricing is based on the actual resources consumed rather than pre-purchased capacity. Moreover, as entreprises address big data analytics, it is important for them to have flexible infrastructure capacity that address occasional spikes in transaction volume. Many analysts and researchers points out that public cloud is the platform of choice for big data engineering.

Migrating your current data platform to cloud needs a well calibrated approach that can be developed in steps. Some of these steps will need custom development and solutions provided by third party players. All these steps will need to consider constraints that can limit data flow and find ways to overcome them. Organizations who have undertaken this journey have improved their success rate by changing data models, rebuilding data management processes, choosing collaborative platforms for data analyses, and ensuring proper access and security. Finally, cloud's pay-as-you-go approach enables a more innovative and flexible approach that increases the focus on business decision making rather than the underlying IT infrastructure.

# About the Authors

**Himansu Sekhar Tripathy**

Himansu Sekhar Tripathy is a Data Management consultant with over 18 years of experience in consulting and delivery of data solutions. His interest areas include enterprise data strategy, cloud data engineering, big data engineering, data integration, quality, metadata management, MDM, and data governance. As a technology evangelist, he believes in leveraging emerging technologies in pushing the boundaries on real time next-gen analytics. He has a master's degree in Business Administration and a Bachelor's degree in Computer Science Engineering.

**Deep Sharma**

Deep Sharma is an Associate Consultant in Cognitive & Analytics Practice unit with more than 2 years of experience in technology consulting, analytics market research and offerings creation on emerging hybrid technology trends across the Data and Analytics technology stack. He has a keen interest in various building blocks of Data & Analytics like Data Integration, Data Quality, Data Governance and Data Visualization. He has earned a Master's Degree in Business Analytics.