LTIMindtree

Whitepaper

Data as a Service Powered by Data Mesh Architecture

Over the past decade, the role of data has undoubtedly changed within businesses globally. Organizations such as Facebook, Alphabet (the parent company of Google and its subsidiaries), LinkedIn, and Twitter, began harnessing the power of data by getting insights, finding correlations, and using artificial intelligence imparted with continuous learning. Together with the technology-enablers and hyper-scalers such as Amazon, Google, and Microsoft, these companies were able to store data on an unprecedented scale and harness the tremendous power from it. The evolution of these businesses has proved how data has become the primary asset in the digital economy.

Data as an Asset

Data, without a doubt, is the most vital part of a digital economy, where the entire business foundations are based on data platforms with cutting edge innovative arrangements; Amazon CTO Werner Vogels said, "data are (sic) at the core of value creation, whereas physical assets are losing their significance in business models." Data centric business models are the key to unlock potential across various business verticals. Every industry depends on insights from data for key decisions.

The Future of Data-Driven Businesses

As data and its derived insights become accessible to business, the data driven culture will step by step pervade through the worldwide business landscape. The Evolution of Analytics with Data implies that data and analytics are the two frequently used words in the last few years. With the abrupt development of multi-type data and analytics, it is important that organizations realize the trends and act on it to stay afloat.

For a business to accomplish a venture wide data-driven culture, it is crucial to get a buy-in from the C-Suite executives and those at the entry-level. Awareness must be created within the business culture, and the leadership team should clarify the reasoning behind utilizing data.

Data Growth and the Challenges ahead

As more data becomes available, it becomes a considerable challenge to process all of it into one platform, given the diversity in data domain types & subtypes due to new use cases, volume, and velocity. With a variety of personas such as analysts, data engineers, and data scientists using data for many use cases, the old paradigm of maintaining a central data store to hold many data domains becomes a huge overhead. The ever-evolving business processes are asking for richer data from more domains and with newer agenda for businesses as an ongoing trend. This trend translates into a need to test new use cases implies the growing number of transformations of the data - aggregates, projections, and slices, in essence, data transformations that can satisfy the current need and adapt to align with the future, as part of the cycle of innovation. As expectations from businesses grow, the response time from IT has always been lagging to the extent that newer ways are being sought in modern data platform architecture.

De-Coupled Pipeline

Traditional data processing consists of processes involving ingestion, cleansing, aggregation, storing, and serving. To meet the needs of adding new data sources, architects scale the system by breaking the process into further smaller deployable components interacting with one another to achieve a functional objective. The motivation behind breaking a system down into its architectural components is to create independent teams able to build and operate the quantum. In a typical monolithic architecture that caters to one primary data domain, adding a sub-type can involve reusing the majority of components with the addition of a few new elements to meet the objective. This capability could potentially reduce the velocity and scale of the data in response to new consumers.

Why Data Mesh

The answer to the many issues and roadblocks in the traditional organizational setup and the monolithic architectures is addressed by a Data Mesh that focuses on building modern distributed architecture at scale - techniques that the industry has embraced and created successful outcomes. The ability to decentralize data storage using logical borders based on the business domains has unlocked the potential to add more variety and operate at scale. With the data team and operations team working on their areas of development, the collaboration of the data and business operations happening at the data storage level combined with the application of the reusable design at the storage level enables the power of data to be unlocked for a variety of domains. When the data scope and purpose are well defined, cataloged for discovery, orchestrated through web services/ APIs, access-controlled using identity management systems, and user access logged, it paves the way for modern architecture designed for scalability, flexibility, and organizational governance.

In addition to timed events, source data domains should also provide easily consumable historical snapshots of the relevant datasets, aggregated over a time interval that closely reflects the interval of change for their domain. Data sets should be indexed based on different dimensions and organized for consumption with very low latency and higher throughput so that the consumer can stitch it to many other data sources to join, filter, and aggregate for intelligent analysis.

Distributed Domain Driven Architecture

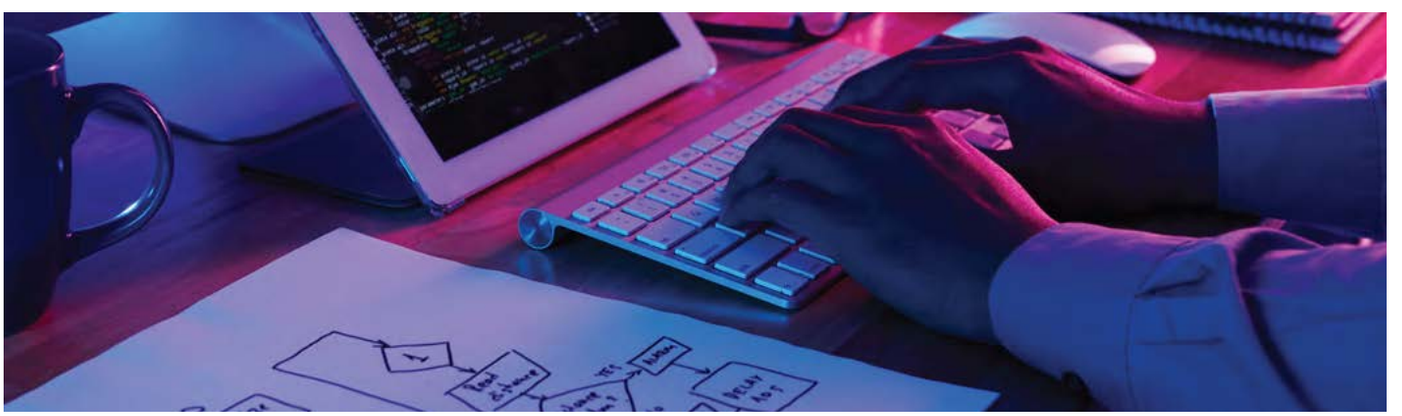
Borrowing the concept from distributed computing, Data Mesh focuses on disseminated data products aligned around enterprise domains and possessed by cross-functional teams with data engineers and product owners, utilizing standard foundational infrastructure as a platform to host, prep, and serve their data-driven information.

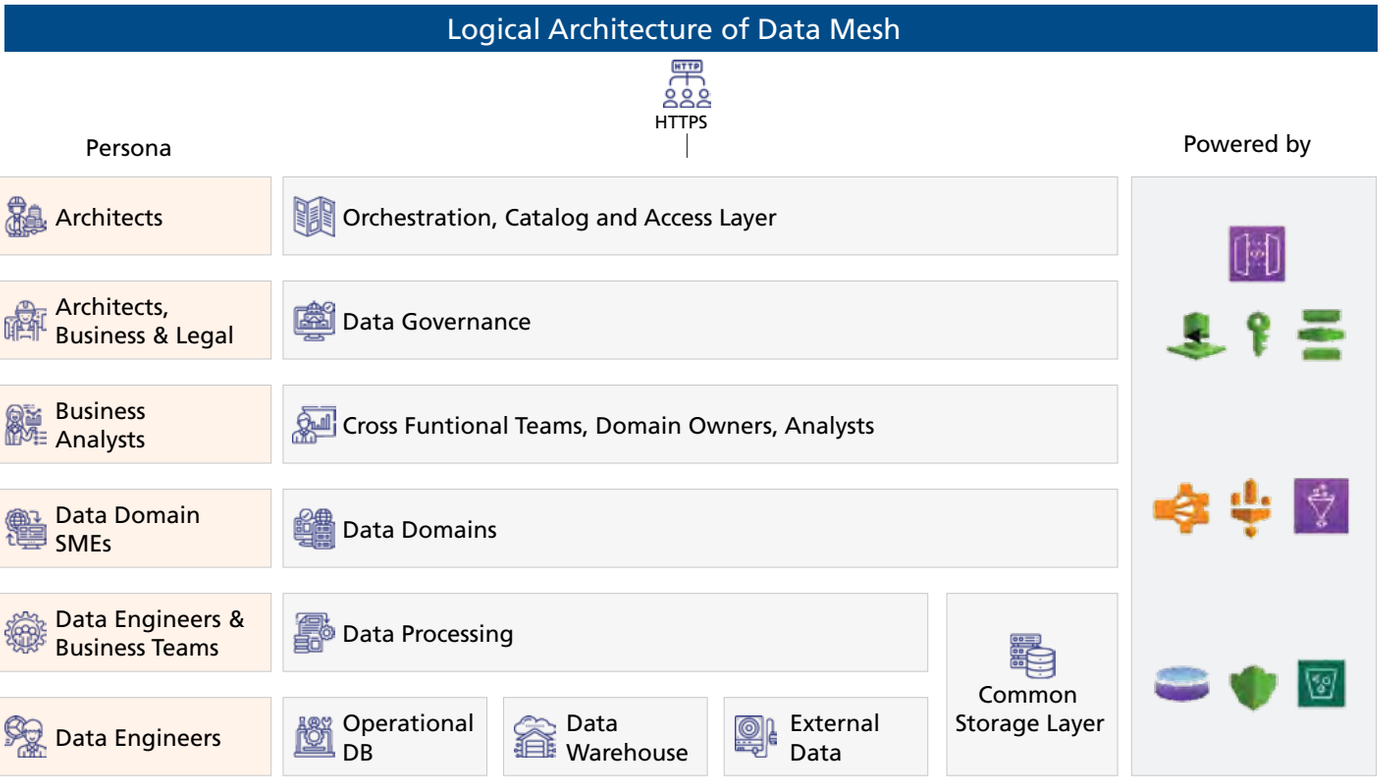
The Data Mesh platform is a distributed data architecture, used under centralized governance and interoperability standards and empowered by a common and orchestrated self-served data infrastructure. It is a long way from a scene of fragmented silos of inaccessible data.

The needs are real, and tools are ready. It is up to the engineers and leaders to realize that the existing paradigm of big data and one true big data platform or data lake will only repeat past failures, only this time, just using new cloud-based tools.

In the above architecture, the operational data and the external data become immutable points of data access, and where data transformation, if needed, is pulled, massaged, and stored as a data mart.

Whether internal or external, every data point becomes a source for a next higher level of abstraction. The vital thing to note here is the various data domains that serve diverse users with data pulled from various sources. The datalake to store monolithic data takes second preference to the ecosystem of various well-cataloged data, orchestrated for access through discovery process makes it simpler to manage it and scale for the future. Use cases for the domains of data identified by a cross-functional team of analysts and data are logically separated and owned by many teams. All sources of various data then become a node and are distributed when the need arises.





The Mesh and cataloging together with access control and governance makes the difference here as it moves away from a fragmented silos of inaccessible data.

Data Mesh claims that for big data to fuel innovation, its ownership must be federated among domain owners, accountable for providing their data as products. Decentralization and interoperability among data from various domains focusing on an end-user experience are key to data democratization. The Data Mesh idea is born out of modern distributed architecture: considering domains as the first-class data repositories, integrating data across many platforms, and providing self-serve data.

A Data Mesh holds the advantage over a traditional data warehouse as a central ETL pipeline gives teams less control over increasing volumes of data. Also, with different data use cases, it becomes difficult to cope with data transformations tilting the balance to relying on providers that specialize in a specific data set, pulling them on demand, and doing a mashup before putting it to use.

Technology Enablers

Hyper-Scalers like AWS, Azure, and Google Cloud can provide cloud infrastructure for storage, and infinite compute on-demand and pay-as-you-use, which can substantially reduce the operating cost and provide the ability to scale operations where needed and adjust domains and data as needed.

Security is paramount to the organization in the context of strict regulations for financial institutions, insurance, and others that maintain sensitive data: PII, PHI, and PCI. Hence, it becomes important to safeguard them. AWS offers KMS and HSM to encrypt data and role-based access for sensitive data using IAM policies.

For a business to achieve an enterprise-wide data-driven culture, data technology awareness must be carefully fostered within the business culture to truly democratize data science.

- Amazon S3/Amazon S3 Glacier provides the data foundation as it is an efficient and cost-effective way to store data. It can scale data based on the needs and archive data to retrieve it with an SLA of a few hours.
- On-Demand computes by AWS EMR provide virtually infinite computing for data-intensive operations based on parallel processing using spark.
- AWS Glue for ETL operation involving cleansing, formatting, mapping, and storing to the target system. AWS Glue DataBrew provides a valuable tool for data science users to manage data. AWS Glue Elastic Views can power the relevant analytical insights a Data as a Service (DaaS)-using-enterprise requires – with high performance.
- Amazon Kinesis can live stream at high velocity and perform analysis in real-time.
- Amazon Sagemaker is the tool for AI/ML, a key offering for any AWS-powered DaaS platform, as it is their one-stop solution for all things AI/ML on AWS.
- Data enrichment via third-party data pulls using APIs expands the scope for adding unlimited data domains and using the join, filter, and aggregate processes.
- API Gateways for scalable and secure data exchange expose data through API gateways powered by Lambda to perform data processing.
- Authorization/role-based access is addressed by identity providers such as AWS Cognito/ OpenAM.

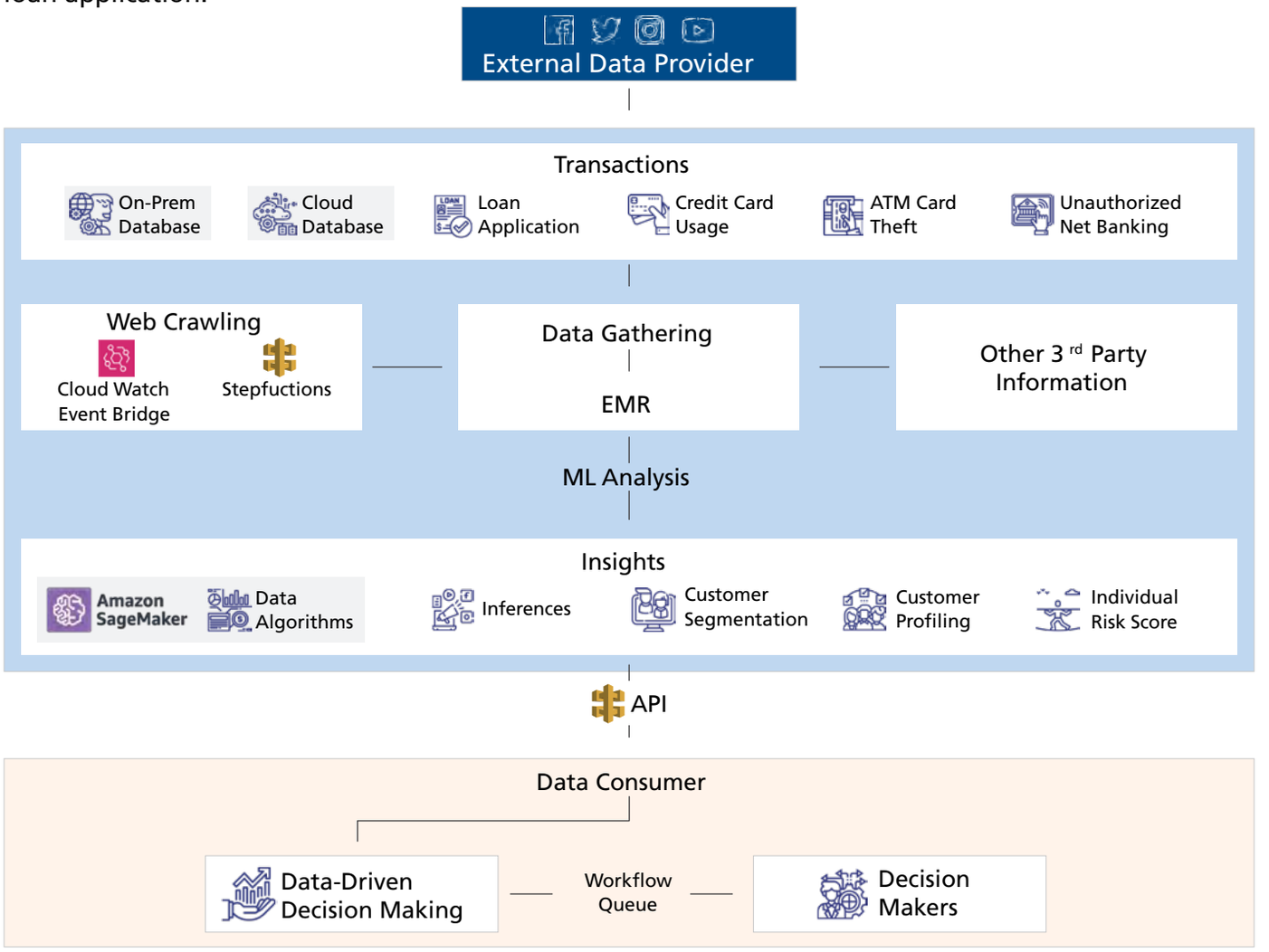
Data Mesh as a Foundation for Data as a Service

We propose a Data Mesh architecture that focuses on a federated way to provide data from multiple sources, even for analytical workloads for Data-as-a-Service (DaaS). For enabling all operational data residing in the organization's data stores, RDS and S3 can be leveraged. External data stores outside of the network firewall providing data on-demand, as a service can also be included as part of the data layer. The data from various sources should be cataloged using Glue Catalog and a semantic layer created for the metadata orchestration. Security can be applied to the metadata and data from multiple sources – internal and external – made available through APIs using AWS API Gateway powered by Lambda for data processing. The authentication can be enabled by an identity provider such as Microsoft AD/AWS Cognito/OpenAM using the OAUTH2 framework. Also, depending on the accepted latency for the request-response cycle, the number of users querying the data, and the volume of data needed, preprocessing can be done on the data using EMR, on-demand for heavy lifting. EMR can be orchestrated using AWS Step Functions and stored in a data mart based on Redshift or Amazon Aurora.

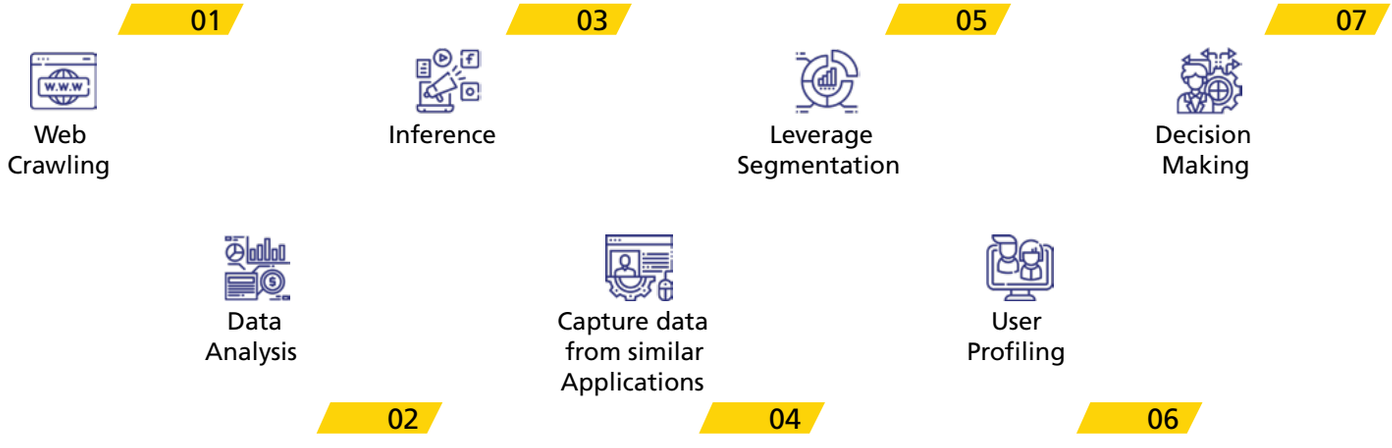
Use Case: Early Fraud Detection

Fraud detection and prevention is one of the many major objectives of banks. Modern analytics has the answer to leverage data driven analytics to proactively monitor using cognizant AI driven process to screen data for suspicious behavior and report possible fraudulent activity to a human for subsequent action. Preventing fraudulent activities reduces the risk of threats and helps gain new clients resulting from sustainable portfolio growth. Included is a workflow that takes a 360-degree approach to assign a risk score for credit loan applicants using data on different types of fraud, unacceptable behavior, and users' intentions, and applicant segmentation based on profiling. Forwarding such a recommendation to the human underwriter will facilitate a well-informed decision for the successful processing of an application and the interest rate.

The below sections explain the steps for a bot to provide recommendations upon receipt of the loan application.



The above diagram is an implementation of a Data Mesh architecture providing a service with data coming from different domains. After mashup, the value-added data is served over an API. The following section details the various pieces of the Data Mesh specific to this use case.



Web crawling: Web crawls made through social media data to gather time device footprint for the applicant using the device id.

Data analysis: Analytics added to traditional methods to enhance fraud detection capabilities. Using statistics and ML techniques, data is analyzed, and anomalies are detected. By analyzing the category of sites, duration visited, related links visited, the interest expressed by the data subject over a gamut of metrics is obtained using a proprietary algorithm.

Inference: Signs of randomization and other anomalies on devices are identified, deviation from the virtual user's behavior is scrutinized to remove false positives. Using an integrated case management system leveraging social media, a trained engineer can derive relevant findings for analysis that can be either from structured or unstructured data.

Capture data from similar applications: Data mining could be used to classify, cluster, segment data, and automatically find associations and rules to signify interesting patterns related to fraud. New business rules and pattern recognition can be set to identify fraudulent behavior. This valuable data helps organizations detect fraud more efficiently than those that rely on traditional methods.

Leverage segmentation: Perform segmentation of the requests flow by disposable income level, user profiling, fraud outcome of similar users, and compute a risk score. The absence of such segmentation can prove harmful to the decisions made using the information because clusters showing specific patterns might not be the characteristic of most of the entities of the group.

User profiling: A user's behavior, network place, etc., allows effective assessment of applicants who are hard to evaluate through conventional data sources, enhancing the resolution of the decision-making process, identifying low-risk groups for comprehensive product offers and increasing the approval rates in general. With the help of data mining tools, customer behavior can be analyzed to derive patterns from extensive customer records that can be used as a predictive tool for future behavior of customers for fraud detection.

Decision making: Data-Driven Decision making (DDDM) is the most critical element of success in an organization. After the data has been transformed into information, customers will be profiled into different buckets according to their risk score, thus allowing the underwriter or relevant department responsible to make an informed decision.

References:

TIBCO; 2021; What is Data as a Service (DaaS)?; Available from: www.tibco.com/reference-center/what-is-daas (Accessed 18 th Jun 2021)

Narayanan, Karpagam; 2018; The Evolution of Data Available from: www.forbes.com/sites/forbestechcouncil/2018/07/17/the-evolution-of-data/?sh=5323d235c95f (Accessed 18 th Jun 2021)

Bisgaard-Bohr, Mikael; 2018; Is Data Really an Asset? Available from: www.teradata.com/Blogs/Is-Data-Really-an-Asset (Accessed 19 th Jun 2021)

Marr, Bernard; 2017; What Is Data Democratization? A Super Simple Explanation And The Key Pros And Cons; Available from: www.forbes.com/sites/bernardmarr/2017/07/24/what-is-data-democratization-a-super-simple-explanation-and-the-key-pros-and-cons/?sh=402c68fc6013 (Accessed 20 th Jun 2021)

Debanjan Saha; 2020; How the world became Data-Driven and what's next. Available from: www.forbes.com/sites/googlecloud/2020/05/20/how-the-world-became-data-driven-and-whats-next/?sh=45f487be57fc (Accessed 19 th June 2021)

Deghani, Zhamak; 2020; Data Mesh Principles and Logical Architecture; Available from: www.martinfowler.com/articles/data-mesh-principles.html (Accessed 20 th June 2021)