Data-as-a-Product using distributed data architecture and smart data platform on GCP: POV

Augmenting data lakes with a modern distributed scale data architecture can help democratize data to derive more insights and business value.

By: Mansi Maharana, D&A specialist

Imagine data as a product that can be easily found, trusted, and acquired securely just like shopping for retail products. Data-as-a-product can help gain insights and drive business value. Data can be turned into a high quality consumable product through adaptation of a set of principles along with design thinking. Today, incumbent technologies and data architecture hinders such transformation. However this can be addressed with a combination of distributed data architecture, unified data fabric and data exchange technologies

Data Mesh architecture

A domain-driven, distributed & decentralized approach to promote data as a product, which goes beyond data sharing, to guarantee quality and ownership.

Smart Data platform

Smart data platform is convolution of incumbent data lakes and data warehouses with data fabric to bolsters data mesh implementation and building data products.

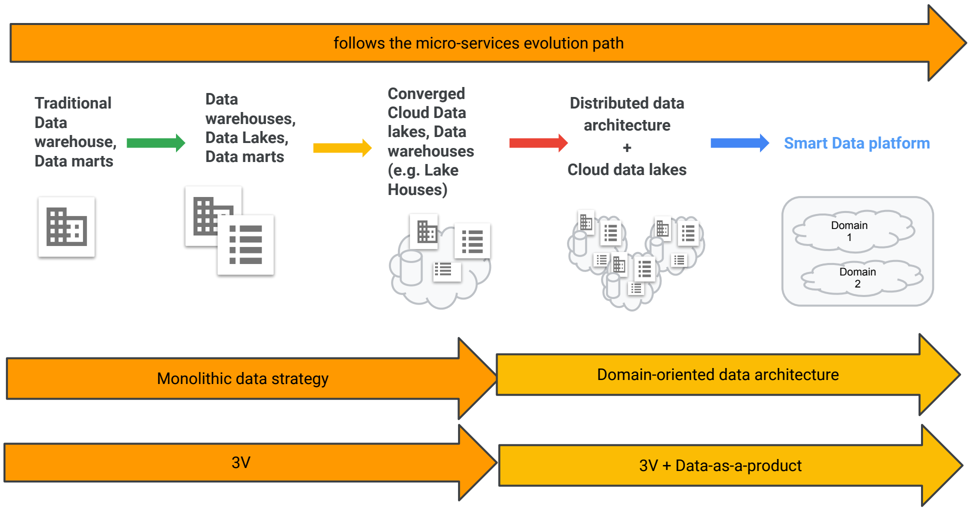
The challenge and the approach[1]

Data has evolved at an unprecedented pace. Big data technology has revolutionized the way data can be captured, stored, and processed. Organizations have evolved from traditional Data warehouse to on-premise Data lakes to Cloud based Data Lakes and Lake Houses. However, extracting value out of data using the existing monolithic Datalakes (even ones on Cloud) with no standard data architecture still remains a challenge.

Zhamak Dehgani describes this evolution of data, the challenges and a way to solve it by using a new data architecture called "Data Mesh".

Data lakes here to stay, by augmenting Data lakes with a modern distributed scale data architecture like Data Mesh, you can democratize data to derive more insights and business value out of it. It will help pave way for innovation with data and analytics for example - open banking, data and analytics monetization etc..

This paper will describe how we can apply the Data mesh techniques - distributed domain driven architecture, self-service IaaS and Data-as-a-product principles by using state-of-the-art products, tools, and services on GCP. This platform will be referred to as "Smart Data Platform", a platform capable of delivering high quality data-as-a-product for both internal and external consumption.


follows the micro-services evolution path
Traditional Data warehouse, Data marts
Data warehouses, Data Lakes, Data marts
Converged Cloud Data lakes, Data warehouses (e.g. Lake Houses)
Distributed data architecture + Cloud data lakes
Smart Data platform
Domain 1
Domain 2
Monolithic data strategy
Domain-oriented data architecture
3V
3V + Data-as-a-product

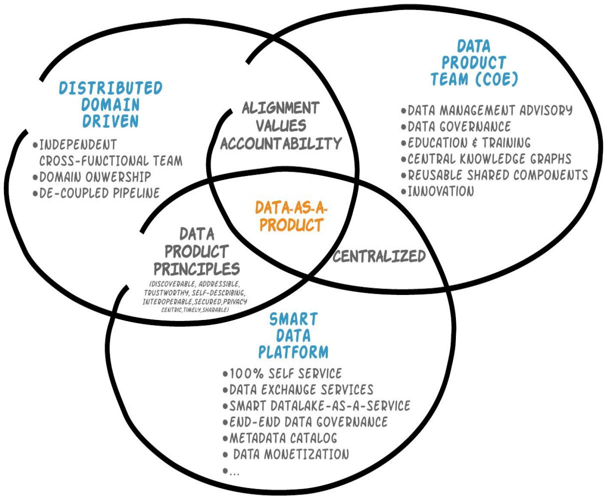
Underpinning components and its key characteristics

Distributed domain-driven data architecture*

A data architecture where the business or internal operations owns, processes using standardized decoupled pipelines, hosts, and serves their domain datasets in a secured and easily consumable way. A domain is responsible for providing high quality data products to its consumers who can be internal or external to the organization. You can treat each logical group in your business or internal operations team that serves or consumes data as a domain. Each domain can consist of one or more data products. A data product can be—a dataset, a file shared over ftp or shared drive, a data-based report, one or more data elements consumed by API, or a stream of data. A domain should ensure data-as-a-product principles are adhered by each of their data products. Typically a domain can be source or consumer or internal operations oriented. A source oriented domain should be system-of-reality and close to data sources and can be limited to being producers only. Each domain should be operated by an independent cross functional team which at a minimum should include a data product owner, a data engineer, a data analyst, and a data steward.

Use smart data platform

Using the right data platform will help catalyze the data-as-a-product adoption. The idea of building all the tools and services to support data-as-a-product can be taxing especially in a monolithic data platform. For example: data quality or cataloging metadata KPIs are often remain undefined and lacks proper ownership. Most of the time developers are deemed responsible of doing this as part of their application or pipeline development. This hard to implement as they are not data experts and does not scale as data grows. Such problems can be solved by a Smart Data Platform, which can provide turn key solution for DQ as data is ingested into domain specific data lake. A smart data platform can: 1. Provide a unified and intelligent data fabric capabilities to ease the data management despite where the data resides and without the need of data movement or duplication 2. Facilitate data exchange/sharing and monetization within and outside the organization. 3. Can be centrally aligned and governed. A data fabric can be responsible for providing capabilities such as turn key solutions for Data Quality, automate discovery and compliance jobs, unified Data governance etc. These capabilities will help ease implementation of data-as-a-product principles.


DISTRIBUTED DOMAIN DRIVEN
•INDEPENDENT CROSS-FUNCTIONAL TEAM
•DOMAIN ONWERSHIP
•DE-COUPLED PIPELINE
ALIGNMENT VALUES ACCOUNTABILITY
DATA PRODUCT TEAM (COE)
•DATA MANAGEMENT ADVISORY
•DATA GOVERNANCE
•EDUCATION & TRAINING
•CENTRAL KNOWLEDGE GRAPHS
•REUSABLE SHARED COMPONENTS
•INNOVATION
DATA-AS-A-PRODUCT
DATA PRODUCT PRINCIPLES
(DISCOVERABLE, ADDRESSIBLE, TRUSTWORTHY, SELF-DESCRIBING, INTEROPERABLE, SECURED, PRIVACY CENTRIC, TIMELY, SHARABLE)
CENTRALIZED
SMART DATA PLATFORM
•100% SELF SERVICE
•DATA EXCHANGE SERVICES
•SMART DATALAKE-AS-A-SERVICE
•END-END DATA GOVERNANCE
•METADATA CATALOG
• DATA MONETIZATION
•...

Data-as-a-product principles*

Distribution of data and ownership can raise concerns around harmonization: this can be overcome by incorporating these design principles into each of the data products— Discoverable, Addressable, Trustworthy, Self-describing, Interoperable, Secured, Privacy-centric, Timely and Shareable. This will help build best-in-class data products for both producers and consumers.

Establish a Data Product CoE

While data mesh is centered around de-centralization, there is a need of a central team to ensure alignment, share values and accountability. Responsibilities at high level can include—ensuring uniform adoption of principles, best practices, and standards across domains, conducting education and training, defining and enforcing data governance policies and compliance (especially with external data sharing), creating & maintaining global knowledge graph, provide domain agnostic reusable components (e.g. CI/CD), provisioning tools, ingestion frameworks, data connectors, standard api interfaces and documentation templates, etc. d last but not least pave the path for innovation.This is beyond the scope of this paper.

Smart Data Platform key building blocks

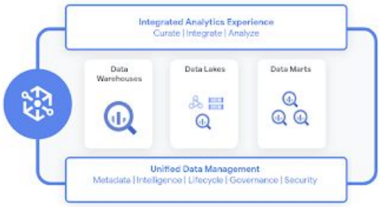
Self service data infrastructure

With more than 100+ products to offer, Google Cloud Platform provides **infrastructure as a service**, **platform as a service**, and **serverless computing**. All of which is empowered by Google cloud's trusted **global** presence, secured and efficient **sustainable data centers**, fast and reliable global **network**, **multi-layered** security, **highly available.**

Learn more

Data fabric - Dataplex(in Preview)

Intelligent data fabric that enables organization to centrally manage, monitor, and govern data across data lakes, data warehouses, and data marts. Dataplex is built for distributed data and enables data unification through a logical layer without data movement or duplication.It provides data intelligence by automating data discovery, data classification, schema detection, global data quality checks and many other capabilities.It also provides centralized security and governance for global control with distributed ov


Integrated Analytics Experience
Curate | Integrate | Analyze
Data Warehouses
Data Lakes
Data Marts
Unified Data Management
Metadata | Intelligence | Lifecycle | Governance | Security

Data Sharing - Analytics Hub

Analytics hub enables you to share data and insights across organizational boundaries without losing control and authority. It is the foundation for commercial data & analytics monetization.

Learn more

Data Experiences - Looker

Apart from all other advanced ML & BI capabilities Looker's embedded analytics is an effective and proven way of monetizing data. This encompasses areas around internal and external data sharing.

Learn more

API Management - Apigee

Design, secure, analyze, and scale APIs anywhere with visibility and control. Provides the ability to build Monetize API products and maximize the business value of digital assets.

Learn more

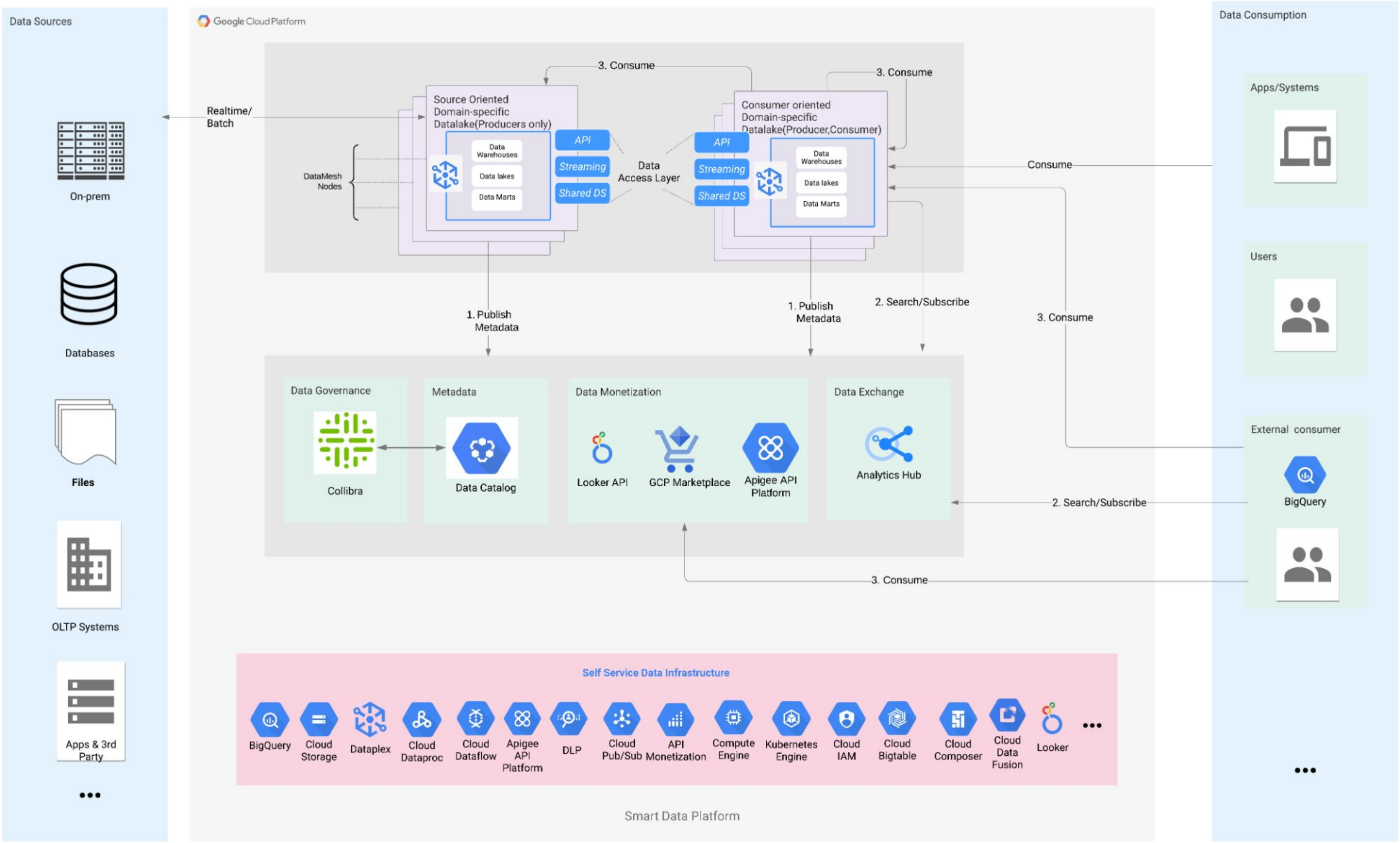
Data governance - Data Catalog and Collibra(SaaS)

Collibra and GCP provide a strong foundation for data governance at scale. With Data Catalog you can enforce data security policies and maintain compliance through Cloud IAM and Cloud DLP.

Learn more

> ”Gartner argues that although some organizations have begun to invest in big data technologies in relation to their customers, with a view to direct or indirect monetization, many organizations lack business models to monetize their data.

Google Cloud



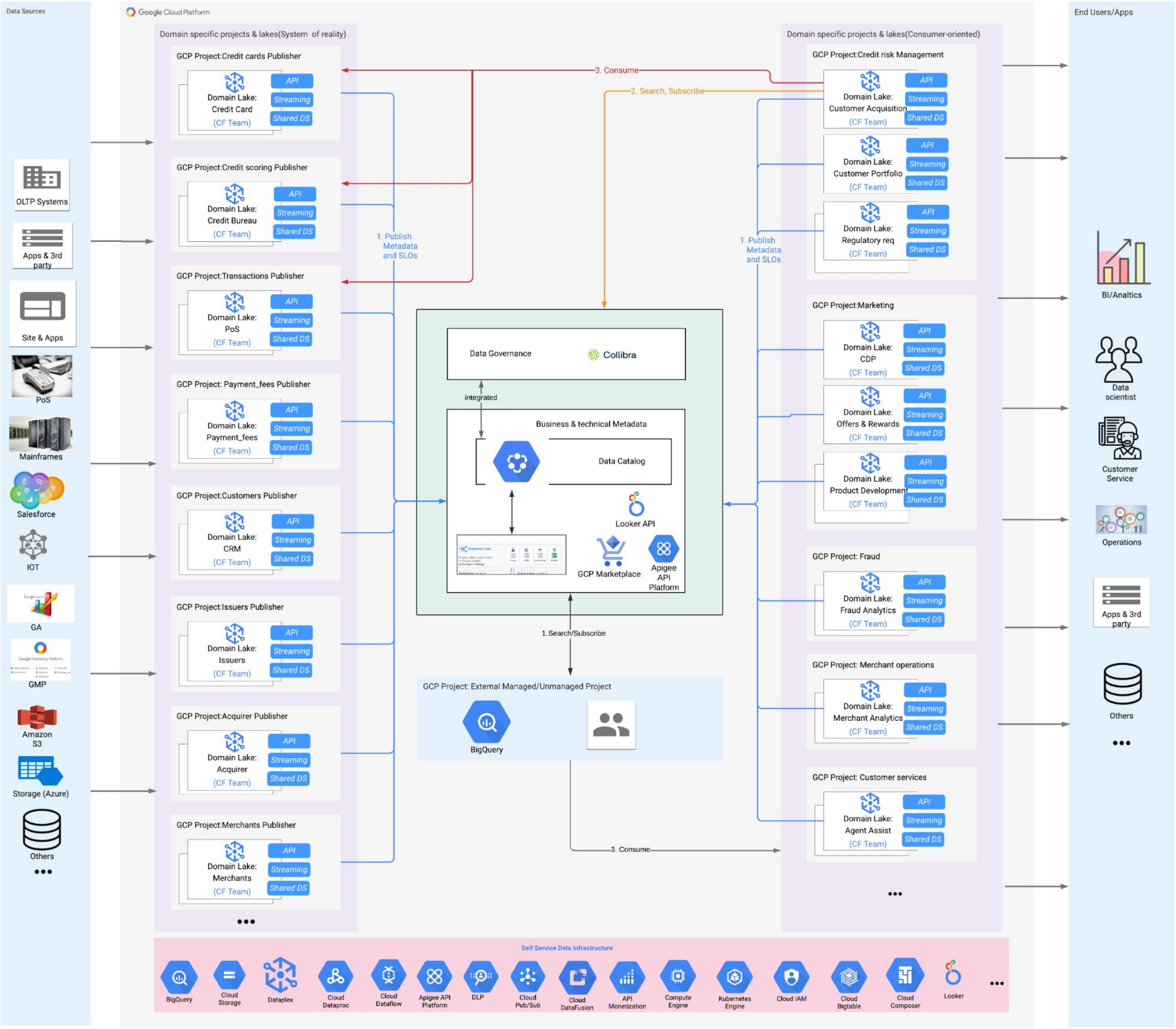
High-level reference architecture on GCP

> “If you’re embarking a journey to Cloud or looking to drive value out of your data or reduce data management overhead then it’s the right time to rethink and modernize your data architecture!”

A Sample Case Study Example

Here is a sample reference architecture of Data Mesh Implementation for a FSI organization that provides credit cards as well as payment services. Let's assume there are a few Line of Businesses(LOB) or Business Units(BUs): Credit cards, Credit scoring, Transactions,Customer, Issuer, Acquirer, Merchant, Payment and Fees, Marketing, Risk, Fraud.

The LOBs can have a source-oriented domain lake(producer only) closest to source of data, owned by a independent Cross-Functional (CF) team. BUs can be more consumer-oriented domain lakes driving the end user/app consumption both external and internal.



[1] https://learning.oreilly.com/library/view/data-management-at/9781492054771/ch01.html#introduction
[2] https://martinfowler.com/articles/data-monolith-to-mesh.html#DataAndDistributedDomainDrivenArchitectureConvergenc
[3] https://medium.com/intuit-engineering/the-intuit-data-journey-d50e644ed279e
[4] https://www.sec.gov/Archives/edgar/data/1393612/000139361219000011/dfs1231201810k.htm#s01DB45DF8F2E53E6873B7A87FBA58FC6
[*] Derived from Data Mesh,originally coined by Zhamak Dehghani
[5]Eric Evans, introduced the concept in 2004, in his book Domain-Driven Design: Tackling Complexity in the Heart of Software.