zensar

Integration Data Hub

Whitepaper

An RPG Company

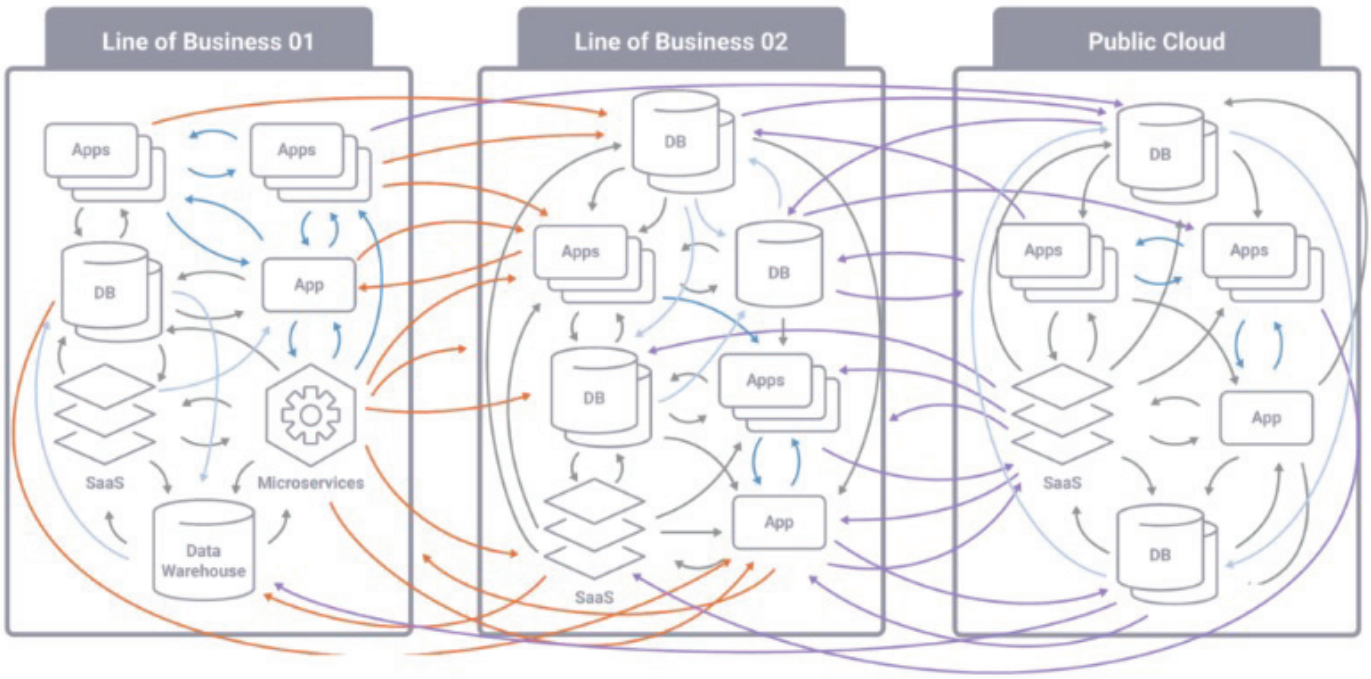
Abstract

This whitepaper is produced to provide a practitioner's point of view on integration data hub (IDH), an integration paradigm that provides a broader guideline on how data systems across an enterprise should connect and exchange data with each other.

Point-to-point integrations

Modern organizations generate a wide range of data as a result of conducting their day-to-day business. Multiple applications for their operational needs may consume the data produced by an application. The number of data-generating and data-consuming systems may vary from organization to organization, depending on the nature and size of the business. However, data exchange, data alignment (the transformation of data to make it useful for consumers), and data distribution among these systems are common needs across all industries, regardless of business vertical or organization size.

While organizations have always had the capability of integrating and aligning the data between systems and applications, these project-focused point-to-point (P2P) integrations have resulted in a spaghetti mesh architecture as shown in the diagram below –



Challenges emanating from P2P integration architecture

The extension of on-the-ground data centers to the cloud and the replacement of bespoke, home-grown applications with SaaS products have transformed almost all organizations into hybrid, multi-cloud data organizations. This, combined with the fact that P2P data integrations have been created over decades, poses the following challenges –

- Data exchange through the legacy integrations is clock-driven, slowing down the rate of information exchange within and outside an organization.
- Because the integrations have been built over decades in a project-funded and isolated manner, there is a multitude of integration technologies. Over time, these integrations become increasingly challenging to understand and maintain.
- Extreme dependencies between all components, which make it practically impossible to modify one component without affecting others.
- Not all the producers and consumers of data are compatible with each other's speed of producing and consuming data. In a hybrid, multi-cloud data center landscape, this becomes a major hindrance to an organization's goal of real-time data integration and processing.
- Often, data systems have their staging areas for data transformation, where they align data from other systems to make it suitable for their consumption or change the shape of data from their system to make it suitable for consumers' needs. These staging areas are isolated from one another, and there may be instances where different teams curate data from the same systems, using the same or similar rules. This leads to coding debt and increased cost of delivery.
- The integration pipelines are laden with complex data transformation rules created over two decades. This makes modernizing integration pipelines very difficult and risky.
- With the potential upgrade of legacy applications to SaaS products, the following challenges will need to be addressed –
    - SaaS applications generally expose data through APIs. Any significant changes to these APIs may be expensive, time-consuming and difficult to maintain with upgrades.
    - SaaS applications are generally reluctant to significant customizations. The customers can raise a ticket for customizations, but this is a time-consuming process.
    - Bespoke data transformations within SaaS environments can prove to be a tedious process with increased cost of ownership.

Core capabilities required for becoming a modern data-driven enterprise

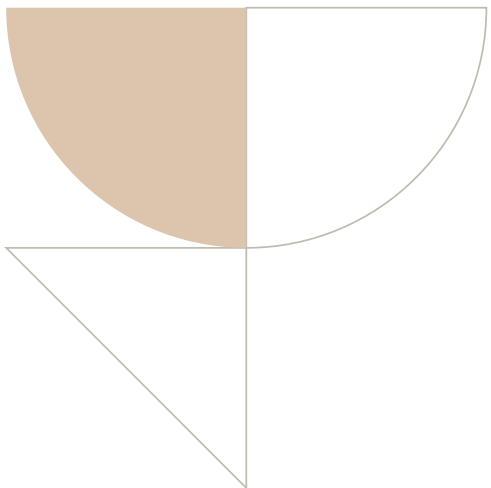
In the previous section, we saw some of the typical challenges that traditional ways of integration and data transformation pose to an organization. In this section, we look at some of the critical integration and data transformation capabilities that an organization must have to fulfill its dream of becoming a future-fit, data-driven enterprise –

- Data integration should be able to connect both operational (run-the-business) and analytical (observe-the-business) applications
- Integration should support a wide range of latency requirements, ranging from RT to batch
- Integration should be able to handle all data types without limitations – structured, semi-structured, unstructured
- Data processing should be able to handle both bounded and unbounded sets of data
- Elasticity to handle massive enterprise-wide data volumes
- The platform should have the ability to store data and metadata together for lineage and governance
- The platform should have robust governance and easy data access policies for safe and secure data sharing
- The platform should have infrastructural modularity, portability, and independent scalability of storage and compute.

Almost every organization that we work with faces the challenges posed by legacy integrations and data transformation. Additionally, every organization aims to become a data-driven and data-enabled organization, which requires all the capabilities listed above.
If we combine the challenges posed by legacy integrations and data transformation, and the goal of an organization to become a data-driven and data-enabled enterprise, the following three points summarize this section -

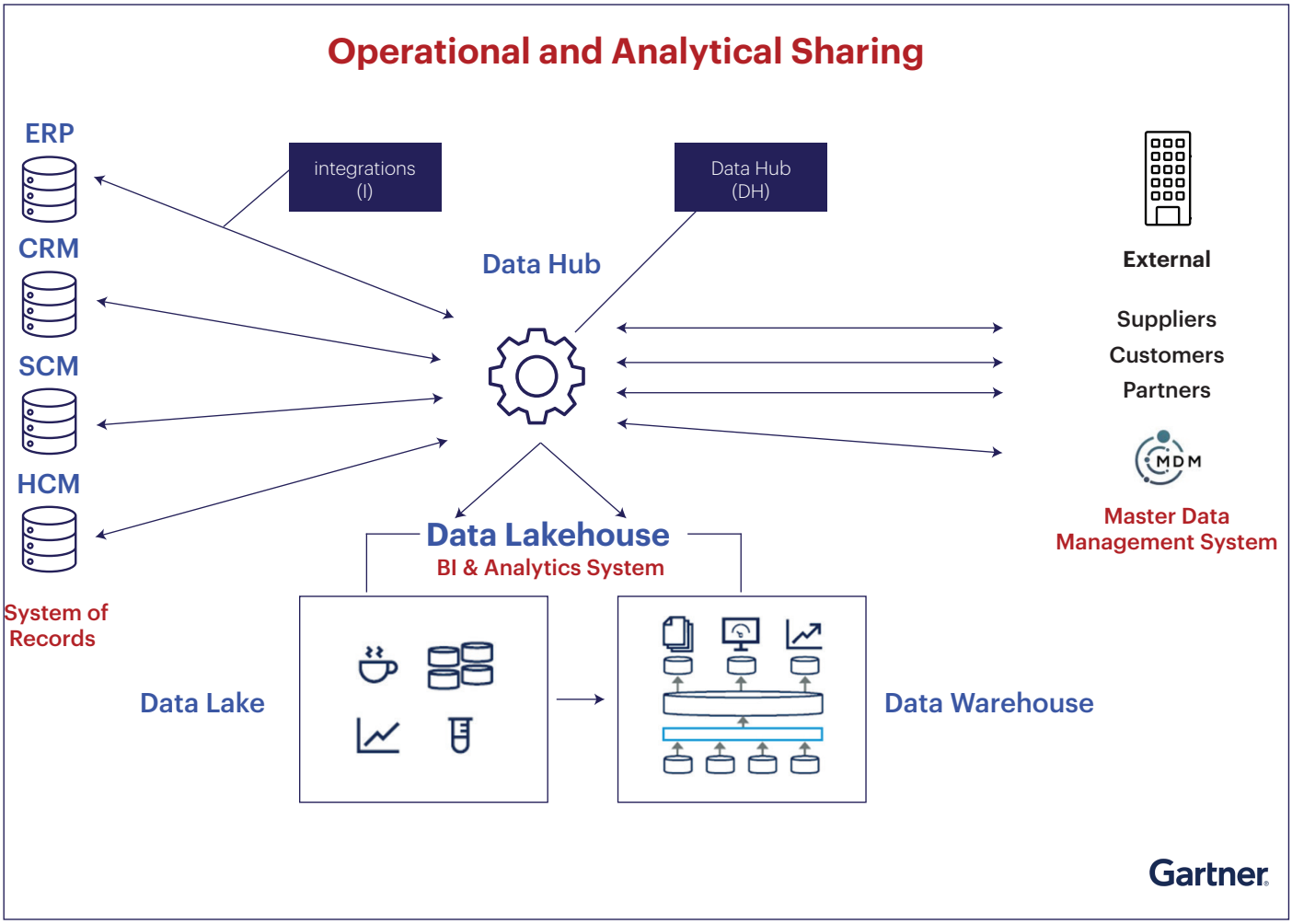
- Decouple producers and consumers of data by breaking P2P integrations
- Provide logically isolated staging areas to different teams for them to carry out data transformation
- Find a platform that supports all the integration needs of an enterprise, from legacy batch to modern real-time

This is where IDH, as a data integration and transformation architectural framework, comes to the rescue.



What is IDH

Gartner defines IDH (Integration Data Hub) as "the logical architecture that enables data sharing by connecting producers of data (applications, processes, and teams) with consumers of data (other applications, processes, and teams) in a hub and spoke model. Endpoints interact, either provisioning data into them or receiving data from them. The hub provides a point of mediation and governance, and visibility into how data is flowing across the enterprise."



IDH consists of two components: Integrations (I) and Data Hub (DH). Integrations faithfully carry data to and from the hub, without altering the data's shape or meaning. They are the 'homing pigeons' in the data exchange phenomenon between a producer and a consumer. The Data Hub, on the other hand, is a physical platform that decouples producers and consumers of data by breaking the peer-to-peer (P2P) integrations between them. More importantly, it acts as a multi-tenant platform that provides producers and consumers of the data with logically isolated staging areas where data can be aligned (transformed) to suit different consumers.

IDH complements enterprise data systems

IDH is an architectural framework that enables an organization's data systems to exchange data seamlessly with each other. It houses logically separated staging areas, empowering data engineers from source or target IT teams to transform the shape of raw data coming from producers, aligning it with the operational needs of the consumer systems.

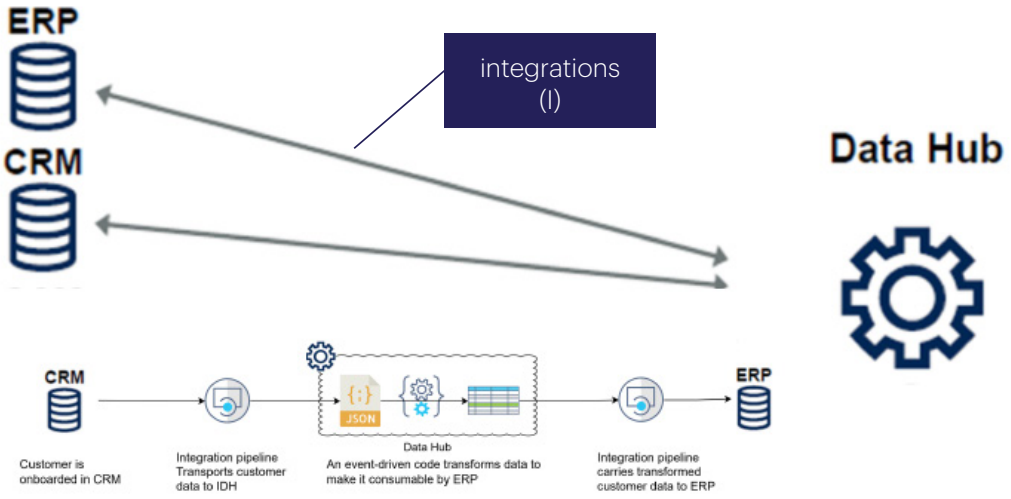
Another interpretation of the above diagram is that IDH serves as a transient repository of data, meaning it should not be used to persist historical data, which is typically stored in a lakehouse. The urge to persist data beyond the needs of data exchange must be resisted. Persisting data on IDH may lead to increased costs due to data redundancy and governance overload.

As we have seen in the previous section, IDH is merely an intermediary between data systems within an organization. However, we have observed that there is often misalignment among various stakeholders within a client organization, where different leaders perceive IDH differently, leading to conflict and confusion. Therefore, it is essential to clarify how IDH differs from other data systems and how it complements them.

System of records

These are applications where the data originates and is mastered. These systems are typically transactional or operational by nature. Some examples include CRM, POS, HCM, Planning etc.

IDH serves as an intermediary between the enterprise's data systems. While it provides individual staging areas for application teams to meet their operational data transformation needs, it must not be viewed as a system of records.



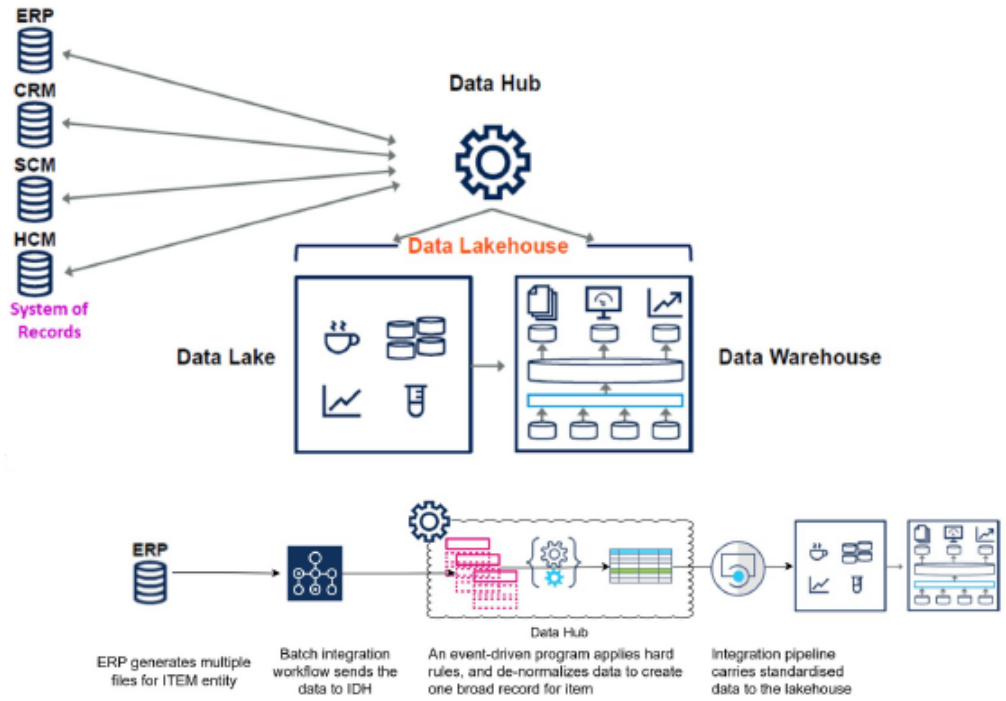
An extension of the above point is that transactional applications should not rely on IDH as a backend data repository for running their operations.

- **System of business observation or BI and analytics systems**

These systems receive facts and context of the business events from transactional systems via IDH. These systems store a vast amount of historical data to enable business intelligence (BI) reporting and analytics. A traditional EDW, or a modern lakehouse, is an example of such a system.

An IDH can be leveraged to left-shift the raw zone and implement hard rules required for cleansing the raw data. This means that the first layer within the enterprise lakehouse is the standardized / silver layer, which contains clean, historical data. This makes the data lake much cleaner.

.



Similarly, IDH can be leveraged to left-shift some of the logic required to prepare data for systems like MDM and ODS.

.

IDH reference capability model

The term IDH is a combination of 'Integration' and 'Data Hub'. In a hub-and-spoke model, ‘Integration’ refers to the two-way data transport capability between the hub and data systems. The ‘Data Hub’ should possess the necessary data engineering capabilities to transform streams of data.

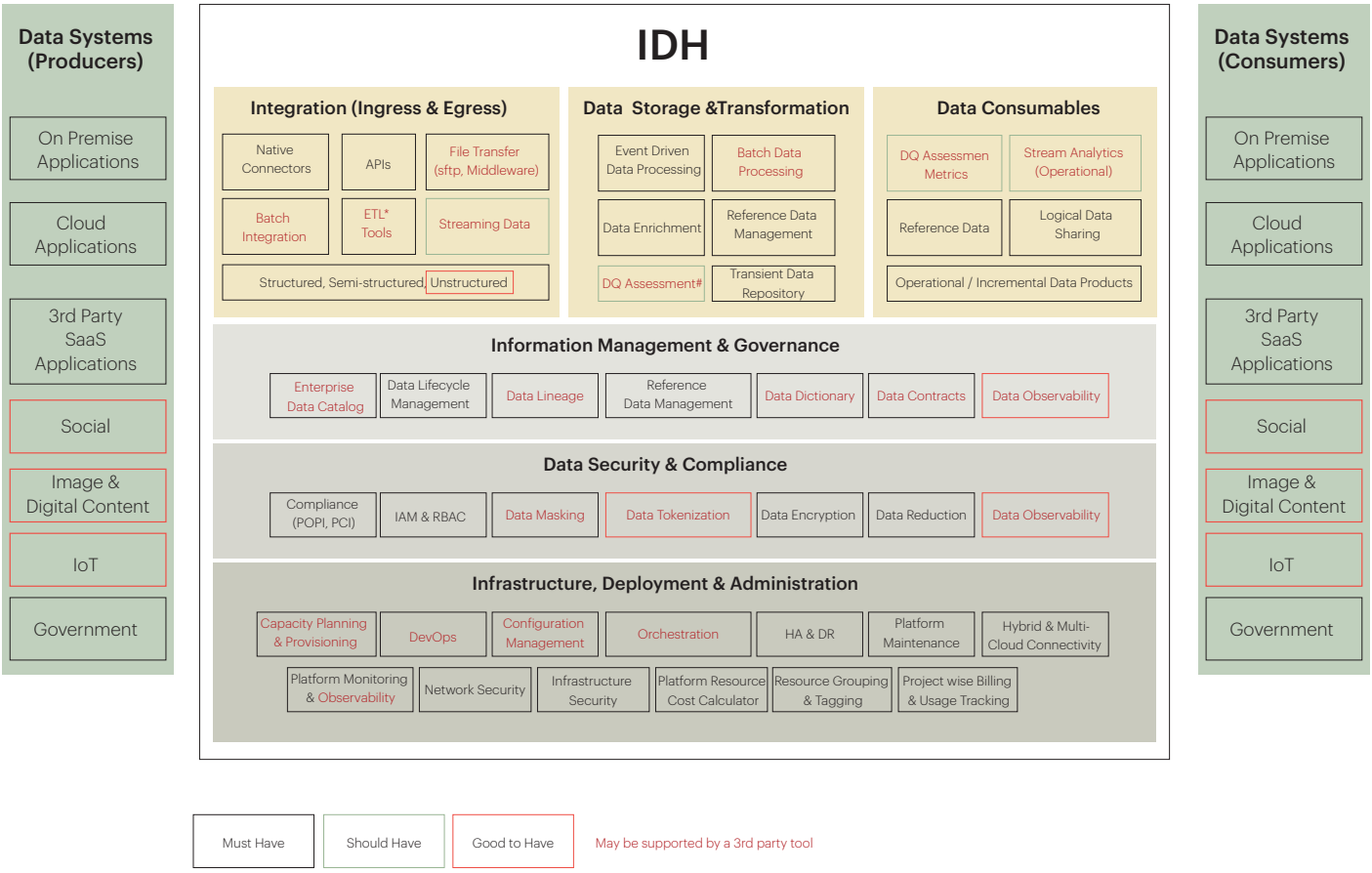
Keeping the above two aspects of IDH in mind, we have developed a reference capability model for IDH. This capability model is based on our experience of implementing IDH.

.

These systems receive facts and context of the business events from transactional systems via IDH. These systems store a vast amount of historical data to enable business intelligence (BI) reporting and analytics. A traditional EDW, or a modern lakehouse, is an example of such a system.

An IDH can be leveraged to left-shift the raw zone and implement hard rules required for cleansing the raw data. This means that the first layer within the enterprise lakehouse is the standardized / silver layer, which contains clean, historical data. This makes the data lake much cleaner.

.



* Whereas 3rd party ETL tools can be leveraged as a means of integrating data between systems and hub, as much as possible, data transformations should not be performed in the Integration layer
\# Only for the source systems that may not have the technical capability of integration with a DQ tool. Assessment does not include writing / implementing the code for fixing the data on EIDH

Reference capability model – conceptual view

While all the building blocks of IDH have their importance, we believe that information management and governance is the most critical aspect towards the success of IDH as an enterprise platform. A data catalog, for example, is a register of all the entities on IDH. A tenant can browse through the catalog to see if what they need is already available, verify through lineage that the source of information is correct, reach out to the entity's owner, and access the entity after a formal data contract has been established.

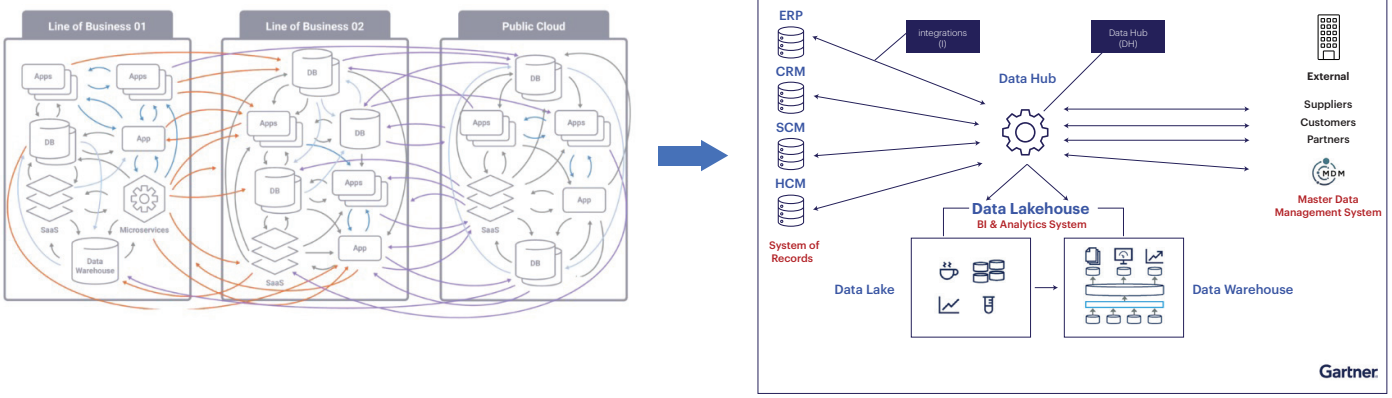
Various platforms can be leveraged as IDH. Stream integration is the future of data exchange, and any platform that supports real-time data streaming, along with data transformations, can be leveraged as an IDH. There are various such platforms available in the market, from open-source to cloud-native and cloud-agnostic PaaS products that can be considered.

Based on your organization's needs and budget, we can help you select the most suitable tools and platforms for implementing IDH.

From spaghetti mesh to hub-spoke integration

The idea of IDH is refreshing, and if implemented correctly, it has the potential to transform a legacy data organization into a modern, future-proof data-driven one. However, the journey from legacy spaghetti mesh architecture to modern IDH can be overwhelming.

.



The following are some of the major challenges that organizations face while trying to move from spaghetti mesh to a hub-spoke model:

.

- The integration pipelines in a P2P architecture are built over decades, using every tech stack possible, having complex transformation logic, and connecting mission-critical applications. When people who create the integration pipelines leave the organization, there is often a lack of adequate documentation to understand the complexity of the pipelines.
- 'Don't fix if it is not broken' is another mental wall that we have encountered during an enterprise-grade change program.
- Business sponsors are generally reluctant to invest in a transformation program if the business benefits that the transformation can bring are not articulated properly. It is also crucial to link these benefits to the way business is conducted and to align the program with the overall business and data strategy.
- There is a transition phase where both P2P integrations and IDH will need to co-exist. Depending on the length of the transition period, significant resources — human, capital, and financial — may be required to keep both the cities running.

The journey to conceptualize, create, and mature IDH can be daunting. Many times, enterprise-focused initiatives like IDH fizzle out due to a lack of planning, resources, and budget.

Having worked with some of the largest organizations, this is what we have seen works –

- See how stream processing contributes to an organization's business goals and objectives
- Create or modify business and IT data strategies to include stream processing
- Include IDH as the de facto architectural block for all the data projects going forward
- Conduct roadshows to spread awareness of IDH and how it can benefit businesses
- Choose an appropriate platform based on your use cases and budget

- Stand up a team of SMEs for the platform. This team will work as the data engineering CoE on the platform
- Work iteratively to create new integration and data transformation pipelines leveraging IDH. Work packets can be created in multiple ways. Licence renewals for all legacy integration and transformation platforms require critical review. New projects should be encouraged and incentivized for using IDH within their data architecture, etc.
- Adopt new ways of working. DataOps and DevOps should be leveraged together to deliver data pipelines to IDH in a continuous manner
- Invest in an enterprise catalog platform. Our experience tells us that a well-maintained catalog is the best advert for popularizing any data platform.
- Zensar is a thought leader in the field of data architecture and governance across hybrid and multi-cloud data centres. With IDH as one of its key focus areas, Zensar has been helping clients across the US, Europe, and South Africa on their data estate modernization journey, starting from data strategy and roadmap development to program and portfolio management, and delivering data analytics and data science use cases that leverage best-in-class data architecture and data engineering.

Case in point

Zensar is currently working with a marquee retail client in South Africa that aims to modernize its IT systems. Replacing legacy custom business and IT applications with state-of-the-art equivalents, upgrading COTS products with SaaS solutions, migrating on-premises data centers to the cloud, and introducing modern data paradigms, architectures, and frameworks are some of the initiatives the organization has undertaken.

Zensar has been a trusted partner on this journey, serving as advisor, planner, and executor. We collaborate with sponsors from various business units and their representatives, as well as CXOs, Heads of data engineering, and architecture, to co-create data strategies, implementation roadmaps, architectural frameworks, and principles for modern data paradigms, including data mesh, data fabric, data lakehouse, and AI lakehouse.

IDH is one of the key capabilities that we have conceptualized and implemented as part of the data strategy. In the legacy landscape, data systems scattered across hybrid, multi-cloud data centers were integrated in a peer-to-peer manner. These data systems include legacy mainframes, on-the-ground databases, AWS, Azure, and third-party SaaS applications.

After due diligence, Confluent Cloud was chosen as IDH for the following reasons –

- Confluent cloud offers out-of-the-box connectors that support the extraction and delivery of data from and to various platforms. It also supports third-party integration and EiPaaS platforms.
- Apache Flink, the open-source framework for both stream and batch data processing, is now fully integrated with Confluent Cloud. In addition, Java-based UDFs are also supported natively.
- Confluent Cloud offers native information management and governance capabilities, including stream catalog, stream lineage, and data contracts. It also supports integration with an enterprise catalog platform, such as Collibra.

The following are the benefits that the client has already started reaping after the first phase of IDH implementation –

- Because the integrations with the data producers are data-focussed, there is a huge reduction in the number of data extraction pipelines. On average, we see 30%-40% fewer pipes to maintain.
- Data-focussed integrations mean that resource needs do not constrain the source systems. This has led to improved performance of source systems.
- The introduction of the catalog has ensured that the data available on IDH is visible to everyone within the organization. This has resulted in faster delivery of projects as the consumers don't need to lay down a pipe all the way to the source system for fetching data
- Newly adopted SaaS applications have begun pushing data to IDH through webhooks. Real-time data processing and consumption are improving inventory management.
- Because the data transformation is carried out within the hub, it relieves the operational and business teams of the additional burden of maintaining a data platform. This has resulted in significant cost savings for tenants.

Conclusion

Stream data processing and real-time information exchange are the core capabilities that a modern data organization must possess. IDH, as the backbone for stream integration, processing, and data exchange, should be part of a modern data estate. It not only helps in decluttering and decoupling P2P data integration pipes but also reduces the burden of housing complex data transformations within source systems, pipes, and target systems, while facilitating real-time data exchange between applications and systems throughout the enterprise.

There are various products available in the market for implementing IDH. The choice of product should be based on the organization's specific use cases, maturity level, and budget. The implementation and adoption of IDH should be gradual yet continuous. Modern implementation approaches, such as DataOps and DevOps, should be employed to deliver integrations consistently and reliably.

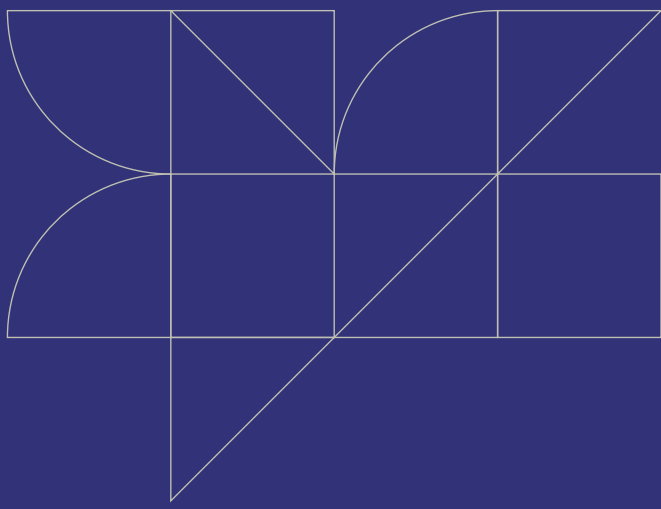
Zensar is an Industry leader in the area of data engineering and analytics. We would be more than happy to engage with organizations in conceptualizing, developing, implementing, and maintaining IDH as part of their enterprise data strategy.

Author:

Vikas Yadav
Enterprise Data Architect
Data Engineering and Analytics
v.yadav@zensar.com

Reviewer:

Rahul Athalye
Head – Enterprise Solutions
Data Engineering and Analytics
r.athalye@zensar.com



zensar
An RPG Company

At Zensar, we're 'experience-led everything.' We are committed to conceptualizing, designing, engineering, marketing, and managing digital solutions and experiences for over 145+ leading enterprises. Using our 3Es of experience, engineering, and engagement, we harness the power of technology, creativity, and insight to deliver impact.

Part of the $4.8 billion RPG Group, we are headquartered in Pune, India. Our 10,000+ employees work across 30+ locations worldwide, including Milpitas, Seattle, Princeton, Cape Town, London, Zurich, Singapore, and Mexico City.

For more information, please contact: **info@zensar.com | www.zensar.com**