Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 12, Issue 12, December 2023

Holistic Approach for Data Extract & Publish Strategy for Enterprises

Vishal Vikas Javalkar
IT Dept ,Cencora Dallas,USA

***Abstract* – Data integration plays a crucial role in enterprises that rely on multiple software applications. It involves the extraction of data from one application and making it available for consumption by other applications, enabling seamless data sharing. However, many enterprises often adopt tactical approaches to address data integration challenges, resulting in non-standardized and inconsistent practices. This white paper emphasizes the importance of taking a holistic approach towards data extraction to overcome these challenges effectively. It provides guidance to Enterprise Architects, Software Architects, and Technology Leaders on developing strategies for data extraction that align with the enterprise's overall objectives. By adopting a holistic perspective, enterprises can ensure standardized and consistent practices, leading to improved accuracy and consistency of shared data while reducing management costs. The focus of this white paper is primarily on the data extraction & publish aspect of data integration. It aims to assist enterprises in understanding the significance of a comprehensive approach and its impact on addressing data integration challenges successfully.**

***Keywords*—Data Integration, Data Streaming, Integration Patterns, Anti-Patterns, ETL (Extract Transform Load), Architecture Principles**

I. INTRODUCTION

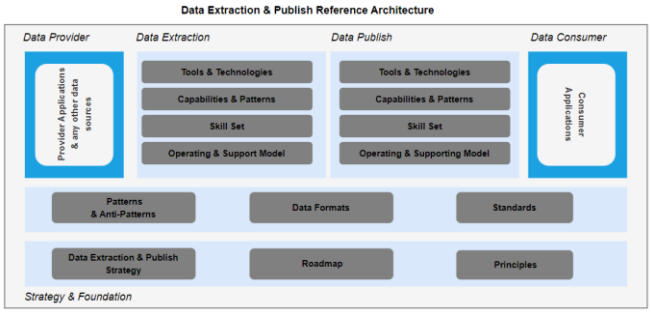
Data integration is a ubiquitous requirement for enterprises, yet many struggle to address it holistically, resulting in common challenges such as poor data consistency, untimely availability of data, and unreliable information. This whitepaper aims to provide enterprises with a comprehensive approach to tackle these issues by focusing on the data extraction and publishing strategy. By adopting the suggested holistic approach outlined in this whitepaper, enterprises can ensure that data is made available to consumer applications in near real-time in consistent manner. Overall, this whitepaper serves as a valuable resource for enterprises seeking to enhance their data integration capabilities and overcome common challenges through a holistic approach centered around efficient data extraction and publishing.

A. Approach

Strategic approach instead of Tactical approach Data Integration requirements come all the time as new solutions are implemented in an enterprise and hence it gernerally looked from a perspective of a given solution or project where the larger context of Data Integration is often missed and tactical architecture options are selected. Its important to approach as Data Extraction as critical capability of Data Integration and must be approached strategically. A holistic Data Extraction Strategy should be established where there are agreed prinicples , patterns , standands supported enterprise aligned tools and technologies are leveraged.

A. Reference Architecture for Data Extract & Publish

Typical Data Extraction & Publish Reference Architecture should consider all of the below aspects as depicted in diagram to approach the data extract and publish strategy.



1) Strategy & Foundation : Its critical to establish a long term strategy for data extract and publish. Ideally have a strategic outlook for 3-5 years.

Some of the foundations aspects to be considered while developing the strategy :

a) Principles : Develop and align principles with all the key stakeholders. Principles should be non-ambigous , aligning with the strategy and should be agreed within the architecture group.

Examples of principles

Principle 1 : Reuse - Discover before build

Principle 2 : No point to point Integration

Principle 3 : Choose Near Real time Iintegration over Batch Integration

b) Roadmap : Assess the current capabilities and make a clear roadmap of capabilities around Data Extract and Publish solutions. You can decide to use Cloud PaaS services of any public cloud or choose a SaaS product for data extract & publish layer

c) Patterns & Anti-Patterns : Define Patterns and Anti-Patterns. Ensure its socialized with all the application development teams and architects within the enterprise. Please refer ‘section C’ for details around Patterns and Anti-Patterns.

d) Data Formats : There are many data formats available when it comes to data integrations. Agree upon the data formats and choose what best suit your enterprise. Some of the data formats best suited for data integration within enterprises are Avro , Parquet , Json and CSV.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 12, Issue 12, December 2023

e) Standards : Prefer use of standards when it comes to connectivity between the applications for example use secure protocols like https and TLS 1.2+, Encryt sensitive data use standard libraries.

2) Data Provider & Consumer: Data Provider in an enterprise are typically transactional applications or databases for applications like ERP , CRM , HR, Finance which are source of truth for the data. Data Consumer are generally downstream system which depends on the source of truth like Data warehouse, Data Lakes, Data Lakes House , any other applications.

3) Data Extraction : Data Extract Layer is one of the critical layer responsible for extracting data from provider systems. There are primariliy two ways to extract data from provider systems "Push" or "Pull" . Push mechanism is used to get the data from provider systems as soon as the data is created or modified in the provider system. This method is best suited when the data needs to shared as soon as the data is created or modified and is the basis for real time or near time data push, this is also called as Change Data Capture (CDC). In Push mechanism the control is with the Provider system. Pull mechanism is when Consumer system dictates when the data is required , this mechanism is suitable for small dataset and control here lies with Consumer system.

Data Extraction layer can be an ETL tool which can connect to variety of data sources and applications on the provider end. This will ensure a single tool can be leveraged to extract data from variety of data sources which helps keep the cost low and enhances reuse. Evaluate various capabilites , technology/tools, skillset and operating/support model for extract strategy. Some of the ETL tools avaialble in the market Azure Data Factory, IBM DataStage, SAP DataSphere, Informatica etc can be used for extracting data. You can select the one which best suit your requirements.

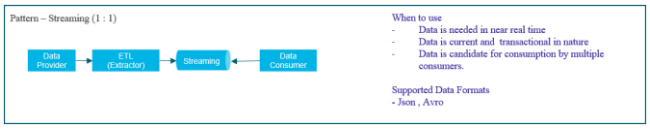
4) Data Publish : Once data is extracted enteprises often end up delivering the data directly to the Consumer applications , this creates multiple different problems like

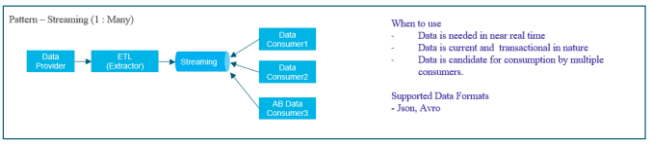
a. Point-to-Point Integration
b. Consumer application Availability constraint
c. Reprocessing of failed data delivery

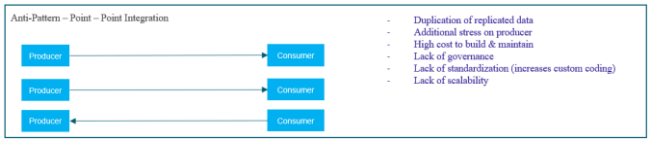
To avoid the above issues, it's important to create a headless and decoupled architecture by adding distinct publish layer in the architecture. Typically, a streaming platform or a Message broker is best suited for this purpose. This enables an Event Driven or Pub-Sub based architecture which decouple the Data provider and Data consumer and creates decoupled, high scalable and resilient architecture for Data Integration. Evaluate various capabilites , technology/tools, skillset and operating/support model for publish strategy Some of the Streaming or Message broker can be used Kafka, Azure Event Hub, RabbitMQ, Amazon Kinesis, Google Pub Sub etc. You can select the one which best suit your requirements.

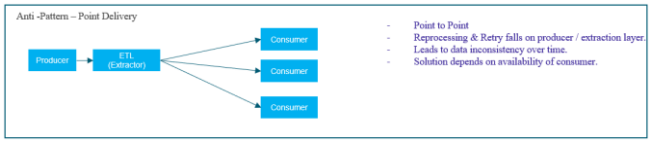
B. Patterns & Anti-Pattern

It's important to define the patterns & anti-patterns and socialize within various application development team to ensure agreed patterns for Data Integration are used and Anti-Patterns are avoided. Here are some of the common patterns and anti-patterns.









Patterns evolve over period, hence make a provision to periodically review the suitability of patterns or any updates which are required to the patterns.

C. Artificats

It's important to develop and maintain certain artifact throughout the journey of Data Extract & Publish Strategy. Some of the artifacts recommended to be maintained are:

a. Principles Catalog
b. Strategy & Roadmap Document
c. Extract & Publish Architecture
d. Patterns & Anti-Patterns Catalog

II. CONCLUSION

In conclusion, this white paper emphasizes the importance of taking a holistic approach to data extract and publish to overcome challenges in data integration effectively. Many enterprises adopt tactical approaches that result in non-standardized and inconsistent practices. By establishing a strategic outlook for data extraction and publishing, enterprises can ensure standardized practices and improve the accuracy and consistency of shared data while reducing management costs. The suggested approach focuses on developing a comprehensive strategy, including principles, roadmap, patterns, anti-patterns, and standards. Furthermore, the paper emphasizes the need for a decoupled architecture with a distinct extract and publish layer, such as streaming platforms or message brokers, to avoid common issues like point-to-point integration and reprocessing of failed data delivery. By following these guidelines and maintaining necessary artifacts throughout the process, enterprises can enhance their data integration capabilities and successfully address common data integration challenges. Overall, this white paper serves as a valuable resource for Enterprise Architects, Software Architects, and Technology Leaders seeking to establish effective strategies for data extraction within their organizations.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Volume 12, Issue 12, December 2023

REFERENCES

[1] Enterprise Integration Patterns
https://www.enterpriseintegrationpatterns.com/
[2] Enterprise Integration Patterns
https://en.wikipedia.org/wiki/Enterprise_Integration_Patterns
[3] Publish Subscribe Pattern
https://learn.microsoft.com/en-us/azure/architecture/patterns/publisher-subscriber