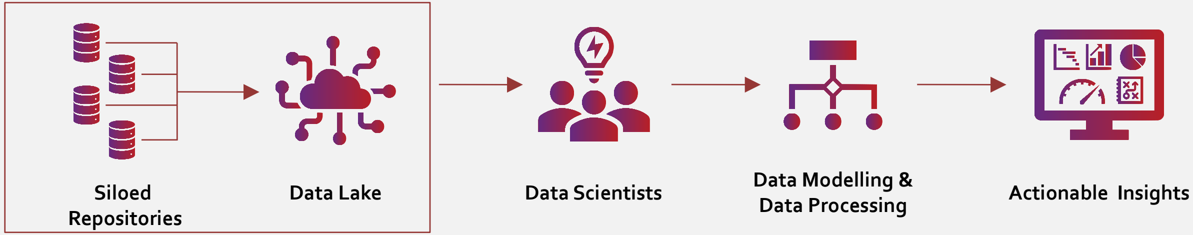
Prodapt powering global telecom

Prevent your data lake from turning into a data swamp

Build a light-weight efficient data lake on Cloud

Credits | Manoj Kumar Sriram V Sathya Narayanan

Highly Efficient Data Lakes Drive Agile and Data-driven Decision Making in DSPs



Procedure to derive actionable insights from data assets

DSPs of the future will be driven by agile and data-driven decision making. Integrating and storing their massive, heterogeneous, siloed volumes of data in a centralized data lake is a key imperative.

A data lake is essential for DSPs to store both structured and unstructured data coming in all formats and from a range of sources, in a single repository and launch analytics programs quickly.

Two different approaches to build a cloud-based data lake

- A *server-based data lake* on the cloud is an agile and highly scalable central repository to store data of any format and structure which benefits from the cloud
- DSPs manage the servers, applications, autoscaling rules on the data lake
- Additionally, integration with analytical tools is a tedious task in server-based data lakes

- Serverless data lakes feature autonomous maintenance and architectural flexibility for diverse kinds of data and abstract away complexity
- Serverless data lake accelerates the data lake building and integration process
- Serverless data lake accelerates integration with analytics engine and improves time to insights

According to Pathfinder, 75% of organizations use or plan to use *serverless technologies* within the next two years, to deliver products faster, save money, and scale efficiently.

Critical Parameters to Build an Efficient AWS Serverless Data Lake

DSPs’ data lakes are not living up to expectations due to reasons such as – data lake becoming a *data swamp, lack of business impact, and complexities in data pipeline replication*. The critical parameters listed here can help DSPs mitigate these challenges and implement a high-performance and efficient data lake

According to Gartner, 80% of Data Lakes do not include effective metadata management capabilities, which makes them *inefficient*

1

Data architectural workshop to Implement *Business-Value-First* approach

Consultant-facilitated workshop, bringing IT & business leaders together to put business value first and drive stakeholder alignment

2

Interface control template to accelerate integrations between serverless services

An interface control template to define and verify necessary integrations between various serverless applications. This helps to avoid much of rework in data lake implementation

3

Infrastructure as code (IaC) to accelerate data pipeline building and scaling

Infrastructure as code is a new paradigm of thinking about cloud-native architecture that’s highly suited for the serverless world. Data lakes scale up resources as well as pipelines multiple times - IaC can be used to accelerate the scaling up of pipelines using customizable configuration files

4

Data cataloging approach to avoid data swamps

DSPs end up hoarding massive amount of data into data lakes without any organization or structure, turning data lakes into data swamps. Strong governance, including data cataloging, is required during the early phase to avoid creating a data swamp.

5

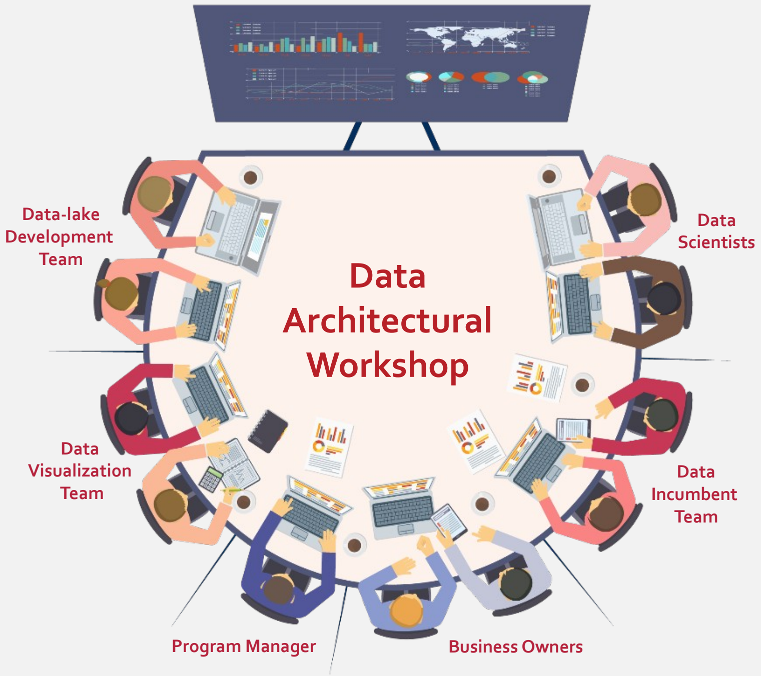
Event-driven orchestration to automate data transformations

Common pitfalls of data transformation in the data lake are time-based batch processing which is ineffective. Event-driven orchestration ensures end-to-end automation of data flow and data transformations negating the time dependency aspect.

Data Architectural Workshop to Implement *Business-Value-First* approach and Effective Architectural Blueprint

1 2 3 4 5

IT and business leaders jointly outline and address relevant technology, design questions and prioritize business cases



Pit falls to avoid

- Most DSP data lake implementations are driven by the IT organization, which may lead to:
  - Highest priority business cases not being addressed.
  - Reduced value creation & adoption rates for the data lake.
- Not creating a long-term blueprint for the data lake with considerations such as performance, scalability, ease of integration, and addition of data sources.
  - Changes to data volumes, complexity and lineage can impact data lake usability.
  - Architectural vetting should not be done without due consideration to the long-term needs or without clear plans for the data lake's use cases.

Recommendations for DSPs: Workshop - Business value-first approach

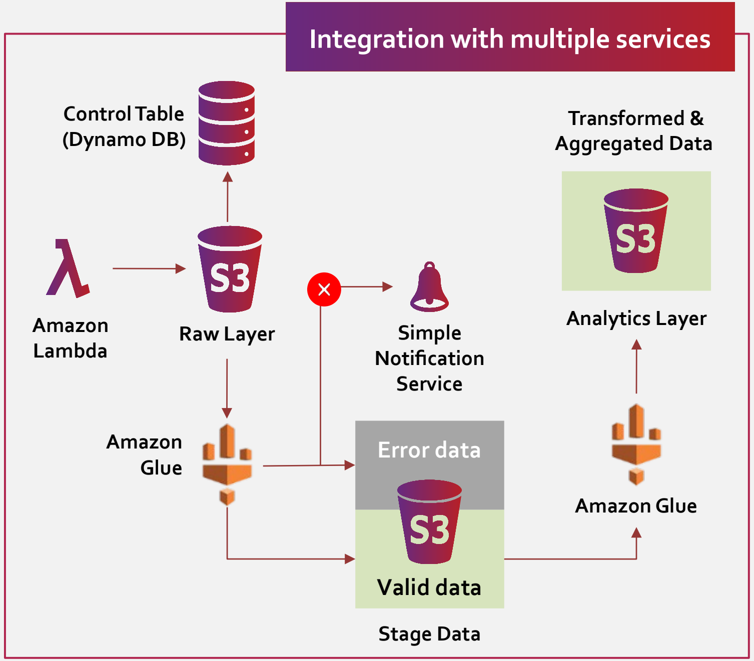
Consultant-facilitated workshop, bringing IT & Business leaders together

- Use a structured approach addressing business needs, cloud and serverless capabilities, design questions.
- Identify the highest priority business cases with strategic cost/benefit analysis and incorporate these elements into architecture and design.
- Build flexibility into the design to address the changing needs of the business, IT and other functions as requirements and use cases evolve.
- Long term architecture blueprint with serverless technology elements, integrations, performance and scalability.
- Identify how the data lake will be populated in detail – what data and events, how and when.

Conducting a consultant-facilitated workshop at an early stage accelerates the requirements gathering phase and improves data lake & pipeline building time by 20%.

Interface Control Template to Accelerate Integrations Between Serverless Services

By the time, the architecture is complete, the necessary integrations between the multiple serverless services need to be intact and verified.


Integration with multiple services
Control Table (Dynamo DB)
Transformed & Aggregated Data
S3
Amazon Lambda
S3
Raw Layer
Simple Notification Service
Analytics Layer
Amazon Glue
Error data
S3
Valid data
Stage Data
Amazon Glue

Pit falls to avoid

- Integrations between serverless applications should be done to accommodate different specifications and needs of the applications.
- The choice of the services and the integrations will depend on the data and events being handled.
- Nature of integration will vary based on the source and target. Interfaces may need to be chosen carefully.
- Once integrations are chosen, DSPs should have a fool-proof mechanism to verify all the integrations.

Recommendation: Interface Control Template

S.No.	Resource	Target	Interface
1	S3	Lambda	S3 events configuration to automatically pass S3 metadata
2	Lambda	S3, SQS, Step Function	Programmatic interface using Boto Library
3	SQS	Lambda	SQS Lambda trigger configuration by allowing send message
4	Step Function	Lambda	Step function template 'Task' & 'Resource' configuration
5	Code Commit	Code Pipeline	Cloudwatch rule to trigger pipeline upon commit

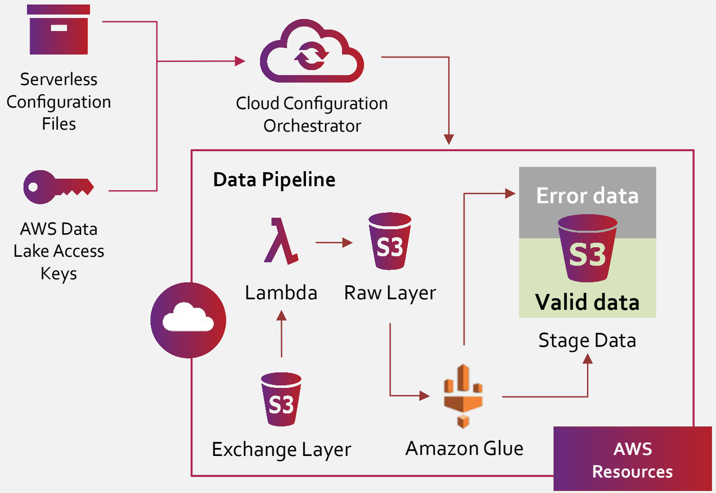
- DSPs benefit from interface control templates that track the essential integrations between different services.
- The template also needs to capture how the interactions between these services will be carried out and how they work in tandem with each other.
- Having ICT is critical to establish an event-driven orchestration in the data lake.

Interface control template reduces the rework in data lake implementation and accelerates the building time by 57%

Infrastructure as Code (IaC) to Accelerate Data Pipeline Building & Scaling

When architecture and integrations are vetted, cloud engineers need ways to rapidly build, orchestrate and manage data pipelines

A data pipeline encompasses how data travels from point A to Point B. It includes the entire process from collection, storing, refining, and analysis of data.



Limitations of using console

- Traditional console applications on cloud platforms are used to orchestrate and manage services.
- The effort and time required to add a new data pipeline are considerably high for the infrastructure team when using consoles.
- For making minor architectural changes, we often have to recreate entire pipelines in the console, which is a tedious task.

Recommendation: Serverless Infrastructure as Code

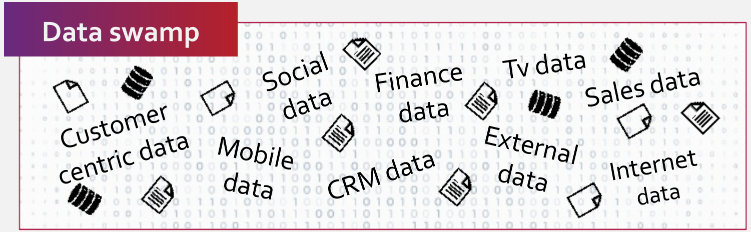
- **IAC:** Infrastructure as Code is the process of provisioning and managing the entire infrastructure (serverless applications) through a series of software using a cloud configuration orchestrator.
- **Cloud configuration orchestrator:** DSPs need to use a cloud configuration orchestrator to spawn new pipelines with all the necessary AWS resources. Resources can be effectively spawned by changing a few configuration files in the orchestrator.
- DSPs need to define the configurations of serverless applications as configuration files and with the help of a serverless framework, AWS resources can be provisioned & the new pipeline will also be deployed.

Using Infrastructure as Code (IaC) mechanism in a serverless data lake accelerates the scaling time for a new pipeline by 70%

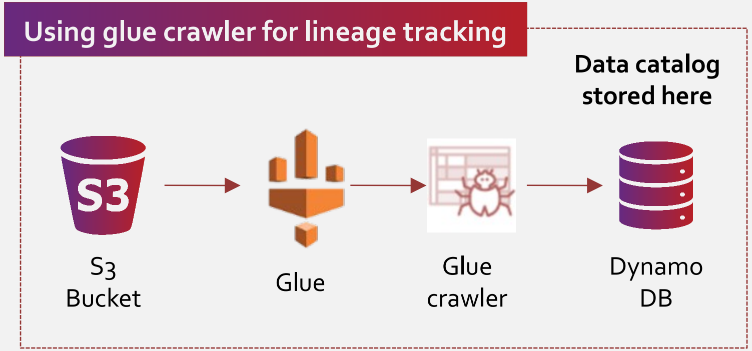
Leveraging Data Cataloging Approach to Avoid Data Swamps

Deployment of services should be followed by careful, well-managed data ingestion and management in the data lake

Data swamp



Using glue crawler for lineage tracking



How do data swamps arise?

- While building a data lake, DSPs generally migrate their entire data stream. Also, due to multiple handoffs and stakeholders, the data lake ends up having different hierarchy in storing data.
- Once the data lake is created, this ends up adding to the complexities of maintaining multiple levels of hierarchies and permission levels thus leading to a data swamp.

Recommendation: Data cataloging approach

To avoid data swamps, DSPs must follow these steps:

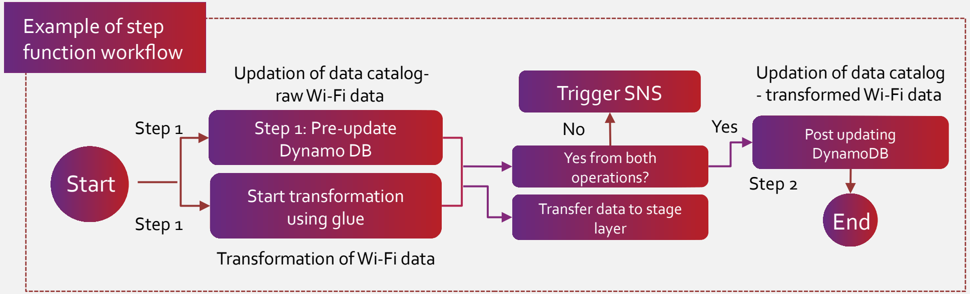
- **Metadata:** Creating metadata using a glue crawler.
- **Data Catalog:** Data catalog of where the data lies, and the path traveled by the data. This is used for lineage tracking.
- **Single hierarchy structure:** Ensure that a single hierarchy and the naming convention is used across the data sources to avoid data dumping.
- **Data Governance:** Proper identity and access management with stringent rules on access. Time-based data archiving rules on AWS glacier.

These recommendations help monitor data lakes long after the initial architecture and avoid data swamps and misuse of the data lake.

Having metadata and data lineage accelerates the time to integrate with BI tools and ML models by 60% which in turn improves the time to insights

Event-driven Orchestration to Automate Data Transfer and Transformation

Once the data lake is set up, it is highly recommended to automate the data transformations and aggregations in it to improve efficiency.



Example of step function workflow

Start

Step 1

Step 1

Updation of data catalog- raw Wi-Fi data

Step 1: Pre-update Dynamo DB

Start transformation using glue

Transformation of Wi-Fi data

Trigger SNS

No

Yes from both operations?

Transfer data to stage layer

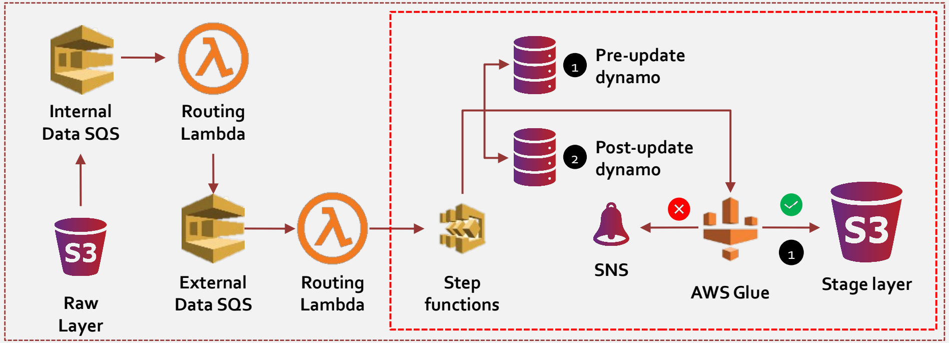
Yes

Updation of data catalog - transformed Wi-Fi data

Post updating DynamoDB

Step 2

End



Pitfalls

- Managing the complexity (transformation, orchestration, latency) of multiple pipelines is a key initial hurdle in managing data lakes.
- DSPs usually have multiple data pipelines and data sources for different use cases, which exacerbates this problem.
- **For example,** a batch processing pipeline coupled with a data arrival latency might be set up to begin without required data, resulting in poor veracity.

Recommendation: Event-driven orchestration

- **Step functions:** Event-driven orchestration using AWS Step Functions enable serverless queries & serverless polling - this helps in automating the end-to-end workflow of the data pipeline.
- Step functions help in orchestrating multiple ETL jobs with minimal intervention, based on the occurrence of specific events. Here, it orchestrates the sequence in which the jobs need to be run (1,1,2 in the figure).
- **Parquet data format:** It is highly recommended to transform the final data in the data lake to the parquet format. Parquet is a columnar store that accelerates the computation time and reduces the cost of storage as well as computing in the later stages.

Event-driven orchestration in a serverless data lake improves the total data processing time by 40%

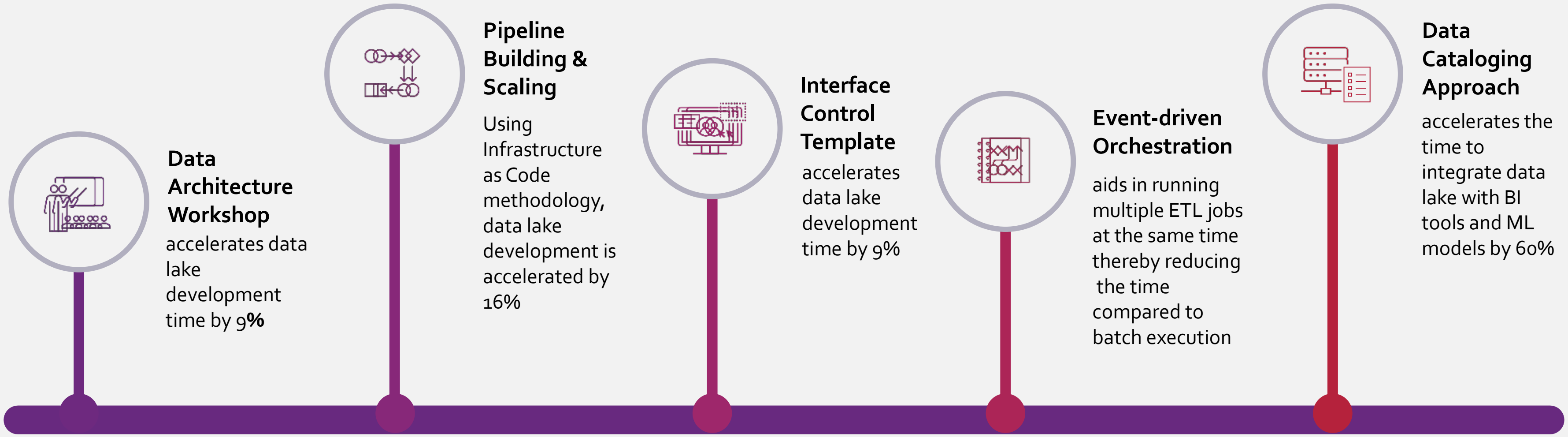
Results achieved by a leading DSP in Latin America by leveraging the methodology described in this insight

Implementing **the five-step methodology** provided in this insight

accelerated the serverless data lake building time by 33%

Benefits - Agility

Other benefits



Data Architecture Workshop

accelerates data lake development time by 9**%**

Pipeline Building & Scaling

Using Infrastructure as Code methodology, data lake development is accelerated by 16%

Interface Control Template

accelerates data lake development time by 9%

Event-driven Orchestration

aids in running multiple ETL jobs at the same time thereby reducing the time compared to batch execution

Data Cataloging Approach

accelerates the time to integrate data lake with BI tools and ML models by 60%

Get in touch

USA

Prodapt North America, Inc.
Oregon: 10260 SW Greenburg Road, Portland
Phone: +1 503 636 3737

Dallas: 1333, Corporate Dr., Suite 101, Irving
Phone: +1 972 201 9009

New York: 1 Bridge Street, Irvington
Phone: +1 646 403 8161

CANADA

Prodapt Canada, Inc.
Vancouver: 777, Hornby Street,
Suite 600, BC V6Z 1S4
Phone: +1 503 210 0107

PANAMA

Prodapt Panama, Inc.
Panama Pacifico: Suite No 206, Building 3815
Phone: +1 503 636 3737

UK

Prodapt (UK) Limited
Reading: Suite 277, 200 Brook Drive,
Green Park, RG2 6UB
Phone: +44 (0) 11 8900 1068

IRELAND

Prodapt Ireland Limited
Dublin: Suite 3, One earlsfort centre,
lower hatch street
Phone: +44 (0) 11 8900 1068

EUROPE

Prodapt Solutions Europe &
Prodapt Consulting B.V.
Rijswijk: De Bruyn Kopsstraat 14
Phone: +31 (0) 70 4140722

Prodapt Germany GmbH
Münich: Brienner Straße 12, 80333
Phone: +31 (0) 70 4140722

Prodapt Digital Solution LLC
Zagreb: Grand Centar,
Hektorovićeva ulica 2, 10 000

Prodapt Switzerland GmbH
Zurich: Muhlebachstrasse 54,
8008 Zürich

Prodapt Austria GmbH
Vienna: Karlsplatz 3/19 1010
Phone: +31 (0) 70 4140722

SOUTH AFRICA

Prodapt SA (Pty) Ltd.
Johannesburg: No. 3, 3rd Avenue,
Rivonia
Phone: +27 (0) 11 259 4000

INDIA

Prodapt Solutions Pvt. Ltd.
Chennai: Prince Infocity II, OMR
Phone: +91 44 4903 3000

"Chennai One" SEZ, Thoraipakkam
Phone: +91 44 4230 2300

IIT Madras Research Park II,
3rd floor, Kanagam Road, Taramani
Phone: +91 44 4903 3020

Bangalore: "CareerNet Campus"
2nd floor, No. 53, Devarabisana Halli,
Phone: +91 80 4655 7008

THANK YOU!

info@prodapt.com | www.prodapt.com

Prodapt powering global telecom