# Data Integration for Cloud Data Lakes: Architecture and Best Practices


PDG

Having the ability to make accurate data-driven business decisions in real-time is no longer a "nice to have" in today's fast-paced business environment. However, many organizations are hampered by data silos. The time it takes to collect all the data from disparate sources can lead to missed opportunities. Meanwhile, it's challenging to foster trust in data across an organization without a straightforward and transparent way to maintain data integrity and enforce data governance.

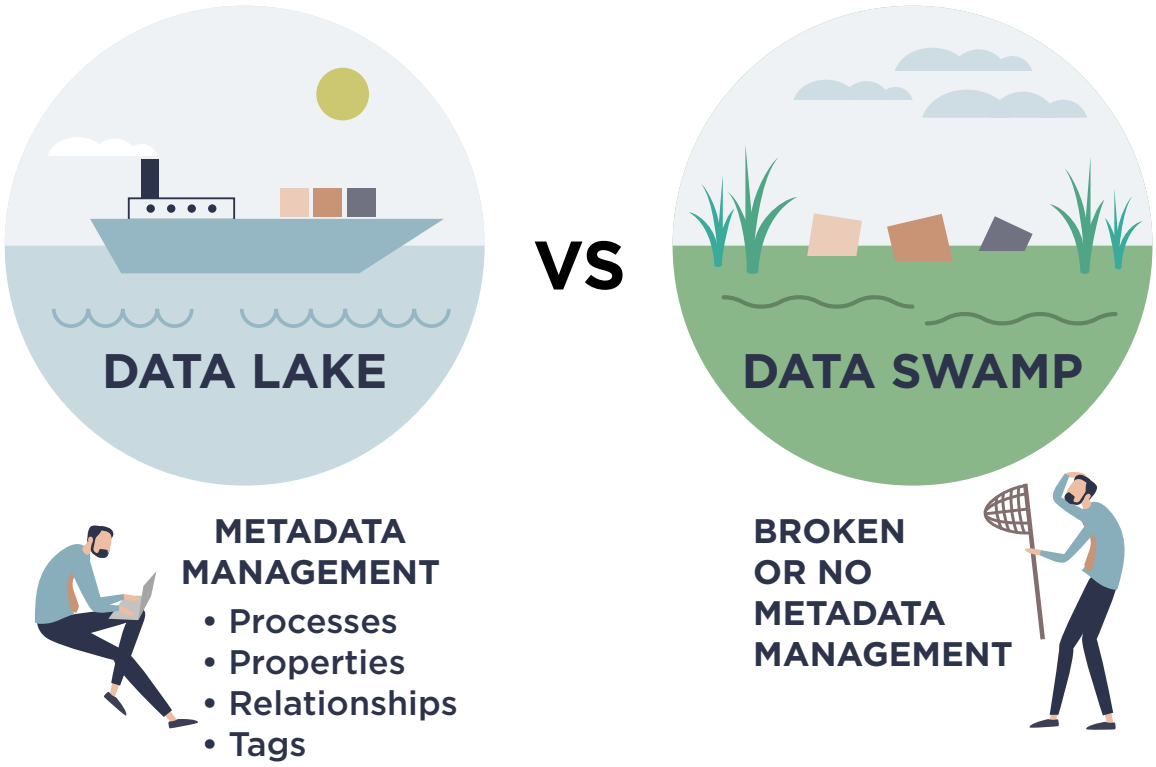
A cloud data lake provides a centralized repository accessible to everyone. This single source of truth allows companies to effectively leverage data in all business activities and open up new possibilities.

For example, we built a data lake for a client to support targeted TV advertising. The data lake collects customer viewership data, advertising campaign data, and customer targeting information, which are then used to estimate and optimize advertising campaigns' performance and reach. They also help the client understand consumer behaviors, predict results, and determine if existing ad campaigns are performing as expected.

The ability to leverage such a vast amount of information starts with having a centralized repository where data from disparate sources can be filtered, queried, analyzed, and put into action. Let's review the data lake architecture and best practices PDG Consulting applies to this and other projects to achieve impactful business outcomes.

## What Is a Data Lake?

A data lake is a centralized repository that stores structured, semi-structured, and unstructured data from many disparate sources in a raw and granular format. The data lake associates each piece of information with identifiers and metadata tags to make the data queryable and quickly retrievable. Compared with a data warehouse, which requires rigid data modeling and definitions upon ingestion, a data lake is more flexible because it can store different types of data in full fidelity.




*A data swamp is the result of poor data management and governance. It enforces little to no data organization or system.*

Data lakes can reside on-premises or in the cloud. On-prem data lakes may be required for specific use cases (e.g., regulatory compliance.) But most businesses can benefit from a cloud data lake solution that offers instant elasticity to accommodate a fast-growing amount of data while minimizing upfront investment for hardware and software.

## The Benefits of a Data Lake

A data lake provides organizations with several noteworthy benefits, including:

- **Flexibility and scalability:** A data lake allows organizations to ingest data from any source at any speed, whether it's coming from on-prem systems, the cloud, or the edge. It can also store any type or volume of data to be processed and analyzed using various applications later. For example, the architecture we used for the targeted TV advertising project, named ASSIST, collects data from an on-prem ad management system, plus customer profile and program metadata from different cloud systems. The data is then run through the EMR-Spark capacity calculator, impression predictor, and viewership predictor to generate insights.
- **All data in one place:** A centralized repository means that decision-makers no longer have to go to different systems and sources to gather information. They can access the data they need to perform detailed analyses and understand big-picture business impacts all in one place. Since everyone's action is informed by a single source of truth, teams don't have to second-guess the foundation on which they make their decisions.
- **Reports and analytics:** A data lake is essentially a one-stop shop for data analytics teams to perform their functions (e.g., identifying data trends to improve business performance.) The ability to use metrics from across the organization helps streamline reporting at all levels to generate timely insights and ensure that everyone is rowing in the same direction.

# Cloud Data Lake Architecture: Design Principles and Best Practices

Data lake architecture comes in many flavors, and there's no one "right" way to build a data lake. That's why our team takes the time to understand each client's business objectives and data requirements.

Here are the key principles and best practices we applied to the ASSIST targeted TV advertising project mentioned above:

## Use Snappy-Compressed Parquet File Format

Snappy allows for fast data compression and decompression. It helps save storage space and cost without impacting performance, a criterion for business intelligence (BI) users. Meanwhile, the parquet format is suitable for handling complex data in large volumes, thanks to its performant data compression and ability to process various encoding types. The binary format optimizes data storage and enables queries of non-relational data for added flexibility.

We used different parquet files/buckets to store targeting profiles, campaigns, ad breaks, impressions, viewership, channel mapping, and program schedules in the ASSIST targeted TV advertising project. The structure helped organize the data and facilitate queries.

## Define Schemas Explicitly

A schema is a plan that defines how data in your data lake is organized to ensure efficient day-to-day operations, data integrity, and consistency. A well-documented schema gives you the foundation for implementing access control, ensuring data security, enhancing internal communication, and monitoring regulatory compliance.

## Set a Retention Policy

For the ASSIST targeted TV advertising project, we used Amazon S3 buckets for object storage, thanks to their ability to store both structured and unstructured information. Like many other applications, we only need to retain data in these buckets for a finite amount of time. A data retention policy allows the system to delete information that's no longer needed to free up storage space for new data, so the client can lower storage costs and increase computing speed.

## Leverage Cloud-Based Managed Services

We used cloud-based managed services—AWS Glue ETL as a service, Glue Data Catalog for persistent metadata storage, Amazon Athena to provide interactive query service, and Amazon EMR to run distributed computing jobs. Since the cloud provider manages these services, we don't have to reinvent the wheel or worry about maintenance. Our team can focus on value-add components instead of platform maintenance, to shorten time to value.

## Partition the Data

Any big data project comes with the need to process high volumes of data. If we have to query all the data, it would take a very long time to find an answer. Partitioning the data helps improve scalability, reduce contention, and optimize performance. For example, we used a time-based partition for the targeted TV advertising project. You can also set other conditions to divide data by usage pattern (e.g., archive data not accessed frequently into cheaper data storage).

## Use Recommended File Sizes

An algorithm has to go through many files to find an answer for a query. File size and the number of files impact how quickly the system can return an answer for a query. To balance the performance trade-off, we need to find the sweet spot. With Athena, the recommended file size is 128MB. We optimize our files to achieve the best performance.

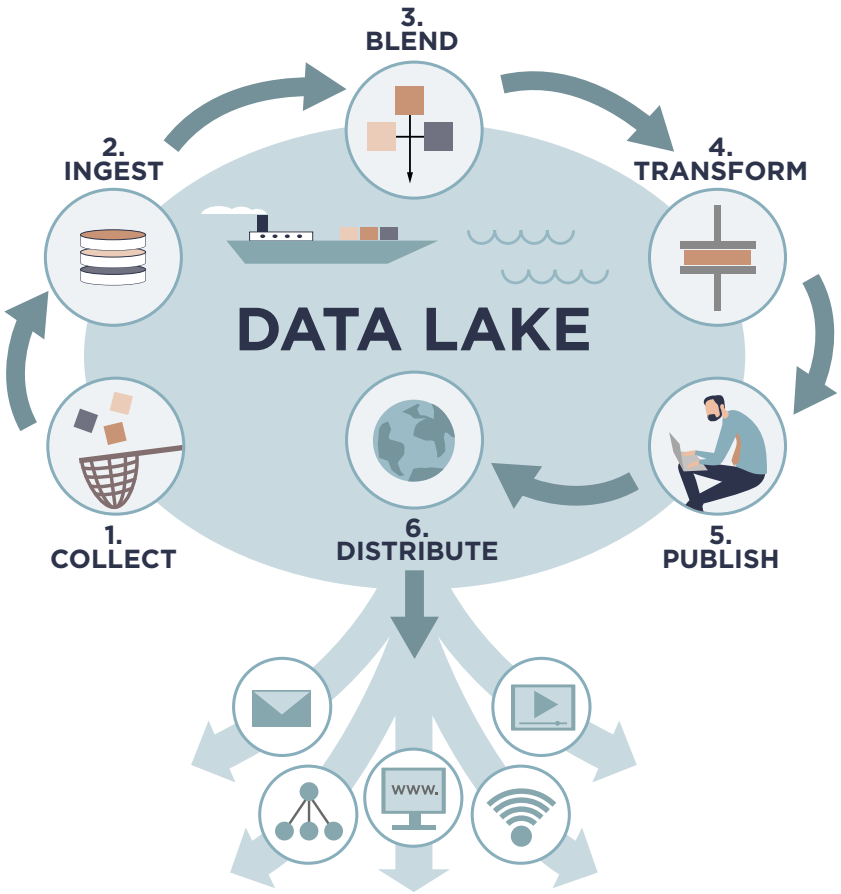
## Building Cloud Data Lakes: A Client-Centric Process

While we adhere to certain basic principles and best practices when designing a solution, we recognize that every client and organization has unique requirements.

We start each project by assessing existing systems and understanding desired business outcomes. Then, we develop data analytics and business intelligence tools to create tailored solutions, often with a data lake as the foundation to support a comprehensive data strategy.

Although it's possible to move from one cloud-based service to another, the process can be time-consuming and costly. Therefore, we take the time to choose the right technologies based on a sound strategy and our team's extensive expertise to build a data lake to meet a client's needs for years to come.

Our engineers then take stock of the data sources and design an ETL (extract, transform, load) path into the data lake. Our team will choose the appropriate data lake ETL tools to support evolving schema-on-read and optimize object storage for optimal query performance. We set up integration with metadata catalogs to make the data queryable by different services and build the capacity to update tables over time.




*ETL is a data integration process that combines and refines data from multiple sources, then delivers the data to a destination or target system.*

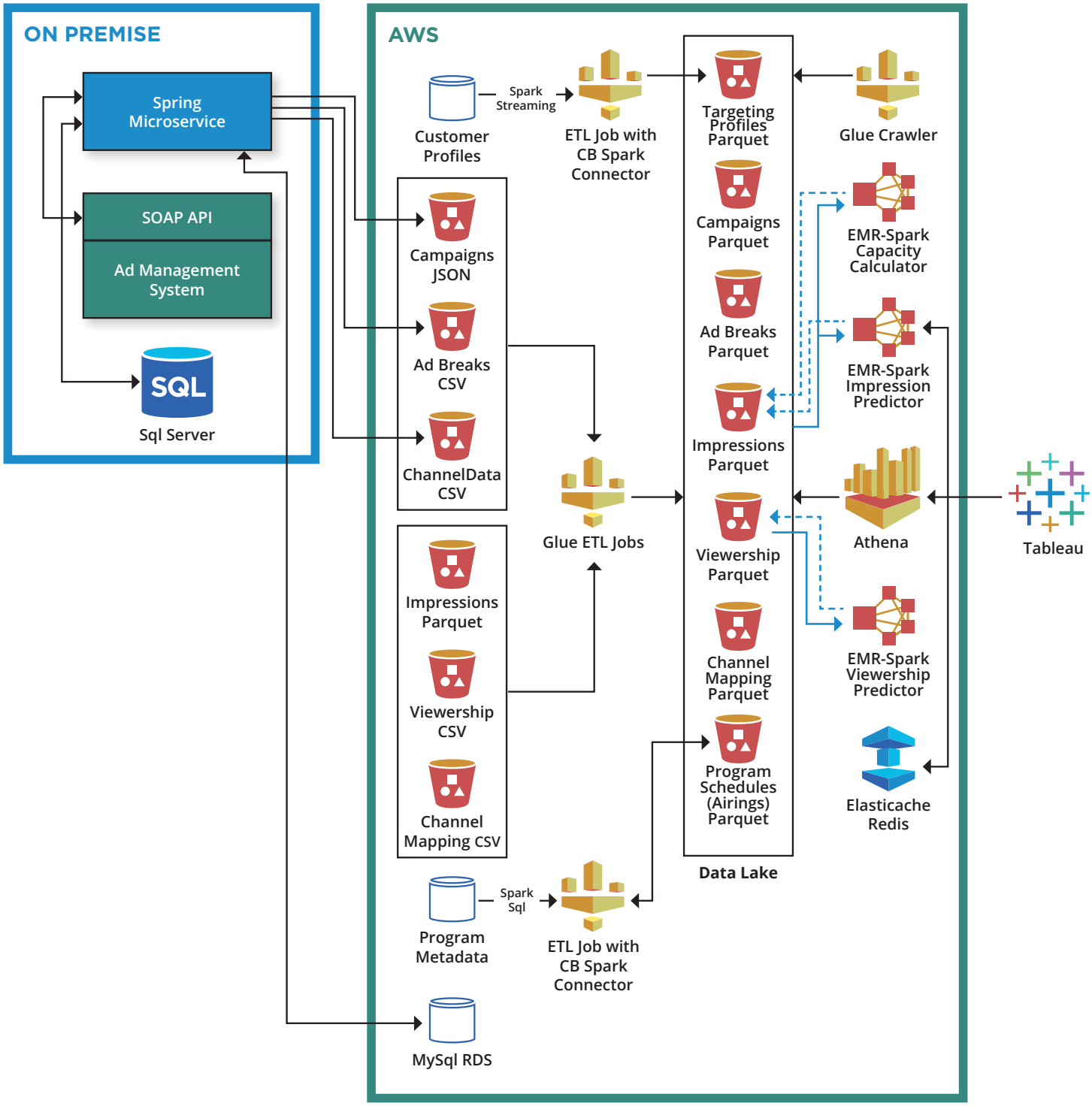
We also help our clients select scalable technologies, such as the right cloud-based managed services to optimize performance and flexibility. We ensure data security and data integrity by encrypting all data at rest and in transit. For example, all personally identifiable information (PII) data is hashed to comply with various data privacy regulations.

## Data Lake: Example of a PDG Big Data Project Leveraging a Data Lake

For the ASSIST targeted TV advertising project, we created an actionable, centralized data lake for all things ad-related. It's main value resides in providing predictive algorithms that leverage big data and Spark distributed computing, which lets users visualize the predicted performance of upcoming advertising campaigns.
This Business Intelligence tool gives our client's advertising business stakeholders and Tech Ops teams high-value dashboards that let users drill down and slice and dice across huge volumes of historical ad performance and customer behavior events.

**See the next page for a diagram illustrating PDG's big data architecture for the ASSIST targeted TV advertising project.**

# ASSIST Architecture




*The above diagram illustrates PDG's big data architecture for the ASSIST targeted TV advertising project.*

## Data Lake: The Foundation of Your Data Story

The advances in cutting-edge, scalable, and cloud-native big data technologies enable organizations to undergo data-led transformations, leveraging real-time insights to make accurate data-driven decisions.

Data lakes offer a solid foundation for the value realization process by collecting, managing, and offering near-real-time availability of structured and unstructured data. We also help our clients tell their data stories to the right people with various analytics, BI, and reporting solutions to achieve maximum business value and capture growth opportunities.

PDG has expertise in various big data and business intelligence technologies—including Apache Spark, Hadoop, Kafka, Amazon Glue, EMR, Kinesis, Firehose, Athena, Microsoft SQL Server Integration Services and Analysis Services, Hive, Presto, NoSQL databases, and Elasticsearch. We can help you implement a tech stack based on your data strategy and business objectives to maximize your ROI.

**Ready to have your data work harder for you?**
**Get in touch to see how we can help.**

www.pdgc.com | sales@pdgc.com | +1 (323) 347-4640

## ABOUT PDG

PDG is a team of innovators and technologists delivering large-scale software application development services, end-to-end big data solutions, and SAAS tools for Fortune 500 and mid-size clients. We specialize in enterprise software development, business intelligence, and digital transformation—with expertise in translating complex business requirements into application design. Customers like WarnerMedia, Sony Pictures, Apple, DIRECTV, CBS, 21st Century Fox, and Amazon trust PDG to increase productivity, drive business growth, and maximize opportunities.


PDG Consulting | sales@pdgc.com | +1 (323) 347-4640

714 S. Hill Street, M1 Los Angeles, CA 90014