accenture

ACCELERATING YOUR DATA JOURNEY

The SAP Data Lake Accelerator from Accenture and AWS

Reaching safe harbor with your SAP data lake

Most SAP enterprise customers are looking for ways to manage multiple data types coming from a wide variety of sources. Faced with massive volumes and heterogeneous types of data, organizations are finding that, to deliver insights in a timely manner, they need a data storage and analytics solution in the cloud that offers more agility and flexibility than traditional data management systems.

An enterprise-wide data lake is a new and increasingly popular way to store and analyze data that addresses many data challenges. A data lake enables you to store all of your structured and unstructured data in one, centralized repository in the cloud. Since data can be stored as-is, there is no need to convert it to a predefined schema and you no longer need to know what questions you want to ask of your data beforehand.

It sounds straightforward, but of course it's not. An SAP data lake implementation must be planned and executed properly, with the right methodology, technical architecture and accelerators.

To provide those essentials to companies, Accenture and AWS have created an SAP Data Lake Accelerator that is based on leading practices and our joint experience, and which offers faster deployment at a lower cost.

Where are you on your data voyage?

Many companies are stuck on shore when it comes to data. To reach their desired business destination, they need to be able to analyze vast amounts of data across silos, sometimes in real time, to fuel growth and product innovation, and deliver world-class customer service.

In an SAP environment, data is sometimes trapped in SAP and many processes need to be created to get it out. The data structure may be fractured, resulting in differing points of view on reporting requirements across functional domains.

Data warehouses have gone a long way toward solving the problem of fragmented data. Yet, data warehouses are primarily geared toward managing pre-defined, structured, relational data for existing analytics or common use cases—data that is processed and ready to be queried.

Today, leading edge data warehouses, which are leveraging public cloud services like Amazon Redshift, are being enhanced with the addition of a data lake leveraging other AWS services. In fact, a typical organization will require a data warehouse as well as a data lake closely connected to it, because each serves different needs and use cases. With a data lake, companies can apply analytics to new data sources such as log files, click-steams, social media and internet connected device without the need to structure the data.

On-premises data lakes present several challenges, including maintenance costs and scalability restrictions. Building your data lake in the cloud enables you to lower your engineering costs and focus on business value by using managed services that help you scale as needed, leveraging automatically updated technologies with high reliability and availability.

The importance of data lakes

Based on Accenture and AWS experience, a data lake should support the following capabilities:

- Collecting and storing any type of data, at any scale and at low cost

- Securing and protecting all data stored in the central repository

- Searching for and finding relevant structured and unstructured data

- Quickly and easily performing new types of data analysis on data sets

- Querying the data by defining the data's structure at the time of use

Data lakes with these capabilities will become increasingly essential. According to research and consulting firm MarketsandMarkets™, the global market for data lakes is expected to grow from $7.9 billion in 2019 to $20.1 billion by 2024, at a compound annual growth rate (CAGR) of 20.6 percent. This growth reflects the need companies have to extract in-depth insights from multiple systems and siloed functions.[1]

1. Marketsandmarkets, Data Lake Market, 2019.
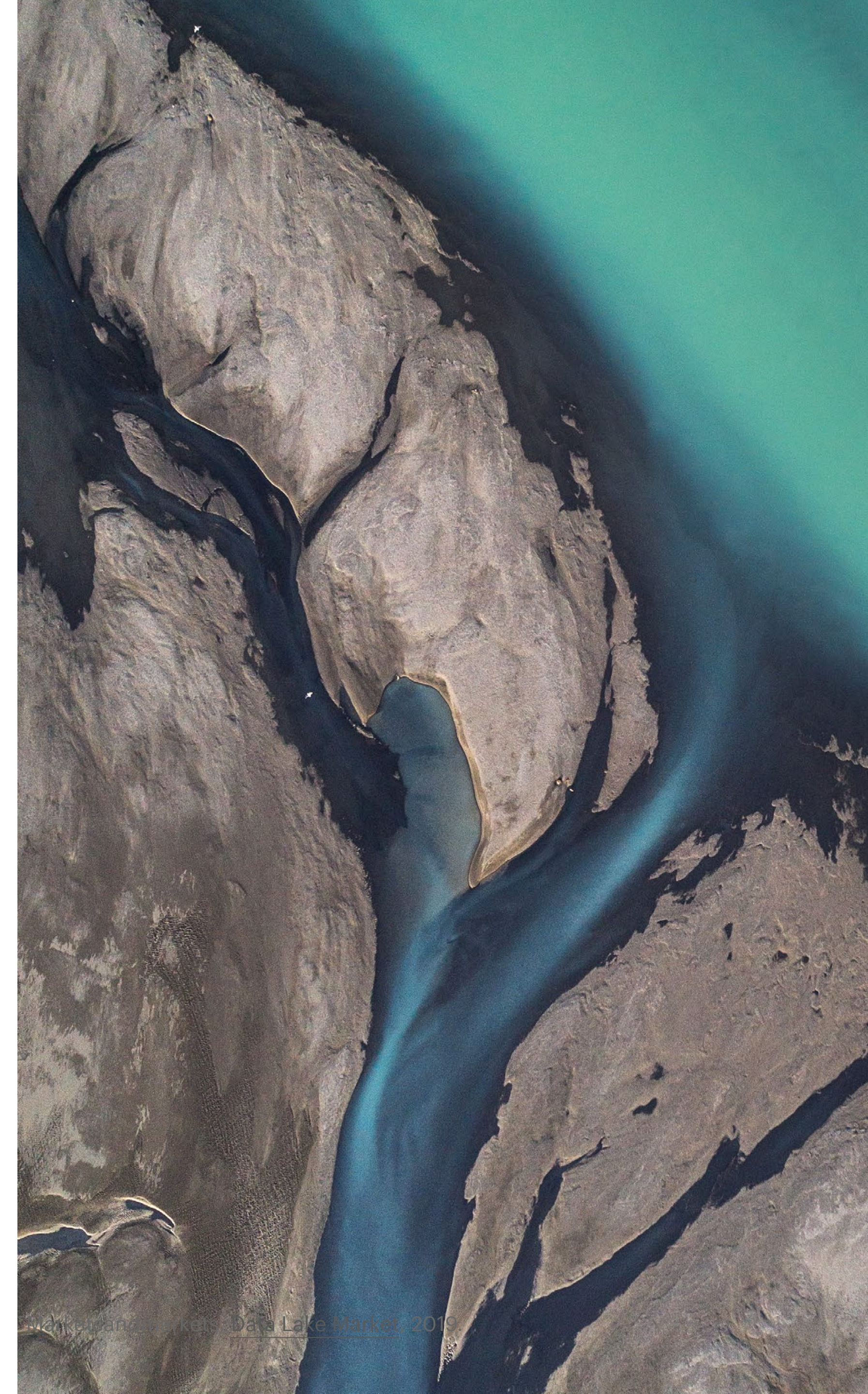
# Features of the SAP Data Lake Accelerator from Accenture and AWS

The SAP Data Lake Accelerator helps minimize time and effort required to set up, configure and maintain a data lake by providing a variety of data ingestion and basic curation methods in a well-defined zonal and secure architecture. This facilitates fully automated movement of data using AWS native services.

Once the data lake is established, it can be reported on using a wide variety of reporting tools as an open architecture—for example, Tableau, Quicksight and Power BI. This means that you can continue to use your preferred tools, or select a new one, and maintain continuity in reporting. In addition, machine learning and advanced analytics can be used to obtain new insights from your data.

Data that is needed only for reporting, not transactional use, can be archived out of the source systems to reduce their database footprint. This would be of particular interest in HANA-based systems because they store data in memory, and archiving helps reduce costs by downsizing the source systems.
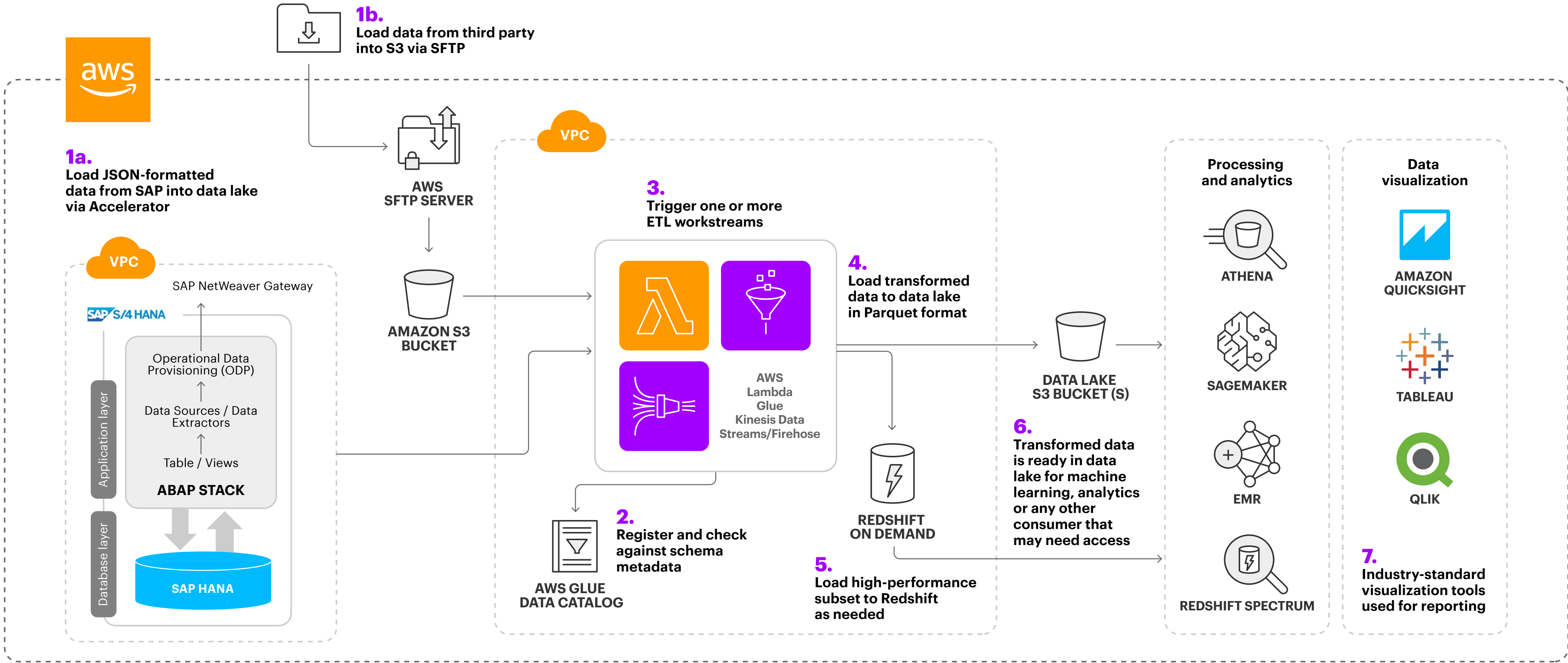
The accelerator includes:

- **Automated analysis of SAP source systems and automated set-up of the data extraction:** The accelerator helps eliminate the need for lengthy analysis cycles and the need to rely merely on people knowledge or outdated documentation. Moreover, the accelerator helps automatically configure the SAP extraction and shares the configuration with the AWS data lake for automated setup of the extraction jobs. The value here is in potentially being able to drive business change very quickly in an agile fashion for a low cost of entry. You can unlock the power of your data faster possibly without spending millions of dollars.

- **Automated generation of SAP data extraction:** Once the mapping to the necessary BW extractors is done, the accelerator automatically generates an oData provider for each extractor and publishes it on the SAP Gateway server.

- **Automated zone creation based on best practices:** At the data level, the accelerator contains strong governance for folder structures. It creates four zones (landing, raw, curated, and transformed & provisioned), and runs scripts that create associated folder structures automatically in minutes. Automation is essential to a data lake that is scalable and sustainable.

- **Parameterized framework:** Users can create one job or task, and then parameters are passed into them allowing for code reuse and orchestration based on configurations rather than having to write code from scratch.

- **Automated file movement:** When creating a brand-new data source, one must be able to trust the data that comes in from the raw zone. Any changes must be noted as such in a new report. Automated file movement helps eliminate the chance of unauthorized changes in raw data.

- **Scheduled data crawling:** The accelerator creates a metadata catalog and self-service so there is a presentation layer for the customers of the datasets. As soon as an event happens, the data starts collecting the metadata immediately rather than days later.
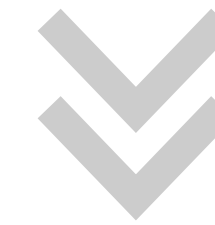
- **Reduced deployment lifecycle:** The focus here is on continuous improvement and continuous deployment using small teams (e.g., a couple of teams with three to six people each) managing at scale, or even into hyperscale. A comparable deployment using other methods might require well over 100 people.

- **File-level quality controls:** With the accelerator, the risk of accepting a partial or corrupted file, or an incomplete database pull, is mitigated. Falsified records cannot damage the data environment because the accelerator performs automatic file verification and check. If the check sums don't match, the file is rejected.

- **Stable reference architecture:** The accelerator uses a reference architecture that is scalable, stable, extensible and market-agnostic, and is also built with industry leading practices. (See Figure 1.)

Figure 1: SAP Data Lake Accelerator—Reference Architecture

1b.
Load data from third party into S3 via SFTP

aws

AWS SFTP SERVER

VPC

1a.
Load JSON-formatted data from SAP into data lake via Accelerator

3.
Trigger one or more ETL workstreams

Processing and analytics

Data visualization

VPC

SAP NetWeaver Gateway

SAP S/4 HANA

Operational Data Provisioning (ODP)

Data Sources / Data Extractors

Table / Views

ABAP STACK

Application layer

Database layer

SAP HANA

AMAZON S3 BUCKET

4.
Load transformed data to data lake in Parquet format

AWS
Lambda
Glue
Kinesis Data
Streams/Firehose

ATHENA

AMAZON QUICKSIGHT

SAGEMAKER

TABLEAU

DATA LAKE S3 BUCKET (S)

6.
Transformed data is ready in data lake for machine learning, analytics or any other consumer that may need access

EMR

QLIK

2.
Register and check against schema metadata

REDSHIFT ON DEMAND

REDSHIFT SPECTRUM

AWS GLUE DATA CATALOG

5.
Load high-performance subset to Redshift as needed

7.
Industry-standard visualization tools used for reporting

Benefits of the SAP Data Lake Accelerator

The accelerator leverages extensive knowledge and assets from Accenture and AWS to help companies create and manage their data lakes effectively:

- **Proven configurations:** Setting up and configuring a data lake can be time consuming. It is difficult to build and extend complicated curation logic. Because the data lake has no native data quality service, basic data checks are required to ensure that data is trustworthy. There is also no native lineage service, so tracing the lineage of data is a challenge.

- **Accepted reference architecture:** Data lakes are new, so there is not yet a market-agreed reference architecture —indeed, there are thousands of permutations. Companies need a proven, out-of-the-box architecture to speed the creation of their data lake.

- **Deep skills:** Broad SAP aptitude is required, including functional, business warehouse, and interface development skills. An SAP interdisciplinary team must know how to work in a complementary way with AWS technical architects. AWS has many services available that comprise a data lake solution, and the knowledge within the accelerator helps companies architect and integrate those services.

- **Extensive experience:** Deep experience is needed to expedite a data lake implementation—across analysis, design, build and deploy. Many challenging questions are involved: What are the pros and cons of different extraction mechanisms? What data is relevant in the source systems?

Accelerating the journey to business value

The SAP Data Lake Accelerator can deliver multiple benefits before and during a data lake implementation in an SAP environment. Because the accelerator facilitates rapid buildout of a data lake, companies can improve speed to market as well as time to insight. They can also embed more flexible architectures into the data lake.

At the scaling phase, the data lake supports better data-led decision making because the data is more trustworthy and because the data lake supports on-demand data consumption while also providing strong governance of data access and delivery. Faster innovation is made possible by enabling ready analysis of many varied forms of data. The accelerator helps companies drive business change quickly in an agile fashion for a low cost of entry.

Get in touch

For more information about the SAP Data Lake Accelerator from Accenture and AWS, please write to:

SAPDataLakeAWS@accenture.com

About Accenture

Accenture is a leading global professional services company, providing a broad range of services in strategy and consulting, interactive, technology and operations, with digital capabilities across all of these services. We combine unmatched experience and specialized capabilities across more than 40 industries—powered by the world's largest network of Advanced Technology and Intelligent Operations centers. With 513,000 people serving clients in more than 120 countries, Accenture brings continuous innovation to help clients improve their performance and create lasting value across their enterprises. Visit us at www.accenture.com.

About Amazon Web Services

For 14 years, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud platform. AWS offers over 175 fully featured services for compute, storage, databases, networking, analytics, robotics, machine learning and artificial intelligence (AI), Internet of Things (IoT), mobile, security, hybrid, virtual and augmented reality (VR and AR), media, and application development, deployment, and management from 76 Availability Zones (AZs) within 24 geographic regions, with announced plans for nine more Availability Zones and three more AWS Regions in Indonesia, Japan, and Spain. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—trust AWS to power their infrastructure, become more agile, and lower costs. To learn more about AWS, visit aws.amazon.com.

Disclaimer

This document is intended for general informational purposes only and does not take into account the reader's specific circumstances, and may not reflect the most current developments. Accenture disclaims, to the fullest extent permitted by applicable law, any and all liability for the accuracy and completeness of the information in this presentation and for any acts or omissions made based on such information. Accenture does not provide legal, regulatory, audit, or tax advice. Readers are responsible for obtaining such advice from their own legal counsel or other licensed professionals. This document may contain descriptive references to trademarks that may be owned by others. The use of such trademarks herein is not an assertion of ownership of such trademarks by Accenture and is not intended to represent or imply the existence of an association between Accenture and the lawful owners of such trademarks.