Brochure

HPE Storage

Leverage a modern data lake to build a cloud-native analytics data pipeline

Use case example: Financial data analysis and processing for fraud detection

HPE GreenLake

Data analytics pipeline challenges

Data analytics and artificial intelligence (AI) have become a general-purpose technology. Enterprises of all sizes and industries use machine learning (ML) and data analytic techniques to improve their products and services, streamline processes, reduce costs, and create a competitive edge. To deploy data analytics and AI/ML at full scale, enterprises must solve two conflicting issues: running multiple parallel AI projects for different business units and centralizing as much as possible on a common storage infrastructure.

Infrastructure plays a central role in the success of AI initiatives. According to IDC, inadequate infrastructure or a lack of purpose-built infrastructure capabilities is often the cause of project failures.[1] Hewlett Packard Enterprise has implemented a modular end-to-end, full-stack data pipeline that covers all phases of the AI and analytical workloads to address infrastructure challenges.

The foundation is a cloud-native platform based on microservices that provides not only the flexibility to run multiple heterogeneous workloads in parallel but also the ability to spin up new analytical tools as needed. The platform leverages HPE Server, Storage, and Networking solutions, along with an integrated but flexible software stack, to build dynamic, scalable, and performant data pipelines based on customer requirements.

This data pipeline leverages a modern data lake based on the HPE Storage object store (Storage Solutions for Scality) running on HPE Alletra Storage Servers 4120. The choice of an object store aims to solve the limitations of legacy Hadoop and distributed file systems environments.

This technical brief describes the first part of a multi-phase project. It focuses on the advantages of using object storage solutions to create end-to-end data pipelines supporting heterogeneous analytics workloads. A financial use case has been implemented to analyze and process financial data for fraud detection.

Architecture overview

The underlying architecture is based on an end-to-end data pipeline framework, which consists of data ingestion (Apache Airbyte), data processing (Apache Spark and AI/ML libraries), data store (a NoSQL DB), data lake (HPE Solutions for Scality), and data visualization layers to manage the data journey for enabling data-driven insights and actions. The end-to-end data pipeline is deployed on a Kubernetes cluster running on Red Hat® OpenShift Container Platform 4.10, orchestrated and automated with Helm.

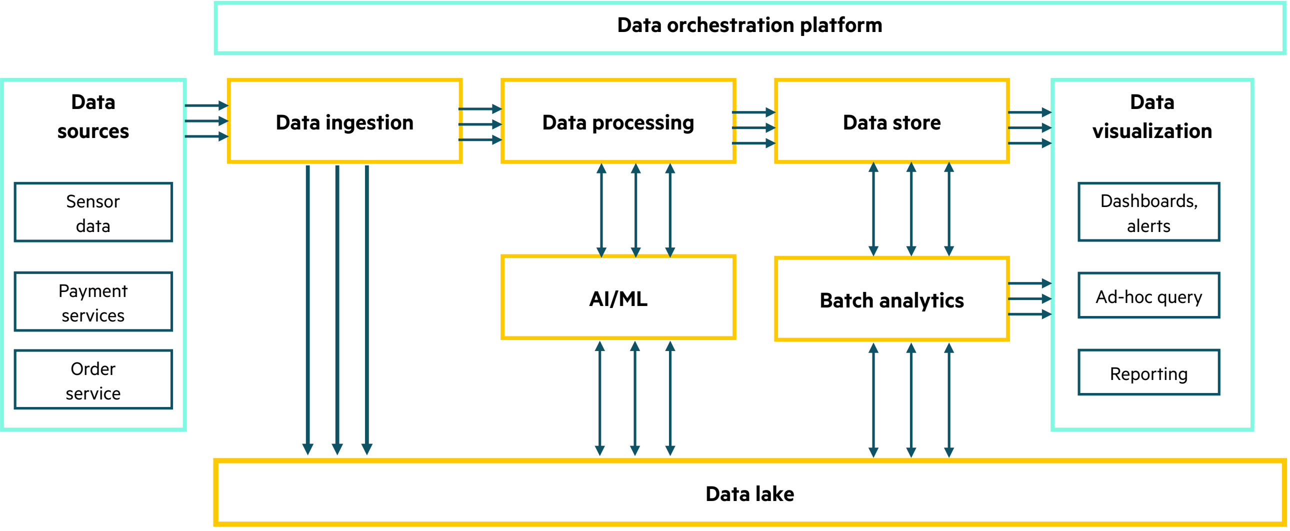

Figure 1. Cloud-native end-to-end data pipeline based on microservices

[1] AI InfrastructureView 2022: Premium

Each functional layer is a module that can be used individually and scaled out, up, or down independently. The pipeline can be deployed as a whole, or it can be plugged into existing architecture.

Components	Description
Data ingestion	With an extensive catalog of connectors (more than 300 connectors for data sources and destinations) and support for batch and streaming modes, Apache Airbyte is the basis for the data ingestion layer. It transfers data from their sources to the data lake to run batch analytics or from the data processing layer to run real-time analytics.
Data store	The data store layer serves as a repository for the data collected from the data ingestion layer. It is based on a NoSQL database or a graph DB, depending on the use case, and serves as the persistent layer for batch or real-time processing.
Data lake	Based on HPE Solutions for Scality and serving all data pipeline layers, the data lake is the centralized storage repository that holds the raw data in its native format and the elaborated data in its format, such as Parquet.
Data processing • **AI/ML libraries** • **Batch analytics**	The data processing layer is based on Apache Spark 3.3.2 and is used to process the data in real time or batch mode before passing it to the AI/ML frameworks or keeping it in the data store. The use of Spark allows the leverage of GPUs to accelerate data processing and ML tasks. Indeed, Spark enables you to schedule and allocate GPUs as a resource for Spark applications, along with the RAPIDS Accelerator for Apache Spark. This plug-in overrides the physical plan of a Spark job by supporting GPU operations, enabling Spark users to run their existing SQL and DataFrame workloads on GPUs without any code changes and to achieve significant speedups and cost savings.

Use case description

The pipeline was built for fraud analytics services that must analyze and process multiple sources and different data types (such as POS data, credit card information, location data, or merchant data) in real time to identify fraud activities immediately and batch mode to analyze and aggregate data from various sources.

Data is ingested and processed in real time by AI models. After the processing, the data and its analysis results are loaded into the data lake in both its native format (schema-on-read) and a retrieval-optimized file format such as Parquet to facilitate later queries or to store results.

Figure 2 shows the use case data flow, the role of each data pipeline module, and the data formats used.

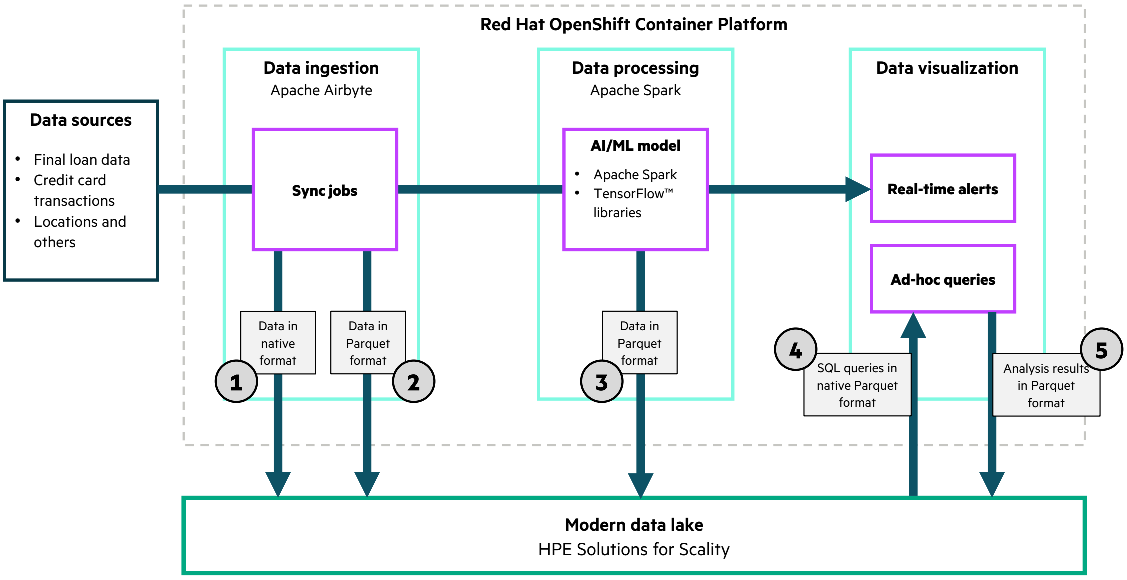

Figure 2. Data flow and processing

Apache Airbyte is part of the data ingestion layer and extracts data from the following sources:

- Point of sales (POS) data → File-based interface
- Location information → API-based connector
- Customer's information (including buying profile) → DB connector
- Merchant's information → DB connector

The processing data flow, which follows the extract, load, and transform (ELT) paradigm, includes the following steps:

1. Apache Airbyte uses the S3 interface as a destination to sink the extracted data into the data lake in the raw format.
2. Airbyte transforms the data from the different sources into a complex data structure loaded into the data lake as an S3 object in both Parquet and JSON format.
3. Apache Airbyte ingests the data into the AI model running on Apache Spark 3.3.2 in real time. The results of the AI model are displayed as alerts and are loaded into the data lake as S3 objects with JSON and Parquet format.
4. Fraud analysts use Apache Spark to analyze data in S3 using standard SQL.
5. Analysis results and new data structure are loaded back into the data lake using S3 and Parquet.

Key takeaways

A data flow of financial data for a fraud analytics service might be deployed as a workload example on top of the cloud-native data pipeline. Scality object store based on the HPE Alletra Storage Server 4120 has been used as a central repository along the pipeline. The server gives the data an optimal balance of economic capacity and performance required to build a modern data lake and manage huge amounts of unstructured data. Apache Airbyte was used to move the data into the central data lake. The ELT paradigm was used because the source datasets were unstructured and large by volume.

About a legacy Hadoop-based architecture, our tests demonstrated that object storage provides multiple advantages, such as:

- **Facilitation of data set preparation for AI models.** The native support for semi-structured and unstructured data in object storage is ideal for data lake use cases. Data from IoT devices, websites, mobile apps, social media, and other sources can be stored durably and securely in its native format, avoiding expensive ETL processes upfront. Schema-on-read is supported, applying schema only when the data is being analyzed, making it easier to handle heterogeneous and high-velocity data streams.
- **Flexibility of consuming the right level of computing resources.** The data lake can act as a centralized repository where any analytics engine can query the data as needed, which provides flexibility to use the most appropriate computing options, such as Big Data engines or serverless data analysis options. New data sources can be added, and analytics workloads can be changed without altering the underlying storage, accelerating analytics initiatives and time to insight.

- **Enhancement of security, privacy, and regulatory compliance.** Data lakes based on object storage provide comprehensive data governance capabilities. Features such as object locking, versioning, and cross-region replication help comply with security, privacy, and regulatory requirements. Granular access controls help ensure that data assets are secured while enabling self-service access.
- **Easy portability to/from cloud environments.** The self-contained data objects, RESTful APIs, scale-out architecture, and network-based access make object storage ideally suited for seamless data migration to and from the cloud. This portability accelerates hybrid cloud adoption and flexible data management.
- **Unlimited scalability within a unified namespace.** Object storage offers unlimited scalability to handle any data volume and can scale seamlessly up to dozens of petabyte-scale in a single namespace. This scalability removes concerns about running out of storage space as data grows. Teams can ingest data first and figure out processing later. Data lakes built on object storage are future-proofed for ever-growing enterprise data.

Conclusion

With exponential data growth in recent years, organizations need efficient and scalable ways to store and analyze large volumes of structured and unstructured data. This need has led to the increased adoption of data lakes and object storage solutions to replace traditional storage systems and Hadoop-based architecture.

Using object storage systems as the foundation for building a data lake provides several benefits over traditional data management approaches. Indeed, object storage delivers efficiency, agility, limitless scalability, and enterprise-grade governance capabilities required for modern data lakes. Building analytics data pipelines with object storage as the data lake foundation accelerates the ability to derive business value from massive datasets. It provides a future-proof, cost-effective platform for advanced analytics initiatives. The separation of storage and compute enables flexibility, and the native handling of semi-structured data makes ingestion seamless. Object storage satisfies the storage requirements for complex cloud-native analytics pipelines and builds the foundation for a modern data lake.

Visit **HPE GreenLake**

Chat now (sales)

a50009546ENW