SnapLogic eXtreme - For Data Lake and Cloud Data Warehouses

Summary

Half of all data lake projects fail because organizations don't have enough skilled resources who can code in Java, Scala, Python and do processing on large amounts of data. Additionally, skilled data engineers need a lot of time and effort to operationalize workloads that prepare, cleanse, aggregate data to cloud data warehouses for analytics and insights. SnapLogic eXtreme helps organizations get to insights quickly by significantly reducing the time needed to build and operationalize big data workloads by enabling even the citizen integrators to efficiently build and support petabyte-scale data integrations.

Big data integration at your fingertips

Pockets of data, buried in the enterprise, go unexplored due to the complexity of connecting large amounts of structured and unstructured data. Many companies that have "lifted and shifted" their big data environments to the cloud are still waiting for the promised benefits. Rather than just shifting to a cloud data lake environment as the volume, importance, and demands on the use of data grow, many companies are also moving to a managed big data framework to more fully realize the benefits of reduced CapEx and OpEx. SnapLogic eXtreme extends the accessible and easy-to-use SnapLogic Intelligent Integration Platform (IIP) to build and submit powerful Spark-based pipelines to Big Data as a Service (BDaaS) providers, such as Amazon EMR. SnapLogic eXtreme thus lowers the overall cost of operations while increasing data governance, SLA, and manageability (higher ROI).

Scalable, cost-effective, and fast

With SnapLogic eXtreme, you can realize the benefits of no CapEx, lower your OpEx by decreasing cluster management time, and reduce the need for hard-to-find skills to manage big data clusters. Also, as the AWS infrastructure changes, we will support major updates. SnapLogic eXtreme is built to support the processing of very high data volumes (peta/exabyte) with high complexity. It also has built-in support for ephemeral, managed big data clusters.

Democratize enterprise data using a visual paradigm for Spark development

With SnapLogic eXtreme, you can accelerate the development of your enterprise's modern data lake, break down silos, simplify the nurturing (cleanse, sample, summarize, and aggregate) of the data lake, and power your team to unearth new insights. SnapLogic visual

Enabling Modern Data Journey
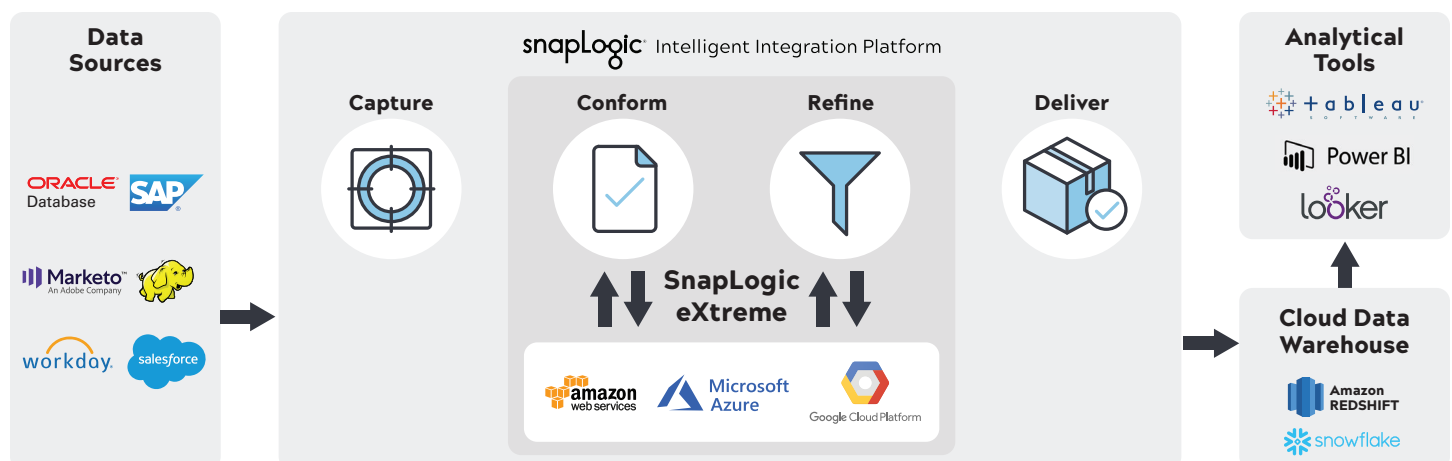
programming interface eliminates the need for error-prone manual coding procedures, leading to quicker time-to-value without the traditional dependence on complex IT or data-science organizations. Unlike other data integration solutions that require integrators to have detailed knowledge on how to build and submit Spark jobs, SnapLogic eXtreme allows business users with domain expertise to perform complex processing and transformations on extremely large volumes of data within the enterprise's existing big data infrastructure.
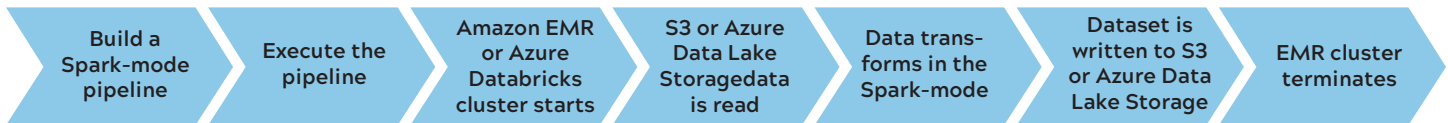
Key Features of SnapLogic eXtreme

- Develop Spark Pipelines using visual design (zero code).

- Fully managed and automated big data runtime environment with a serverless architecture.

- Lifecycle management of cloud-based, transient, big data clusters in AWS/Azure Cloud environments.

- Run complex and high-volume data transformation routines at elastic scale for terabytes or petabytes of data.

- Perform read-write operations using Cloud Data Warehouses like Redshift and Snowflake with data stored in data lakes like AWS S3 and Azure WASB and support for Databricks Delta Lake for better reliability

- Pushdown Optimization support for Cloud Data Warehouses like Redshift and Snowflake.

- Execute PySpark and Java Spark applications.

Comparing SnapLogic eXtreme against the conventional approach

Task	Conventional Approach		With SnapLogic eXtreme	
Install & Configure a secure Big Data Runtime Environment	100 Lines of Code (LOC)	4 Hours	Automated (**)	0.25 Hour (15 Mins)
Develop Spark Pipeline to process data in Data Lake (*)	2000 LOC	80 Hours	0 LOC + 30 Snaps	4 Hours
Develop mechanism for Spark Pipeline Job submission in Big Data Runtime Environment	100 LOC	4 Hours	Automated	0 Hours
Set up Life cycle Management policies for Big Data Runtime Environment (Termination, Auto-scaling, Monitoring, Instance Type Selection)	500 LOC	40 Hours	Automated	0 Hours
Develop mechanism to handle schema changes (and software version upgrades)	500 LOC	40 Hours	0 LOC + 5 Snaps	0.25 Hour (15 Mins)
Total	**3200 LOC**	**168 Hours/21 Days**	**0 LOC + 35 Snaps**	**4.5 Hours**

Workflow with SnapLogic eXtreme

| Build a Spark-mode pipeline | → | Execute the pipeline | → | Amazon EMR or Azure Databricks cluster starts | → | S3 or Azure Data Lake Storagedata is read | → | Data transforms in the Spark-mode | → | Dataset is written to S3 or Azure Data Lake Storage | → | EMR cluster terminates |

(*) Pipeline includes multiple transformations over Store Sales, Household Demographics and Customer data; Final LOC will actually be much higher due to the iterative nature of development

(**) Fully automated after completion of a one-time configuration