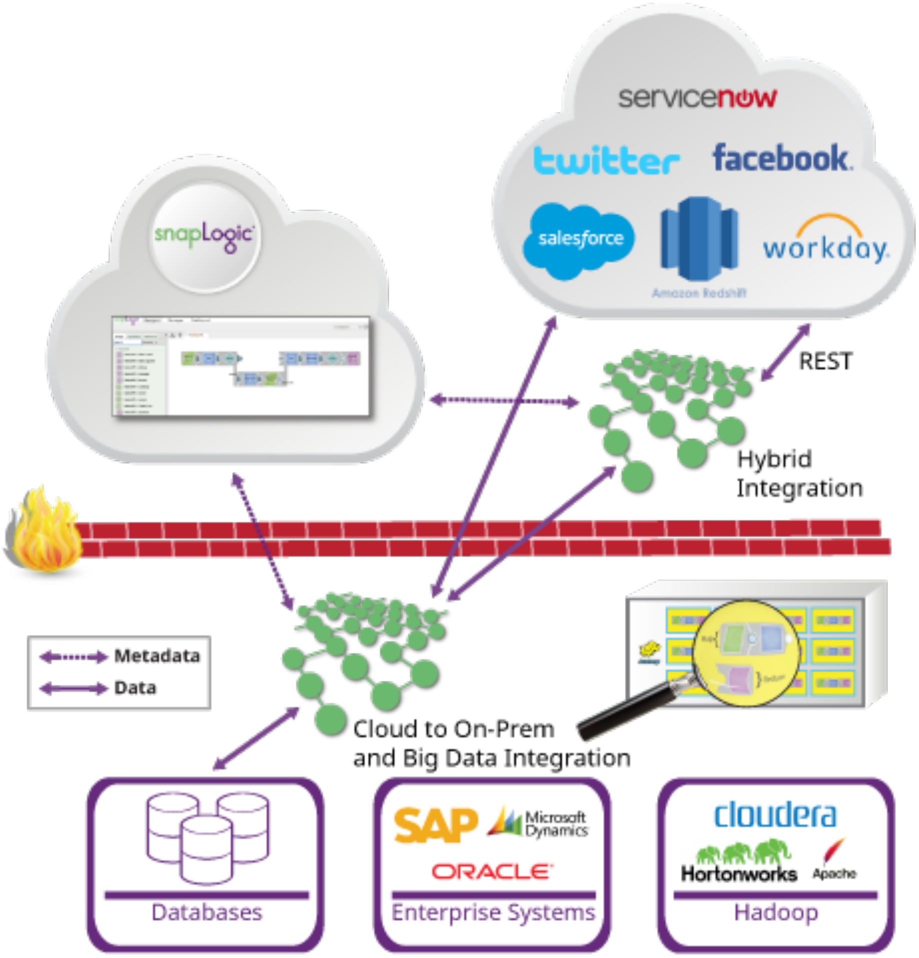
SNAPLOGIC BIG DATA INTEGRATION PROCESSING PLATFORMS



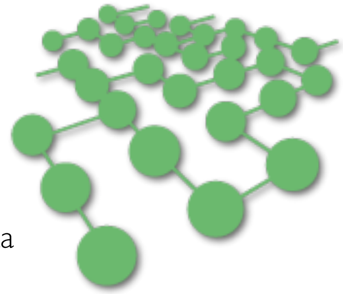
Big Data Integration Processing Platforms

One of our goals at SnapLogic is to match data flow execution requirements with an appropriate execution platform. Different data platforms have different benefits. This paper reviews the nature of SnapLogic data flow pipelines and how to choose an appropriate data platform. In addition to categorizing pipelines, this paper also explains our currently supported execution targets and our planned support for Apache Spark.

First, some preliminaries. All data processed by SnapLogic pipelines is handled natively in an internal JSON format. We call this *document-oriented processing*. Even flat, record-oriented data is converted into JSON for internal processing. This lets us handle both flat and hierarchical data seamlessly. Pipelines are constructed from Snaps. Each Snap encapsulates specific applications or technology functionality. The Snaps are connected together to carry out a data flow process. Pipelines are constructed with our visual designer. Some Snaps provide connectivity, such as connecting to databases or cloud application. Some Snaps allow for data transformation such as filtering out documents, adding or removing fields or modifying fields. We also have snaps that perform more complex operations such as sort, join and aggregate.

Given this setup, we can categorize pipelines into two types: streaming and accumulating. In a streaming pipeline, documents can flow independently. The processing of one document is not dependent on another document as they flow through the pipeline. Such streaming pipelines have low memory requirements because documents can exit the pipeline once they have reached the last Snap. In contrast, an accumulating pipeline requires that all documents from the input source must be collected before result documents can be emitted from a pipeline. Pipelines with sort, join and aggregate are accumulating pipelines. In some cases, a pipeline can be partially accumulating. Such accumulating pipelines can have high memory requirements depending on the number of documents coming in from an input source.

Now let's turn to execution platforms. SnapLogic has an internal data processing platform called a Snaplex. Think of a Snaplex as a collection of processing nodes or containers that can execute SnapLogic pipelines. We have a few flavors of Snaplexes:



- A **Cloudplex** is a Snaplex that we host in the cloud and it can autoscale as pipeline load increases.
- A **Groundplex** is a fixed set of nodes that are installed on-premises or in a customer VPC. With a Groundplex, customers can do all of their data processing behind their firewall so that data does not leave their infrastructure.

We are also expanding our support for external data platforms. We have recently released our Hadooplex technology that allows SnapLogic customers to use Hadoop as an execution target for SnapLogic pipelines. A Hadooplex leverages YARN to schedule Snaplex containers on Hadoop nodes in order to execute pipelines. In this way, we can autoscale inside a Hadoop cluster. Recently we introduced SnapReduce 2.0, which enables a Hadooplex to translate SnapLogic pipelines into MapReduce jobs. A user builds a designated SnapReduce pipeline and specifies HDFS files and input and output. These pipelines are compiled to MapReduce jobs to execute on very large data sets that live in HDFS. (Check out the demonstration in our recent cloud and big data analytics webinar.)

Finally, as we announced as part of Cloudera's real-time streaming announcement, we've begun work on our support for Spark as a target big data platform. A Sparkplex will be able to utilize SnapLogic's extensive connectivity to bring data into and out of Spark RDDs (Resilient Distributed Datasets). In addition, similar to SnapReduce, we will allow users to compile SnapLogic pipelines into Spark codes so the pipelines can run as Spark jobs. We will support both streaming and batch Spark jobs. By including Spark in our data platform support, we will give our customers a comprehensive set of options for pipeline execution.

Choosing the right big data platform will depend on many factors: data size, latency requirements, connectivity and pipeline type (streaming versus accumulating). Here are some guidelines for choosing a particular big data integration platform:

Cloudplex

- Cloud-to-cloud data flow
- Streaming unlimited documents
- Accumulating pipelines in which accumulated data can fit into node memory

Groundplex

- Ground-to-ground, ground-to-cloud and cloud-to-ground data flow
- Streaming unlimited documents
- Accumulating pipelines in which accumulated data can fit into node memory

Hadooplex

- Ground-to-ground, ground-to-cloud and cloud-to-ground data flow
- Streaming unlimited documents
- Accumulating pipelines can operate on arbitrary data sizes via MapReduce

Sparkplex

- Ground-to-ground, ground-to-cloud and cloud-to-ground data flow
- Allow for Spark connectivity to all SnapLogic accounts
- Streaming unlimited documents
- Accumulating pipelines can operate on data sizes that can fit in Spark cluster memory

Hortonworks CERTIFIED TECHNOLOGY PARTNER

cloudera CERTIFIED

Note that recent work in the Spark community has increased support for out-of-core computations, such as sorting. This means that accumulating pipelines that are currently only suitable for MapReduce execution may be supported in Spark as out-of-core Spark support becomes more general. The Hadooplex and Sparkplex have added reliable execution benefits so that long-running pipelines are guaranteed to complete.

At SnapLogic, our goal is to allow customers to create and execute arbitrary data flow pipelines on the most appropriate data platform. In addition, we provide a simple and consistent graphical UI for developing pipelines which can then execute on any supported platform. Our platform agnostic approach decouples data processing specification from data processing execution. As your data volume increases or latency requirements change, the same pipeline can execute on larger data and at a faster rate just by changing the target data platform. Ultimately, SnapLogic allows you to adapt to your data requirements and doesn't lock you into a specific big data platform.

About SnapLogic

SnapLogic connects enterprise data and applications in the cloud and on-premises for faster decision-making and improved business agility. With the SnapLogic Elastic Integration Platform, organizations can more quickly and affordably accelerate the "cloudification" of enterprise IT with fast, multi-point and modern connectivity of big data, applications and things. Funded by leading venture investors, including Andreessen Horowitz and Ignition Partners, and co-founded by Gaurav Dhillon, co-founder and former CEO of Informatica, SnapLogic is utilized by prominent companies in the Global 2000. For more information about SnapLogic, visit www.SnapLogic.com.

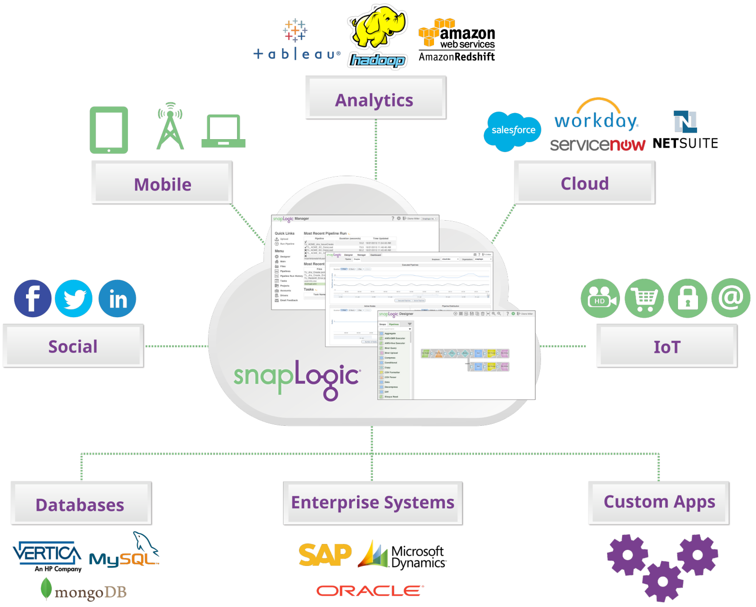


tableau
hadoop
amazon web services
Amazon Redshift
Analytics
Mobile
salesforce
workday.
servicenow
NETSUITE
Cloud
Social
IoT
snapLogic
Databases
Enterprise Systems
Custom Apps
VERTICA An HP Company
MySQL
mongoDB
SAP
Microsoft Dynamics
ORACLE