



## SnapLogic: From ETL to VVV

Squeezing value out of data in the enterprise has always pushed the limits of the available technology. Furthermore, when business needs exceed available capabilities, vendors push to innovate within the processor, storage, and network constraints of the day.

This inherent stress between enterprise demands and vendor innovation gave rise to the Extract, Transform, and Load (ETL) marketplace over twenty years ago. Business realized that building complex, ad hoc SQL queries on increasingly large databases would grind them to a halt, thus requiring an alternate approach to gaining essential business intelligence.

The best solution given the hardware limitations of the time required controlled, pre-planned extraction of data from various databases of record, followed by complex, time-consuming transformation steps, and then loading the transformed data into separate reporting data stores (dubbed data warehouses and data marts) specially optimized for a range of analytical queries.

As available storage and memory ramped up, ad hoc data transformations became increasingly practical, allowing for the transform step to take place as needed, subsequent to the load step – and Extract, Load, and Transform (ELT) became a popular alternative to ETL.

The transition from ETL to ELT represents an important stepping stone to real-time data analysis. ELT still wasn't truly real-time, as businesses had to extract and load their data ahead of time, but much of the analysis work depended on the now-accelerated and increasingly flexible transformation step.

### Hadoop and ELT

Today, all the buzz is about Big Data and the most important technology innovation on the Big Data analysis scene: Hadoop. In spite of all the commotion around Hadoop, this open source platform and its various add-ons are little more than the next generation of the transform capability of ELT, albeit at cloud scale.

The core motivations that drove the Hadoop community to create this tool were the increasing size of data sets (leading to the awkward *Big Data* terminology), as well as the need to process data of diverse levels of structure – in particular, a mix of unstructured (content-centric) and semi-structured (generally XML-formatted), as well as structured (relational) information.

In other words, traditional ETL and ELT tools weren't up to the challenge of dealing with the volume and variety of data that enterprises increasingly produced and wished to analyze. Hadoop addressed these

TRADITIONAL ETL AND ELT TOOLS WEREN'T UP TO THE CHALLENGE OF DEALING WITH THE VOLUME AND VARIETY OF DATA THAT ENTERPRISES INCREASINGLY PRODUCED AND WISHED TO ANALYZE.

challenges with a horizontally scalable, highly redundant file system (the *Hadoop Distributed File System*, or HDFS), as well as *MapReduce*, an algorithmic approach to analyzing data appropriate for processing the necessary volumes of data on HDFS.

The first version of Hadoop, however, was essentially a batch analytics platform. Data analysts had to surmount the significant challenges of extracting data from their source locations and loading them properly into HDFS, only to run arcane MapReduce jobs to produce useful results. As a result, the hype surrounding Hadoop 1.0 exceeded its actual usefulness for most organizations brave enough to implement it.

As an open source project, however, Hadoop had enough backing from the community to drive the development of version 2, which offers a resource negotiator for MapReduce tasks dubbed YARN as well as fledgling real-time processing capabilities. Today, real-time Hadoop is at the cutting edge, as various tools in an expanding Hadoop ecosystem mature to address the velocity requirements for real-time data analytics.


### **Hadoop's Missing Pieces**

In terms of the maturation of ETL technologies, therefore, the current version of Hadoop can be thought of as a modern transformation engine running on a horizontally scalable file system that in theory offers the “three V’s” of Big Data: [volume](#), [variety](#), and [velocity](#). In practice, however, many capabilities are missing from the open source distribution.

As a result, other open source projects as well as commercial software providers have an opportunity to fill in the gaps that Hadoop leaves in the areas of enterprise data integration in the context of modern enterprise infrastructures. Today, such integration scenarios typically fall within hybrid cloud environments that combine on-premise and cloud-based capabilities.

In the enterprise context, the extract and load steps of ELT require organizations to leverage diverse data sources both on-premise and in the cloud. Those data sources may be a mix of relational, hierarchical, or content-centric. Furthermore, the business may require real-time (or near real-time) analysis of data from such diverse data sources.

To address these challenges, [SnapLogic](#) has built a data and application integration platform that resolves many of Hadoop's shortcomings. As I wrote about in [a previous BrainBlog post](#), SnapLogic separates their technology into a *Control Plane* and a *Data Plane*. The Control plane resides in the cloud and contains the *Designer*, *Manager*, and *Dashboard* subcomponents which manage the Data Plane, allowing the Data Plane to act as a cloud-friendly abstraction of the data flows or Pipelines that users can create with the SnapLogic Designer.



THE HYPE SURROUNDING  
HADOOP 1.0 EXCEEDED ITS  
ACTUAL USEFULNESS FOR  
MOST ORGANIZATIONS  
BRAVE ENOUGH TO  
IMPLEMENT IT.

The data integrations themselves run as Pipelines, which are sequences of atomic integration steps that SnapLogic calls *Snaps* – because people literally snap them together. Snaps support the full gamut of data types and levels of structure, facilitating the ability to send the full variety of enterprise data to Hadoop.

SnapLogic has also recently rolled out *Hadooplexes*, which are Snaplexes (data processing components) that run as YARN apps in Hadoop, as well as *SnapReduce*, SnapLogic's support for Big Data integrations that leverage Hadoop to process large amounts of data across large clusters.

SnapReduce enables Pipelines to generate MapReduce jobs and scale them across multiple nodes in a Hadoop cluster. Each Hadooplex then delegates MapReduce-based analytic operations automatically across all Hadoop nodes, thus abstracting the horizontally distributed nature of the Hadoop environment from the user.


The result is an elastic, horizontally scalable integration fabric that provides the extract and load capabilities that Hadoop lacks. Each data integration can be run manually, on a preset schedule, or via a trigger – and SnapLogic exposes such triggers as URLs (either on the Internet or a private network), allowing any authorized piece of software to kick off the integration.

### End-to-End Modern ELT

In summary, SnapLogic modernizes each element of ELT for today's modern, cloud-centric, Big Data world. Instead of traditional extraction of structured data, SnapLogic allows for diverse queries across the full variety of data types and structures by streaming all data as JSON documents. Instead of simplistic, point-to-point loading of data, SnapLogic offers elastic, horizontally scalable Pipelines that hide the underlying complexity of data integration from the user. And within Hadoop, Hadooplexes simplify the distribution of YARN-based MapReduce algorithms, allowing users to treat the Hadoop environment as though it were a traditional reporting database.

Furthermore, SnapLogic can perform each of these steps in real-time, in those situations where the business requires real-time analytics. Each pipeline simply streams the data from the acquisition point to the delivery point, handling the appropriate operations statelessly along the way. The end result is a user-friendly data integration and analysis tool that adroitly hides an extraordinary level of complexity behind the scenes – opening up the power of Big Data to an increasingly broad user base.

*SnapLogic is an [Intellyx](#) client. At the time of writing, no other organizations mentioned in this article are Intellyx clients. Intellyx retains full editorial control over the content of this article.*



THE END RESULT IS A USER-FRIENDLY DATA INTEGRATION AND ANALYSIS TOOL THAT ADROITLY HIDES AN EXTRAORDINARY LEVEL OF COMPLEXITY BEHIND THE SCENES – OPENING UP THE POWER OF BIG DATA TO AN INCREASINGLY BROAD USER BASE.