White Paper

Intelligent ETL (Extract, Transform, Load)

90% of your Big Data problem isn't Big Data.

It's the ability to handle Big Data for better insight.

By Arjuna Chala

LexisNexis®

Risk Solutions
HPCC Systems

Introduction

LexisNexis is a leader in providing essential information that helps customers across all industries and government verify identity, assess risk, and predict and detect fraud. To manage, sort, link, and analyze billions of records within seconds, LexisNexis designed a data-intensive supercomputer based on high performance cluster computing (HPCC) technology. The supercomputer, called HPCC Systems, has been proven for the past ten years with customers who need to process large volumes of data in critical 24/7 environments. Customers such as leading banks, insurance companies, utilities, law enforcement and federal government, leverage the HPCC Systems platform technology through various LexisNexis products and services. Originally conceived in response to the need of LexisNexis to manage its own big data challenges, the HPCC Systems platform helped LexisNexis Risk Solutions scale to a $1.4 billion information solutions division, and has evolved to become a mainstream big data analytics system leveraged across many industries.

HPCC Systems® Introduction

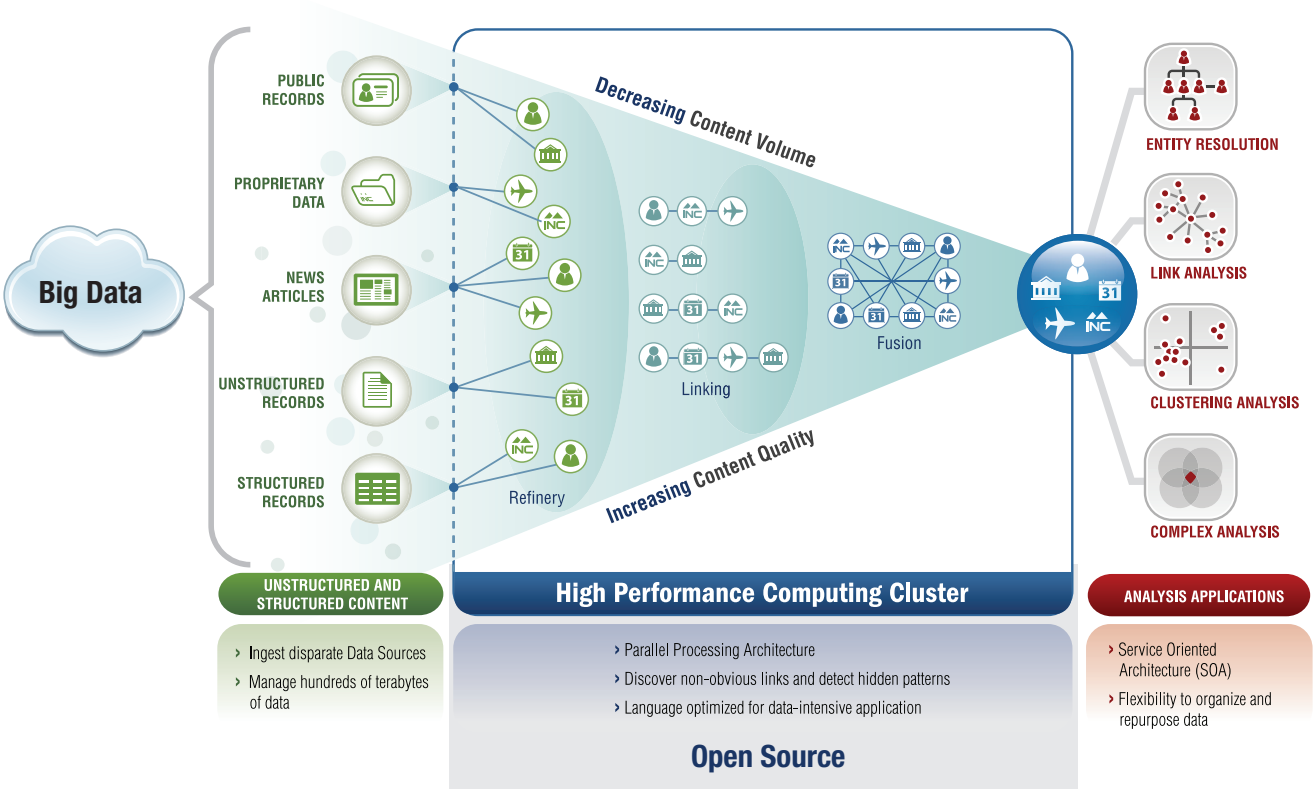
The **conceptual vision** for this computing platform is depicted in the figure below. The HPCC Systems platform was built specifically to analyze large volumes of data in minutes rather than days or months to solve complex problems. The platform can deliver the data and reports to more people and provide fast reporting. Since only small development teams are needed – this reduces investment in large teams and keeps processes agile.

The HPCC Systems platform has three distinguishing factors that make it an effective choice for big data analytics and processing:

The HPCC Systems Data Refinery engine (Thor) helps clean, link, transform and analyze Big Data. Thor supports ETL (Extraction, Transformation and Loading) functions like ingesting unstructured/structured data out, data profiling, data hygiene, and data linking out of the box. In addition, Thor supports flexible record oriented data structures.

The HPCC Systems Data Delivery engine (Roxie) provides highly concurrent and low latency real time query capability. The Thor processed data can be accessed by large number of users concurrently in real time fashion using the Roxie. The Roxie queries are typically complex and could include embedded rules logic.

The Data Scientist friendly programming language, called Enterprise Control Language (ECL), is used to program both the data processing jobs on Thor and the queries on Roxie. ECL is a declarative, implicitly parallel and data flow oriented programming language that abstracts complex data processing tasks by providing a simple programming interface.

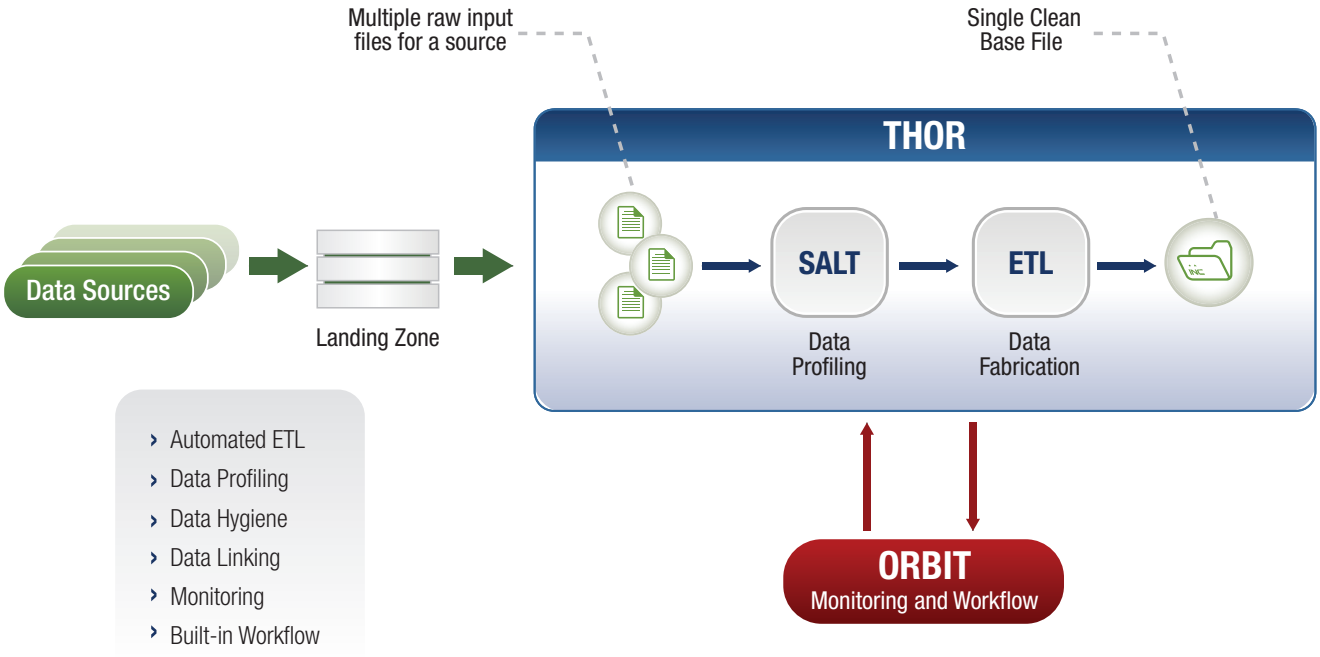


HPCC Systems and ETL

The Problem

The typical data ingest process where data is gathered from multiple sources; loaded, cleaned, linked and analyzed is both complex and resources intensive. In the LexisNexis data factory, it has been observed that 60-80% of the effort in trying to analyze or mine data happens in the ETL process. This is where data is ingested from several thousand data sources daily and transformed into cleaned data that can then be used for data mining. Traditional methods of ETL using drag and drop user interfaces fail to address the needs around raw data that is constantly in flux. Manual intervention and remapping of processes would be needed to account for the changes in the data.

The HPCC Systems ETL Platform



The HPCC Systems ETL platform is built to be flexible and scalable at the same time. The underlying ECL programming language and built in components provides the foundational tooling to build fully automated ETL jobs that can quickly ingest new data and adjust to changes in schema. The key components are described below:

THOR

Thor (the Data Refinery Cluster) is responsible for consuming vast amounts of data, transforming, linking and indexing that data. It functions as a distributed file system with parallel processing power spread across the nodes. A cluster can scale from a single node to thousands of nodes.

ECL

ECL (Enterprise Control Language) is the powerful declarative programming language that is ideally suited for the manipulation of Big Data. The declarative nature of the language makes it well suited for a Data Scientist.

SALT

SALT is an acronym for Scalable Automated Linking Technology. It is a programming environment support tool which functions as an ECL code generator to automatically produce ECL code for a variety of data integration applications, addressing common data processing tasks:

- data ingest
- data profiling
- data hygiene
- record linking and clustering
- data source consistency monitoring
- file comparison to determine delta changes between versions of a data file
- data parsing and classification

ORBIT

The ORBIT component provides the monitoring, rules and workflow management. ORBIT monitors all the jobs and reports on job execution status (start, in process, success and failure). Built in alerting will help manage the jobs that fail and notify people to take action. A complete audit trail helps ETL developer quickly identify and fix issues. In addition, ORBIT ensures that all incoming data meets field level criteria's set using rules.

For more information:

Call 877.316.9669
or visit http://hpccsystems.com

About HPCC Systems®
HPCC Systems® from LexisNexis® Risk Solutions offers a proven, data-intensive supercomputing platform designed for the enterprise to process and deliver Big Data analytical problems. As an alternative to Hadoop and mainframes, HPCC Systems offers a consistent data-centric programming language, two processing platforms and a single architecture for efficient processing. Customers, such as financial institutions, insurance carriers, insurance companies, law enforcement agencies, federal government and other enterprise-class organizations leverage the HPCC Systems technology through LexisNexis® products and services. For more information, visit http://hpccsystems.com.

About LexisNexis® Risk Solutions
LexisNexis® Risk Solutions (www.lexisnexis.com/risk/) is a leader in providing essential information that helps customers across all industries and government predict, assess and manage risk. Combining cutting-edge technology, unique data and advanced scoring analytics, we provide products and services that address evolving client needs in the risk sector while upholding the highest standards of security and privacy. LexisNexis Risk Solutions is part of Reed Elsevier, a leading publisher and information provider that serves customers in more than 100 countries with more than 30,000 employees worldwide.