HPCC SYSTEMS

LexisNexis® RISK SOLUTIONS

WHITE PAPER

HPCC Systems: The End-to-End Data Lake Management Solution

Data lakes are helping leading organizations solve the problem of extremely large, unstructured datasets, allowing them to increase responsiveness and scalability while reducing costs.

INTRODUCTION

Today, most organizations recognize that data is key to the ability to innovate and remain competitive in a rapidly changing business landscape. A key challenge:

As datasets become larger and more complex, it's impossible to quickly respond to changing business needs using traditional relational data store such as data warehouse.

That's largely because it's difficult, time-consuming and often expensive to add new data and access paths to relational data stores. The problem is becoming increasingly acute as businesses use more unstructured information that relational databases simply weren't designed to handle, such as data from Internet of Things devices, the Web, and social media.

To overcome this challenge, many organizations — including some of the world's largest companies — are successfully using a proven alternative approach: a data lake. Data lakes support datasets that are extremely large, complex and diverse, and they easily accommodate new data sources such as IoT. They allow IT groups to quickly create new applications that support changing business needs, unlocking the power of complex data for all users within the organization. They also scale much more easily and cost-effectively than relational databases. As a result, data lakes enable greater responsiveness to business groups and external customers, reduced costs, and greater scalability.

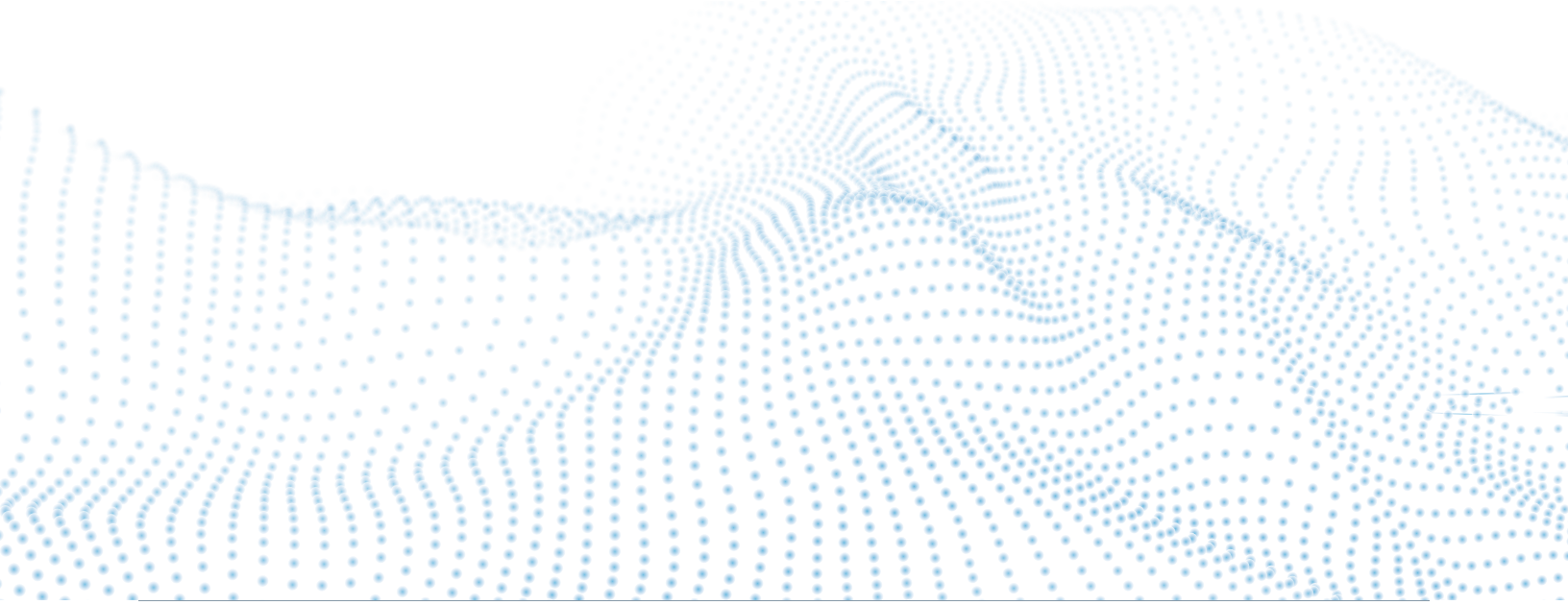
UNLOCKING THE VALUE OF DATA

Data lakes embrace the little-known truth that the key to unlocking the value of data lies in combining and synthesizing data from multiple sources and perspectives.

The value of the information expands exponentially as the organization discovers and exploits the synergies between data elements.

However, although the potential synergies in any large, complex set of data are almost infinite, they are not immediately obvious — they are discovered over time through insight and experimentation.

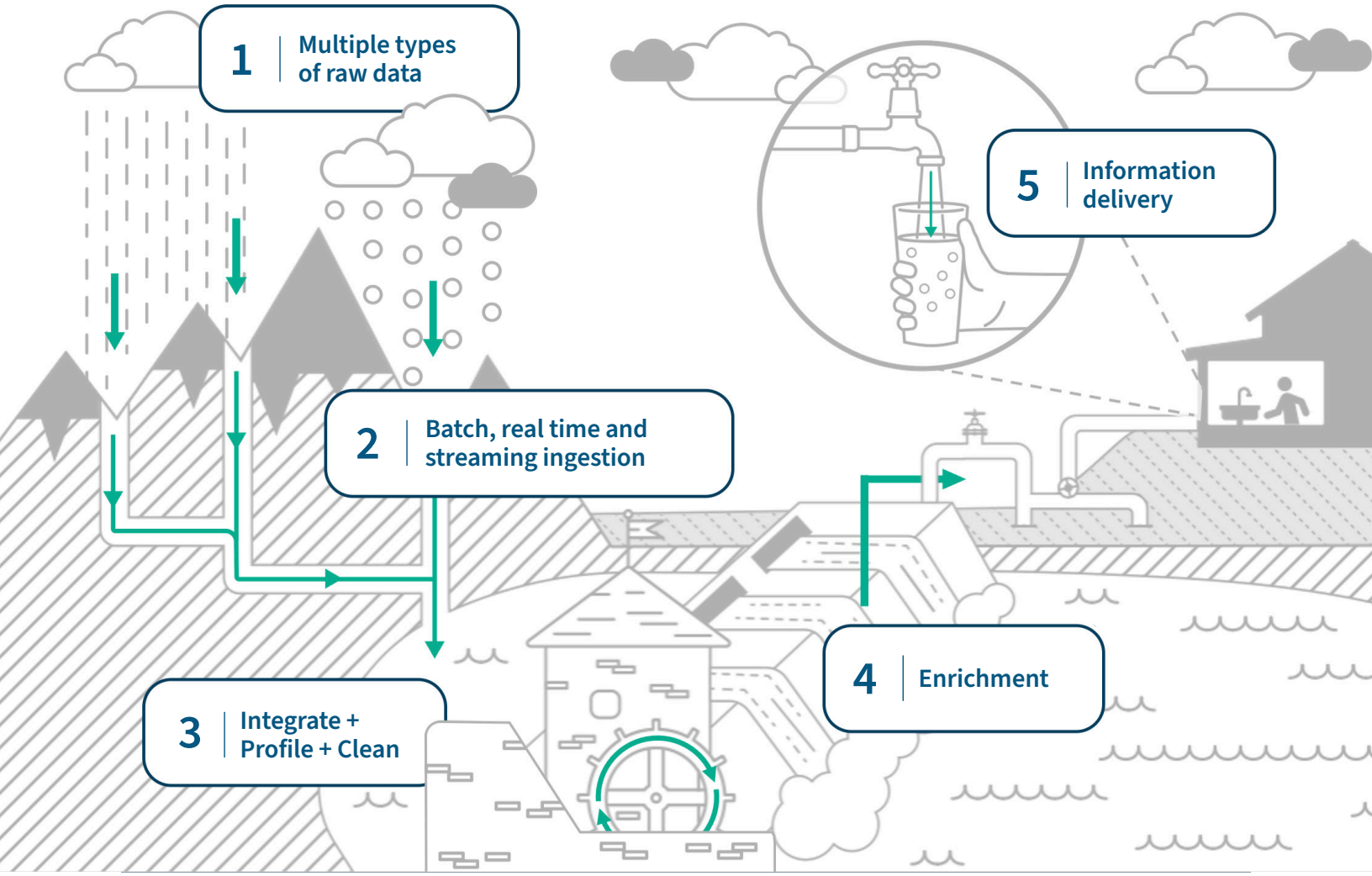
This means that any data store that excludes most of the raw data and includes only a final subset — like a typical data warehouse — also excludes most of the data's potential value. In a data lake, the ability to rapidly and repeatedly enrich and refine the raw data to exploit the synergies is critical to capitalizing on the value of the data.



WHAT IS A DATA LAKE?

A data lake is a distributed data store that can hold all an organization's data — including unstructured data such as text and images as well as the structured tabular data that's traditionally stored in relational databases. Data lakes can accommodate data from any source, including IoT sensors, social media, and weblogs as well as operational applications. Data lakes are typically distributed across clusters of hardware nodes that process data in parallel; you can increase the size or performance of the data lake by simply adding nodes.

A key feature of data lakes is that they retain data at all stages of processing and refinement — from the raw incoming data to enriched application-ready data. Developers and users can take advantage of data at any stage of enrichment.



HOW A DATA LAKE DIFFERS FROM A DATA WAREHOUSE

The biggest advantage of a data lake over a data warehouse is that it lets an organization respond to changing business needs more quickly and easily. In addition, data lakes are much easier to scale as the volume of data grows. Those advantages stem from fundamental differences in the way that data lakes store and access data, compared to data warehouses.

Data warehouses are repositories of data collected from business applications and used for analysis and management reporting. Based on relational databases, they are highly structured and optimized for specific applications and queries. In general, the data warehouse holds only the data required to support these applications. The structure of the database is determined at the outset and defined in a database schema. Raw data typically must be transformed to match the schema before it can be imported into the database.

This makes it difficult to respond to new business requirements or to add new data to the database. The schema must be centrally controlled, which can become increasingly complex as the database grows over time. Developers cannot build new applications unless the data they need is available in the schema and they are provided with the necessary access paths. If the data isn't already available, they need to request extensions to the schema, which is typically a lengthy project. This approach also introduces very expensive dependencies into the development process, because changes to the schema can affect other applications that use the data.

Furthermore, developers and users can't take advantage of raw data or interim data transformation stages, because the warehouse typically contains only the application-ready final data.

The biggest advantage of a data lake over a data warehouse is that it lets an organization respond to changing business needs more quickly and easily.

With a data lake, in contrast, all the organization's raw structured and unstructured data is imported into the data lake in a simple flat format. Raw data is progressively enriched and transformed to produce enhanced datasets for specific applications. Each time the data is enhanced, the new data layer remains in the data lake; nothing is lost. There may be many of these layers produced by different developers for different purposes. In fact, there is no concept of a "final" form of the data.

Data curation is used to make all of these data layers visible and available for reuse, and to provide developers and users with an understanding of all the layers and how they were produced. The data transformations between data sets are also documented.

DATABASE CENTRIC PERSPECTIVE

Stores only the structured data required for a specific set of applications and queries.

Data structure, schema and access paths are defined in advance to support predetermined uses.

Data must be transformed to match schema before it can be imported into data warehouse.

Only the final transformed data is available to developers.

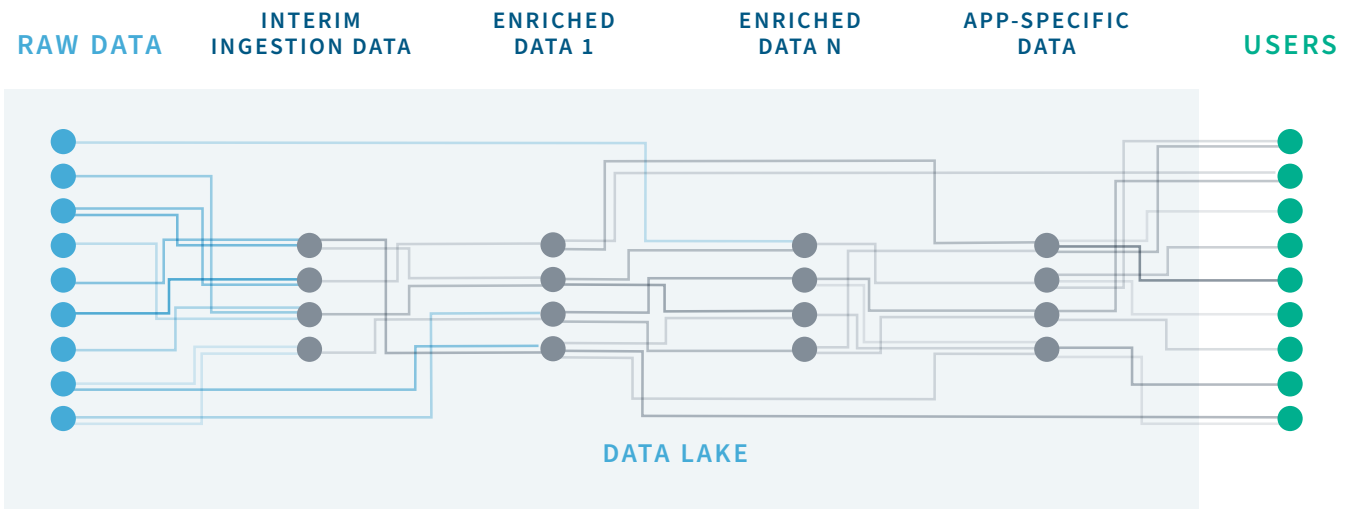
DATA LAKE PERSPECTIVE

Stores all the organization's data, including structured and unstructured raw data from any source.

All data is stored in a simple flat format. No need to decide in advance how data will be accessed.

No need to transform data before adding it to data lake; all stages of transformation and processing are stored within the data lake.

Developers have access to all data, including raw data, all layers of processing and final application-ready data.



HOW A DATA LAKE ENABLES GREATER RESPONSIVENESS TO THE BUSINESS

Adopting a data lake has profound implications for the ability to respond quickly to changing business needs. Developers can easily and quickly bring up new applications, because they have access to all of the organization's data within the data lake, and because new enrichment and access paths can be created as needed.

There's no need to anticipate all data needs and usage paths up front. A large number of developers, each focused on their own needs, can develop concurrently against the data without unnecessary dependencies — so they create more new applications, faster. New applications can reach into the data lake anywhere, accessing any of the data layers, including the raw data and intermediate data produced by other applications.

HOW A DATA LAKE PROVIDES GREATER SCALABILITY

The hardware architecture supporting a data lake differs substantially from the typical data warehouse architecture. Data warehouses typically run on expensive high-speed and redundant storage arrays (such as RAID). It is expensive to expand storage, and the architecture imposes fixed limits on storage bandwidth and storage space. In contrast, data lakes are distributed across clusters of relatively low-cost commodity hardware, making it possible to scale a data lake incrementally and at relatively low cost by adding hardware nodes.

The language of data lakes: Data puddles, swamps and oceans

The increasing popularity of data lakes has spawned a handful of related terms describing different types of data store.

Data puddle
A small data store containing data used for a specific application.

Data swamp
An unmanaged data lake that delivers little value, often because lack of curation means the data is hard to find and use.

Data ocean
An amalgam of data from multiple independently curated data lakes.

MANAGING A DATA LAKE: STEPS, CHALLENGES AND SOLUTIONS

There are three main aspects of managing a data lake:

Data acquisition and enrichment: Importing and transforming data for use by different applications.

Data delivery: Delivering data to many users across the organization.

Data curation: Managing and cataloging the contents of the data lake so that developers and other users can see what's in the data lake and use the data in applications.

Each of these steps can present challenges, largely because a data lake may have to manage extremely large datasets and support many users. A successful data lake requires technology that can rapidly ingest and enrich vast amounts of data, enable rapid development by multiple development teams, and scale to enable speedy access by large numbers of users across the organization. A single, powerful declarative language (ECL) is used to define operations for both Thor and Roxie, accelerating and simplifying development.

Procedural vs. Declarative languages

How a declarative language like HPCC Systems ECL differs from procedural languages:

Procedural
You tell the computer what steps you want it to perform. The programmer needs to be aware of the distributed processing environment.

Declarative
You define the source, transformations, and form of the desired data. The language determines how to get that job done. It can handle parallelization and sophisticated optimizations transparently to the user.

DATA ACQUISITION AND ENRICHMENT

This stage includes all the steps involved in importing data into the data lake and transforming it for use by different applications.

For many organizations, the ability to quickly import data into the data lake and make it available to applications is critical to using that data for competitive advantage. For example, a business group may rely on new sales or operational data becoming available for analysis by the next day. If the data acquisition and enrichment steps take too long, the information can become stale and lose its business value. Therefore, a data lake platform must be able to import vast amounts of data and conduct extremely complex enrichment operations at high speed, reliably and within a predictable timeframe.

How HPCC Systems handles data acquisition and enrichment

HPCC Systems' data enrichment engine is designed to solve these challenges. It transparently takes advantage of parallel processing to accelerate each data acquisition and enrichment operation. The dataset size is limited only by the available storage space, and processing capacity is elastic based on the resources assigned. Datasets are automatically streamed from memory or non-volatile storage as appropriate, allowing HPCC Systems to handle virtually unlimited dataset sizes. HPCC Systems Super files simplify management of frequently updated information, such as log files, by allowing a base file and its incremental updates to be treated as a single file, including cross-file indexing.

All acquisition and enrichment tasks, of any level of complexity, are defined as data transformations in ECL, a language that was specifically designed to make it easier to specify data transformations without worrying about how they will be executed. The system then automatically compiles these transformations into high-speed, production-ready parallelized execution streams.

HPCC Systems handles all of the parallelization and execution complexities for you, automatically distributing tasks among the available resources. The platform ensures non-stop operation even in the event of node failure, ensuring the timely availability of data to the users and applications that depend on it.

DATA ACQUISITION

Data acquisition steps include:

Data retrieval: Bringing in data from many different data sources. Some organizations have thousands of different internal and external data sources, each in its own format and often updated at different frequencies.

Merging incremental updates and similar data sources as early steps toward integration.

Cleaning the data to remove obvious errors or omissions like a missing name.

Normalization: Reconciling differences in similar datasets and producing a smaller set of more homogenous data.

DATA ENRICHMENT

Once the data has been added to the data lake, a wide range of data enrichment steps can be applied to prepare the data for use. These steps include:

Data refinement: Incrementally refining data for use by different applications. There may be many refinement steps, each producing a new layer of data. Each layer is retained in the data lake and becomes available to any application.

Entity resolution: Mapping all related data to a single entity. This may involve sophisticated probabilistic matching across multiple record types and fields.

Cross-entity linking: Establishing links between related entities. One key limitation of traditional data warehouse environments is that these relationships are resolved in a non-transparent way before the data reaches the warehouse, using hard-coded software modules. With HPCC Systems, all of the transformations are visible and documented, and the intermediate steps are available to developers to build alternate resolutions when one format does not fill all uses.

Analytics and machine learning: Many types of analytics can be applied to the data to uncover trends and relationships that are valuable to the organization. This includes the use of machine learning algorithms to identify patterns in historical data and use the information to make predictions about the future. HPCC Systems provides a robust library of parallelized analytics and machine learning algorithms that can be used to analyze data, text, images, videos, and sensor data. These capabilities, along with the powerful expressiveness of ECL, allow organizations to fully capitalize on the value inherent in their data.

Indexing: Creating indexes to speed data access, based on an understanding of how users will access the data. It's essential to be able to add indexes for new applications to ensure rapid response for users. The ability to add indexes on an ad-hoc basis is a key advantage of data lakes over relational databases.

DATA DELIVERY

The data lake provides a source of data that may be used by many different applications and internal and external users. However, even the richest data repository can only be effectively monetized if that data is rapidly made available to users in a scalable, reliable, secure, and high-performance fashion.

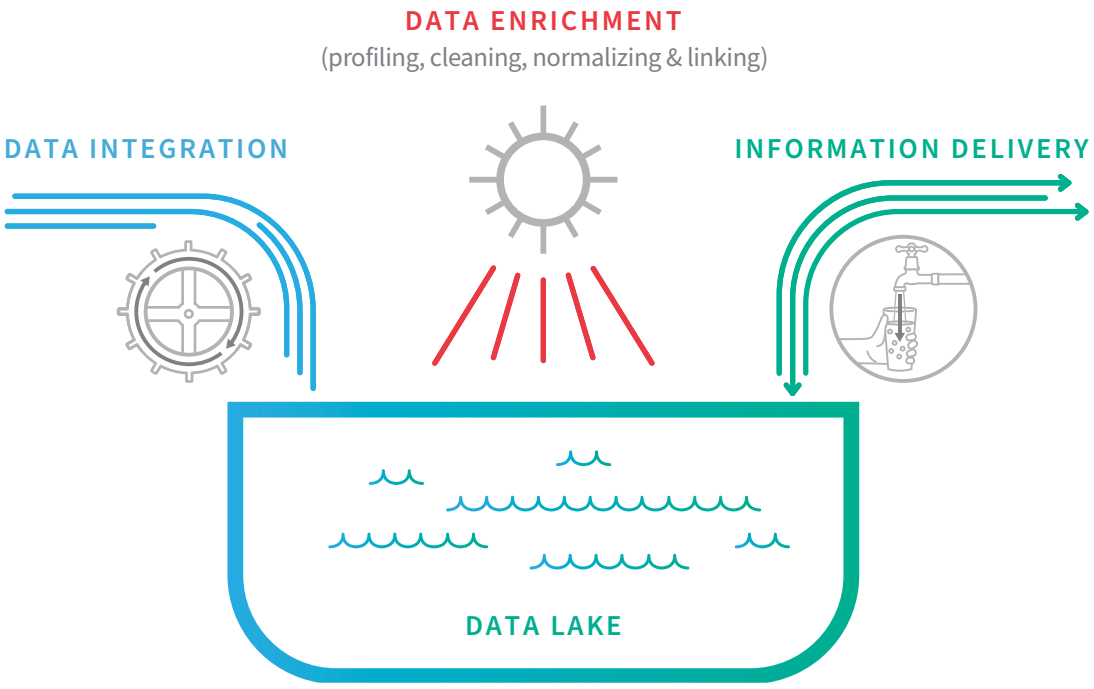
Therefore, the data lake should not impose hard limits on scalability. It must be able to scale to deliver data to an extremely large audience while continuing to provide fast response times. At the same time, it must provide the ability to securely segregate information so that only the right people have access to a given slice of data.

How HPCC Systems solves data delivery challenges

HPCC Systems provides a very robust information delivery engine (Roxie) that is scalable, fault-tolerant, and extremely fast. A sophisticated security capability ensures that the data resources are fully protected and segregated to prevent unauthorized access.

The platform serves data directly from the data lake (or optional database caches) using parallel processing, enabling it to support large numbers of users and client devices. The system automatically manages the complexities of parallel processing. Developers can create new indexes as needed to accelerate the response to user queries. Data can be delivered to users through a variety of APIs and formats, including HTML, XML, SOAP, and REST.

A significant advantage of HPCC Systems is that data delivery is defined using the same declarative language (ECL) that is used to enrich the data. Applications can dip into any authorized data within the Data Lake. Using a common language and environment dramatically reduces the development time since it allows developers not only to present existing data but also perform further enrichment in cases where the existing data is not a perfect match for the application's needs.



DATA CURATION

Data curation — the management of information about the data lake — is an essential function for the success of any data lake. Data curation includes creating catalog information describing each dataset, its provenance, and its security requirements. It also includes the ECL definitions of the transformations that have been performed on each dataset, and the definition of the client APIs.

This information allows developers to find the information they need when developing an application. Just as important, developers can gain a greater understanding of the data and use the data with greater confidence, because they can see exactly how it was created and how it has been transformed. HPCC Systems uses the open-source Tombolo framework for its data catalog. While Data Lake architectures provide the capability to ingest data rapidly and transform it, tracking and documenting datasets (data dictionaries), capturing compliance (HIPAA, GLB, DPPA, GDPR etc.), managing compliance to transformation rules, recording lineage of data and ownership is a non-trivial task. Tombolo helps in keeping a record of all your assets in the Data Lake and how they are being used. This enables both developers and stakeholders (product owners and auditors) manage data assets quickly and efficiently.

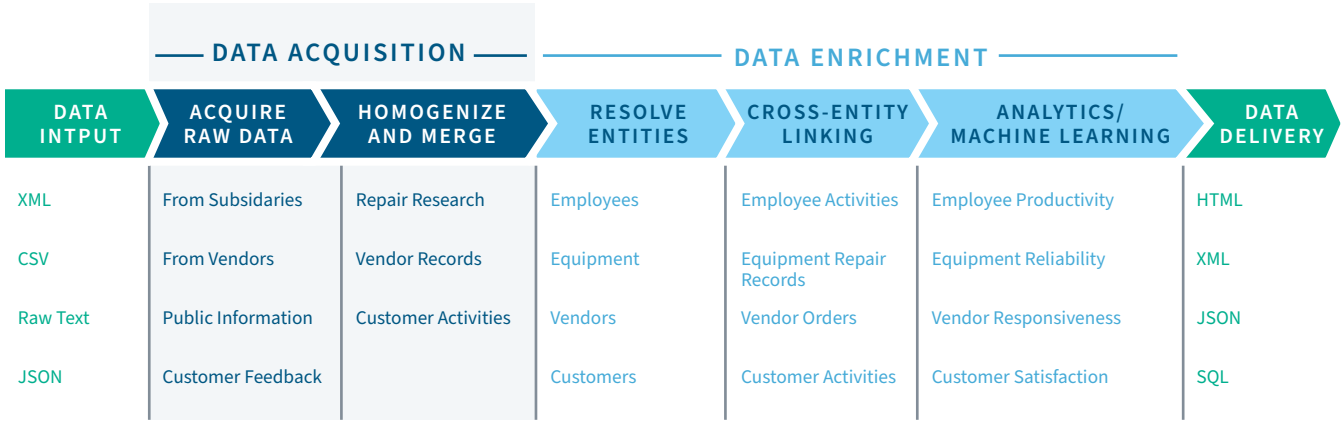
Example: How a maintenance company used a data lake to unlock the value in its information

Acme Services, an equipment maintenance and repair company, has expanded aggressively by acquiring other companies and allowing them to operate independently using their existing systems. Together, these subsidiaries now service more than 50,000 commercial customers and use equipment and parts from over 3,000 vendors. The subsidiaries hold a tremendous amount of data that could benefit Acme's executives, customers, and vendors, and potentially could be aggregated and resold. However, Acme couldn't harness this data because it was buried in silos within each subsidiary.

To solve the problem, Acme decided to use HPCC Systems to aggregate its vast information resources into a data lake. As a result, it can now rapidly respond to the information needs of its diverse stakeholders — and it's increasing revenue and profitability by unlocking the value in its data. A few of the benefits:

- Acme's executives can more easily analyze profitability and the source of revenue problems across all the company's subsidiaries and product lines.
- Marketing can set consistent pricing across all products and analyze factors that influence customer satisfaction and retention.
- Operations can identify reliability problems and increase field maintenance efficiency.
- Customers have a unified view of all their equipment from different vendors.

Here's how Acme created the data lake and uses it to deliver business value



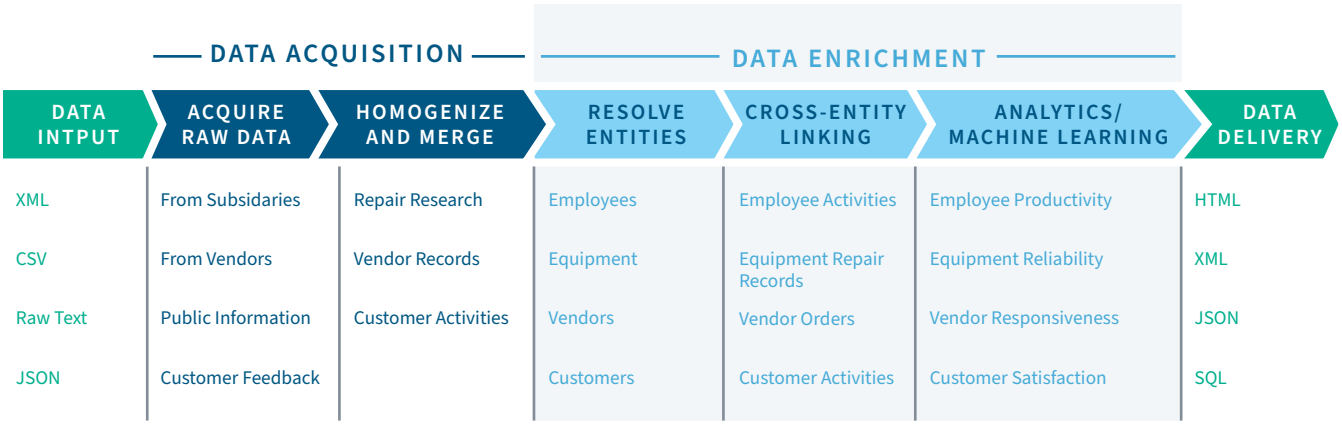
Data acquisition activities:

Acquire raw data

Related data comes from many different data sources in many different formats, with different fields and subtly different meanings for similar fields. Using HPCC Systems, Acme organizes the data, merges updates, and stores all original and updated data.

Homogenize and merge

Acme converts data from different sources into a single form with consistent meaning, creating a single stream of updates that can be processed in a consistent manner. Data cleaning actions include detecting errors and filling in missing information. These are intensive processes requiring complex transformations. Acme simply defines the required transformations using the ECL declarative language; HPCC Systems transparently manages the parallel processing required to execute them.



Data enrichment activities:

Resolve entities

Real-world entities do not typically have a single identity across diverse data sources. For example, a single customer may be known by different names, varying abbreviations, or stock symbols. Resolving these entities often requires multi-attribute or statistical matching. ECL provides the power and expressiveness to define these difficult resolution processes.

Cross-entity linking

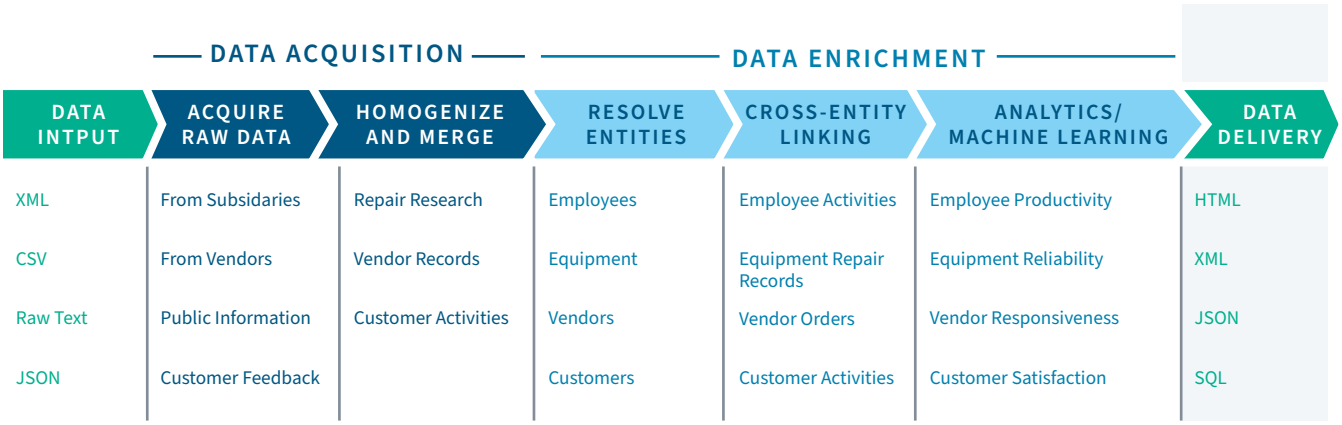
Understanding the relationships between different entities expands the value of the data exponentially. Different data sources typically reflect different relationships: one may include purchases from vendors, another tracks sales to customers, etc. With cross-entity linking, Acme can now see which vendors provided which products to which customers, and which salespeople sold them. All the transformations are visible and documented, and developers can use any interim transformation stages.

Analytics and machine learning

Resolving entities and understanding their relationships opens the doors to creating additional value through statistical analysis and machine learning. Acme uses the HPCC Systems robust library of analytic and machine learning algorithms that to analyze data, text, images, videos, and sensor data. By doing so, it can predict the likelihood of retaining a customer, determine the viability of vendors, project future revenues, and understand equipment reliability and failure patterns.

Indexing

Unlike relational data stores, HPCC Systems allows new indexes to be created as needed to accelerate information delivery for each of the diverse audiences that use the data lake.



Data delivery:

The information in the data lake is delivered to users in a scalable, reliable, secure, and high-performance fashion. Acme achieves this with the HPCC Systems information delivery engine (Roxie). Using a common language (ECL) and environment for both data acquisition and delivery dramatically reduces the development time. Acme delivers data through a variety of APIs and formats to meet the needs of executives, marketing, operations, and customers.

CONCLUSION

A data lake enables organizations with large and complex datasets to respond faster to changing business needs. Compared with traditional data warehouses, data lakes enable greater responsiveness to business groups and external customers, reduced costs, and greater scalability. They also accommodate a much wider range of data, including unstructured information and data from IoT devices.

HPCC Systems is the most proven data lake solution available. Used in production environments for more than a decade, HPCC Systems is a high-performance open-source platform that can scale to support very large numbers of users and datasets of unlimited size, enabling organizations to capitalize on big data for competitive advantage.

THE HPCC SYSTEMS DATA LAKE SOLUTION

Free and open source: You can test and implement HPCC Systems without making a big investment.

Proven and reliable: The most proven data lake solution available, used in demanding production environments for more than a decade.

High-performance: Processes data and queries at high speed. Lightweight core architecture supports batch, real-time and streaming information processing.

Standards-based: Runs on clusters of commodity hardware. Supports many tools and interfaces including SOAP, XML, REST, and SQL.

Scalable: Dataset size is limited only by storage capacity; supports very large numbers of users and clients. Scale easily and transparently by adding nodes.

Fault-tolerant: Highly available and redundant.

Easy to learn: Single data language, used across the platform, is easy to learn and enables high productivity.

Cloud-compatible: Runs on commodity hardware or in the cloud.

For more information, call 877.316.9669 or visit www.hpccsystems.com

HPCC SYSTEMS™

LexisNexis® RISK SOLUTIONS

About HPCC Systems®

HPCC Systems® from LexisNexis® Risk Solutions is a proven, comprehensive, dedicated data lake platform that makes combining different types of data easier and faster than competing platforms — even data stored in massive, mixed schema data lakes . It's also open source, free to use, and easy to learn. You can acquire, enrich, deliver and curate information faster using HPCC Systems — and the automation of Kubernetes in our cloud-native architecture makes it easy to set-up, manage and scale your data to save time and money, now and in the future. HPCC Systems offers a consistent data-centric programming language, two processing platforms and a single, complete end-to-end architecture for efficient processing. To learn more, visit us at hpccsystems.com.

About LexisNexis® Risk Solutions

LexisNexis® Risk Solutions harnesses the power of data and advanced analytics to provide insights that help businesses and governmental entities reduce risk and improve decisions to benefit people around the globe. We provide data and technology solutions for a wide range of industries including insurance, financial services, healthcare and government. Headquartered in metro Atlanta, Georgia, we have offices throughout the world and are part of RELX (LSE: REL/NYSE: RELX), a global provider of information-based analytics and decision tools for professional and business customers. For more information, please visit www.risk.lexisnexis.com and www.relx.com.