HPCC™ SYSTEMS

LexisNexis® RISK SOLUTIONS

WHITE PAPER

Keep the TCO of Your Big Data Platform Under Control with HPCC Systems

INTRODUCTION

In today's enterprise, a successful big data strategy can mean the difference between success and failure. For example, Netflix reports the company is able to save $1 billion a year from customer retention thanks to its use of big data analytics, and enterprises in every other vertical market are following suit. Market research firm Statista forecasts big data analytics software spending will hit $68 billion by 2025. But adopting a big data strategy is a big undertaking for enterprises, and there are a host of questions an IT team must answer before they can decide on the best big data platform for their needs. Should the enterprise use an on-premises or cloud-based datacenter for data analytics? Which data analytics software best fits the organization's use case? Does the IT team have the requisite experience and expertise to implement a big data solution? Will the chosen big data platform still meet an enterprise's business needs in 12 months? What about in 5 years?

This paper will examine multiple criteria an enterprise IT team should consider before they adopt any big data platform. By rigorously evaluating a potential platform in each of these categories, IT teams will have a better understanding of the total cost of ownership (TCO) of their chosen platform. The paper will then apply each of those criteria to reveal how well an HPCC Systems data lake platform addresses the criteria. Finally, the paper will examine how an actual HPCC Systems customer evaluated HPCC Systems TCO and decided the platform was the best fit for their big data needs.

For more information about the performance and features of the HPCC Systems platform, please read the Taming the Data Lake: The HPCC Systems Open Source Big Data Platform whitepaper.

For a comprehensive comparison between the HPCC Systems platform and other big data solutions, please read Understanding HPCC Systems and Spark: A Comparative Analysis.

About HPCC Systems

HPCC Systems is an open source data lake platform designed to continuously acquire data from many data sources in both structured and unstructured formats. Data lakes are ideal for use with big data applications because they support extremely large, complex, and diverse datasets, and they easily accommodate new data sources such as IoT. They allow IT groups to quickly create new applications that support changing business needs by unlocking the power of complex data for all users within the organization. They also scale more easily and cost-effectively than relational databases and offer the huge storage and compute resources needed for data analytics. As a result, data lakes enable greater responsiveness for users and external customers, reduced costs, and greater scalability.

A typical HPCC Systems implementation begins with just a few data sources and some initial analytical and reporting tools, but the size, complexity, and capability of the HPCC Systems data lake can grow quickly. Once data is added to the data lake, the process of data enrichment begins. Data enrichment is an evolving, iterative process that extracts as much knowledge as possible from data sources. Once that knowledge is extracted, it's available to other data lake users that need it via a process known as data delivery. During data delivery, HPCC Systems ensures that data is transferred to data lake users in a responsive, secure, and reportable manner.

The illustration on the next page captures the data lifecycle of an actual HPCC Systems data lake currently in use. Moving left to right, the data sources deliver data to the HPCC Systems data lake for refinement, enrichment, indexing, and analysis. HPCC Systems can generate reports or dashboards about the data at any step in the process, depending on what the report's consumer needs. All of these processes occur within the data lake environment to produce consistent results.

HPCC Systems Data Enrichment Pipeline

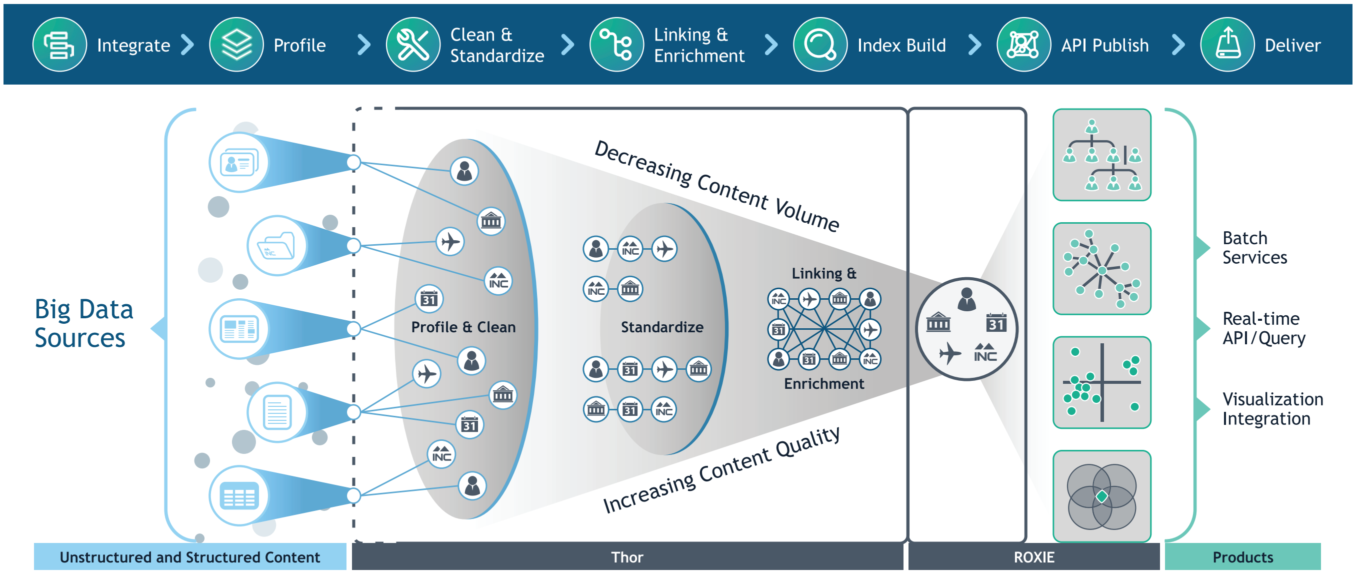

Illustration 1: The HPCC Systems data pipeline above follows data from the source, to its ingestion into the HPCC Systems cluster, where it is formatted, enriched, and then made available to applications hosted in the cluster.

An HPCC Systems Data Lake is comprised of three primary components:

- The **ECL programming language**, which has been in use since 2000, is a data-oriented, declarative programming language developed for use on HPCC Systems data lakes. In layman's terms, ECL lets developers tell the system what they want, but leaves it up to the system to determine the best way to go about doing it. This results in a smaller, more efficient code base.
- **Thor** is a bulk data processing cluster that cleans, standardizes, and indexes inbound data for use by the data lake. Once data has been refined by Thor it can then be used by the Roxie cluster.
- **Roxie** is a real-time API/Query cluster for querying data after refinement by Thor. Roxie queries execute in sub-second times and provide for very high concurrency.

The HPCC Systems Engines

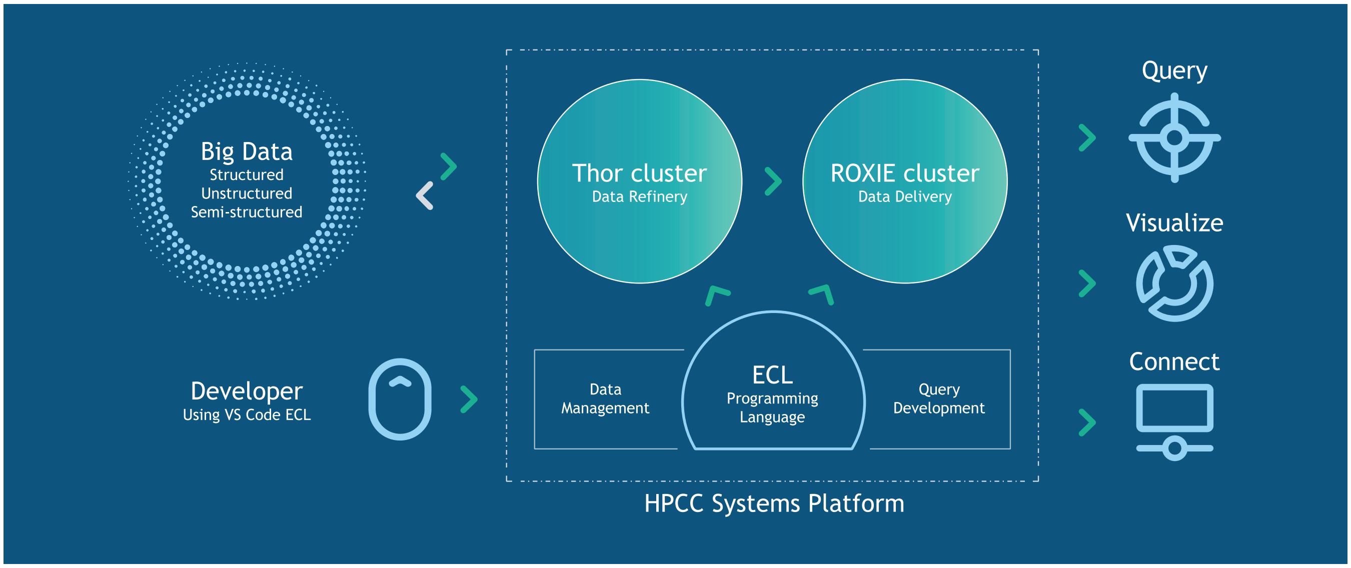

Illustration 2: An HPCC Systems diagram featuring a Thor cluster (for bulk data processing) and a Roxie cluster (for handling data queries)

Establishing Total Cost of Ownership for a Big Data Platform

In addition to confirming a potential big data platform has the performance and features needed, IT teams should also consider other factors if they are to gain a true understanding of the capital and operational expenses required to keep a big data platform operational and scalable.

The list below captures twelve criteria to consider. There may be additional factors to consider based on the specific needs of the enterprise looking to adopt a big data platform.

1. **Fees for seat licenses** – Proprietary big data platforms require enterprises to purchase licenses, often sold on a per seat basis. As the platform grows and matures, additional IT staff and seat licenses may be necessary, so license fees could increase over the long term.

2. **Technical support** – If the platform isn't performing as required, enterprises need access to technical support resources that can quickly and effectively solve the problem. But identifying which vendor is responsible for providing a technical fix in mixed-vendor datacenter environments is difficult, and it's common for customers to be stuck without a fix for their problem because their vendors can't agree which component of the data lake is at fault. Technical support is particularly important for security. If a platform is orphaned by its vendor, security patch development ceases, leaving the platform's users and data at risk.

3. **Third-party software** – Many big data platforms are sourced from multiple vendors; each vendor providing a solution for a specific stage in the big data pipeline. Enterprises need to be sure they understand all the software they'll need to install and support to get the functionality they require.

4. **Hardware** – The amount of processing and storage hardware a big data platform needs will vary over time. Many businesses have seasonal data workloads that rise and fall throughout the year. This can lead to overprovisioning: purchasing extra hardware to manage spikes in datacenter capacity only to watch that same hardware go unused during periods of low activity. Furthermore, more power efficient processor architectures are now available to help enterprises keep datacenter electricity consumption low but using them is only possible if their datacenter software supports these new architectures.

5. **Cloud support** – Cloud-based data centers allow enterprises to scale their compute and storage capacity up or down in real time to keep overprovisioning under control. That said, cloud computing has security risks that may require specific security capabilities to meet regulations and SLAs around data privacy, sovereignty, and security, and IT teams will need to make sure any cloud-based big data solutions comply with those requirements.

6. **Staffing** – IT managers need to determine if their choice of big data platform will require adding additional staff to address any skill gaps or increase code output. What skills will those new hires need? As big data grows in popularity, potential hires with expertise in big data and cloud computing will be in high demand and their salaries will reflect this. IT managers must keep staff payrolls in mind when considering a big data platform's TCO.

7. **Implementation time** – After a platform is selected, how long will it take to get the platform up and running? Sourcing software and hardware from different sources can cause compatibility issues that must be addressed before the platform goes live, potentially delaying the platform's launch date.

8. **Ongoing maintenance** – Once a big data platform is operational, how much ongoing opex will it cost to keep it running? How much electricity does it consume? Will the datacenter require more square footage to accommodate additional hardware? If the platform's processing and storage capacity need to expand in the future, how long will that expansion take and how cost-effective will it be?

9. **Flexibility** – If an IT team requires its big data platform to support specific features, what resources are available to provide that feature if the platform's vendor is unable or unwilling to build it?

10. **Developer ecosystem** – Is there a robust, global network of developers working on value-added projects for the platform? Does an enterprise need their big data platform to support a specific vertical industry? Or a particular application? The larger a big data platform's developer community, the more likely software for specific industries or use cases is already available.

11. **Reliability/maturity** – Is the platform's technology new and without extensive real-world testing? Is the vendor a startup who may not be around to support their technology, or currently not able to scale to meet demand for support? Do they have good technology AND good customer service? Can they provide localized support resources for different regions?

12. **Data support** – Does the platform process data in any format? Does data in different formats work well together? Data in different formats often end up siloed in separate databases that don't communicate with one another, which can lead to inaccurate or incomplete data analysis.

A TCO Analysis of HPCC Systems

Now let's revisit the TCO criteria list and see how HPCC Systems addresses each one.

1. **Fees for seat licenses** – As an open source software platform (released under v2.0 of the Apache License), HPCC Systems is available to use by any organization without a licensing fee.

2. **Technical support** – As HPCC Systems is a complete data lake platform, its software supports all stages of the data processing pipeline. Accordingly, support is available for the entire platform from HPCC Systems global network of developers and system consultants (Infosys, for example). HPCC Systems developers also maintain multiple active community forums and a JIRA database for cataloguing bugs/fixes and new features added to the platform.

3. **Third-party software** – HPCC Systems is a complete big data solution, but if an enterprise needs additional software not currently available in the HPCC Systems stack, it can source the needed software from HPCC Systems global developer network. This open source model has been used by other technologies with great success. For example, Google developed the Kubernetes code base and made it available to a global network of developers so they could continue to expand and improve upon it, and Google would be able to leverage the developers' innovations on its own cloud platform.

4. **Hardware** – There is interest among enterprises in leveraging alternative processor technologies to help keep power consumption and hardware costs low. Fortunately, HPCC Systems is compatible with alternative processor architectures like ARM and RISC-V. In fact, HPCC Systems was originally designed to run on commodity hardware solutions to help keep implementations of the platform more cost effective. And as available memory is a key aspect of cluster configuration, HPCC Systems focuses on available memory to provide a more accurate estimate of necessary compute resources than other platforms, which means there is less risk of hardware overprovisioning.

5. **Cloud support** – HPCC Systems has offered cloud support since 2020. Cloud-based instances of HPCC Systems provide hardened security, end-to-end encryption, authentication, authorization, and other important security measures. The goal is to deliver state-of-the-art security features that protect customer data in the cloud, while still providing the excellent data management and analysis performance seen in on-premises HPCC Systems deployments.

6. **Staffing** – Because HPCC Systems uses the ECL language across the entire data workflow, it's possible to maintain an HPCC Systems data lake using a smaller IT team. Furthermore, since ECL is more efficient for coding data-related tasks, a smaller development team can generate more data analysis algorithms and reports than a larger team using a less efficient programming language could in a similar time frame.

7. **Implementation time** – The completeness of the HPCC Systems data workflow means users can deploy a complete data lake platform in just a few hours. Creating a data lake with capabilities like HPCC Systems requires a combination of individual solutions from different vendors that could require weeks of compatibility testing before it is fully operational.

8. **Ongoing Maintenance** – Because HPCC Systems can run on commodity hardware and requires a smaller cluster footprint in comparison to competing platforms, if an HPCC Systems data lake needs additional processing or data storage capacity, that capacity can be added at lower cost. Additionally, the platform has native support features like user authentication, authorization, reliability monitoring, performance monitoring, and job management that make managing the data lake less burdensome in terms of time and number of IT staff needed.

9. **Flexibility** – The HPCC Systems roadmap uses a monthly feature release plan that consists of release candidates and gold releases. As the platform is open source, outside developers can contribute directly to the feature release plan. This balance creates a flexible development model to quickly bring new features to market.

10. **Developer ecosystem** – HPCC Systems has a global developer network actively working to customize the platform for use in a range of vertical markets, including agriculture, finance, healthcare, insurance, IoT, and others.

11. **Reliability/maturity** – The HPCC Systems data lake platform was created in 1999. Since then, there have been numerous successful deployments of the platform that serve as proof of its reliability and efficacy. LexisNexis Risk Solutions and their customers use HPCC Systems to execute production workloads daily using a platform that is battle tested and production ready. Several of these deployments are detailed in case studies available on the HPCC Systems website.

12. **Data support** – HPCC Systems is very adept at handling and storing data, no matter the format (CSV, XML, JSON, plain text, and binary files are all supported). Raw data files are maintained in standard Linux filesystems, and data storage is implemented via a modernized version of Indexed Sequential Access Method (ISAM) files to quickly create search indexes that make it easier to find individual records within that dataset.

Case Study: DataSeers

Founded in 2017, DataSeers is a B2B SaaS company with a mission to address the challenges experienced by the banking and payments industry. Instead of updating legacy banking software built for the age of checks and wire transfers, DataSeers reimagined back-office software from the perspective of ACH payments and prepaid cards. More specifically, DataSeers looked into the key pain points of fraud, compliance, and reconciliation, eventually broadening this to include KYC/KYB and onboarding.

Adwait Joshi, CEO and founder of DataSeers, was intimately involved with the development of his company's big data platform. Joshi and his team evaluated several different options before selecting HPCC Systems.

Joshi cited several factors that informed DataSeers' decision to go with HPCC Systems: the open source license, ease of setup, support for commodity hardware, the ECL programming language, its flexibility, and its support for multiple data formats.

> "When we were first starting up as a company, our budgets were very tight. TCO was a critical part of our decision making process around which big data platform we chose. The fact that HPCC Systems was available for free through its open source license made it very compelling to us."

"Setting up our first HPCC Systems cluster took very little time; I believe it was up and running in a matter of hours. We were also thrilled to discover we could run HPCC Systems on commodity hardware, which also helped our budget. And now that more of our customers are using cloud and Kubernetes, we're happy to report to our clients looking to leverage the cloud that HPCC Systems runs very well on bare metal servers."

Joshi said there was a bit of a learning curve when it came to his team learning how to code in ECL, but it was well worth their time. “ECL is a very powerful language for data analysis. After a few days working with ECL, the team quickly understood that if they could master about six key processes, there was no program or application they couldn’t develop. ECL is also helpful in that the entire HPCC Systems stack uses ECL. This means it only takes one DataSeers programmer to get a complete HPCC Systems environment up and running for a client. This helps keep our headcount low and profit margins high.”

Joshi praised HPCC Systems flexibility and compatibility with other big data technologies. “We use other technologies besides HPCC Systems in our big data environment, like Elasticsearch, for example. But they all work well with HPCC Systems. And if there’s a particular application or feature not readily available in HPCC Systems, I still have the option of leveraging solutions developed in C++ or Python. I can embed their code into ECL to easily integrate them into our HPCC Systems environment.”

Finally, Joshi credited HPCC Systems ability to manage multiple data formats as a deciding factor in their choice of HPCC Systems. “With HPCC Systems support for structured and unstructured data, we can ingest and format data of any type and use it to help inform our analysis. CSV, XML, JSON, plain text, or binary files? Doesn’t matter. HPCC Systems Thor cluster easily ingests, formats, and enriches all of it, regardless of file type.”

CONCLUSION

As enterprises look to leverage the business benefits of big data, they must be selective in their choice of big data platform. In addition to vetting a platform's technical specifications and features, enterprises should also consider a host of other criteria to gain a more informed understanding of how well a platform meets their specific needs and what its total cost of ownership will be. By carefully evaluating a big data platforms strengths and weaknesses, businesses can choose a platform that best fits their data use case today and in the future.

For more information, visit www.hpccsystems.com

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 covers 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 includes seven brands that span multiple industries and sectors. We harness the power of data, sophisticated analytics platforms and technology solutions to provide insights that help businesses and governmental entities reduce risk and improve decisions to benefit people around the globe. 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 LexisNexis Risk Solutions and RELX.