HPCC Systems: Powerful Open Source Big Data Analytics Platform

Born from LexisNexis® Risk Solutions deep data analytics history, HPCC Systems® helps businesses of all sizes find the answers they need by making data easier to process, analyze, and understand.

HPCC Systems platform is a set of easy-to-use software components enabling developers and data scientists to process and analyze data at any scale. With a strong commitment to the open source community, the HPCC Systems platform is available free of licensing and service costs.

HPCC Systems is cost-effective, comprehensive, fast, powerful, and scalable. Ultimately, it makes managing big data easier.

The HPCC Systems stack is comprised of:

**ETL** - Extract, Transform, and Load your data using a powerful programming language (ECL) specifically developed to work with data.

**Data Management Tools** - Data profiling, Data Cleansing, Snapshot Data Updates and consolidation, Job Scheduling and automation are some of the key features.

**Query and Search** - An indexed based search engine to perform real-time queries. SOAP, XML, REST, and SQL are all supported interfaces.

**Predictive Modeling Tools** - In place (supporting distributed linear algebra) predictive modeling functionality to perform Linear Regression, Logistic Regression, Decision Trees, and Random Forests.

---

**HPCC Systems. It's easier:**

- Easier to update
- Easier to learn
- Easier to program
- Easier to integrate data
- Easier to manage clusters

End to end big data in a massively scalable super computing platform

---

**Features:**

- Standard hardware, operating system, and protocols
- High redundancy and availability
- Practical tools and extensions
- Efficient programming with a declarative, modular, extensible language
- End-to-end configuration
- Optimized distributed file systems (DFS)
- Massive scalability and performance
Thor - Big Data ETL Made Easy

Thor executes data manipulation scripts to transform and analyze raw data.

Known as "Thor" after the hammer-wielding god of thunder, this cluster is designed to execute big data workflows including extraction, loading, cleansing, transformations, linking, and indexing.

Thor uses a master-slave topology in which slaves provide localized data storage and processing power, while the master monitors and coordinates the activities of the slave nodes and communicates job status information.

Middleware components provide name services and other services in support of the distributed job execution environment.

ROXIE – High Performance Real-Time Queries

ROXIE is an index based search engine to perform real-time queries.

The ROXIE rapid data delivery cluster provides separate high-performance online query delivery for big data. ROXIE (Rapid Online XML Inquiry Engine) utilizes highly optimized distributed B-tree indexed data structures conceived for high concurrent use.

Each ROXIE node runs a Server process and an Agent process. The Server process handles incoming query requests from users, allocates the processing of the queries to the appropriate Agents across the ROXIE cluster, collates the results, and returns the payload to the client.

Queries may include joins and other complex transformations, and payloads can contain structured or unstructured data.