Whitepaper

Scaling Data Science Capabilities - Leveraging HPCC Systems® to Build a Homogeneous Big Data Ecosystem

An Industry Insight series from ClearFunnel®, a Big Data Analytics as-a-Service provider

Authors:

Rohit Verma, CEO/Co-Founder of ClearFunnel, LLC, rohit@clearfunnel.com

Raj Chandrasekaran, CTO and Chief Data Scientist/Co-Founder of ClearFunnel, LLC, raj@clearfunnel.com

September 2018

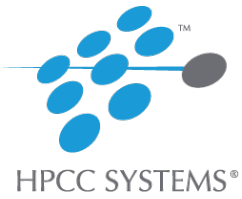

Table of Contents

Note from the authors: We are grateful to the leaders at HPCC Systems who have consistently provided the guidance and support during the last several years in developing our expertise and experience in advanced applications of Big Data and Data Science technologies.

Summary

Big Data is not a new phenomenon but its adoption, execution, and successful ecosystem management continue to challenge enterprises, big and small. Fast forward: the fusion of three major technological revolutions of our time: Big Data, Machine Learning, and Cloud capabilities have vastly expanded the demand from businesses for all encompassing data lakes, sophisticated analytics, as well as opened the doors to developing new Artificial Intelligence (AI) based solutions in every sector of the economy.

However, while organizations across the board continue to struggle to build enterprise-wide data lakes and capture value from data analytics, there is also a plethora of tools and technologies flooding the market, which make it even more complicated and unwieldly for businesses to take advantage of Big Data and Deep Learning capabilities.

This study is a collaboration between HPCC Systems and ClearFunnel for bringing into focus the real world, multi-year experience of a cloud-based Big Data and Data Science startup in successfully building an advanced analytics business based upon using a homogeneous technology stack.

This study aims to help organizational leaders understand the direct impact that choice of Big Data and Machine Learning technology stack has on the successful execution of their Big Data and Advanced Analytics vision.

1. Introduction

The more widely known story of Big Data processing started with Map Reduce. Internally, Map Reduce needs to read and write the file from the persistence store during each operation. Then came along Spark, which avoided the read and write by keeping the data in memory. Spark also introduced something which the Spark community called 'cool' – the Directed Acyclic Graph (DAG), which basically allowed you to process and navigate a graph, but with the caveat - as long as the graph data did not have any circular reference. The reason I start with reference to Map Reduce, Spark, and DAG is because all these capabilities are really version 1.01 in the HPCC Systems world. The elegance of HPCC Systems lies not only in the fact that it was already in production use before even the Map Reduce whitepaper was introduced, but also primarily because for a Big Data practitioner, HPCC Systems represents a paradigm shift in Big Data engineering. With the perfect combination of declarative programming, object-oriented organization, expressive syntax of its ECL language, native and powerful data transformation capabilities, and masking of the complexities from the programmer of how the distributed computation is executed, HPCC Systems varies from Hadoop where the user needs to care about map-shuffle-aggregate-reduce complexities through all steps in the data preparation and processing operations.

2. Machine Learning and Advanced Analytics

Today, the Big Data industry has scaled-up to incorporate Machine Learning based computations within the Big Data platforms. Traditionally, R and Python were widely used languages to express Machine Learning algorithms. Spark provided support for these languages by offering what it seemed as a quick gateway for processing Machine Learning libraries at scale. The only small problem was that the R and Python constructs needed to 'conform' to the Spark paradigm before existing code written in R and Python would successfully execute on a Spark cluster. The other problem with this approach is that Python, which is primarily a middle-tier scripting language retro-fitted with OOP capabilities, is now being used for complex data manipulation, transformation, and computation problems. In contrast, ECL is designed from the ground-up for handling complex data engineering scenarios with built-in native support for almost all types of data computation and data manipulation operations. In the HPCC Systems world, there is no need to fit the data into a specific paradigm before executing complex data transformation and computation operations. Additionally, people often point to writing Spark applications in Scala and that brings-up two more issues: (1) the need to master multiple skills to be able to operate a Big Data ecosystem outside of HPCC Systems, and (2) inability to use the proven Python libraries of NumPy or scikit-learn that simply aren't in Spark.

3. Cloud adoption

After Machine Learning and Advanced Analytics, the next stage of revolution in Big Data affairs has been the maturity of Cloud and its industry-wide adoption, including for deploying enterprise-wide, self-managed big data lakes. The 'big daddy' of cloud service providers – Amazon Web Services ("AWS") started providing data adaptors for faster loading of data to Map Reduce clusters. Given that Map Reduce itself is built over JVM and AWS accelerators are built using a similar category of programming languages (Perl and Python), there are considerable overheads due to layered encapsulations in these utilities which are used to operate Map Reduce clusters on cloud. Here again, the simplicity of HPCC Systems becomes its strongest advantage. HPCC Systems is devoid of any fluff, it is purpose built for massive data operations, and it natively supports C++ libraries. These advantages of HPCC Systems has helped ClearFunnel to build APIs (right from ECL itself and also using C++) to achieve I/O speeds of reading/writing between AWS S3 storage and HPCC Systems cluster deployed on EC2 instances faster than what can be achieved using existing Map Reduce and AWS accelerators. In a medium sized test bed, ClearFunnel has successfully achieved I/O speeds of 2 TBPS in transferring data between AWS S3 and HPCC Systems clusters. That is a lot of I/O speed which is of significant advantage in efficiently processing several hundred terabytes of Big Data. With ClearFunnel's framework and design of extending HPCC Systems capabilities using AWS and by implementing innovative methods to spray/despray files from/to S3, the I/O speed between HPCC Systems (on EC2 instances) and S3 can be even further increased to the point where the issue of file I/O bottlenecks becomes irrelevant in Big Data performance considerations.

4. Speed of deployment

For running a SaaS business like ClearFunnel, the simplicity of solution engineering and speed of deployment becomes a key success factor. Key operational requirements in our business are to successfully execute complex Big Data and Machine Learning use cases on tight budgets and build and operate several advanced analytics use cases at Big Data scale for various clients at the same time. One of the inherent advantages present in HPCC Systems that comes in very handy in managing this pace and complexity is that its core architecture is very light-weight. HPCC Systems does not have multiple layers of complex 'add-ons' like Zookeeper, MR, Hive, Impala, HBase, Yarn, Mesos, RDD, Cluster Manager, DFS, GraphX, MLib, SparkSQL, and others. These add-ons only make the whole Hadoop and Spark technology stacks and production deployment complex, time consuming, and very expensive to maintain – not to mention the myriad variety of talent that is needed to design, build, operate, and maintain such a complex cluster. In contrast, the HPCC Systems architecture is extremely nimble and agile and offers the right level of flexibility to perform the full spectrum of Big Data and complex Machine

Learning operations without requiring any of the countless add-ons to implement new data engineering features. The biggest advantage of its architecture is the speed and cost of operations that it provides for companies and use cases of all sizes. Consider this – a fully functional, production-grade HPCC Systems cluster consisting of several hundred nodes can be spun-up within a few minutes ready to fire away at the most complex and demanding data engineering tasks, all of which can be easily programmed and managed with just ECL!

5. Use cases

Some of the capabilities that make HPCC Systems a perfect candidate for a full-spectrum of data engineering use cases:

5.1 ECL lends itself very elegantly to abstraction.

This provides virtually endless possibilities of extending HPCC Systems capabilities, such as:

i. Building fully-parallel machine learning libraries,
ii. Large graph traversal across nodes,
iii. Image processing,
iv. Operational and Monitoring analytics,
v. Bioinformatics and genetic dataset processing,
vi. NLP and advanced text processing,
vii. Near real-time stream processing,
viii. IOT based sensor-data integration and analytics,
ix. Advanced search and querying – there really is no equivalent in other Big Data stacks to ROXIE's B-Tree based massively parallel, high-volume, fast querying engine. To give you an example: Some of our ROXIE indices in production clusters are 12+ TB in size and the ROXIE still returns results for queries with sub-second speed!
x. Seamless integration with cloud dev ops.

5.2 Near real-time stream processing

Let us explain one near real-time stream processing extension that we at ClearFunnel have built on HPCC Systems to support micro-batching for Big Data. Leveraging the native hooks provided by ECL and the micro-services of AWS cloud, we have designed and implemented a micro-batching solution that processes incoming stream data and updates query indices in ROXIE with 10-second intervals. This capability includes automated roll-up of sub-indices, coordinated housekeeping of super-files and dependent sub-files, and hot-swap of new index version in true Zero Downtime Deployment (ZDD)

mode. For all the near real-time use cases that we support, with only a couple of exceptions, a 10-second interval for analytics update is an acceptable outcome for our clients' businesses. Using HPCC Systems, we have successfully tested micro-batching runs of less than 10-second update interval, but the cost involved in running such a solution at scale, 24/7, and to keep the clusters and indices in-sync usually defeats the business case of operating at such a narrow time window. Our key learnings here are:

i. Any type of micro-batching system at Big Data scale that operates at less than 10-second update interval leads to other issues, which we have also seen in Spark. Even though HPCC Systems is not the suitable architecture for near real time analytics, a micro-batching solution with 10 – 15 second update time window can be cost-effectively implemented and continuously operated using HPCC Systems and without introducing a new technology just for this purpose in the Big Data solution stack.

ii. We have consolidated the results of our micro-batching test cycles into a summarized lookup table that effectively maps delta (incremental) index sizes to an optimal micro-batching frequency. This is a key reference table used in our fully autonomous dev-ops routine, where based on the size of an index in each micro-batch processing cycle, the system dynamically determines the optimal wait time before hot-swapping a specific ROXIE index during that update cycle. This avoids race conditions and other random micro-batching issues and maintains the sub-second response time serviceability of the high-volume, concurrent query requests.

5.3 Code re-use as a paradigm

In an ECL job, every data operation step is a 'graph' – essentially, a string of graphs constitute a pipeline. This is a great model when it comes to either extending the pipeline, breaking the pipeline into sub-parts for more efficient management, executing different graphs on different sized clusters for maximizing cost efficiency, and integrating the pipeline with third-party capabilities like IOT and cloud-provided micro-services, etc. This simple, yet a very powerful paradigm works great for rapid solution deployment using HPCC Systems because code re-use (read 'graph re-use') and its adaptations across different data processing pipelines becomes very easy to design and implement.

5.4 Fast data transfer between clusters of different sizes

The concept of operating different stages ('graphs') of the same data processing job on different-sized hardware clusters is unique to HPCC Systems. The seamless integration with Cloud services like AWS S3 storage and EC2 compute instances allows for very fast data transfer between clusters of different

sizes to process various graphs of the same overall job. This model, apart from being orders of magnitude more cost effective as compared to operating all the stages ('graphs') of a job on the same-sized cluster, is also a great fault-tolerance architecture. With this execution model, immediate recovery from a cluster failure at any stage of the job is always guaranteed, is blazing fast, and can be easily managed using automation.

6. Enabling fail-fast, fail often

Fail-fast, fail-often: In the Big Data Analytics world, a platform that efficiently supports this concept becomes a key driver of success.

i. By its very nature, solutions for complex Big Data and Machine Learning based analytics use cases are iterative and, therefore, they evolve and refine over time. This characteristic necessitates redesigning solutions multiple times over, rewriting pieces of code several times, and sometimes overhauling the entire solution a few times.

ii. SaaS based Big Data Analytics service providers, like ClearFunnel, build and operate several custom use cases for multiple clients across different industry domains and solution landscape. To remain agile, responsive to changing client needs, and to incorporate best Big Data engineering practices at all times, we often need to redesign our solutions several times during every solution's lifecycle.

iii. In the SaaS model of service delivery, ClearFunnel not only builds the initial Big Data Analytics solutions for its clients, but also continues to maintain, enhance, and operate those solutions throughout their respective lifecycles. With this end-to-end responsibility of managing the Big Data solutions, ClearFunnel has to constantly innovate and refactor its code base to not only take advantages of advancements in Big Data, Cloud, and ML technologies, but also to provide the best value for money to its clients for their custom use cases.

All of these above-mentioned scenarios require a lot of constant code refactoring behind-the-scenes, which would have been seriously complicated, cost-prohibitive, and an endlessly time-consuming exercise in any existing Big Data technology stack other than HPCC Systems.

7. Operating at scale with minimal costs

Let us cover the aspect of 'cost of operations' of Big Data ecosystem in a little more detail:

7.1 Platform simplicity requires fewer specialized resources

When ClearFunnel operates multiple massive clusters of several hundreds of nodes in each cluster, it involves serious Big Data computing for mission-critical, revenue-generating business applications of its clients. If these solutions were on Spark or Hadoop-based architectures, any business would have to deploy a sizeable team of Big Data and infrastructure engineers to deal with:

i. Numerous memory issues,
ii. Small files-related problems in HDFS and Spark,
iii. Expensive repartitioning and reshuffling of RDDs in Spark,
iv. Random debugging issues with large streaming pipelines operating non-stop at scale 24/7, and
v. Other indiscriminate errors.

Solving all of these requires intimate knowledge and experience of everything from the implementation language down to the OS kernel.

Compare this to the small engineering team at ClearFunnel, which has no dedicated cloud engineer, infrastructure engineer, network engineer, production support engineer, dev ops engineer, or tech ops specialist even though ClearFunnel is continuously operating scores of Big Data clusters at massive scale for several clients. This only goes to prove the industrial-grade robustness of HPCC Systems, which has been in production use supporting billions of dollars of enterprise businesses at LexisNexis Risk Solutions as well as other government and commercial corporations for more than a decade.

Nothing comes close to the simplicity and the dependence that HPCC Systems provides for continuously operating at scale with minimal costs and hand-holding.

7.2 Reduced server footprint reduces infrastructure costs

Additionally, given the architectural homogeneity and simplicity of HPCC Systems, the server footprint of a HPCC Systems based solution deployment is comparatively lesser than the footprint of a complex Spark or Hadoop based cluster deployment. This may seem like a small difference for a single use case, but this advantage quickly increases and results in big infrastructure cost difference when multiple solutions are operated at maximum scale.

8. Extending ECL language to support proprietary algorithms

ClearFunnel works with many startups that have their own secret sauce and proprietary algorithms to win in their respective markets. These models cannot be replicated using standard machine learning models and classifiers. To be able to seamlessly implement these use-case specific and very custom ML models using ECL becomes a quick and simple effort given the rich set of data processing constructs and modular organization already available in ECL. Not only are we able to recreate those complex, custom algorithms in ECL, but ECL then automatically executes it at scale leveraging the power of the HPCC Systems distributed computing capabilities.

9. High Availability techniques for HPCC Systems clusters (leveraging AWS capabilities)

With some of ClearFunnel's ROXIE indices in a single use case over 12+ TB in size, we had to design a reliable and a cost-effective recovery approach in case of failures of both, the primary and the backup clusters. At a nominal add-on cost, this powerful High Availability capability provides our clients with an assured recovery time window for their critical production use cases. Based on the cost and recovery time objectives, there are four design options that ClearFunnel has devised to implement High Availability for ROXIE clusters:

i) Files in S3: Store sorted and pre-split files in S3. To execute Disaster Recovery, bring-up a Thor cluster, build indices and push to ROXIE. This will take anywhere from 2 – 20 hours depending on the number and size of the indices, but it provides almost a zero cost (excluding the S3 storage cost) option for High Availability and is recommended for non-critical queries.

ii) Replicate production indices in a separate EBS volume (or volumes, based on cost). The recovery procedure involves a simple restore of indices from the replicated EBS volume attached to a new ROXIE cluster. This reduces the recovery time to between 10 – 40 minutes.

iii) The entire production EBS volume (containing only the ROXIE indices) is binary-backed up on S3. Since ROXIE index files cannot be separately copied or backed-up, this provides a cost beneficial recovery option for ROXIE indices without keeping a separate EBS volume live for replication purpose.

iv) Maintain two ROXIE clusters (in Active-Passive) configuration. This option provides sub-second Disaster Recovery, recommended for critical queries, and includes the highest cost. Based on actual use case requirement and costs, the Passive ROXIE cluster can be a slightly smaller infrastructure as compared to the Active ROXIE cluster and in the event of a disaster, while the lower-capacity Passive cluster holds the fort and avoids service outage, a new Active cluster on a full-scale infrastructure can be rebuilt within a couple of hours using backed-up files or replicated EBS volumes to take-over from the Passive cluster.

10. General-purpose nature and suitability of ECL

This is a key aspect in this discussion about using HPCC Systems for a broad set of use cases without having to complicate the architecture stack to meet each and every technical requirement of any complex data engineering use case. Consider the following example:

i. One of ClearFunnel-implemented solutions required extreme Graph traversal at one specific step in the overall analytics pipeline. The specific engineering need in that step was to recursively mine a Graph of several hundred billion nodes and vertices and so, the query needed to very quickly go very deep in this massive Graph dataset. The solution to this use case was similar to the one provided by the Dijkstra's algorithm, which ClearFunnel was able to quickly (and with relative ease) implement it grounds-up in ECL. The reason to mention this use case is not to compare ECL's performance with other commercial Graph traversal solutions, but to highlight the point that ECL is general-purpose enough to tackle a range of complex Big Data engineering scenarios (including large-scale, Graph traversal use cases), which may otherwise need dedicated Graph databases and specialized programming skills.

ii. In this case, given the cost-benefit analysis of adding a new technology into an otherwise relatively homogeneous and simple stack, it was more cost-effective and a lot quicker to write the solution in ECL by leveraging its graph and loop constructs than introducing new technologies to solve this challenge. If a dedicated Graph database technology was used, then in addition to the cost, time and skills required to integrate it into the overall solution, the technology stack would have become more complex and the analytics processing pipeline would have necessitated transforming and moving very large datasets between two Big Data systems – and all of these complexities and cost was effectively avoided by building a simple, purpose-built solution in ECL. Since this specific use case necessitated Graph traversal step as part of a larger analytics computation job, it did not make technical, financial, and operational sense to introduce a dedicated Graph processing technology into the existing homogeneous stack.

11. Maintaining agility and edge in providing Big Data Analytics

Finally, ClearFunnel's SaaS model of providing fast and cost-effective Big Data Analytics services to its clients' custom use cases require leveraging of the common and core platform capabilities, libraries, and accelerators across different use cases – to rapidly implement new solutions at minimal costs. At the same time, for new and unique requirements of different clients, there is a constant need to extend and customize the existing common frameworks and libraries. Here, too, the native support in ECL for OOP paradigm allows the efficient balance of both of these above-mentioned technology requirements, which are critical for ClearFunnel to maintain its competitive edge, continue to deliver fast results, and provide the best value for price to its clients. The ability to implement complex Big Data Analytics solutions with relative ease by reusing, extending, and customizing core libraries of the HPCC Systems based ClearFunnel platform for different clients' use cases is a key advantage in maintaining superior service delivery with minimal re-write and code duplication.

About HPCC Systems (https://hpccsystems.com)

HPCC Systems helps businesses of all sizes find the answers they need by making data easier to process, analyze, and understand. Born from the deep data analytics history of LexisNexis® Risk Solutions, HPCC Systems provides high-performance, parallel processing and delivery for applications using big data.

The open-source platform incorporates a software architecture implemented on commodity shared-nothing computing clusters for resilience and scalability. It is configurable to support both parallel batch data processing and high-performance data delivery applications using indexed data files. The platform includes a high-level, implicitly parallel data-centric declarative programming language that adds to its flexibility and efficiency.

Developers, data scientists and technology leaders adopt HPCC Systems because it is cost-effective, comprehensive, fast, powerful, and scalable.

About ClearFunnel, LLC (https://clearfunnel.com)

ClearFunnel is a US-based Big Data and Data Science solutions provider that has pioneered a new business model of delivering analytics results as a subscription-based service for customer-specific business challenges. The company leverages the best of advancements in Machine Learning, hyper scale Big Data technologies, and Cloud capabilities to rapidly develop end-to-end advanced analytics solutions for its clients' custom use cases.

By leveraging a homogeneous Big Data Analytics platform, adopting a SaaS business model, and with focus on developing advanced decision making systems, ClearFunnel has successfully built and has been operating solutions for its customers in the Biotechnology, Predictive Marketing, Information Services & Content Analytics, Maritime Analytics & Security, and Electronic Toll Collection industries.

Specifically, ClearFunnel's advanced analytics solutions span across the domains of:

- Predictive analytics on Geospatial and operational data streams from satellite, IOT devices, and machine sensors,
- Deep Learning based image analytics (automatic license plate recognition),
- Genome sequencing analysis to detect known and unknown genetic mutations,
- Natural Language Processing (NLP) driven content analysis and recommendations,
- Web-scale data analysis for user behavior prediction, and
- 2-way integration with Salesforce data for closed-loop B2B campaign analysis.

The opinions expressed within this whitepaper represent ClearFunnel's opinions. LexisNexis and ClearFunnel believe this whitepaper experience generally represents the experience found with other similar customer situations.

However, each customer will have its own subjective goals and requirements and will subscribe to different combinations of services to suit those specific goals and requirements. This whitepaper may not be deemed to create any warranty or representation that any other customer's experience will be the same as the experience identified herein.

About LexisNexis Risk Solutions
At LexisNexis Risk Solutions, we believe in the power of data and advanced analytics for better risk management. With over 40 years of expertise, we are the trusted data analytics provider for organizations seeking actionable insights to manage risks and improve results while upholding the highest standards for security and privacy. Headquartered in metro Atlanta USA, LexisNexis Risk Solutions serves customers in more than 100 countries and is part of RELX Group, a global provider of information and analytics for professional and business customers across industries. For more information, please visit www.risk.lexisnexis.com.