Technical Review

Optimizing Analytics Workloads with the Cisco Data Intelligence Platform

Date: June 2020 **Author:** Alex Arcilla, Validation Analyst

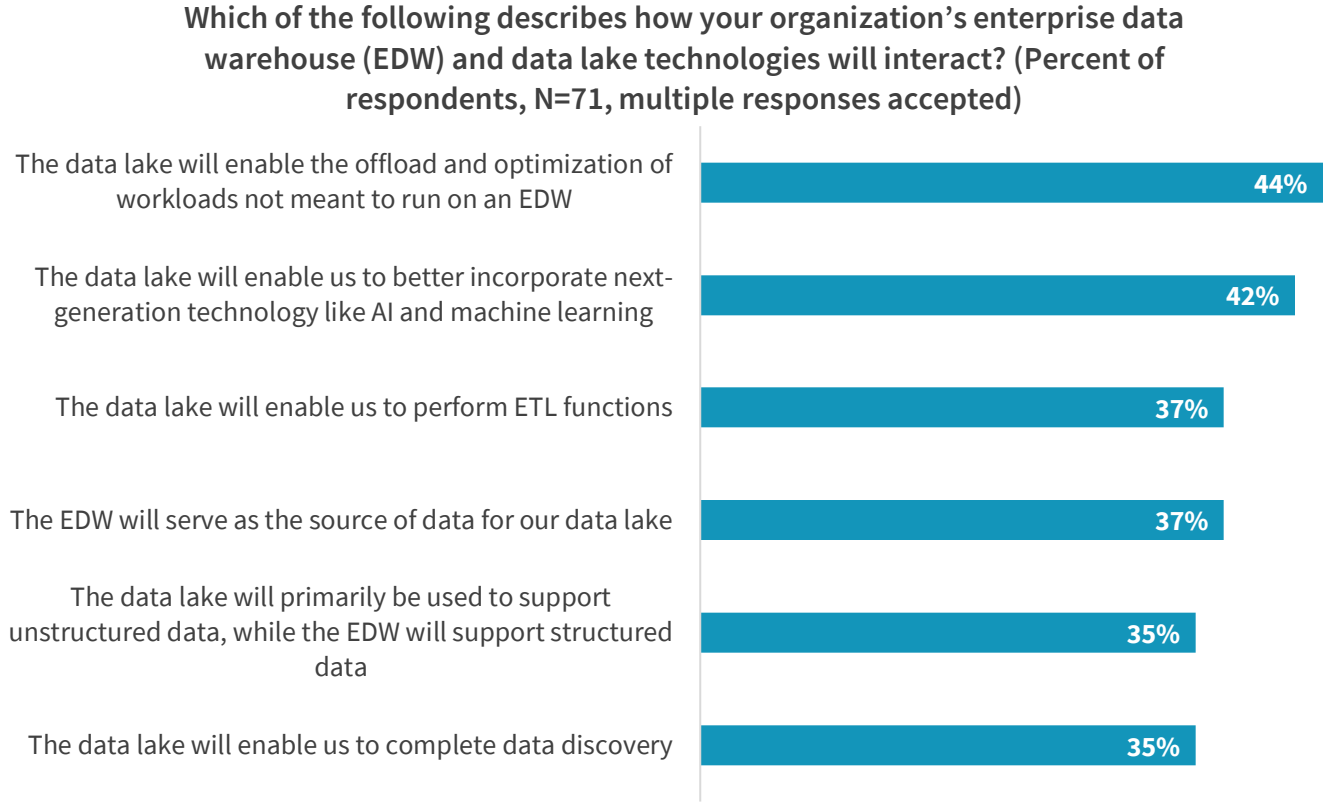
Abstract

This ESG Technical Review documents testing of the Cisco Data Intelligence Platform with a goal of validating workload performance and the ease of manageability and scalability.

The Challenges

Companies are leveraging more applications using artificial intelligence (AI) and machine learning (ML). In fact, ESG research has found that 50% of respondents are already using AI/ML technology, while the other 50% plan to use it in the foreseeable future.[1] ESG research has also uncovered that 42% of organizations view data lakes as a way to better incorporate next-generation technology such as artificial intelligence and machine learning,[2] as data lakes process and store all types of data—structured, semi-structured, and unstructured (see Figure 1)—as opposed to using an electronic data warehouse (EDW).

Figure 1. Data Lakes Expected to Enable Next-gen Technologies



Source: Enterprise Strategy Group

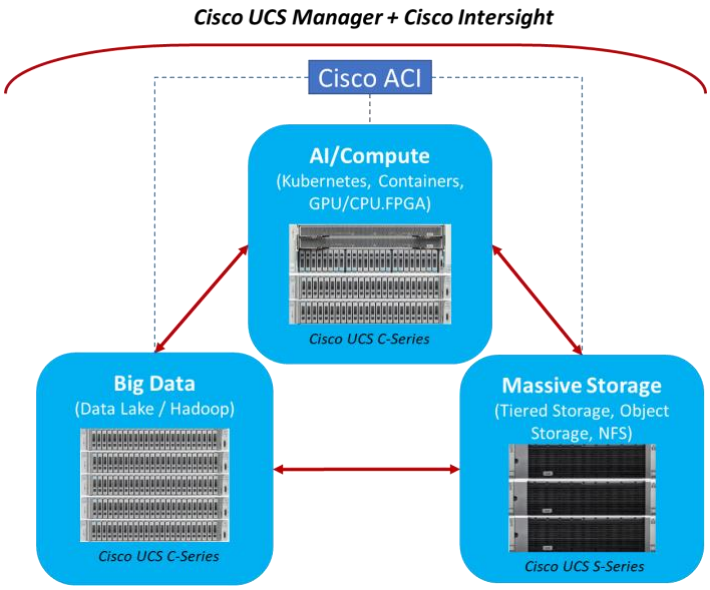
[1] Source: ESG Master Survey Results, *Artificial Intelligence and Machine Learning: Gauging the Value of Infrastructure*, March 2019.
[2] Source: ESG Master Survey Results, *The State of Data Analytics*, August 2019.

AI/ML applications and data lakes have typically resided on separate platforms, making it more difficult for IT to scale each platform adequately so that analytics workload performance is optimized. As more AI/ML workloads are working more on data contained within a data lake, the need for both technologies to be integrated into a unified architecture increases so that workload performance can scale and be optimized while easing the burden of overall management and administration.

The Solution: Cisco Data Intelligence Platform

The Cisco Data Intelligence Platform (CDIP) is a solution that enables enterprises to integrate technologies enabling data lakes, AI/ML, and object storage that work together to support analytics workloads at terabyte scale in one unified architecture. Leveraging the Cisco UCS Infrastructure, the CDIP brings together Cisco Unified Computing System (UCS) C-Series servers for compute and data lake workloads along with Cisco UCS S-Series for tiered storage. Organizations can independently scale both compute and storage resources non-disruptively, scaling the CDIP up to thousands of nodes as business needs dictate.

With the CDIP, organizations can now perform fast data ingest and data engineering at the data lake level. The CDIP can support different AI/ML frameworks (Kubernetes and Hadoop Yarn) and compute types (GPU, CPU, and FPGA with containers) depending on application demand. The integrated SAS-based storage, which can scale to multiple petabytes, enables organizations to retire data—in hot, warm, or cold tiers—from the data lake for future reference and analytics when required, thus helping to optimize data lake performance continuously.



To scale the network of compute and storage nodes, organizations employ the Cisco Application Centric Infrastructure (ACI). With Cisco ACI, the entire network of nodes and racks is treated as a single entity. Organizations can configure, manage, and monitor this single entity via pre-set application policies, simplifying scaling of compute and storage resources for any analytics workload. Starting with a single rack of compute and storage resources, the Cisco ACI is designed to help organizations add multiple nodes to the CDIP as needed, while managing the entire network with a single management interface.

With Cisco Intersight, a cloud-based management tool, organizations can gain visibility and receive alerts regarding compliance issues as well as proactive alerts for upgrade recommendations. Since Cisco Intersight has integrations with the Cisco Technical Assistance Center (TAC), organizations can upload automatically generated tech support files from the customer.

Using Cisco Intersight in conjunction with Cisco UCS manager, organizations can deploy compute and storage resources with automated, policy-driven workflows that minimize configuration errors, avoid configuration drifts, and decrease overall system downtime. Organizations can also manage multiple instances or domains across geographically dispersed locations. Thus, organizations can increase their business agility and gain valuable business insights with minimal delay.

To support organizations with CDIP implementation while minimizing risk, Cisco has collaborated with multiple independent software vendors (ISV) and semiconductor companies to offer several Cisco Validated Designs (CVD). All CVDs have been developed to help organizations address issues with running big data analytics workloads on CDIP such as performance, scalability, availability, and data tiering. Included in these CVDs are reference architectures with Cloudera and Hortonworks for big data use cases, Red Hat for compute farm use cases with OpenShift, and object storage use cases with MinIO and Scality. Each CVD is also released with the latest CPUs/GPUs from its partners, Intel, AMD, and Nvidia.

ESG Validated

ESG evaluated the CDIP to validate that the solution can help organizations ingest and process data quickly with predictable and scalable performance. Our evaluation also considered the ease of managing the CDIP to minimize downtime so that the CDIP can deliver continuous and near-instantaneous business insight. We focused our evaluation on the CDIP implementation as defined by the CVD for CDIP implementation with Cloudera.

Predictable and Scalable Performance

ESG began by reviewing the results of performance testing against industry standards set out by the Transaction Processing Performance Council (TPC), an industry-based, vendor-neutral organization that defines transaction processing and database benchmarks for a variety of workloads and applications. We specifically considered the TPC Express Benchmark HS (TPCx-HS) version 2.0.3, a data input/output (I/O) and network-intensive benchmark specifically designed to measure the performance of big data analytics workloads—typically at terabyte scale—leveraging Hadoop. Measuring against this benchmark enables organizations to evaluate Hadoop-based systems against one another in a neutral manner.[3]

ESG audited the results of Cisco's latest TPC submission to validate that tests were performed as prescribed in the TPCx-HS benchmark. The test bed consisted of the 17 Cisco UCS C240 M5 rack servers (one Name Node, 16 Data Nodes), each connected to redundant Cisco UCS 6332 Fabric Interconnects via 40GE links, populating 38 rack units. The test bed leveraged Red Hat Enterprise Linux Server version 7.6 as the operating system and Cloudera Enterprise Basic 6.3.0, based on Apache Spark.

ESG considered four metrics characterizing how well a vendor performs against the TPCx-HS benchmark:

- Scale Factor - size of the physical storage present in the nodes within the test bed.
- HSph@SF - a normalized performance metric that reflects effective sort throughput at a specified storage size.
- $/HSph - a price-performance metric.
- System availability date - date on which the tested system has been made generally available to customers.

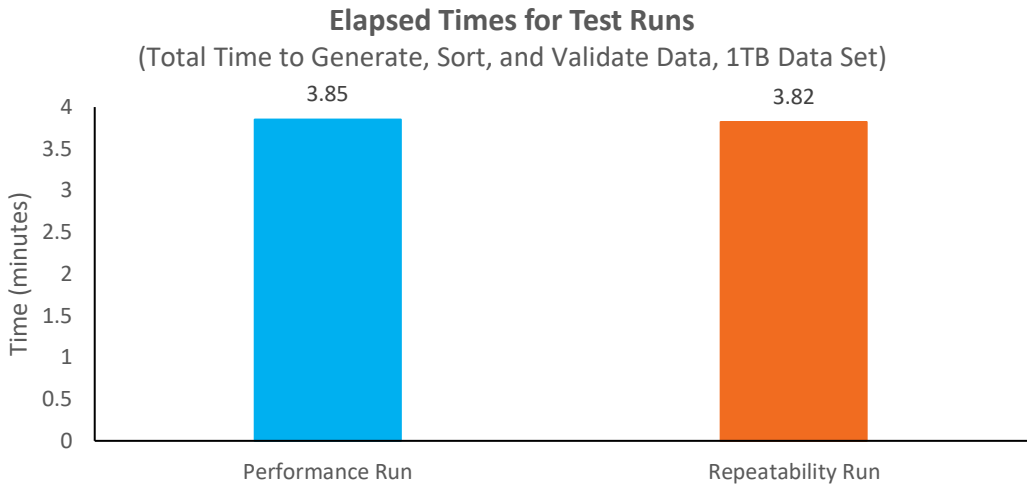
The scale factor defines the second and third metrics obtained during testing. At the time of report publication, ESG audited results for a test using a scale factor of 1 TB.[4] A 1TB data set is typically considered a minimum data set size for training an ML model.

ESG should note that the scale factors allowed by the TPC to be tested against the TPCx-HS benchmark represent generic systems that can be deployed for big data analytics workloads.[5] When evaluating these results, organizations need to consider how they apply to their own deployments and use cases.

Benchmark testing required two test runs (a performance and repeatability run) that simulate Hadoop data processing with Apache Spark, including generating, sorting, and validating the data. The limit of two runs prevents an organization from performing multiple runs and choosing the best numbers to submit for auditing by the TPC. Times to complete each run were measured to be 3.85 minutes and 3.82 minutes, respectively (results shown in Figure 2).

[3] *TPC EXPRESS BENCHMARK HS (TPCx-HS) Standard Specification Version 2.0.3*, March 2018.
[4] ibid.
[5] ibid.

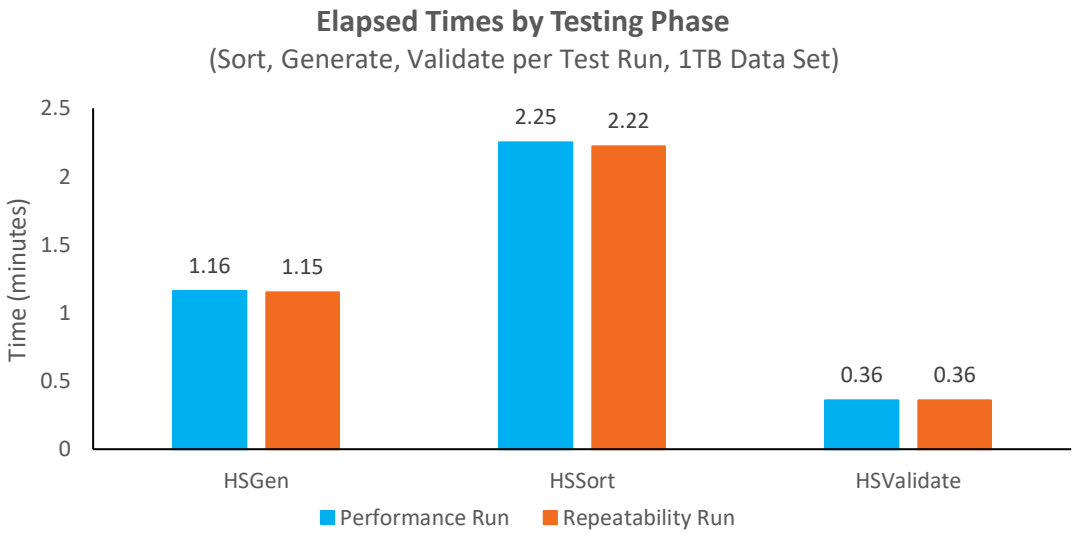
Figure 2. Elapsed Times for Performance and Repeatability Runs in TPCx-HS Apache Spark Benchmark Test



Source: Enterprise Strategy Group

When we audited the total run times, ESG noted that intermediate times for generating, sorting, and validating the data exhibited similar performance across runs (see Figure 3).

Figure 3. Elapsed Times for Completion of Intermediate Testing Phases – Generate, Sort, Validate



Source: Enterprise Strategy Group

Because the HSph@SF is determined by using the elapsed time associated with the performance run, we calculated an HSph of 15.6.[6]

ESG then reviewed the bill of materials (BOM) for the test system. To build this BOM, we included both the hardware and software components, as well as a service contract, to approximate a total cost of ownership should a customer purchase a CDIP. Accounting for a discount of 61% for hardware and 35% for a three-year maintenance contract, the total cost for this CDIP was $1,014,330. The price-performance ratio ($/HSph@SF=1) was calculated to be $65,021.06.

[6] Ibid.

Combined with the system availability date of December 13, 2019, both the calculated HSph and $/HSph for this CDIP implementation earned Cisco the top slot in TPCx-HS performance rankings at the time of report publication.[7] (A higher HSph and lower $/HSph for a given scale factor is considered better.) Rankings change over time as new submissions are reviewed and accepted by the TPC.

TPCx-HS Metrics for CDIP

(Scale Factor = 1)

HSph = 15.6

$ / HSph = $65,021.06

Highly performant, cost-effective

With a calculated HSph of 15.6, this metric represents how quickly the CDIP can process and sort data for analysis. For organizations to leverage the most out of their AI/ML-based applications, the faster that data can be ingested and prepared, the faster ML models can be trained so that AI implementations can provide near-instantaneous and relevant business insight. It is also worth mentioning that the integration of the AI/compute and data lake platforms within the same architecture further helps to reduce the time to insight, as processed data from the data lake no longer needs to be fed into a physically separate AI/ML platform.

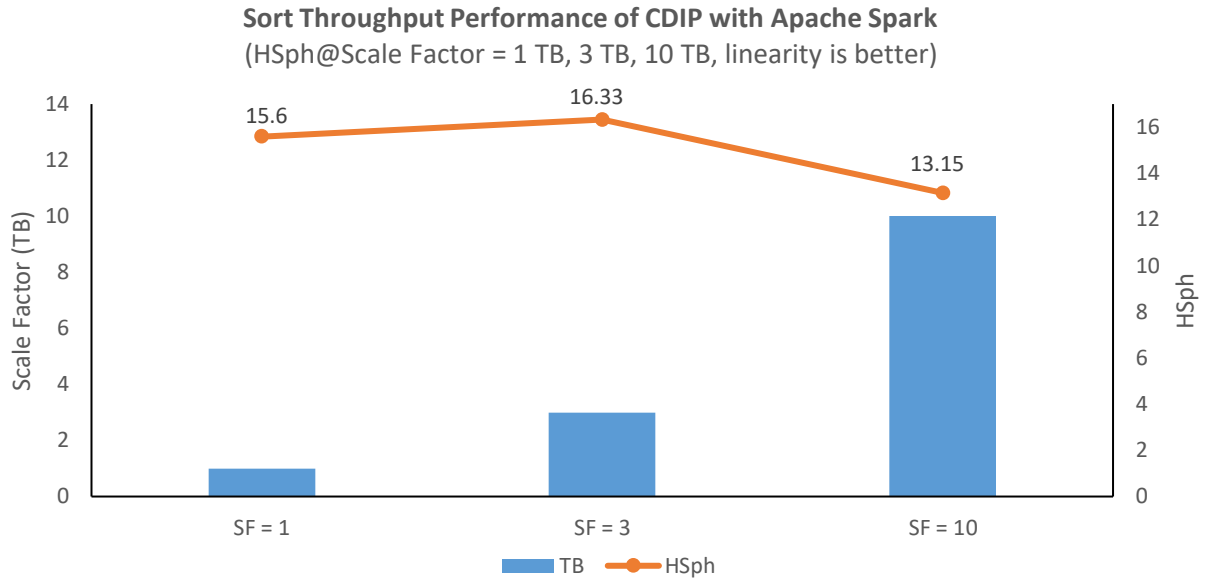
Considering both the total elapsed times of the two performance runs, as well as the times to complete each testing phase, ESG could also see that CDIP performance can be repeatable. Since organizations want to extract the most value out of their analytics initiatives, ESG can see how consistent performance of the underlying platform becomes an important factor to consider. Our examination of the test methodology and the results show that the CDIP can provide that level of consistency and predictability.

The price-performance ratio (with SF=1) indicates that an organization can extract more value out of the CDIP with a smaller investment. As organizations continue to invest in AI/ML initiatives, they must continue to evaluate how quickly and cost-effectively value can be extracted. The current price-performance ratio demonstrated by the CDIP can indicate that high value can be extracted from making such an investment in this platform.

We then audited additional results obtained when running the same benchmark test while increasing the scale factor to 3 TB and 10 TB, with the goal of observing linear performance of CDIP. According to the TPCx-HS standard, linear performance is exhibited when the HSph remains fairly constant (with some allowable deviation) as the scale factor increases. We used the same test bed used in the previously discussed test and replaced Hadoop MapReduce with Apache Spark, a faster data processing engine designed for workloads beyond batch processing, such as interactive, iterative, and streaming workloads. As seen in Figure 4, CDIP performance remains fairly consistent as the scale factor increases.

[7] www.tpc.org/tpcx-hs

Figure 4. Results of TPCx-HS Benchmark Test with Scale Factor at 1 TB, 3 TB, and 10 TB using Apache Spark



Source: Enterprise Strategy Group

ESG observed that the deviation between calculated HSph was minimal as the scale factor grew. We noted that the small variance of these normalized metrics—HSph@SF=3 was 5% more than HSph@SF=1, while HSph@SF=10 was 16% less than HSph@SF=1—was exhibited across a change in scale factor larger than 1 TB by an order of magnitude. ESG concluded that CDIP performance can be fairly consistent when using a similar CDIP configuration to the utilized test bed, and thus predictable, as the amount of CDIP storage grows.

ESG also audited the calculated $ / HSph for 1 TB, 3 TB, and 10 TB (see Table 1). We found that the price/performance ratios at 3 TB and 10 TB also exhibited minor deviations from the ratio calculated for 1 TB.

Table 1. TPCx-HS Benchmark Results for CDIP with Apache Spark - HSph and $ / HSph

Scale Factor (SF)	HSph	$ / HSph
1 TB	15.60	$65,021.16
3 TB	16.33	$62,114.52
10 TB	13.15	$77,135.37

Source: Enterprise Strategy Group

Given ESG's audit of the test results, ESG can see that CDIP can achieve fairly constant performance when increasing the size of the cluster. Simultaneously, we also see that organizations can extract value out of additional resources to the CDIP infrastructure in light of the incremental estimated three-year costs. The lower the $/HSph as the scale factor increases, the more value that can be extracted from the CDIP compared to other systems with a higher calculated $/HSPh.

Why This Matters

Any delay in processing vast amounts and types of data, particularly those collected and stored in a data lake, may prevent organizations from extracting the most value out of AI/ML applications they are increasingly employed to extract continuous and near-instantaneous business insight. System performance issues or downtime can hamper an organization's efforts to gain competitive advantage when running big data analytics workloads.

ESG validated that the CDIP can help organizations to maximize the performance of their analytics workloads. After auditing the results obtained from TPCs-HS benchmark testing, ESG found that the CDIP can support generalized data processing workloads with predictable, repeatable, and linear performance. Given the observed consistent performance, ESG believes that the CDIP can support organizations who want their data to be ingested and prepared quickly for AI/ML-based applications. By evaluating the price/performance ratios calculated for the CDIP, we also believe that organizations can extract high value from investing in this platform.

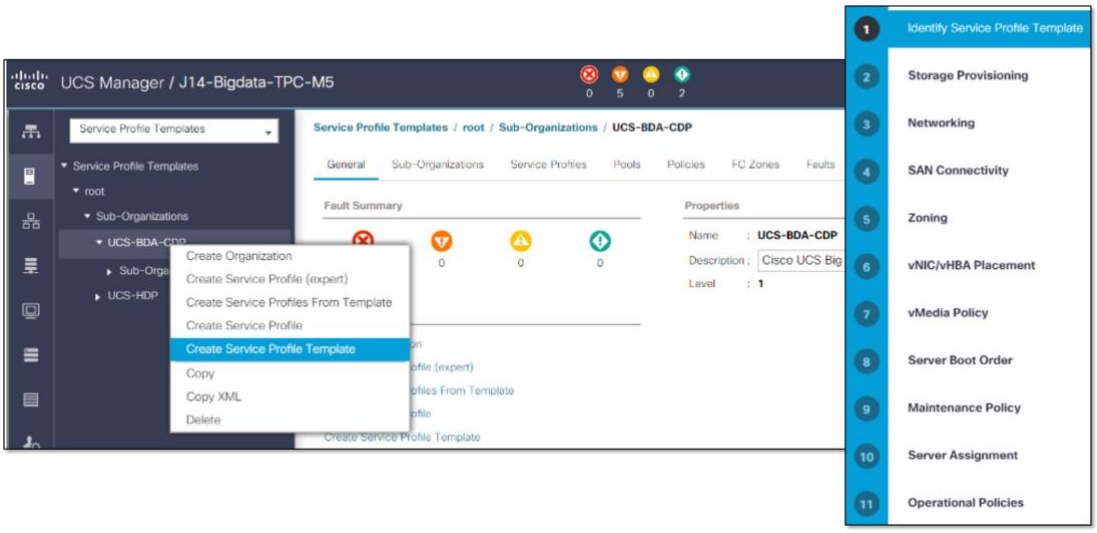
Manageability

ESG then examined how the CDIP eases overall management and administration to reduce downtime, thus enabling data analytics and AI/ML applications to offer continuous business insight. Specifically, we examined the creation of service profiles to automate hardware configurations for a given workload, the level of visibility to be obtained with Cisco Intersight, and the integration of Cisco Intersight with the Cisco TAC.

ESG began by creating a service profile, a software definition of a server and its LAN and SAN network connectivity. When a service profile is deployed to a server, the Cisco UCS Manager automatically configures the server, adapters, and fabric interconnects to match the configuration defined within the profile. Since the automation decreases the number of manual steps typically completed to configure hardware for a given workload, an IT administrator can deploy the infrastructure for a Hadoop cluster without waiting for individual components to be allocated and deployed separately and manage configuration drift. Time to insight decreases, which puts organizations at a competitive advantage.

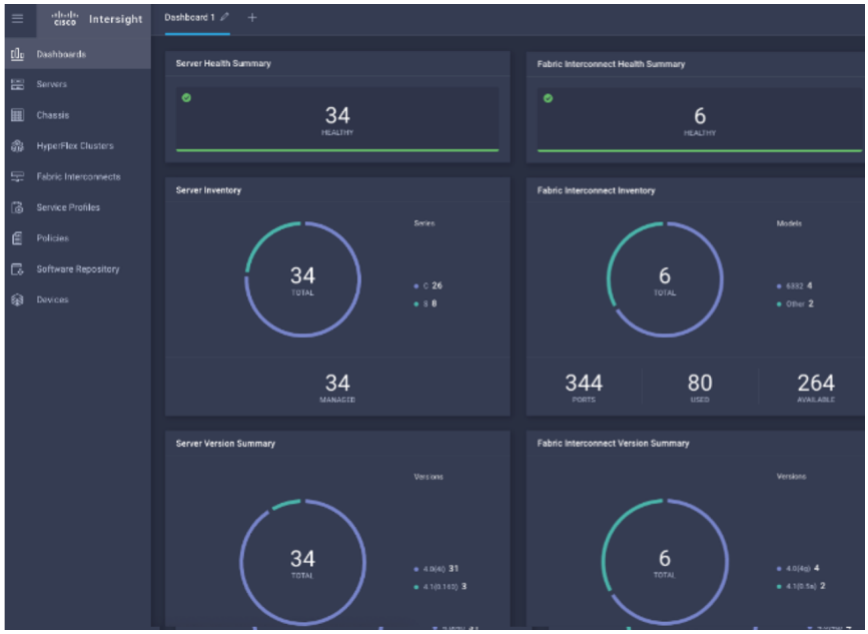
ESG first navigated to the Cisco UCS Manager interface and selected "Create Service Profile Template" (see Figure 5). By creating a template, an IT administrator can deploy hardware resources quickly and repeatedly, particularly when workloads require similar resources. We were then directed to define the service profile (right-hand side of Figure 5), such as amount of storage, network connectivity, and assignment of peripherals (NICs or HBAs). We could also assign a maintenance policy that directs how the CDIP would handle disruptive events such as reboots or network interruptions.

Figure 5. Creating a Service Profile Template via Cisco UCS Manager



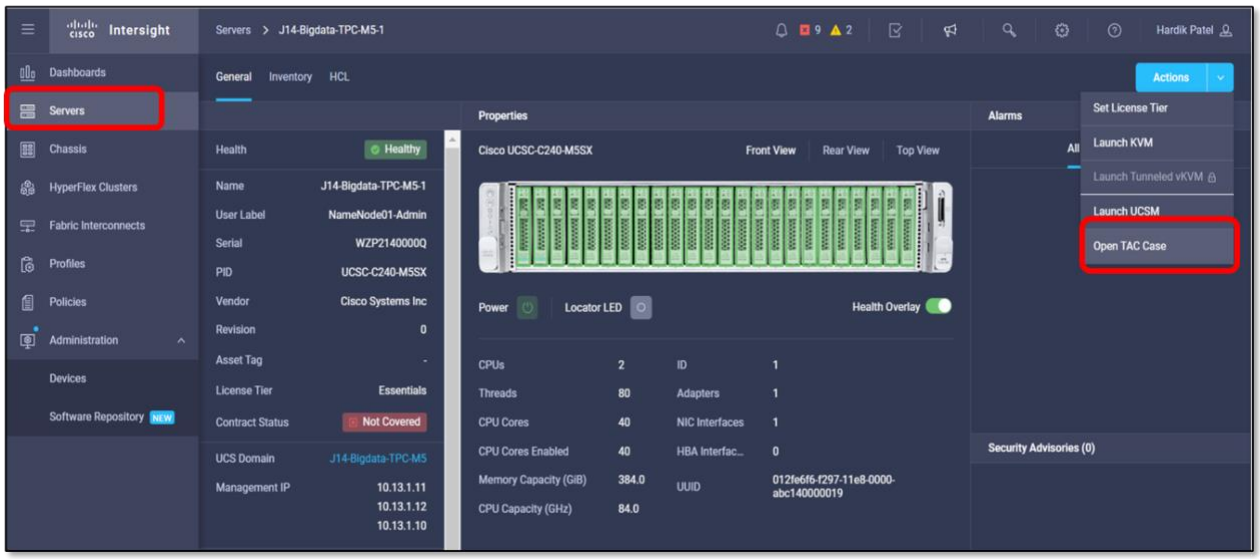
ESG then navigated to the dashboard window of Cisco Intersight. We saw how this view can help an IT administrator to determine resource inventory and the overall health of the CDIP infrastructure on one screen (see Figure 6).

Figure 6. Dashboard View of Cisco Intersight



To view a specific resource, we could select it on the left-hand side of the screen. We selected *Servers* and the screen (displayed in Figure 7) appeared.

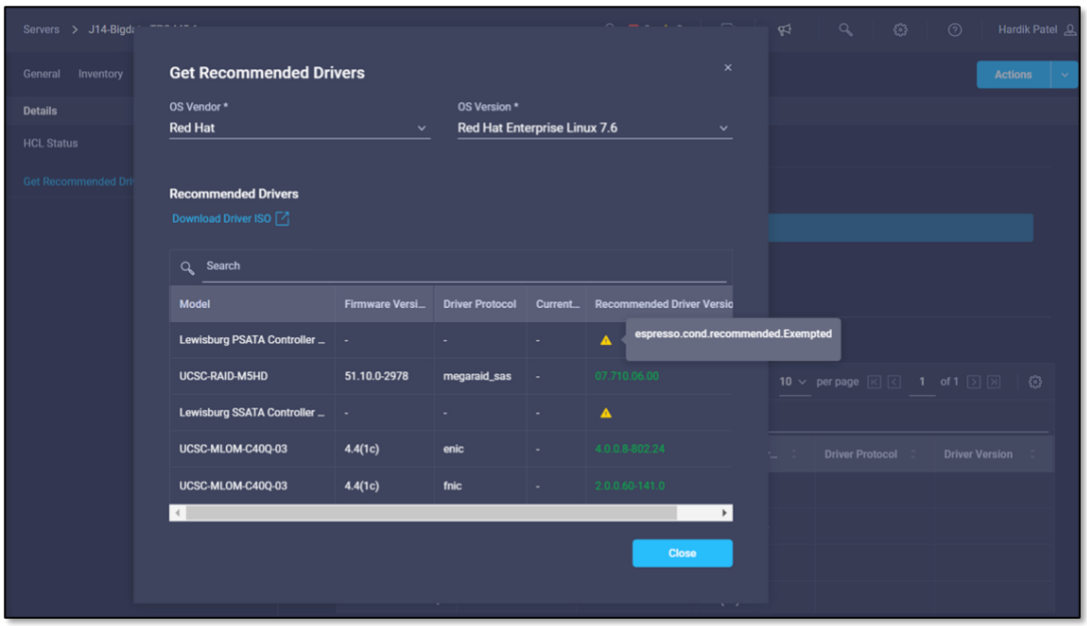
Figure 7. Opening a Cisco TAC Case via Cisco Intersight



To show the integration of Cisco Intersight with Cisco TAC, we navigated to the *Actions* tab in the upper right-hand corner of the screen, clicked to reveal the drop-down menu, and selected "*Open TAC Case*." Another window popped up so that we could confirm our choice. We were then directed to submit a case, which would then flag Cisco Intersight to respond automatically by sending the proper technical support files from TAC to troubleshoot the issue. We could see how this helps to minimize CDIP downtime as an administrator can receive information automatically to solve the case without waiting for in-person assistance.

ESG also viewed how Cisco Intersight can flag hardware compatibility issues. In Figure 8, we navigated to the *Details* menu on the *Servers* page and selected "*Get Recommended Drivers*." Should any drivers need to be updated, Cisco Intersight would show its recommendations. ESG noted that the automated polling can help an administrator to minimize configuration drift, ensuring that the CDIP infrastructure is optimized to support the organization's analytics workloads.

Figure 8. Automated CDIP Hardware Compliance



Why This Matters

For organizations that wish to obtain continuous business insight from their data analytics and AI/ML applications, easing the management and administration of the underlying platform is critical to minimize downtime associated with configuring hardware for workloads and troubleshooting service-affecting issues.

ESG validated that organizations can leverage the combination of Cisco Intersight and Cisco UCS Manager to maximize uptime via automated workflows for configuring and maintaining the CDIP. We also verified that the combination of these tools provides organizations with comprehensive visibility into the CDIP. ESG also saw how the integration of Cisco Intersight with TAC can expedite the resolution of CDIP issues, as the TAC can send the proper technical support files to troubleshoot submitted issues, removing the time spent on waiting for real-time assistance.

The Bigger Truth

Extracting the most value out of data without delay is paramount for any organization in the current business environment. If they can't extract that value, organizations are at risk of losing revenue and emerging business opportunities.

The CDIP helps organizations to extract maximum value out of data by combining the necessary AI/ML compute, data lake, and storage technologies into one platform. Instead of cobbling together siloed hardware and software resources, organizations can allocate the necessary compute and storage resources automatically from a unified resource pool to support AI/ML-based analytics workloads. Ongoing CDIP management and administration via automated workflows helps to maintain high system performance and availability so that organizations can extract maximum value out of AI/ML-driven analytics.

ESG validated that the CDIP can support Apache-Hadoop-based workloads requiring predictable data sort throughput performance. We audited the results obtained from TPCx-HS benchmark testing and found that the CDIP offers predictable performance when the scale factor was set at 1 TB. We also found that the calculated price-performance ratio revealed that organizations can extract maximum value from the CDIP in a cost-effective manner. ESG also found that CDIP performance remains consistent as the scale factor increases.

By examining the ways in which Cisco Intersight and Cisco UCS Manager automates workflows related to resource deployment and system maintenance, ESG verified that organizations can minimize overall CDIP downtime, enabling AI/ML-based workloads to provide continuous and near-instantaneous business insight.

While ESG has validated that the CDIP can support AI/ML-based workloads requiring high performance and availability, primarily via our audit of results obtained with TPCx-HS benchmark testing, we encourage potential prospects to conduct their own evaluations to determine if the CDIP meets their specific business requirements and objectives.

If you wish to leverage a data analytics platform that helps to maximize the value from data and wish to leverage both AI/ML and data lake technologies, ESG recommends that you place the CDIP on your evaluation list.