Cisco UCS Integrated Infrastructure for Big Data and Analytics with IBM BigInsights

Bring the power of Apache Hadoop, Apache Spark, and SQL to a flexible and open analytics platform

The Challenge

Over time, big data applications and analytics have advanced from batch processing to real-time data processing. SQL on Hadoop enables the user to process big data on Hadoop systems with SQL-style queries, thus simplifying the data querying, retrieval, and analysis processes. This solution should not only be stable and scale with the number of users, but also be able to respond well to varying data sizes.

IBM BigInsights over Cisco UCS Integrated Infrastructure

Cisco UCS® Integrated Infrastructure for Big Data and Analytics integrates computing, network, storage, and management resources into a cohesive programmable infrastructure that can scale as workloads demand. It provides an industry-leading solution that transparently integrates with enterprise applications.

IBM BigInsights, built on IBM Open Platform (IOP), is designed to provide analytics, operational excellence, and data security.

Cisco UCS Integrated Infrastructure for Big Data and Analytics

Cisco UCS Integrated Infrastructure for Big Data and Analytics (Figure 1) is a popular choice for enterprise deployments. It provides a highly scalable architecture designed to meet a variety of scale-out application demands. It offers transparent data and management integration capabilities for the enterprise applications that are deployed on the hardware.
Cisco UCS 6200 and 6300 Series Fabric Interconnects
Cisco UCS 6200 Series Fabric Interconnects provide high-bandwidth, low-latency connectivity for all Cisco UCS. The fabric interconnects provide integrated, unified management for all connected devices with embedded Cisco UCS Manager. The fabric interconnects are deployed in redundant pairs to deliver full active-active redundancy, performance, and the exceptional scalability needed to support the large number of nodes that are typical in clusters that service big data applications.

The Cisco UCS 6300 Series is the latest version of this technology. The Cisco UCS 6300 Series Fabric Interconnects are a core part of Cisco UCS, providing low-latency, lossless, 10 and 40 Gigabit Ethernet, Fibre Channel over Ethernet (FCoE), and Fibre Channel functions with management capabilities for systems deployed in redundant pairs. Cisco® fabric interconnects offer the full active-active redundancy, performance, and exceptional scalability needed to support the large number of nodes that are typical in clusters serving big data applications.

Cisco UCS Manager
Cisco UCS Manager enables rapid and consistent server configuration using service profiles. It also automates ongoing system maintenance activities such as firmware updates across the entire cluster as a single operation. It enables fast and accurate configuration of computing, networking, and storage resources. Cisco UCS Manager also offers advanced monitoring with options to raise alarms and send notifications about the health of the entire cluster.

Cisco UCS C-Series Rack Servers
Cisco UCS C240 M4 Rack Servers support a wide range of computing, I/O, and storage-capacity demands in a high-density, compact design. The server uses dual Intel® Xeon® processor E5-2600 v4 series CPUs and supports up to 1.5 TB of main memory and a range of hard-disk drive (HDD) and solid-state disk (SSD) drive options. The performance-optimized option supports 24 small-form-factor (SFF) disk drives, and the capacity-optimized option supports 12 large-form-factor (LFF) disk drives. This server can be used with the Cisco UCS Virtual Interface Card (VIC) 1227 or 1387, depending on the fabric interconnect that is being used. The VIC 1227 is designed to optimize high-bandwidth, and low-latency cluster connectivity. The VIC 1387 offers dual-port enhanced Quad Small Form-Factor Pluggable (QSFP+) 40 Gigabit Ethernet and FCoE in a modular-LAN-on-motherboard (mLOM) form factor.
IBM BigInsights

With IBM BigInsights 4.2, IBM provides the full range of analytics for Hadoop, Spark, and SQL over an open and flexible platform (Figure 2). This platform cohesively combines batch processing, SQL, streaming and complex analytics for any use case to handle a wide range of data processing scenarios.

The Lambda architecture in Figure 3 shows the most typical use cases for streaming and batch analytics. It shows the flow of data from various sources to Fog and Apache Kafka nodes, and then to Spark, and then further downstream to Hadoop Distributed File System (HDFS), NoSQL, SQL databases, Apache Solr, and other systems for additional processing.

Figure 2: IBM BigInsights

Figure 3: Common Use Cases for Streaming and Batch Analytics
IBM Open Platform

IOP with Apache Hadoop and Apache Spark is IBM’s big data platform. IOP is built on 100 percent open-source Apache ecosystem components. It is designed for flexible and efficient analytics and operations.

The main features of IOP with Hadoop include:

- Apache Ambari operational framework for provisioning, managing, and monitoring Hadoop clusters
- Native support for rolling upgrades for Hadoop services
- Support for long-running applications within YARN with:
  - Greater cluster utilization
  - Lower operating costs
  - Less data in motion

Integration with Apache Spark

IOP includes integration with Apache Spark 1.6.1. The benefits include fast processing from the Spark core, near-real-time analytics with Spark streaming, built-in machine learning libraries that are highly extensible using the Spark machine learning library (MLlib), querying of unstructured data, and more value from free-form text analytics with Spark SQL and graph computation and graph analytics with Spark GraphX.

Built-in Security

IOP now supports Apache Ranger. It provides a centralized security platform for managing authorization, access control, auditing, and data protection.

IBM Big SQL

Big SQL is the ultimate platform for relational database management system (RDBMS) offload and consolidation. It offers standards-compliant SQL as well as support for many vendor-specific extensions. It enables faster and easier offloading of old data from existing enterprise data warehouses or data marts to free capacity while preserving most of the familiar SQL from those platforms. Big SQL’s SQL engine for Hadoop can work with Apache Hive, HBase, and Spark concurrently for best-in-class analytic capabilities.

As with any RDBMS, performance is a critical factor, and this is certainly the case for Big SQL. Hence, one of the most significant improvements in Big SQL 4.2 is improved out-of-the-box performance with features such as partitioning capabilities and built-in default processing plans.

Main features of IBM Big SQL include:

- Easy installation and administration using Ambari
- Improved performance with concurrent query processing, partition options, and optimized default configurations
- Statistics collection and measurement with gathering statistics on table and data for query optimization
- Enhanced security through impersonation support to allow a service user to securely access data in Hadoop on behalf of another user
- Metadata integration with automatic synchronization of Big SQL metadata with Hive
- Resource management with optimal distribution of resources for high-demand (enterprise) environments
- Big SQL disaster recovery and support for online backup of the Big SQL metastore and data (local tables) and offline restoration from a remote disaster-recovery site
- Regular backup and restore processes configured to meet the user’s recovery window requirements

IBM Text Analytics

IBM Text Analytics is a powerful system for extracting structured information from unstructured and semistructured text by defining rules to create extractors. It includes an all-new powerful web-based visual text analytics framework that allows developers to easily build high-quality applications that can process text in multiple written languages and derive insights from large amounts of native textual data in various formats.
Cisco UCS Integrated Infrastructure for Big Data and Analytics offers several configurations to meet a variety of computing and storage requirements (Table 1).

### Table 1: Cisco UCS Integrated Infrastructure for Big Data and Analytics Options

<table>
<thead>
<tr>
<th>Performance Optimized Option 1 (UCS-SL-CPA4-P1)</th>
<th>Performance Optimized Option 2 (UCS-SL-CPA4-P2)</th>
<th>Performance Optimized Option 3 (UCS-SL-CPA4-P3)</th>
<th>Capacity Optimized Option 1 (UCS-SL-CPA4-C1)</th>
<th>Capacity Optimized Option 2 (UCS-SL-CPA4-C2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Connectivity:</strong></td>
<td><strong>Connectivity:</strong></td>
<td><strong>Connectivity:</strong></td>
<td><strong>Connectivity:</strong></td>
<td><strong>Connectivity:</strong></td>
</tr>
<tr>
<td>• 2 Cisco UCS 6296UP 96-Port Fabric Interconnects</td>
<td>• 2 Cisco UCS 6296UP 96-Port Fabric Interconnects</td>
<td>• 2 Cisco UCS 6332 Fabric Interconnects</td>
<td>• 2 Cisco UCS 6296UP 96-Port Fabric Interconnects</td>
<td>• 2 Cisco UCS 6296UP 96-Port Fabric Interconnects</td>
</tr>
<tr>
<td><strong>Scaling:</strong></td>
<td><strong>Scaling:</strong></td>
<td><strong>Scaling:</strong></td>
<td><strong>Scaling:</strong></td>
<td><strong>Scaling:</strong></td>
</tr>
<tr>
<td>• Scales up to 1000s of servers using Cisco Nexus 9000 or 7000 series switches</td>
<td>• Scales up to 1000s of servers using Cisco Nexus 9000 or 7000 series switches</td>
<td>• Scales up to 1000s of servers using Cisco Nexus 9000 or 7000 series switches</td>
<td>• Scales up to 1000s of servers using Cisco Nexus 9000 or 7000 series switches</td>
<td>• Scales up to 1000s of servers using Cisco Nexus 9000 or 7000 series switches</td>
</tr>
<tr>
<td><strong>16 Cisco UCS C240 M4 Rack Servers (SFF), each with:</strong></td>
<td><strong>16 Cisco UCS C240 M4 Rack Servers (SFF), each with:</strong></td>
<td><strong>16 Cisco UCS C240 M4 Rack Servers (SFF), each with:</strong></td>
<td><strong>16 Cisco UCS C240 M4 Rack Servers (LFF), each with:</strong></td>
<td><strong>16 Cisco UCS C240 M4 Rack Servers (LFF), each with:</strong></td>
</tr>
<tr>
<td>• 2 Intel Xeon processor E5-2680 v4 CPUs (14 cores on each CPU)</td>
<td>• 2 Intel Xeon processor E5-2680 v4 CPUs (14 cores on each CPU)</td>
<td>• 2 Intel Xeon processor E5-2680 v4 CPUs (14 cores on each CPU)</td>
<td>• 2 Intel Xeon processor E5-2620 v4 CPUs (8 cores on each CPU)</td>
<td>• 2 Intel Xeon processor E5-2620 v4 CPUs (8 cores on each CPU)</td>
</tr>
<tr>
<td>• 256 GB of memory</td>
<td>• 256 GB of memory</td>
<td>• 256 GB of memory</td>
<td>• 128 GB of memory</td>
<td>• 256 GB of memory</td>
</tr>
<tr>
<td>• Cisco 12-Gbps SAS Modular RAID Controller with 2-GB flash-based write cache (FBWC)</td>
<td>• Cisco 12-Gbps SAS Modular RAID Controller with 2-GB FBWC</td>
<td>• Cisco 12-Gbps SAS Modular RAID Controller with 2-GB FBWC</td>
<td>• Cisco 12-Gbps SAS Modular RAID Controller with 2-GB FBWC</td>
<td>• Cisco 12-Gbps SAS Modular RAID Controller with 2-GB FBWC</td>
</tr>
<tr>
<td>• 24 x 1.8-TB 10,000-rpm SFF SAS drives (691 TB total)</td>
<td>• 24 x 1.8-TB 10,000-rpm SFF SAS drives (691 TB total)</td>
<td>• 24 x 1.8-TB 10,000-rpm SFF SAS drives (691 TB total)</td>
<td>• 12 x 8-TB 7,200-rpm LFF SAS drives (1536 TB total)</td>
<td>• 12 x 8-TB 7,200-rpm LFF SAS drives (1536 TB total)</td>
</tr>
<tr>
<td>• 2 x 240-GB 6-Gbps 2.5-inch enterprise value SATA SSDs for boot &amp; Cisco UCS VIC 1227 (with 2 x 10 Gigabit Ethernet SFP + ports)</td>
<td>• 2 x 240-GB 6-Gbps 2.5-inch enterprise value SATA SSDs for boot &amp; Cisco UCS VIC 1227 (with 2 x 10 Gigabit Ethernet SFP + ports)</td>
<td>• 2 x 240-GB 6-Gbps 2.5-inch enterprise value SATA SSDs for boot &amp; Cisco UCS VIC 1227 (with 2 x 10 Gigabit Ethernet SFP + ports)</td>
<td>• 2 x 240-GB 6-Gbps 2.5-inch enterprise value SATA SSDs for boot &amp; Cisco UCS VIC 1227 (with 2 x 10 Gigabit Ethernet SFP + ports)</td>
<td>• 2 x 240-GB 6-Gbps 2.5-inch enterprise value SATA SSDs for boot &amp; Cisco UCS VIC 1227 (with 2 x 10 Gigabit Ethernet SFP + ports)</td>
</tr>
</tbody>
</table>

Please note: The storage disks on the nodes being used for Big SQL installation must have sector size of 512 bytes. Else, the user will be returned an error “SQLCODE-902” during the creation of the Big SQL database. SQLCODE-902 states that a system error has occurred.
Conclusion

Cisco UCS Integrated Infrastructure for Big Data and Analytics with IBM BigInsights offers an enterprise-class platform for processing, analyzing, and visualizing Internet-scale data volumes. It combines the cohesive programmable infrastructure of Cisco UCS with the open-source distributed computing offered by IBM BigInsights. Thus, this joint solution is optimized to deliver faster insights into the data and greater performance efficiency while reducing the user's total cost of ownership (TCO).

For More Information

- For more information about Cisco UCS big data solutions, see www.cisco.com/go/bigdata.

- For more information about Cisco UCS Integrated Infrastructure for Big Data and Analytics, see http://blogs.cisco.com/datacenter/cpav4.

- For more information about IBM BigInsights, see https://www.ibm.com/analytics/us/en/technology/biginsights/.