Cray cluster supercomputers are turnkey solutions purpose-built for realizing the potential of Hadoop. Built on optimized configurations of the Cray CS300™ cluster supercomputer series, these systems feature Linux™ OS, workload management software, Advanced Cluster Engine™ (ACE) management software and the Intel® Distribution for Apache Hadoop software—all integrated, optimized, validated and supported by Cray.

- **Turnkey.** Integrated solutions incorporate hardware, software, services and support for rapid, hassle-free deployment.
- **Performance.** Validated and optimized technologies provide performance for the most demanding Hadoop requirements.
- **Reliability.** Reliable service levels leverage proven technologies like ACE management software as well as integrated resource management and job scheduling.
- **Maintainability.** Easy to maintain with management tools and services for the entire solution.
Developing the Fourth “V” — High Value Hadoop

Volume, velocity and variety are often cited as the three “Vs” of Big Data. But a fourth “V” — value — is emerging as the desired result of high performance data analytics. Some analysts predict Hadoop’s importance will equal or surpass the value of ERP systems within a few years, prompting Hadoop users to take forward views of their implementations, looking beyond immediate projects to consider its longer term potential.

Once used only as a warehouse for unstructured data and batch reporting, organizations are increasingly coming to expect more from Hadoop.

<table>
<thead>
<tr>
<th>Store</th>
<th>Report</th>
<th>Analyze</th>
<th>Monitor</th>
<th>Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type or Users</td>
<td>• Data Scientists</td>
<td>• Data Scientists</td>
<td>• Data Scientists</td>
<td>• Data Scientists</td>
</tr>
<tr>
<td></td>
<td>• Analysts</td>
<td>• Analysts</td>
<td>• Analysts</td>
<td>• Analysts</td>
</tr>
<tr>
<td># of Users</td>
<td>Few</td>
<td>Few</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Algorithms</td>
<td>• Few Crude</td>
<td>• Few Basic</td>
<td>• Many Complex</td>
<td>• Many Complex</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Third Party Apps</td>
<td>• Third Party Apps</td>
</tr>
<tr>
<td>Currency</td>
<td>Infrequent Batch</td>
<td>Frequent Batch</td>
<td>Frequent Batch</td>
<td>Frequent Batch</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Some Real-time</td>
<td>• Much Real-time</td>
</tr>
<tr>
<td>Data Types</td>
<td>• Unstructured</td>
<td>• Unstructured</td>
<td>• Unstructured</td>
<td>• Unstructured</td>
</tr>
<tr>
<td></td>
<td>• Binary</td>
<td>• Binary</td>
<td>• Binary</td>
<td>• Semi-Structured</td>
</tr>
<tr>
<td></td>
<td>• Semi-Structured</td>
<td>• Semi-Structured</td>
<td>• Semi-Structured</td>
<td>• Big Table-RDBMS</td>
</tr>
<tr>
<td>Data Volume</td>
<td>Medium-High</td>
<td>Medium-High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Value</td>
<td>Low</td>
<td>Medium-High Value</td>
<td>High Value</td>
<td>High Value</td>
</tr>
</tbody>
</table>

Expanding Hadoop beyond basic storage and reporting increases complexity in a multi-dimensional way.

Organizations expanding Hadoop beyond basic storage and reporting begin to push the performance limits achievable with ad-hoc infrastructures. Furthermore, as they become increasingly dependent on the value that Hadoop provides, system reliability and maintainability become critical.
High Value Requirements: Performance, Reliability and Maintainability

Organizations looking to expand their Hadoop framework to take advantage of business intelligence, real-time monitoring and prediction must evaluate the following capabilities: Performance, reliability and maintainability.

**Performance**

**Support More Data**
- **Variety** – Unstructured, semi-structured, Big Table/RDBMS and binary can all be incorporated
- **Volume** – Explosive data growth
- **Velocity** – Near real time required with monitoring and predictive uses

**Support More Users**
- **Population Growth** – User-friendly tools and evolving business requirements will extend use beyond data scientists to non-technical users
- **User Expectations** – Users will require more frequent and faster queries/joins

**Support More Complexity**
- **Big & Fast** – Oceans of data will likely be correlated with streams
- **Algorithms** – Complexity increases with expanding analysis and predictive uses
- **Cluster Sprawl** – New Hadoop clusters will sprawl if performance gaps exist

**Reliability**

- **Deliver Fast ROI** – Need quick and reliable production schedules
- **Instill Confidence** – Need to reliability meet future objectives
- **Reliable Service Levels** – Integrated resource management and job scheduling

**Maintainability**

- **Support** – Relying on numerous vendors can result in indeterminate issues
- **Management** – Needs integrated and holistic management
- **Change** – Controlled risk with inevitable upgrades and configuration changes
- **Staffing** – Must not overburden exiting staff or require erratic ramp-ups
Cray Delivers Turnkey High Value Hadoop

With its Hadoop solutions Cray provides a fast and reliable path to using Hadoop applications as the basis for critical decisions and strategic moves. Instead of worrying about the system, organizations can focus on accessing more data, posing more questions and choosing the best results.

Features

Flexible, Hadoop-Optimized Configurations
Cray CS300 cluster supercomputers have been purpose built for a variety of high value Hadoop environments. Users can deploy native Hadoop Distributed File System (HDFS) clusters with direct-attached disk configurations or Lustre® with the Cray Sonexion® scale-out Lustre storage system. Cray CS300 systems are liquid- or air-cooled and can be configured with a variety of disks, processors and interconnects. In addition to Hadoop, users can extend the system for scientific computing or other workloads.

Intel® Distribution of Apache Hadoop Software
The Intel® Distribution of Apache Hadoop software incorporates HPC-centric enhancements, including hardware performance optimization, improved management and enhanced security. Cray has extended this distribution to include InfiniBand support and tuning and improved resource management. This combination provides an advanced Hadoop distribution in an HPC-customized package for higher value Hadoop workloads.

Integrated Resource Management and Job Scheduling
Cray cluster supercomputers for Hadoop incorporate out-of-the-box resource management and job scheduling. Tightly integrated with Hadoop, the Cray-optimized Simple Linux Utility for Resource Management (SLURM) suite ensures that resources are optimally allocated to maintain reliable service levels. Varying Hadoop libraries and extensions by job enables users to service a wide variety of Hadoop workloads and avoid adding new clusters.

Hadoop-Optimized Advanced Cluster Engine (ACE)
ACE is a complete management software suite designed to eliminate the complexity of managing a cluster while providing all the tools to run large, complex Hadoop implementations. It includes command line (CLI) and graphical user interfaces (GUI) providing flexibility for the cluster administrator. An intuitive and easy-to-use GUI connects directly to the ACE daemon on the management server and can be executed on a remote system running Hadoop services or other integral services. The management modules include network, server, cluster and storage management.

Service and Support
Cray provides service and support for the entire solution, from installation to end of life. By leveraging one partner for the entire Hadoop software and hardware stack, organizations can avoid the issues often associated with unsupported or multiple vendor-supported solutions.