Actionable Insights through Next-Gen Analytics

As the business world grows ever more connected, enterprise business users and IT are challenged with getting the most value from today’s large data sets. IT managers struggle with silos of data, needing to enable users to interact with a variety of data sources in a seamless manner and with familiar tools. Solving this problem is critical to long-term success in data analytics and, in turn, fostering business success.

Apache Hadoop* paired with the Oracle® Big Data Connectors software suite helps IT provide users with a unified, seamless data resource and high-performing analytics platform. Building on this to enable an even wider range of data analysis of Apache Hadoop with enterprise-grade capabilities, Intel has developed the Intel® Distribution for Apache Hadoop* (IDH) software, which is now certified to work with Oracle Big Data Connectors. This enables IT to fully integrate data stored in existing infrastructure with data stored in the Hadoop Distributed File System.

Forging an Information Delta

To derive true business value from big data, you need the right tools to capture and organize a wide variety of data types from different sources, and be able to easily analyze it within the context of all your existing enterprise data. For business enterprises leveraging Oracle Database, Hadoop allows large data volumes—both structured and unstructured—to be organized and refined without needing to move these volumes to a separate analytics cluster. Oracle Big Data Connectors software acts as a bridge for data processing in Hadoop along with the rest of an enterprise’s data residing in Oracle® Database 11g Release 2 or Oracle® Database 12c. As a result, database administrators can use familiar tools to integrate big data with existing data stored in Oracle engineered systems or Oracle Database.

By incorporating Oracle Big Data Connectors, database administrators can transfer data from a Hadoop cluster with the Intel Distribution for Apache Hadoop into Oracle Database; bypass the need to learn new querying languages; and perform analysis on significantly larger data sets while minimizing data movement. The software suite also allows database administrators to quickly deliver data discovery applications to business users while reducing the complexities of Hadoop through graphical tooling.
Components of the Oracle Big Data Connectors software suite include:

- **Oracle® SQL Connector for Hadoop Distributed File System (HDFS):** This connector allows for fast access to data in Hadoop from your Oracle Database, users can load data into a database with Oracle SQL, or query the data in-place in Hadoop via external tables. After automatic generation of the external table, users simply use Oracle SQL to query or load the table into the database. Full SQL access allows users to join data in Hadoop with data stored in Oracle Database, apply analytic functions, and more. Furthermore, Oracle SQL Connector for HDFS loads and queries data in parallel, making it extremely fast for bulk loads.

- **Oracle® Loader for Hadoop:** This connector utilizes Hadoop compute resources to sort, partition, and convert data into Oracle-ready data types before loading. In addition to greatly reducing database CPU usage through pre-processing, Oracle Loader for Hadoop intelligently distributes data across tasks that load data into the database in parallel, helping minimize the effects of data skew.

- **Oracle® Data Integrator (ODI) Application Adapter for Hadoop:** Specific ODI Knowledge Modules optimized for operations on Hadoop are included within the ODI Application Adapter for Hadoop. These knowledge modules can be used to build Hadoop metadata within ODI, load data into Hadoop, transform data within Hadoop, and load data into Oracle Database using Oracle Loader for Hadoop and Oracle SQL Connector for HDFS.

- **Oracle® R Advanced Analytics for Hadoop:** This package offers users of the open-source statistical environment R the ability to analyze data stored in Hadoop, as well as run R models efficiently against large volumes of data leveraging MapReduce processing, without requiring R users to learn yet another API or language. This package also enables users to take advantage of massive amounts of data to build complex predictive models.

- **Oracle® XQuery for Hadoop (OXH):** Enabling the use of XQuery to process and transform XML, JSON or Avro content stored on a Hadoop cluster, OXH can be used to perform web log analysis, transformation operations on XML, JSON, and Avro content, and process large volumes of content prior to loading into a database.

**Business transformation through big data**

With a complete Intel platform-based solution, organizations can shorten the time to process a 1 TB sort, from > 4 hours to < 10 minutes, opening the door to improved insights and greater innovation.

**Greater data insight with IDH: Built from the silicon up to deliver balanced performance**

Achieving the fastest, most-reliable analytics via Apache Hadoop requires the latest technology. Using the latest generation of hardware components comprising the systems’ compute, memory, storage, and networking resources delivers significant gains in performance.
Tested and proven in data-intensive environments, IDH can meet the most demanding requirements of an enterprise. Intel has worked closely with some of the largest telecommunications, e-commerce, and government agencies in China to develop and deploy data intensive solutions. In the US, the IDH is deployed at a number of sites in production as the platform for big data applications ranging from genomic analysis to information management for financial services. For organizations already leveraging Intel® Xeon® processors in their data centers, the IDH is the only distribution built with the hardware in mind to deliver balanced performance. Along with the Intel Xeon processor, Intel® Solid-State Drives (Intel® SSD), and 10Gb Intel® Ethernet networking, the Intel Distribution offers a robust platform where optimized performance drives faster insights.

- **Processor:** The Intel Xeon processor E5-2600 v2 product family delivers up to 50 percent higher performance than the previous generation.\(^1,2,3,11\)
- **Storage:** Performance improvements in storage I/O are ongoing, and the costs to gain the performance advantages of SSDs keep dropping. The use of Intel SSDs with Intel® Cache Acceleration Software can reduce the processing time by 80 percent for some workloads.\(^1,4\)
- **Network:** 10Gb Intel Ethernet (10 GbE) networking demonstrates its value in the form of high levels of network utilization in the Hadoop cluster. The full use of greater bandwidth can reduce time to ingest and to export data by 80 percent.\(^5\) Moreover, the cost per gigabit of bandwidth with 10 GbE is now much lower than 1GbE, making it a natural choice for big data.

### Security without compromising performance

IDH is enterprise-ready, offering enhanced performance for Hadoop cluster data processing without sacrificing security. Security-sensitive files are encrypted by external applications before they arrive at the Hadoop cluster and are loaded into HDFS. Each file arrives with the corresponding encryption key. When an encrypted file enters the Apache Hadoop environment, it remains encrypted in HDFS. It is then decrypted as needed for processing and re-encrypted before it is moved back into storage. The results of the analysis, including intermediate results, are also encrypted. Data and results are always encrypted, while in transit and at rest. And with assistance from Intel® Advanced Encryption Standard New Instructions (Intel® AES-NI),\(^8\) a technology built into Intel Xeon processors, businesses can accelerate data encryption and decryption in Hadoop by 20x.\(^1,9\) Enterprises and government agencies can now store and process encrypted data on Apache Hadoop and HDFS without having to trade off performance by deploying the IDH.

### Committed to open source analytics innovation

In addition to these hardware integration benefits, Intel delivers platform innovation in open source and is committed to supporting the Apache developer community with code and collaboration. Intel is working on open source projects to establish a comprehensive security framework for Hadoop to help organizations protect sensitive and legally protected consumer data from identity theft and fraud (code name Project Rhino) and enable full SQL-92 coverage for OLAP applications on Hadoop while prioritizing both open source coverage and real-time performance of analytic SQL queries on Hadoop (code name Project Gryphon). Additionally, Intel is working to make Hadoop analytics faster (code name Spark and Shark).

---

**IDH paired with Oracle® Big Data Connectors can help better organize big data for easier and more informed analysis.**

Integrating the IDH and Oracle Big Data Connectors can benefit enterprises across industries. Converting machine data into call detail records for telecommunications, analyzing RFID scans and similar machine code output in manufacturing, or performing retail social media sentiment analysis are just a few instances where IDH and Oracle Big Data Connectors can turn massive amounts of data into actionable analytics.
Better Together: The Intel Distribution for Apache Hadoop and Oracle Big Data Connectors

Better business insights start with a fast, reliable, and holistic flow of data in data centers. Yet one of the greatest challenges facing organizations working with Apache Hadoop is integrating data stored in existing infrastructure with data stored in HDFS. Solving this problem is critical to long-term success in data analytics. Intel has partnered with Oracle to certify the compatibility between IDH and Oracle Big Data Connectors, providing enterprise organizations with a complete solution to this data integration challenge with full support from both companies. In a business world where connectivity to data is critical, pairing IDH and Oracle Big Data Connectors together helps enterprise IT achieve a complete database strategy.

### Intel® Distribution for Apache Hadoop® + Oracle® Big Data Connectors: Functional Testing

<table>
<thead>
<tr>
<th>IDH</th>
<th>Oracle Database</th>
<th>Oracle Big Data Connectors</th>
<th>OSCH (SQL)</th>
<th>OLH (Loader)</th>
<th>ODIA</th>
<th>ORCH (R)</th>
<th>OXH (XQuery)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDH 2.5</td>
<td>11.2.0.4</td>
<td>2.3</td>
<td>Certified</td>
<td>Certified</td>
<td>Certified</td>
<td>Certified</td>
<td>Certified</td>
</tr>
<tr>
<td>IDH 2.5</td>
<td>12.1</td>
<td>2.3</td>
<td>Certified</td>
<td>Certified</td>
<td>Certified</td>
<td>Certified†</td>
<td>Certified</td>
</tr>
</tbody>
</table>

†Operating System Support: Redhat Enterprise Linux 6 and CentOS 6. IDH 2.5 is based on Apache Hadoop 1.x.

*ORCH connector requires update of Apache Hive to version 0.11 in IDH. Contact Intel for support of this update.

**Take advantage of our partnership**
- To learn more about the IDH for your datacenter, visit intel.com/hadoop
- To learn more about Oracle Big Data Connectors for your datacenter, visit oracle.com/us/products/database/big-data-connectors/overview/index.html

**Learn more about complementary Intel solutions**
- Intel Solid-State Drives - intel.com/go/SSD
- Intel Xeon Processors - intel.com/go/Xeon
- intel.com/bigdata
Intel, the Intel logo and Xeon are trademarks of Intel Corporation in the U.S. and/or other countries. Oracle is a registered trademark of Oracle and/or its affiliates. *Other names and brands may be claimed as the property of others. 

1. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

2. Processor numbers do not measure performance. Processor numbers differentiate features within each processor family, not across different processor families: Go to: http://www.intel.com/products/processor_number


4. Intel estimated reduction using Base Cluster Configuration (see below) with Intel® SSD 520 Series vs. 10K SAS HDD with Intel® Xeon® Processor E5-2690 and Intel® Ethernet 1 Gigabit running Apache® Hadoop 1.0.3

5. Intel estimated reduction using Arista 7050T 48-port 10 Gbps switch, Hadoop cluster configuration (1 head node [name node, JobTracker], 9 worker nodes [Data nodes, TaskTracker]), Network interface cards (1 Intel® Ethernet 10 Gigabit Converged Network Adapter [10GBASE-T], 1 Intel® Ethernet Gigabit Server Adapter [1000BASE-T per server]), Hardware: 10 SuperMicro SuperServer® 1026T-URF servers (1U, two-socket); Intel® Xeon® procesors 5690, 48 GB RAM, 5 700 GB SATA hard drives @ 7200 RPM, Software: CentOS 6.2.

6. Intel estimated reduction using Apache® Hadoop 1.0.3 with running Intel® Ethernet SFP+ vs. 1Gb Intel® Ethernet with Intel® Xeon® Processor E5-2690 and Intel® SSD 520 Series running Apache® Hadoop 1.0.3

7. Intel estimated reduction using Base Cluster Configuration (see below) with running Intel Distribution for Hadoop 2.1.1 vs. Apache® Hadoop 1.0.3 with Intel® Xeon® Processor E5-2690, Intel® SSD 520 Series, and 10Gb Intel® Ethernet SFP

8. Intel® AES-NI requires a computer system with an AES-NI enabled processor, as well as non-Intel software to execute the instructions in the correct sequence. AES-NI is available on select Intel® processors. For availability, consult your reseller or system manufacturer. For more information, see http://software.intel.com/en-us/articles/intel-advanced-encryption-standard-instructions-aes-ni/

9. Encryption and decryption - Test configuration: Intel® Xeon® E5-2690 0 @ 2.90GHz (32 core, only 1 core used), 32 GB of memory, 240 GB SSD (INTEL SSDSC2CW240A3), CentOS 6.3, AES-NI: OpenSSL 1.0.1c AES-NI enabled, HDFS: name node, datanode, and test program ran on the same machine, Test input: 1 GB text file. Results: No AES-NI-MB enabled resulted in 84 MB/s encryption time and 64 MB/s decryption time. AES-NI-MB enabled resulted in 1567 MB/s encryption time and 1258 MB/s decryption time.

10. The IDH version 2.5 is certified on Oracle Big Data Connectors version 2.3 and Oracle Database versions 11.2.0.4 and 12.1 with Apache 1.0.

11. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. Intel does not control or audit the design or implementation of third party benchmark data or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmark data are reported and confirm whether the referenced benchmark data are accurate and reflect performance of systems available for purchase. For more information go to http://www.intel.com/performance.

Original Cluster Configuration:

Hardware: Arista 7050T, 10 x SuperMicro 1U servers (Intel processor: 2 x 3.46 GHz Intel® Xeon® processor 5690, Memory: 48GB RAM, Storage: 5 x 700 GB 7200 RPM SATA disks), 10Gb Intel® Ethernet Server Adapters (10GBASE-T), Intel® Ethernet Gigabit Server Adapter (1000BASE-T)
Software: CentOS 6.2, Cloudera's Distribution of Apache Hadoop, Oracle JDK 1.7.0, Cluster: 1 client machine, 1 head node (name node, Job Tracker, 9 Workers (data nodes, task trackers).

Base Cluster Configuration:

1 Head Node (name node, job tracker), 10 Workers (data nodes, task trackers), 10-Gigabit Switch: Cisco Nexus 5020; Head Node Hardware: 1 Dell r710 1U servers (Intel: 2x3.47GHz Intel® Xeon® processor E5-2690, Memory: 128G RAM, Storage: 10K SAS HDD with Intel® Xeon® Processor E5-2690 and Intel® Ethernet 1 Gigabit running Apache® Hadoop 1.0.3)

10. IDH version 2.5 is certified on Oracle Big Data Connectors version 2.3 and Oracle Database versions 11.2.0.4 and 12.1 with Apache 1.0.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

Copyright © 2014 Intel Corporation. All rights reserved.