ORACLE BIG DATA CONNECTORS

BIG DATA FOR THE ENTERPRISE

KEY FEATURES

- Tight integration with Oracle Database
- Leverage Hadoop compute resources for data in HDFS
- Enable Oracle SQL to access and load Hadoop data
- Fast and very efficient load from Hadoop into Oracle Database
- Partition pruning of Hive tables during load and query
- Graphical user interfaces of Oracle Data Integrator drive data transformation workflows on Hadoop
- Automatically transform R programs into Hadoop jobs
- Process large volumes of XML files in parallel and load XQuery results into the database
- Access data in HDFS securely with Kerberos authentication

KEY BENEFITS

- Quickly deliver data discovery applications to business users

  Query data in-place in Hadoop with Oracle SQL
- Extremely fast data loading between Hadoop and Oracle Database while minimizing database CPU utilization during load
- Enable data scientists to use R on data in Hadoop and combine with advanced analytics in the database
- Process extremely large volumes of XML data in Hadoop
- Reduce the complexities of Hadoop through graphical tooling
- Integrated and tested on Big Data Appliance
- Easy-to-use for Hadoop and Oracle Developers

Oracle Big Data Connectors is a software suite that integrates processing in Hadoop with operations in a data warehouse. Designed to leverage the latest features of Apache Hadoop, Big Data Connectors connect Hadoop clusters with database infrastructure to harness massive volumes of structured and unstructured data for critical business insights. Big Data Connectors greatly simplify development and are optimized for efficient connectivity and high-performance between Oracle Big Data Appliance and Oracle Exadata. Oracle Big Data Connectors 3.0 delivers a rich set of new features, increased connectivity, enhanced performance, and security for Big Data applications.

Oracle Big Data Connectors

Large volumes of data are increasingly collected and processed in Hadoop, while enterprise IT systems are centered on relational data warehouses. Oracle Big Data Connectors bridges data processing in Hadoop with Oracle Database, providing the crucial ability to unify data across these systems. Combining pre-processing of large data volumes of raw and unstructured data in Hadoop with the advanced analytics, complex data management, and real-time query capabilities of Oracle Database, Oracle Big Data Connectors deliver features that support information discovery, deep analytics and fast integration of all data in the enterprise. The components of this software suite are:

- Oracle SQL Connector for Hadoop Distributed File System
- Oracle Loader for Hadoop
- Oracle Data Integrator Application Adapter for Hadoop
- Oracle R Advanced Analytics for Hadoop
- Oracle XQuery for Hadoop

Oracle Big Data Connectors work with Oracle's engineered systems - Oracle Big Data Appliance and Oracle Exadata - as well as with supported Hadoop distributions and database versions on non-engineered systems.

Oracle SQL Connector for Hadoop Distributed File System

Oracle SQL Connector for Hadoop Distributed File System (HDFS) is a high-speed connector for loading or querying data in Hadoop from Oracle Database. Oracle SQL Connector for HDFS pulls data into the database; the data movement is initiated by selecting data via SQL in Oracle Database. Users can load data into the database, or query the data in-place in Hadoop, with Oracle SQL via external tables. The load speed from Oracle Big Data Appliance to Oracle Exadata is 15 TB/hour. Full query access using all of Oracle SQL enables users to apply the richest SQL in the industry to data in stored both in Hadoop and in Oracle Database.

Oracle SQL Connector for HDFS can query or load data in text files or Hive tables over text files. Partitions can be pruned while querying or loading from Hive partitioned tables. Oracle SQL Connector for HDFS has the ability to query or load Oracle Data Pump files generated

RELATED PRODUCTS

The following are related products available from Oracle:

- Oracle Big Data Appliance
- Oracle Exadata
- Oracle NoSQL Database
- Oracle Exalytics
- Oracle Business Intelligence Enterprise Edition
- Oracle Endeca Information Discovery
- Oracle Data Integrator

by Oracle Loader for Hadoop. When compared to simple text files, loading and querying Data Pump files delivers a 4x reduction in use of database CPU resources, as data has been transformed into Oracle binary types while generating the Data Pump files.

Oracle SQL Connector for Hadoop Distributed File System	
Features	
Oracle SQL access to data in Hadoop	Query Hive tables and text files in Hadoop directly from Oracle Database
Partition-aware access of Hive partitioned tables	Load or query only partitions of interest from Hive partitioned tables
Parallel query and load	Fast, efficient parallel query and load into Oracle Database
Security	Authenticated access with Kerberos on Oracle Big Data Appliance
Flexible and easy to use	Automatic creation of external tables
Input Formats	Text files, Hive tables over text files, Oracle Data Pump files generated by Oracle Loader for Hadoop

Oracle Loader for Hadoop

Oracle Loader for Hadoop is a high performance and efficient connector to load data from Hadoop into Oracle Database. Oracle Loader for Hadoop pushes data into the database; data transfers are initiated in Hadoop. Oracle Loader for Hadoop takes advantage of Hadoop compute resources to sort, partition, and convert data into Oracle-ready data types before loading. Pre-processing data on Hadoop reduces database CPU usage when loading data. This minimizes impact on database applications and alleviates competition for resources, a common issue when ingesting large data volumes. It makes the connector particularly useful for continuous and frequent loads.

Oracle Loader for Hadoop uses an innovative sampling technique to intelligently distribute data across reducer tasks that load data into the database in parallel. This minimizes the performance effects of data skew, a common concern in parallel applications.

Oracle Loader for Hadoop can load data from a wide range of input formats and input sources. Natively it can load data from text files, Hive tables, log files parsed by a regular expression, and Oracle NoSQL Database. When loading from Hive partitioned tables partitions of interest can be selectively loaded. Through integration with Hive, Oracle Loader for Hadoop can load from a variety of input formats accessible to Hive (example, JSON files) and input sources (example, HBase). In addition, Oracle Loader for Hadoop can read proprietary data formats through custom input format implementations provided by the user.

Oracle Loader for Hadoop	
Features	
Offload data pre-processing to Hadoop	Minimized impact on database CPU during load
Parallel load	Load into the database in parallel from nodes in the Hadoop cluster
Load balancing	Automatic even distribution of load across reducer tasks if there is data skew

Security	Authenticated access with Kerberos on Oracle Big Data Appliance
Online and offline load option	Connect to the database for online load or create Oracle Data Pump files for copy and offline load to non-local database
Input formats	Load data from text files, Hive tables (any input format or source accessible in Hive), log files parsed by a regular expression, Oracle NoSQL Database, and custom formats.
Partition-aware load	Load only partitions of interest from Hive partitioned tables

Oracle Data Integrator Application Adapter for Hadoop

Oracle Data Integrator (ODI) Application Adapter for Hadoop provides native Hadoop integration within ODI. Specific ODI Knowledge Modules optimized for operations in Hadoop are included within ODI Application Adapter for Hadoop. The 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.

Hadoop implementations oftentimes require complex Java MapReduce code to be written and executed on the Hadoop cluster. Using ODI and the ODI Application Adapter for Hadoop, developers use a graphical user interface to create these programs. ODI generates optimized HiveQL which in turn generates native MapReduce programs that are executed in Hadoop.

Oracle Data Integrator Application Adapter for Hadoop	
Features	
Optimized for Developer Productivity	• Familiar ODI graphical user interface • End-to-end coordination of Hadoop jobs • MapReduce jobs created and orchestrated by ODI
Native Integration with Hadoop	• Native integration with Hadoop using Hive • Ability to represent Hive metadata within ODI • Transformations and filtering occur directly in Hadoop • Transformations written in SQL-like HiveQL
Optimized for Performance	• Optimized Hadoop ODI knowledge modules • High Performance load to Oracle Database using ODI with Oracle Loader for Hadoop and Oracle SQL Connector for HDFS

Oracle R Advanced Analytics for Hadoop

Oracle R Advanced Analytics for Hadoop runs R code in a Hadoop cluster for scalable analytics. Oracle R Advanced Analytics for Hadoop accelerates advanced analytics on Big Data by hiding the complexities of Hadoop-based computing from R end users. The connector integrates with Oracle Advanced Analytics for Oracle Database, to execute R and in-database Data Mining computations directly in the database.

Oracle R Advanced Analytics for Hadoop delivers faster insights by including a rich collection of high performance, scalable, parallel implementations of common statistical and predictive techniques, leveraging the Hadoop cluster without requiring data duplication or data movement. A complete list of supported techniques is in the table below. Transparent scalability is enabled by executing R code from stand-alone desktop applications, developed in any IDE the R user chooses, in parallel in Hadoop. Oracle R Advanced Analytics for Hadoop enables rapid development with R-style debugging capabilities of parallel R code on user desktops, supported by simulating parallelism under the covers.

The connector enables analysts to combine data from several environments - client desktop, HDFS, HIVE, Oracle Database and in-memory R data structures - all in the context of a single analytic task execution, greatly simplifying data assembly and preparation. Oracle R Advanced Analytics for Hadoop also provides a general computation framework for execution of R code in parallel. The I/O performance of R-based MapReduce jobs matches that of pure Java based MapReduce programs with the support of binary RData representation of input.

Oracle R Advanced Analytics for Hadoop	
Features	
Scalable, distributed analytics for Big Data	• Native distributed R analytics in Hadoop for transparent execution of R code in parallel • Support for Hive and text input data stores
Ease-of-use and rapid deployment without requiring new skill sets	• Developer productivity: R code developed and debugged in a familiar R environment on a user's desktop without the need for parallel computing skills • Simplified interfaces allow R users to leverage Hadoop's map-combine-reduce data flows • Support for Hybrid data assembly and scalable data preparation
Native distributed R analytics	• Statistics and Advanced Matrix Computation ○ Covariance and Correlation matrix computation ○ Reservoir Sampling ○ Principal Component Analysis ○ Matrix completion using low rank matrix factorization ○ Non negative matrix factorization • Regression Models ○ Linear regression ○ Single layer feed forward Neural Networks ○ Generalized linear models • Classification Models ○ Logistic regression based on generalized linear models • Segmentation using k-Means clustering

Oracle XQuery for Hadoop

Oracle XQuery for Hadoop enables XQuery to be used to process and transform text, XML,

JSON and Avro content stored in a Hadoop Cluster. Oracle XQuery for Hadoop takes full advantage of the large numbers of CPUs present in a typical cluster, evaluating XQuery operations in a massively parallel manner.

Oracle XQuery for Hadoop is based on a Hadoop optimized Java implementation of Oracle Database's proven XQuery engine. The XQuery engine automatically evaluates standard W3C XQuery expressions in parallel, leveraging the MapReduce framework to distribute an XQuery expression to all nodes in the cluster. This enables XQuery expressions to be evaluated by taking the processing to the data, rather than bringing the data to the XQuery processor. This method of query evaluation delivers much higher throughput than is available with other XQuery solutions.

Typical use cases for Oracle XQuery for Hadoop include web log analysis and transformation operations on text, XML, JSON, and Avro content. After processing data can be loaded into the database or indexed with Cloudera Search.

Oracle XQuery for Hadoop	
Features	
Scalable, Native, XQuery Processing	XQuery engines are automatically distributed across the Hadoop cluster, so XQueries execute where the data is located
Hadoop Input Data Stores	Process data stored in HDFS, Hive or Oracle NoSQL Database
Integration with Hadoop technologies	Execute Oracle XQuery for Hadoop jobs from Apache Oozie workflows Cloudera Search
Parallel XML Parsing	Very large XML documents can be processed extremely efficiently
Fast Load of XQuery Results into Oracle Database	Fast load of XQuery results into Oracle Database using Oracle Loader for Hadoop

Contact Us

For more information about Oracle Big Data Connectors, visit oracle.com or call +1.800.ORACLE1 to speak to an Oracle representative.

Oracle is committed to developing practices and products that help protect the environment

Copyright © 2014, Oracle and/or its affiliates. All rights reserved.

This document is provided for information purposes only and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Cloudera, Cloudera CDH, and Cloudera Manager are registered and unregistered trademarks of Cloudera, Inc. Other names may be trademarks of their respective owners.

Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. UNIX is a registered trademark licensed through X/Open Company, Ltd. 0611

Hardware and Software, Engineered to Work Together