ORACLE DATA SHEET

Oracle Big Data Connectors

ORACLE
BIG DATA

Oracle Big Data Connectors is a software suite that integrates processing in Apache Hadoop distributions with operations in Oracle Database. It enables the use of Hadoop to process and analyze massive volumes of data and to use it with database data to derive new and critical business insights. Optimized for Hadoop and Oracle Database, Oracle Big Data Connectors include products for fast data load, data access, and R and XML processing of data in Hadoop. They are easy to use with existing skill sets, simplifying development of Big Data solutions. Available for on-premise and cloud deployments, Oracle Big Data Connectors deliver a rich set of features, security, and high speed connectivity for Big Data applications.

"Oracle Big Data Appliance was especially attractive because of its extensibility, and by using Oracle Big Data Connectors, we could move data seamlessly between Oracle Big Data Appliance and Oracle Database Appliance, which supports high throughput and low latency."

CRAIG FRYAR
HEAD OF BUSINESS INTELLIGENCE
WARGAMING

ON-PREMISE AND CLOUD

- Connect Hadoop with Oracle Database on-premise and in the cloud.
- Optimized for Oracle Big Data Appliance and Oracle Big Data Cloud Service.

KEY BUSINESS BENEFITS

- Quickly deliver data discovery applications to business users.
- High performance data load from Hadoop to Oracle Database.
- Access Hadoop data from Oracle

Oracle Big Data Connectors

Big Data solutions enable enterprises to process large volumes of raw data and exploit new analyses to get multi-dimensional insights into their businesses. These solutions use Hadoop for large-scale data processing, data transformation, and exploratory analysis of data, integrating the results with business critical data in Oracle Database for real-time queries, advanced analytics, and complex data management. Oracle Big Data Connectors connect Hadoop with Oracle Database, providing an essential infrastructure for Big Data solutions. They include tools for fast data load from Hadoop to Oracle Database, data access between Hadoop and Oracle Database, and R and XML analytics in Hadoop, enabling information discovery, deep analytics, and fast integration of all data in the enterprise. The components of this software suite are:

- Oracle Datasource for Apache Hadoop **(new)**
- Oracle SQL Connector for Hadoop Distributed File System (HDFS)
- Oracle Loader for Hadoop
- Oracle Advanced Analytics for Hadoop
- Oracle XQuery for Hadoop
- Oracle Data Integrator[1]

Oracle Big Data Connectors work with Oracle engineered systems, connecting Oracle Big Data Appliance with Oracle Exadata, Oracle SuperCluster, and Oracle Database Appliance, and with certified Hadoop distributions and Oracle Databases on commodity

[1] Oracle Big Data Connectors includes a restricted use license for Oracle Data Integrator only on Oracle Big Data Appliance. Additional licensing is required on other Hadoop platforms.

ORACLE

Database and Oracle Database tables from Hadoop.

- Enable data scientists to use their favorite R IDE and run R commands on Hadoop and Spark for extremely fast analytics.
- Simplify machine learning in R by calling Spark MLib APIs with a single line of R code.
- Allow XML application developers to use the familiar XQuery language to process XML data in Hadoop.
- Reduce the complexities of Hadoop through graphical tooling.
- Quick development and deployment of Big Data solutions with easy-to-use familiar tools for Hadoop and Oracle Developers.

KEY FEATURES

- Tight integration with Oracle Database.
- Leverage Hadoop compute resources.
- Enable Hive to directly access data in Oracle Database tables.
- Use Oracle SQL to access and load data in HDFS.
- Fast load from Hadoop into Oracle Database with minimal use of database CPU.
- Partition pruning of Hive tables during load and query.
- Graphical user interfaces of Oracle Data Integrator to drive data transformation workflows in Hadoop.
- Automatically transform R programs into Spark or Hadoop jobs.
- Process large volumes of XML files in parallel and load XQuery results into Oracle Database.
- Secure access of HDFS data with Kerberos authentication.
- Certified with multiple database and Hadoop versions on Oracle's engineered systems and commodity hardware.

hardware. Click here for the latest certifications or search for Oracle Big Data Connectors in your favorite search engine and click on the certifications tab.

Oracle Big Data Connectors in the Cloud

Oracle Big Data Connectors can be used with on-premise and cloud installations of Hadoop and Oracle Database. It is part of Oracle Big Data Cloud Service.

Oracle Datasource for Apache Hadoop

Oracle Datasource for Apache Hadoop (formerly Oracle Table Access for Apache Hadoop) turns Oracle Database tables into a Hadoop data source (i.e., external table) enabling direct and consistent Hive QL/Spark SQL queries, as well as direct Hadoop API access. Applications can join master data or dimension data in Oracle Database with data stored in Hadoop. Additionally data can be written back to Oracle Database after processing.

Oracle Datasource for Apache Hadoop optimizes a query's execution plans using predicate and projection pushdown, and partition pruning. Database table access is performed in parallel based on the selected split patterns, using smart and secure connections (Kerberos, SSL, Oracle Wallet), regulated by both Hadoop (i.e., maximum concurrent tasks) and Oracle DBAs (i.e., max pool size).

FEATURES	
Hive access to Oracle Database tables	Query Oracle Database tables directly with Hive
Optimized query execution plans	Predicate pushdown and partition pruning to retrieve only relevant data
No intermediate data type conversion	Directly convert SQL data types to/from Hadoop data types
Parallel access to data	High speed access to Oracle Database tables with intelligent split patterns
Secure access	Support for Kerberos, SSL, Oracle Wallet
Write data back to Oracle Database	Save the results of Hadoop processing in Oracle Database for use in database analytics and BI applications

Oracle SQL Connector for Hadoop Distributed File System (HDFS)

Oracle SQL Connector for HDFS allows you to query of Hadoop resident data from the database using Oracle SQL. The data is accessed via external tables, which can be queried like any other table in the database. Data can also be loaded by selecting data from the external table and inserting it into a table in the database. The load speed from

Oracle Big Data Appliance to Oracle Exadata is 15 TB/hour.

Oracle SQL Connector for HDFS can query or load data in text files or Hive tables over text files. When querying from a Hive partitioned table, Oracle SQL Connector for HDFS can be restricted to access a subset of Hive partitions, minimizing the data accessed for faster performance. Oracle SQL Connector for HDFS can also query files in Oracle Data Pump format, such as data files created by Oracle Loader for Hadoop for offline loading and data files copied from the database to Hadoop for archival.

FEATURES	
Oracle SQL access to data in Hadoop	Query Hive tables and text files in HDFS directly from Oracle Database
Partition-aware access of Hive partitioned tables	Load or query only relevant Hive partitions, enabling partition pruning
Parallel query and load	High speed parallel query and load into Oracle Database
Security	Authenticated access with Kerberos
Flexible and easy to use	Tool to create external tables
Access Oracle Data Pump files in HDFS with Oracle SQL	Archive database data as data pump files in Hadoop and query from Oracle Database

Oracle Loader for Hadoop

Oracle Loader for Hadoop is a high performance load tool for fast movement of data from Hadoop to Oracle Database. Oracle Loader for Hadoop takes advantage of Hadoop compute resources to sort, partition, and convert data into Oracle types on Hadoop before loading into the database. Transforming the data into Oracle types reduces database CPU usage during the load, minimizing impact on database applications. It alleviates competition for resources, a common issue when ingesting large data volumes. This feature makes this connector particularly useful for continuous and frequent loads.

Oracle Loader for Hadoop uses an innovative sampling technique to intelligently distribute data across Hadoop nodes while loading data in parallel. This minimizes the effects of data skew, a concern in parallel applications.

Oracle Loader for Hadoop loads into tables with any database compression: Basic Table compression, Advanced Row compression, and Hybrid Columnar Compression.

Oracle Loader for Hadoop can load a wide variety of input formats. Natively it can load data from text files, Hive tables, log files (parse and load), Oracle NoSQL Database, and more. Through Hive it can also load from input formats (e.g.: Parquet, JSON files) and input sources (e.g.: HBase) accessible to Hive. In addition, Oracle Loader for Hadoop can read proprietary data formats through custom input format implementations provided by the user.

ORACLE BIG DATA CONNECTORS

Oracle Big Data Connectors provides high performance integration between Hadoop and Oracle Database.

RELATED PRODUCTS

The following are related products available from Oracle

- Oracle Big Data Appliance
- Oracle Exadata
- Oracle NoSQL Database
- Oracle Big Data Discovery
- Oracle Exalytics
- Oracle Business Intelligence
- Oracle Data Integrator

RELATED SERVICES

The following services support Oracle Main Product:

- Update Subscription Services
- Product Support Services
- Professional Services

Features	
Offload data type conversion to Hadoop	Minimize impact on database CPU
Parallel load	High speed parallel load leveraging Hadoop nodes
Load balancing	Automatic load balancing to address data skew
Security	Authenticated access with Kerberos
Input formats	Load from a variety of input formats: text files, Hive, Parquet, JSON, sequence files, compressed files, log files, and more
Partition-aware load	Load only partitions of interest from Hive partitioned tables

Oracle R Advanced Analytics for Hadoop

Oracle R Advanced Analytics for Hadoop runs R code in Spark and Hadoop for extremely scalable analytics. It hides from the R user the complexities of using Spark MLib and Hadoop-based distributed computing. R users develop applications using R and Spark MLib in any IDE (R client) of their choosing, on their desktop or laptop, and the connector transforms and executes the code in parallel in Hadoop or in Spark. Several R functions are implemented in Spark, for a 200-300x speedup over MapReduce implementations of R analytics.

Oracle R Advanced Analytics for Hadoop additionally integrates with Oracle Advanced Analytics in Oracle Database, to execute R and in-database Data Mining computations directly in the database.

A unique innovation integrates Spark MLib with R, enabling one line of R code to replace dozens of lines of Java code in machine learning applications.

The connector enables analysts to combine data from several environments – client desktop, HDFS, Hive, Oracle Database and in-memory R and Spark data structures – all in the context of a single analytic task execution, greatly simplifying data assembly and preparation.

Oracle R Advanced Analytics for Hadoop enables faster insights with the rich collection of scalable, high performance, parallel implementations of common statistical, predictive, and machine learning techniques in Hadoop, without requiring data movement to any other platform. A complete list of supported techniques is in the table below.

Features	
Scalable, distributed analytics for Big Data	Native distributed R analytics in Hadoop for transparent execution of R code in parallel
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 parallelism Support for hybrid data assembly and scalable data preparation
Spark MLib available in R	Regression Models • Linear regression • Least Absolute Shrinkage and Selection Operator • Ridge Regression • Logistic Regression Decision Trees Random Forest Support Vector Machines k-Means Clustering Principal Component Analysis
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 the use of XQuery to process and transform text, XML, JSON and Avro content stored in Hadoop. It takes advantage of the parallelism of Hadoop to evaluate W3C XQuery expressions, using all nodes in the cluster.

Oracle XQuery for Hadoop is based on a Java implementation of the proven Oracle Database XQuery engine that has been optimized for Hadoop. The XQuery engine transparently evaluates XQuery expressions in parallel in Hadoop, distributing an XQuery expression to all nodes in the cluster. This takes XQuery 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 on the Hadoop platform.

Features	
Scalable, native, XQuery processing	XQuery engines are automatically distributed across the Hadoop cluster, so XQueries execute where the data is located
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 XML/XQuery extensions for Hive
Parallel XML parsing	Process very large XML documents extremely fast
Fast load of XQuery results into Oracle Database	Fast load using Oracle Loader for Hadoop

Oracle Data Integrator Enterprise Edition

Oracle Data Integrator (ODI) has a comprehensive set of knowledge modules for native Hadoop integration within ODI. The modules allow complex data transformations and data movement operations to be specified and executed through a familiar graphical interface, greatly simplifying running jobs in Hadoop. Data movement from one source to another can be graphically defined (e.g.: from HBase to Oracle Database, HBase to Hive, Hadoop to third party databases and from third party databases to Hadoop, etc.). Knowledge modules for Oracle Loader for Hadoop and Oracle SQL Connector for HDFS provide high speed data movement from Hive and HDFS to Oracle Database. With graphical tooling and declarative specifications Oracle Data Integrator knowledge modules eliminate the need to write complex code typically required for Hadoop applications.

Oracle Big Data Connectors includes a restricted use license for Oracle Data Integrator when licensed on Oracle Big Data Appliance. Additional licensing is required on other Hadoop platforms.

Features	
Optimized for developer productivity	Familiar ODI graphical user interface End-to-end coordination of Hadoop jobs Hadoop jobs created and orchestrated by ODI
Native integration with Hadoop	Ability to represent Hive metadata within ODI Transformations and filtering occur directly in Hadoop
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

.

CONTACT US

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