ORACLE DATA SHEET

Oracle Big Data SQL

Release 4.1

ORACLE BIG DATA

The unprecedented explosion in data that can be made useful to enterprises – from the Internet of Things, to the social streams of global customer bases – has created a tremendous opportunity for businesses. However, with the enormous possibilities of Big Data, there can also be enormous complexity. Integrating Big Data systems to leverage these vast new data resources with existing information estates can be challenging. Valuable data may be stored in a system separate from where the majority of business-critical operations take place. Moreover, accessing this data may require significant investment in re-developing code for analysis and reporting - delaying access to data as well as reducing the ultimate value of the data to the business.

Oracle Big Data SQL enables organizations to immediately analyze data across Apache Hadoop, Apache Kafka, NoSQL, object stores and Oracle Database leveraging their existing SQL skills, security policies and applications with extreme performance. From simplifying data science efforts to unlocking data lakes, Big Data SQL makes the benefits of Big Data available to the largest group of end users possible.

KEY FEATURES

- Seamlessly query data across Oracle Database, Hadoop, object stores, Kafka and NoSQL sources
- Runs all Oracle SQL queries without modification – preserving application investment
- Smart Scan on Hadoop, Kafka, NoSQL and object store enhance scalability and performance by processing data using fan-out parallelism
- Enables Oracle Database 19c access to leading Hadoop distributions
- Oracle Database Security features provide single access control to sensitive data across Oracle Database, Hadoop, object stores, Kafka and NoSQL data
- Easily copy data from Oracle

Rich SQL Processing on All Data

Oracle Big Data SQL is a data virtualization innovation from Oracle. It is a new architecture and solution for SQL and other data APIs (such as REST and Node.js) on disparate data sets, seamlessly integrating data in Apache Hadoop, Apache Kafka, object stores and a number of NoSQL databases with data stored in Oracle Database. Using Oracle Big Data SQL, organizations can:

- **Use Oracle SQL to query and analyze data** in Apache Hadoop, object stores, Apache Kafka and NoSQL
- **Maximize query performance** on all data using advanced techniques like Smart Scan, Aggregation Offload, Partition Pruning, Storage Indexes, Bloom Filters and Predicate Push-Down in a distributed architecture
- **Integrate big data** analyses into existing applications and architectures
- **Extend security** and access policies from Oracle Database to data in Apache Hadoop, object stores, Apache Kafka and NoSQL

ORACLE

Database to Hadoop using Copy to Hadoop

KEY BENEFITS

- Transparently analyze data sets across Hadoop, object stores, Kafka, NoSQL and Oracle Database
- Achieve fast query performance by leveraging data local processing
- Use your existing SQL skills to analyze data across big data sources
- Current SQL-based applications can seamlessly integrate new data
- Seamlessly extend Information Lifecycle Management strategy to leverage lower cost Hadoop and object storage
- Use Oracle Database security policies to keep all sensitive data safe

Enhanced External Tables

When dealing with large data sets stored in disparate systems, it can be difficult to know where your data is, let alone understand how the data is structured. Big Data SQL uses Oracle Database 19c big data-enabled external tables, which give users a single location to catalog and secure data in Hadoop, object stores, Kafka and NoSQL systems: the Oracle Database. Big Data SQL keeps track of the metadata about external data sources – both clusters and the tables within them – without moving or copying data. External tables for Big Data SQL provide:

- **Seamless metadata integration and queries** which join data from Oracle Database with data from Hadoop, object stores, Kafka and NoSQL databases
- **Automatic mappings** from metadata stored in HCatalog (or the Hive Metastore) to Oracle Tables
- **Multiple cluster support** to allow one Oracle Database to query multiple Hadoop clusters
- **Enhanced access parameters** to give database administrators the flexibility to control column mapping and data access behavior

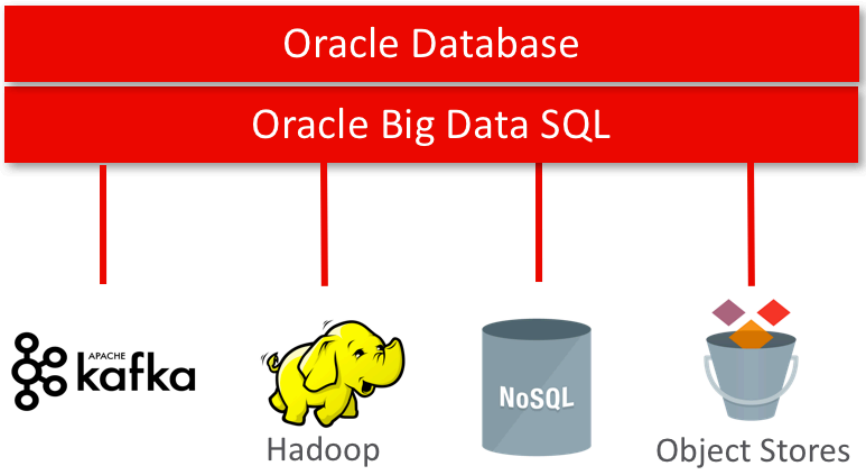

Figure 1. Oracle Big Data SQL enables Oracle SQL queries to span Oracle Database, Apache Hadoop, Apache Kafka, selected NoSQL data stores and object stores

Smart Scan: Data-Driven Parallel Processing

Finding insights from Big Data can mean sifting through an extraordinary amount of data. With the massive increase in data volumes that Big Data brings, analytical performance can only be achieved by moving the analytics to the data, not the other way around. Big Data SQL applies the power of Smart Scan, first introduced in Oracle's best-in-class Exadata Database Machine, to big data stores. Smart Scan enables Oracle SQL operations to be pushed down to the storage tier of the Big Data system. Paired with the horizontal scalability of these storage systems, Smart Scan automatically provides parallel processing equal to your biggest data set, enabling:

- **Locally filtered data**, so that only the rows and columns relevant to your query are transmitted to Oracle Database

- **Join optimization** via Bloom filters and key vectors, speeding up joins between data in Oracle Database and massive amounts of external data
- **Distributed Aggregation,** utilizing the compute capacity of the Hadoop cluster to aggregate data locally and returning summarized data back to the Oracle Database
- **Scoring** for data mining models and enhanced processing for querying document data sets in for example JSON or XML
- **Oracle-native operators** providing complete fidelity between queries run with Big Data SQL and Oracle Database alone

Storage Indexing: More Effective I/O

In addition to the set of Smart Scan features, Oracle Big Data SQL provides Storage Index technology to speed up processing before any I/O occurs. As data is accessed, Oracle Big Data SQL automatically builds local, in-memory indexes that capture where relevant data is stored. On subsequent queries of the same data, Storage Index technology ensures that data blocks that are not relevant to the query are not read. Because data blocks in Big Data systems can be very large (up to hundreds of megabytes), this "I/O skipping" strategy can improve performance on some queries by orders of magnitude.

Predicate Push-Down: Harness External Storage Systems

Oracle Big Data SQL not only enables easy integration of data from Hadoop and NoSQL sources, Big Data SQL also leverages the underlying storage mechanisms to provide the best possible performance. Big Data SQL's *predicate push down* technology allows predicates in queries issued in Oracle Database to be executed by remote systems, and to be pushed into certain file formats. Using predicate push down, Big Data SQL enables you to:

- **Prune partitions** from tables managed by Apache Hive
- **Minimize I/O** on files stored in Apache Parquet and Apache ORC formats
- **Enable remote reads** on data stored in Oracle NoSQL Database or Apache HBase

Distributed Aggregation: Faster Summary Queries

Oracle Big Data SQL utilizes Oracle In-Memory technology to push SQL aggregations to the Oracle Big Data SQL cells. This enables Oracle Big Data SQL to leverage the processing power of the Hadoop cluster for distributing aggregations across the cluster nodes. This has the potential for significant performance gains – oftentimes exceeding an order of magnitude improvement.

Distributed aggregations apply to queries involving both single tables and multi-table joins. Common "star" and "snowflake" queries typically join smaller dimension tables to large fact tables. Oracle Database In-Memory Aggregation utilizes efficient key vector processing as part of Smart Scan to aggregate data as part of the scan operation – greatly accelerating the processing times for eligible queries.

Query Streams: SQL Access to Kafka Topics

Apache Kafka is a distributed, scalable, fault tolerant messaging system. Organizations utilize Kafka as a central hub for delivering real-time streams of data. Instead of systems communicating directly with one another, applications publish messages to a Kafka topic – and these messages are then consumed by other applications. Big Data

SQL supports direct access to Kafka topics – enabling SQL queries to combine near real-time events with data from Oracle Database and big data stores.

Query Server Simplifies SQL Access to Hadoop Data

Oracle Big Data SQL Query Server enables applications to query data in Hadoop without requiring a separate Oracle Database. Query Server is an Oracle Database SQL engine that is automatically installed and configured on an edge node of your Hadoop cluster. It requires zero maintenance; metadata and authorization rules are inherited from the Hive metastore and HDFS. Big Data SQL external tables are automatically synchronized with your selected Hive databases. Data authorization leverages Apache Sentry and HDFS access controls.

Query Server is intended for SQL on Hadoop processing. It augments Big Data SQL deployments that are integrated with the Oracle Database; it does not persist data as all queries are targeted at data stored on the Hadoop cluster. To run queries that combine data in Oracle Database with external sources, then Oracle Database with Big Data SQL is the right alternative. Query Server is an ideal solution when all your data resides in Hadoop and you want to leverage Oracle Database's rich SQL language and query execution capabilities.

Extend Information Lifecycle Management to Hadoop

For many years, Oracle Database has provided rich support for Information Lifecycle Management (ILM). Numerous capabilities are available for data tiering – or storing data in different media based on access requirements and storage cost considerations. These tiers may scale from 1) in-memory for real time data analysis, 2) Database Flash for frequently accessed data, 3) Database Storage and Exadata Cells for queries of operational data and 4) Hadoop and object stores for infrequently accessed raw and archive data:

Copy to Hadoop

Copying data from Oracle Database to Hadoop can be complicated. Oracle Big Data SQL includes the Oracle Copy to Hadoop utility. This utility simplifies copying Oracle data to the Hadoop Distributed File System (HDFS). Data copied to the Hadoop cluster by Copy to Hadoop is stored in Oracle Data Pump format. This format optimizes queries thru Big Data SQL: 1) the data is stored as Oracle data types – eliminating data type conversions and 2) the data is queried directly – without requiring the overhead associated with Java SerDes. Native Hadoop tools like Hive can easily access these same Oracle Data Pump export files using optimized input format classes.

Hybrid Partitioned Tables

Oracle Partitioning is the enabling technology that allows a single table's data partitions to be stored on the various tiers. This enables immutable archive data within a table to reside in Hadoop or object stores using a variety of file formats (Apache Parquet, Apache ORC, Apache Avro and text). Because the data is in open formats, it is not reserved for Oracle Database processing. The data may be shared with any other application in your data lake. Database queries seamlessly access this archive data as they would any other data.

In addition, Big Data SQL's Smart Scan capabilities enable compound performance

benefits. Big Data SQL Smart Scan utilizes the massively parallel processing power of the Hadoop cluster to filter data at its source – greatly reducing data movement and network traffic between the cluster and the database.

Oracle Database Security on Big Data

Oracle Big Data SQL's unique approach to integrating data enables applications to automatically leverage underlying data authorization rules (i.e. access privileges on files in HDFS and Apache Sentry policies on Hive metadata) and then layer on top of that advanced Oracle Database Security policies. This approach both simplifies secure implementations and enables the utilization of Oracle security features that are unavailable on underlying stores. Using Oracle security mechanisms, you can secure Big Data using:

- Standard Oracle Database roles and privileges to govern access to data
- Data redaction, to ensure that sensitive information is obscured when accessed by unauthorized users
- Virtual Private Databases to better enforce governance policies
- Oracle Database Vault to protect sensitive data from privileged accounts
- Oracle Database Security Assessment Tool to identify potentially sensitive data in underlying stores – in addition to the overall database security status

Supports a Range of Big Data Deployments

Oracle Big Data SQL is designed to support a wide range of deployment options and platforms. Big Data SQL requires 1) Oracle Database running Enterprise Linux and 2) leading Apache Hadoop distributions from Cloudera and Hortonworks. Big Data SQL achieves highest performance when paired with Oracle Engineered. Big Data SQL takes full advantage of the power of Oracle Exadata and Oracle Big Data Appliance to create a best-in-class Big Data Management System, unifying the power of big data and Oracle Database

Oracle Database Version	Database Hardware	Hadoop Cluster Hardware	Hadoop Distribution and Version
19c	Oracle Exadata (Linux OL6, OL7) or Any Intel x86 64-bit system (Linux OL6, OL7, RHEL6, RHEL7)	Any Intel x86 64-bit system (Linux OL6, OL7, RHEL6, RHEL7)	• CDH* 5.x (5.5 and higher), 6.x • HDP ** 2.x (2.3 and higher)
19c	Oracle Exadata (Linux OL6, OL7) or Any Intel x86 64-bit system (Linux OL6, OL7, RHEL6, RHEL7)	Oracle Big Data Appliance (Linux OL6 and OL7)	• CDH* 5.x (5.5 and higher), 6.x

* CDH: Cloudera's Distribution Including Apache Hadoop

** HDP: Hortonworks Data Platform

Getting Started

Try using Oracle Big Data SQL – as well as other components of Oracle's big data platform – in Oracle Big Data Lite Virtual Machine (http://www.oracle.com/technetwork/database/bigdata-appliance/oracle-bigdatalite-2104726.html). Big Data Lite allows you to test drive Oracle's big data capabilities from your laptop or desktop computer. Capabilities include CDH, Oracle Big Data Spatial and Graph, Oracle Big Data Discovery, Oracle Big Data Connectors, Oracle Data Integrator, Oracle Golden Gate and more.

ORACLE

CONTACT US

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