Oracle Big Data SQL brings SQL and Performance to Hadoop

Jean-Pierre Dijcks
Oracle
Redwood City, CA, USA

Keywords:

Big Data SQL, Hadoop, Big Data Appliance, SQL, Oracle, Performance, Smart Scan

Introduction

Oracle Big Data SQL is an innovation from Oracle only available on Oracle Big Data Appliance. It is a new architecture for SQL on Hadoop, seamlessly integrating data in Hadoop and NoSQL with data in Oracle Database. Using Oracle Big Data SQL, organizations can:

- Combine data from Oracle Database, Hadoop and NoSQL in a single SQL query
- Query and analyze data in Hadoop and NoSQL
- Integrate big data analysis into existing applications and architectures
- Extend security and access policies from Oracle Database to data in Hadoop and NoSQL
- Maximize query performance on all data using Smart Scan

Oracle Big Data SQL radically simplifies integrating and operating in the big data domain through two powerful features: newly expanded External Tables and Smart Scan functionality on Hadoop.

Overcoming Barriers to the Adoption of Big Data Technologies

While a host of new technologies have emerged to help process big data, both their rapid evolution and the inherently challenging task of application development for naively parallel distributed systems has led to a series of barriers to reliably producing value from big data. For those organizations that have found business cases likely to produce value to the business, significant barriers to actualizing this value exist: both in organizational skill sets and the integration of big data systems with existing enterprise information architectures.

To this end, there has been a resurgence of interest in the SQL language for managing and manipulating big data, particularly for manipulating data stored in the Hadoop ecosystem. Beyond its linguistic maturity, broad adoption, and deep integration with tools and applications, SQL is the natural language for working with data. The declarative, data-oriented nature of SQL requires that users describe the shape of the answers they seek from data, without having to deal with the complexities of how that data is accessed or processed. This enables SQL users to be much more productive, and focus their efforts on business problems, rather than complicated control flow.

Given the inherently complicated nature of programming for large-scale distributed systems, it is little wonder that there is tremendous interest for SQL access to data in Hadoop. Many groups are working to provide SQL access to data stored in Hadoop. However, while these efforts alleviate some of the skills barriers to realizing value from big data, a fundamental problem remains with respect to integration of the big data technologies within the broader enterprise architecture.

Specifically, in a big data environment, data lives in many places. While increasingly more data is stored in Hadoop, most business-critical data is stored in a relational database. To that end, what is needed is not simply SQL access to data stored in Hadoop, but seamless SQL access to data stored in Hadoop, various NoSQL databases, and relational databases. Moreover, to truly integrate value from big data, the roles, policies and information governance which already exist must be extended to data residing in these new data stores.

Oracle Big Data Management System

When Hadoop and NoSQL technologies were first becoming popular, some of their early supporters talked them up as replacements for the relational database. More recently, though, it's become clear that they are complementary.

So maximizing value from big data requires that you have Hadoop, NoSQL and relational database systems all nicely integrated together. And everybody from pioneers of Hadoop to large enterprises, from analyst firms to vendors, are saying pretty much the same thing: you need all these things working together. So when you have all these things working together, what do you have?

We think there's a simple answer. You've gone from managing data to managing big data, from a relational database management system to a big data management system that seamlessly incorporates Hadoop, NoSQL and your relational data warehouse (possibly other sources as well).

"Big Data Management System" is a totally generic term: it's what many organizations need to run their business in this new era of big data; and it's what vendors need to deliver or help their customers to acquire and build.

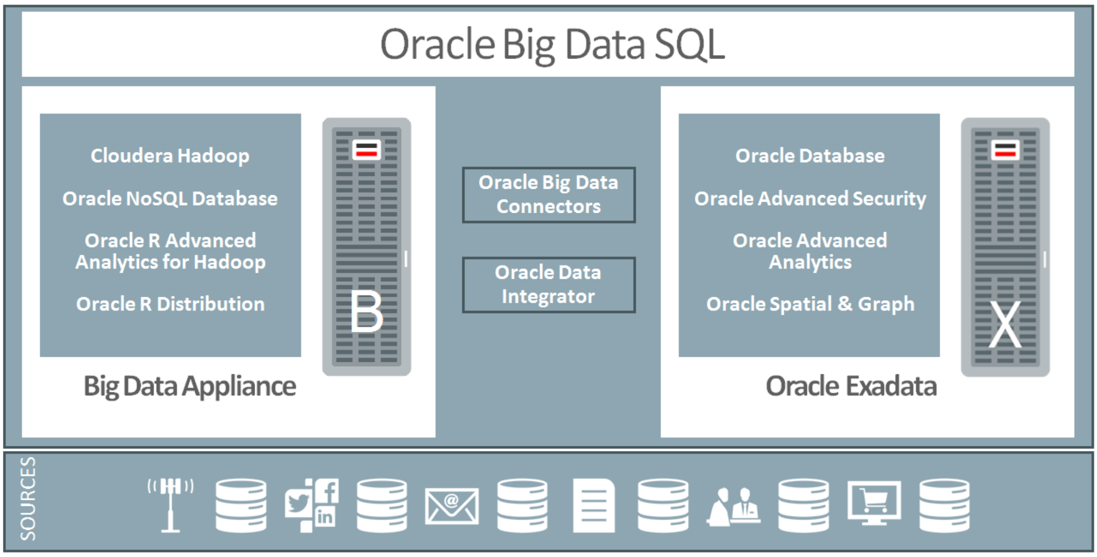


Figure 1. Oracle Big Data Management System

Putting a label on this thing, then, is not hard. What is harder, of course, is building one. The key to building the Big Data Management System is however provided by Engineered Systems and Oracle Software innovations.

In Oracle’s case Big Data Appliance combined with Oracle Exadata delivers a turn-key solution to lay the foundation of the Big Data Management System. Through the usage of InfiniBand networking, all nodes seem local to each other. Through the use of the same management console, Oracle Enterprise Manager, administration is very much unified and simplified.

But being able to reduce complexity through integration, through administration etc. is not sufficient. Data is still in addressed differently, security is not uniform and often spotty on Hadoop systems. In short, it is not just integration of hardware and networking that makes a Big Data Management System work. Software, and more specifically SQL integration is needed on top of the systems.

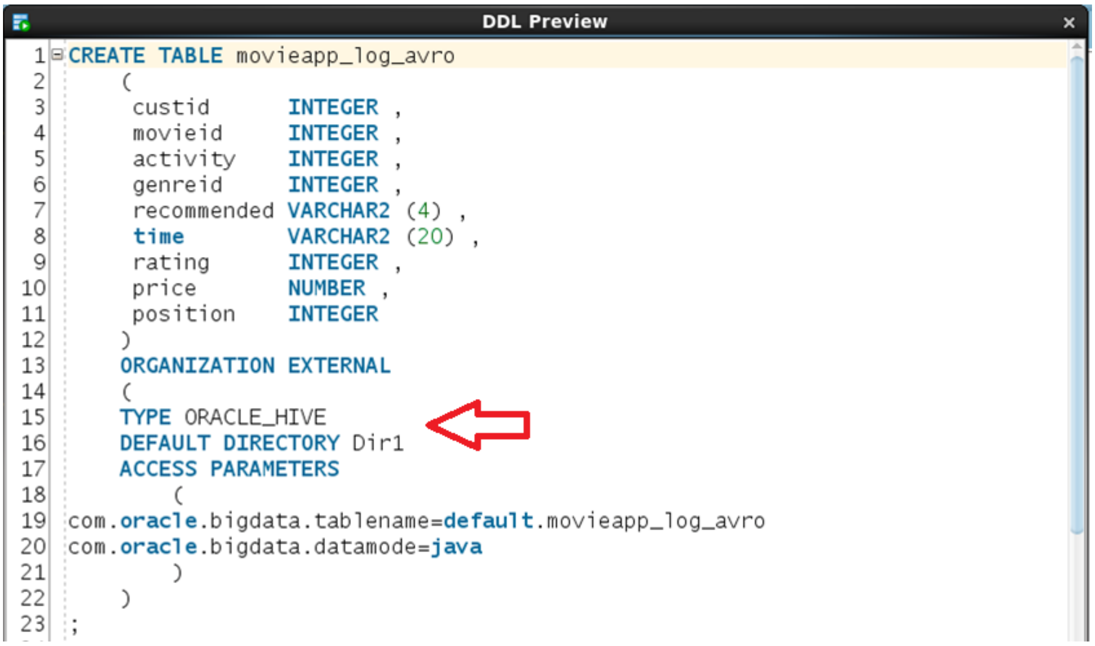
That is what Oracle Big Data SQL uniquely provides.

Oracle Big Data SQL – External Tables

Big Data SQL is Oracle’s breakthrough approach to simplifying access and integration to big data sources. Oracle Big Data SQL provides the ability to query all data – in Hadoop, NoSQL datastores, or Oracle Database – in a single SQL statement. Oracle Big Data SQL presents Hadoop and other sources as enhanced external tables, available as of Oracle Database 12.1.0.2. These tables are engineered to transparently map the external semantics of data access – horizontal parallelism, location, and schema – to Oracle internals.

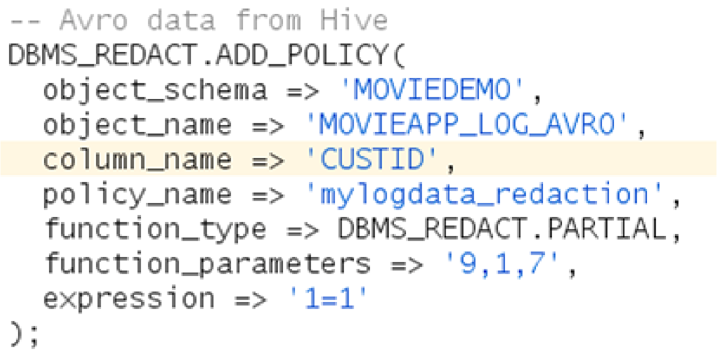
This mapping ensures the best possible optimizations for access and native processing throughout. Oracle Big Data SQL enables users to:

- Express their queries on all data using the world’s richest SQL dialect
- Integrate big data quickly into reports or applications using existing interfaces
- Extend existing Oracle security and access control policies to data stored in Hadoop



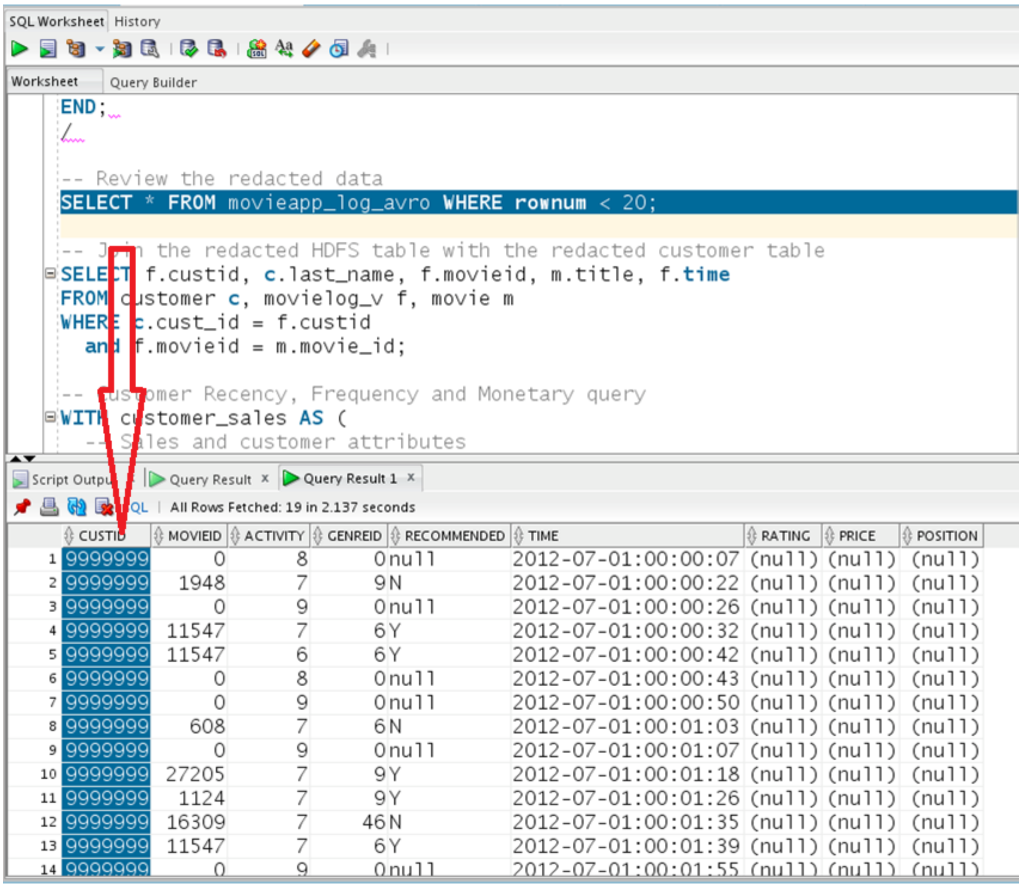
```
CREATE TABLE movieapp_log_avro
    (
     custid      INTEGER ,
     movieid     INTEGER ,
     activity    INTEGER ,
     genreid     INTEGER ,
     recommended VARCHAR2 (4) ,
     time        VARCHAR2 (20) ,
     rating      INTEGER ,
     price       NUMBER ,
     position    INTEGER
    )
    ORGANIZATION EXTERNAL
    (
    TYPE ORACLE_HIVE
    DEFAULT DIRECTORY Dir1
    ACCESS PARAMETERS
        (
com.oracle.bigdata.tablename=default.movieapp_log_avro
com.oracle.bigdata.datamode=java
        )
    )
;
```

Figure 2. New External Table Types link to Hive Metadata



```
-- Avro data from Hive
DBMS_REDACT.ADD_POLICY(
  object_schema => 'MOVIEDEMO',
  object_name => 'MOVIEAPP_LOG_AVRO',
  column_name => 'CUSTID',
  policy_name => 'mylogdata_redaction',
  function_type => DBMS_REDACT.PARTIAL,
  function_parameters => '9,1,7',
  expression => '1=1'
);
```

Figure 3. Redaction Policy on Avro Data in Hadoop



```
END;
/

-- Review the redacted data
SELECT * FROM movieapp_log_avro WHERE rownum < 20;

-- Join the redacted HDFS table with the redacted customer table
SELECT f.custid, c.last_name, f.movieid, m.title, f.time
FROM customer c, movielog_v f, movie m
WHERE c.cust_id = f.custid
  and f.movieid = m.movie_id;

-- Customer Recency, Frequency and Monetary query
WITH customer_sales AS (
  -- Sales and customer attributes
```

All Rows Fetched: 19 in 2.137 seconds

	CUSTID	MOVIEID	ACTIVITY	GENREID	RECOMMENDED	TIME	RATING	PRICE	POSITION
1	9999999	0	8	0	null	2012-07-01:00:00:07	(null)	(null)	(null)
2	9999999	1948	7	9	N	2012-07-01:00:00:22	(null)	(null)	(null)
3	9999999	0	9	0	null	2012-07-01:00:00:26	(null)	(null)	(null)
4	9999999	11547	7	6	Y	2012-07-01:00:00:32	(null)	(null)	(null)
5	9999999	11547	6	6	Y	2012-07-01:00:00:42	(null)	(null)	(null)
6	9999999	0	8	0	null	2012-07-01:00:00:43	(null)	(null)	(null)
7	9999999	0	9	0	null	2012-07-01:00:00:50	(null)	(null)	(null)
8	9999999	608	7	6	N	2012-07-01:00:01:03	(null)	(null)	(null)
9	9999999	0	9	0	null	2012-07-01:00:01:07	(null)	(null)	(null)
10	9999999	27205	7	9	Y	2012-07-01:00:01:18	(null)	(null)	(null)
11	9999999	1124	7	9	Y	2012-07-01:00:01:26	(null)	(null)	(null)
12	9999999	16309	7	46	N	2012-07-01:00:01:35	(null)	(null)	(null)
13	9999999	11547	7	6	Y	2012-07-01:00:01:39	(null)	(null)	(null)
14	9999999	0	9	0	null	2012-07-01:00:01:55	(null)	(null)	(null)

Figure 4. Enacting Transparent Redaction via SQL on Avro Data in Hadoop

Oracle Big Data SQL – Smart Scan for Hadoop

While big data may be massive, very often the amount of data that is relevant to a given query is smaller than the total data volume by an order of magnitude or more. This provides an opportunity for tremendous optimization in query performance. Smart Scan for Hadoop – based on Exadata Storage Servers Software – maximizes the performance of Oracle Big Data SQL by providing:

- Data-local scanning: data is read and processed at the point of storage
- Predicate evaluation and projection: only relevant data is transmitted from Hadoop
- Complex parsing: data such as JSON and XML are processed locally at the source
- Bloom Filters: Optimized joins through conversion to Bloom Filter on Hadoop

Conclusion

Oracle Big Data SQL is a new architecture for driving business value from ALL data in an organization. Big Data SQL delivers the full capability of Oracle SQL to data stored in Hadoop and NoSQL data stores while processing the data locally on the distributed nodes. Not only does Big Data SQL deliver all Oracle SQL to unstructured data, Big Data SQL also enables a single set of security policies across all data, regardless of their storage system. This enables access to larger data sets by more people in the organization without risking compliance issues or security breaches. Oracle Big Data SQL uniquely enables big data success in any organization.

Contact address:

Name
Jean-Pierre Dijcks
Oracle
500 Oracle Parkway
MS4op7
Redwood City, CA 94002, USA

Phone:	+1 650 607 5394
Email	jean-pierre.dijcks@oracle.com
Internet:	www.oracle.com/bigdata