# Oracle Big Data SQL – Technical Update

**Jean-Pierre Dijcks**
**Oracle**
**Redwood City, CA, USA**

**Keywords:**

Big Data, Hadoop, NoSQL Databases, Relational Databases, SQL, Security, Performance

**Introduction**

This technical session focuses on Oracle Big Data SQL. It is however one component in a large product stack that Oracle provides for Big Data. The paper therefore introduces a number of new components and go through some updates and then will focus on the SQL innovations Big Data SQL brings to Hadoop and NoSQL data stores.

**Introducing the updates to Oracle Big Data Management System**

Here at Oracle we take a holistic approach with respect to managing big data, so we include Hadoop, NoSQL and Relational Databases in our Big Data Management System, as well as deliver a complete Big Data Management in the Oracle Public Cloud.

**Big Data Cloud Service**

Big Data Cloud Service is one of the big-ticket items. With the general availability of Oracle Big Data Cloud Service now a fact, customers can leverage the full power of our Big Data Management System both on-premises and in the cloud. One of the key differentiators is the architecture where Oracle delivers the same platform you use in your data center in Oracle Public Cloud.

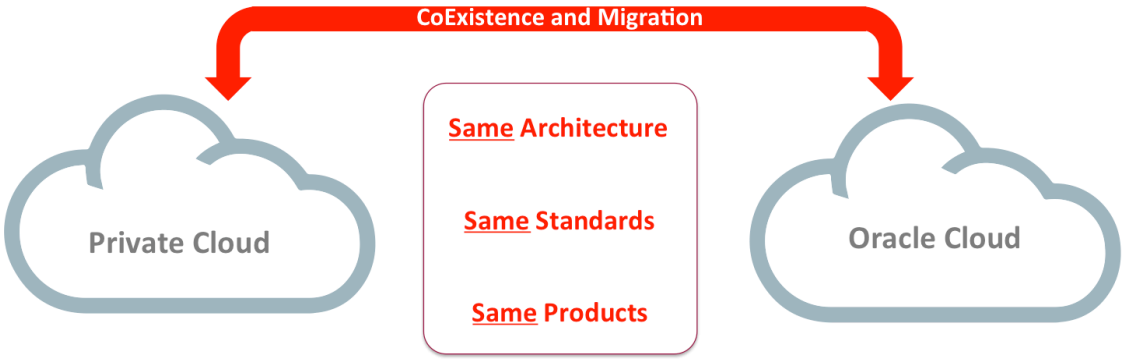


Figure 1. Oracle Big Data Cloud Service - Private and Public Cloud

With Oracle Big Data Cloud Service now generally available, a next set of highly anticipated services will soon be ready for GA, notably Oracle Big Data Discovery Cloud Service and Oracle Big Data Preparation Cloud Service.

**Big Data Appliance**

There was no update to the major version of the hardware: it is still X5-2. However, we did expand the storage capacity by moving from 4TB SAS drives to 8TB SAS drives, giving the BDA Full Rack roughly 1.7PB of raw disk. A short – updated – comparison with the previous generation now shows a massive spike on all components.

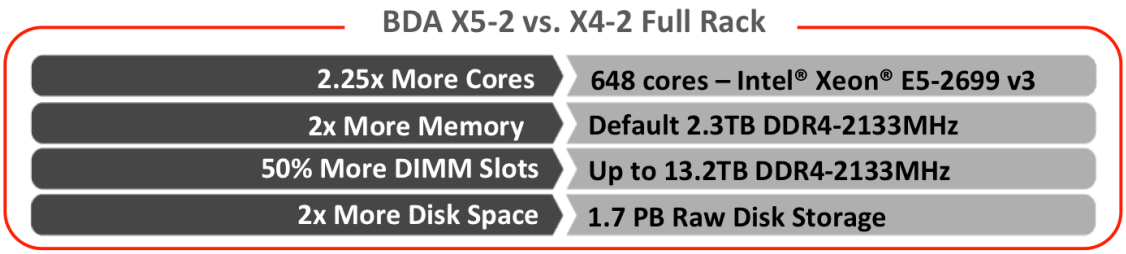


Figure 2. Comparing X4-2 with the latest X5-2 system

Although the disk capacity stands out, the most interesting change is the CPU. configuration. Big Data workloads are becoming much more mixed. These mixed workloads, MapReduce for ETL, SQL for ad-hoc analytics and Spark for some data processing, require more CPU power to adequately run on a system. By moving to the latest Intel Xeon processor BDA is better able to run these mixed workloads, while delivering a massive amount of storage.

On the software side of Big Data Appliance, it is now possible to leverage an Oracle feature to access Oracle resident data from BDA based queries. As an example, a Hive query can now do a join, or a lookup into an Oracle table without moving the data first into Hadoop. This dramatically simplifies building a Big Data Management System. Oracle Table Access for Hadoop is available on BDA with BDA 4.3. Within the same update of the software, an improved version of Copy to BDA is available, where a lot more automation is provided when moving data in batch from Oracle Database to Hadoop on BDA. A SQL Developer interface will also be provided to further simply the workflows for Copy to BDA.

**Big Data Spatial and Graph**

While Big Data Spatial and Graph is a very new product, it has grown a little in functionality and is now sporting a brand new component – multimedia analytics. This brand new component enables you to scan for example video or images and detect shapes. Faces are an example that comes to mind as well as out of the box with this functionality. Of course there are many other applications for shape detection in manufacturing or in working with other types of images. Multimedia analytics leverages the power of the distributed Hadoop compute cluster to scale to large data volumes and still analyze this complex data in reasonable time.

For those who have not heard about Big Data Spatial and Graph, it is well worth a look.

Big Data Spatial enables the enrichment of data with spatial elements using the full power of distributed computing on data in HDFS. By enriching the data it is far simpler to now visualize large quantities of data in maps and derive insights much faster. Big Data Spatial can also be used for proximity and containment analysis and vector and raster preparation of the data for further analysis.

Big Data Graph brings a new graph database to the ecosystem, which enables in-memory analytics on large graphs with the underlying data stored in HDFS. This graph database leverages Oracle NoSQL Database or HBase to serve up the large graph for analytics. While the obvious use case is social networks, it's also applicable in the Internet of Things world, where interactions between things are becoming as critical as interactions between humans. Various other use cases in cyber security and industrial engineering are also interesting.

**Big Data SQL Update**

The key goals of Big Data SQL are to expose data in its original format, and stored within Hadoop and NoSQL Databases through high performance Oracle SQL being offloaded to Storage resident cells or agents. The architecture of Big Data SQL closely follows the architecture of Oracle Exadata Storage Server Software and is built on the same proven technology.

**Retrieving Data**

With data in HDFS stored in an undetermined format (schema on read), SQL queries require some constructs to parse and interpret data for it to be processed in rows and columns. For this Big Data SQL leverages all the Hadoop constructs, notably InputFormat and SerDe Java classes optionally through Hive metadata definitions. Big Data SQL then layers the Oracle Big Data SQL Agent on top of this generic Hadoop infrastructure, as can be seen in Figure 3.

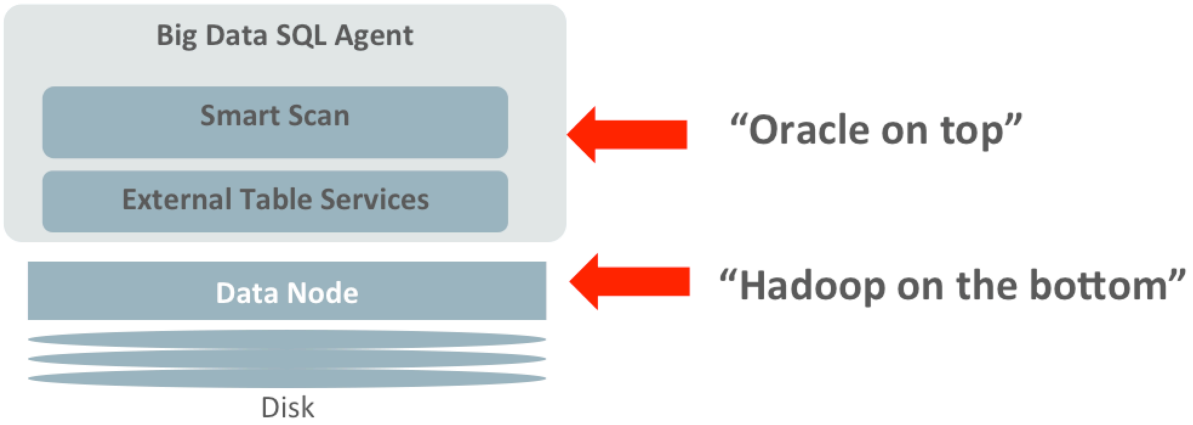


Figure 3. Architecture leveraging core Hadoop and Oracle together

Because Big Data SQL is based on Exadata Storage Server Software, a number of benefits are instantly available. Big Data SQL not only can retrieve data, but can also score Data Mining models at the individual agent, mapping model scoring to an individual HDFS node. Likewise querying JSON documents stored in HDFS can be done with SQL directly and is executed on the agent itself.

**Smart Scan**

Within the Big Data SQL Agent, similar functionality exists as is available in Exadata Storage Server Software. Smart Scans apply the filter and row projections from a given SQL query on the data streaming from the HDFS Data Nodes, reducing the data that is flowing to the Database to fulfill the data request of that given query. The benefits of Smart Scan for Hadoop data are even more pronounced than for Oracle Database as tables are often very wide and very large. Because of the elimination of data at the individual HDFS node, queries

across large tables are now possible within reasonable time limits enabling data warehouse style queries to be spread across data stored in both HDFS and Oracle Database.

**Storage Indexes**
Storage Indexes (SI) provide the same benefits of IO elimination to Big Data SQL as they provide to SQL on Exadata. The big difference is that in Big Data SQL the SI work on an HDFS block (on BDA – 256MB of data) and span 32 columns instead of the usual 8. SI is fully transparent to both Oracle Database and to the underlying HDFS environment. As with Exadata, the SI is a memory construct managed by the Big Data SQL software and invalidated automatically when the underlying files change.

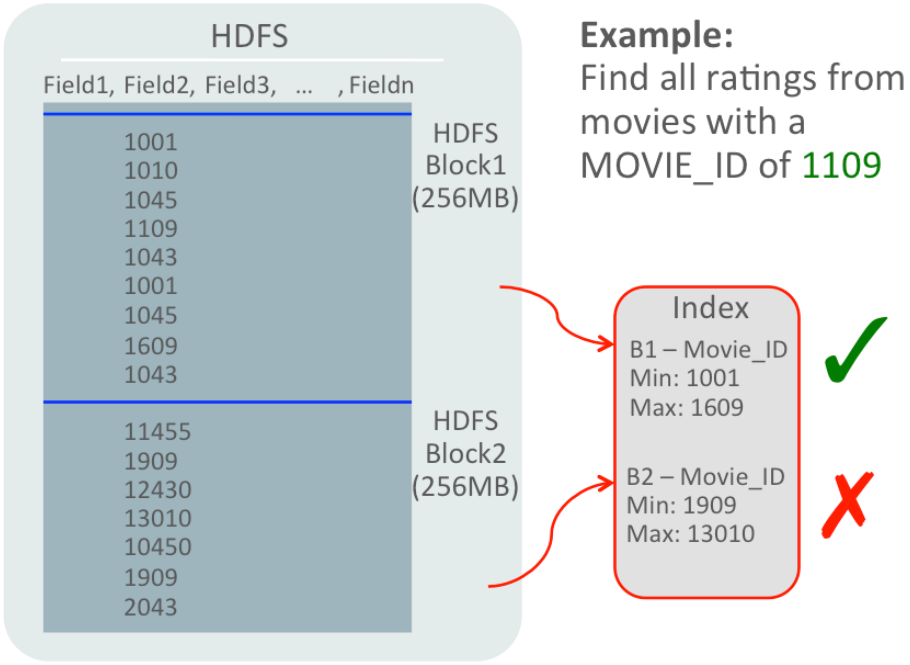


Figure 4. Storage Indexes work on HDFS Blocks and speed up IO by skipping blocks

SI works on data exposed via Oracle External tables using both the ORACLE_HIVE and ORACLE_HDFS types. Fields are mapped to these External Tables and the SI is attached to the Oracle columns, so that when a query references the column(s), the SI - when appropriate - kicks in. In the current version, SI does not support tables defined with Storage Handlers (ex: HBase or Oracle NoSQL Database).

**Compound Benefits**
Both Smart Scan and Storage Index features deliver compound benefits. Where Storage Indexes reduces the IO done, Smart Scan then enacts the same row filtering and column projection. This latter step remains important as it reduces the data transferred between systems.

**Virtual Machine to try out Big Data SQL and all other components**
Of course, having all these new features in the platform is a lot of fun. But how do I get to try all of this? It's simple. Oracle packages all of these features into a virtual machine called

Oracle Big Data Lite VM. This is updated for each new version of BDA software and picks up the components in the stack. This VM not only includes all of Oracle Data Integration – including the Big Data add-on – but it now also includes Oracle Big Data Discovery. It is the perfect client for a BDA, and the perfect test bed for any of your investigations. You can find the VM here: http://www.oracle.com/technetwork/database/bigdata-appliance/oracle-bigdatalite-2104726.html

**Contact address:**

**Jean-Pierre Dijcks**
Oracle
500 Oracle Parkway
MS 4op7
Redwood City, CA 94065

Phone: +1 650 607 5394
Email Jean-Pierre.Dijcks@oracle.com
Internet: www.oracle.com/bigdata