ORACLE
Business Process
Services

Big Data and Enterprise Data, Bridging Two Worlds with Oracle Data Integration

WHITE PAPER / JANUARY 25, 2019

ORACLE

Table of Contents

INTRODUCTION

Volume, Velocity, Variety - Big Data is all the vogue. Why? Enterprises know that there is a treasure trove of information they should be tapping into. This information is predominantly less structured data consisting of weblogs, social media, email, sensors, and photographs that can be mined for useful information. Whether it is healthcare, financial, manufacturing, government or retail, Big Data presents a big opportunity. A recent McKinsey study (see figure 1) illustrates that the US retail industry could realize a 60% increase in operating margins – just by fully exploiting Big Data. And this is just the beginning. Big Data turns traditional information architecture on its head, putting into question commonly accepted notions of where and how data should be aggregated processed, analyzed, and stored. This is where low cost options of Hadoop and NoSQL come in – new technologies which solve new problems for managing unstructured data. But one shouldn't forget what's already running in today's IT. Today's Business Analytics, Data Warehouses, Business Applications (ERP, CRM, SCM, HCM), and even many social, mobile, cloud applications still rely almost exclusively on structured data. This dilemma is what today's IT leaders are up against: what are the best ways to bridge enterprise data with Big Data through a common set of unified data integration tools? And what are the best strategies for dealing with the complexities of these two unique worlds?

Figure 1: Real World Use Cases for Big Data

	NEW DATA	WHAT IS POSSIBLE	WHY
Health Care Improve Quality And Efficiency	Practitioner's Notes; Machine Statistics	Best Practices; Reduced Hospitalization	Increase industry value by $300Bn per year
Retail One Size Fits All Marketing	Weblogs; Click Streams	Micro-Segmentation; Recommendations	Increase net margin by 60%
Banking Fraud Detection; Risk Analysis	Weblogs, Transaction , Systems, Fraud Reports	Semantic Discovery; Pattern Detection	Billions of dollars lost annually in bank fraud
Location Based Services Based On Home Zip Code	Personal Location Data	Geo- Advertising, Traffic, Local Search	Increase Revenue For Providers By $100B+
Utilities Resilient And Adaptable Grid	Smart Meter Reading, Call Center Data	Real Time And Predictive Utilization Analysis	Energy use expected to increase by 22% by 2030.

"The advantage of Big Data comes into play when you have the ability to correlate Big Data with your existing enterprise data. There's an implicit product requirement here in consolidating these various architecture principles into a single integration solution. The advantages of a single solution allow you to address not only the complexities of mapping, accessing, and loading Big Data but also combining it with your enterprise data – and this correlation requires integrating across mixed application environments. The correlation is key to taking full advantage of Big Data and requires a single unified tool that can straddle both environments."

Chai Pykmikulla
Senior Director, Product Management
Oracle Corporation

HARNESSING THE POWER OF BIG DATA – BEYOND THE SQL WORLD

Big data is having a tremendous impact on the data integration space. This has to do primarily with the fact that Big Data poses new questions for the best ways to process volumes and varieties of data at higher speeds and at faster velocities while keeping operating costs to a minimum. But before we look at the impact in more detail, let's first look at the current state of data integration. Data integration in a 'traditional sense' solves the issues of bulk data movement, replication, synchronization, virtualization, transformation, data quality, and data services.

These capabilities serve as a key technology component for moving data between Data Warehouses, Business Analytics, Master Data Management, Enterprise Applications, and Custom Applications. But now there's more to be moved. Much, much more. In fact, it's not just the scale, but it's the complexity of new technologies. Data Integration tools have evolved to support integration to and interoperation with technology offerings like Hadoop Distributed File System ("HDFS") using innovative techniques like MapReduce and NoSQL. Oozie, Sqoop and new access methodologies like Kafka, Spark, Spark Streaming and HiveQL have all merged the realms of Data Integration into the Big Data world. And in many cases to effectively leverage the power of this technology effectively it has to run natively in the technology. Much like E-LT runs natively on relational database systems, so too data integration technology needs to run its transformation, loading natively in Big Data environments

BIG DATA TECHNOLOGIES – WHAT YOU NEED TO KNOW

- Kafka – Apache Kafka® is an open-source platform for stream-processing of real-time data feeds.
- Spark – is an open-source cluster computing framework to implement iterative algorithms (e.g. training algorithms for machine learning) for repeated data queries.
- Spark Streaming – is an extension of Spark that enables scalable, high-throughput processing of live data streams. It lets you ingest data from Kafka (and others like: Flume, Twitter, ZeroMQ, etc.) and pushes data out to filesystems, databases and live dashboards.
- Hadoop - is an open source software project that supports data-intensive distributed applications. It enables applications to work with thousands of computational independent computers and petabytes of data.
- MapReduce - is a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers
- Hive – is a query language similar to standard SQL statements. The query language is called Hive Query Language (HQL). Hive is executed on Hadoop clusters based on MapReduce operations.
- NoSQL – is a term used for a wide variety of database management systems whose one thing in common is that they don't use SQL. Oracle NoSQL Database, for example, is a key value database, where data is stored under and accessed by a unique key, and variable schemas are easily supported.
- Hadoop Distributed File System (HDFS) - is a distributed, scalable, and portable file system written in Java for the Hadoop framework. In order to scale a Hadoop cluster to hundreds and thousands of nodes, HDFS is critical.

- MongoDB – is a NoSQL open source Database designed for scale, flexible data aggregation and to store files of any size. It has rich querying, high availability and full indexing support and is fast being adopted by many businesses.
- HBase – is an Apache open source, distributed, versioned project that provides real - time read/write access to Big Data atop clusters of commodity hardware. Apache HBase provides these capabilities on top of Hadoop and HDFS.

CHALLENGES POSED BY BIG DATA

- A shift from familiar Data Integration tools to new technologies to deal with Big Data problems will impact productivity of the organization and increase costs to invest in new skills and tools.
- Performance constraints of the Data Integration tools to deal with larger and more complex data within ever shortening time windows and real time analytics.
- Costly network traffic in moving data within and outside the business and storage areas, and lastly
- Redundancy of existing costly IT investment which become obsolete if they are not "Big Data Enabled."

Oracle's approach to Big Data solves each of these concerns. Oracle Data Integrator 12c (ODI12c) addresses productivity and skill requirements without leaving familiar programming and IT environments. Oracle Data Integrator for Big Data helps bring the data together while Oracle's Integrated Platform delivers optimal Big Data flow ensuring not just data size and complexity, but also data speed and delivers Fast Data.

BEST TECHNOLOGY APPROACHES FOR INTEGRATING BIG DATA

Integrated Design Tools: Improve Productivity

The advantage of Big Data comes into play when you have the ability to correlate Big Data with your existing enterprise data. There's an implicit product requirement here in consolidating these various architecture principles into a single integration solution. The advantages of a single solution allow you to address not only the complexities of mapping, accessing, and loading Big Data but also combining it with your enterprise data – and this correlation requires integrating across mixed application environments. The correlation is key to taking full advantage of Big Data and requires a single unified tool that can straddle both environments.

In addition, Big Data sources consist of many different types and in many different forms. How can anyone be sure of the quality of that data? In the Big Data scenario, Data Quality is important because of the multitude of data sources. Multiple data sources make it difficult to trust the underlying data. Being able to quickly and easily identify and resolve any data discrepancies, missing values, etc in an automated fashion is beneficial to the applications and systems that use this information. Oracle Enterprise Data Quality integrated with Oracle Data Integrator 12c (ODI12c) enables data analysts to investigate, identify and resolve data quality issues by discovering and analyzing anomalies.

Oracle Data Integrator 12c (ODI12c): Loading and Transforming Big Data

Oracle Data Integrator 12c (ODI12c) leverages unified tooling for both Big Data and enterprise data which translates into a faster learning curve as well as seamless usability so that the data scientist or data analyst can focus on integration versus usability. Flow based declarative designs in Oracle Data Integrator 12c (ODI12c) helps build complex expressions that are easy to maintain and support. Two representations of the same ELT mappings, the logical representation and the physical representation, provide customized working environments for business analysts and data scientists.

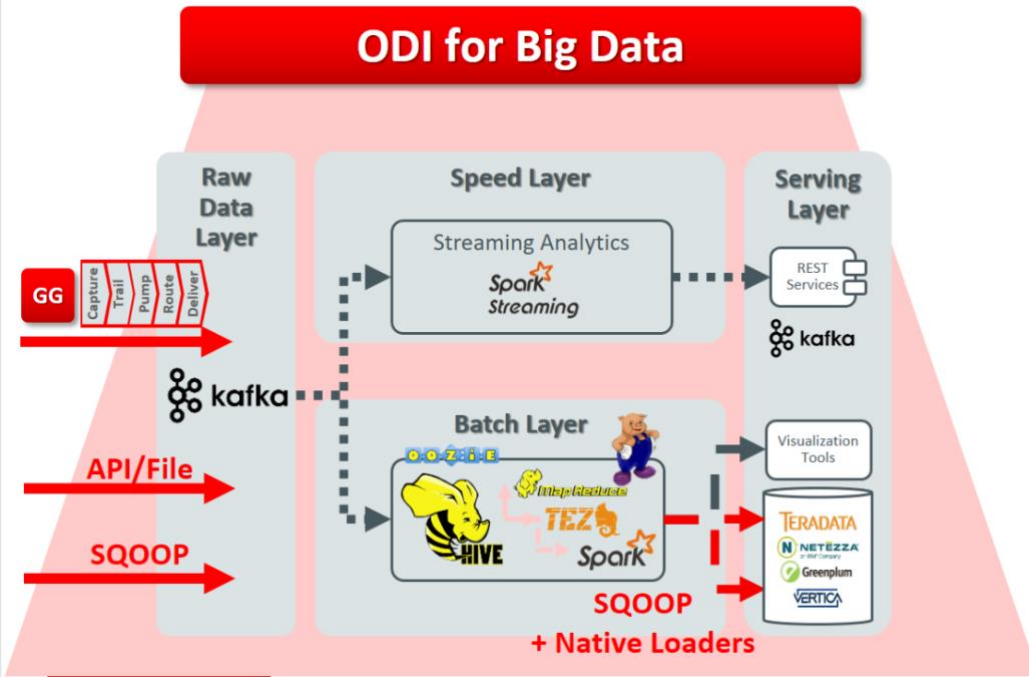

Figure 2: Loading and Transforming Big Data using Oracle Data Integrator 12c

Oracle Data Integrator Application Adapters for Hadoop

As a component of Oracle Big Data Connectors, Oracle Data Integrator Application Adapter for Hadoop provides native Hadoop integration together with Oracle Data Integrator 12c (ODI12c). Knowledge Modules can then be used to build Hadoop metadata within Oracle Data Integrator 12c (ODI12c), load data into Hadoop, transform data within Hadoop, and load data easily and directly into Oracle Database utilizing Oracle Loader for Hadoop. Once the data is processed and organized on the Hadoop cluster, Oracle Data Integrator 12c (ODI12c) loads the data directly into Oracle Database utilizing the Oracle Loader for Hadoop.

Oracle Data Integrator Application Adapter for Hadoop simplifies data loading and movement between Hadoop and an Oracle Database through Oracle Data Integrator 12c (ODI12c)'s easy to use prebuilt interfaces. Oracle Data Integrator Application Adapter for Hadoop simplifies data integration from Hadoop and an Oracle Database through Oracle Data Integrator 12c (ODI12c)'s easy to use interface. By providing efficient connectivity between Oracle Database and Hadoop, Oracle Data Integrator for Big Data enables analysis of all data, both structured and unstructured, in enterprise data warehouses. Additionally, enterprises that are already using a Hadoop solution integrate data from HDFS using Oracle Data Integrator for Big Data as a standalone software solution as well.

Integrated Platform: Simplify and Optimize

Taking all the miscellaneous technologies around Big Data – which are new to many organizations - and making them each work with one another is challenging. Making them work together in a production- grade environment is even more daunting. Integrated systems can help an organization radically simplify their Big Data architectures by integrating the necessary hardware and software components to provide fast and cost-efficient access, and mapping, to NoSQL and HDFS.

Combined hardware and software systems can be optimized for redundancy with mirrored disks, optimized for high availability with hot-swappable power, and optimized for scale by adding new racks with more memory and processing power. Take it one step further and you can use these same systems to build out more elastic capacity to meet the flexibility requirements Big Data demands.

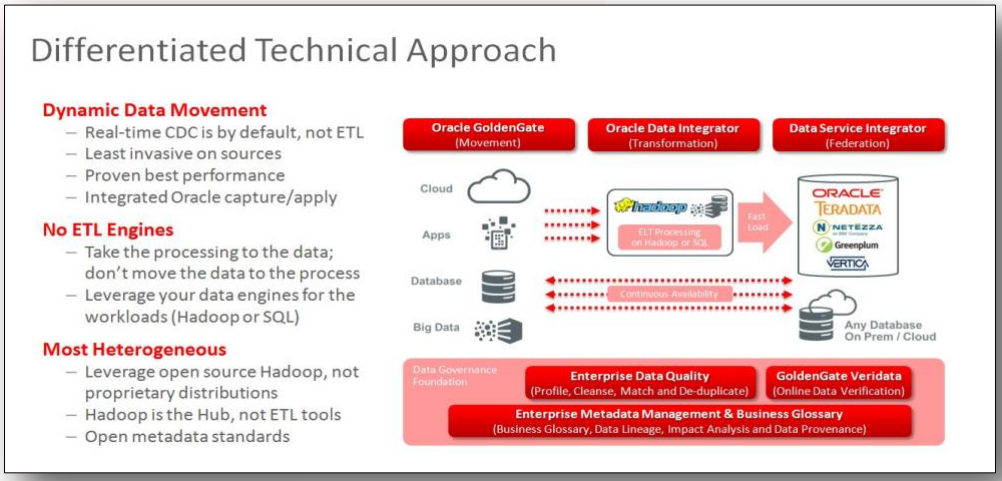

Figure 3: Oracle's Big Data Story

Oracle is uniquely qualified to combine everything needed to meet the Big Data challenge –including software and hardware – into one engineered system. The Oracle Big Data Appliance is an engineered system that combines optimized hardware with the most comprehensive software stack featuring both the Cloudera Distribution including Apache Hadoop and specialized solutions developed by Oracle to deliver a complete, easy-to-deploy solution for acquiring, organizing and loading Big Data into Oracle Database. It is designed to deliver extreme analytics on all data types, with enterprise-class performance, availability, supportability and security. With Oracle Big Data Connectors and Oracle Data Integrator 12c (ODI12c), the solution is tightly integrated with Oracle Exadata and Oracle Database, so that data can be analyzed data with extreme performance.

Real-time Business Analytics: Extends the Value of Big Data

One of the key expectations for Big Data is to yield real-time analytics that improve business insights. But how timely is the data is a question that will still need to be answered for both Big Data or traditional enterprise data.

While Big Data on its own has no means of applying 'traditional' change data capture [since there are no log files in NoSQL or HDFS], it's still an important requirement to implement real-time solutions in conjunction with big data. Otherwise the speed advantage to indexing realms of Big Data will be undone by sluggish ETL processing that it's dependent on. Big Data can be processed at high volume with high velocity. In fact, this is the entire strategy behind the invention of MapReduce – providing

search responses of the highest quality in close to the blink of a human eye. Combine this power with real-time solutions in replication, change data capture, synchronization, and the integration to business analytics tooling, and you have what amounts to the compelling advantages of real-time business analytics.

The combination of Oracle Data Integrator and Oracle GoldenGate pack a powerful punch when it comes to achieving true real-time business analytics. Oracle GoldenGate is an integral part of the real-time business analytics use case to accomplish real- time data replication and capture ensuring applications have the data they need immediately.

CONCLUSION

Big data continues to be the center of attention - take away the hype and there's a key takeaway. For years, companies have been running their critical business infrastructure and building business insights based on transactional data stored in relational databases. Beyond that critical data, however, is a potential treasure trove of less structured data: web and call logs, social media, email, sensors, and photographs [and more] that can be mined for useful information. Companies that are seeking ways to capitalize on the hidden potential of Big Data need to consider data integration technologies to help bridge the gap and correlate that data across the enterprise.

Bridging the two worlds of Big Data and enterprise data means considering solutions that are complete, based on traditional as well as emerging Hadoop technologies, and are poised for success through integrated design tools, integrated platforms, and real-time analytics. Leveraging these best practices ensures improved productivity, lowered TCO, IT optimization, and better business insights. Only Oracle provides the most complete, integrated, and real-time solution data integration that helps bridge the two worlds of Big Data and enterprise data to accelerate adoption and yield greater returns for these emerging technologies.

For more information on Oracle Data Integrator go to our website:
www.oracle.com/goto/dataintegration.

ORACLE CORPORATION

Worldwide Headquarters
500 Oracle Parkway, Redwood Shores, CA 94065 USA

Worldwide Inquiries
TELE + 1.650.506.7000 + 1.800.ORACLE1
FAX + 1.650.506.7200
oracle.com

CONNECT WITH US

Call +1.800.ORACLE1 or visit oracle.com. Outside North America, find your local office at oracle.com/contact.

 blogs.oracle.com/oracle

 facebook.com/oracle

 twitter.com/oracle

Integrated Cloud Applications & Platform Services

White Paper Title
January 2017
Author: [OPTIONAL]
Contributing Authors: [OPTIONAL]

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

ORACLE