ORACLE
FUSION MIDDLEWARE
DATA INTEGRATOR

The Five Most Common Big Data Integration Mistakes To Avoid

ORACLE WHITE PAPER | APRIL 2015

ORACLE

Executive Summary

Big Data projects have fascinated business executives with the promise of higher business returns and greater customer understanding. Many high profile successes have also grabbed the headlines, hastening big data investments. Those who adapted their business and IT models to take advantage of this data deluge have derived envious returns through revenue growth, increased fact based decisions and innovative customer engagement campaigns. However, far less talked about is the large number of big data projects that have hit a plateau within businesses that have not been able to deliver the promised pot of gold. Business and technology reasons combine to bring about these sub-optimal results.

The success of any big data project fundamentally depends on an enterprise's ability to capture, store and govern its data. The better an enterprise can provide fast, trustworthy and secure data to business decision maker's the higher the chances of success in exploiting big data, obtaining planned return on investments and justifying further investments. In this paper, we focus on big data integration and take a look at the top five most common mistakes enterprises make when approaching big data integration initiatives and how to avoid them.

The Five Vital Big Data Integration Mistakes

1. *Not choosing an enterprise grade Hadoop foundation and data integration technology*
2. *Retaining outdated data warehousing models instead of focusing on modern Big Data architecture patterns*
3. *Not prioritizing efficient data integration principles*
4. *Underestimating the importance governance, and finally*
5. *Ignoring the data processing power of Hadoop/NoSQL when handling complex workloads.*

Understanding and fixing these five principles will kick start your big data integration project.

---

1. Invest In Enterprise Grade Hadoop And Data Integration Technology

There are a number of Apache projects that are exciting and aim to fill niche data processing requirements. However when selecting Hadoop for the enterprise, it has to stand up to rigorous stresses of security, administration and monitoring. In addition there should also be continued dedicated development and support resources that enhance the chosen technology.

Similarly, the data integration technology that is chosen should lend itself to scale with the Hadoop technology. User adoption and developer productivity should be high if the enterprise needs to make this a strategic initiative. The data integration technology should deliver tools to streamline development, enforce quality and shorten time to

Key To Big Data Success

ETL vs ELT

To Extract, Load and Transform (ELT) data using the big data platform's capabilities is more efficient than to Extract, Transform and Load (ETL) data because it results in

- ✓ *Minimal middleware*
- ✓ *Fast performance with set based processing, and*
- ✓ *Reduced n/w traffic*

implement. This reduces custom coding from proliferating, a common precursor to tough maintenance and data transparency.

As with any enterprise technology standard big data integration tools should be able to work with a variety of heterogeneous big data languages and sources while abstracting the user base from the complexity of implementation. . The data integration technology should allow enterprises to work with newer Hadoop standards as and when they mature and switch between multiple underlying big data standards without business risks and disruption.

2. Focus On Modern Big Data Architecture Patterns

Many enterprises approach big data architecture as an extension of their existing data warehouses. Big data architectures (including data reservoirs, data lakes etc..) frequently co exist with traditional data warehouses, but to build one along similar principles of economic data storage will restrict the value of data within the big data store.

Specialist storage and engineered platforms build for performance complement the big data reservoirs which are mainly used for data exploration. Properly engineered, big data reservoirs can hand off subsets of often used data to engineered platforms to improve speed and performance.

Modern big data architectures emphasize data streaming for real time data ingestion into the big data platform, data enrichment and transformation using native big data query languages (some examples are like Pig Latin, HiveQL, MapReduce etc...) and are fully orchestrated and governed to minimize risk.

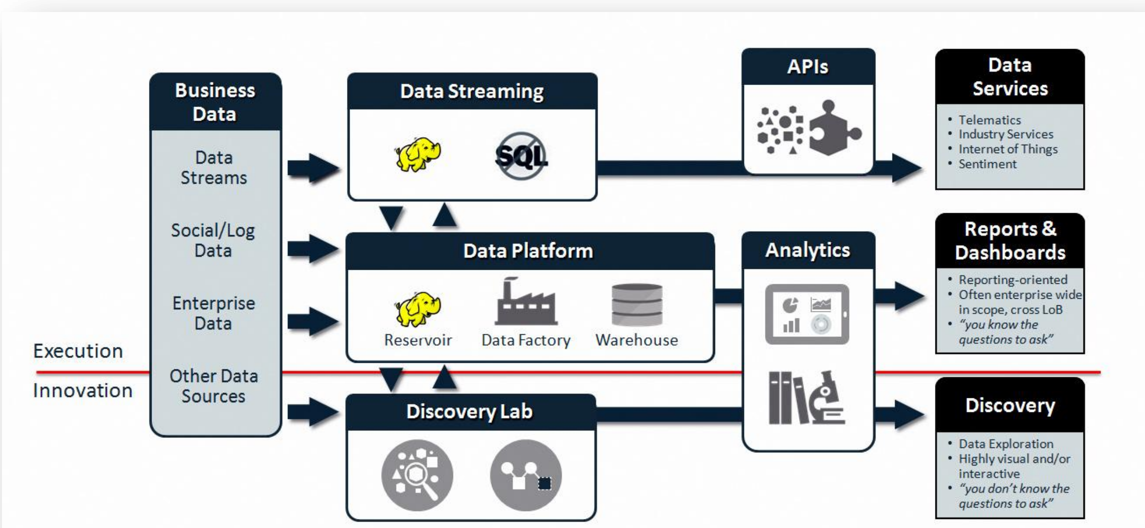

FIGURE 1: 4TH GENERATION DATA ARCHITECTURE FOR BIG DATA IS EMPHASIZES INTEGRATION WITH CRUCIAL DATA REQUIREMENTS INCLUDING DATA STREAMING, DISCOVERY, ENRICHMENT, TRANSFORMATION AND GOVERNANCE.

3. Prioritize Efficient Data Ingestion And Data Transformation

Choosing the right data integration technology depends on key criteria.

It starts with the ability to ingest real time data into the data reservoir. This ensures that the data used for decision making is up-to-date and business analytics reflects the latest data. Sub second latency differentiates average user experience from excellent customer experience providing timely insights.

These data ingestion tools should be non-invasive when capturing data and not impact source technology performance. Once within the data reservoir, the transformation technology should be transparent and not inject proprietary code onto the Hadoop nodes. It should provide facilities for modular, team based development. It should be portable across platforms, or in other words, abide by the “design once, run anywhere” mantra.

Some of these criteria are satisfied by traditional data management technologies. However to see success in big data projects all these criteria are necessary specifications for the selected tool.

4. Incorporate Pervasive Data Governance

Big data reservoirs are generally considered the blackbox playpen of data scientists. While this was true in the first wave of Hadoop projects this is no longer the case. In fact, proper emphasis should be laid in ensuring transparency in Hadoop clusters. If the upsides of storing full raw data in the data reservoirs are profits and customer experience, the downside of a data leak to large amounts of data is costly litigation and irreparable reputation damage.

Managing metadata across every technology in the data management landscape is key to govern data. It allows complete data provenance, allowing business and IT accountability for the data that passes through the systems and business decisions.

Good governance depends as much on technology as on the organization’s culture and business processes. However selecting the right governance technology is critical in enabling the business govern its data. A good governance technology brings data transparency, accountability and helps identify areas of process and performance improvements.

In the integrated big data platform, it is important that the governance tool cuts across multiple technologies (data bases, data warehouses, data quality and enrichment technologies, data integration technologies, business intelligence and analytics technologies) to efficiently fulfill governance requirements. The governance technology should service both the business user and the technology users.

5. Get The Most Out Of Your Hadoop Cluster

If you think of Hadoop and NoSQL as just commodity data stores you miss the big advantage they provide through their compute capabilities. The gains that you achieve through data storage are lost if you do not efficiently utilize the big data platform for processing. To do this, you should offload compute intense queries into the big data store by generating code that is native to the underlying big data standard. This allows you to use your big data investment both for storage and processing, and as your data volume and storage scales up, you do not have to invest in additional processing hardware.

Tech Clues

A data integration tool isn’t native to Hadoop if it

- ✓ *Runs outside the cluster,*
- ✓ *Requires installing proprietary software on the cluster,*
- ✓ *Requires proprietary monitoring technology, and*
- ✓ *Does not let you “design once, deploy anywhere natively”.*

To do this, the data integration technology should not use middleware, or a processing platform that is proprietary and exists outside of target/source data bases. Traditional Extract Transform and Load (ETL) technologies that are designed for relational databases normally have middleware based architecture which counteracts and nulls any big data advantages. Modern tools that have an ELT (Extract, Load and then Transform on target/source) based technology are best suited for big data integration.

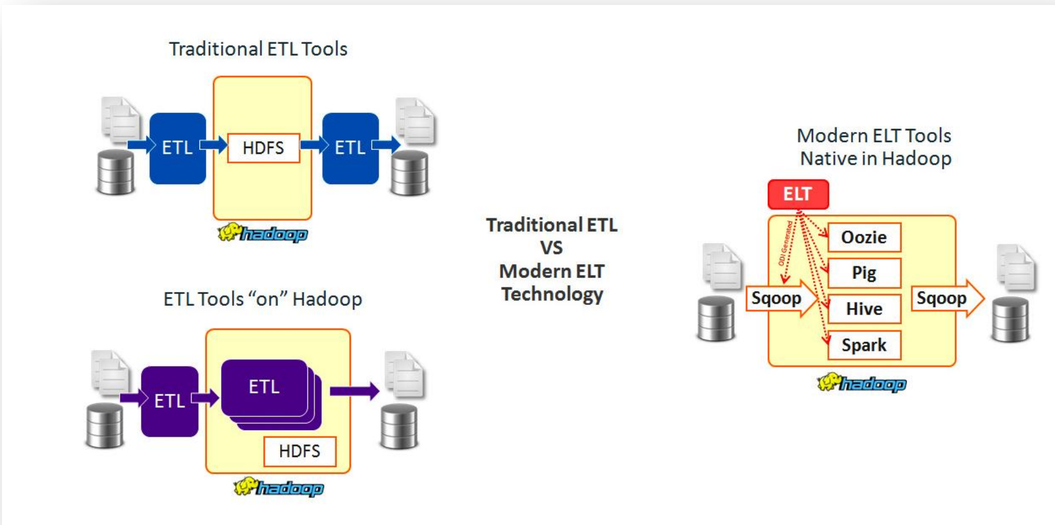

Figure 2: Big Data Integration technologies should be able to offload workload into the Hadoop database through native code generation.

Future Proof Big Data Integration

Oracle Data Integrator

- *Native Hadoop*
- *Design Once Deploy Anywhere*
- *Conceived around ELT architecture*

Oracle GoldenGate

- *Real time ingest*
- *Hadoop based*

Oracle Enterprise Metadata Management

- *Big Data Governance*
- *Complete Data Provenance*

The Oracle Advantage

Oracle Data Integration solutions provide a future proof, modern portfolio of solutions that helps customers succeed in their big data projects.

Oracle Data Integrator and Oracle Data Integrator Enterprise Edition Advanced Big Data Option provide an ELT based, heterogeneous enterprise grade data integration tool. It enables streamlined development and faster time to value by providing out of the box support for big data functionalities for a variety of big data standards including MapReduce, Pig Latin, HiveQL and Spark.

Heterogeneous and Optimized

- *Certified with all leading technologies, Oracle and non-Oracle*
- *Supports leading Big Data query languages and platforms*
- *Optimized to deliver the best performance on Oracle's Engineered Exadata platform*

Oracle GoldenGate for Big Data is a Hadoop-based technology that allows customers to stream real-time data from heterogeneous transactional systems into big data systems, including targets such as Apache Hadoop, Apache Hive, Apache HBase and Apache Flume. It allows customers to enhance big data analytics initiatives by incorporating existing real-time architectures into big data solutions, while ensuring their big data reservoirs are up to date with production systems.

Oracle Enterprise Metadata Management helps govern data integration technologies through metadata management. It provides a business user friendly, search driven interface that provides data performance, accountability and transparency across traditional and big data systems.

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

Oracle Corporation, World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065, USA

Worldwide Inquiries
Phone: +1.650.506.7000
Fax: +1.650.506.7200

CONNECT WITH US

- blogs.oracle.com/oracle
- facebook.com/oracle
- twitter.com/oracle
- oracle.com

Hardware and Software, Engineered to Work Together

The Five Most Common Big Data Integration Mistakes to Avoid
April 2015
Author: Oracle