Building a Robust Big Data QA Ecosystem to Mitigate Data Integrity Challenges

With big data evolving rapidly, organizations must seek solutions to ensure robust processes for quality assurance around big data implementations.

Executive Summary

Harvesting relevant information from big data is an imperative for enterprises seeking to optimize strategic business decision-making. Opportunities that were traditionally unavailable are now a reality, with new and more revealing insights extracted from sources such as social media and devices that constitute the Internet of Things. Consequently, emerging technologies are enabling organizations to gain valuable business insights from data that is growing exponentially in volume, velocity, variation of data formats and complexity.

Leading industry analysts forecast the big data market to reach U.S.$25 billion by 2015. As a consequence, organizations will require newer data integration platforms, fueling demand for QA processes that service new platforms, leading to the necessity of big data testing.

For big data testing strategy to be effective, the “4Vs” of big data – volume (scale of data), variety (different forms of data), velocity (analysis of streaming data in microseconds) and veracity (certainty of data) – must be continuously monitored and validated. In addition to the large volumes, the heterogeneous and unstructured nature of big data increases the complexity of validation, rendering sampling-based traditional QA strategy infeasible. Setting up a QA infrastructure to manage these volumes itself is a challenge. The absence of robust test data management strategies and a lack of performance testing tools within many IT organizations make big data testing one of the most perplexing technical propositions that business encounters.

Meeting the big data testing challenge requires utilities and automation solutions to improve test coverage, particularly when sampling-based traditional QA strategies are inadequate. This white paper outlines our proposed big data testing framework, with a focus on identifying the key processes in data warehouse testing, performance testing and test data management.
Challenges in Big Data

Since the mid-1990s, organizations have become accustomed to handling data contained in relational databases and spreadsheets; this data is structured. However, with the advent of big data, information can reside in semi-structured or unstructured formats, which are cumbersome to interpret and manage as the data resides in database rows and columns. With the phenomenal explosion in the IT intensity of most businesses, data volumes and velocities have accelerated, creating a need for real-time big data testing. This has heightened concerns over how to assure quality across the big data ecosystem.

At present, testers process clean and structured data. However, they also need to handle semi-structured and unstructured data. Key issues that require relatively more attention in big data testing include:

- Data security.
- Performance issues and the workload on the system due to heightened data volumes.
- Scalability of the data storage media.

Data warehouse testing, performance testing, and test data management are the fundamental components of big data testing. Addressing these challenges is tantamount to verifying the entire big data testing continuum (see Figure 1).

Streamlining Processes to Overcome Challenges in Big Data Testing

Given the ever-evolving technology landscape, today’s necessity becomes obsolete tomorrow. As a result, it is important to establish streamlined processes that will stay the course despite changing technologies and evolving platforms. Software testing follows the same evolutionary cycle.

At a Glance: Challenges in Big Data Testing

<table>
<thead>
<tr>
<th>Test Data Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management of test data during automated testing process.</td>
</tr>
<tr>
<td>Anticipating the acquisition and management of test data during different phases of the software testing lifecycle.</td>
</tr>
<tr>
<td>Test data setup in relation to test coverage, accuracy and types of big test data.</td>
</tr>
<tr>
<td>Investment in servers utilized for performance testing and small-scale companies may not be cost-effective.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generation of greater workload for performance testing of big data.</td>
</tr>
<tr>
<td>Test results in the form of reports, charts and graphs are at least twice as big in comparison with traditional BI reports.</td>
</tr>
<tr>
<td>Interpretation of results and identifying bottlenecks.</td>
</tr>
<tr>
<td>Performance tuning.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Warehouse Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decreased test coverage due to complex organization of big data requirement.</td>
</tr>
<tr>
<td>Test data supports limited normalization.</td>
</tr>
<tr>
<td>4 Vs – variety, velocity, volume and veracity – are not monitored.</td>
</tr>
</tbody>
</table>
To address the dynamic changes in the big data ecosystem, organizations must streamline their processes for data warehouse testing, performance testing and test data management.

**Strengthen Data Warehousing Processes**

While data warehouse testing is performed in a controlled environment, the unpredictable nature of the big data testing environment presents a unique set of challenges. Data warehouse and business intelligence testing require highly complex testing strategies, processes and tools pertaining specifically to the 4Vs of big data. Recommendations to refine test strategies and processes include:

- **Make “big” things simple through a “divide-and-test” strategy.** Organize your big data warehouse into smaller units that are easily testable, thus improving the test coverage and optimizing the big data test set.

- **Normalize design and tests.** Achieve a more effective generation of normalized test data for big data testing by normalizing the dynamic schemas at the design level.

- **Enhance testing through measuring the 4Vs.** Data warehouse test environments that are specifically designed to handle the 4Vs of big data will result in improved test coverage.

**Strengthen Performance**

Performance testing is an integral part of system testing that focuses on volumes, workloads, real-time scenarios and end users' navigational habits. The performance of a system depends on variable factors such as network, underlying hardware, Web servers, database servers, hosting servers, number of peak loads and prolonged workloads. However, addressing these requirements — and maintaining big data test systems performance — requires the organization’s full attention. Recommendations to implement the big data framework in performance testing include:

- **Simulate a real-time environment with distributed and parallel workload distribution.** Testing should be carried out in parallel in a distributed environment. The scripts generated by performance testing tools should be distributed among the controllers to simulate a real-time environment.

- **Integration with distributed test data:** Performance testing strategies depend predominately on the scenario set of the controllers. The spreadsheets and the back-end databases that typically hold test data often lack the ability to hold unstructured big data. To overcome this obstacle, the controller should be provided with an interface that can be used to integrate with the existing distributed test data.

- **Parallel test execution:** Enabling distributed virtual users to execute tests in parallel is an effective way to handle test execution.

**Strengthen Test Data Quality**

Recommendations for addressing the pain points in test data management of big data include:

- **Planning and designing:** Automated scripts cannot be scaled to test big data. Scaling up test data sets without adequate planning and design will lead to delayed response time and possibly timed-out test execution. Performing action-based testing (ABT) will help mitigate this issue. In ABT, tests are treated as actions in a test module. These actions are pointed toward keywords along with the parameters required for executing the tests.

- **Infrastructure setup:** Test automation consumes enormous resources to generate workloads. However, investing in dedicated servers is not cost-efficient for the small-scale operations that process big data. Renting infrastructure as infrastructure as a service delivered via the cloud can help mitigate costs. Alternatively, the generation of higher workloads for performance testing of big data can be effectively handled with virtual parallelism on numerous virtual machines.

**Big Data Testing Is No Longer a Distant Chimera**

Big data testing strategy is pivotal for the success of big data initiatives. As a logical extension, testing and QA teams will not be exempted from handling big data. Yet, big data testing remains in a nascent stage and lacks a defined manual testing framework to transition to automated testing. Moreover, QA processes, customized frameworks and tools used in various specialized testing services will require a significant upgrade to effectively and efficiently handle big data.

What was once called garbage data is now known as big data. Nothing is wasted, deleted or removed.
The risks and challenges illustrated in this paper are just the tip of the iceberg. Big data is getting bigger by the day. With every passing moment there are bigger challenges of scalability and increased usage of cloud resources to ramp up testing of big data ecosystems. Unprecedented risks and issues may still emerge as the testing community starts working with big data. Leveraging the expertise of big data testing consultants can ease the pain and accelerate the learning curve in three important ways:

• Designing an end-to-end QA strategy to contend with the 4Vs.
• Gaining advice on the use of tried-and-true and appropriate tools.
• Mitigating anticipated (and unanticipated) risks and related issues.

What was once called garbage data is now known as big data. Nothing is wasted, deleted or removed. All data sets, structured, semi-structured and unstructured, are of paramount importance for businesses interested in making informed and timely decisions on strategies that drive business success today and tomorrow.

Footnote

About the Author
Sushmitha Geddam is a Project Manager at Cognizant’s Data Testing Center of Excellence (CoE) R&D team and leads the big data initiatives. She has over 10 years of experience handling projects in specialized testing areas such as database, DW/ETL and data migration across numerous industry sectors, including investment banking, healthcare, insurance and telecom. Sushmitha can be reached at Sushmitha.Geddam@cognizant.com.

About Cognizant QE&A Data Testing Center of Excellence
Cognizant QE&A Data Testing CoE breeds expertise around industry-leading technologies. Our experts help you to accelerate and optimize the QA of your data testing needs by using state-of-the-art products, solutions and frameworks. For more details, reach us at QEAinBigData@cognizant.com.

About Cognizant
Cognizant (NASDAQ: CTSH) is a leading provider of information technology, consulting, and business process outsourcing services, dedicated to helping the world’s leading companies build stronger businesses. Headquartered in Teaneck, New Jersey (U.S.), Cognizant combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. With over 75 development and delivery centers worldwide and approximately 187,400 employees as of June 30, 2014, Cognizant is a member of the NASDAQ-100, the S&P 500, the Forbes Global 2000, and the Fortune 500 and is ranked among the top performing and fastest growing companies in the world. Visit us online at www.cognizant.com or follow us on Twitter: Cognizant.