



## The Problem

With the increasing use of high speed, large volume of complex data in a variety of formats flowing via old and new platforms (NoSQL, Hadoop, Cloud) has increased the prevalence of poor Data Quality. Poor quality of Big Data results in compliance failures, manual rework cost to fix errors, inaccurate insights, failed initiatives and lost opportunity. Data Validation has become more important than ever in the age of Big Data. Companies that need to move large quantities of data on a daily basis or during M&A or Divestiture will find this very useful in validating their Big Data Quality in a fraction of the time previously needed.

## The Bigger Problem

The best practice when moving critical data to ensure its integrity is to validate it. Data Validation is complex and costly. Traditional tools are not focused on Big-Data Validation solutions due to their older, incompatible architecture.

## Our Solution

DataBuck™ is an **Ultra-Fast, Cross- Platform, Big-Data Validation** solution. Leveraging Apache Spark and proprietary algorithms, DataBuck makes it easy to validate massive volumes of data as it moves across different platforms. It enables organizations to profile Big Data, check for quality issues, identify anomalies, and match two or more data sources. It can also keep track how a single data source changes unexpectedly over time.

## Functional Overview

- **Data Profiling:** Profiles data and compares it with prior runs and also across different applications for batch and streaming data processes, to determine unexpected changes in different instances of the same data and over time
- **Big Data Quality:** Performs standardized and custom rules-driven Big Data Quality tests
- **Anomaly Detection:** Identifies data outlier using standard and custom algorithms
- **Matching:** Performs Data Completeness test across multiple Big Data sources

## Why Data Buck?

- Simple, Fast, and powerful DataBuck offers the only end-to-end, cross-platform, Big Data Validation capabilities that drastically reduce the risk of data errors in the downstream processes
- **Big Data Input Sources:** For 20+ Big Data sources commonly found in Banks, Financial Services and Insurance organizations, wizard-led connection eliminates the need for coding to validate Big Data as it flows via multiple platforms
- **Built in Algorithm:** Over 100 pre-built data validation functions from Null Checks to Matching are available with DataBuck
- **Code-Free:** Wizard based user interface eliminates the need for sophisticated technology resources. Validation checks can be set up within minutes
- **Ultra-fast performance:** Multi node in-memory parallel computing coupled with proprietary data validation algorithm resulting in superior performance.

## Summary

DataBuck™ is a SaaS or On-Premise software that validates Big Data Quality, matches and detects Anomalies in Big Data. Leveraging Spark and proprietary algorithms, DataBuck performs Cross-Platform-Validation (Hadoop/ SQL/no-SQL/Cloud) and is 10x faster than any traditional approach. Web-based, code-free, intuitive interface enable setting up Data Quality check in under 15 mins

For more information:

 +1-385-393 4436

[Contact@FirstEigen.com](mailto:Contact@FirstEigen.com)

<http://firsteigen.com/databuck/>

# DataBuck

<p><b>Server System Requirements</b></p> <p><b>Recommended</b></p> <ul style="list-style-type: none"> <li>• 12-24 1-4TB hard disks in a JBOD (Just a Bunch Of Disks) configuration</li> <li>• 2 quad-/hex-/octo-core CPUs, running at least 2-2.5GHz</li> <li>• 64-512GB of RAM</li> <li>• Bonded Gigabit Ethernet or 10Gigabit Ethernet (the more storage density, the higher the network throughput needed)</li> <li>• 4-12 nodes</li> </ul> <p><b>Software Requirements Recommended</b></p> <ul style="list-style-type: none"> <li>• Linux / EC2 (64 bit)</li> <li>• LAMP</li> <li>• HBASE</li> </ul> <p><b>Supported File Formats</b></p> <ul style="list-style-type: none"> <li>• Flat Files</li> <li>• ASCII</li> <li>• CSV – Comma Separated Value</li> <li>• Parquet</li> </ul> <p><b>Reporting Formats</b></p> <ul style="list-style-type: none"> <li>• CSV</li> <li>• Integration with Tableau and Quilkview</li> </ul>	<p><b>Supported Databases</b></p> <ul style="list-style-type: none"> <li>• Apache Hive</li> <li>• Amazon Redshift</li> <li>• Amazon S3</li> <li>• Cassandra</li> <li>• Cloudera Impala</li> <li>• HP Vertica**</li> <li>• IBM DB2 • IBM Netezza®*</li> <li>• Microsoft SQL Server®*</li> <li>• MongoDB</li> <li>• MySQL®</li> <li>• NetSuite</li> <li>• Oracle*</li> <li>• Teradata® • Teradata Aster</li> </ul> <p><b>Data Source Support</b></p> <ul style="list-style-type: none"> <li>• Amazon EC2</li> <li>• Microsoft Azure</li> <li>• Hortonworks</li> <li>• Cloudera</li> </ul> <p><b>Integration</b></p> <ul style="list-style-type: none"> <li>• Scheduling system</li> <li>• Hadoop Platform</li> </ul>
---	--