Why Validate Data?

“Untrustworthy, low quality data causes organizations to reduce returns on information investments and exposes them to increased business risks, just at the time when such constraints limit the value of a growing array of newly available data and sources.

Data silos seriously cripple organizations' ability to compete, comply with regulations and govern themselves. If the data silo blight cannot be contained, the costs tend to be excessive.”

Gartner “Cool Vendors in Information Innovation and Governance, 2017”, May 2017, Andrew White, Svetlana Sicular, Saul Judah
The Big Data Validation Challenge

Large volume of data in a variety of formats (NoSQL, Hadoop, SQL, etc.) flowing via old and new platforms has decreased reliability of Data Quality. Big Data users are struggling to answer these basic questions:

- "How do I validate data at rest (ex., Data Lake)?"
- "How do I validate data in motion (regular incoming data)?"

Data Quality: “Is my Big Data complete, current, unique, consistent, reasonable, and accurate?”

Data Matching- Target vs Source: “Has the data truly landed from the source to the target location?”

Traditional tools were architected for regular/small data and their older architecture cannot process Big Data efficiently. NoSQL data sources are not easy to handle and have no solution. A large number of variables that have to be validated do have clear quality guidelines which lead to known and unknown data quality risks.

New Paradigm

- **Big Data Architecture**- Tool must not be propped up on old architectures
- **New data source compatibility**- Must handle NoSQL, Hadoop, SQL, etc.
- **Intelligence**- Must detect unknown and unexpected data quality risks/threats, by scanning a combined >100k columns in 1000’s of Tables continuously

DataBuck– Autonomous Data Quality Validation

DataBuck eliminates anticipated and unanticipated data quality errors, without coding. It’s powered by Machine learning to autonomously create 1000’s of Data Quality Fingerprints and validation checks without any manual intervention for Data Quality testing and Data Matching. Spark and specialized algorithms make it >10x faster than any approach, home-grown or otherwise. Errors can be filtered out autonomously from all major platforms (SQL/no-SQL/Hadoop/Cloud) in just 3 Clicks.

**BE UP-AND-RUNNING IN FOUR HOURS**