Turing Award winner's Insights on Data Reliability

Turing Award Winner and MIT Professor, Dr. Michael Stonebraker wrote a white paper outlining his transformative view on data (1). He believes real digital transformation must start with clean, accurate, consolidated data sets. These ideas are already driving major change at GE, HPE, Thomson Reuters, and Toyota. This is a summary of his paper.

Q: Why are companies struggling with unclean, inaccurate and unreliable data?
Data Quality inconsistencies and errors creep in as data moves between IT systems. Companies are struggling to sieve out Data Quality issues because they still use tools that were designed for the 90's to handle simple data problems, which cannot cope with the enormity and complexity of today's data.

According to Prof. Stonebreaker, key elements that will be required for producing good data are:

- Automatic Operations: Any scalable system must perform the vast majority of its operations automatically.
- Parallel Processing: To scale, any unification computation must be run on multiple cores and multiple processors.
- Superior Algorithms: parallel algorithms with lower complexity than N ** 2 is required for truly scalable applications.
- Machine Learning: A rule system implementation will not scale. Only machine learning systems can scale to the level required by large enterprises

As a practitioner, we were delighted to read Prof. Stonebreaker was echoing exactly what we at FirstEigen saw firsthand. Interacting with our customers we saw that looking for errors in vast amounts of data was like looking for a needle in a haystack. It's a very complex problem for large data sets, flowing at high speeds, from many different sources, via many different platforms. It's a nightmare for the coders and for the people who want to make decisions based on that data.

From our experience we concluded, similar to Prof. Stonebraker, that the hallmarks of the best-of-breed data validation tool are:

- Highly efficient in validating Data Quality and ensuring its fit for use.
- Must understand what data is expected with very little or no training of models it builds. Unsupervised Machine Learning algorithms have to be preferred to minimize manual coding.
- Must handle both small and big data elegantly. Nearly every tool available today runs into a time-out issue for even medium sized data sets. They read Big Data formats but processing under the engine is still very, very slow.
- Must trap both expected and unexpected data threats with no programming needed. No more trying to predict how or where data will mutate or data errors will creep in and writing code to trap those errors.

We spotted this problem 2 years ahead of Prof. Stonebreaker and developed DataBuck, the first and only Autonomous Data Quality Validation tool powered by Machine learning (2). It autonomously learns data's expected behavior and creates and applies 1000's of data validation checks across 1,000's of tables with minimal manual intervention, for Data Quality and Data Matching. At the same time, it allows enterprises to keep a balance between accuracy and amount of manual involvement needed in data validation. Built on Spark platform with specialized algorithms, it delivers incredible processing speed over any other tool. Errors can be filtered out autonomously from multiple data sets in just 3 Clicks.

DataBuck is the only tool in the market that can do what Prof. Stonebreaker has laid out so eloquently. DataBuck was honored by Gartner as the most innovative Data Quality tool ("Cool Vendor") (3), and by IDC with the "IDC Innovator" award for being the first to leverage cutting edge Machine Learning in Data Quality validation. IDC also recognized FirstEigen as one of the top vendors for Data Lake Quality assurance solutions in their report ("Are Your Data Lakes Failing — What Can You Do About It?"):

References

(1) https://www.tamr.com/dr-stonebrakers-seven-tenets-scalable-data-unification/
(2) htt://www.firsteigen.com/databuck/
(3) https://www.gartner.com/doc/3725017
(4) https://www.idc.com/getdoc.jsp?containerId=prUS43636218
(5) https://www.idc.com/getdoc.jsp?containerId=US43199117