Big Data Validation Case Study *

Chunli Xie
Department of Computer Science & Technology
Jiangsu Normal University,
Xuzhou, China
xcl_bhb@13.com

Jerry Gao
Department of Computer Engineering
San Jose State University
San Jose, USA
jerry.gao@sjsu.edu

Chuanqi Tao
College of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing, China

***Abstract*—With the advent of big data, data is being generated, collected, transformed, processed and analyzed at an unprecedented scale. Since data is created at a fast velocity and with a large variety, the quality of big data is far from perfect. Recent studies have shown that poor quality can bring serious erroneous data costs on the result of big data analysis. Data validation is an important process to recognize and improve data quality. In this paper, a case study that is relevant to big data quality is designed to study original big data quality, data quality dimension, data validation process and tools.**

Keywords—big data quality, big data validation, data checklist , big data tool, case study

I. INTRODUCTION

The use of big data and analytics has received significant attention in industry, retail, banking, government and education. More and more companies and researchers recognize the importance of data and begin to utilize big data to make better business decisions. Big data is the collection of large amounts of data from places like web-browsing data trails, social network communications, sensor and surveillance data that is stored in computer clouds. As we know, Google offers Gmail, a free email service, consisting of a personal account, a phone number, and an email title. Google may recommend advertisements to you according to your email information or email receivers. Some companies are utilizing external consumer data for the primary purpose of helping to improve their respective business operations. For example, the retail industry is increasingly employing data mining techniques to analyze the buying behavior of its customers and using some predictive analytics look for future consumers by utilizing external consumer data from some social network websites. Almost everyone has realized the importance of big data in our lives.

Due to the huge volume of generated data, the fast velocity of arriving data, and the large variety of heterogeneous data, the quality of data is far from perfect [1][2]. As we know, poor data quality will make significant effects on company decisions. So, the original big data quality becomes an important and critical issue in academic research topics [3][4][5]. Data validation is an important step to improve data quality. Almost all kinds of enterprises have begun to pay attention to big data validation. For example, sensor data from the flash memory cards is initially processed by software tools provided by the sensor manufacturer. The data should be converted from a storage format used on the memory card to a text format for post processing [6]. To help analysis some complex big data, such as clinical trials data, except machine learning methods [7][8], some visualization-aided validation methods are applied to improve data quality [9]. For thermodynamic big data, an automated framework was proposed to identify which data are consistent [10]. A data validation component is generated plugs-in a big data application to measure the amount of unpredictable data during storage [11]. A framework for big data quality assurance was presented which implemented data management, analysis, discovery, applications for genetic big data [12]. Although these published papers addressed data validation in the past, most of them focus on some special data quality dimensions. There is a need in research work to study big data validation.

This paper is written to conduct a series of big data quality case study for one big data source. The case study reflects data validation tool surveys, data validation process, data validation result. It presents our informative study on big data validation for quality. The rest of this paper is organized in three sections. Section II presents big data validation process and checklist. Section III describes case study and lists data validation tools. Case study findings are presented in Section IV. Finally, conclusion and future work are presented in last section.

II. BIG DATA VALIDATION PROCESS AND CHECKLIST

A. Big Data Validation Process

Data validation is to define data validity, data completeness and data consistency and validate data is trustworthy, accurate and meaningful. It's been reported that more than half of the time spent in big data projects goes towards data cleansing and preparation. This section discusses the validation process for big data. As shown in Fig.1, data collection, data cleaning, data transformation, data loading and results report are the necessary data validation process. The detailed illustration for data validation process is as follows.

This work is supported in part by the National Natural Science Foundation of China under Grant No.61502212, the National Natural Science Foundation of China under Grant No.61402229, the Postdoctoral Fund of Jiangsu Province under Grant No.1401043B, the Open Fund of the State Key Laboratory for Novel Software Technology (KFKT2015B10) and the Natural Science Foundation of Jiangsu Normal University under Grant 14XLA01.

978-1-5090-6318-5/17 $31.00 © 2017 IEEE
DOI 10.1109/BigDataService.2017.44

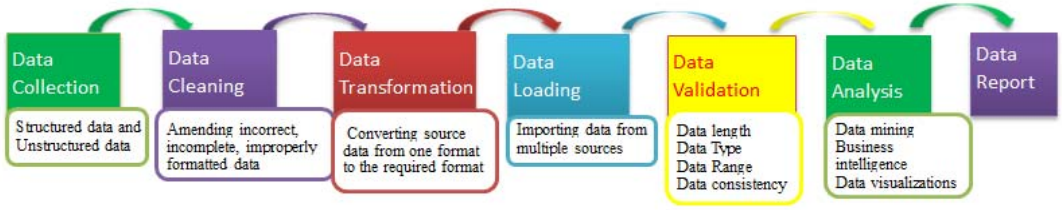

Fig.1. Big data validation process.

- ***Data collection***

  This stage involves collection of data from several types of data sources, data marts and data warehouses, such as from Emails, sensors and smart mobile devices.

- ***Data cleaning***

  Correct or remove corrupt or inaccurate records from a record set, table or database. The major purpose is to detect and identify incomplete, incorrect, inaccurate, irrelevant data parts in data sets [13].

- ***Data transformation***

  Big data types are diversity: structured, unstructured, geographic, real-time media, natural language, time series, network and linked. It is necessary to convert a set of data values from the data format of a source data system into the data format of a destination data system [14], transforming data to specific set of values, data types, timestamp, range of a certain constraint, etc.

- ***Data Loading***

  Data are loaded into a big data repository, for example, a Hadoop environment or a NoSQL Big database.

- ***Data Validation***

  Write customized expressions to check data quality, for example, check the data type, formats, phone numbers, numerical values, positive/negative values, greater than previous, greater than range, less than range, lesser than next/previous, etc.

- ***Data Analysis and Report***

  The last step is to write a document to report data validation results. How much valid records or invalid records in the data? Data consistency, data completeness and other data quality dimensions are completely analyzed. The limitations of tools are also reported in the report.

B. Big Data Checklist

In TABLE I, we list the big data validation checklists for different types of big data. The listed items from data type check to data integrity check are the basic data validation which is applicable to all kinds of data. In practice, there are also different special checks for different data. For example, the items from communication reliability check to data location check are the special check list for sensor data, and the items from coordinates validation check to point clouds check are the special check list for geospatial data and so on.

III. CASE STUDY

Other case studies in this investigation deal with initiatives that are already considered for big data, such the size of their data sets, the technology being performed, or the analyzed results [14][15][16][17]. The intent of our case study, however, is to show the quality of original data sources, the validation process of big data and the results of some existing data validation tools. Participants in the two case studies included the students of San Jose State University at computer engineering department associated with the big data quality project, and the participants are divided into four teams which choose different data validation tools, including advanced ETL processor professional, Talend Open Studio, Data Cleaner, Pentaho, Query Surge, Splunk, etc.

A. Case Study Description

Data source is the Public Weather of North California, which has 8601 raw records. There are some sensors per weather station measuring phenomena, such as city, county, highest temperature, lowest temperature, max 24hr precipitation, max 24hr snowfall, mean temperature, coldest maximum and average minimum. In addition to location attributes such as latitude, longitude, and elevation, there are also links to locations in GeoNames that are near each weather station. We set some basic data validation criteria for the data set, such as data type check, data location check, data consistency check, data limitation check, null value check and et al.

B. Big Data Validation Tools

We study some data validation tools and list the selected tools in the four teams as shown in TABLE II. The first row is their makers, and the last row is the team who selected the tool as their validation tool in this case study. The supported data validation process are listed which can help you select your required tools for your business.

Supported OS - As shown in TABLE II, Windows, Linux and Mac are the three most popular platforms for most data validation tools. A few of these tools, such as Datameer also support other operation platforms, such as OSX, Debian, Centos, Ubuntu and Solaris.

TABLE I. BIG DATA QUALITY CHECKLIST

Data Check List	Sensor Data	Geospatial Data	Health Record Data	Description
Data Type Check	√	√	√	Checks the data type of the input, e.g., In an input box accepting numeric data.
Data Limit Check	√	√	√	Data are checked for one limit only, upper or lower, e.g., data should not be greater than 2 (<=2)
Data Logic Check	√	√	√	Checks that an input does not yield a logical error, e.g., an input value should not be 0 when it will divide some other number somewhere in a program.
Data Range Check	√	√	√	Checks that the data is within a specified range of values, e.g., the month of a person's date of birth should lie between 1 and 12
Data Presence Check	√	√	√	Checks that important data is actually present and have not been missed out, e.g., customers may be required to have their telephone numbers listed.
Data Format Check	√	√	√	Checks that the data is in a specified format, e.g., dates have to be in the format DD/MM/YYYY.
Null Value Check	√	√	√	Check for nulls. Are there mandatory values, or are null / empty values allowed?
Spatial Accuracy Check		√		Check the accuracy of the spatial component, measured by dimensionality. For example, for points, data accuracy can be defined as the discrepancy between the encoded location and the location as defined in the specification.
Temporal Accuracy Check		√		Check the agreement between encoded and actual temporal coordinates.
Data Completeness check	√	√	√	Check the amount of valid data obtained from the overall measurement system compared to the amount of data collected and submitted for analysis.
Data Consistency check	√	√	√	Check the relevant uniformity in data.
Data Validity check	√	√	√	Check a process for ensuring the value is consistent with respect to its context.
Data Correctness check	√	√	√	The value is valid but may not be correct, for example, a ZIP code is a valid number, but it is not consistent with a street.
Data Conformity check	√	√	√	Check how well data adheres to standards and how well it's represented in an expected format.
Data Duplication Check	√	√	√	Check for duplicates, such as a unique key.
Data Integrity Check	√	√	√	Ensures that all data in a database can be traced and connected to other data.
Communication Reliability Check	√			Check the reliability of communication in the sensors.
Data Timeliness Check	√			Check the time expectation for the accessibility of data. Timeliness can be measured as the time between when data is expected and when it is readily available for use.
Data Location Check	√			Check the sensor where the collected data is located.
Coordinates Validation Check		√		In the case of geometry with a texture, check for associated texture coordinates.
Solid Boundaries Validation Check		√		Check the unclosed boundaries, invalid projection, incorrect face orientation, unused vertices, free faces
CAD Data Check		√		Ensure the robust extraction of layers, geometry, text, line types, blocks, extended entity data, etc.
XML / JSON Check		√		Validate the syntax or schema.
Point Clouds Check		√		Check for correct components and values.
Representational Adequacy Check			√	The extent to which an operationalization is consistent with/differs from the desired concept (validity), including but not limited to imprecision or semantic variability, hampering interpretation of data.
Information Loss/Degradation Check			√	Including but not limited to reliability, change over time, and error.
File Format Check				Check the files are DNG or write-once medial.
Data Transfer Error Check				Check common cause of missing or corrupted files occurs in transfer.

Supported File Format – All of the listed tools have a wide range support for different types of data files formats. The commonly supported file formats include: CSV/TSV, TXT Files, Fixed Width Text, HTML, and Server Log File.

Supported Database –On the other hand, all of these selected tools support different types of databases, including both relation databases and non-relational databases. The most popular supported relational databases include MySQL, DB2, Oracle, PostgreSQL, Vertica, and Teradata. The commonly used non-relational databases include Hive and Hbase. In addition, Datameer also supports Windows Azure Blob Storage and Amazon Redshift. If you want to know more information about the tools, you can read the paper written Gao [1].

C. Case Study Results

Basic data validation criteria, such as null value check, data type check, data value check and data logic check are set by almost all teams for the data set. The test results are shown in tables or figures. We only count the results provided by some teams who display the result in tables. TABLE III shows the results. From the table, we can see the results are almost the same when do null value check, data type check, data value check and data presence check. However, there is a big difference when check data logic even if the criterion is the same which is set to be the mean temperature cannot be empty if the minimum temperature and maximum temperature are not null. The result may be caused by the two different testers or by the tools.

TABLE II. BIG DATA TOOLS LIST

Features	Tool Name	DataCleaner	Datameer	Tableau	Pentahao	Advanced ETL Processor	Talend	Zoho	Query surge	Splunk
Maker		DataCleaner Team	Datameer	Tableau	Pentaho	ETL-Tools	Talend	Zoho	IBM	Splunk
Supported OS	Windows	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Linux	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
	Mac	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Supported Validation Process	Data Collection	No	No	No	No	No	No	No	Yes	Yes
	Data Cleaning	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Data Transformation	Yes	No	No	No	No	No	Yes	Yes	No
	Data Loading	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Data Analysis	Yes	Yes	No	Yes	No	No	Yes	Yes	No
	Data visualization	No	Yes	Yes	Yes	No	No	Yes	Yes	No
File Format	CSV/TSV	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Server Logs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Fixed Width Text	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	HTML	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	XML	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	spreadsheets	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	NoSQL	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes
Supported Database	RDB	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Cassandra	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes
	HBase	Yes	Yes	No	Yes	No	Yes	Yes	Yes	No
	MongoDB	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes
	CouchDB	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
	Hive	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Team		3	1,4	1,2,3,4	1,2,3	4	1,3	2		2

TABLE III. BIG DATA VALIDATION RESULTS

Tools / Data Checklist	Zoho	Pentaho	Tableau	Talend
Null Value	28722	28543	28435	28560
Data Type	4415	4234	4302	4412
Data Value	8926	6858	7523	7843
Data Presence	0	0	0	0
Data Logic	2718	2858	2753	2718

TABLE IV. BIG DATA VALIDATION RESULTS

Tools / Data Checklist	DataCleaner	Tableau	Talend	Pentaho
Data Duplicate	0	0	4	0
Data Inconsistency	28697	28697	29704	28698
Data Incompleteness	2834	11410	1652	2832
Out of Boundary	162	2817	162	162
Data Range Difference	1476	0	1400	1475

TABLE IV is the validation results got from team 3 by four different tools. They set four type data validation: data duplicate, data inconsistency, data incompleteness, and data range check. The reason why we show the results provided by team 3 is that this team give a complete results for this four types test. There are some abnormal results in this table, for example, the number of data incompleteness, out of boundary and data range check by Tableau are too large or too small. So, we think some problems occur either in the tool or in the test operation. Maybe the set range by the tester is unreasonable. The Fig.2 shows the analysis results from four teams. We try to find valid row count, distinct count, duplicate count and null value count using two same tools by the four teams. For example, the valid row count is 8601, distinct count is 383 and duplicate count is 318 from advanced ETL by team 1. Each team has six columns and the first three columns show the results of Advanced ETL and the next six columns show the results of Talend. From Fig.2, we can see that the valid row count from advanced ETL performed by three different teams is almost the same but the results of the distinct count are different, 383 distinct records are found by team 1 but 0 distinct count is found by other three teams. In a word, any two teams didn't get the same result.

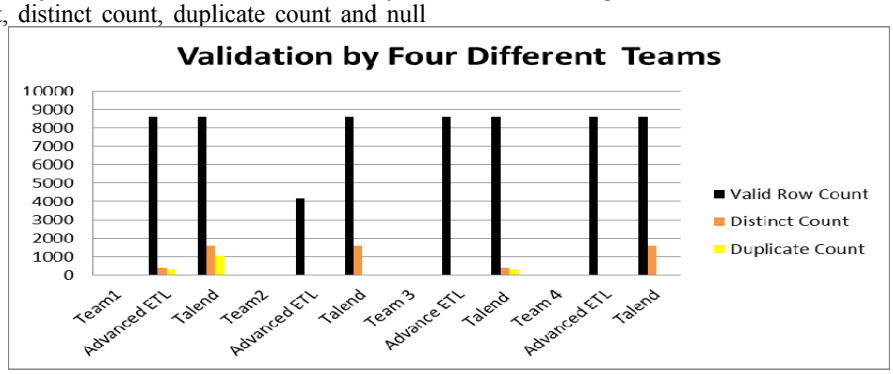

Fig.2. Data validation results by four teams.

IV. CASE STUDY FINDINGS

This section presents findings about data quality validation for sensor big data.

Finding 1: Too much null values in big data

The case study shows about 1/3 original data is null. It is reported that due to lack of data quality management, on the web, most of the data are stored in XML documents, among which only one third documents are valid [18]. A simple error of missing tags or missing values make the data lack consistency, accuracy, completion, and other problems associated with lack of quality. Because of this, the importance of data management has come to be realized. There has been increasing demand in industries and academia for developing data quality management systems, aiming to effectively detect and correct errors in the data, and thus to improve data accuracy and other data quality dimensions.

Finding 2: Existing data validation tools can help users to clean the invalid data, improve data quality

Due to the large amount of data, a manual process for data validation is time consuming and inaccurate. Data quality tools can help industries or companies to detect and correct data errors. Thus, people need to develop data quality tools to assure data quality in business processes. Indeed, the market for data quality tools is growing at 16% annually, over the 7% average forecast for other IT segments [19]. These data validation tools

can do domain validation, rule validation, business validation. These tools indeed help do data validation, some advantages are listed below:

1) Before applying the validation rules, the out of range criteria could not be easily identified but with the criteria the data falling beyond the range could be easily identified.

2) The value range is easily identified with the criteria specified.

3) The completeness criteria are easily identified when the missing data is easily represented in the graph.

4) The conformance criteria make the representation of the graph easy as the data that meet the standards can be easily identified.

5) The unique data can be found through the data duplication check.

Finding 3: Limitation of existing data validation tools

The whole data process includes data collection, data integration, data loading, data analysis, data warehousing and data visualization. Splunk is designed to validate data in the process of data collection. Data integration is the important process, so most of the tools, such as Datameer, Pentaho, Advanced ETL Processor, Talend and Querysurge, perform data validation using this process. Data analysis is the most important data validation process, in this process, you can set some business rules such as the rule for comparing price, the rules for adjusting the price and so on. DataCleaner, Datameer, Pentaho and Querysurge are designed to improve your data quality in the process of data analysis.

Finding 4: Lack of big data engineer

In our case studies, at the first time, we asked all teams to do data validation without any information. We asked them to do it again after they finished the first test and we found that most of the teams without training only did a part of the validation test, and the trained teams did multiply validation tests. For example, the tester missed consistency check and conformance check in the first time and did duplication check, missing value check, inconsistency check, incompleteness check, boundary check, and range check in the second test. So, if you want to get a good result from data validation, you need to know how to well define big data quality [20].

V. CONCLUSION AND FUTURE WORK

In this paper, some big data validation criteria are set, and some different data validation tools are selected to do a case study on public weather sensor data by four teams. The functions, features and limitations of tools are discussed in detail. About ten selected data quality dimensions are checked and the results are presented and discussed. Some open data quality issues have not been solved and should be studied in the future. Due to lack of available research results on big data quality models and quality evaluation metrics. The check done is not complete. We are going to define and develop big data quality models and metrics for big data validation tools so they will be useful.

REFERENCES

[1] J. Gao, C. Xie, and C. Tao. “Big Data Validation and Quality Assurance-- Issuses, Challenges, and Needs”. In Proc. of 2016 IEEE Symposium on Service-Oriented System Engineering (SOSE), Oxford, UK, 2016, pp.433-441.

[2] P.Woodall P, J.Gao, A. Parlikad A, and A. Koronios. Classifying Data Quality Problems in Asset Management. Springer International Publishing, 2015.

[3] JJ. Gassman, WW.Owen, T.E. Kuntz, JP. Martin, and WP. Amoroso. “Data quality assurance, monitoring, and reporting”. Controlled Clinical Trials, 1995, vol.16,no.2, pp.104-136.

[4] N. Laranjeiro,S.N. Soydemir, J. Bernardino. “A Survey on Data Quality: Classifying Poor Data”. In Proc. of IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC), IEEE, 2015, pp. 179-188.

[5] D.Becker, T. D. King, and B. Mcmullen. "Big data, big data quality problem." In Proc of IEEE International Conference on Big Data IEEE, 2015, pp.2644-2653.

[6] B.Wallace, R.Goubran, F. Knoefel, S. Marshall, M. Porter, and A. Smith, "Driver Unique Acceleration Behaviours and Stability over Two Years", In Proc of 2016 IEEE International Congress on Big Data (BigData Congress), 2016, pp. 230-235.

[7] P. A. Traganitis, K. Slavakis and G. B. Giannakis, "Big data clustering via random sketching and validation," In Proc of 48th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, 2014, pp. 1046-1050.

[8] KI. Noguchi, Y. Sato, H. Shiohara. “Advanced, high-performance big data technology and trial validation” NTT Technical Review, 2016, vol. 14, no. 2.

[9] Z. Zhang, H. Fang and H. Wang, "A New MI-Based Visualization Aided Validation Index for Mining Big Longitudinal Web Trial Data," IEEE Access Practical Innovations Open Solutions, vol. 4, 2016, pp. 2272-2280.

[10] P. Buerger P, J. Akroyd J, JW. Marti and M. Kraft."A big data framework to validate thermodynamic data for chemical species." Combustion and Flame, vol.176 ,2017, pp. 584-591.

[11] S. Barahmand and S. Ghandeharizadeh, "Benchmarking Correctness of Operations in Big Data Applications," In Proc of *IEEE* 22nd International Symposium on Modelling, Analysis & Simulation of Computer and Telecommunication Systems, Paris, 2014, pp. 483-485.

[12] J. Ding, D. Zhang and X. H. Hu, "A Framework for Ensuring the Quality of a Big Data Service," 2016 IEEE International Conference on Services Computing (SCC), San Francisco, CA, 2016, pp. 82-89.

[13] T, Nan. Big data cleaning, Asia-Pacific Web Conference, Springer International Publishing, 2014, pp.13-24.

[14] P.Rajkumar, “17 important case studies on Big Data”. Retrieve at:http://bigdata-madesimple.com/17-important-case-studies-on-big-data.

[15] R. Petersen, “37 Big Data Case Studies with Big Results”. Retrieve at:https://www.businessesgrow.com/2016/12/06/big-data-case-studies.

[16] D. Becker, TD.King, B. McMullen. ”Big data, big data quality problem”. In Proc of IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, pp.2644-2653.

[17] D. Becker, TD.King, B. McMullen. “Big Data Quality Case Study Preliminary Findings”. Bedford, MA: National Security Engineering Center, 2013.

[18] S. Grijzenhou, M. Marx. “The quality of the XML web”. Web Semantics: Science, Services and Agents on the World Wide Web, 2013, vol.19, pp. 59-68.

[19] P. Paygude, PR. Devale. “Automated data validation testing tool for data migration quality assurance”. The International Journal of Mechanical Engineering and Research (IJMER), vol.3, no.1,2013, pp.599-603.

[20] C.Tao, and J. Gao. "Quality Assurance for Big Data Applications–Issues, Challenges, and Needs". In Proc of the Twenty-Eighth International Conference on Software Engineering and Knowledge Engineering, Redwood City, San Francisco Bay, California, 2016.