Pioneering Data Quality and Security in Smart Grid

Bright Roy, Prashant Tripathi, and Rahul Goel

Abstract The essence of a data lies in the hidden information inside it. If the data is not of good quality or not sufficiently protected, the outcome will undoubtedly be harmful. Quality and Security are two essential aspects that add value and meaning to the data and their implementation has become a real need and must be adopted before any data exploitation. Due to the high volume of data generated every day, the effective implementation of such systems requires well thought out mechanisms and strategies. This paper provides a detailed analysis and solutioning of Data Quality and Data Security in the context of Smart Grid. Through this paper we want to highlight the proposed solution and challenges that may exist during the implementation of data security and data quality management systems.

Keywords Smart Grid · Data quality · Security · Consistent · Accuracy · Machine learning · Utility

1 Introduction

With over 1000+ serving Customers in the Utility Sector, all Smart Grid utilities is ever growing with enormous amount of data which gets generated in the Smart Grids due to the huge number of endpoints, higher rates of measurement and various types of hops in between. Smart Grid data consists of some critical and important information about the grid. Applications which are driven by data are being constantly developed and improvised for better monitoring, operation and planning of Smart Grids. The outcome of data analytics heavily depends on the quality data coming of

B. Roy (✉) · P. Tripathi · R. Goel
GDS, Landis+Gyr, Noida, India
e-mail: bright.roy@landisgyr.com

P. Tripathi
e-mail: prashant.tripathi2@landisgyr.com

R. Goel
e-mail: rahul.goel@landisgyr.com

R. K. Pillai et al. (eds.), *ISUW 2021*, Lecture Notes in Electrical Engineering 843,
https://doi.org/10.1007/978-981-16-8727-3_12

the attached Meters, Collectors. However, up till now we haven't seen much work that has been reported on the smart-grid data's quality assessment.

This whitepaper addresses the objective assessment of Smart Grid data quality and data security. Different Smart Grid data quality dimensions and security are identified in this whitepaper. Few mathematical formulas have also been proposed to quantify our Smart Grid data quality and the data quality metrics have been proposed to be applied on a newly built service: **Data Quality and Security Services (DQSS)**.

DQSS will be a potential service which will be capable to improve Data Quality of any data of Smart Grid utilities, imposed into it. From data security perspective, it will have additional customization features to put inbound and outbound port/protocol restrictions to the data as part of vulnerability test, apart from providing encrypting capabilities.

2 Problem Statement

The modern Smart Grid reflects a combination between Information and Communication Technologies (ICT) and Internet of Things (IoT), whereby data services such as aggregation of sensor data and analysis of voltage consumption from our smart meters offer a foundation for the concept of smartness.

The current data quality problems in Smart Grid are addressed still in an ad-hoc style. For example, we in smart utilities focused on the outlier detection of electricity consumption data through several application reports. Their solution tackles a specific quality aspect of electricity consumption data. However, this will obstruct our engineers to foresee the other data quality problems and delay the reaction on time for potential data quality problems.

Example: In the past utilities have experienced multiple firmware reset events in the field for a significant number of meters in hot/humid environments. The issue was initially looked at the outlier approach and engineers kept digging the problem in the network and work on the workarounds to address the issue whereas later it found, it's the behavior where a super capacitor shows symptoms of venting under certain hot/humid climate conditions. This venting can cause elevated counts of processor reset incidents within the meter or inhibit communications altogether. For example, an outlier in the energy consumption data may be caused by missing data items or data corruption.

Therefore, focusing on specific quality aspects can mislead the root causes of the data quality problems. Based on our review, there is a lack of a systematic framework for managing data quality in Smart Grids. Also, data quality is critical in the Smart Grid domain, as invoices of end users depend for example on the collected power consumption data.

Similarly, communication networks in Smart Grid bring increased connectivity right from meters, routers, collectors and interconnected and independent products like Meter Data Management (MDM), Head End System (HES) which involves increased severe Security vulnerabilities and challenges. Smart grid can be a target for cybercrime because of its critical nature. However, most of the classified attacks are based on confidentiality, integrity, and availability. These exclude the attacks caused due to compromise of accountability.

Data Quality and its Security are the two main pillars to focus as the utilities are now stressing on these components as their prime requirement and are must to survive in this competitive environment.

3 Problem Impact

Based on research by Gartner, “the average financial impact of low data quality on any organizations is $9.7million per year”. Poor data impact financial resources, as well as it will also negatively impacts your efficiency, productivity, and credibility.

In fact, “IBM estimated that poor quality data cost the company $3.1 trillion in the U.S. alone in 2016”.

3.1 Less Productivity and Growth

Bad data quality can obstruct growth of business and decrease the productivity all over the Smart Grid utilities. Even a single percent bad data can lead to many other issues and it can be very difficult to trace back to find out the issue and get rid of that cause. Even this will interrupt related processes and cause a lot of unwanted efforts. All these things will reduce the overall the productivity, as it needs a lot of effort to neutralize all the adverse effects.

3.2 Increased Financial Costs

Poor data quality will not impact only on business strategies, even it will increase the financial cost for the productivity, customer support and so on.

3.3 Data Security Breach

Poor data quality is the biggest threat to data security. As, characteristics of data quality: confidentiality, integrity, and availability can be the reason of the data breaches in any system like smart meters, collectors, and related products.

3.3.1 Revenue/Financial Loss

In Addition, considering the Cost of Data Breach Study in 2018, “the normal cost of a data breach in the U.S. is $7.91 million”.

There are the following most significant consequences of data breaches. A non-functional product of Smart Grid utilities may trigger potential customers to explore other alternatives.

As per the analysis that “29% of any businesses which cope with a data breach end up losing revenue”.

3.3.2 Damage to Brand Reputation

Apart from revenue loss, data security breach can affect the reputation of our organization which is the long-term practice for any organization.

Customers value their confidentiality and privacy too. New and existing customer can be uncertain to trust a energy business because of poor data security.

4 Mitigating the problem—The L&G Solution DQSS

4.1 Purpose

To create a service named as **DQSS** which will be able to validate the data quality and security of any kind of dataset as per the configured data quality dimensions and allow the user to optimize the data quality of any product/application of Smart Grid utilities. This whole process can be automated and integrated with any product/application of our organization.

To accomplish the purpose, we can divide the solution prominently into 4 layers as shown in Fig. 1.

4.2 Solution

4.2.1 Data Integration Layer

L+G system consists of multiple devices including smart meters, routers, collectors, HES and independent products.

Therefore, This **DQSS** product will be having the capability to integrate as Input data/output data layer with other existing Smart Grid utilities product. Even output of any process of MDM, HES can also be integrated to **DQSS**.

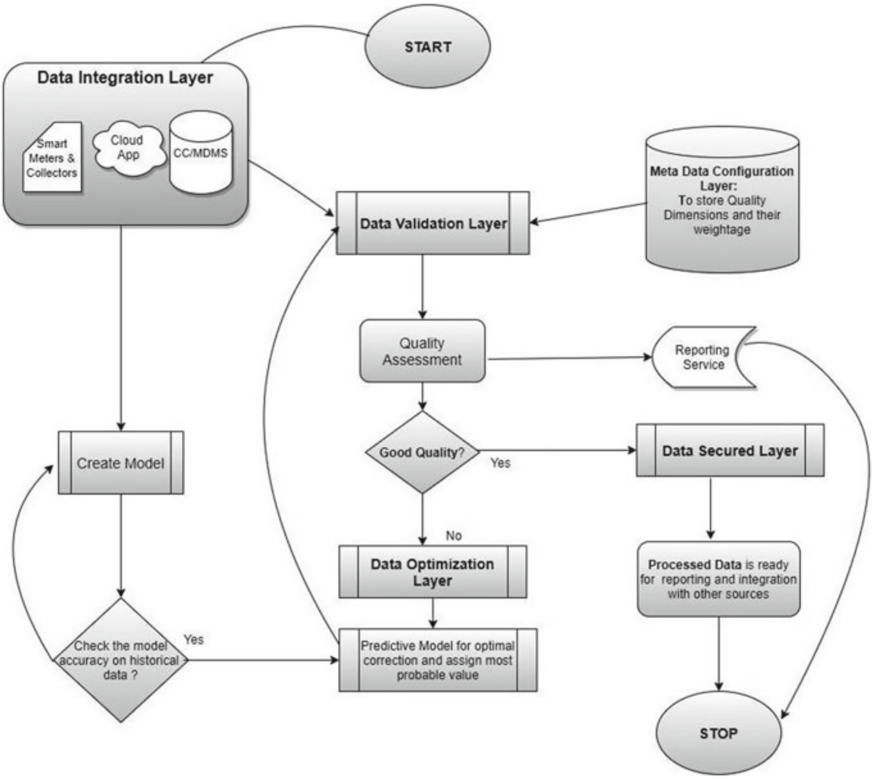
START
Data Integration Layer
Smart Meters & Collectors
Cloud App
CC/MDMS
Meta Data Configuration Layer:
To store Quality Dimensions and their weightage
Data Validation Layer
Quality Assessment
Reporting Service
Good Quality?
Yes
Data Secured Layer
Create Model
No
Data Optimization Layer
Processed Data is ready for reporting and integration with other sources
Check the model accuracy on historical data ?
Yes
Predictive Model for optimal correction and assign most probable value
STOP

Fig. 1 DQSS solution flow diagram

To achieve this, **DQSS** will support connections to consume/produce structured, unstructured and semi structured data with the product/application. Whole data integration layer will support real time/batch streaming of any data set.

DQSS will be having UI capabilities to provide connectors for real time/batch streaming with the applications or products.

4.2.2 Metadata Configuration Layer

To check the data quality, **DQSS** will require data quality dimensions which can be configured as Metadata. We will create/configure standard dimensions set which will be applicable for all datasets. According to the relevance in the products/applications, we can also configure and assign the appropriate weights to each of the dimensions.

(1) Applications like MDM will give more weightage to accurate/complete/consistent dimensions data

(2) Smart Meter based application will give more weightage to validity/missing/consistent dimensions.

(3) Product like smart meters/collectors will give more weightage to timeliness dimension.

DQSS will provide UI to view the standard dimensions set and their weightage according to applications. Even with the help of UI, user can add, remove, and reorder all kinds of dimensions sets and respective weightage. The **DQSS** UI will even highlight the details of each dimensions as well.

The order of the dimensions in the metadata configuration will trigger the order of evaluation of the data quality Assessment layer.

4.2.3 Data Quality Assessments Layer

The objective of data quality assessments for Smart Grid data are based on the evaluation of the quality dimensions. This layer will generate the unified data quality index of the integrated source after evaluating each of the dimensions.

Internally, this process will apply the quality dimension sets to the source data and generate the statistical summary as an output for complete dimensions sets and individual dimension. Descriptions of statistical summary are shown in Table 1.

If source data like smart meters, collector are generating the logs then very first text mining will be used to convert freeform text into structured information, then quality dimensions sets will be used to generate the analysis.

(a) *Evaluation of each dimensions*

Validation of quality dimension can be evaluated based on its computation. Consider, for example, in any source data of application/product has m total records and n are the parameters (columns) like reads, voltage, rate, frequency etc.

- Consistency: Any application of Smart Grid utilities have data set can will be consistent if two or more values do not conflict with each other. The consistency will be checked for each parameter or with respect to total number of parameters at a certain time. The number of ambiguous instances will be equal to the number of such repeated values.

Table 1 Statistical summary of quality

Statistical summary		
Counts	Passed	Failed
A statistical summary of the records that passed and failed the rules defined in the ruleset	A list of all the records that passed all rules/individual defined in the ruleset	A list of all the records that failed any of the rules/specicfic rule defined in the ruleset, with detail giving the result of each rule for that record

Let t^k be the number of ambiguous entries of parameters denotes by k in dataset of m^k records, then quality of Dataset in terms of consistency is given by

$$_{con}dqi_{ds} = 1/n \sum_{k=0}^{n} \left(1 - t^k/m^k\right)$$

- Timeliness: Timeliness calculates the promptness in storing/updating data. It provides an estimate of the delay involved in capturing the data with respect to the time at which the value got expected/changed in the system. quality in terms of data timeliness is given by

$$_{time}dqi_{ds} = v_{expected} - v_{storage}$$

Consider, for example, smart meter application expects the register reads at the mid-night ($v_{expected}$), but its stored in the database ($v_{storage}$).

- Completeness: Any application of utilities has data set will be complete if all information is present in the respective database. The dataset will be incomplete if it is missing, missing data refers null values.

Let t^k be the number of missing entries of respective parameters denotes by k in dataset of m^k records, then quality of Dataset in terms of completeness is given by

$$_{comp}dqi_{ds} = 1/n \sum_{k=0}^{n} \left(1 - t^k/m^k\right)$$

- Availability: Any application of utilities has data set will be called available if the processed information is ready for the use.

Let t_{req} be the time at which the processed data was requested by the application/user, t_{delay} is the time deadline within which the application/user must get the data and $t_{delivered}$ be the time at which the data was delivered to the user, then quality in terms of data availability is given by

$$_{avail}dqi_{ds} = \left(t_{delivered} - -t_{req}\right) / \left(t_{delay} - t_{req}\right)$$

- Interpretability: Any application of utilities has data set will be interpreted if all information is represented using an appropriate notation in the respective database. The dataset will be not appropriate if data being entered under a wrong column or data values having wrong characters.

Let t^k be the number of uninterpretable data points of respective parameters denotes by k in dataset of m^k records, then quality of Dataset in terms of Interpretability is given by

Table 2 Quality index of each dimesnion

Dimension	Index value
Consistency, accuracy, timeliness, validity, consistency, etc.,	It will provide the measurement of each quality dimension. As per this value, Smart Grid utilities set this value as threshold to optimize the data

$$_{\text{inter}}\text{dqi}_{\text{ds}} = 1/\text{n}\sum_{\text{k}=0}^{\text{n}}\left(1 - \text{t}^{\text{k}}/\text{m}^{\text{k}}\right)$$

For example, between integration of two system, mapping of two fields of sperate is wrong.

(b) *Evaluation unified data quality index*

As per the individual quality index and assigned appropriate weight to each dimension according to their relevance in the application (smart meters, collectors)/product like MDM, HES we will propose unified data quality Index. Let $_{\text{i}}\text{QI}^{\text{k}}$ denote kth quality dimension among n chosen dimensions and W_{k} be the corresponding chosen weights then unified data qualified index of the dataset DS, denoted by $^{\text{u}}\text{DQI}_{\text{ds}}$ is given by

$$^{\text{u}}\text{DQI}_{\text{ds}} = \sum_{\text{k}=0}^{\text{n}} {_{\text{i}}\text{QI}^{\text{k}}}\text{W}_{\text{k}}$$

where, $\sum_{\text{k}=0}^{\text{n}} \text{W}_{\text{k}} = 1$. $^{\text{u}}\text{DQI}_{\text{ds}}$ gives a measure of the data quality of any dataset irrespective of its size or number of parameters.

This complete assessment layer will generate the individual/unified data quality index as per the current dataset of integrated source. With new dataset, will come up by real time/batch streaming, these quality index will keep updating.

DQSS will be having UI capabilities to run the data quality assessment on the dataset of integrated source like MDM, HES, smart meters, collectors etc. and able to provide the interactive reporting analysis of statistical summary of dimensions, individual/unified quality index (Table 2) with other users.

4.2.4 Data Optimization Layer

The objective of this layer to enrich the quality of each dimensions. To enhance the quality dimensions, data analytics and data science will be the standard in **DQSS**.

To enrich the quality of dataset, there is the standard data quality cycle which has the process:

(a) *Analyze Data*

Data quality assessment layer will evaluate and create a matrix of the individual quality index for each dimension.

(b) *Clean Data*

Before enriching the data, it will be cleaned and standardized to meet the data quality goals according to the existing dataset of MDM, HES etc. Consider, for example, text mining will be used to convert freeform text into structured information, which will be used further for analytical methods.

(c) *Create Predictive model*

In current era, analytics and data science do not only demands on data quality, even they will be the best source to improve the quality dimensions. To use the analytics, we will create predictive model using machine learning (ML) for each quality dimension for the dataset of smart meters, MDM, HES etc. there are key points to implement the model in include:

Required historical data for the predictive modeling

Very first examines that whether the amount of data is sufficient for the analysis or not. Required data quantity will be managed in the Smart Grid utilities easily. In case if there are small samples of data, analytics also provides methods for modelling are events.

Consider, for example, for time series forecasting predictive model, analytics has so-called intermittent demand models that can be implemented on small samples of data.

Required variables for predictive modeling

As per the target quality dimension, selection of variables will be different from same input dataset, which will be required variables having strong relationship with target dimension.

Analytics provides number of methods for the selection of variables. Simple metrics like R-square and advanced metrics like LARS, LASSO and ELASTICNET are the methods to select the variables. Consider, for example, forward, backward, and stepwise model selection in regression modelling.

Using the selective variables over the Historical data, **DQSS** will train the model and test the model. Further it will be deployed for respective dataset of MDM, HES.

Monitoring of predictive model quality

Analytics tools are designed to create/trained the predictive model. We need to use the analytics also to assess the model quality time to time.

Re-training of predictive model quality

In case if the assessment of predictive model is low then **DQSS** need to re-train the model.

Even, with new data each time, we can schedule the re-training of the model to enhance the quality of the predictive modeling.

(d) *Enrich Data*

To enrich the quality dimension of data, **DQSS** will use created predictive model of respective dataset of MDM, HES, collectors etc. Let see the quality dimensions and how and which predictive model will be applicable:

Accuracy/Data outliers

With the help of predictive models and time series methods, the calculation of validation limits, optimal correction and most probable value will be done for data outlier dimension.

Missing values/completeness

Using computation algorithm, which are based on analytics methods like decision trees or spline interpolations for time series will compute the incomplete/missing data for average-based or individual values of respective application like MDM, HES, smart meters, collectors.

Consistency/Data standardization

This identification and elimination of duplications can be easily achieved using database analytics. Even measure of closeness and similarity between records will be also evaluated based on business information.

(e) *Monitor and Check Data*

Data must be regularly monitored and checked by data quality assessment layer to guarantee that it maintains the applicable data quality.

4.2.5 Data Secured Layer

The objective of this layer is to secure the qualified data. Sensitive data is encrypted before integrating to other system like MDM, HES etc. To encrypt the data most popular algorithm will be used: "AES, RSA, TRIPLE DES, TWOFISH". Encryption of data can be done on the file systems, block level, bare-metal server, virtual machine, or virtual disks.

Data Security is also applied by applying ports and protocols restrictions to the data.

DQSS will provide the UI capabilities to create a predictive model for the integrated source of MDM, HES etc. with the help of UI this model can be tested and deployed to optimize the data set of integrated sources.

Time to time with the **DQSS** UI, it can monitor, assess and re-train the model.

5 Achitecture and Design

DQSS has been conceptually designed in a 3-tier architecture. This means that the entire application is composed of mainly 3 tiers or layers which provides many benefits to production and development environments by modularizing the user interface, storage, and the business logic into different units.

In standard terms, these three tiers have been named as Presentation, Application and Data Tier. Figure 2 in the next page shows a clear bifurcation of these 3 layers based our product perspective. Let us map these 3 tiers (or layers) with our application.

5.1 *Functional Tier*

As shown in Fig. 2, the middle-represented box is the principal module which contains the business logic where algorithms to qualify and secure the data are written.

The backend language used here is Python with JavaScript for client-side interaction. The code has been modularized into independent pieces which makes the entire application easy to debug for errors.

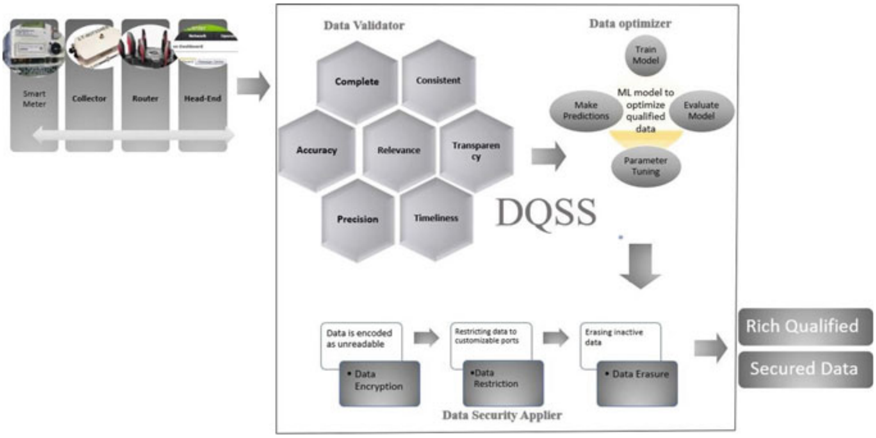

Fig. 2 DQSS design and architecture

5.2 Data Tier

The data tier comprises of the database storage and access layer. In this case, since the data source can be both relational, non-relational as well as cloud, hence the Data-Integration layer has been designed in a flexible way.

For example, Command Center of different utilities are built on different databases like Oracle, SSMS, Postgres. Other data sources also include Big Data.

5.3 UI Tier

The User interface tier also called the Presentation layer is the front end of 3-tier architecture. The UI tier is mostly the graphical one accessible either through web browser or a web-based application.

This tier is built on a web browser supported front end which displays content and information useful to the utilities.

DQSS UI tier is built on simple HTML and CSS templates which communicates to the Functional Tier through API calls using the Data tier.

6 Benefits of DQSS

DQSS as an organization level product is conceptualized keeping in mind the Smart Grid business. By improving the raw data quality and making the data more secured, will provide an economical as well as intangible value-add to the Smart Grid business.

6.1 Product Feature Benefits and Use Cases

6.1.1 UseCase 1: Smart Meter and Collector Logs

Smart meters and collectors generate the logs, which are the best source to optimize the quality of both smart devices. This DQSS will give Smart Grid utilities the capability to analyze the logs to ensure the quality of the smart devices. As per the assessment, it will be easy to find out the root cause of issues of meters/collectors and even help in knowing the root cause of missing/incomplete reads. In general, Smart Grid utilities manufactures the best smart meters worldwide, however DQSS will help the production unit to improve the quality and security of smart meters. Better quality is directly proportional to better security.

6.1.2 UseCase 2: MDM Data Quality

As per the requirement MDM has been able to enhance the smart meter data quality up to certain extent. There are few processes within MDM which has been able to achieve the true capability of the MDM system, though these processes have created redundancy of the data which are causing the issues in reporting and other area. DQSS can provide a second quality and optimization check over the same processes of MDM.

Even internal process had been built using rule-based approach only, which can also be optimized using ML approach of DQSS optimization layer.

6.1.3 UseCase 3: HES and Independent Portals

HES, which are first database source to hold reads, programs, location, rates data directly from smart devices, Therefore, there is a huge scope to enhance the dimension of each quality and security of smart meter data, which will be an additional edge to expand the business, reporting, decisions etc. with respect to Utility and End-Users. The ability of independent portals, which is offering the automated customize reporting, and other services will automatically be enriched.

6.2 Deploying DQSS as a Product to Any Organization

DQSS will be having the plug-play features according to customize requirement, therefore this will not only be effective to Smart Grid Utilities, even it can be deployed as a product to any organization to enrich their data quality and security.

6.3 New Business Leads

Due to capability and features of DQSS, this new generation product will also help getting new business leads for Smart Grid utilities in the Energy Consuming Sector and different domains as well.

7 Challenges and Issue

Although there are no major foreseen challenges in implementing thiss solution still there are few areas which needs to be considered while implementation.

- Testing effort might be high initially due to complex nature of Smart grid, HES and other products linked with HES. Thorough validation is needed to check if all the test cases are covered and data is validated at each interim step without escape.
- Because of the diversity and tremendous data volume generated by smart meters and ingested into command center HES, it is difficult to judge data quality within a reasonable amount of time. This could still be optimized to put minimal impact on overall performance.
- Due to the huge number of available data sources, variety of data types and complex data structures, there can be difficulties in data integration.
- There is an imminent need to setup a unified and approved data quality standards in Smart Grid utilities across all the products.

8 ROI from Proposed Solution

"Data quality: It's a journey, not a destination."

Though it is easy to assume that better data will benefit the organization, quantifying that benefit is critical to securing investment. By quantifying the impact of data quality in a methodical way, you can measure the impact of effort, the value to the business and a tangible return on investment. Later, this ROI can drive future investments and further promote data quality within the organization.

A robust data quality solution automates many profiling and discovery processes, so the business and IT team can accurately assess project risks and measure the enhancement of information performance.

Acknowledgements This whitepaper is a result of the research and development of three writers involved in this project. In the accomplishment of this whitepaper successfully, many people have helped us and whom we would like thank whole heartedly.

Firstly, we would like to thank all the internal teams of Landis+Gyr who assisted us with their extensive domain knowledge and attended to our queries as and when required.

We are also very fortunate enough to get constant encouragement from our office colleagues. So last but not the least, special thanks to our team mates for their secondary support.

We hope this whitepaper finds the readers in good health and spirit and motivate us further to research and write more on booming IT technologies.

References

1. Tahla M, El Kalam A (2019) Trade-off between data quality and data security, April 29–May 2 2019
2. SpotlessData The importance of data quality for data security