Journal of Applied Statistics and Machine Learning
Volume 1; Number 1; 2022 : pp. 27-38

LAYER BASED ARCHITECTURE FOR DATA MANAGEMENT

Siddhartha Roy[1], Girdhar G. Agarwal[2] & Dieudonne Phanord[3]

[1]*Associate Director (Advanced Analytics) Gainwell Technologies, Bangalore, India. E-mail: arpsid@gmail.com; corresponding author*
[2]*Ph.D. Ex-Professor, Department of Statistics, Lucknow University, India. E-mail: girdhar1751@gmail.com*
[3]*Professor, Department of Mathematical Sciences, University of Nevada Las Vegas, USA. E-mail: Dieudonne.phanord@unlv.edu*

ARTICLEINFO

Received: 9 August 2021
Revised: 19 August 2021
Accepted: 6 September 2021
Online: 11 January 2022

To cite this paper:
Roy, S., Agarwal, G.G. & Phanord, D. (2022). Layer based Architecture for Data Management. *Journal of Applied Statistics and Machine Learning*. 1(1): pp. 27-38

ABSTRACT

Regardless of the size or complexity of a research study or reporting, there is usually more than one approach for managing the data that is collected to answer the questions posed by the investigators. In this paper, we suggest Smart Information Assimilation & Aggregation System (SIAAS) that follows a layer-based architecture to handle most of the data issues. The architecture and applications of SIAAS in the COVID-19 situation is discussed in detail.

Keywords: Data discrepancies, Data diagnosis, Reliability, Layer based Architecture, Analytical Layer

REQUIREMENT ANALYSIS

To carry out any research, reporting, analysis, and model building activities, the data needs to be relevant, complete, accurate, meaningful, and actionable. Initiating a study prematurely without proper analysis of requirements harms the employee efficiency, productivity, and overall business. Organizations and professionals are spending a lot of time in planning, controlling, delivering, and enhancing the quality of data. Data is coming from structured, unstructured, and semi-structured sources like web applications and databases on a real-time basis. Collating a variety of data can be complex and time-consuming. A Veritas study reveals that IT professionals spends two hours daily to identify relevant data (Helpnet, 2019). A lot of reputed enterprises lose money in managing operational efficiencies arising due to data quality concerns. According to a Veritas news release, a survey of around 1,500 IT decision makers and data managers across 15 countries reveals that the estimated loss for the organizations is close to $2million a year due to persistent data challenges (Veritas, 2019).

INHERENT ISSUES

Data of poor quality results in higher maintenance costs, incorrect results, lower customer satisfaction and retention, distorted campaign effectiveness metrics, loss of revenue, fraudulent transactions, indecisions, loss of time and productivity, inefficiency, operational delay, and increased operational costs (Vasudev, 2015). This kind of data can emanate in forms of missing values or non-numeric values, duplicate entries, inaccurate entries, format issues, outliers, mapping concerns, consistency issues, lag issues, irrelevance, and junk information. The data used for analysis and reporting should be accurate, complete, consistent, timely, and unique.

There have been inherent issues of data format consistency while integrating heterogeneous data. Integrating data from multiple sources is a real challenge. Among all issues, challenges related to dates, numbers, character sets and encoding, multiple languages and human errors are quite prominent ones (McCarthy, 2017). A large amount of data is generated through the sensor network these days in the field of scientific applications such as those in astronomy, earthquake science, gravitational-wave physics, military medicines, military surveys, and environmental studies. It becomes extremely critical to manage and aggregate the data at an appropriate level to reduce redundancy.

In this age of digital supremacy, it has become extremely critical to manage the data effectively to maintain productivity and a competitive edge. In a study conducted by Bourne (2019), it was found that there was a 16% drop in workforce efficiency due to the loss of two hours per employee each day in searching the data for analysis and reporting related to performance, revenue, profit, etc. (Forbes, 2017). The study estimated that their organization loses over two million dollars a year due to challenges faced with managing their data.

With the continuous increase in the number of cyber-attacks, it has become essential to secure the data safely. If the data falls into the hands of hackers or corrupt officials, there is a remarkably high chance of it being misused. Layered based architecture has been used by NASA earlier (Israel, Hooke, Freeman, and Rush, 2006).

VALIDATION CRITERIA

Data is shared with various stakeholders on a real-time basis; hence reliability is a key factor. It is important to keep a constant check that the data should be updated, appropriate, free from human biases, errors, malware, and bugs. With the evolution of data in form of design, content,

and technologies it is crucial to define a clear set of goals to be achieved. The validation techniques employed should justify the success of the analysis performed and determine if the data management system complies with the requirements and attains intended outputs. For instance, the most important criterion for a successful laboratory experiment is to identify the outliers within two days in the receipt of the specimen. Then, this criterion becomes the testing base when the system is developed. The development team must test both normal and exception paths to determine whether the criteria are met in each instance. Normal paths are the data types and formats that the system expects to process. Exception paths are unexpected data items or formats, such as missing values or non-numeric values.

REQUIREMENTS OF DIFFERENT ORGANIZATIONS

In the brand and marketing world, poor and incomplete data may lead to inaccurate outputs, impacts revenue sales, and profitability, increase in operation cost, results in customer churn, loss in Return of Investments (ROI) of marketing campaigns, lack of brand growth, and much more.

In *the Banking and Insurance sector*, the issues of inaccurate customer segmentation might lead to customer churn. Incorrect or old contact details of customers may lead to operation costs in scenarios where customers have defaulted. There are Know Your Customer (KYC) challenges when mobile numbers or date of birth does not match with Aadhaar information. Huge volume, data privacy, data security and compliance are also key issues faced by banks (Amol, 2016). Data siloes hinder the ability to create a customer 360-degree view thus losing out revenue and profit by offering incompetent products. Data quality breaks down occurs mainly because of frequent updates to the regulatory demands, regulatory inquisitions and resources needed to keep legacy architecture intact for reporting purposes (Ransome, 2017). Data is stored in different systems, hence data linking between two systems is a critical aspect, the absence of which leads to inaccuracy and wrong outputs. Heale (2014) suggests a seven-step data quality process. It is critical to have accurate and accessible information to verify information, trace transactions for anti-money laundering purposes. There is also a need for structural as well as semantic consistency. This will ensure that the data formats are consistent and there is no ambiguity in the information collected. So, validation at the source of data collection also becomes an important dimension of data quality.

The main challenge that the *manufacturing industries* are facing today is how to leverage and integrate their data. A huge amount of data is

generated at every stage of the product cycle. But the complexity lies in collating all the relevant information in a central system that can be accessible to the entire enterprise. As per studies, 80% of enterprise information is either unstructured or is held within business silos and generally inaccessible.

There have been major concerns for the quality and value of the data used for *clinical trials***.** Fraud, mishandling, intentional or unintentional noncompliance, and carelessness can lead to loss of money and cause catastrophic setbacks that demand reruns of clinical trials. In 2017, 60% of the warning letters issued by the Food and Drugs Administration (FDA) were the result of a lack of data integrity.

The Healthcare sector is experiencing the explosive growth of data – from 500 petabytes in 2013 to 25,000 petabytes by end of 2020 (Gandhi and Wang, 2020). It was observed that around 92% of the medical records are duplicate (McClellan, 2009). The industry is adopting big data techniques to improve the quality of healthcare delivery. But the big data processing approaches still encounter lag in giving real-time alerts and making accurate predictions. With the ongoing crisis due to COVID-19, the world has more data that the data that is recorded in form of reports. Every healthcare organization today is collating and analyzing data from various online & offline sources like hospitals, mortuaries, government websites, and portals to identify the virus spread, information on infected cases, fatality trends, and recoveries across countries. Proper patient identification has always been a major factor affecting patient safety causing a perilous and expensive issue for hospitals and health care centers (Bittle et al., 2007). The inconsistency of the data published on various portals has crippled the governments and aid organizations across the globe to fill in the gap that has developed between the resource needed and resources available to combat the COVID-19 pandemic.

A huge amount of data is generated by firms operating in the *retail sector* and is generally driven by multiple stakeholders, that are not updated regularly, and many times lead to data redundancy or inconsistencies. This also leads to missing opportunities to earn profits as much of the time and effort goes extracting, modifying, combining, validating, and further load the data for in-depth analysis and visualization. Poor data management leads to error in pricing, discrepancies across channels, shipment and delivery errors, lack of visibility into inventory, damaging legal action from customers, and regulators with serious cost implications. Conflicts in promotions across channels and inadequate customer knowledge are adding to customer dissatisfaction. Critical customer data is vulnerable to

threats and ransomware. Several consumer companies reported data breaches in the past (Green and Hanbury, 2019). Many of them were caused by flaws in payment systems either online or in stores. A report published by cybersecurity firm Shape Security showed that 80-90% of the people who log in to a retailer's e-commerce site are hackers using stolen data.

Telecom Sector is going through a transformation, the active mobile broadband subscriptions globally have increased from 3.3 billion to 7.7 billion in just 5 years. Growth in the devices has increased the challenge of handling unstructured data on the parameter of volume, velocity, and variety. According to a survey, 80% of today's data is unstructured data which makes retrieving and accessing the quality data complex and time-consuming. Besides, there are no standard guidelines for the retention and use of data as well as metadata. In recent years, Telecoms use the Internet of Things (IoT) for connecting everything to the Internet. It is the core technology behind smart homes, self-driving cars, smart utility meters, and smart cities. IoT products have made cyberattack permutations unpredictable.

In *Hospitality and Tourism* sector, with the rise in the digitalization, hospitality and tourism sector has been in the radar of frequent cyber-attack as most of the hotel and online travel company holds customer personal information like full names, address, passport number, and credit card details. This makes this industry vulnerable to data breaches. In 2019, there have been numerous data breaches in the hospitality industry, the latest being the "Choice Hotels data breach". This breach left 700,000 guest records vulnerable, disrupting customer trust, and potentially leading to a major drop in customer loyalty, impacting their revenue as well.

In recent months, a significant amount of hike has been noticed in the issues related to data governance and compliance driving in new data management and data privacy regulations such as the General Data Protection Regulation (GDPR), which places tough new standards on how personal data is held. Due to digitalization, an enormous amount of private information, transitional data, and sensitive data concerning individuals or organizations are profusely collected by a great number of applications and the Internet of Things which may pose a great threat to individuals or organizations. According to a new Identity Theft Resource Center report (Helpnet, 2019), the number of U.S. data breaches tracked in 2019 increased 17 percent from the total number of breaches reported in 2018.

The evolving world where data is the new oil, there is a critical need for a solution that is capable to process, refine, transform, and aggregate the data in a very methodical and timely fashion. Here we suggest a

perspective on some approaches that will help validate data quickly and accurately, thus providing the platform for any kind of analysis.

SUGGESTED APPROACH

In the era of evolution, data has become the new source. There is a critical need for a solution that is capable to process, refine, transform, and aggregate the data in a very methodical and timely fashion. Here we suggest a perspective on some approaches that will help validate data quickly and accurately, thus providing the platform for any kind of analysis.

Smart Information Assimilation and Aggregation System (SIAAS)

SMART INFORMATION ASSIMILATION AND AGGREGATION SYSTEM (SIAAS) is layer-based architecture to handle most of the data issues mentioned previously. We talk about five key layers:

1. Dumping Layer
2. Refinery Layer
3. Aggregation Layer
4. Analytical Layer
5. Application Layer

1. Dumping Layer: This layer is the dumping ground of data from various data sources. The dumpling layer acts as a warehouse where data is accumulated in varied forms like Excel, CSV, structured and unstructured pdf, image, and videos. This layer has algorithms that connect to the client databases and fetches all relevant data with no specific filtering exercise. It creates a matrix of data tables and fields. This layer also can create a Level-1 knowledge dictionary capturing information about the source, column names, size, rows, columns, and few other attributes. It also comprises of few powerful algorithms which provide an intuitive Exploratory Data Analysis (EDA) of the entire data.

2. Refinery Layer: This layer runs a series of validation on the initial set of data captured in the Dumping layer. This layer filters relevant information, checks file formats, find outliers, inconsistencies, mapping issues, missing information, and a lot more. This layer encompasses a bunch of refinery algorithms that can study the data characteristics and inherent challenges. It also can carry out multiple automated simulations to scan the data fields and publishes diagnosis reports. These reports can be studied to understand the magnitude of data issues and explore the data. Once the diagnosis reports are published then another set of refinery algorithm will

be triggered to refine and code the data. The concept of Dynamic step validation can be implemented in this layer that will help to validate data as it enters the refinery engine.

3. Aggregation Layer: This layer is particularly important and executes a set of mapping and aggregation algorithms to transform and summarize the data in an appropriate way. The key concept here is information once captured need not be repeated for a future set of aggregation, until and unless there is some change in the data point. This enables smart aggregation of data such that it could be utilized for future trends. An aggregated data mart created in this layer comprises summarized tables in line with the views required for the final outputs. Only the latest summarization will be added to the already existing data-marts. This concept is incredibly useful in cases where the past data does not change dynamically.

4. Analytical Layer: This acts as the platform for any analysis and reporting needs. Analysts and modelers will require to connect to this layer to carry out their final analysis. This layer acts as a highly Intelligent Centralized Master Schema and caters to all critical business and project needs. This will enable the user to perform all the reporting and analysis in a noticeably short span. This will also ensure the quality and accuracy of the schema. This will act as a Linkage Based Schema, where the centralized master table will be linked to all the sub-tables in the schema. This will have an Integrated Algorithm that can scale up and eases out change management along with an integrated customer view. This layer is updated on the regular frequency and timely intervals but in an accurate way. When new data arrives or data is refreshed, there is a minimal movement of data values, and appending is done in a very smart way with blocks of data. This uses the block appending concept as it works on the blocks of data rather than playing with the entire data columns. This implies it targets the block where necessary data management operations are being carried out. Each block of data is represented by a unique ID so that it can be picked with ease. The resulting changes could be studied through a few advanced simulations and mathematical predictions of the resulting tables. This also has an automated validation engine that runs and keeps checking the sanctity of data after every update. Thus, this layer will be a one-stop solution for all the critical reporting and analysis requirements and is extremely adaptable and fast-paced.

5. Application Layer: This is a front layer and will get connected to the Analytical schema and portrays all relevant analysis and reporting. This can be any available application layer like Spotfire, Tableau, or ClickView.

APPLICATIONS IN DIFFERENT SECTORS

SIAAS can be implemented in any domain. Here are couple of examples:

1. Banking: Banking executive offering credit, wealth management, investment banking products to a high net worth individual with limited product brochures and customer background. SIAAS can assist by providing recent credit history, the value proposition to this customer, and other data-driven insights by extracting a summary from assorted different functions databases and enterprise data warehouse in close to real-time.

2. Hospitals: A patient visiting a large multi-specialty hospital regularly faces a lot of challenges. He must explain to various specialist doctors the history and various diagnoses whenever he consults a different specialist or visits after a while (must carry a good number of reports and documents). Various diagnosis reports are not available to doctors in real-time and from a single system. The patient must make payment for different facilities like diagnosis, medicine, consultation etc. separately. SIAAS can help provide an integrated patient view to the doctor in quick time along with past history for a long time: diagnosis reports, medicines prescribed, consultation from various specialists, and others, which will enable with analytical insights to doctor to provide a more effective prescription.

Special Issues in Covid-19 Situation

Coronavirus has become a big problem and many countries across the globe are struggling to control this pandemic. A Bloomberg City Lab study points out that that data used for making decisions is incomplete and not perfect for reporting purpose (Patino, 2020). A huge amount of data is being published like details of people affected, negative cases, patients hospitalized, ventilator shortfalls, beds occupancy in hospital, deaths, morbidity ratios, forecasts, and much more. The data is available in various formats – excel sheets, pdf files, government websites, news portals, social platforms, and various web applications. Analysis and reporting of these numbers enable epidemiologists, health officials, government departments to make informed and timely decisions to fight this pandemic. The authenticity of the data is a big question as inconsistencies have been identified in the data which is not matching at the ground level (Shinkman, 2020). A recent news article stated that a Lancet paper has been retracted after investigation found inconsistencies in the data (The Guardian, 2020). Here are typical inconsistencies that have been observed:

1. Number mismatching
2. Lag in the data collection

3. Representation of data
4. Inconsistency in death data and the reason for death
5. Inconsistency in bed numbers and ventilators available
6. Varied sources
7. Testing speed
8. Inconsistencies in tagging red, orange, and green zones in different states
9. Definition and calculation of the reported cases
10. The veracity of the testing kits
11. Changing case definitions
12. False negatives

These shortcomings have been published in some reports where it has been mentioned that there are accuracy issues in the data reported by a few countries. Death data that is being reported has its limits as it has lag issues. The definition and calculations in reported cases vary from country to country as in some cases number of asymptotic cases has not been reported. The inability to track the asymptotic cases may lead to a wrong representation of the data on the infected population. This situation gets aggravated by the concealment of data as well. The speed of testing is also a vital factor in gaining an understanding of the actual spread of the infection. If the testing ratio is less than the numbers reported will always be a fraction of what is reality. There is also an impact of changing case definitions on the number of reported cases as outlined by a Chinese study in the reported cases of mainland China (Tsang *et al.,* 2020).

There is a lot at stake in the analysis and reporting of COVID data, hence the data credibility is a key concern. False or incomplete data will lead to

1. Incorrect decision making
2. Inability to take adequate measures
3. Inaccurate and incomplete planning
4. Mismanagement in hospitals
5. Incorrect public messaging
6. Resistance to policy formation

How SIAAS can help

SIAAS will assist in building Data Diagnosis and Validation platform for COVID data that will smartly identify and eliminate the inconsistencies.

It will further aggregate the data that will be reported in quick time intervals.

With this approach first, all the key sources of information will be identified, and raw data will be pulled in the "dumping layer". This will contain data and information on COVID cases for a specific period. This data may be in the form of data files, articles, or reports. The dumping layer will also have a Master summary file which will include the source information, type of data (structured or unstructured), format, period, and more format related information. The folder structure of this dumping layer will be designed in such a way that the data in folders and subfolders follow a similar pattern, type, and period. Every time the data are refreshed, and new data are pulled these will be sent to a defined folder structure. This layer will act as a data warehouse layer and incoming files and documents will be positioned according to the pre-defined structures methodically.

Once all the data for a specific period is accumulated, the data from the dumping layer will be then shipped to the Refinery layer where most of the data refinement will take place. Initially, a series of validation algorithms will scan the data and will perform a thorough health check-up of the new set of information that has arrived in the form of raw excel files, pdf docs, articles, images, videos, etc. Once the scanning is done the diagnosis report published will outline information on the inconsistencies mentioned above. Based on the inconsistencies and discrepancies identified, another set of algorithms will run and clean up the irrelevant data and narrow down on a refined set of data that will be structured, filtered, and comparatively cleaner. This layer will focus on each unstructured file (articles, pdf, images, etc.), filter irrelevant stuff, transform it into structured data with relevant information. The next layer of cleaning will take into consideration clean it up taking into consideration the information collected in other files. Basically, another set of algorithms shall run and identify duplicity of information across all available files and then shall do some further refinement to ensure each of those files are structured and have meaningful information. In some scenarios, there might be a necessity to do some meaningful imputations to create a more realistic transformation of raw files. In summary, this layer will perform:

1. Data Diagnosis operation
2. Filtering operations
3. Transformation operations
4. Refinery operations
5. Mapping operations

In the next layer, the refined data will be aggregated and summarized in a meaningful way. For example, for a specific period, the COVID data can be summarized at the city, state, country level and then it will be transformed in shape so that it is ready to be appended in the Final Analytical Layer.

The final analytical layer will have a schema-based architecture so that the analysis can be carried out at different levels (in this case City, State, Country-level) in a noticeably short period. All-important dashboards, reports, and analysis can be quickly carried out through this layer. Policymakers, State health officials, Researchers, Scientists, Doctors, Lawmakers, Police officials who relate to this pandemic in one way or the other can make informed decisions in a quick time. This will help in tackling the problem in a much more effective manner and can plan more strategically.

CONCLUSION

Though SIAAS has a lot of benefits in terms of managing the data and structuring data extremely methodically, there may be some associated challenges in the implementation of a SIAAS system. Although the layer-based approach would be standard across domains, there would be a need to customize it, keeping in the end objectives in mind. In some situations, it may be a bit complex and cumbersome to handle some of the raw unstructured, or structured data. The layers' building will require quite a lot of effort in the initial phase, but the process will become seamless once the pillars are built. Refreshing the data in specific time intervals will have to be monitored closely until the process is stable. Continuous learning will be the key to ensure that the SIAAS system is useful and trustworthy. But once the SIAAS architecture is built methodically, it will have a lot of advantages. This will reduce operational inefficiencies significantly and will help the decision-makers with some real quick summary statistics to make informed and fruitful decisions. Implementing this system in real-world problems will require some additional research and efforts in building intelligent algorithms to make this approach a long-term success.

References

1. Shinkman, P. (2020), The Flaws in Coronavirus Case Reporting Data.
2. Helpnet Security Article (2019): Data management challenges are having a severe impact on profitability (https://www.helpnetsecurity.com/2019/03/13/data-management-challenges/).
3. Vasudev, M. (2015), What is Bad Data and its Side-Effects.

4. Bourne, V. (2019) Veritas Study Report.
5. Forbes Press Releases (2017), Poor-Quality Data Imposes Costs and Risks on Businesses https://www.forbes.com/sites/forbespr/2017/05/31/poor-quality-data-imposes-costs-and-risks-on-businesses-says-new-forbes-insights-report/#4951021d452b.
6. Amol, S. K. (2016), The Importance of Data Quality Management and Data Cleansing for Banks (https://www.nelito.com/blog/importance-data-quality-management-data-cleansing-banks.html).
7. Heale, B. (2014), Data Quality is the Biggest Challenge.
8. Ransome, O. (2017), Data quality issues hurt banks. Stuff worth knowing from The Bankers' Plumber (https://3cadvisory.com/data-quality-issues-really-hurt-banks-stuff-worth-knowing-from-the-bankers-plumber/).
9. McCarthy, K. (2017), The top 5 most common data quality issues (https://www.edq.com/blog/the-top-5-most-common-data-quality-issues/).
10. Patino, M. (2020): Bloomberg City Lab. Coronavirus Data in the U.S. Is Terrible, and Here's Why (https://www.bloomberg.com/news/articles/2020-05-01/why-coronavirus-reporting-data-is-so-bad).
11. The Guardian (2020): Covid-19: Lancet retracts paper that halted hydroxychloroquine trials (https://www.theguardian.com/world/2020/jun/04/covid-19-lancet-retracts-paper-that-halted-hydroxychloroquine-trials).
12. Veritas New Release (2019): Data Management Challenges Cost Organizations $2 Million a Year (https://www.veritas.com/news-releases/2019-03-12-data-management-challenges-cost-millions-a-year-reveals-veritas-research).
13. Gandhi, M., Wang, T. (2020)The Future of Personalized Healthcare: Predictive Analytics (https://rockhealth.com/reports/predictive-analytics/).
14. Bittle M.J., Charache P., Wassilchalk D.M. (2007) Registration-associated patient misidentification in an academic medical center: causes and corrections. Joint Commission Journal on Quality and Patient Safety/Joint Commission Resources. 2007;33:25–33 (https://pubmed.ncbi.nlm.nih.gov/17283939/).
15. McClellan, M. (2009) Duplicate Medical Records: A Survey of Twin Cities Healthcare Organizations (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2815491/).
16. Green, D., Hanbury M. (2019). Business Insider: If you bought anything from these 10 companies in the last year, your data may have been stolen. (https://au.finance.yahoo.com/news/bought-anything-10-companies-last-075945374.html).
17. Helpnet Security Article (2019): 2019 saw more data breaches, fewer sensitive records exposed (https://www.helpnetsecurity.com/2020/01/29/more-breaches-fewer-records-exposed/).
18. Israel, D.J., Hooke, A.J., Freeman, K. and Rush, J.J. (2006). The NASA Space Communications Data Networking Architecture. NAS Technical Report; NAS-06-014.
19. T K Tsang, P Wu, Y Lin BM, E H Y Lau, G M Leung, B J Cowling (2020). Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study. Lancet Public Health 2020; 5: e289–96.