Application of Google Cloud and Google Apps in Data Structuring

Shruti B, Sachin H, Naina J, Luvesh T, Sunita Suralkar, Richard Joseph

Department of Computer Engineering
*Vivekanand Education Society Institute of Technology **Mumbai, India***

Abstract : *In today's world one of the most integral part is the data. There is lots of scattered (unstructured) data around that is available. Enterprises deal with lots of data and there are requirements that need to be imposed on such data in terms of access, control and maintenance. In order to achieve these functionalities it is very important to structure the data. Structuring of data is part of data governance and this can be achieved by various technologies. Since any enterprise has large amount of data we need cloud for storing this data. This paper presents the way of using Google cloud services and google apps in ordered to obtain uniform good quality data that can be further used for analysis purpose.*
Keywords: *Data governance, Data structuring, Google apps, Google Cloud storage, MVC Architecture*

I. Introduction

Nowadays the most important asset of any organization is the data. Every organization needs to insert, store and update huge amount of data. Thus cloud can be effectively used in these purposes. Google provides reliable and scalable cloud storage services. There are many advantages in using cloud storage, like files stored in the cloud can be accessed at any time from any place as long as you have Internet access. Also Cloud storage can provide the benefits of greater accessibility and reliability; rapid deployment; strong protection for data backup, archival and disaster recovery purposes; and lower overall storage costs as a result of not having to purchase, manage and maintain expensive hardware. [1]

The enterprises needs to maintain the information about their Employees, business details, etc. For example if we consider an Education Institute, the data which needs to be stored will be: Student details, Faculty details, Departments, courses provided etc. The entries in these fields keep increasing and also the data needs to be updated periodically. Thus cloud can be used. Google provides multiple services and applications for collection of data and storing it on the cloud. Data can be of any form like documents, pdf, images, videos, data extracted from application form, data in excel sheets and all of this data is in scattered format so it is important to govern the data. If any member of the organization wants to access or view any particular data it becomes a tedious task because it is to be searched in a big ocean of unstructured data. Thus it is very important to start with structuring the data and thus govern it according to the organization's/institution's requirement. The main objective of this paper is to introduce an efficient way of storing all important data in a sorted order on the cloud and thus securing the data and simultaneously making this data available to everyone. We can start with using rule based system for entering the data:

Rule-based data entry method:

A rule can be defined as a statement that tells what is or is not allowed in a particular process, situation etc. [9]. Thus rules can be used to restrict which valid data can be inserted in which domain or relevant part of the structure.

Rules can be either pre-defined or dynamically defined depending on the data which is to be entered. Some examples of the pre-defined rules are:

1. Gender value should be either M (male) or F (female) or T (transgender) and not anything else
2. First-name or Last-name should start with capital letter.
3. Email address should be of valid type like "abcd@email.com".

The rules above are generic rules that is it can be used in any organization since name, email-id, gender are common data stored everywhere. On the other hand there are few organization specific data like in a finance company's shares, investments, bank details, loan amount etc. are important entities while in an education institution student details, courses, departments, faculty details etc. are important entities. So such data keep varying and thus rules cannot be predefined and thus there is requirement for dynamic definition of rules.

Some examples of dynamically defined rules can be as follows:

1. In a Finance organization if we want to enter the information of only those companies whose share price is greater than Rs50,00,000 then the rule will be as follows:Value of company share)>=5000000.
2. In a Medical organization if we need to store the information of medicines containing only vitamin A then the rule will be as follows: ((Percentage of vitamin A>=1) && Percentage of other vitamins=0))

3. In an E-commerce organization if we need to store information of customer who always have bill less than Rs.500.(Customer- bill < 500)

Dynamic rules are completely dependent on the scenario of the system.

Implementing rules using Google App [3]

Google Sheets is a very convenient option for storing data in a systematic manner.Google sheets is also called as spreadsheets. In google sheets we can apply queries on the data and these queries act like rules.

Consider in an Education Institute we have a spreadsheet storing details about "courses conducted" and we need to extract information year wise so we can fire a query and accordingly only that relevant entries will be displayed and it can stored in a separate spreadsheet. Thus if the admin wants to know only about the courses conducted in the year "2016" he can directly access the sub-spreadsheet which is formed by querying the master spreadsheet. Consider the following example of a query:

query('Form Responses 1'!1:100002, "Select B,C,D,E,F,G,H,I Where G = 2016",1)

Where, Form Responses means the data is collected from google forms which are also one of the important google apps. B,C,D,E,F,G,H,I stands for the columns which need to be shown in the new sub-spreadsheets. "G=2016" is the main part which causes the filtering.

This same filtering can also be done using Google add-ons. Add-ons run inside Google Sheets, Docs, and Forms, but they're built developers. [4] Add-ons make the task easier. For example there is are many add-ons for generating queries for us and they have interactive UI so we can easily communicate and request for what we want.

Consider the Fig. 1, it shows the UI of "EZ Query" Add-on. Using Add-ons makes our work easier. Also we can form complex queries by using multiple options given in this add-on.
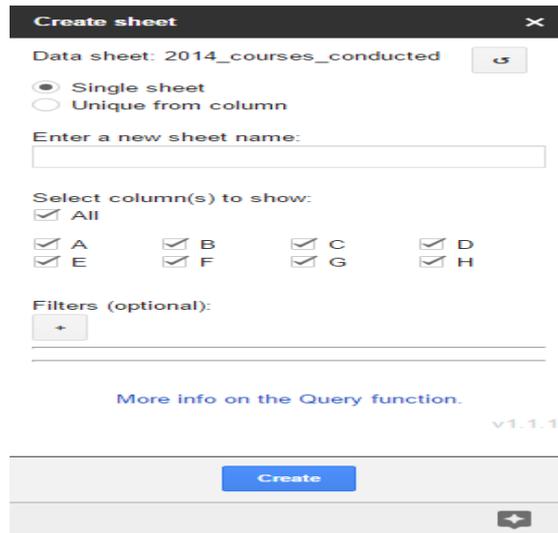
Fig. 1.1

From the figure Fig. 2 we see that there are various filters available and thus we can select any one and query will be generated and new spread sheet will be generated.
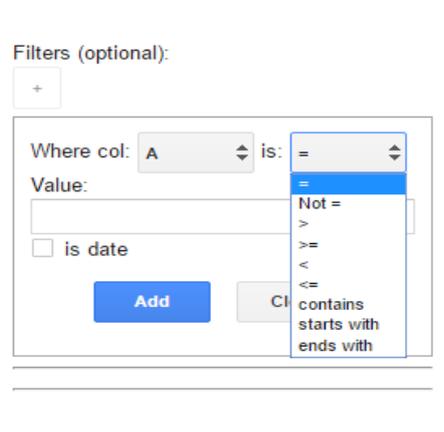
Fig. 1.2
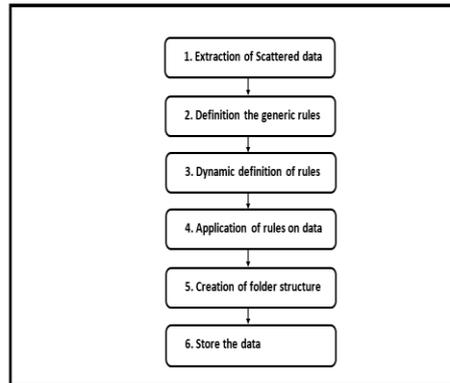
Fig. 1.3

Process:

After the above steps are followed we can easily access any data by just accessing the correct folder. The data is well sorted and organized and thus this data can now be used for data analysis, data mining or can be displayed on the website of the organization.

II. Methodology

As explained above the whole procedure is divided into six steps. We will consider an Education Institute as an organization example in order to understand the whole process.

Step 1- Extraction of Scattered data [2]

Extraction is the initial step where all the data which is available should be gathered. This can be done using google forms. Google forms allow us to generate customize forms. There is a drag and drop method for building forms. Google forms can be shared with other people and the responses gets automatically recorded in a spreadsheets. For example nowadays in an education institute the academic details of all the students, can be collected by sharing forms with the students and they will enter their details and submit the form. Thus google forms is a convenient option for collection of data.

Step 2:

Forming the rules is the most important part of the process since the data is classified using these rules. Thus the rules should be accurate enough to filter the data. For example in the above example if we want a rule that can filter eligible and non-eligible students for a particular exam, then it can be something like this:

(Marks obtained in 10th grade should be greater than 60%) && (Marks obtained in 12th grade should be greater than 65%). Thus such rules are really handy in classifying the data efficiently.

Step 3:

We can also dynamically define rules and again filter the data as per the requirement.

Step 4:

Applying the rule will segregate the data into different categories and thus can be easily accessed.

Step 5:

The structure of organization is developed depending upon the hierarchical structure in the organization. For example in the above example the structure can be as follows:
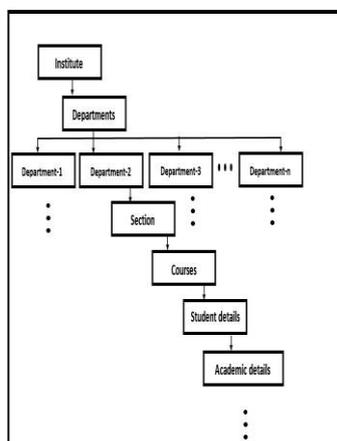
Fig. 2.1

This Hierarchical structure can be developed using Google Drive. In Google Drive we have an option of developing folder structure and also access rights can be given to authorized people. We can store any data, files etc. in these folders.
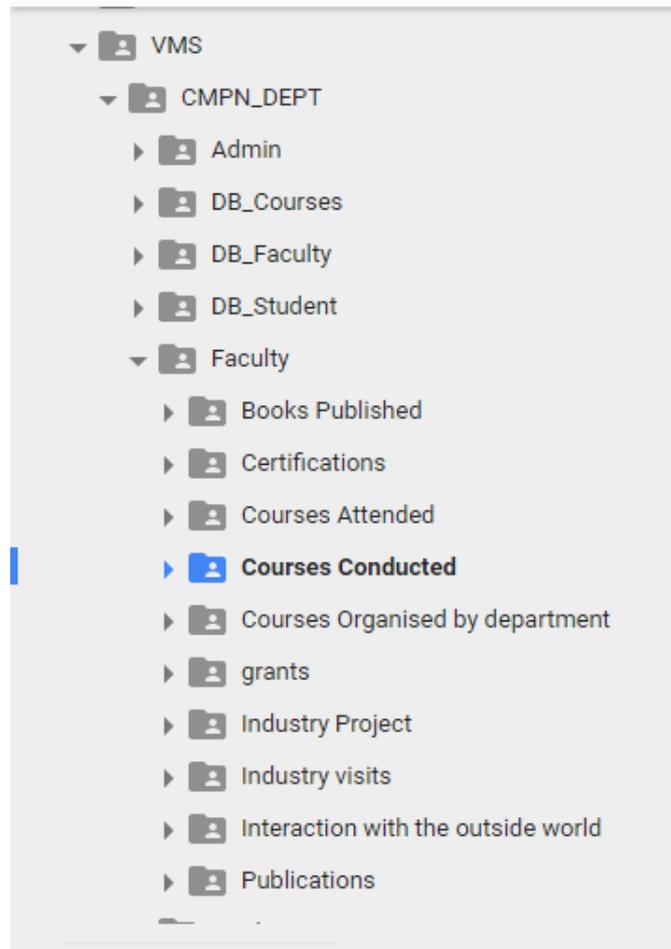
Fig 2.2
Fig 2.2 shows a hierarchical folder structure generated in Google Drive.

Step 6:
Since the skeleton is ready now the classified data can be stored in the relevant folder. And thus data can be easily retrieved.

III. Implementation

The implementation of this project depends upon where the data is stored and till what level the data is scattered. If the data is stored in a list format then it will be easy to extract but if it is stored in a random way where the previous data is not linked with the next data then it will be tedious and difficult task to extract the data. Since the above methodology is explained using an example of education institute, implementation will be explained assuming that all the data is stored on the google cloud storage but is in unstructured format. So the task is to extract the data from cloud storage using necessary scripting language and then follow the other steps specified in the methodology column.

The overall approach this paper want to propose towards this scenario is a "model-view-controller" (MVC) architecture. The block diagram of this architecture is as given in Fig 4.

MVC architecture [5]

In Fig 4 there are four major Blocks namely:
- User
- View
- Controller
- Model

User:

Since the architecture is for an education institute the users of the system can be either of the following:
• Students
• Parents
• Faculty Members.

View:
        The data can be viewed in any format and also on a website developed using HTML and CSS languages. Since google provides an option of drag and drop method for generating websites thus google sites [7] are more convenient for such projects where the data is more important than prototypes and look of the project.
Also the data can be viewed in google sheets or PDF format accordingly.

Controller:
        Controller is the module or block where all the code or functions are stored. The Controller consist of functions responsible for extracting data from the cloud storage, Defining pre-defined/generic rules, Dynamically defining the rules, Filtering the data according to the rules and then again storing the data but in a structured manner.Since we have taken an example of google cloud, google has provided a scripting language in order to code such functions that is Google AppScripts [6]. This Scripting language can do all the above specified task and many more.

Model:
The model is basically the big ocean of data from where we are accessing, storing and manipulating the data. The Model can be a cloud storage [8], Database or anything.
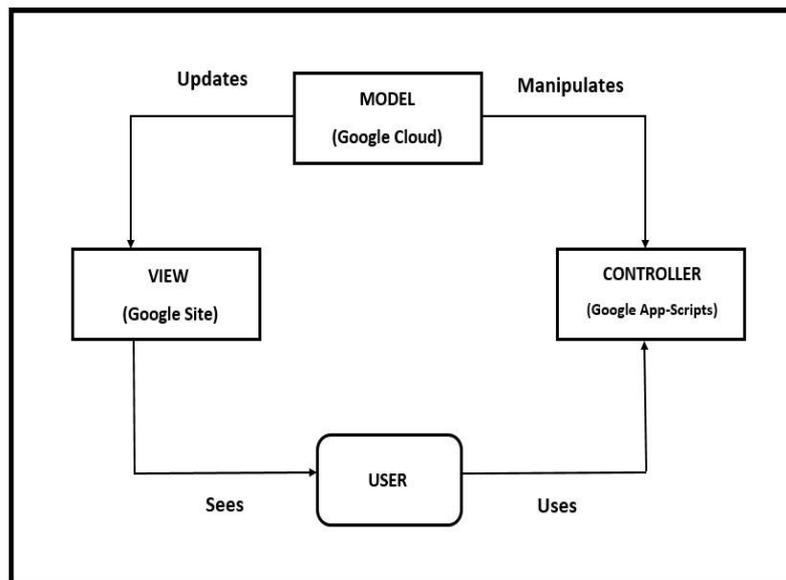
Fig. 3.1

        Thus from Fig. 3.1 we see that the user uses the controller by using the view (Websites or an App) and when the user enters any data via forms, the data is filter according to the rules which are defined by the admin using the scripting language, and then the filtered, good quality data is stored on the cloud in the relevant folder.

IV. Conclusion
        Unstructured data is the key asset in this whole scenario, because it is the input to the process and accordingly the operations are performed. The main objective of this paper was to put a light on the current data related problems and how it can be resolved using the latest Technologies like the google applications and google cloud storage and a good level of structured format for the organization. Also the rule-based system is an effective method since it will classify the data and thus searching any data will be easier and such well classified and good quality data can be used for data analysis and other mining algorithms. Also such data can be easily used for report generation in many scenarios. Google provides many add-ons for generating reports from spreadsheet data and these reports can be in word format or pdf format. Thus it is very important to manage, store and govern the data because nowadays all the organization are data centric and without data no operations are possible.

References

[1].   http://www.webopedia.com/TERM/C/cloud_storage.html
[2].   MOHAMMAD FIKRY ABDULLAH and DR. KAMSURIAH AHMAD, "The Mapping Process of Unstructured Data to Structured Data" 3rd International Conference on Research and Innovation in Information Systems – 2013.
[3].   TATYANA SVIRIDOVA, LARISA SVIRIDOVA AND BOGDAN TYMOSHENKO "GOOGLE APPS AS SOLUTION OF COMMUNICATION ISSUES IN EDUCATIONAL PROCESS" MEMSTECH'2011, 11-14 May 2011, Polyana-Svalyava (Zakarpattya), UKRAINE
[4].   https://developers.google.com/apps-script/add-ons/
[5].   https://en.wikipedia.org/wiki/Model%E2%80%93view%E2%80%93controller
[6].   https://developers.google.com/apps-script/
[7].   https://www.google.com/sites/help/intl/en_GB/overview.html
[8].   https://cloud.google.com/
[9].   https://www.merriam-webster.com/dictionary/rule