AidData.org: A Donation Analysis

Harini Musunuru, Patrick Kinnicutt, Roger Lee
Department of Computer Science
Central Michigan University, Mount Pleasant, MI, USA
{musun1h@cmich.edu, kinni1p@cmich.edu, lee1ry@cmich.edu}

Abstract --Every year, in order to improve the lives of people the government and people are spending a lot of money in developing countries. For better country future, the government can take better decisions if they have relevant and accessible data like citizens who are willing to donate and invest few dollars to increase the developments. Aiddata.org is a kind of website which provides data to visualize, and analyse data on $40 trillion in financing for development. Aid Data is a research lab which maintains development finance related activities that help to improve profitable outcomes by providing development finance data to government and people at their fingertips. Aid Data dataset includes over 1 million aid activities funded by more than 90 donors from the 1700s to present. The analysis of this dataset mainly help us to understand the process about the donation analysis between the countries and many different organizations which undertook this donation activity.

Key words – Hadoop, SAS Enterprise Guide, OLAP cube, SQL Server

I. INTRODUCTION

Each year, billions of dollars are spent to improve the lives of citizens in developing countries. With accessible and relevant data at their fingertips, governments can make better decisions to plan for their country's future, citizens can hold their leaders to account for providing public goods, and donors can invest aid dollars to maximize development results. Aiddata.org provides data to visualize, and analyse data on $40 trillion in financing for development.

Aid Data [2] is a research and innovation lab that seeks to improve development outcomes by making development finance data more accessible and actionable. Aid Data dataset includes over 1 million aid activities funded by more than 90 donors from the 1700s to present. The analysis of this dataset helped us to understand the business process about the donations between the countries and the organizations which undertook this activity.

The analysis of this dataset mainly help us to understand the process about the donation analysis between the countries and many different organizations which undertook this donation activity. We have implemented a Hadoop data mart in order to analyse aspects like Number of Donations a country or a sector made in a particular year and the results are displayed using Tableau tool.

The paper includes general background information regarding AidData.org donation analysis and the results that are obtained by using few tools like Tableau a Hadoop tool used for analysing big data. Section 3 in this paper describes the methods that are involved in doing analytical comparisons done for donation analysis. It also includes the data mart implementation and data mart population. Section 4 in this paper presents analytical results using Hadoop [4], SAS enterprise guide [7], OLAP cube [6] and SQL server [8]. Finally Section 5 concludes the paper.

II. BACKGROUND

About Aid Data:

Due to rise and rapid growth of Social and Corporate Responsibility towards society, Aid Data [1], an online website have come up with a collaborative initiative providing products and services that promote the dissemination, analysis, and understanding of development finance information. The Aid Data website provides access to development finance activity records from most official aid donors. The Aid Data portal provides access to development finance activities from 1700 to the present from more than 95 donor agencies.

Using the Donation Database Data from AidData.org we can successfully analyse the Business Process data like Donations made between countries.

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.119

Data analysis includes both Donation analysis and Casual Analysis. The analysis starts with Data Profiling [5], identifying the analytical themes and also understanding the business story about the data. We have implemented a Hadoop [4] data mart in order to analyse aspects like Number of Donations a country or a sector made in a particular year and the results are displayed using Tableau tool [9].

III. METHODS

A. Characteristics of Donations

Now with the help of Aid Data which has the characteristics of Donations like

* Donor

* Disbursement Amount

* Financial Agency

* Sector

We aim on extracting useful information. Basically, we intend to convert raw data into suitable format and push them to Data warehouse, using which we can perform various analysis for better understanding of the consumer and provide them with the best service possible. We have raw data of Aid data starting from 1700.The aid data basically gives us information of the donations made across the countries in a particular sector and details about the financial agency that was involved. The dataset will help governments to make better and quick decisions to plan for their own country's future benefit and citizens can hold their leaders to account for providing public goods and many other purposes, and donors who can donate more can invest on aid dollars to maximize development results. Once the data has been re-structured and significant patterns have been found, we can specifically analyze which country has donated the maximum in a particular sector.

B. Dataset

The data set used in here is the Aid research provided by the Aid Data Organization, which gives us the information of the Aid provided to different recipient countries by different donation countries and agencies. Different donation transactions, the amount donated for specific purposes are mentioned.

The data set used in here is the Aid research provided by the Aid Data Organization, which gives us the information of the Aid provided to different recipient countries by different donation countries and agencies. Different donation transactions, the amount donated for specific purposes are mentioned.

C .Data Quality

Data Quality is the most important aspect of any data mining project. In the data set available from the Aid Data, we measure the quality of the data based on these controls:

* Severity of inconsistency

* Incompleteness

* Accuracy

* Missing data

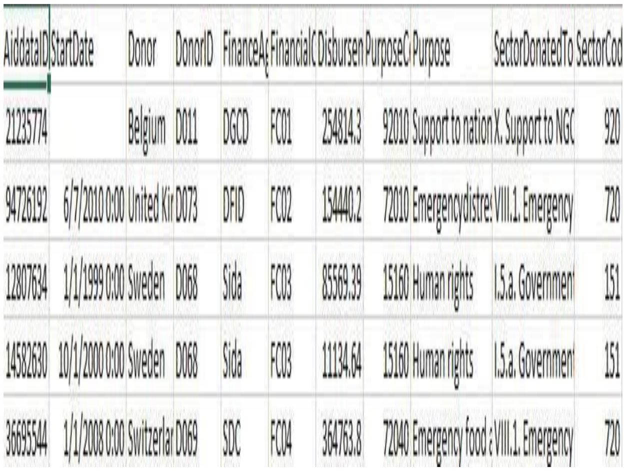
If there are too many missing values or inconsistency then we either remove that particular field or try to give our own values so we don't deviate much from the desired results. The usage of incorrect data could crucially impact output. Thus, establishing Data Quality Control process provides the protection of usage of data control and establishes safe information usage.

Data cleansing is the process of cleaning the data by filling up the unknown value and correcting (or removing) corrupt or inaccurate records from a record set. After cleansing, a data set will be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores.

Changes made to the Dataset:

We removed many unwanted columns from the above Dataset. We have generated unique IDs for the data in the Column Donor and Finance Agency using excel functions for better analysis of our data. We used excel function vlookup to find the corresponding Donor and assign the corresponding unique ID to each Donor. Similar process is followed for Finance Agency. We faced a lot of issues while doing this as the data wasn't well formatted but after a lot of manipulations and customized formulas, we were able to get the below result in Excel.

After performing the ETL tasks shows the below dimensional model we followed for loading the data into the SQL Server for developing the OLAP Cube for Data Analysis from the csv file(flat file).



AiddataID	StartDate	Donor	DonorID	FinanceAg	FinancialC	Disbursen	PurposeC	Purpose	SectorDonatedTo	SectorCod
21235774		Belgium	D011	DGCD	FC01	254814.3	92010	Support to nation	X. Support to NGC	920
94726192	6/7/2010 0:00	United Kir	D073	DFID	FC02	154440.2	72010	Emergencydistre:	VIII.1. Emergency	720
12807634	1/1/1999 0:00	Sweden	D068	Sida	FC03	85569.39	15160	Human rights	I.5.a. Governmen	151
14582630	10/1/2000 0:00	Sweden	D068	Sida	FC03	11134.64	15160	Human rights	I.5.a. Governmen	151
36695544	1/1/2008 0:00	Switzerlar	D069	SDC	FC04	364763.8	72040	Emergency food a	VIII.1. Emergency	720

Figure 1: Aid Data after Basic Removal of Unwanted Data

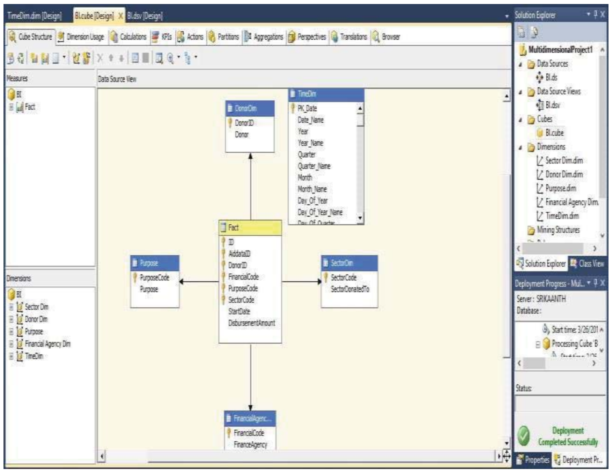
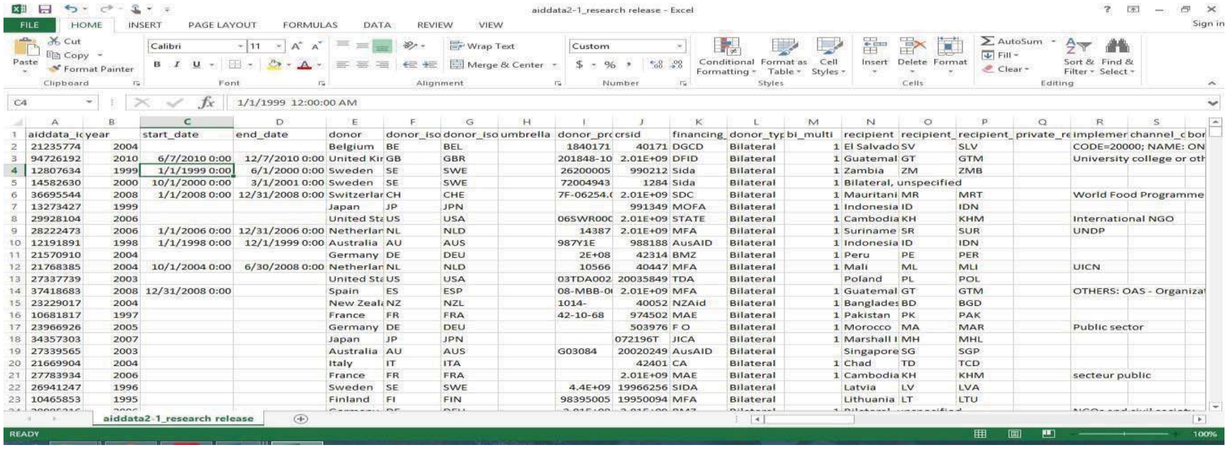


Figure 2: Dimensional Model

D. DataMart Implementation

Raw Data:

The below shown is the raw data that we used for donation analysis which was taken from AidData.org



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	aiddata_ic	year	start_date	end_date	donor	donor_iso	donor_iso	umbrella	donor_prc	crsid	financing_	donor_typ	bi_multi	recipient	recipient_	recipient_	private_re	implemer	channel_c	bor
2	21235774	2004			Belgium	BE	BEL		1840171	40171	DGCD	Bilateral	1	El Salvado	SV	SLV		CODE=20000; NAME: ON		
3	94726192	2010	6/7/2010 0:00	12/7/2010 0:00	United Kir	GB	GBR		201848-10	2.01E+09	DFID	Bilateral	1	Guatemal	GT	GTM		University college or oth		
4	12807634	1999	1/1/1999 0:00	6/1/2000 0:00	Sweden	SE	SWE		26200005	990212	Sida	Bilateral	1	Zambia	ZM	ZMB				
5	14582630	2000	10/1/2000 0:00	3/1/2001 0:00	Sweden	SE	SWE		72004943	1284	Sida	Bilateral	1	Bilateral, unspecified						
6	36695544	2008	1/1/2008 0:00	12/31/2008 0:00	Switzerlar	CH	CHE		7F-06254.(	2.01E+09	SDC	Bilateral	1	Mauritani	MR	MRT		World Food Programme		
7	13273427	1999			Japan	JP	JPN			991349	MOFA	Bilateral	1	Indonesia	ID	IDN				
8	29928104	2006			United Sta	US	USA		06SWR00C	2.01E+09	STATE	Bilateral	1	Cambodia	KH	KHM		International NGO		
9	28222473	2006	1/1/2006 0:00	12/31/2006 0:00	Netherlar	NL	NLD		14387	2.01E+09	MFA	Bilateral	1	Suriname	SR	SUR		UNDP		
10	12191891	1998	1/1/1998 0:00	12/1/1999 0:00	Australia	AU	AUS		987Y1E	988188	AusAID	Bilateral	1	Indonesia	ID	IDN				
11	21570910	2004			Germany	DE	DEU		2E+08	42314	BMZ	Bilateral	1	Peru	PE	PER				
12	21768385	2004	10/1/2004 0:00	6/30/2008 0:00	Netherlar	NL	NLD		10566	40447	MFA	Bilateral	1	Mali	ML	MLI		UICN		
13	27337739	2003			United Sta	US	USA		03TDA002	20035849	TDA	Bilateral		Poland	PL	POL				
14	37418683	2008	12/31/2008 0:00		Spain	ES	ESP		08-MBB-0(	2.01E+09	MFA	Bilateral	1	Guatemal	GT	GTM		OTHERS: OAS - Organiza		
15	23229017	2004			New Zeal:	NZ	NZL		1014-	40052	NZAid	Bilateral	1	Banglades	BD	BGD				
16	10681817	1997			France	FR	FRA		42-10-68	974502	MAE	Bilateral	1	Pakistan	PK	PAK				
17	23966926	2005			Germany	DE	DEU			503976	F O	Bilateral	1	Morocco	MA	MAR		Public sector		
18	34357303	2007			Japan	JP	JPN			072196T	JICA	Bilateral	1	Marshall I	MH	MHL				
19	27339565	2003			Australia	AU	AUS		G03084	20020249	AusAID	Bilateral		Singapore	SG	SGP				
20	21669904	2004			Italy	IT	ITA			42401	CA	Bilateral	1	Chad	TD	TCD				
21	27783934	2006			France	FR	FRA			2.01E+09	MAE	Bilateral	1	Cambodia	KH	KHM		secteur public		
22	26941247	1996			Sweden	SE	SWE		4.4E+09	19966256	SIDA	Bilateral		Latvia	LV	LVA				
23	10465853	1995			Finland	FI	FIN		98395005	19950094	MFA	Bilateral		Lithuania	LT	LTU				

Figure 3: Raw Data

E. DataMart Development

We have implemented the Data ware house developed in Hadoop with the necessary tools. The data warehousing of the given dataset in the SQL server [8] with Business Intelligence tools for analysis is very tedious process. Hence we considered Hadoop for the implementation which is more subtle and ease for the analysis.

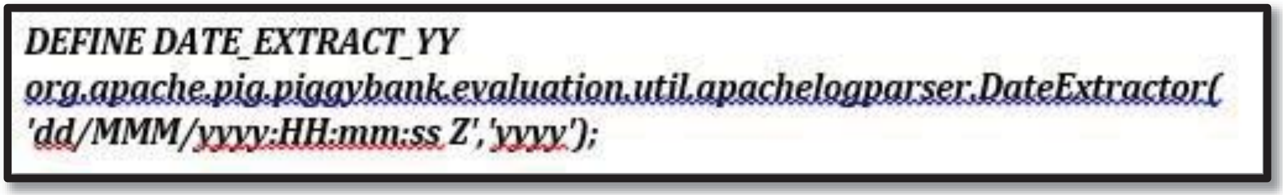
Hadoop [4] is an open source framework that allows users to process very large data in parallel. It's based on the framework that supports Google search engine. The Hadoop core is mainly divided into two modules:

- HDFS is the Hadoop Distributed File System. It allows you to store large amounts of data using multiple commodity servers connected in a cluster.
- Map-Reduce (MR) is a framework for parallel processing of large data sets. The default implementation is bonded with HDFS.

F. DataMart Population

We used Apache Sqoop [3] for importing data from the relational DB SQL server to Hive. For our Hadoop implementation we have used PIG for the data pre-processing and the Analysing of data in various dimensions. We use date extractor to decompose date variable and take out individual components of date, month, year, and so on.

We used the HIVE Metastore manager data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL where we can create the tables manually according to the tables in the SQL Server along with the appropriate data types for each column.



```
DEFINE DATE_EXTRACT_YY
org.apache.pig.piggybank.evaluation.util.apachelogparser.DateExtractor(
'dd/MMM/yyyy:HH:mm:ss Z','yyyy');
```

Import data into HDFS

We import data into HDFS by using Sqoop import command, where in we specify our relational DB table and the connection parameters. This will import the data and store it as a CSV file in a directory in HDFS.

Import data into HIVE

To import data into HIVE, we use the Sqoop import command and specify the 'hive-import'. This will import the data into a Hive Table with the appropriate data types for each column.

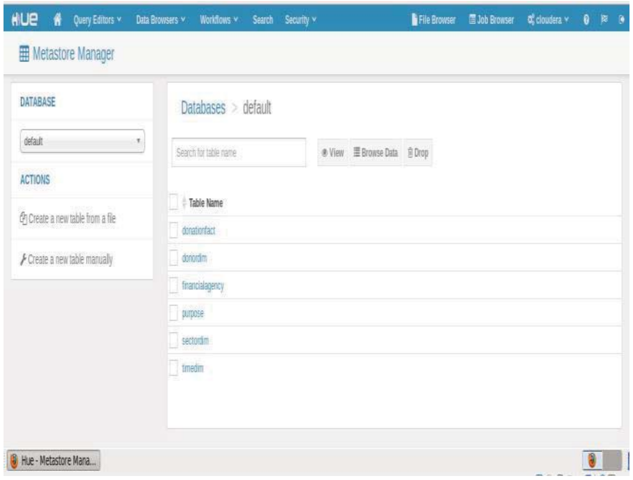


Figure 4: List of Tables in HIVE Metastore Manager

IV. ANALYSIS RESULTS

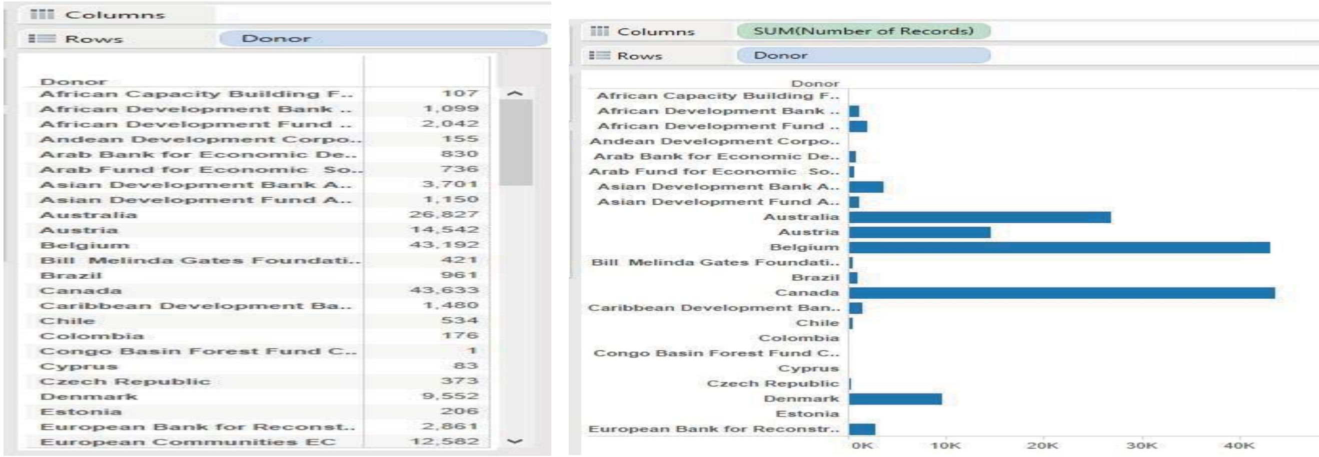
In this section, we will list all the finding from our project as per the requirement. It will cover general results, results from SAS Enterprise Guide, OLAP Cube Browser results and Hadoop analysis results. The Donations based on the country is shown in every step. In the same way we need to do for remaining cases like Number of Donations a country made in a particular year, Number of Donations in each sector, Number of Donations made by particular Financial Agency for particular purpose and Number of Donations made in particular Date.

A. Data Analysis Using Hadoop [5]

We connected our HIVE database to Tableau to perform data analysis using Hadoop. Our results are as follows:

No of Donations Vs Country

By using Hadoop we can perform Big Data analytics as it is a open-source software which is created and maintained by a network of developers from around the globe. The below statistics shows the no. of donations made by a particular Country (Donor). From all the countries that are listed below Canada is the country which offered more amount of donations.



Donor	
African Capacity Building F..	107
African Development Bank ..	1,099
African Development Fund ..	2,042
Andean Development Corpo..	155
Arab Bank for Economic De..	830
Arab Fund for Economic So..	736
Asian Development Bank A..	3,701
Asian Development Fund A..	1,150
Australia	26,827
Austria	14,542
Belgium	43,192
Bill Melinda Gates Foundati..	421
Brazil	961
Canada	43,633
Caribbean Development Ba..	1,480
Chile	534
Colombia	176
Congo Basin Forest Fund C..	1
Cyprus	83
Czech Republic	373
Denmark	9,552
Estonia	206
European Bank for Reconst..	2,861
European Communities EC	12,582

Columns
SUM(Number of Records)
Rows
Donor
Donor
African Capacity Building F..
African Development Bank ..
African Development Fund ..
Andean Development Corpo..
Arab Bank for Economic De..
Arab Fund for Economic So..
Asian Development Bank A..
Asian Development Fund A..
Australia
Austria
Belgium
Bill Melinda Gates Foundati..
Brazil
Canada
Caribbean Development Ban..
Chile
Colombia
Congo Basin Forest Fund C..
Cyprus
Czech Republic
Denmark
Estonia
European Bank for Reconstr..
0K
10K
20K
30K
40K

Figure 5: No of Donations Vs Country

B. Analysis Using SAS Enterprise Guide [7]

No of donations vs Country Donor

SAS Enterprise Guide is a Windows client application with an easy-to-use graphical user interface (GUI). The GUI consists of pull-down menus, dialog boxes, and windows that display and organize data, and perform numerical and graphical tasks. In this section, we will present basic results we obtained by using SAS Enterprise Guide. Before you can do anything in SAS Enterprise Guide, you need to add the data that you want to analyse to your project. The below statistics shows the no of donations made by a particular Country (Donor). From all the countries that are listed below United States is the country which offered more amount of donations.

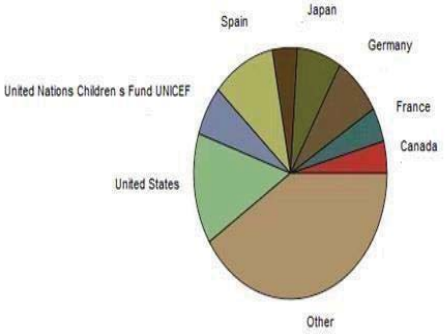


Figure 6: No of donations vs donor

C. Analysis Using OLAP Cube [6]

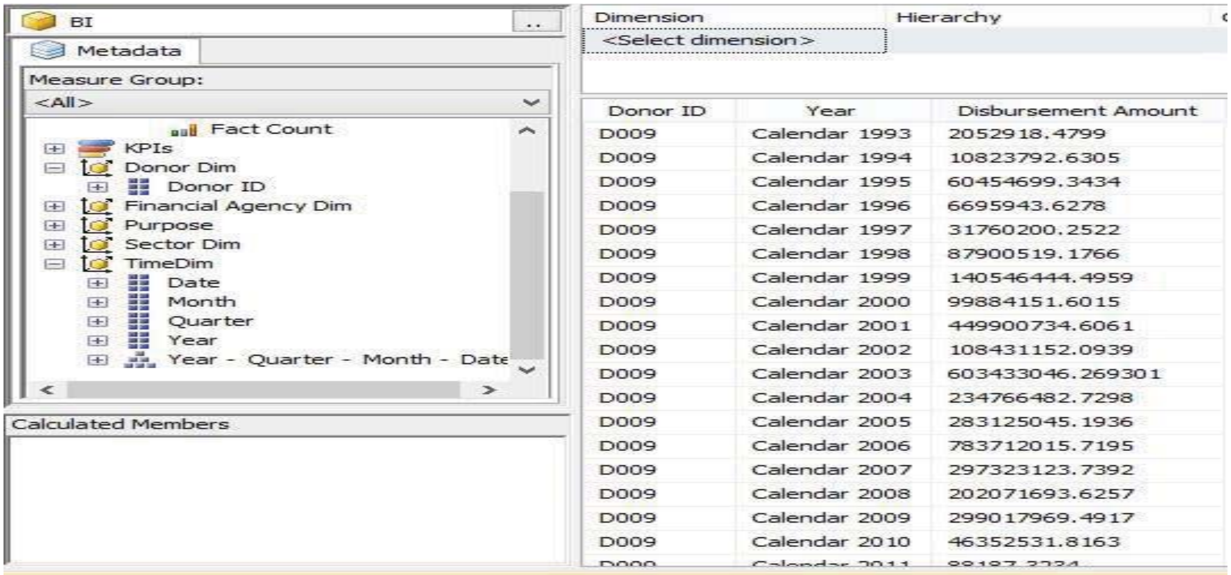
Amount Donated vs Country

An OLAP cube is a term that typical refers to multi-dimensional array of data. OLAP is an acronym for online analytical processing, which is a computer-based technique of analysing data to look for insights.

The Dimensional modelling is very important to be considered when constructing an OLAP Cube as in understanding what all should be the dimensions and what should be the measures and facts in the fact table. Dimensional data model is most often used in data warehousing systems.

After performing the ETL tasks a dimensional model is created which consists of fact table and dimension table which we follow for loading the data into the SQL Server for developing the OLAP Cube for Data Analysis.

In the fact table we will have start date column. So based on that we can analyse like how much amount is donated in that particular year. The below statistics shows the amount that was donated by each country in a particular year. From the time dimensional table we can get Date, Month, Quarter and Year grouping the particular country and financial Agency.



Donor ID	Year	Disbursement Amount
D009	Calendar 1993	2052918.4799
D009	Calendar 1994	10823792.6305
D009	Calendar 1995	60454699.3434
D009	Calendar 1996	6695943.6278
D009	Calendar 1997	31760200.2522
D009	Calendar 1998	87900519.1766
D009	Calendar 1999	140546444.4959
D009	Calendar 2000	99884151.6015
D009	Calendar 2001	449900734.6061
D009	Calendar 2002	108431152.0939
D009	Calendar 2003	603433046.269301
D009	Calendar 2004	234766482.7298
D009	Calendar 2005	283125045.1936
D009	Calendar 2006	783712015.7195
D009	Calendar 2007	297323123.7392
D009	Calendar 2008	202071693.6257
D009	Calendar 2009	299017969.4917
D009	Calendar 2010	46352531.8163
D009	Calendar 2011	[illegible]

Figure 7: Amount Donated vs Country

V. CONCLUSION

Using the Donation Database Data from AidData.org and with the help of Microsoft SQL Server 2012, Business Intelligence Development Studio, Visual Studio 2010, Microsoft Excel 2013, Cloudera 5.3, Oracle VM Virtual Box, SAS EG 6.1 and Tableau 9 we successfully analysed the Business Processes like Donations made between countries. Our analysis included both Donation analysis and Casual Analysis. Our analysis started with Data Profiling, identifying the analytical themes and designing the dimensional model and also understanding of the business story about the data. We implemented Hadoop data mart and our analysis results from all the three shows aspects like Number of Donations a country made in a particular year, Number of Donations in each sector, Number of Donations made by particular Financial Agency, Number of Donations made for particular purpose and Number of Donations made in particular Date using Tableau tool.

REFERENCES

[1] “Aid Data” http://aiddata.org/

[2] Aid Data, “Expanding Our Understanding of Aid with a New Generation in Development Finance Information.” special issue World Development, eds. J. Timmons Roberts, Michael G. Findley and Darren G. Hawkins 39 (11), 2011

[3] “Apache Sqoop” http://sqoop.apache.org/

[4] Elagib, S.B.; Najeeb, A.R.; Hashim, A.H.; Olanrewaju, R.F. "Big Data Analysis Solutions Using Map Reduce Framework", Computer and Communication Engineering (ICCCE), 2014 International Conference

[5] Integrating Hadoop into Business Intelligence and Data Warehousing, Second Quarter 2013, Philip Russom

[6] “OLAP Cube” https://support.office.com/en-us/article/Overview-of-Online-Analytical-Processing-OLAP-15d2cdde-f70b-4277-b009-ed732b75fdd6

[7] “SAS Enterprise Guide” http://support.sas.com/software/products/guide/

[8] “SQL Server” Alexander Rubin Principle Architect, Percona Apr 18, 2015 - Using Hadoop. Together with MySQL ... Big Data Analytics with Hadoop

[9] “Tableau tool” http://www.tableau.com/