International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 02, February-2021

Application of Data Virtualization

Siddhant Garg
School of Computing; DIT University,
Dehradun (U.K.); India.

***Abstract*: In this project, I take different data from different data sources (Eg: PostgreSQL, Excel, Redshift, etc.) and integrate them through Data Virtualization applying tool-Denodo to make a Data Warehouse. It is a simple process to deliver a holistic view of the enterprise.**

Key Words: Amazon Simple Storage Service (S3); CSV; Data sources; Data Virtualization; Data Warehouse; JSON; PostgreSQL; XML.

1. INTRODUCTION

Data is first and foremost thing in running an IT industry. Data allows companies to more felicitously determine the cause of problems and even find solutions to problems. Good data allows companies to establish criterions, benchmarks, and goals to keep moving forward. Nowadays organizations are facing different issue regarding Big Data but foremost is handling enormous historical data and analyzing to make data useful.

To overcome these problems, Data Warehouse acts as a central repository of integrated data from a company's operational databases as well as external sources. Data warehouses are merely intended to perform queries and analysis and often contain large proportion of historical data. Its analytical capabilities allow organizations to obtain valuable business insights from their data to improve decision-making.

The popular techniques for creating Data Warehouse are ETL (Extract, Transform and Load) and Data Virtualization [1]. These are in demand to integrate data from different data sources.

In today's scenario, the driven data enterprise (Eg: Oracle, IBM) store data in different data sources or servers like pdf, NoSQL, JSON, PostgreSQL, Excel, Cloud and other RDBMS sources. The problem of diversity is - ***It is difficult to integrate data from different sources and to use it in the system.*** Data Integration tool perform mapping, transformation, and data cleansing. The diversity is caused when each and every user has a different requirement for these data. According to them, data is stored in different data sources and servers which creates challenge in integrating the different sources of data for a single channel, which is major challenge faced by data driven enterprise (Data Scientist or Data Consumers).

2. BASIC REQUIREMENTS

2.1 Hardware Requirements:

- Physical Memory (RAM) of 4GB or more.
- Graphic card minimum 4GB.
- Minimum 5GB disk space free.
- Minimum i5 7th Gen processor.

2.2 Software Requirements:

- Denodo (version 6.0 and above, Express License) [2]

Limitations of Virtual Data Port with Denodo Express	
Number of user accounts	1 Virtual Data Port includes a default administrator account ("admin"). As the license only allows you to have one user account, you will always have to use "admin".
Maximum number of simultaneous requests	3
Maximum number of rows returned by a query	10,000
ODBC adapters allowed	Microsoft Access and Excel
Sources not allowed	Multidimensional databases: SAP BI, SAP BW, Mondrian, Microsoft Analytical Services and Oracle Essbase. Denodo Aracne Google Search Custom wrappers The Denodo Browser client cannot be used to retrieve data from CSV, JSON or XML files.
View parameters	You cannot set view parameters in a derived view.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 02, February-2021

Importing extensions (jars) is not allowed	This implies not being able to develop custom connectors to sources, custom policies or custom functions.
Version Control System	Disabled. With Denodo Express you cannot store your work in a Version Control System such as Subversion, GIT, etc.
JMS listeners cannot be created	
Limitations of Scheduler with Denodo Express	
Maximum number of jobs	1

- PostgreSQL (version 12.1)
- Visual Paradigm (version 16.1)
- Excel (version 16051.12827.20336.0)
- Excel (csv file) to XML converter
- Excel (csv file) to JSON converter
- Amazon Simple Storage Services(S3)

Limitations for Amazon S3 free usage tier:
AWS customers receive 5 GB of Amazon S3 Standard storage. 20,000 Get Requests, 2,000 Put Requests, 15GB of data transfer in and 15GB of data transfer out each month for one year.

➢ DVD Rental Data

3. AUXILIARY TECHNIQUES

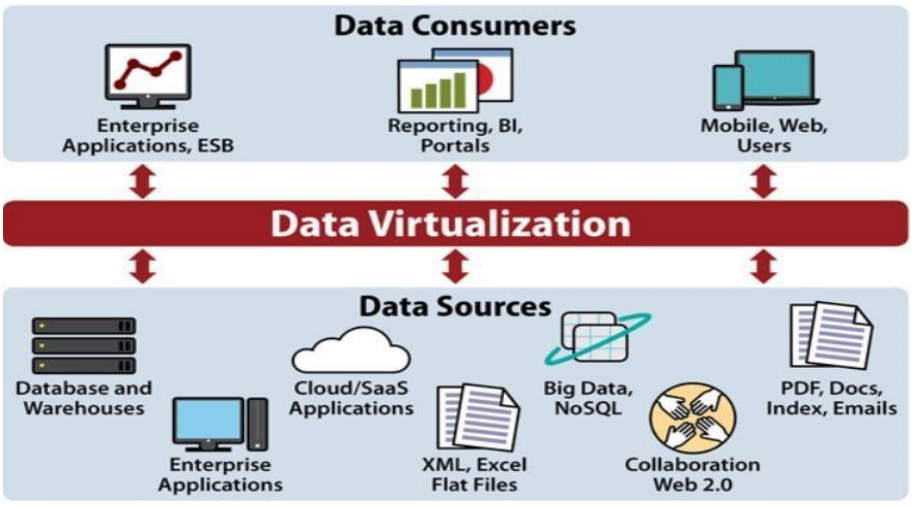
Tech 1:

For example, if we take two data sources- (1) Excel file, (2) SQL. In case of integrating data of both the sources, ETL is a needed platform through which data pipelines are created which firstly extracts the data from excel files and ingest them into SQL server.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 02, February-2021

Thus, if we apply the same scenario in a data driven enterprise which has diverse data then it is a complex process.

Tech 2:



To overcome this challenge, modern technique i.e., Denodo Data Virtualization [3] is used, which works on meta data of the underlying sources of information to develop a single, logical, virtual layer of information and provides them a real-time updated data. This integrates any type of enterprise data siloed which gives a maximum data without any loss or loop - holes. Unlike ETL solutions, which replicate data, Data Virtualization leaves the data in source systems, simply exposed and integrated view of all the data-to-data consumers. Data Virtualization fetches data in real time from underlying source systems and it proves that connecting to data is far superior to collecting it. Currently data storing is not enough to have greater competition, but it is necessary that data should be integrated in a single place so that they cease to be a cost to become business asset. Data virtualization does not normally persist or replicate data from source systems. It only stores metadata for the virtual views and integration logic. Caching can be used to improve performance but, by and large, data virtualization is intended to be very lightweight and agile.

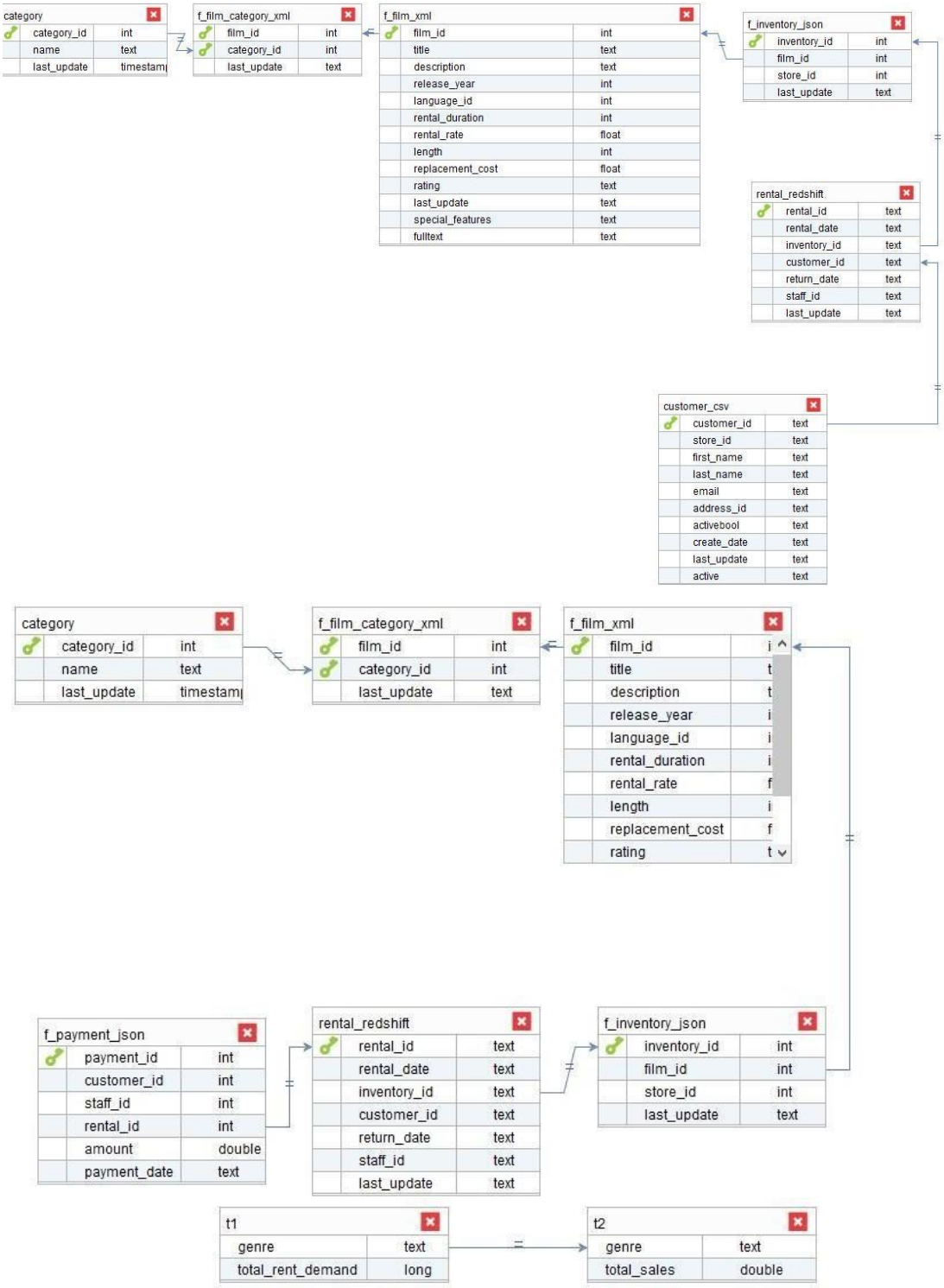
4. PROJECT

Data Virtualization is a critical part of the Logical Data Warehouse [4] architecture enabling queries to be federated across multiple data sources-(1) traditional structured data sources, such as databases, data warehouses, etc., (2) less traditional data sources, such as Hadoop, NoSQL, Web Services, SaaS applications and so on while still appearing as a single 'logical' data source to the user. It is the ultimate in modern data integration because it breaks down silos and formats, performing data replication and federation in a real-time format, allowing for greater speed and agility and response time. It should be noted that data virtualization is not a data store replicator. Data virtualization does not normally persist or replicate data from source systems. It only stores metadata for the virtual views and integration logic. Caching can be used to improve performance but, by and large, data virtualization is intended to be very lightweight and agile.

I used DVD Rental [5] data which consists of 15 tables and divided them into 5 different data sources i.e., Amazon S3, CSV, PostgreSQL, JSON, XML.
The data source files consisting of different distributed tables are integrated through Denodo and it forms a self-service Logical Data Warehouse.
I analyzed the data by taking some questions:

Q1: What are the top and least rented (in-demand) genres and what are their total sales?
In this, I integrated different tables from different data sources through Denodo and by using Join operator got the result for top and least rented genres and their total sales.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 02, February-2021



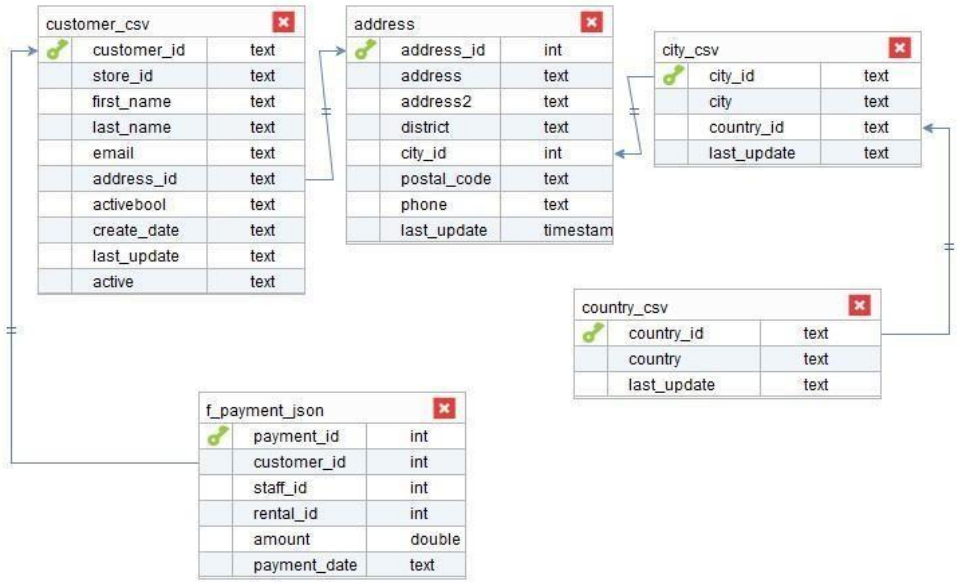
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 02, February-2021

Result:

genre	total_rent_demand	total_sales
Sports	1179	4892.189999999867
Animation	1166	4245.309999999882
Action	1112	3951.8399999998974
Sci-Fi	1101	4336.009999999889
Family	1096	3830.1499999999073
Drama	1060	4118.459999999898
Documentary	1050	3749.6499999999114
Foreign	1033	3934.469999999903
Games	969	3922.179999999915
Children	945	3309.389999999933
Comedy	941	4002.4799999999136
New	940	3966.379999999914
Classics	939	3353.37999999993
Horror	846	3401.269999999934
Travel	837	3227.3599999999433
Music	830	3071.51999999995

Q2: Who are the top 5 customers per total sales and can we get their detail just in case Rent A Film wants to reward them?

In this, I integrated different tables from different data sources through Denodo and got result for top 5 customers per total sales and their details.
It gives real time data which ETL takes time to process.



Result:

full_name	email	address	phone	city	country	total_purchase_in_currency
EleanorHunt	eleanor.hunt@sakilacustomer.org	1952 Pune Lane	354615066969	Saint-Denis	Runion	211.5500000000001
KarlSeal	karl.seal@sakilacustomer.org	1427 Tabuk Place	214756839122	Cape Coral	United States	208.58000000000013
MarionSnyder	marion.snyder@sakilacustomer.org	1891 Rizhao Boulevard	391065549876	Santa Brbara dOeste	Brazil	194.61000000000007
RhondaKennedy	rhonda.kennedy@sakilacustomer.org	1749 Daxian Place	963369996279	Apeldoorn	Netherlands	191.62000000000006
ClaraShaw	clara.shaw@sakilacustomer.org	1027 Songkhla Manor	563660187896	Molodetno	Belarus	189.60000000000005

It is an excellent solution in compare to ETL as data needs to be accessed and delivered in real- time and this is very important for decision support applications. It also leaves the source data where it is and delegates the queries down to the source systems while ETL copies data from source system and stores it in a duplicate data store.

CONCLUSIONS:

Data Virtualization is a critical part of the logical Data Warehouse architecture enabling queries to be associated across multiple data sources-traditional structured data sources and less traditional data sources while still appearing as a single ‘logical’ data source to the user.
The benefits of data virtualization in future for companies are quickly combining different sources of data, improving productivity, accelerating time value, eliminating latency, maintaining data warehouse and reducing the need for multiple copies of data as well as less hardware.

REFERENCES:

[1] community.denodo.com/kb/view/document/Data%20Virtualization%20and%20ETL?tag= Best+Practices
[2] community.denodo.com/docs/html/browse/6.0/platform/installation/appendix/limitations_ of_the_denodo_express_license/limitations_of_the_denodo_express_license
[3] https://www.denodo.com/en/data-virtualization/overview [4]www.tibco.com/reference-center/what-is-a-logical-data-
[4] warehouse#:~:text=A%20logical%20data%20warehouse%20(LDW,logical%E2%80%9D%20data%20source%20to%20users.
[5] [5] https://www.postgresqltutorial.com/postgresql-sample-database/