IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Elucidation Data Migration to Cloud Services

[1]Jyothi Kulkarni, [2]Shaila H Koppad
[1]Student, [2]Assistant Professor
[1]Departmenet of Master of Computer Applications,
[1]RV College of Engineering, Bengaluru, India

Abstract: Cloud computing is a booming technology in today's software world. Cloud data migration is a process of moving data from on-premise setup to cloud. Data migration is most important and all organizations should implement it. It is necessary to move data to cloud as it provides different advantages like security from malicious attacks, elasticity, autoscaling, load balancing and so on. cloud provides different services like software as a service (eg: Google apps, Commerce cloud) and platform as a service (eg : Microsoft azure ,Open Shift) and Infrastructure as a service(google compute engine ,AWS ec2). Cloud technologies provides services like Databases, servers, Documents, storage etc. Using Cloud one can remotely access all the services provided by the cloud provider. Transferring data to cloud include some challenges like more cost is involved in migrating data from on-premise database to cloud and also selecting the perfect cloud technology required for the organization require skills.

***Index Terms* – Cloud Computing, Microsoft Azure, Data Lake, Data Bricks, Elasticity, Autoscaling**

I. INTRODUCTION

As we all know Data is growing day-to-day exponentially. Different types industries like Healthcare industries, Education industries, Transport industries are producing huge amount of data. Data is growing in different varieties, volumes, velocity and veracity which is necessary for analysis of data. So it is important to store all the data in secure place.

Cloud is a most powerful technology used to store data. Cloud can handle huge amounts of data in a most secure way. There are different cloud platforms available in today’s market like AWS, Microsoft azure, google cloud. Organizations can store all its data into different types of cloud which provide different services like Databases, Servers, Networking [3]. So it is important to transfer data into the cloud.

As we all know information is a powerful tool which is used for analysis and decision making. All most all organizations are planning to transform their data into data warehouse in the cloud. As data warehouse has some specific benefits like, it can handle vast amount of data and helps in decision making. Using Data warehouse one can work on heterogeneous data sources and can perform different operations like filters, transformations, cleaning .Then data is used for reporting, analyzing and decision making. Here Data warehouse divided into data marts, Data marts are the parts of data warehouse used for different business lines like sales, finance, marketing etc[2].

There are different types of data like real time data, initial data, derived data. Storing all these data in the data warehouse and storing them in different data marts based on the type of data. By analyzing the data from different data marts one can report the data and draw business insights.

Some of the top benefits of the cloud computing are:

1.Cost: Through cloud one can reduce the cost incurred to perform the specific task. Because cloud provides hardware and various software’s and one can make use of those. Provides pay-as-you-go services.

2.Elasticity: It is the most important advantage of cloud, Cloud has an ability to increase or decrease the resources based on the necessity. It reduces the wastage of resources as well as shortage of resources.

3.Performance: Services of cloud computing usually run on the different networks over the world. So they will be continuously upgraded and provide good performance when compared to the on-premise storage.

4.Back-up of data: Here there is no necessity of thinking about losing data. Because if one uses the cloud to store the data ,Their data security is taken care of by the service providers.

5.Collaboration among employees: By making use of cloud for storing the organization's data, it increases the collaboration among the employees, as all data is stored at one place, they can access any data whatever they want and also store the data.

Data Modernization

Through Data Modernization organization can provide ideal services to their clients. Data Modernization is moving the data from on-premise setup or existing Database to the cloud storage (Data Lakes or the Data warehouse). By doing the Data migration or the Data modernization organization can achieve higher performance, secure Data, Reliability, Productivity.

This section is followed by literature survey which includes different papers about the cloud computing, Building Pipelines, Data Lakes and Analysis of data using Data warehouses. The next section followed by proposed methodology which includes the steps followed in proposed system. Proposed methodology is followed by Results section which explains the results of the process and last section is conclusion of the project.

II. LITERATURE SURVEY

In this section I have explained different research Papers on Data security in cloud, Building pipelines in cloud, Data storage in data lake, analysis using data marts and also about data warehouse which I have referred while doing the project.

In this paper author gives us a brief introduction about the data security in the cloud. Data security in the cloud is the main concept where the on-premise databases lack in. And using cloud storage data security issues can be resolved and proposes ideas against threats and enhances data security in the big data environment and also restrict the access to the unauthorized users.[1]

In this paper author defines the necessity of the data migration to the cloud. As it is difficult to handle big data as well as heterogeneous data. It is necessary to migrate data to the cloud .Here the author explained about the methods to collect data from different resources and transform it to the cloud.[2]

This paper explained about the importance of the cloud computing. In the world of big data , it is difficult to process the big data and store it .So it is essential to use the cloud .Using cloud one can run the entire operating system. This Paper shows the need of the cloud and how it helps in aspects of Data security, Data Modernization etc [3].

This analysis shows the creation of Data Marts (Data marts are the parts of data warehouse which are specific to the specific business lines like sales, finance etc) using data warehouse and storing all necessary data in warehouse. The development of data marts involves identification of dimensions ,facts etc. These papers explained about different types of data warehouses.[4]

ETL (Extract, Transform, Load) is extracting different types of data from different resources and transforming the data using required filters and queries, finally loading to the data warehouse. This paper mainly includes the various technologies used for ETL process and also how the processes is taken place. It includes building the ETL pipelines for batch processing.[5]

This paper proposes the use of Data factory in this era of Big Data, which helps in the Automation of tasks and reduces the cost. Data factory is a Data Analysis solution which contains different activities like Copy Data, Move and Transform and it also contains different filters , through which one can build a pipelines and automatically transform data.[6]

This Papers proposes the use of data lake, In Today’s industry data is generated from different sources and different types of data is generated which a normal on-premise databases cannot handle. So Data lake is most important thing, which can handle heterogeneous data. And also can store huge amount of data which comes from various departments. This Data can be used by Data Analysts, Business Intelligence Developer to analyze the data.[7]

As mentioned in the above papers Data Migration is cloud is important but while transferring data there is chances of data loss during transformations and also it requires more time to transfer data through ETL process. So to transfer data from on-premise setup to cloud it is necessary to build a model where the data storage is done in each and every step and Automation of tasks like triggering pipelines and building JSON’s for parameters can be done to save time and effort.

III.PROPOSED METHODOLOGY

As transferring data to cloud has many advantages like Storage, Autoscaling, Security, Elasticity here we are proposing a new system which can transfer data from on-premise setup to Microsoft azure securely with the help of Data Factory pipelines and By Storing the intermediate data in the data lake one can back up data when lost. This Diagram explains the Proposed System.

Block Diagram:

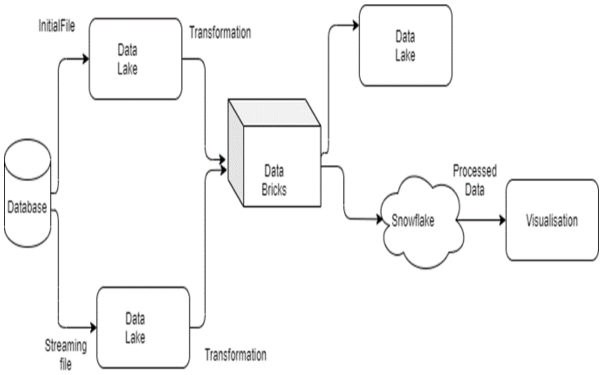


Fig 1: Block Diagram

In fig1 the block diagram consists of Database where currently the data is stored. Data is a storage in the Data Lake which stores the raw data in the same format as it is. Here there are 2 data lakes. One is used to load the initial file and the other is used to store the real time data .Data Bricks is used for the transformation of data. After Transformation data is being stored in a Data warehouse which is used for visualization.

Database is where the current data is stored. ETL (Extract, Transform, Load) Tools extracts data from this layer and implements different functionalities on this data. It undergoes different functionalities like Preparation ,Transformation ,storage are the different processes performed .First data is collected ,then cleaned, transformed and Stored in the Data Warehouse.

By storing data in a Data Warehouse, Data scientists can analyze different datasets and get huge data to analyze, Business intelligence developers can analyze and draw insights from the data.

Proposed System has 5 modules:

1. Data Collection

As data collected from different sources, it is necessary to store it in a particular location. Here Data Collected from different sources is taken as input and Data stored in the Database.

2. Data Storage in Data Lake

It includes transferring of data from database to data lake. There are 2 types of data, initial data and incremental data. Loading those data into 2 different data lake. Initial data loading is one time loading and incremental data loading is a continuous loading.

3. Data Preparation and Transformation

Taking Data from 2 different Data lakes, Data preparation and required transformation has to be done. And storing the transformed or processed data into data warehouse. Data preparation includes cleaning of data, removing null values, Normalizing of some values. Then required transformations are done on that data.

4. Data visualization

Data visualization is a process of deriving business insights from the dataset. Input is taken from the Data Warehouse and different graphs are drawn to get the quick insights of the data .This can be used for further business decisions.

Extract, Transform and Load Workflow

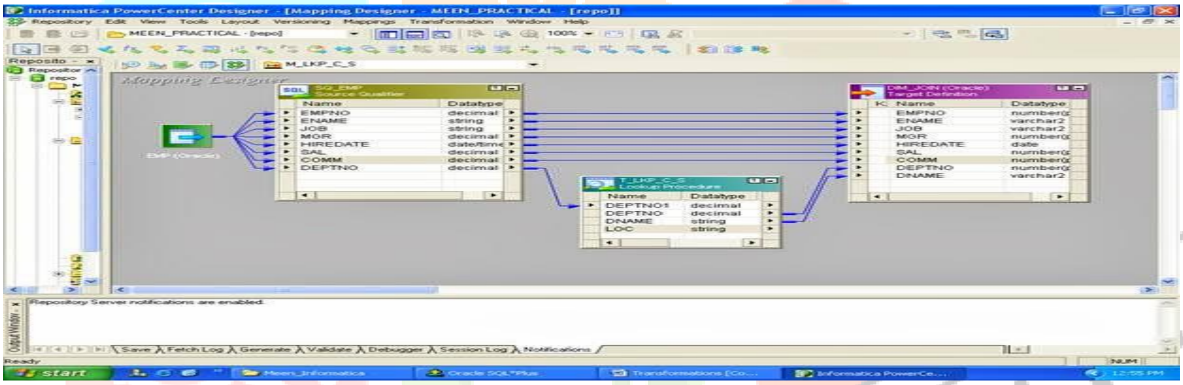
- The data warehouse platform will be enabled through orchestrated data pipelines which automatically refresh, ingest data and Power BI reports with minimal IT overhead. This includes:
- Convert Extract, Transform and Load logic from Informatica to Azure based data engineering components/services like ADF, Azure Databricks etc.
- Data warehouse moved from On premise Database to Snowflake Data Warehouse
- Reporting to be moved from My BI to Power BI.

IV RESULTS

In this section final results of the project are discussed. On migrating data to cloud there are some benefits like Elasticity, Backup. Here I have compared the features of the existing system and proposed system, through which one can get advantages of proposed system over existing system.

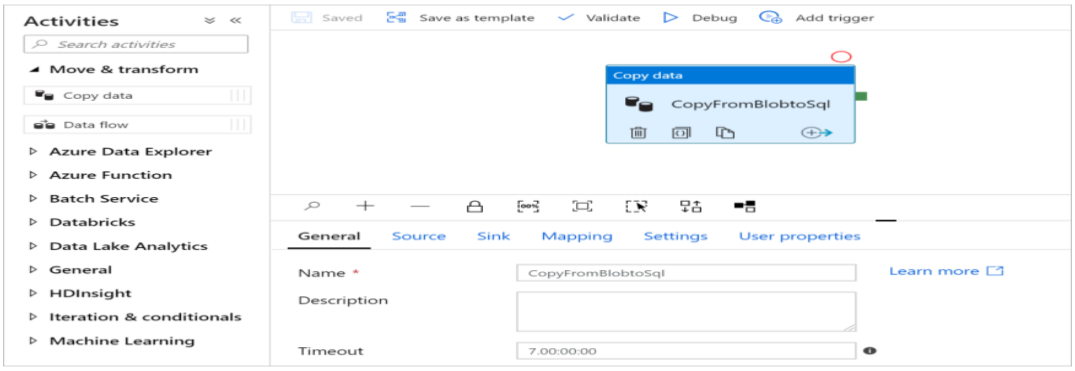
SL NO	Existing System	Proposed System
1	Data is stored in on-premise Database	Data is stored in the cloud
2	Data is not secured. Their is a possibility of malicious attack.	Data is secured as it is stored in remote place and provides different security measures and policies
3	Elasticity is not possible means cannot increase or decrease resources	Elasticity is possible means can increase resources whenever needed and stop them when not in need
4	Back up is not automated and not cost efficient	Back up is automated and it is cost Efficient

Migrating data from Database to Cloud involves different activities such as Backtracking tables in informatica, uploading data to data lake, Performing transformations in Data Bricks and Storing data into Data Warehouse. Here I have included some screenshots of informatica backtracking, Running pipeline in data factory and launching data bricks workspace to write code to perform transformation



Repository Server notifications are enabled.

Fig 2 : Informatica Backtracking : It is process done while migrating data to cloud which involves backtracking of tables which are present in the database to get information like filter applied on the data, source of the data, transformations done on data and also respective column names in the source.

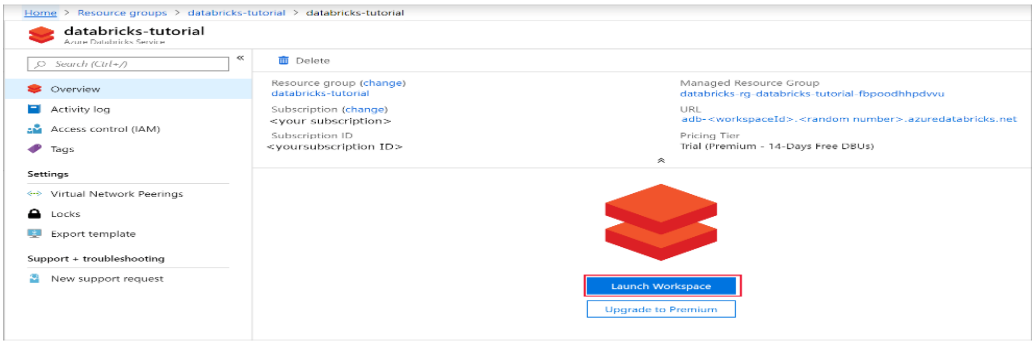


Name * CopyFromBlobtoSql

Description

Timeout 7.00:00:00

Fig 3: Copy Data Activity: To transfer data from one part of the cloud to other like data lake to data bricks notebook or data lake to data warehouse, it should be done using data factory pipelines. Data factory pipelines builds the connection between source and destination to pass the data. And also Data factory pipelines contains different activities like Copy Data activity, Lookup activity etc.. through these activities one can perform the action required.



Resource group (change)
databricks-tutorial

Subscription (change)
<your subscription>

Subscription ID
<yoursubscription ID>

Managed Resource Group
databricks-rg-databricks-tutorial-fbpoodhhpdvvu

URL
adb-<workspaceId>.<random number>.azuredatabricks.net

Pricing Tier
Trial (Premium - 14-Days Free DBUs)

Fig 4: Data Bricks for transformation- Once data is loaded into the data lake, one has to perform required transformations to data. So that it can be uploaded to the cloud data warehouse. Data bricks is also used to write code to perform automation tasks like uploading JSON file, executing queries on daily basis etc.

V.CONCLUSION

As Data is increasing day by day in exponential way and it is important to store the data securely, also it is important to provide a common platform for all the developers to upload the data and access the data. So it is important to migrate data from database to cloud which servers as a common platform and stores data securely. Migrating data to cloud is advantageous as it overcomes problems like security, network traffic and provides some features like disaster recovery, access from the remote place, automation of tasks etc. So Organizations have to move their data to cloud platforms. This paper explains about how to migrate data from data base to cloud data warehouse using data lake and data bricks.

VI. REFERENCES

[1] Fengling Wang. and I. Han Wang, " Research on Data Security in Big Data Cloud Computing Environment", 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)

[2]Koong Wah Yan and Nagendran M. Perumal, "Data migration ecosystem for big data invited paper", 2013 7th IEEE International Conference on Digital Ecosystems and Technologies (DEST).

[3] Shyam Patidar and Dheeraj Rane, "A Survey on Cloud Computing:", 2012 Second International Conference on Advanced Computing & Communication Technologies.

[4]Boon Keong Seah and Nor Ezam Selan"Design and implementation of data warehouse with data model using survey-based services data", Fourth edition of the International Conference on the Innovative Computing Technology (INTECH 2014),

[5] Qin Hanlin and Jin Xianzhen, "Research on Extract, Transform and Load(ETL) in Land and Resources Star Schema Data Warehouse", 2012 Fifth International Symposium on Computational Intelligence and Design.

[6]Yaojun Wang and Yangyang Li, "Data Factory: An Efficient Data Analysis Solution in the Era of Big Data", 2020 5th IEEE International Conference on Big Data Analytics (ICBDA).

[7] Hassan Mehmood and Ekaterina Gilman, "Implementing Big Data Lake for Heterogeneous Data Sources", 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)

[8] Suyel Namasudra and Pinki Roy, "Cloud Computing: Fundamentals and Research Issues **",** 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM)

[9] MISKUF, M. and I. ZOLOTOVA, "Application of business intelligence solutions on manufacturing data", *2015 IEEE 13th International Symposium on. IEEE*, 2015.

[10] MISKUF, M. and I. ZOLOTOVA, "Application of business intelligence solutions from Microsoft and IBM on manufacturing data", *2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE*, 2016.

[11] P. PENIAK and M. FRANEKOVA, "Open communication protocols for integration of embedded systems within Industry 4", *Applied Electronics (AE) 2015 International Conference on. IEEE*, 2015.

[12] T. LOJKA, M. BUNDZEL and I. ZOLOTOVA, "Industrial gateway for data acquisition and remote control", *Acta Electrotechnica et Informatica*, vol. 2, no. 15, pp. 43-43.

[13] L. JAY, B. BAGHERI and H. -A. KAO, "A cyber-physical systems architecture for industry 4.0-based manufacturing systems", *Manufacturing Letters*, vol. 3, pp. 18-23, 2015.

[14] R. BARTLETT, "A practitioner's guide to Business Analytics Using Data Analysis Tools to Improve Your Organization's Decision Making and Strategy", ISBN 978-0-07-180759-3.

[15] P. URBAN and L. LANDRYOVA, "Identification and evaluation of alarm logs from the alarm management system", *Carpathian Control Conference (ICCC) 2016 17th International. IEEE*, 2016.

[16] V. SIMONCICOVA and P. TANUSKA, "Creating a management view on key indicators using business intelligence in small and medium enterprises", *Cybernetics & Informatics (K&I). IEEE 2016*, 2016.