IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 10, October 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 10, October 2021 ||

| DOI:10.15680/IJIRSET.2021.1010078 |

A Case Study on Traditional Vs Cloud EDW

Yedidi Naga Shiva Harish

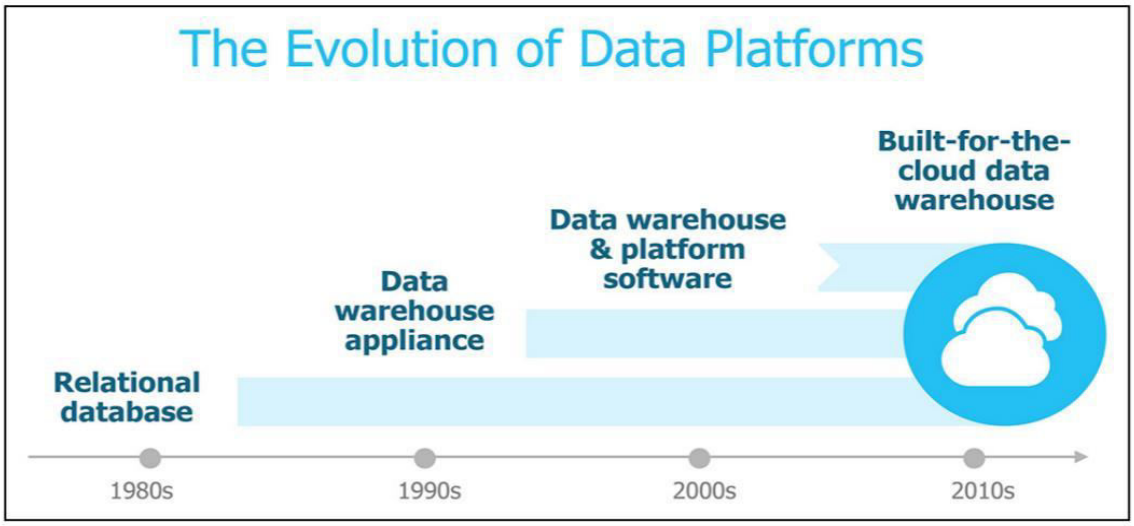
Gayathri Vidya Parishad, College of Engineering, Visakhapatnam, India

ABSTRACT: Data warehousing over the cloud environment has been studied and from the study it was understood that the data warehouse system over the cloud environment has advantages such as flexibility, scalability, dependability and reduced costs, without compromising on the security and compliance standards of the on-premise data warehousing.

I. INTRODUCTION

Currently major business Organization's Enterprise Data Warehousesare on Teradata Intelliflex System (hardware platform for demanding data analytics) which is on-Premise service model. Keeping in view of the operational cost expenditure, there is a necessity to explore alternative avenues to reduce operational cost without compromising the database performance and user experience.
This document elaborates different cloud Data warehouse systems considered for alternative Analytics platforms.
2. Background and Basic Terminology

a). What Is a Data Warehouse?
A data warehouse is a computer system dedicated to storing and analyzing data to reveal trends, patterns, and correlations that provide information and insight. Traditionally, organizations have used data warehouses to store and integrate data collected from their internal sources (usually transactional databases), including marketing, sales, production, and finance.
b). The Evolution of Data Warehousing
Data warehousing has evolved over four decades



c). Issues / limitations of the traditional data warehouse
Major setbacks of the traditional database are

- Availability
- Capacity estimation and forecasting to address exponential growth
- Performance (I/O) and user queries
- System availability and spikes
- system resource maintenance and Operational expenses (upgrades/maintenance and patching)
- Licensing costs

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 10, October 2021 ||

| DOI:10.15680/IJIRSET.2021.1010078 |

- Backup and DBA activity.

Cloud operating platform is emerging as alternative methodology to overcome the limitations of the traditional data warehousing.

3. Introduction to cloud data warehousing

Cloud data warehousing is a cost-effective way for companies to take advantage of the latest technology and architecture without the huge upfront cost to purchase, install, and configure the required hardware, software, and infrastructure. The various cloud data warehousing options are generally grouped into the following three categories

a). Software as a Service (SaaS)
This model allows the customers to use the entire application of the providers running on the cloud environment.
b). Platform as a Service (PaaS)
This model deploys a layer of software or development environment on to the cloud infrastructure and offered as a service. Higher levels of service can be built over this layer. To meet the manageability and scalability requirements of the applications, a predetermined combination of OS and application servers are suggested by the PaaS providers.
c). Infrastructure as a Service (IaaS)
This model provides intrinsic storage and computing proficiencies to the customers as normalized services. The basic computing resources, storage, networking equipment, etc. are shared to customers to handle workloads.
d). Characteristics of Cloud Computing
Cloud is expected to have the following five characteristics that are described below.

- **On-demand self-service:** Without interacting with the service providers, a customer can independently provision computing proficiencies, such as system storage and server time.
- **Broad network access:** Competences that are available over the network are accessed through typical mechanisms that confirm its use by diverse weak or strong client platforms.
- **Resource pooling:** According to the needs of customer, they are dynamically allocated and deallocated with different physical and virtual resources. The cloud providers' computing resources are then shared to various customers' using a multi-tenant model. Due to location independence, the customer has no control or knowledge over the location of available resources.
- **Rapid elasticity:** In some cases, in order to quickly scale out and scale in, automatically the proficiencies are quickly and flexibly provisioned to the customers. The proficiencies can be purchased in any number at any time by the consumer since they are unlimitedly available.
- **Measured Service:** Using a metering facility, cloud systems can automatically control and optimize resource usage based on the type of services at some level of abstraction. The utilization of resources can be controlled, observed and reported which allow both the provider and consumer to make use of their exploited services efficiently.

e). Advantages of cloud computing that organizations are experiencing today

- Do more with less
- Flexible costs
- Always-on availability
- Improved mobility
- Quick Software upgrades
- Improved collaboration
- Cloud computing is more cost effective
- Expenses can be quickly reduced
- Flexible capacity
- Facilitate M&A activity
- Less environmental impact

4. On-premise vs Cloud data warehousing: Pros and Cons
Companies are increasingly moving towards cloud-based data warehouses instead of traditional on- premise systems. Cloud-based data warehouses differ from traditional warehouses in the following ways:

- There is no need to purchase physical hardware.
- It's quicker and cheaper to set up and scale cloud data warehouses.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 10, October 2021 ||

| DOI:10.15680/IJIRSET.2021.1010078 |

In recent years, many Industries' data warehouses are moving to the cloud. The new cloud-based data warehouses do not adhere to the traditional architecture; each data warehouse offering has a unique architecture.

Table below outlines differences between on-premises and cloud platforms

Sl.No.	Feature	On – Premises	Cloud
1	**Capacity**	• Requires continuous monitoring and analysis on capacity and TTL (time to • Capacity is provisioned based on • guessing and periodical growth. • For wrong estimates, we ended up paying for expensive resources and • insufficient capacity • Data Center resource costs	• Capacity Planning is Agile • Capacity can be facilitated in few hours • We can add/delete capacity as temporary/ permanent based on the usage needs • No Data center costs
2	**Agility**	• Infrastructure change Planning and implementation takes time • This delays the implementation speed. • Affects experimentation and slow down	• Experimentation quickly with low cost and risk • Spin up resources in few clicks for innovation and testing • Meets the organization targets in very fast due to dramatic increase in agility in implementation
3	**Scalability & Elasticity**	• Needs more planning and budgeting to address future resource needs • We end up paying infrastructure or data center costs for the resources that are not used	• Auto scaling and elastic load balance enables - Scale the resources up/down quickly and easily with quick turnaround - No additional costs for the resources that are not used (pay as you use model)
4	**Security and Compliance**	• Organization standards and security protocols may avoid security bleach. • Manage more security access layers on Customers PII and PCI data. • Periodic security and compliance checks will helps manage tightest securities on organization sensitive data • Security ownership lyes with the Organization	• Although cloud service providers implement the best security standards and industry certifications, storing data and important files on external service providers always opens risks • Security Ownership shared between Cloud service provider and Organization. Improper handshaking leads to risk of security breach
5	**Availability**	• Need more resources and effort to monitor systems. • During upgrades and maintenance system will not be available	• Systems always accessible with fault tolerant infrastructure • System upgrades are seamless to end users and system is always available

5. Comparison study on Traditional Vs Cloud Data WarehouseSystems

We proactively conducted a proof of concept and analyzed Architectural and Technical comparison between current Traditional Data warehouse systems and different Cloud data warehouses platforms. This document details out the architectural and technical comparison of these platforms.

Following cloud platforms are considered for comparison and assessment for futuristic adoption to facilitate data analytics requirements.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 10, October 2021 ||

| DOI:10.15680/IJIRSET.2021.1010078 |

• Snowflake
• BigQuery (Google Cloud Platform)

Following are the considerations against which above systems and current Teradata System are evaluated.

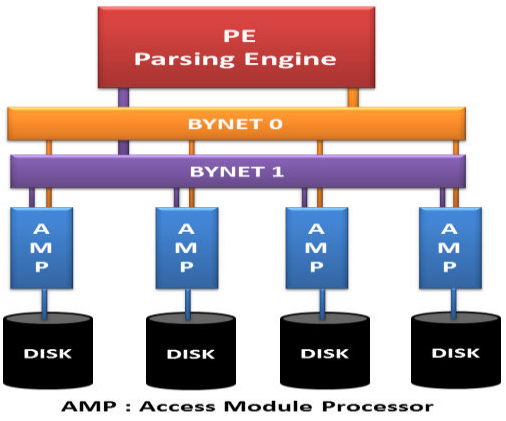
- Architectural -> Teradata Vs Cloud Data warehouse systems
- Features -> Teradata Vs Snowflake Vs Big Query
- Technical Differences -> Teradata Vs Snowflake Vs Big Query

Please make a note that we used Teradata prototype systems is used for this comparison.
1. Architecturaldifferences->TeradataVsSnowflakeVsBigQuery
In this section, we compare the architecture of Teradata to snowflake and BigQuery.
• Teradata

The Teradata Database is a relational database management system (RDBMS) that drives an Organization data warehouse. The Teradata database is not just a software solution, the Teradata RDBMS consists of hardware AND software. The Teradata database is the first RDBMS specifically designed for Massively Parallel Processing (MPP). The Teradata Database is an open system, compliant with industry ANSI standards

Teradata Architecture and Main features

- **ANSI SQL 2011 standard**
- **Single data store**
- **Linear Scalability**
- **Unconditional parallelism (parallel architecture)**
- **Ability to model the business**
- **Mature, parallel-aware Optimizer**



Picture Referred from: http://www.terada tawiki.net/2013/09/Teradata-Architecture.html

• Snowflake

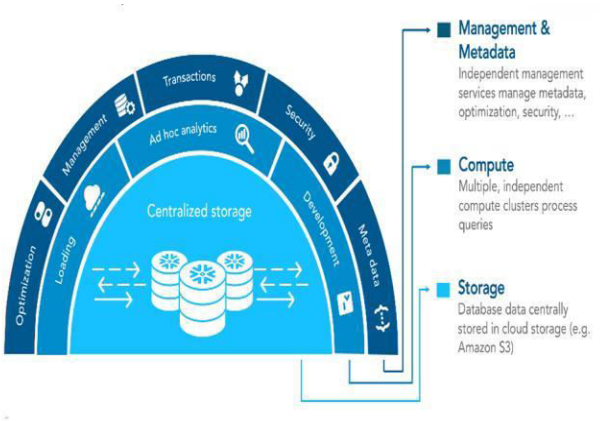
Snowflake is a fully relational columnar database as an opposite to row storage databases like Oracle, MS SQL, MySQL etc. Snowflake is an analytic data warehouse provided as a Software-as-a-Service (Saas). Snowflake provides a data warehouse that is – faster, easier to use, and far more flexible than traditional data warehouses. Snowflake has many similarities to other enterprise data warehouses, but also has additional functionality and unique capabilities.
• Snowflake Architecture and Main features

- Shared and centralized cloud storage
- Independent and flexible compute clusters
- Fully relational SQL data warehouse
- Turn-key service
- Pay what your use
- Standard interfaces for ETL and BI

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 10, October 2021 ||

| DOI:10.15680/IJIRSET.2021.1010078 |



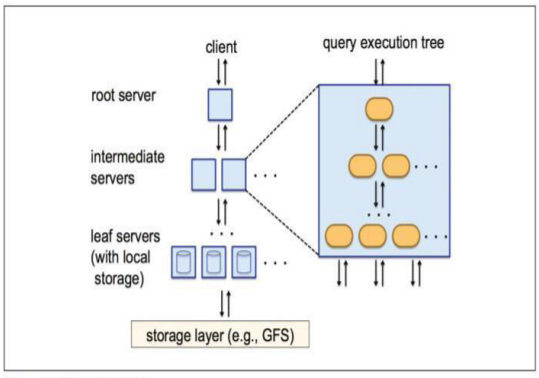
Picture Referred from www.Snowflake.com/architecture

• BigQuery

BigQuery is Google's Enterprise data warehouse enabling super-fast SQL queries using the processing power of Google's Dremel architecture, providing a full-managed and cloud based interactive query service for massive data sets. This is a fast, highly scalable, cost effective and fully managed cloud data warehouse for analytics with built-in machine learning capabilities.

BigQuery Architecture and Main features

- ANSI SQL 2011 standard
- Columnar storage
- Full or partition column scan using tens of thousands of machines over fast Google network
- Tree Architecture is used for dispatching queries and aggregating results across thousands of machines in a few seconds.
- Real-time streaming using API



Tree architecture of Dremel

Picture Referred from cloud.google.com/ igQueryTechnicalWP.pdf

2. Feature Differences Teradata Vs Snowflake Vs Big Query

Sl.No.	Feature	Teradata	Snowflake	BigQuery
1	Initial Release	1984	2014	2010
2	Database Mode	Relational RDMS	Relational RDMS	Relational RDMS
3	Available Platforms	On Premises and Cloud	Only Cloud	Only Cloud
4	Concurrency	Managed Workloads	High concurrency	High concurrency
5	Maintenance	Needs downtime for planned and scheduled maintenances	Seamless maintenance	Seamless maintenance

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 10, October 2021 ||

| DOI:10.15680/IJIRSET.2021.1010078 |

6	Scalability	Linearly scalable	automatic	automatic
7	Elasticity	No	Yes	Yes
8	Parallelism	Yes	Yes	Yes
9	Data Format Support	LATIN & UNICODE	LATIN & UNICODE & Support for diverse data like JSON & PARQUET, XML, Avro	LATIN & UNICODE & Support for diverse data like JSON, Avro Parquet, ORC
10	Data operations	Supports INSERT/UPDATES /DELETES/EXPORTS	Supports INSERT/UPDATES /DELETES/EXPORTS	Append-only, batch operation SQLs are expensive
11	Data Loading	Batch Loading	Batch Loading & Streaming	Batch Loading & Streaming
12	Processing Unstructured data	No	Partially (regular expression matching on text)	Partially (regular expression matching on text)
13	Handling large results / Join large table	High Performance	Low Performance	Low Performance
14	Workload management	Yes	No	No
15	Operating Systems	LINUX/ Windows	Hosted on AWS	Hosted on GCP
16	Availability	High availability, DR built in (No backup) is not required for Intelliflex Systems. All other Teradata systems model required DR (Disaster Recovery)	High availability, DR built in (No backup)	High availability, DR built in (No backup)
17	Tools/ Utilities	Teradata has own data loading Tools BTEQ, FEXP, FLOAD, MLOAD and TPUMP	No Inhouse tools, need use vendor products or Open Source tools	Needs use other vendor products or Open Source tools
18	BI /Reporting	Supports all Analytics and Reporting tools available in the market	Tableau	Google Data Studio
19	Pricing Model	Regular	Pay-as-you-go	Pay-as-you-go
20	Machine Learning and Artificial Intelligence Analytics	No	Yes	Yes
21	Monitoring Tool	Viewpoint	No Monitoring Tool, monitor resource usage using web interface and SQL	Stackdriver

3. Technical Differences Teradata Vs Snowflake Vs Big Query

Sl.No.	Feature	Teradata	Snowflake	BigQuery
1	SQL	ANSI SQL	ANSI SQL	ANSI SQL
2	Stored Procedure	Yes	No	No
3	Triggers	Yes	No	No
4	Data Types	Supports Industry standard datatypes including CLOB and BLOB	Standard datatypes & semi-structured data types like VARIANT OBJECT, ARRAY. It doesn't support CLOB and BLOB	Standard datatypes & semi-structured data types like ARRAY, STRCUT Geography. It doesn't support CLOB and BLOB
5	Programming Language	Yes	Yes	Yes

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 10, October 2021 ||

| DOI:10.15680/IJIRSET.2021.1010078 |

	support			
6	APIs & other access methods	.NET Client API HTTP REST JDBC JMS Adapter, ODBC, OLE DB	CLI Client, JDBC ODBC	RESTful HTTP/JSON API
7	Foreign keys	Yes	No	No
8	Join Indexes	Yes	No	No
9	Secondary indexes	Yes	No	No
10	Geospatial	Yes	No	Yes
11	Compression	Yes	Yes	Yes
12	Columnar	Need define explicitly	By default	By default
13	Control Data Access	Ability to Create Roles and Views and provisioned ad secure data access	Ability to Create Roles and Views and provisioned ad secure data access	BigQuery uses Identity and Access Management (IAM) to manage access to resources
14	Partitions	Supports Single and Multilevel Partitions	micro-partitions and data clustering	Limited partitions
15	Table Types & Views	• Physical Tables • Derived Tables • Volatile Tables • Global Temporary Tables • Views	• Temporary • Transient • Permanent • Views	• Native tables • External tables • Views
16	User-Defined Functions	Yes	Yes	Yes
17	Materialized Views	No	Yes	No

6. Conclusionandrecommendations

Data warehousing over the cloud environment has been studied and from the study it was understood that the data warehouse system over the cloud environment has advantages such as flexibility, scalability, dependability and reduced costs, without compromising on the security and compliance standards of the on-premise data warehousing.

REFERENCES

1. https://cloud.google.com/bigquery/docs/monitoring#view-dashboards https://docs.snowflake.net/manuals/sql-reference/data-types-unsupported.html

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details