A Comparative Study Between Hadoop and Snowflake

Raghavendra S [1], Hrithik Gautham T G[2], R Sindhu Rajendran[3]

[1] Raghavendra S, Dept. of Electronics and Communication Engineering, RV College of Engineering, Bengaluru, Karnataka, India
[2] Hrithik Gautham T G, Dept. of Electronics and Communication Engineering, RV College of Engineering, Bengaluru, Karnataka, India
[3] R Sindhu Rajendran, Assistant Professor, Dept. of Electronics and Communication Engineering, RV College of Engineering, Bengaluru, Karnataka, India

---***---

Abstract - *In the recent community, there have been a lot of enhancements in the Data Analytics and Machine Learning field. This helps various organizations to predict, analyze and tailor to their customer's needs. But, most of their Machine Learning models and Data analysis needs large quantities of processed data. Nowadays, storing large quantities of data comes with a number of challenges. Data integrity, Data protection to name a few. Due to the advancements in cloud computing, the consumer doesn't have to care about all these concerns. They just have to subscribe to the major cloud computing offerings like AWS. This cloud computing platform takes care of all the concerns, including scaling.*

In This paper, we discuss the history of Hadoop and usage of Hadoop in the past decade and how Hadoop is used in big. It also explains about the use of Snowflake and its new multi-bunch, shared-information architecture. The paper features a portion of the key highlights of Snowflake: Data Volumes, deployment complexity, Data quality, Query Latency and minimum data size. It features the contrast among Hadoop and snowflake and the benefit of snowflake over Hadoop which cause the relocation of Hadoop to snowflake lately. It concludes with exercises gained from surveys and a standpoint of work which is going on in recent times.

***Key Words*: Cloud Computing, Internet of Things, Amazon Web Services, Data Analytics, Machine learning, Database.**

1. INTRODUCTION

The approach of the cloud indicates a move away from programming conveyance and execution on close by servers, and toward shared information focuses and programming as-an administration arrangements facilitated by platform suppliers, for example, Big tech companies uses Shared Framework with outrageous adaptability and accessibility, and pay-more only as cost arise cost domain that adjusts to eccentric us-age requests. Be that as it may, these points of interest must be captured as software or hardware itself can scale flexibly over the pool of ware assets that is the cloud. Customary data warehousing arrangements to the cloud. They are de-marked to run on little, static groups of respectful machines, making them a poor architecture. One of the open source programming, Hadoop is utilized for capacity and handling of the big data on the huge clusters of the vendible equipment. Hadoop is one of the high-level projects from Apache which is being conveyed and it is utilized by the top worldwide network of specialist organizations and clients. It gives scalable capacity to all various information which improves the processing power and gives more prominent agility to deal with simultaneous tasks for all intents and purposes. Hadoop is known for its distributed data store model. Data scientists need large amounts of data to make and train these machine learning models. Within the age of Artificial Intelligence, quick and accurate access to data has become a very important competitive mortal. Data management (discovering, securing access, cleaning, combining, and making ready the information for analysis) is often recognized because it is the most time-consuming aspect of the process.

Modern Data warehouse such as Redshift, Big Query, Snowflake are common to use in many industries. Redshift is modern popular Data warehouse which is cloud based and facilitated legitimately with respect to Amazon Web Services, the organization's current cloud Infrastructure. Perhaps the greatest advantage to Redshift. Engineering that can scale in seconds to fulfil changing capacity needs. A significant issue confronting association with quickly changing information necessities is that scaling can be both expensive and complex in nature. Google Big Query is a modern data warehouse which is also cloud-based Enterprise information that offers fast SQL questions and intuitive investigation of enormous datasets. Big Query was developed based on Google's technology and was developed initially to process read-only data. The platform utilizes a columnar storage method that allows for much rapid data scanning as well as a tree like architecture model that makes querying and aggregating of data which stored in database results significantly easier and more efficient. The stage has been used continuously misrepresentation discovery, by utilizing its information gathering and authoritative limits. Snowflake Cloud Data Platform consolidates the intensity of information warehousing, the adaptability of huge information stages, the flexibility of the cloud, and live information sharing at a less expense of customary information stage arrangements. Snowflake conveys the presentation, simultaneousness, and effortlessness expected to store and break down the entirety of your information in one area, both for internal use and to make a data trade.

2. EVOLUTION OF HADOOP AND SNOWFLAKE OVER THE YEARS

The regularly developing innovation has brought about the requirement for putting away and preparing too much information. The current volume of information is huge and is relied upon to repeat more than multiple times in later stage of year 2014, out of which, some greater extent of percentage would be improved.
[1] Data warehouse concept Hadoop first evolved in data science technologies handling framework in 2006 at search engine companies. the thought depends on company's method of set of information and changes over it into another arrangement of information, which was first distributed by company dependent on their restrictive set of information and changes over it into another arrangement of information usage. In the previous few years, Hadoop has become a generally utilized stage and runtime condition for the sending of data applications
Recently, everybody understood that moving to Hadoop was the correct choice. [2] Everything considered, we could even contend this very choice was the one that spared, having seemed a couple of years back with its blindingly quick and negligible pursuit experience, was commanding the hunt advertise, while simultaneously, with its overstuffed landing page resembled a thing from an earlier time. Information about science and research were essentially offered the opportunity to play and investigate the world's information. Having recently been kept to just subsets of that information, Hadoop was invigorating. New thoughts sprung to life, yielding upgrades and new items all through, revitalizing the entire organization.
At this point other enormous, web scale organizations have just acquired a quick look at this new and energizing stage. [3] Around this time, Twitter, Facebook, LinkedIn and numerous others began accomplishing genuine work with Hadoop and contributing back tooling and structures to the Hadoop open source environment. In February, Yahoo! revealed that their creation Hadoop group is running on 1000 hubs.
In 2010, [4] there was at that point a tremendous interest for experienced Hadoop engineers. Still at, at the position of vice president Hadoop Software Engineering. That was a significant issue for major company! and after some thought, they chose to help in propelling another organization. All entrenched Hadoop (Project Management Committee) individuals, devoted to open source. For its unequivocal position that all their work will consistently be open source, Hortonworks got network wide recognition.
In 2012, [5] Hadoop group checks 42 000 hubs of Hadoop donors received at 1200. Before Hadoop got boundless, in any event, putting away a lot of organized information was dangerous. Money related weight of enormous information storehouses made associations dispose of insignificant data, keeping just the most important information. [6][7] Hadoop altered information stockpiling and made it conceivable to keep all the information, regardless of how significant it might be.
[8] Snowflake was established in 2012 in, by three information warehousing specialists: Benoit Dadeville, Thierry Cranes and Marcin Bukowski. Dadeville and Cranes recently functioned as information designers at Oracle Corporation; Bukowski was a prime supporter of the Dutch beginning up Vector wise. The organization's first CEO was a financial speculator at Ventures. The organization's name was picked as a tribute to the authors' adoration for snow sporty.

Snowflake came in 2014 not long after designating the cloud information distribution center turned out to be commonly accessible in June 2015and had 80 associations utilizing it around then. In 2019 and 2020, Hadoop Storage File System is dead, a direct result of its intricacy and cost and in light of the fact that process in a general sense can't scale flexibly in the event that it remains attached to Hadoop Storage File System. For ongoing bits of knowledge, clients need quick and versatile process limit that is accessible in the cloud. Information in Hadoop Storage File System will move to the most ideal and cost-proficient framework, be it distributed storage object stockpiling. Hadoop Storage File System will expire, however Hadoop figure will live on and live solid.
Recently, two mammoths of the enormous information Hadoop time, Cloudera and Hortonworks, detailed they would combine. The affirmation stated it would be a "merger of equivalents. It is enrapturing to see these two memorable pioneers getting together. They enabled associations to take up ventures that weren't conceivable already. They were at the center of the move from settling Tech issues to unravelling business issues, and business pioneers quickly appreciated the ability of these new innovations to pass on new administrations, activated by information, to their customers In 2019, [9] joined Snowflake to improve the disadvantage and improved scalability issues. Company chairman said in a question and answer session that Snowflake may seek after a first sale of stock as ahead of schedule as the mid-year of 2020, however that numerous components could change the planning.

Recently, Snowflake is Improving the item by concentrating on execution, security, and broadness of new capacities with the goal that Snowflake can be utilized by clients for a much more extensive arrangement of information stockroom remaining tasks at hand. Associate By and recursive are progressive Structured Query Language question sentence structures offered by on premises information distribution center arrangements, and are table stakes for enormous endeavors. These highlights further cement Snowflake situation as the Enterprise Data Warehouse worked for the cloud, and they make relocation from inheritance on premises information stockroom items a chance.

3. ARCHITECTURE

Hadoop gives a distributed document framework and a system for the investigation and transformation of huge data collections worldview.

3.1 HADOOP ARCHITECTURE

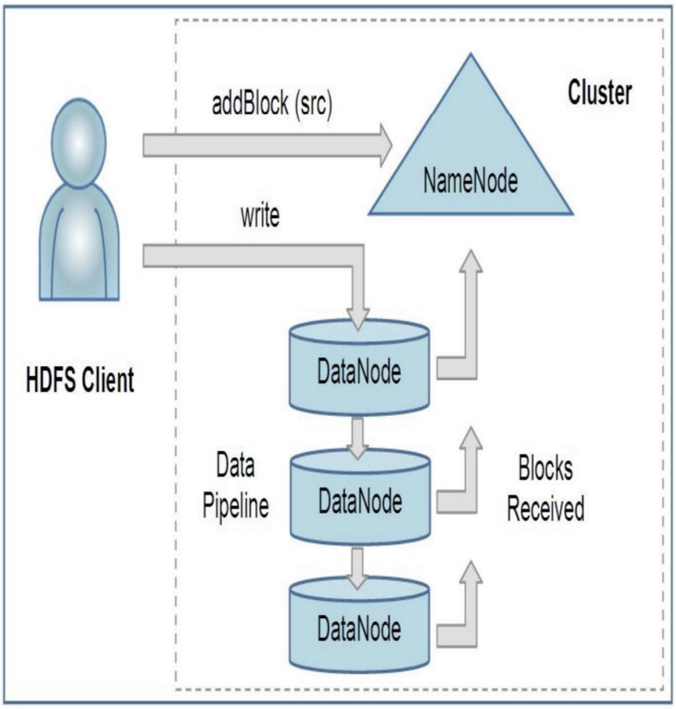

Fig -1: Hadoop Architecture

The figure 1 describes the Hadoop Architecture and each Nodes is explained as follows:

Name Nodes: Files and catalogs are a unit imagined on the Name Node by in hubs that record properties like authorizations, alteration and access times, namespace and plate space amounts. The record content is part into gigantic squares and each square of the document is severally recreated at numerous Data Nodes.

Data Nodes: Each square copy on a Data Node is imagined by 2 records inside the local host's local documenting framework. the essential record contains data itself and thusly the subsequent document is the square's information just as checksums for the square information and in this way the square's age stamp.

Hadoop Distributed File System Client: User applications get to the recording framework exploitation the Hadoop Distributed File System shopper, a code library that trades the Hadoop Distributed File System documenting framework interface.

Picture and Journal: The diary might be a compose ahead submit log for changes to the documenting framework that must be relentless.

Checkpoint Node: The Checkpoint Node periodically combines the prevailing stop associate degreed journal to form a replacement stop and an empty journal.
Backup Node: It keeps up an associate degree in memory, modern picture of the recording framework that is perpetually corresponding with the condition of the Name Node.

3.2 SNOWFLAKE ARCHITECTURE

Snowflake is a cloud information stockroom based on the Amazon Web Services (AWS) cloud foundation and is a genuine SaaS offering. There is no device (virtual or physical) for to pick, present, organize, or administer. There is no item for you to present, organize, or manage. All nonstop upkeep, the administrators, and tuning is dealt with by Snowflake.

Snowflake is intended to be an enterprise prepared assistance. Other than offering high degrees of convenience and interoperability, undertaking status implies high accessibility. To this end, Snowflake is an assistance arranged engineering made out of exceptionally shortcoming open minded and freely adaptable administrations. These administrations impart through Application Programming Interface interfaces and fall into various compositional layers:

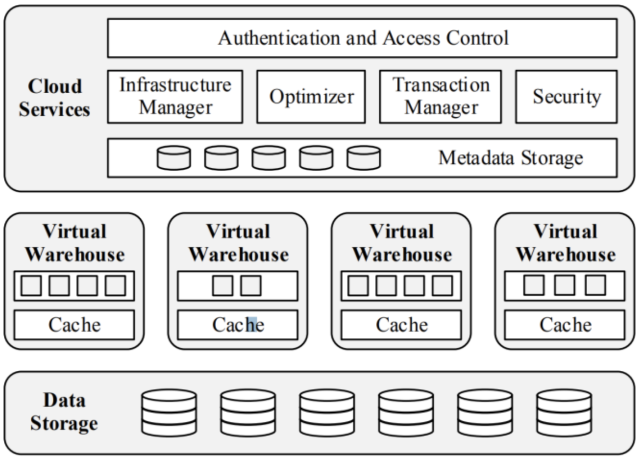

Fig -2: Snowflake Shared data Architecture

The figure 2 describes Snowflake Shared data Architecture. Architecturally, there are three components that made up to the Snowflake data warehouse and it is briefly explained as below:

Data Storage: This layer utilizes S3 to store table information and query results. The basic record framework in Snowflake is upheld by S3 in Snowflake's record, all information is scrambled, compacted, and dispersed to advance execution. In Amazon S3 the information is geo-repetitive and gives brilliant information toughness and accessibility.

Virtual Warehouses: This layer handles query results inside flexible groups of virtual devices, called virtual distribution centers. Snowflake gives the capacity to make "Virtual Warehouses" which are fundamentally register bunches in EC2 that are provisioned off camera. Virtual Warehouses can be utilized to stack information or run inquiries and can do both of these errands simultaneously. These Virtual Warehouses can be scaled up or down on request and can be delayed when not being used to diminish the spend on register.

Cloud Services: This layer is an assortment of administrations that oversee virtual distribution centers, questions, exchanges, and all the metadata that circumvents that: database compositions, get to control data, encryption keys, utilization insights, etc. Coordinates and handles every single other assistance in Snowflake including meetings, validation, SQL arrangement, encryption, and so forth.

4. COMPARISON OF SNOWFLAKE OVER HADOOP

Table -1. Comparison of snowflake and Hadoop

Features	Hadoop	Snowflake
Hardware Option	Hadoop can run on modest product servers with legitimately joined capacity. It does anyway should be conveyed on premises or appointed in the cloud.	Snowflake is a product as-service stage and requires no equipment buy. Basically, load your information and inquiry.
Data Volumes	Hadoop will effectively oversee many terabytes and scale to several petabytes	Some Snowflake clients have single tables more than a petabyte in size.
Tools Supported	Mainly open source with some help for third party apparatuses utilizing Open Database Connectivity and Java Database Connectivity.	A broad cluster of information the board and business insight instruments, numerous with committed interfaces.
Deployment Complexity	Extremely high. Needs profoundly gifted help and frameworks the board.	Extraordinarily easy. Near to zero administration overhead with no records or measurements to oversee.
Data Quality	Not better quality	Better quality
Query Latency	The Hadoop Architecture implies there will be complication and an extra overhead on queries.	Snow flake queries don't give poor solution and can run in few milliseconds.
Minimum Data Size	Although no physical least exists. Hadoop little tables (under 1Gb) ought to be kept away from where conceivable and Hadoop doesn't function admirably with little records.	Snowflake usually Supports queries on information which varies from some kilobytes up to few petabytes.

5. MIGRATION FROM HADOOP TO SNOWFLAKE:

With various advantages of snowflake over Hadoop, many companies are in the process of migration from Hadoop to snowflake. It is less expensive, provides greater scalability and maintains a codebase. The migration from Hadoop to Snowflake must be arranged cautiously and in a precise manner to be effective. Now and again, the movement should deal with advancing things that were not working or working sub ideally in the current arrangement. Likewise, the system for one-time relocation versus gradual synchronizations can be totally extraordinary. There are some set out a methodology that has to follow for one of the relocations.
Any kind of relocation needs arranging and execution through stages. Right now, are following bellow Phases.

1)Discover
2)Demonstrate
3)Deploy

1)Discovery

This is the most vital stage where one needs to extricate and examine the stock of existing procedures. The different exercises incorporate distinguishing the choices accessible for one-time movement versus gradual synchronization of information in the information distribution center. The necessities and information to get examples of the shoppers should be concentrated cautiously to decide the ideal information model in snowflake. The information word reference and mapping of the fields and the datatypes to the objective snowflake model should be worked out. This is the chance to institutionalize the naming shows of the different information objects. Kafka is the bearer of the information from the source to Hadoop. There are in excess of fifty unique articles that get shipped to Hadoop. Different purchasers at that point expend from the Hadoop source.

2)Demonstrate

This is where proof of idea for every one of the choices that have chosen for both one-time relocation and gradual synchronization of information. From that point the records are replicated into the interior stage and afterward to the last tables. There is a requirement for some change before the information gets replicated to the last tables. The python-based arrangement additionally follows a comparable methodology however with Extract, Transform and Load, parcel of functionalities and changes come inbuilt and it quickens the relocation. For steady synchronization, snow pipe and stream with errands alternatives were assessed. Stream is a superior alternative as it assists keep with following of the counterbalances and furthermore the ingestion from stream to definite tables is a finished nuclear exchange. Any disappointment in that pipeline won't erase the changed record in the stream object. The stream object gets flushed just when the ingestion to the last table is fruitful. In any case, both stream and undertakings are in sea mode and are relied upon to be at the end of next quarter.

3)Deploy

Eventually the reality is sending. The arrangement stage is in progress for us. However, one of the most significant actions of the organization procedure is to assemble an approval structure that can approve the fruitful movement of information from source to target. The approval must not limit itself to only approval of number of lines among source and target, it should likewise deal with certain information quality checks. The arrangement ought to follow a gradual methodology where less hazardous items are moved before the least secure ones. This helps in feeding back the learnings to the process and make it better in each phase.

6. CONCLUSION

In this paper, we have discussed the history of Hadoop and snowflake, the origin of Hadoop and snowflake and its development in industry. Also explained about modern data warehouse briefly. Hadoop was used as a data warehouse for a decade because of its less complex structure. Also, there was talk in industry about the downfall of Hadoop. Snowflake was a modern data warehouse which is currently used in many industries. Many disadvantages of Hadoop are solved by snowflake. The architecture of Hadoop and snowflake were discussed. Migration of Hadoop to snowflake is made due to the latest trend of snowflake which is known to be the best data warehouse.

REFERENCES

[1]. L. Jing-min, H. Guo-hui, "Research of Distributed Database System Based on Hadoop", IEEE International conference on Information Science and Engineering (ICISE), pp. 1417-1420, 2016.

[2]. K. Shvachko, H. Kuang, S. Radia, R. Chansler, "The Hadoop Distributed File System", IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Sunnyvale, California USA, vol. 10, pp. 1-10, 2016

[3]. J. Bhimani, J. Yang, Z. Yang, N. Mi, Q. Xu, M. Awasthi, R. Pandurangan, and V. Balakrishnan, "Understanding Performance of I/O Intensive Containerized Applications for NVMe SSDs," in 35th IEEE International Performance Computing and Communications Conference (IPCCC). IEEE, 2016.

[4]. K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on IEEE 2017, pp. 1–10.

[5]. R. Dua, A. R. Raja, and D. Kakadia, "Virtualization vs containerization to support PaaS,"in Cloud Engineering (IC2E), International Conference on. IEEE, 2018, pp. 610–614.

[6]. Debajyoti Mukhopadhyay, Chetan Agrawal, Devesh Maru"Addressing NameNode Scalability Issue in Hadoop Distributed File System using Cache Approach" International Conference on. IEEE, 2018, pp. 310–314.

[7]. Y. Pingle, V. Kohli, S. Kamat, N. Poladia, "Big Data Processing using Apache Hadoop in Cloud System", National Conference on Emerging Trends in Engineering & Technology, pp. 475- 479, 2017.

[8]. Benoit Dageville, Thierry Cruanes, Marcin Zukowski, Vadim Antonov, Artin Avanes,
Jon Bock, Jonathan Claybaugh, "The Snowflake Elastic Data Warehouse" International Conference on. IEEE, 2016, pp. 215-226.

[9]. H. Gao, Z. Yang, J. Bhimani, T. Wang, J. Wang, B. Sheng, and N. Mi, “AutoPath: Harnessing Parallel Execution Paths for Efficient Resource Allocation in Multi-Stage Big Data Frameworks in Snowflake,” in the 26th International Conference on Computer Communications and Networks (ICCCN). IEEE, 2017.

[10]. Lucas, F.J., Molina, F. and Toval A, “A Systematic Review of Consistent UML Model Development and Management”, in Journal of Information and Software Technology, Vol. 51(12), pp. 1 – 15, June 2017.

[11]. T. DeMarco, “Structured Analysis and System Specification”, in Yourdon Press Computing Series, Prentice Hall, Englewood Cliffs, New Jersey, Volume.50, No.3, pp. 168–196, Jan 2018.

[12]. Bache R, Mullberg M, “Measures of Testability as a Basis for Quality Assurance”, in Software Engineering Journal, pp. 86-92, March 2016.

[13]. B. Beizer, “Software Testing Techniques”, in International Journal of Scientific and Research, vol. 2, Issue 2, pp. 547-589, Feb 2016.

[14]. Basili, V.R, Selby R.W, “Comparing Effectiveness of Software Testing Strategies”, in IEEE Transactions SE, vol. 13, pp 1278-1296, May 2019.