[5]. Dageville B, Cruanes T, Zukowski M, Antonov V, Avanes A, Bock J, Claybaugh J, Engovatov D, Hentschel M, Huang J, Lee AW. The snowflake elastic data warehouse. InProceedings of the 2016 International Conference on Management of Data 2016 Jun 14 (pp. 215-226).

The problem being addressed?

The rise of cloud computing has provided a plethora of opportunities for handling huge computing workloads. This has triggered many organizations to innovate more SaaS (Software as a Service) products, which enabled the usability of enterprise-class systems for all ranges of customers. Similarly, data warehousing community was in need of a SaaS product that helps data developers to seamlessly integrate data and use it for analysis. Therefore the authors studied

the problem of classical on-premises data warehousing solutions and proposed an alternative to it.

Limitations of Existing work?

According to the authors, the integration of multiple sources and systems has led to varied data formats, which triggered huge computing needs to process the incoming data into the system. Traditional data warehousing solutions were not able to support and scale as per the rising demands due to hardware limitations. While some parts of the data warehousing community pivoted to big data platforms such as Hadoop and Spark to create data warehouses, these solutions required significant engineering effort to roll out and handle a cluster. In addition the existing systems did not provided support for varied data types. All of these existing problems demonstrate the need for a new system.

Summary of Proposed Strategy?

The authors developed a novel columnar database engine named Snowflake, which consists of 3 layers : a) cloud services layer, b) the virtual warehouse layer, and c) data storage layer.

Cloud Services layer:

The cloud services layer is considered as the brain of a unary snowflake environment, because it contains crucial software services such as a query optimizer, a transaction manager, an infrastructure manager, a metadata storage, and security services.

1) A query optimizer, uses a cascade-style approach, following top-down cost-based optimization. It optimizes the submitted query to the engine by reducing its search space due to lack of indexing and uses delayed evaluation. Similarly, once the optimizer completes its work, the resulting execution plan is distributed across worker nodes, and the cloud services continuously monitor the state of the query.

2) A transaction manager handles concurrency control, and ensures ACID transactions by using Snapshot isolation. It internally uses MVCC (Multi-Version concurrency control, a protocol where a copy of every amended object is stored until a certain period of time) based engine, that enables time travel and cloning whenever required.

3) The infrastructure manager monitors and provisions virtual warehouses as a typical compute cluster in Snowflake which has primary memory and disk storage to process the incoming query.

4) The metadata component tracks the statistics of all tables, and stores them as min-max ranges.

5) The security services layer, oversees the security aspects of the objects stored in Snowflake.

Virtual Warehouse Layer:

1) The virtual warehouse layer leverages the power of compute and storage in a loosely coupled manner, and uses the convention of t-shirt size such as X-small to XX-Large.

2) A proprietary shared nothing layer caches query results data onto local disks.

3) The query results are frequently accessed, over the network, and this is defined as a multi-cluster shared data environment (when a compute cluster has multiple nodes in it).

4) The query results are accessible until the virtual warehouse is up and running. Once a virtual warehouse is suspended then the query results are flushed out from the virtual warehouse.

Data Storage Layer:

1) For persistent table storage snowflake uses existing cloud object stores like S3 (Simple Storage Service).

Summary of Methodology and Result of Evaluation?

The authors mentioned that Snowflake is purely offered as a SaaS product where users are not needed to tune databases or perform physical backups of data because all of these activities are taken care by the engine itself. This not only enhances the user experience but also improves usability. Also, storage services are highly available with a guarantee between 99.99% and 99.999999999% of data durability. In addition, Snowflake extensively supports semi-structured data using the variant data type. It consists of many helper functions that can potentially transform the data in a structured way. Overall the authors found that, the test performance of 22 standard queries on Snowflake using relational and semi-relational data is quite acceptable, but for few queries it took significantly longer time due to sub-optimal join order.

Any Research Directions Provided?

The authors provided various research directions, which include novel query processing methodologies that utilize multi-cluster shared data architecture. Also, they leveraged the power of cloud computing and provided more details regarding MVCC architecture and time travel, cloning. Overall the authors presented a cloud-native based data warehousing solution that has the ability to scale on demand, and briefly discussed its replication and query processing techniques.