**What level of AWS, Azure or GCP expertise is required?**

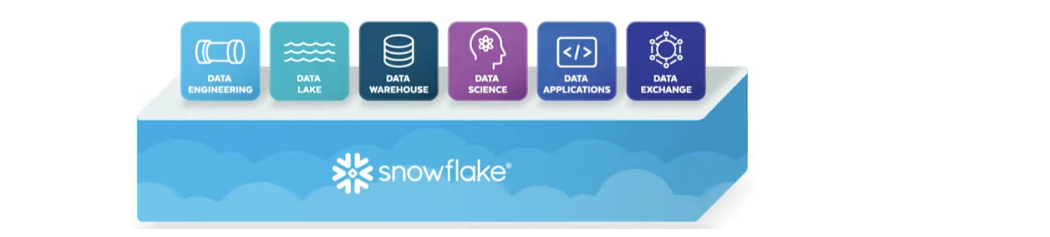
Basic level only cloud storage services (AWS S3, Azure Blob storage and Google cloud storage) will be used primarily for data storage and that will be covered as a part of this session

**Snowflake Purpose**

Snowflake is Ideal for below purposes

Snowflake is a fully managed SaaS (software as a service) that provides a single platform for data warehousing, data lakes, data engineering, data science, data application development, and secure sharing and consumption of real-time / shared data.

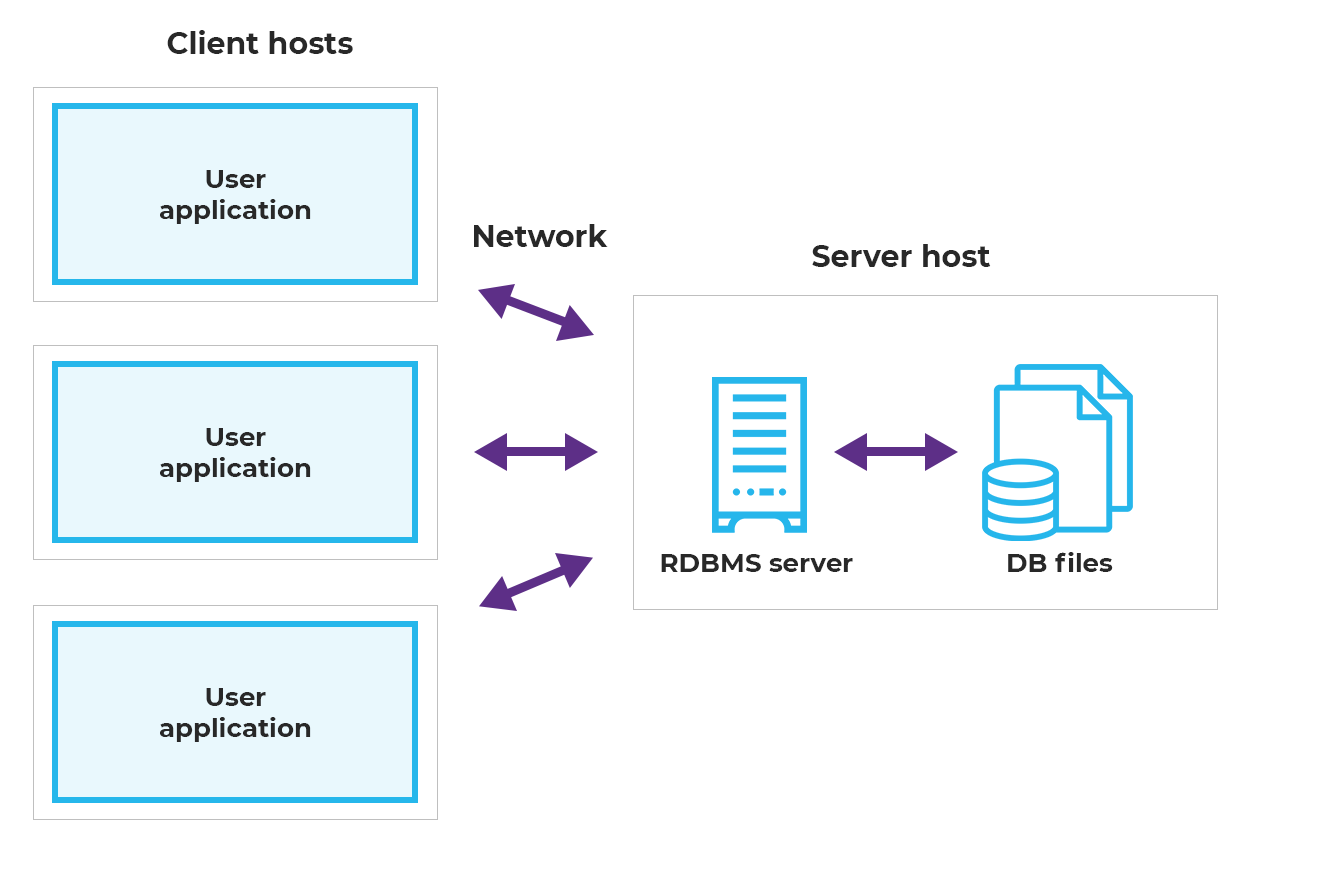
* **Data Warehouse (primary)**
* **Data Lake (primary)**
* Data Exchange
* Data Apps
* Data Science
* Data Engineering
* UNISTORE (OLTP )

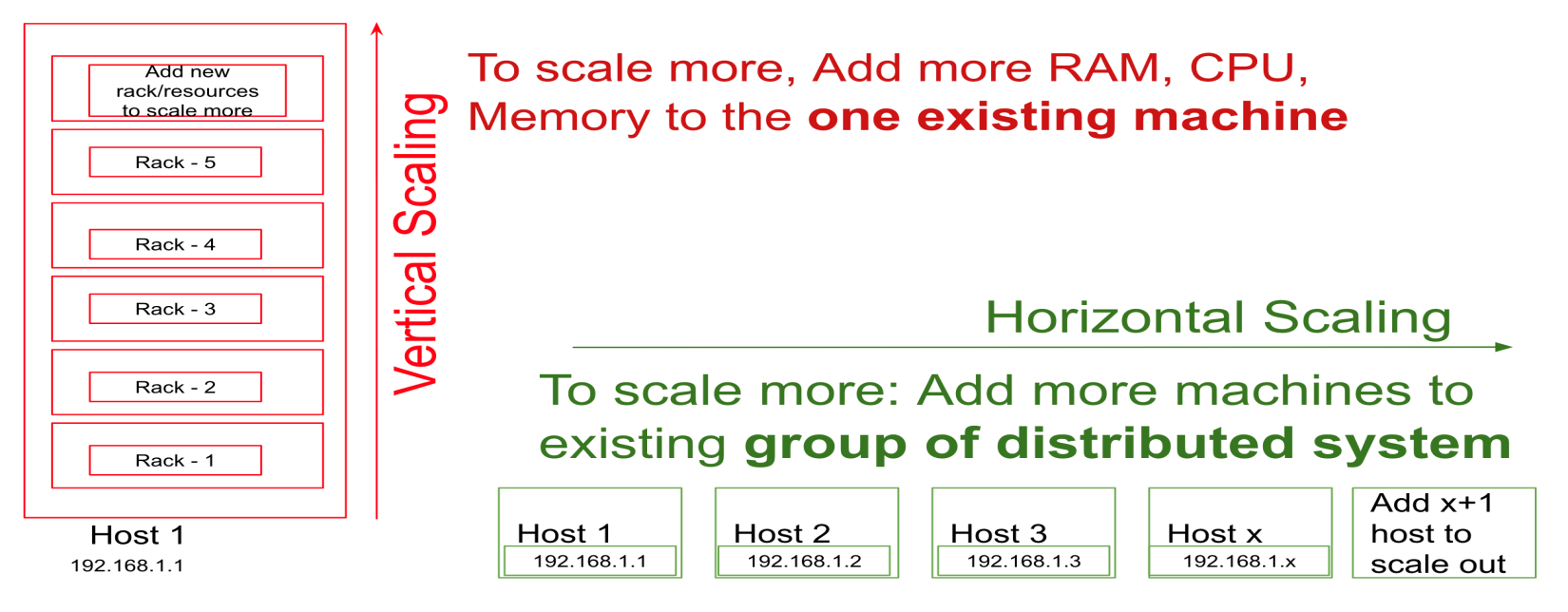


**Relational Databases (RDBMS)** (Oracle, SQL Server, MySQL, PostgreSQL…)

#### Relational databases are designed to run on a single server in order to maintain the integrity of the table mappings and avoid the problems of distributed computing.

Machine/node => computer

**Shared Disk Architecture**



**Scaling vertically** ===> One big machine will do all the work for you.

**Scaling horizontally** ===> Thousands of machines will do the work together for you.

RDBMS can be scaled vertically but not horizontally

In order to run a query on a database it would require compute resources like Processor, RAM

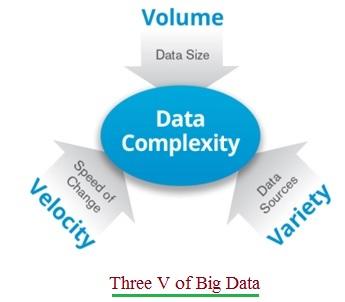
**If Data volumes/users increases we will have to face performance issues since there is limit to scale a machine/computer vertically.**

**There is limit to increase capacity of one machine’s RAM, CPU & Memory**

**Limitations of relational databases**

1. **Scalability, performance and speed**
2. Licensing cost and maintenance over head
3. Concurrency issues (can’t handle large number of users at a point of time)
4. Limited/No support for Semi structured and unstructured data
5. Database Failure
6. Up gradation Costs

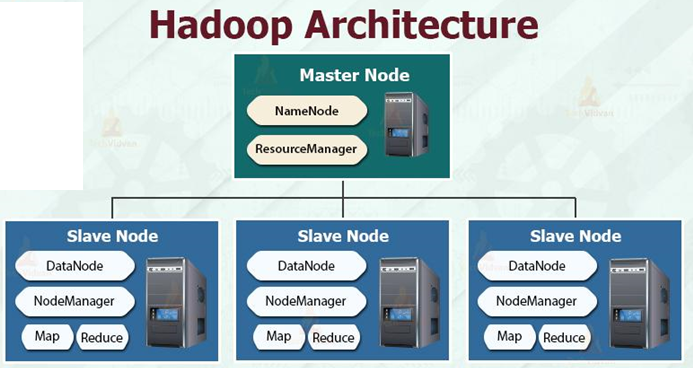
### What is Big Data?

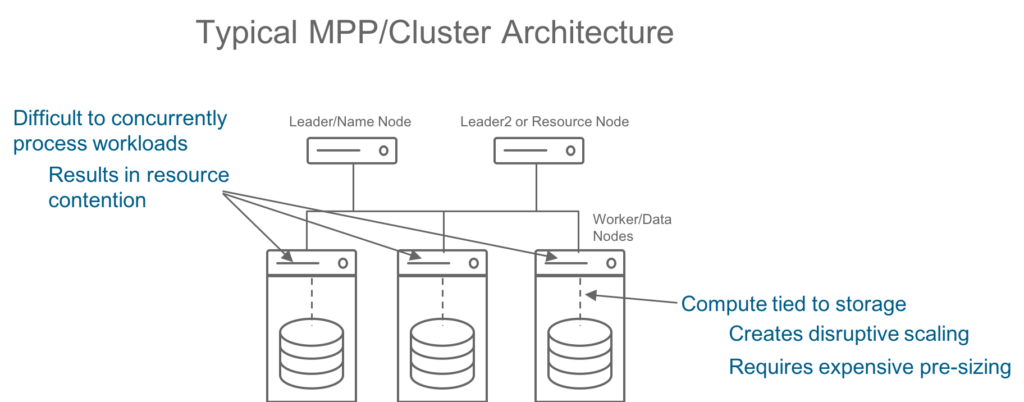


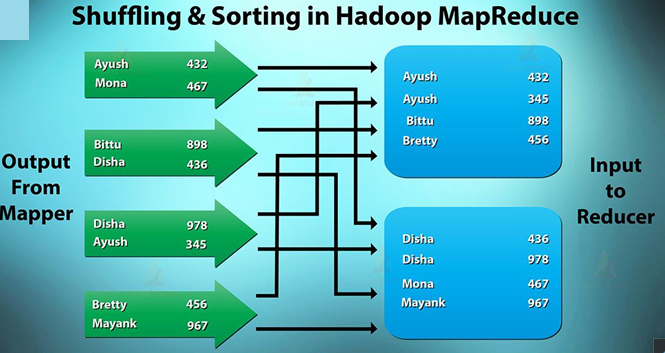
•The word "Big" in big data not just refers to data volume alone. It also refers to the fast rate of data origination, its complex format and its origination from a variety of sources. The three V's of big data are Volume, Velocity and Variety.

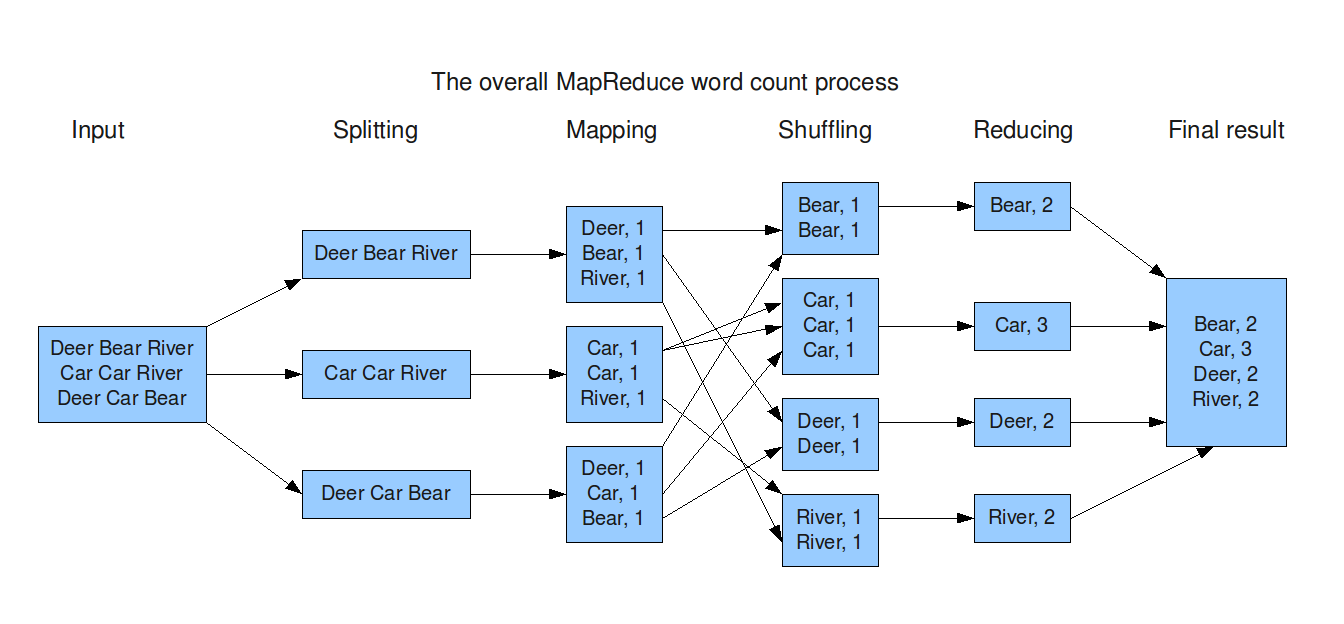
Hadoop architecture consists of two layers.  
• MapReduce a Processing/Computation layer  
• Hadoop Distributed File System (HDFS) as Storage layer (all the data will be distributed across multiple slave nodes/machines as 64/128 MB blocks/files) and master node manages all the data distribution, retrieval, data replication and metadata information

Shared nothing (data and compute will be done at each slave node level)









**Disadvantages of Hadoop**:  
➨It is not suitable for small and real time data applications.  
➨Joining multiple data set operations are complex.   
➨Data retrieval will be slow. Since it has to get data data from multiple slave nodes which involves lot of shuffling and sorting of data that degrades performance

➨It does not have storage or network level encryption.  
➨Cluster management is hard i.e. in cluster, operations like debugging, distributing software, collection logs..  
➨When operated by a single master it will cause difficulty in scaling.  
➨Programming model is very restrictive.

**Snowflake Architecture**

HYBRID OF SHARED-DISK & SHARED-NOTHING architectures

Snowflake’s architecture is a **hybrid of traditional shared-disk and shared-nothing** database architectures.

* **Similar to shared-disk architectures** → Snowflake uses a **central data repository** for persisted data that is **accessible from all compute nodes in the platform**.
* **Similar to shared-nothing architectures** → Snowflake processes queries using virtual warehouses where **each node in the cluster stores a portion of the entire data set locally**.



Data is stored in the cloud storage and works as a shared-disk model thereby providing simplicity in data management

For compute it will take advantage of performance and scale-out benefits of shared nothing architecture

This approach offers the **data management simplicity of shared-disk architecture**, along with the **performance and scale-out** benefits of a shared-nothing architecture.

Snowflake architecture allows storage and compute to scale independently, so customers can use and pay for storage and computation separately

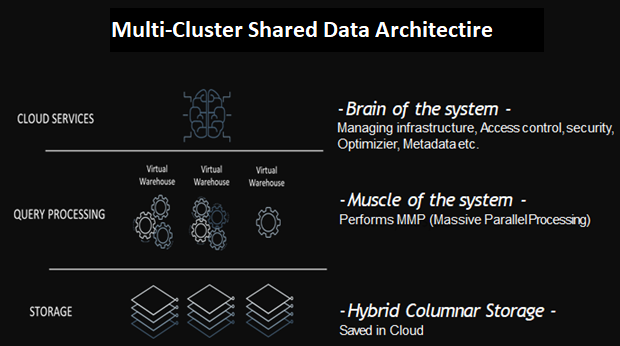
While data is a core asset for modern enterprises, technology’s ability to scale (cheaper storage) has created a surge of big data.

Managing and storing that data has become a critical function for modern business operations.

Most enterprises are already using a cloud data platform, but many are evaluating whether a data migration might be needed in order to stay competitive

**Snowflake Architecture**

**Multi-Cluster Shared Data Architecture**



SNOWFLAKE LAYERS

Snowflake’s unique architecture consists of three key layers, all of them with High Availability. **The price is also charged separately for each layer**.

Each layer can independently scale : storage, compute, and services.

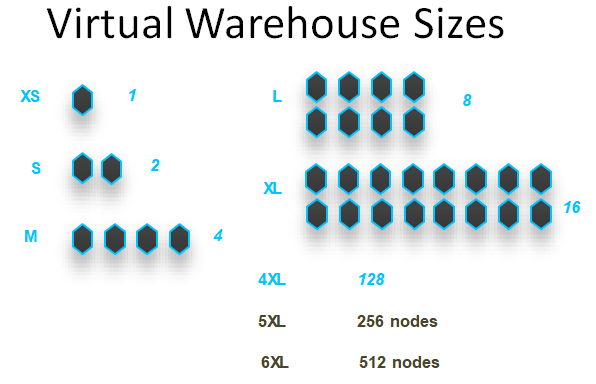
1. **Centralized Storage →**When data (structured. semi structured) is loaded into Snowflake, Snowflake reorganizes that data into its internal optimized, compressed, columnar format in this layer.

Snowflake manages all aspects of how this data is stored like organization, file size, structure, compression, metadata, and statistics. This storage layer runs independently of compute resources.

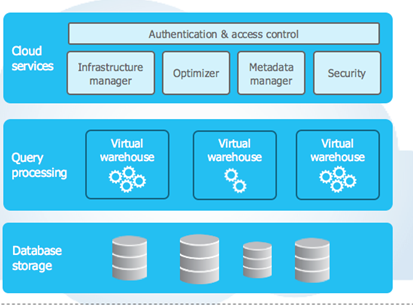
1. **Compute** → The compute layer is made up of virtual warehouses that execute data processing tasks required for queries. Each virtual warehouse (or cluster) can access all the data in the storage layer, then work independently, so the warehouses do not share, or compete for, compute resources.

**Virtual warehouse = cluster of compute nodes/machines**

Query execution is performed in the compute/processing layer using “virtual warehouses”. Each virtual warehouse is an MPP (Massive Parallel Processing) compute cluster composed of multiple compute nodes allocated by Snowflake from a cloud provider.



\*5XL and 6XL are in preview state



1. **Cloud Services** →

The overall brain in the system, this layer is a collection of services that handle query management, optimization, transactions, security and governance, metadata, and sharing and collaboration

The cloud services layer uses ANSI SQL and **coordinates** the entire system. It eliminates the need for manual data warehouse management and tuning Collection of services that coordinate activities across Snowflake. It includes:

* Authentication. (user logins)
* Infrastructure management. (Assigning compute resources..)
* Metadata management. (Table structures, columns, micro partition details)
* Query parsing and optimization (query performance optimization, execution plan)
* Access control. (Access Management)

**Cloud Agnostic Layer** → there is another fourth layer also, known as the Cloud Agnostic Layer. It is used only the first time when we choose a cloud provider.

 “Cloud Agnostic” is generally regarded to refer to applications and workloads that can be moved seamlessly between cloud platforms

Snowflake’s architecture allows flexibility in handling big data.

Snowflake decouples the storage and compute functions, which means organizations that have high storage demands but less need for CPU cycles, or vice versa, don’t have to pay for an integrated bundle that requires them to pay for both.

Users can scale up or down as needed and pay for only the resources they use.

Storage is billed by terabytes stored per month, and computation is billed on a per-second basis.

Warehouse Billing

30 secs – 1 minute

45 secs – 1minute

61 secs – 61 secs

65 secs – 65

90 secs – 90 secs

Snowflake account creation

Its Multi-cluster Shared Data architecture

