AWS RedShift

Why RedShift ?

With digital data growing at incomprehensible way enterprises are finding it difficult to ingest, store and analysis data quickly while keeping the cost low.. with that cause they moving data warehouses to cloud

Why House Data in a Data Warehouse?

A Data Warehouse provides an organization with analytics, deep querying, and reports.

They feature performance and scale not possible with "traditional" Relational SQL databases used in day to day operations

Relational SQL Databases

A relational SQL database excels at storing and retrieving "real time" operational data. In technical terms, a good relational database provides ACID guarantees to data storage:

- **A - Atomicity**, which means "all or nothing". The classic example is a banking transaction. When transferring money between accounts, both accounts should update (or neither should).
- **C - Consistency**, which means "the database is always in good shape". Any operation has to leave the database in a stable state, without any half-baked writes.
- **I - Isolation**, which means "even if several things are happening at the same time, they all succeed". In other words, if there are many cooks in the kitchen, their dishes all turn out perfectly, as if cooked one at a time.
- **D - Durability**, which means "once something completes, it stays completed". The database writes updates to disk. Losing power doesn't matter. And so on.

Relational SQL Databases

- While the above is critical to business software, it forces design decisions ill-suited for data analysis queries. These queries are expensive:
- They search across long time periods.
- They group data according to deeply buried characteristics.
- They join disparate data points to find correlations and trends.
- Theoretically, relational databases can support queries like this, but they struggle with performance. This means slow reports, and can mean detrimental impact to "real time" operations. Unacceptable!

Data Warehouses

- Data Warehouses began appearing in the 1980's as an attempt to solve the problems relational databases couldn't. In a nutshell, they addressed a few problems:
- Large organizations have a multitude of relational databases, and need to run reports across them in a unified fashion.
- Reporting can involve data across several tables, necessitating the use of (expensive) joins to unify the data.
- Expensive queries can kill operational databases.

A data warehouse solution solves these problems by:

- Merging the data from one or more relational databases
- Normalizing ("lumping") the data together in ways that support deep querying.

What is Data Warehouse ?

- You can think of Data Warehouse as a repository that data generated from your organization operational systems and many other external sources is collected, transformed and stored.
- You can host this Data Warehouse on your organization main frame server or on Cloud.

What is Data Warehouse ?

- A data warehouse is a subject - oriented, integrated, time variant and non-volatile collection of data in support of organizations decision making process

Moving to Cloud …

Companies are moving from on- premises to cloud..

Do you know why?

Because of its underline architecture and disadvantages

Traditional Data Warehouse Architecture

Where is the data comes from?

Traditionally data sources are divided into 2 groups

Internal data: data which is generated and

consolidated with in organization from different departments

External data: data which is not generated

with in organization which means it come from external sources

DW follows simple 3 tier architecture

1. Bottom tier

In bottom tier.. We have a warehouse database server or you can say Relation database system.. in this tier using different kind of back end tools and utilities we may extract data from different sources and cleanse the data and transform it before loading it into Data warehouse

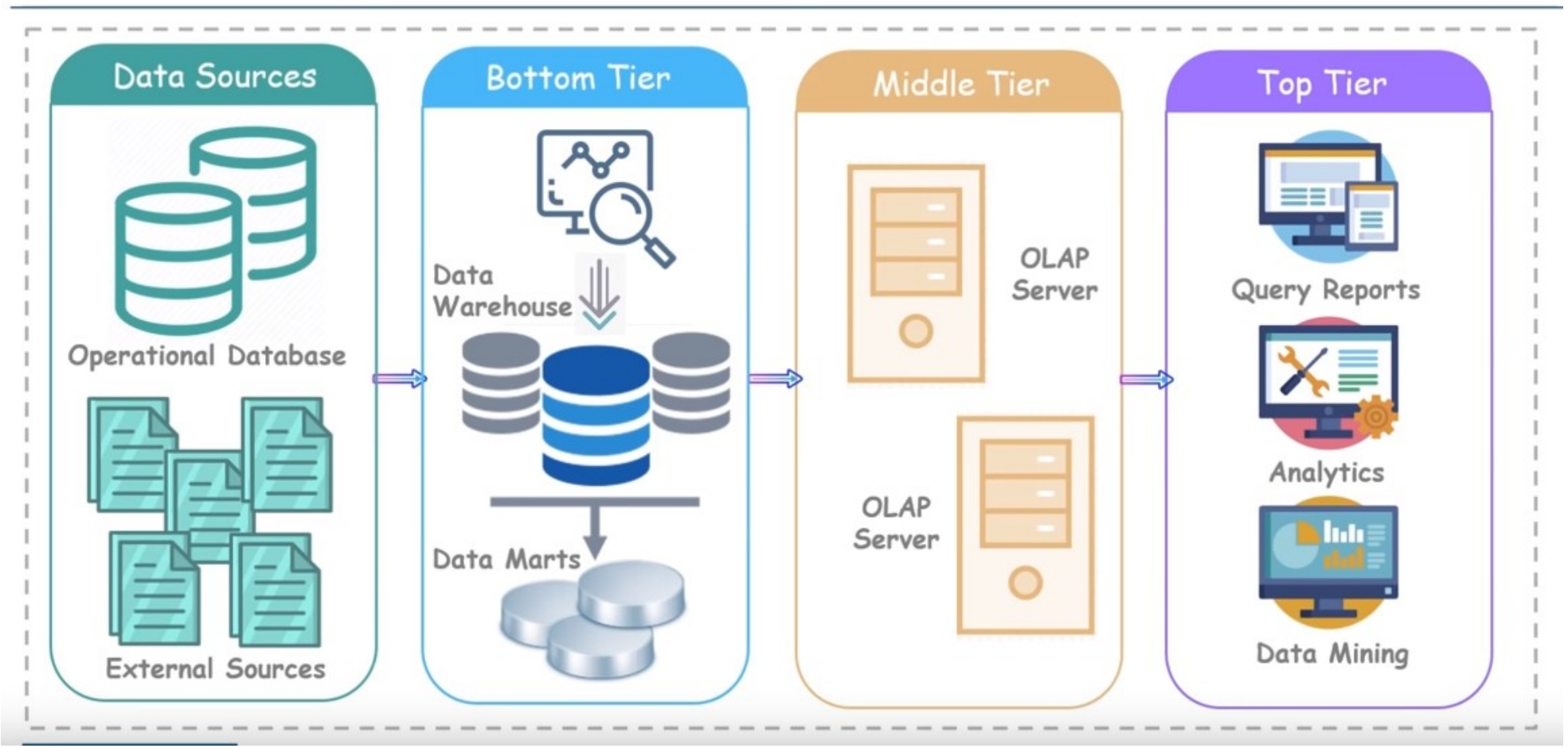
2. Middle Tier

we have OLAP server (online analytical process). This OLAP performs multi dimensional analysis of business data and transform the data into a format such that we can perform complex calculations, analysis and data modeling on this data

3. Top tier

This is like a front end client layer, this tier holds different kind of queries and reporting tools using which the client application can perform data analysis, query reporting and data mining.

Architecture of Traditional Data Warehouse



Summary of Traditional Data Warehouse

So, In summary in traditional data warehouse has simple 3 tier architecture

Bottom: we have backend tools using which we collect and cleanse the data

Middle: we have tools which is OLAP server using which we transform into the data the way we want

Top tier: in which we use different query and reporting tools we can perform data analysis and data mining.

Disadvantages of Traditional Data Warehouse

Business Services Company

Difficult to set-up, deploy & manage

Difficult to scale up or down

Performance Issues

Spiralling costs

Example of Traditional Data Warehouse

There is this leading US business service company which is running a commercial enterprise DW. This DW has data coming from different sources across different regions.

The first problem this company faced was setting up traditional data warehouse.

As we discussed earlier architecture of traditional is not so easy.. it consists of data models which extract data, transform load processes which is called ETL, BI tools sitting on top...

So, this company has to spend lot of money on resources to setup a traditional dw, dw which is initially 5TB which is growing high and higher and it was expected that there is a lot of growth in future... so to meet this continuous need, company wants to upgrade the hardware, again upgrading these hardware wants more resources and more money...

And auto scaling in traditional dw is not an easy job

Example of Traditional Data Warehouse

Company could not meet all these storage and compute needs easily it was facing a lot of performance issues as well and finally company had to deal with increasing costs...

Initially it has to spend a lot on setting up data warehouse. Like has to spend on hardware, manpower, electricity, security and real estate and deployment cost and many other

As data warehouse grew they have to spend again to meet storage and compute needs

So to sum it up.. setting up data warehouse and deploying it and managing it later and was lot of money and resources and more over auto scaling in traditional data warehouse is not easy concept

Because of all these reasons companies are moving to towards cloud based warehouse instead of traditional on premises...

AWS RedShift

Amazon Redshift

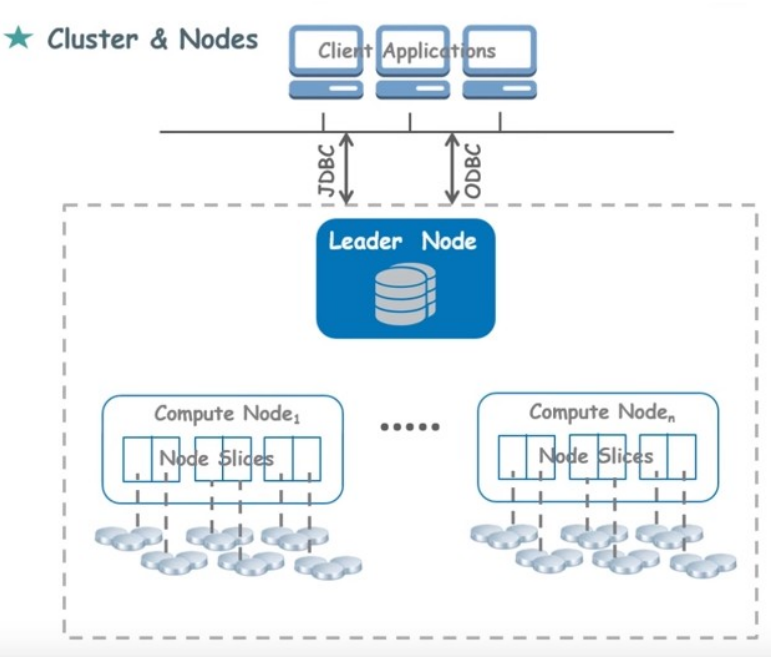
Amazon redshift is a cloud bases data warehouse provided by AWS

Redshift is a fast scalable data warehouse that makes it simple and cost effective to handle all the data

It is massively parallel, column-orientated database deployed on the AWS platform that makes it simple and cost effective to analyze all your data across your data warehouse and data lake

Key Concepts

- Amazon Redshift data ware is a collection of compute resources which we call **NODES**
- These nodes then organized into groups they become **CLUSTERS**
- Each of these clusters run amazon redshift engine and it contains one and more databases
- This cluster has **leader node** and one or more computes nodes

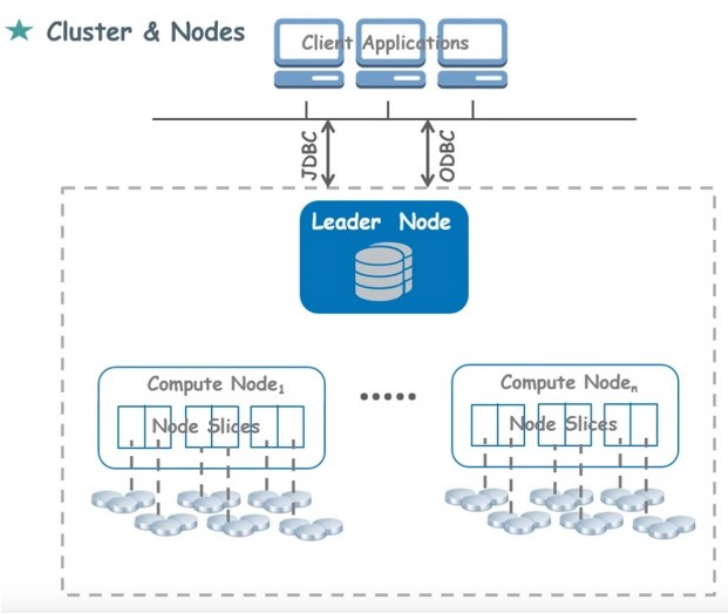


Key Concepts

- As for the leader node it receives queries from client applications and then it passes these queries and develops a suitable query execution plan

- And then it co-ordinate the parallel execution of these plans with one or more computes nodes

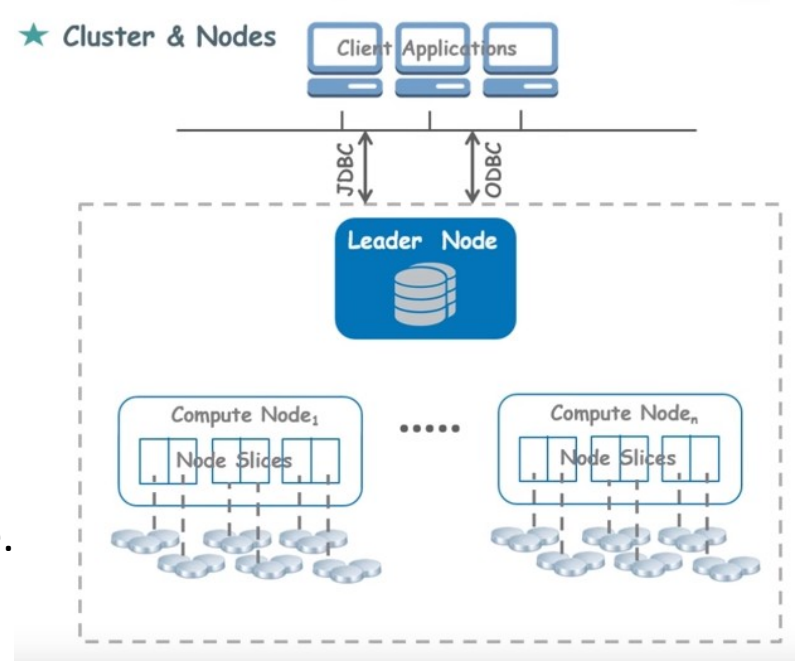
- Once the compute nodes finishes executing these plan again the leader node aggregate the results from all this intermediate compute nodes and then sends it back to client applications



- **Compute nodes**: you can think of these compute nodes as a compute resources that execute the query plan which was develop by leader node.

Key Concepts

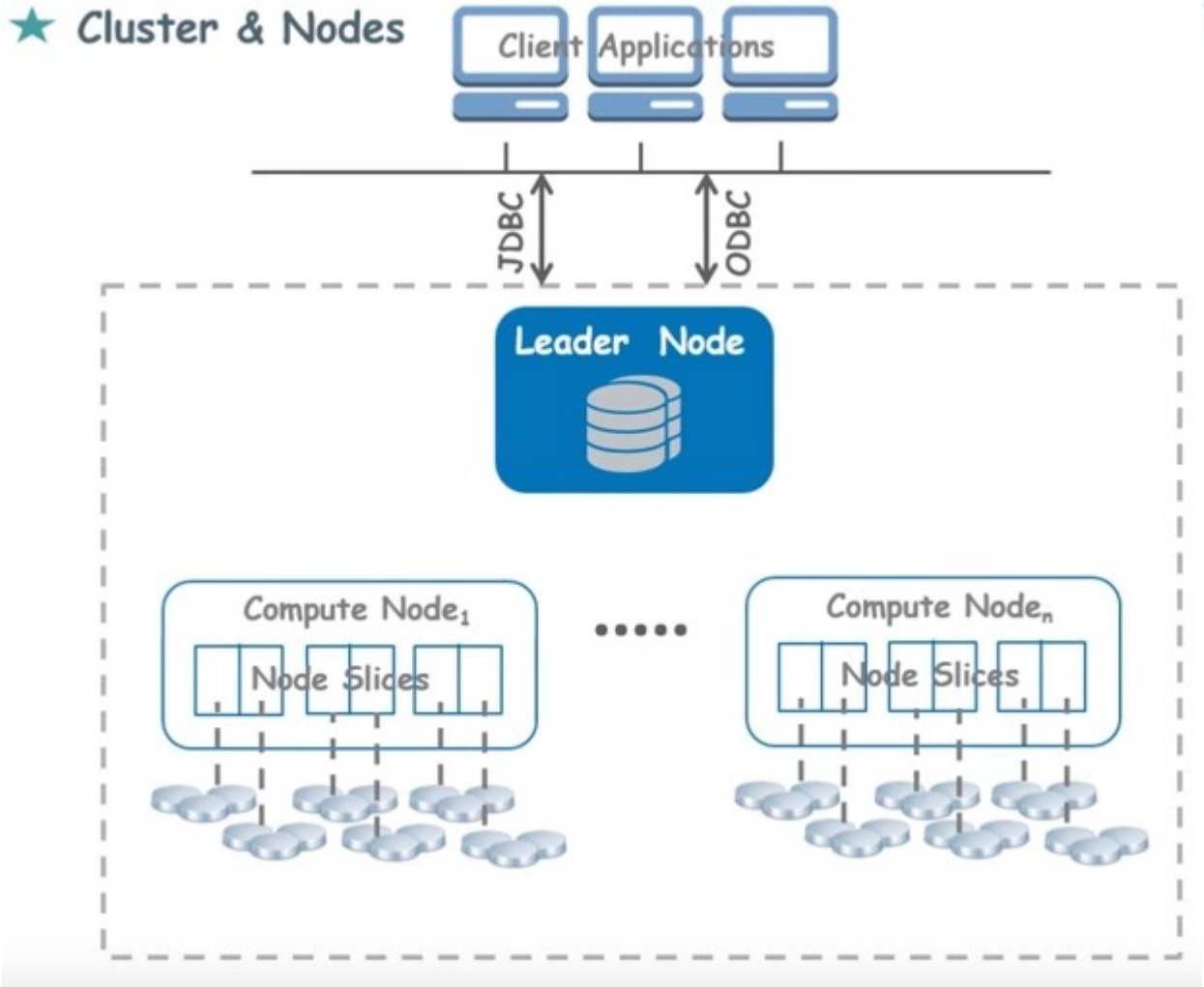
- And when they are executing these plan they transmits data among themselves to solve many queries



- These computes nodes further divided into slices which we called **node slices**..
- Each of these node slices receive part of memory and disk space. so the leader node distribute data and part of user queries that is received from client application to this node slice
- And all these node slices work in parallel to perform operations and increase the performance of your redshift data warehouse.

Key Concepts

★ Cluster & Nodes



★ Nodes

Dense Storage(DS)

Dense Compute(DC)

Node you choose depends on:

1. Data Quantity
2. Complexity Of Queries
3. Downstream Systems

Node Types

There are two types of nodes

- **Dense Storage (DS):** These are storage optimized and they are used to handle huge data workloads and basically they use hard disk drive or HDD type of storage

- **Dense Compute(DC)**: these are compute optimized they use to handle high performance intensive workloads and they mainly use solid state drive or SSD type of storage..

How do they interact with client applications?

I have client applications like BI tools, analytical tools which communicate with AWS redshift using drivers like JDBC (java)and ODBC (sql)

Using these drivers client applications sends a query to leader node

Leader nodes on receiving the client query it passes these queries and develop the suitable execution plan..

Once the plan is setup ..compute node and slices will start to work on this plan and transmit data among themselves to solve these queries so once the execution is done leader node aggregate all the results from all these intermediate nodes and sends it back to client application..

Why Use Amazon Redshift?

For setting up traditional data warehouse involves lot of money and resources it is Difficult to set-up, deploy and manage

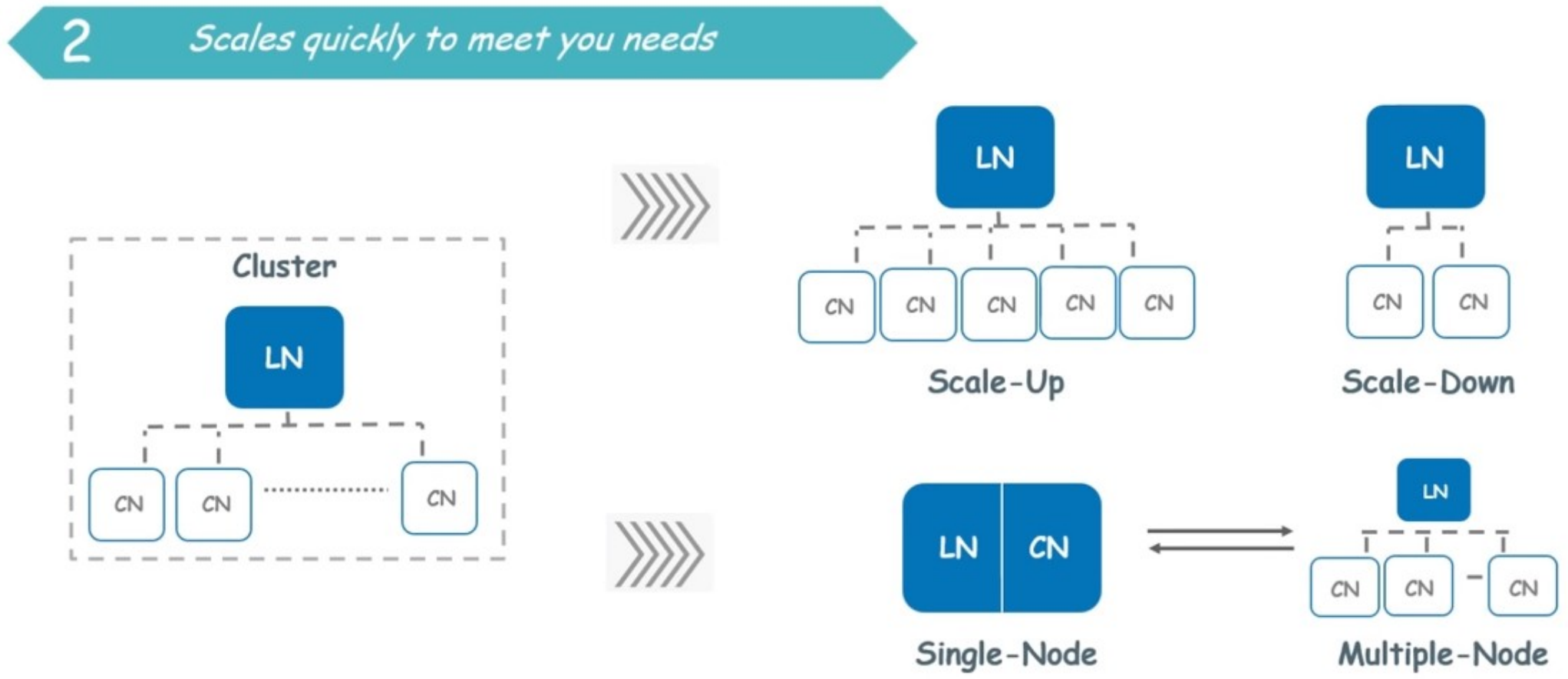
But its very easy to setup deploy and manage data warehouse using amazon redshift

Difficult to scale up and down on traditional system but using amazon redshift it is damn easy to scale quickly to meets your needs..

Decrease in the performance but with AWS redshift its 10 times better and faster performance

Why use Amazon RedShift

Why use Amazon RedShift



Why use Amazon RedShift

3 *10x better & faster performance*

★ Columnar Data Store

Row Storage

SSN	NAME	AGE
107135024	Jenson	25
382634557	Sam	27

107135024 | Jenson | 25 — Block 1

382634557 | Sam | 25 — Block 2

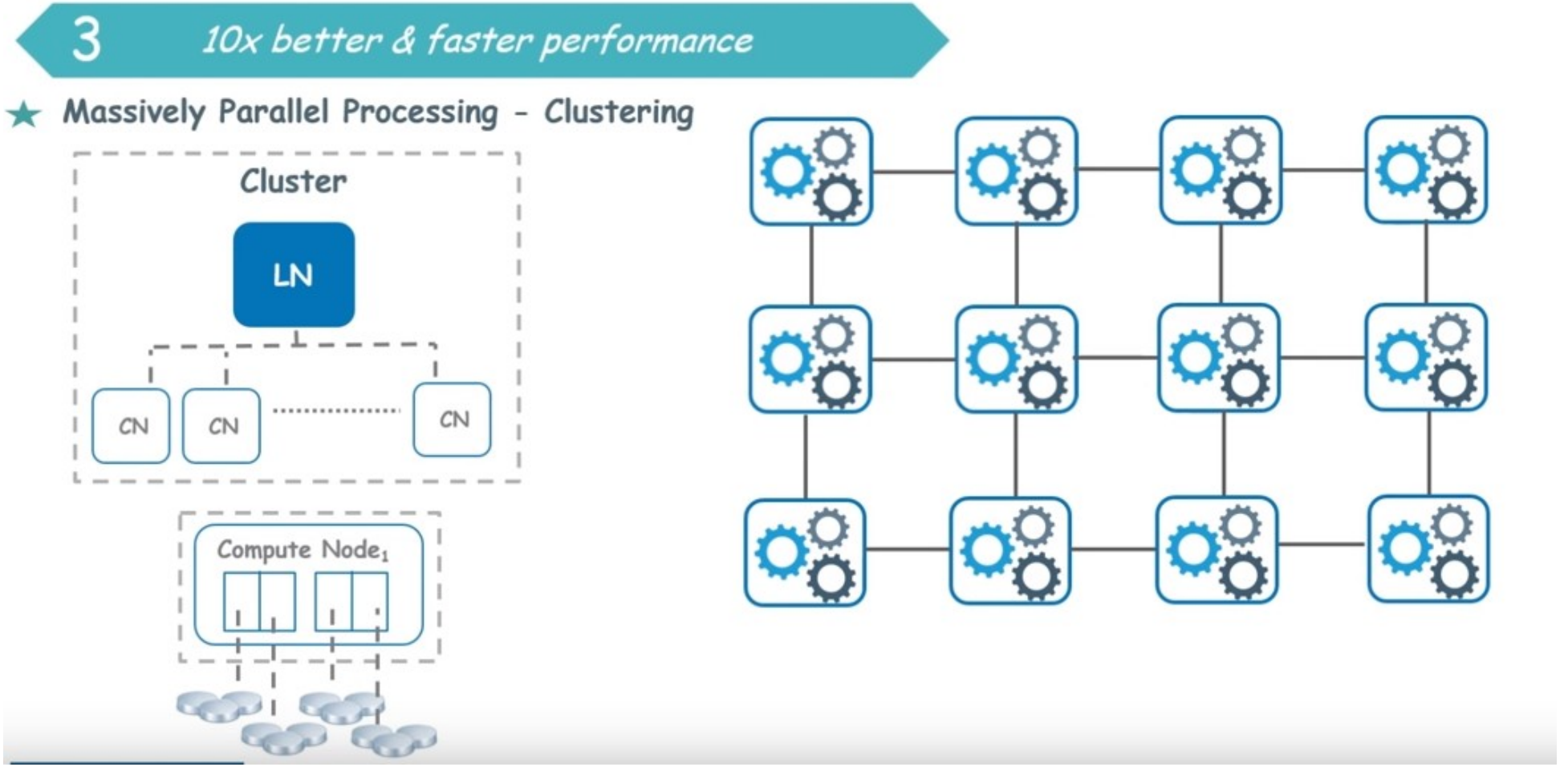
Column Storage

SSN	NAME	AGE
107135024	Jenson	25
382634557	Sam	27

382634557 | 107135024 | ..... — Block 1

Jenson | Sam | ..... — Block 2

Why use Amazon RedShift

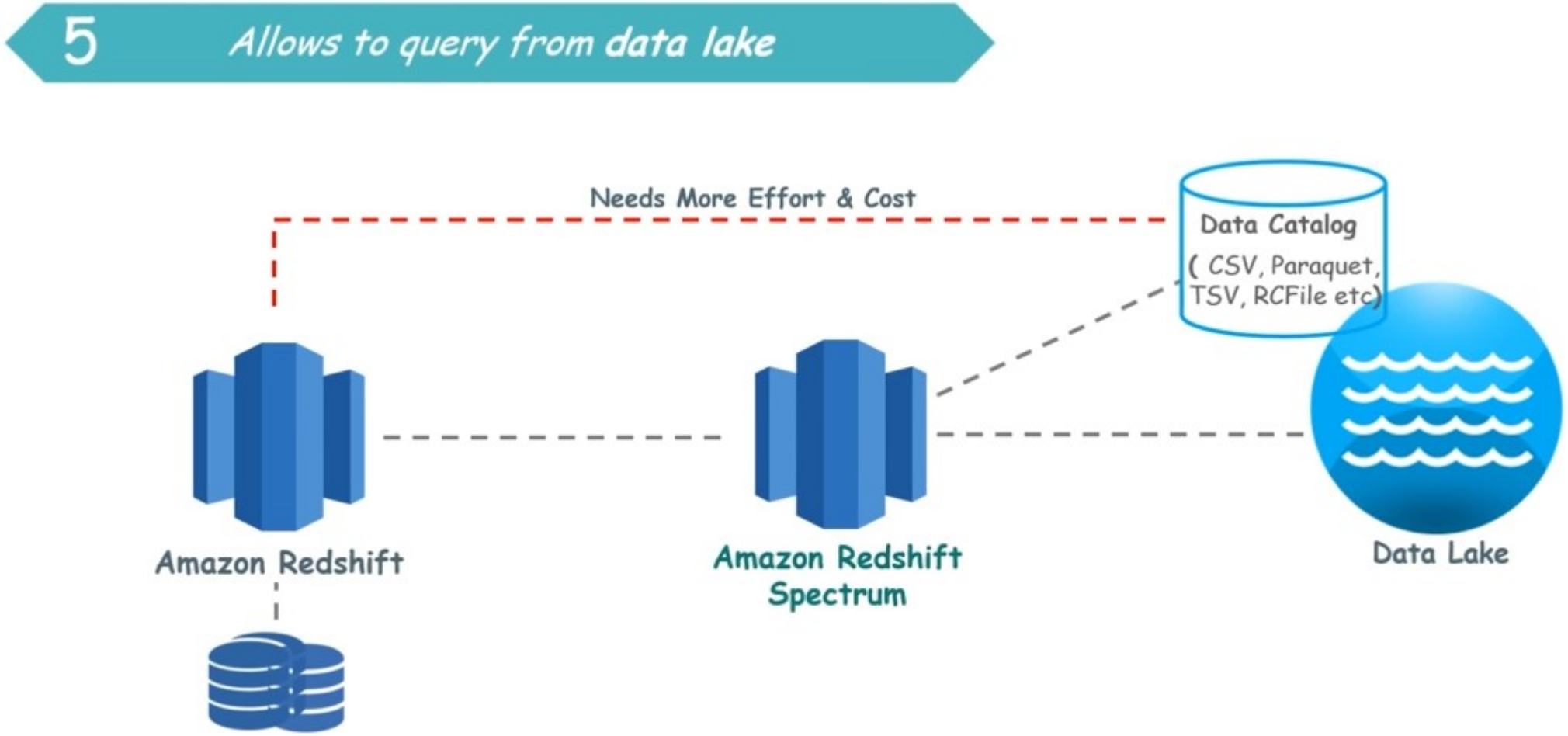


Why use Amazon RedShift

Cost – Effective & 1/10th of traditional data warehouse

Cheaper to set-up because of no hardware or upfront costs

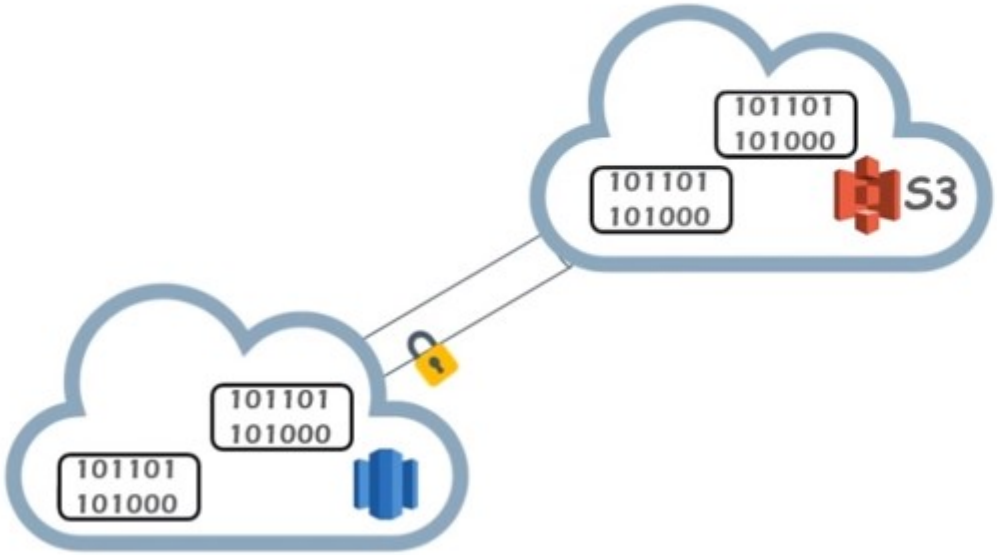
Why use Amazon RedShift



Why use Amazon RedShift

6 *Data is secure in Redshift*

★ Backup & Recovery



★ Encryption

Redshift: Exam Tips

- Redshift is used for Business intelligence.
- Backups enabled by default with a 1 day retention period.
- Maximum retention period is 35 days
- Redshift always attempts to maintain at least three copies of your data(the original and replica on the compute nodes and a backup in S3).
- Redshift can also asynchronously replicate your snapshots to s3 in another regions for DRR.
- Single Node (160GB)
- Multi node
  - Leader node(manage client connections and receives queries)
  - Compute node(store data and perform queries and computations). Upto 128 Compute nodes.
- Currently available only in 1 AZ

DEMO

Setting up A Data Warehouse on AWS Redshift

DEMO

Install SQL WorkBench/J

Install JDBC Driver

Make Sure you have Java enabled on your OS