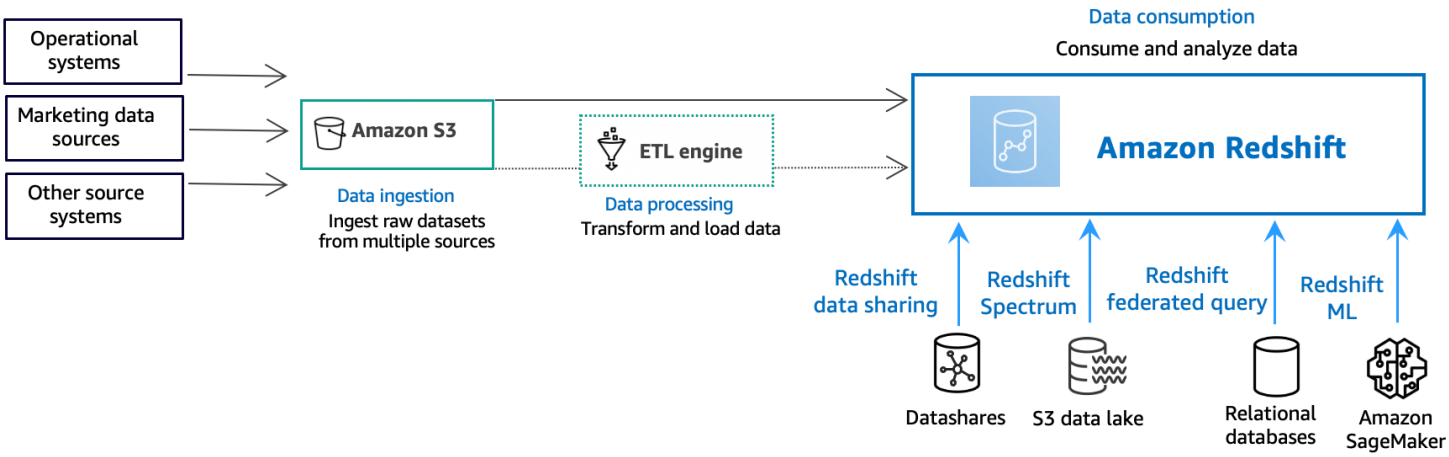
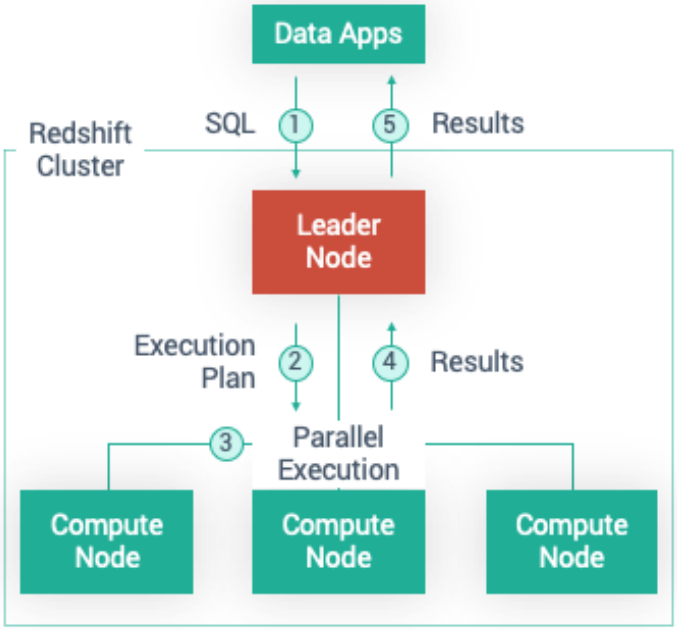
Agenda

a. Optimisations in Hive
b. **Essential features of Redshift**
C. Columnar Data warehouse concepts
d. Architecture
e. Distribution key and sort key



Amazon Redshift Architecture: The Life of a Query



The 5 Steps to Process a Query

1. The cluster receives a query coming from a data app and parses the SQL in the leader node.
2. The leader node creates an execution plan that breaks a query down into a discrete sequence of steps.
3. The leader node distributes the work of executing the steps in parallel across the compute nodes.
4. The compute nodes send the results back to the leader node to merge data into a single result.
5. The leader node addresses any final sorting or aggregation and returns the results to the data app.

A DBA's Guide to the Amazon Redshift Architecture

intermix.io

→ Data lakes (Glue + Nudge)

→ Airflow

→ Spark → RDD
→ Dataframe (1)
→ Streaming → Kafka
→ Dataframe (2)

→ Interview Questions

→ Case study → 19th Sept 2024

Flow
↳ Partitioning

Dynamic
↳ Flow

Create
↳ Partition by (state)

Static
↳ State = "PUNJAB"

29 partitions

↳ Partition by (...)

Cloud

↳ Service based Model

fully furnished room

Tenant

↳ Bang →

⇒ 01th Sept

30th Sept ⇒ Rent

OS

2 weeks

MySql Server

N/W

1.5L

Ma

upgrades

Patches

70K

30K

x

1 minute ⇒ 0.005$

EC2 →

Public

AWS (57)
Google → GCP (20)
Microsoft → Azure (22)
3 Tb
IBM | Snowflake | OCI | Alibaba ...

Private cloud
↳ Ericsson

AWS Redshift
↳ enterprise relational DWH that stores upto PB | EB of data.

(client)

(1) Data Apps :-

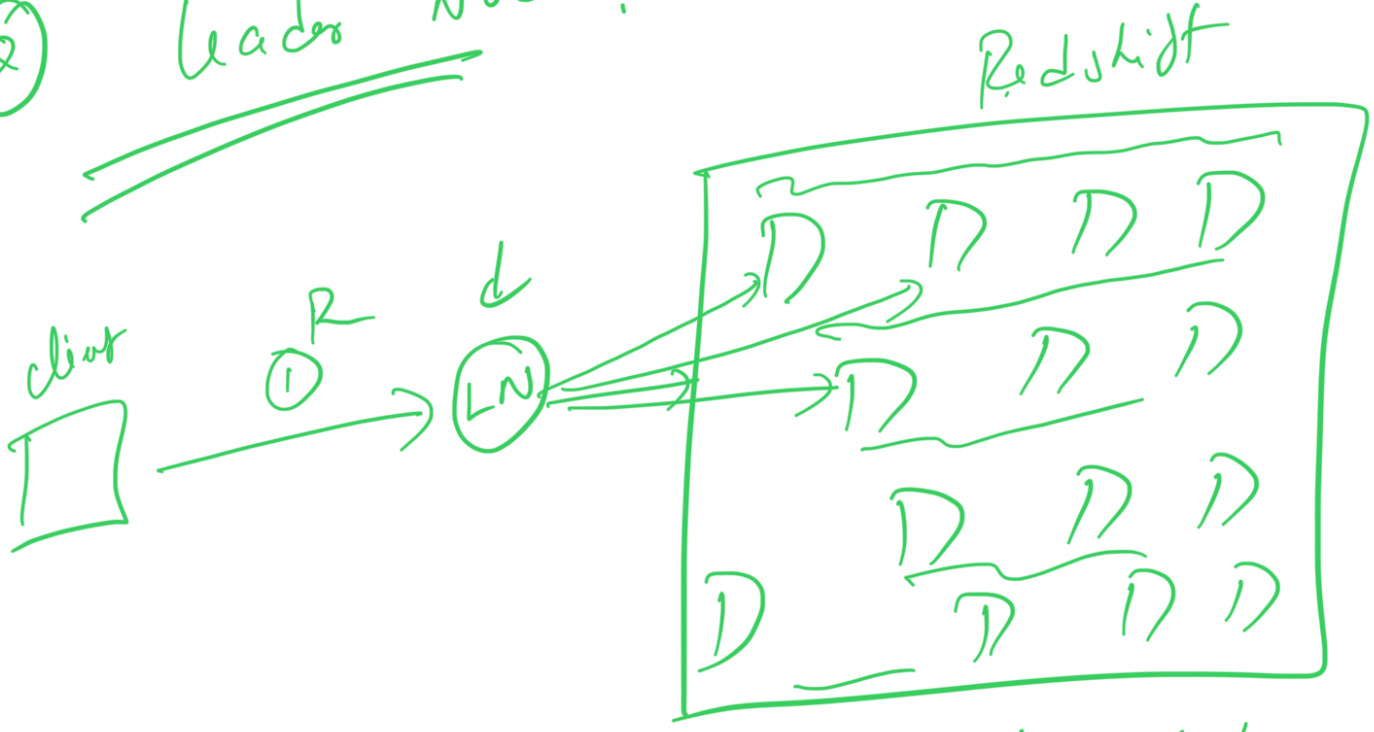
(a) Data Integration (Load)

(b) workflow orchestration
↳ these are the apps that run batch jobs on a schedule.

Analysis :-

© Ad h-oc ...

② Leader node :- ⓐ Communicate with Dataapps.

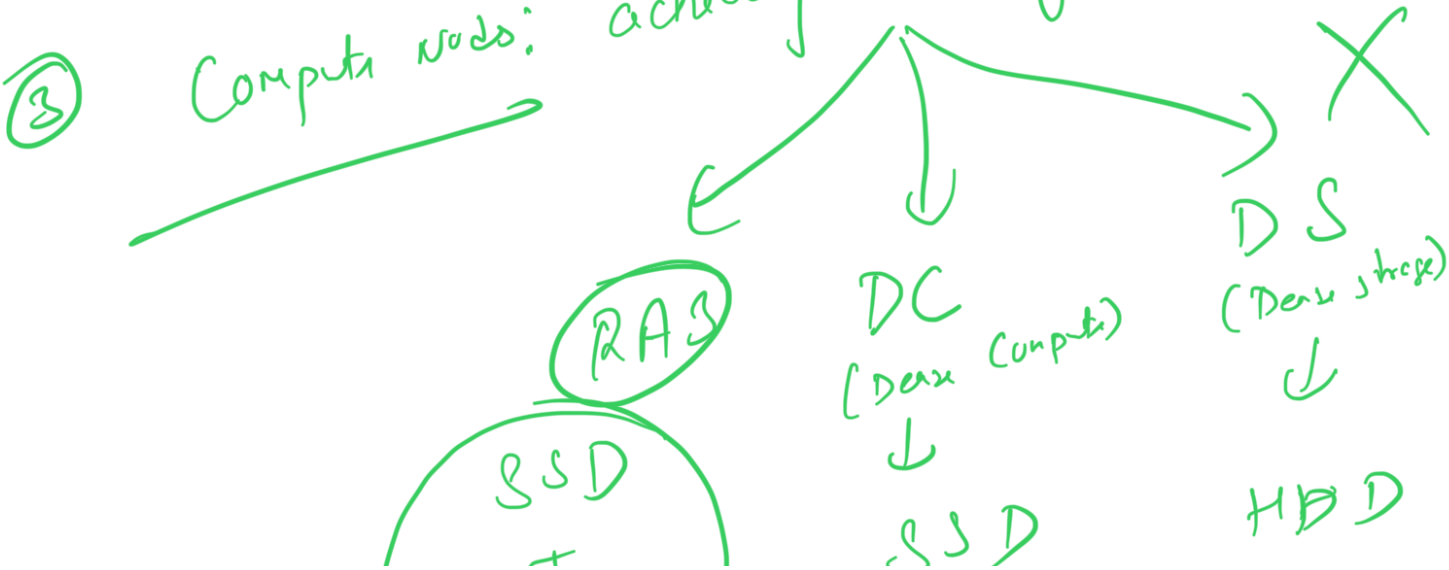


ⓑ Distribution of workloads.

ⓒ Caching of query results.

ⓓ Maintain metadata.

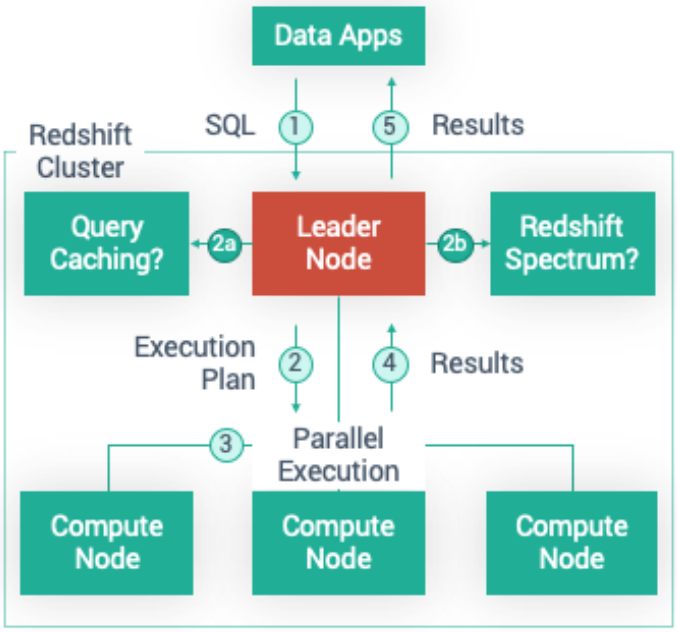
③ Compute nodes: actually store your data



S3

4 Spectrum :-

Amazon Redshift Architecture: Query Caching & Redshift Spectrum



1. The cluster receives a query coming from a data app and parses the SQL in the leader node.
2. The leader node creates an execution plan that breaks a query down into a discrete sequence of steps.
   - 2a Determines if cached result is available
   - 2b Determines if query goes to Redshift Spectrum
3. The leader node distributes the work of executing the steps in parallel across the compute nodes.
4. The compute nodes send the results back to the leader node to merge data into a single result.
5. The leader node addresses any final sorting or aggregation and returns the results to the data app.

A DBA's Guide to the Amazon Redshift Architecture

intermix.io