Recap :

On-premises DW vs Dw on cloud

DW -DM - ?

Cloud computing ;

Fault tolerance

Data availability

Scalable

Distributed storage

Concurrent

-----------------------------------------------------------

Snowflake :

SAAS → no software /hardware

Dw on cloud

Analytical engg

DB --Schema--Tables

AWS /Azure/GCP

DML ? insert/update/delete

Arch :

Data storage : Remote Disk

Central repos:

Shared disk ? : AWS s3 : internal snowflake managed

Compute layer /query processing ;

Virtual warehouse : Ec2

Multicluster ---> employee → 4gb

N1 1gb n2 1gb n3 1gb n4 1gb

MPP --Massively parallel processing

=======================================

10 ---> 5 mb ⇒

=======================================

AWS ---> data loading ⇒

============================

Cloud service layer : foundation db

Brain of snowflake

Metadata

Security

Optimizer

stats⇒

=================================

Special Features of SF :

Micropartition

Zero copy cloning

Data sharing

Time travel and fail safe

Caching

Streams and task

===================================

Snowflake table :

compressed

Optimized

Columnar

Micropartitioned

ETL / ELT ⇒ Extract Load and Transform

Load data => stages ===> snowflake DW

Emp.csv --->emp.csv.gz ⇒ emp

==================================

Row oriented vs column oriented

111 aaa 6000

112 bbb 7000

10m →

Table scan : select \* from emp where sal >20000;

6000 7000 8000 :

Column pruning : ignore the column not in query

10 trillions --worldwide ===> US →

Partition ===> sharding ⇒ chunk/blocks/partition

========================================

**Oracle : value -> hard disk →**

**Partitioning ===> splitting based particular column value**

**year=2001**

**year=2002**

**year=2003**

=======================================

Micropartition :

50 mb -500 mb --uncompressed

Large table ==millions of MP

=====================================

Column pruning : 721000 ===> 7000 ⇒

=======================================

Data clustering also :

Day --1 1-------------------30

Day -2 20-------------------40

Day -3 1--10

Day -4 1- 50

Natural dimension : date / geographical region

Query ; filter , agg , join =frequently used

=====================================

111 aaa 6000

111111 111 aaa 8000

======================================

Reclustering is expensive → snowflake

=======================================

Afternoon session :

Insert

Load --bulk data load

Local ---put ------Stage ---->copy ----Snow table

----------------------------------------

Types of stages

1. User stage
2. Table stage
3. Internal named stage
4. External stage

===========

Internal named stage =>st\_lti\_int ⇒ snowflake managed stage