Feedback /suggestions/recommendations

Concepts

Technical

Pace

Recap :

Types of table

Zero copy cloning :

Sales sales\_1 sales\_2

1 tb 1 tb 1tb

Zero copy

Sales sales\_1 sales\_2

1tb metadata metadata

Only one --unmodified data , no storage , no cost

After cloning , each table -independent object

Modified data of individual table stored in new micropartition

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

Creating environment

dev 10 - test 10 -> 2 → --prod ->

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

Dynamic caching ;

Remote ---disk → first queries ====data storage

Local/ SSD cache /warehouse cache

Cloud CSL => query result cache /

metadata cache ==count(\*)

Result cache ==24 hrs

Identical queries --------24 hrs ====extended

6am 11pm 11pm

Maximum : 30 days

flush/clean===>changes in table

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

Warehouse :

First query (full table scan 10 /10 partition) ⇒ remote

Second query ⇒ local

First query (partial scan 5 / 10 partition ) ==>remote

Second query ⇒ local + remote

======local cache cleared

Warehouse ==suspended / resize

Changes in table

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

Full import → 15 years ⇒ bulk transfer ==.snow

Delta import

Incremental load ⇒ source =insert/update/delete

1-1000

2 -insert

1-update

2-delete snowflake

Change data capture ⇒

Streams

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

Resize / change type ==credits

Auto suspend/ auto resume

Concurrency workload ==scale in ---min -max

Scale out → max to min

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

File formats :

Data engg :

Replication : 1 gb 1gb 1gb

Data available / fault tolerance

Fb --2 gb → photos → thumbnail

Compression → encoding file format

Semi structured data :

Encoded file format : orc /parquet /avro

Json / xml

Compression codec : snappy / deflate/bzip/gzip

Fileformat : source data

Snowflake apply the best compression on the column level

Compression --entire file level

Sales.csv sales.csv.zip

Storage optimization :

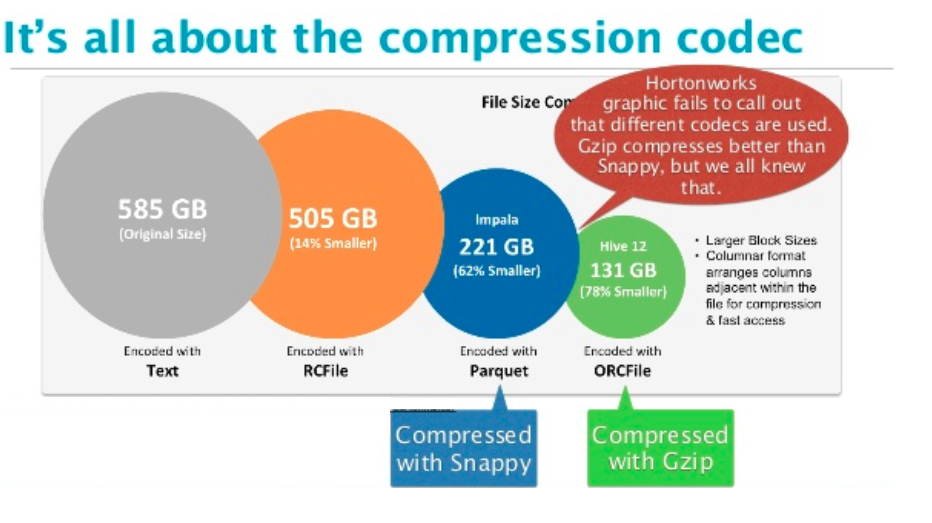
====================================

{ empno: 111

Ename: aaa

Sal:6000

}



Data type :

Variant ->

Object

Array

Dense /sparse →

Employee -10m

Empno ,ename,sal,comm,bonus , deptno,ph1

111,aaa,6000,null,10,11111111

112,bbb,7000,200,null,10

113,ccc,7000,null,300,20

Json

{ empno : 111

Ename: aaa

Sal :6000

Deptno:10

}

{ empno:112

Ename:bbb

Sal:7000

Comm:200

bonus:700

Deptno;10}

{

Empno:113

Ename:ccc

Sal:7000

Bonus:300

Deptno:20

Contact : { 111111, 9888888 }

RDB : fixed schema /rigid schema

Semistructured data :

Schemaless /schemafree

**Schema design :**

1. Upfront schema → create table and we load
2. Schema on read -> external table
3. Schema on write →

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

D:/datasets/sales.txt 200 mb

Oracle --sales table ⇒

1.sql loader ---> load data into table

2. Create a table and map the data into table -salesext

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

External stage snowtable

Data ========>copy ======= metadata/data

Superstore.csv --copy ---columar/optimized/compressed snowflake table (remote)

External table : (Read only)

External stage :

s3://lti935/employees/emp.csv

No data copy ====>metadata (snowflake)

1. If you delete the table , data remains same at your source location

No schema violation errors /notification even if they folders contains different types and kind of file

2021 --------emp.csv ==365 files ==>10 m

yyyy/mm/dd

2021/01/01