VERTICA

Sizing and Configuring Vertica in Eon Mode for Different Use Cases

Sumeet Keswani & Shrirang Kamat
Product Technology Engineers

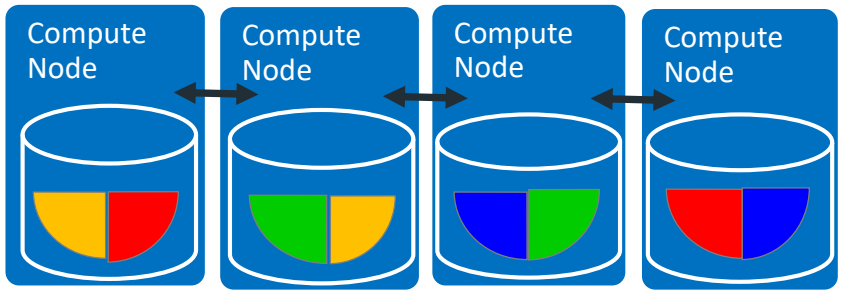
MICRO FOCUS

Agenda

- Eon mode concepts
- Sizing an Eon cluster
- Use Cases for Eon mode
- Tips & Tricks
- Q&A

Enterprise Mode

When Enterprise Mode ?



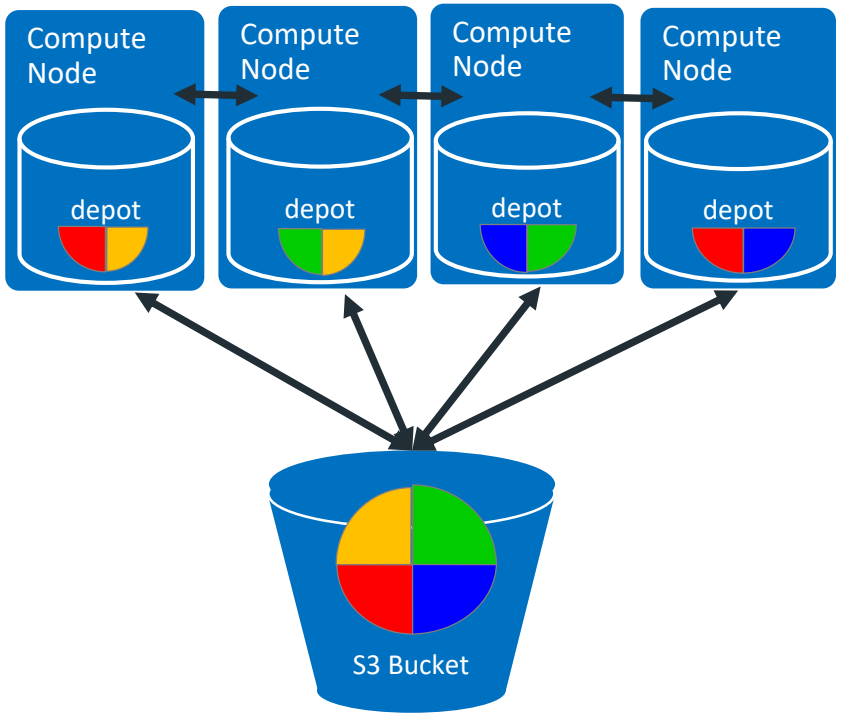
For on premises and embedded deployments. Can also be deployed in cloud and virtual environments

Only need general purpose compute nodes with attached storage

Fast and Reliable performance all the time as entire data is always stored local to the nodes

Eon Mode

When Eon Mode ?



For cloud deployments but can also be deployed on premises with S3/webhdfs compliant communal storage

Scale infrastructure quickly to meet the demands of changing workloads

Isolate analytic workloads to subset of nodes in the cluster

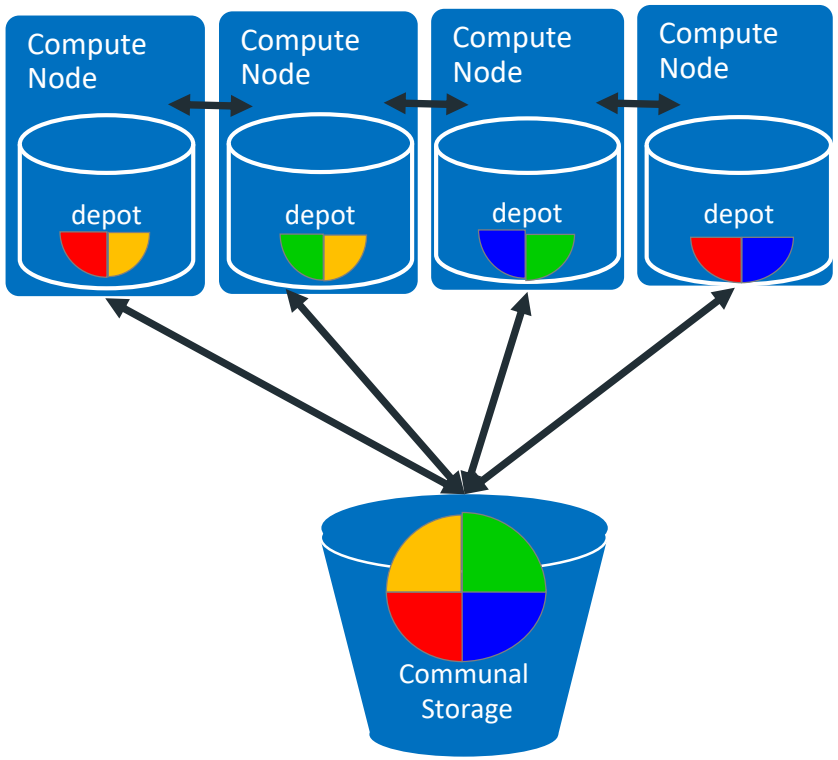
Separation of compute and storage resources means load data without worrying about local disk capacity

Achieve reliable performance by provisioning compute nodes with local disk large enough to hold most frequently queried data

Eon Mode Terminology

Communal Storage - A storage location shared among the nodes of a database (COMMUNITY LIBRARY )

Shard - A segment of the data stored in communal stored in communal storage(BOOKS ARE DIVIDED INTO SECTIONS)

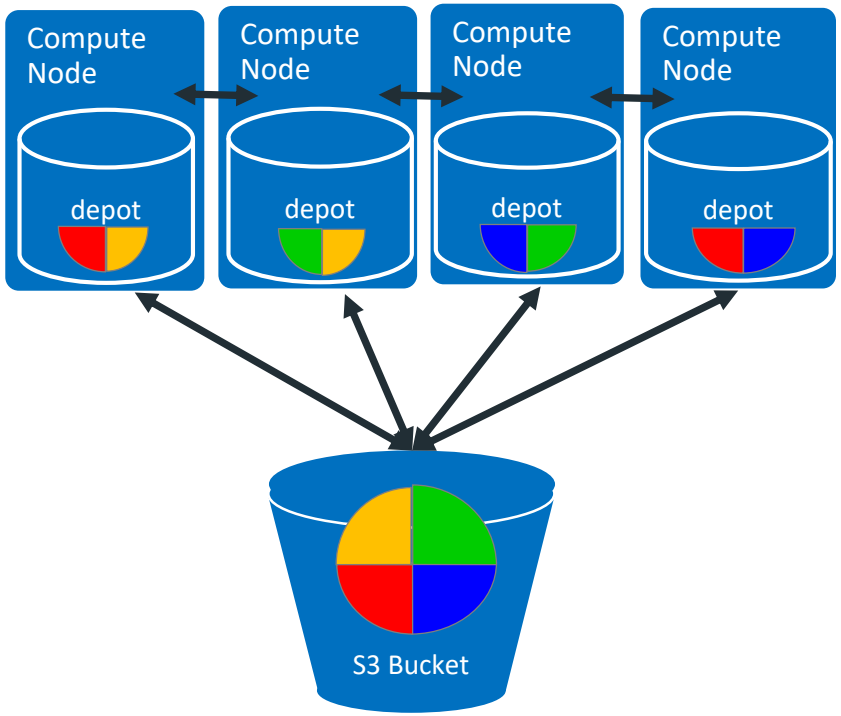


Subscription -The set of shards a node can read from and write to.(CONSUMERS SUBSCRIBE TO ONE OR MORE SECTIONS & MAINTAIN UPTO DATE COPY OF ALL TITLES IN THOSE SECTIONS)

Depot - A local copy of subset of data associated with shards that a node subscribes to (YOUR BOOKSHELF AT HOME)

Shards

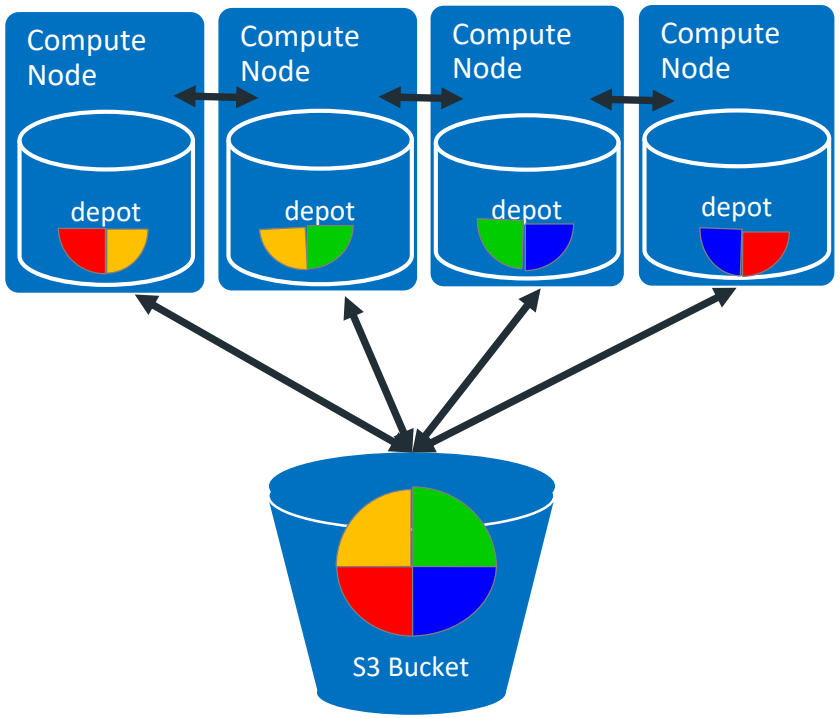
Data segments



- In communal storage data broken into segments called shards
- Shard count is supplied at database creation and cannot be changed
- Shard count will decide maximum number of nodes that will participate in a query
- Default replica shard exist in every database for storing data belonging to replicated projections

System table: SHARDS

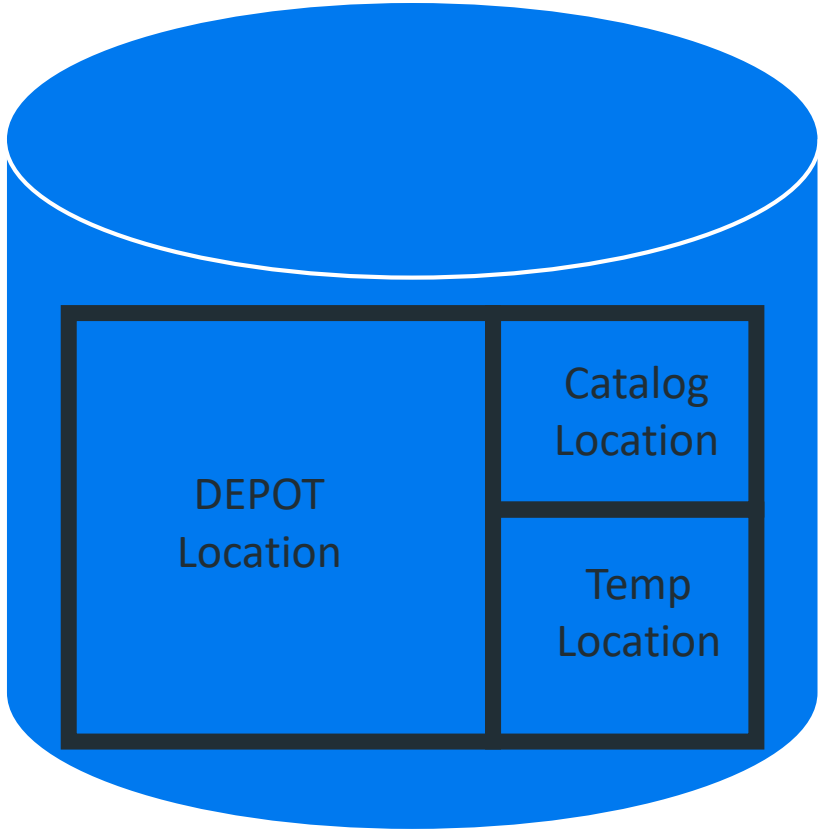
Subscriptions



- Each node subscribes to a subset of the shards and each shard has more than one subscriber node
- Subscribing node has up to date metadata only for shards that it subscribes to
- A session is assigned a list of nodes and participating subscriptions called session subscriptions
- Session subscriptions are used for loading data and running queries

System tables: NODE_SUBSCRIPTIONS & SESSION_SUBSCRIPTIONS

What's on Local Disk

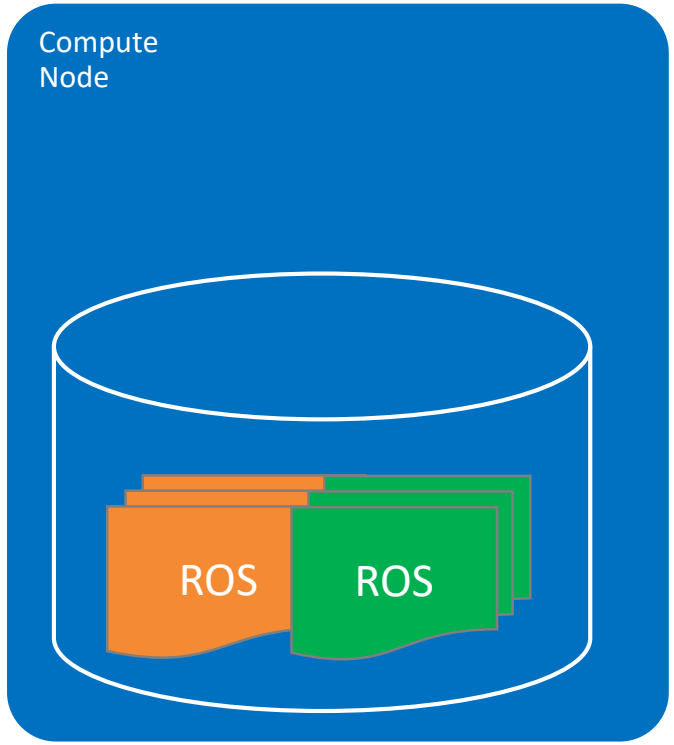


Depot Location- Cached local copy of ROS files and user defined libraries

Temp Location - Temporary table data and data spills from queries. Data for temporary tables is not uploaded to communal storage

Catalog Location - Persistent catalog written at every commit and read at startup, Vertica logs and Data Collector tables. Local persistent copy of catalog is synced to communal storage every 5 minutes

Depot Usage

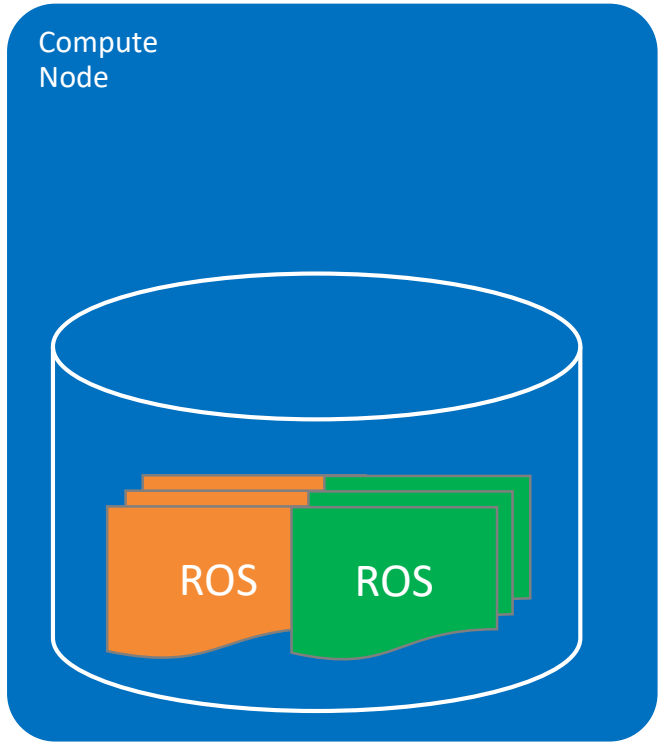


Depot is local disk ROS file cache to improves query performance and reduces network traffic

New ROS file is written to the depot first and then uploaded to the communal storage as part of a transaction. Configuration parameter to skip depot when writing new files exists

Queries make use of files in depot. When a query can't find a file in the depot, it will access from communal storage. Query performance is best when data required to answer query is present on local disk/depot

Depot Fetches & Eviction

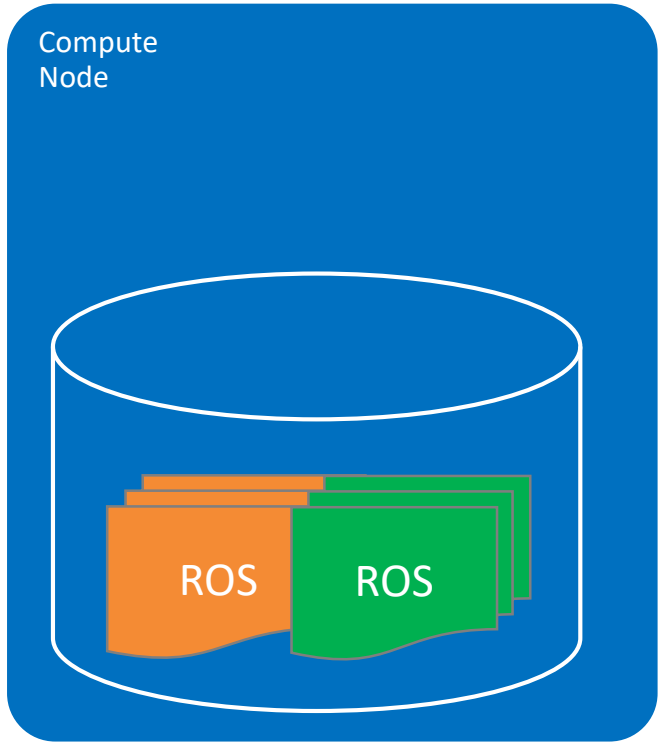


A ROS file is asynchronously fetched in depot if a query could not find a ROS file in the depot. Configuration parameter DepotOperationsForQuery can be set at database, session and query level. You will be able to set it at user level in 10.0

Least recently used (LRU) files are evicted from the depot to make space for the new files as needed.

Pinning Policies can help shape depot content for subcluster or cluster as per your needs

Depot Fetches & Eviction



System tables:

DEPOT_SIZES -Size of depot on each node

DEPOT_FILES - Files that are currently present in the depot

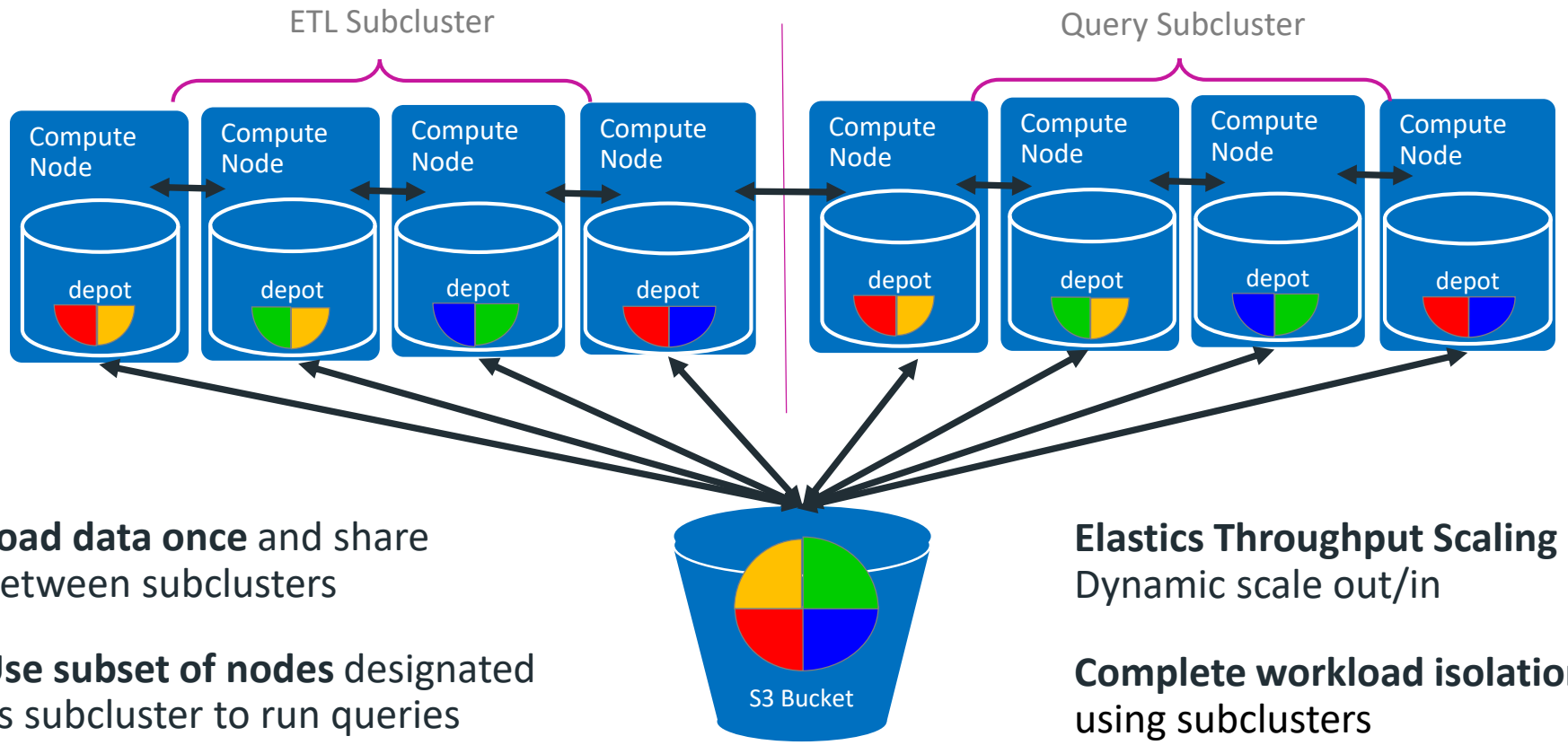
*DEPOT_EVICTIONS - Files recently evicted from the Depot

*DEPOT_FETCHES -Files recently fetched to the Depot

DC_FILE_READS - Files read from the Depot

* system tables based on data collector tables

Eon Mode - Subcluster



Load data once and share between subclusters

Use subset of nodes designated as subcluster to run queries

Elastics Throughput Scaling by Dynamic scale out/in

Complete workload isolation using subclusters

Subcluster Types

Primary and Secondary

- **Dynamic Scaling -** Support multiple subclusters that can be stopped and started on demand without impacting cluster state.
- **Workload Isolation -**Provide ability to stop and start subclusters without impacting workloads running on other subclusters
- **Large Clusters -** Optimize for cluster with hundreds of nodes.

Primary Subcluster

- **Always UP** - A cluster must have at least one primary subclusters that is always be UP
- **Participate in commits** - Nodes in primary subcluster participate in commit operation
- **Persist catalog on disk -** Nodes in primary subcluster persist transaction logs to disk at every commit
- **Participate in quorum** - Nodes in primary subcluster participate in database quorum that is used to decide if database can stay UP and/or restart.
- **ETL workload -** Primary subclusters can run all kinds of workloads but best suited for running ETL workloads

Secondary Subcluster

- **Available on demand**- A cluster can have zero or more secondary subclusters that can be added or removed on demand
- **Designed for dynamic scaling** - Add/remove or start/stop secondary subclusters without impacting workloads running on other clusters.
- **Maintains in memory copy of catalog**, receives catalog changes at every commit but does not write or read persistent copy on disk
- **Highly recommended** - Secondary subcluster is highly recommended for every eon clusters running on cloud.

Sizing Eon Cluster

Sizing Eon cluster

- **Shard Count** for maximum number of nodes that can participate in a query
- **Number of Nodes & Instance Type** for each subcluster
- **Subclusters** for Workload Isolation and Elastic Throughput Scaling

Picking Shard count

- Depends on amount data processed by queries and not total data size
- Most dashboard or multitenant applications don’t require more than 12 nodes
- Pick a number that has multiple factors (not 17!)

Small/Medium – 12

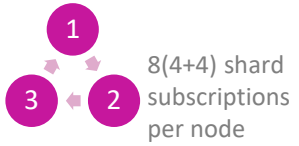
Large – 24

X-Large – 48

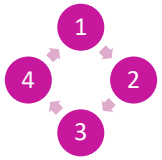
Picking Nodes per Subcluster

- Number of nodes in a subcluster must be equal to factor of shard count so that you have balanced shard subscriptions among nodes
- When number of nodes match shards, each node will have 2 shard subscriptions. 2 or 4 shard subscriptions per node is recommended, avoid more than 8 shard subscriptions per node
- 12 node subcluster has 2x depot capacity and processing power as compared to 6 node subcluster but cost 2x more

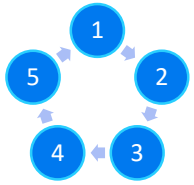
Small - 12 shards (3,4,5,6,12 nodes)



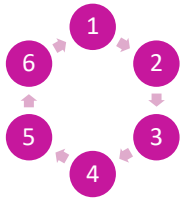
8(4+4) shard subscriptions per node



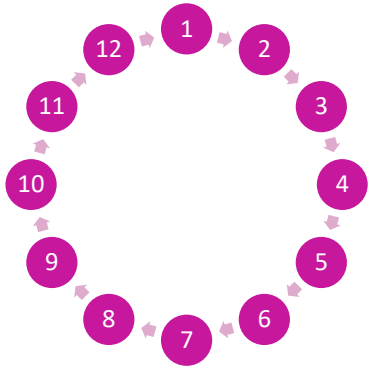
6(3+3) shard subscriptions per node



All Nodes will not have equal shard subscriptions as **5 is not factor of 12**



4(2+2) shard subscriptions per node



2(1+1) shard subscriptions per node

VERTICA

MICRO FOCUS

Picking Subclusters

- Every Eon database has one default primary subcluster
- Add/Remove secondary subclusters on demand to meet workload isolation and/or ETS needs
- Each subcluster can have different number of nodes and instance types
- All nodes within subcluster must be homogeneous

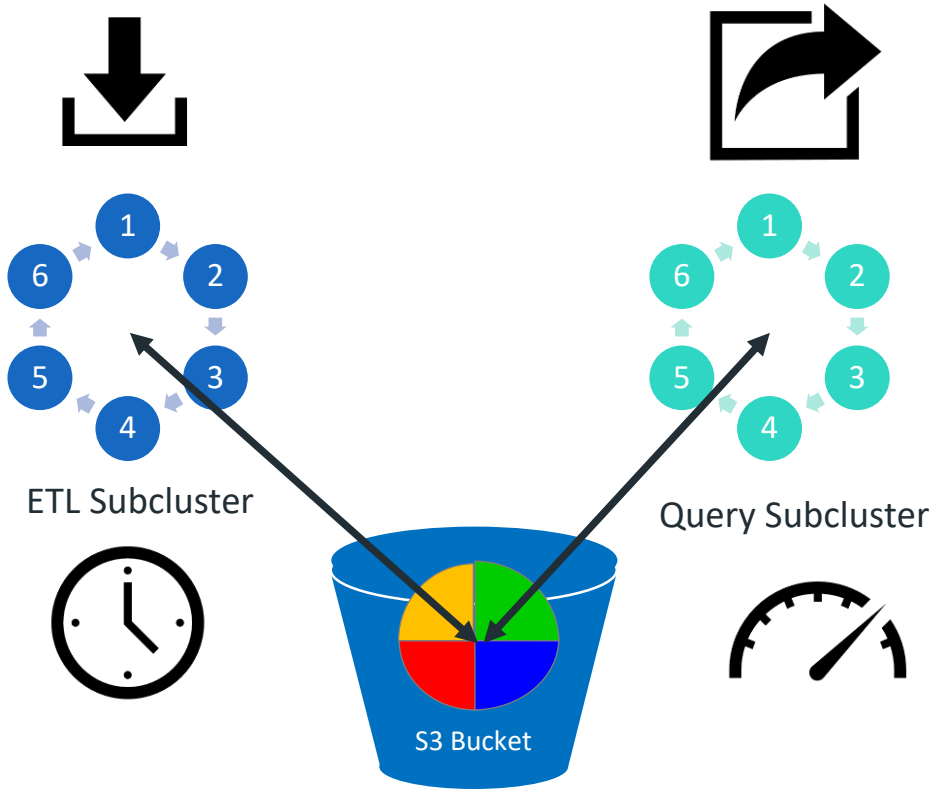
Picking Instance Type

- **Pick instance type** for subcluster depending on the workload and available budget
- **Provision instances** with at least 2TB local disk. Depots are must and work as query performance enhancer for all Eon deployment
- **i3 instances** with (NVMe) SSD-based instance storage works great for Eon on AWS
- **n1-highmem-16** are recommended for Eon on GCP

Use Cases

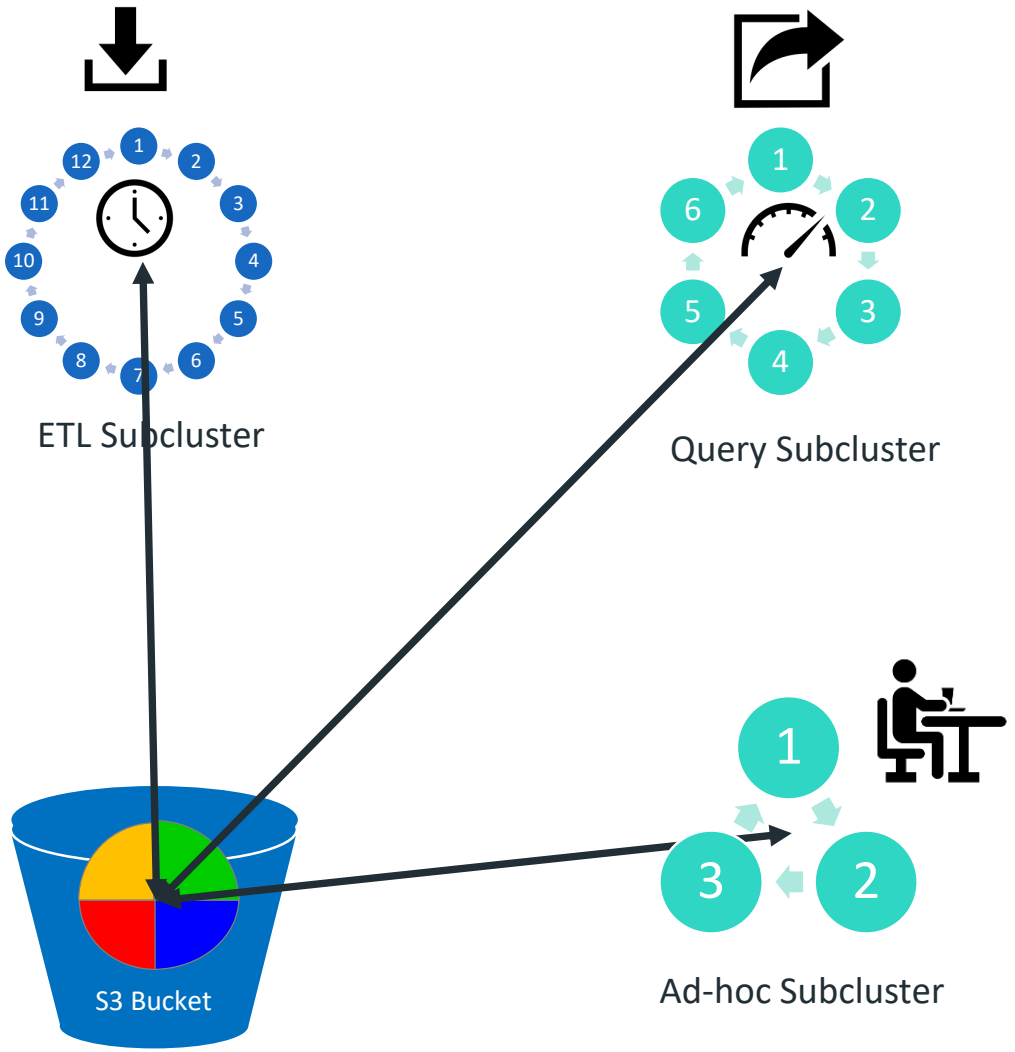
ETL and Query Subcluster

- Basic use case
- Typically the first one users start out with
- Primary Subcluster is used for ETL
  - Always or Mostly ON
- Secondary subcluster is used for queries
  - Sometimes ON (say 9 AM –5 PM)



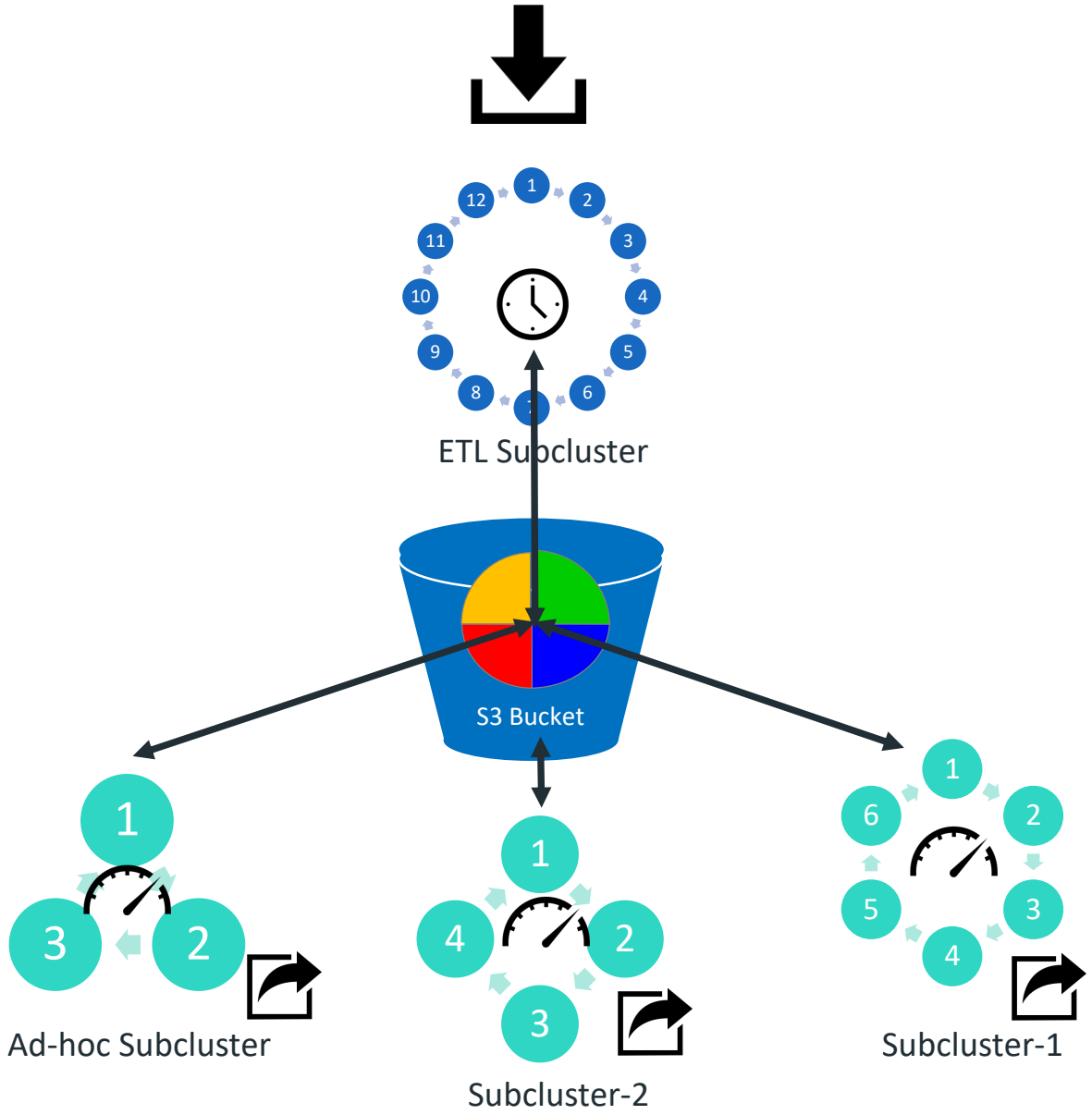
ETL, Query and Ad-hoc Subcluster

- The next level
- Primary Subcluster is used for ETL
- Secondary Subcluster is used for queries
- Ad-hoc secondary Subcluster on demand , provisioned for end of quarter processing or for the CEO



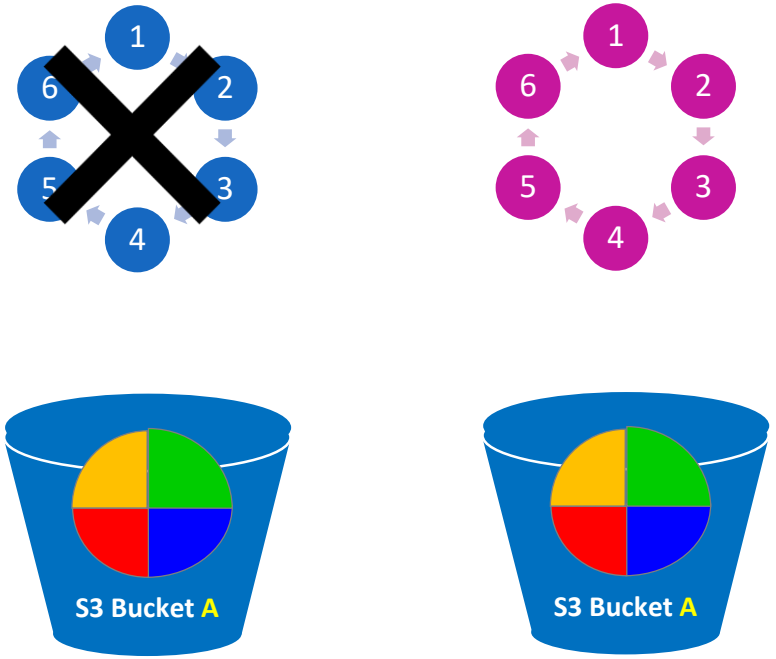
ETL, (n)Subclusters

- The next level,
- Primary Subcluster is used for ETL
- Several secondary subcluster is used for queries, running different workloads
  - Subcluster –1 (analysts, 6 nodes, ON 9AM-5PM)
  - Subcluster –2 (finance, 4 nodes, ON 24/7 at end of Quarter)
  - Subcluster –3 (CEO, 3 nodes, on demand)
  - Subcluster –4 (demanding tenant)



Hibernate and Revive

- Hibernate
  - You can deprovision/delete all instances and subclusters of a database
  - After you hibernate you only pay S3 storage cost
- Revive
  - You can provision new instances, and revive the database
  - You **must** revive with the same topology. (i.e same number of nodes and subclusters)
- Before you hibernate
  - Delete secondary Subclusters
  - Reduce the size of primary subcluster
  - Sync to S3

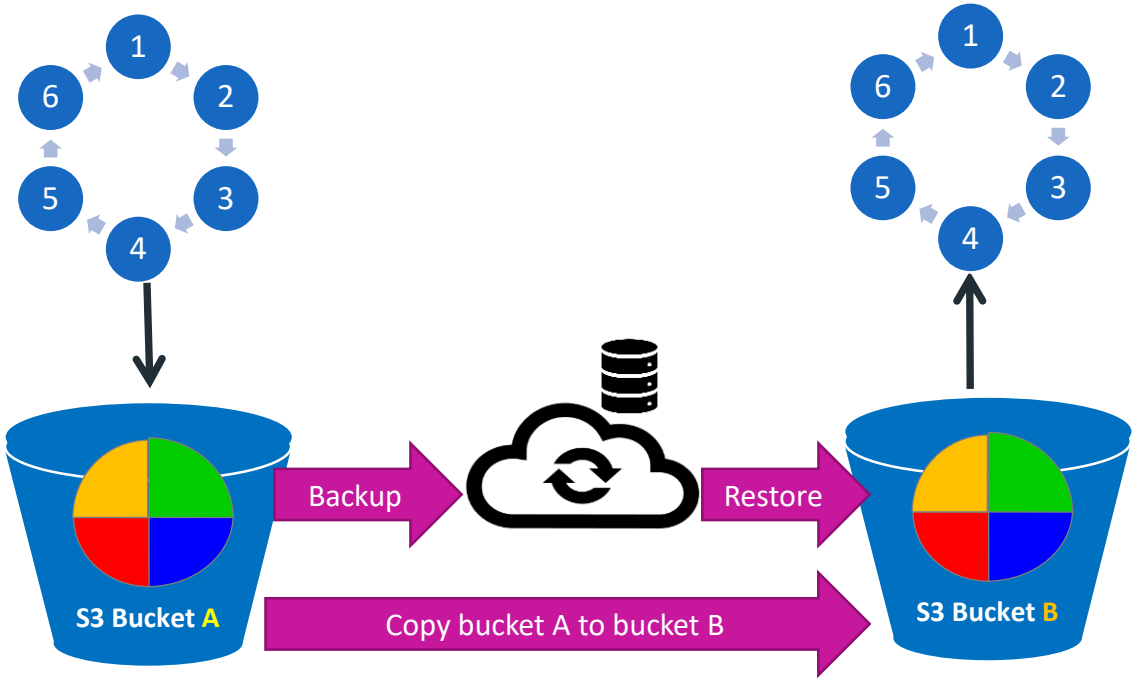


Replication and Backups

- Make a backup, restore it into another bucket

OR

- Make a copy of the bucket and revive



Best Practices

Depot shaping

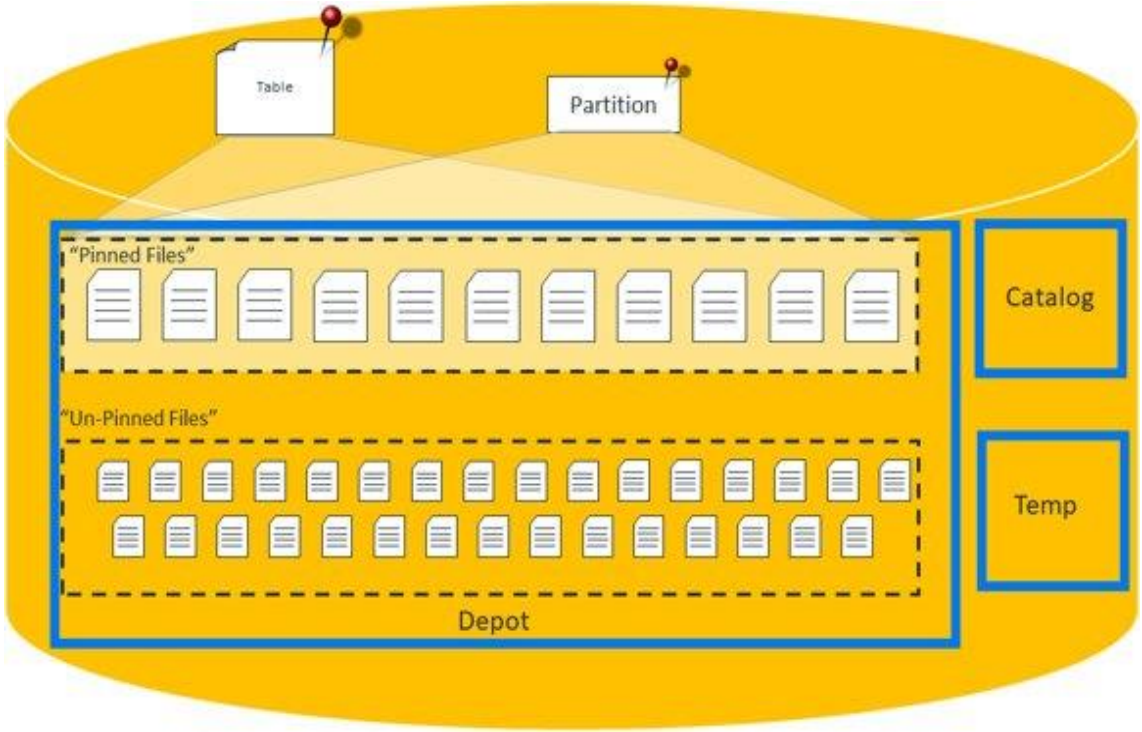
• Add pinning policy to shape your depot. Pinning policy is a ROS file eviction policy

• Can pin tables and/or partitions in depot

• ROS files of unpinned tables are evicted before pinned tables

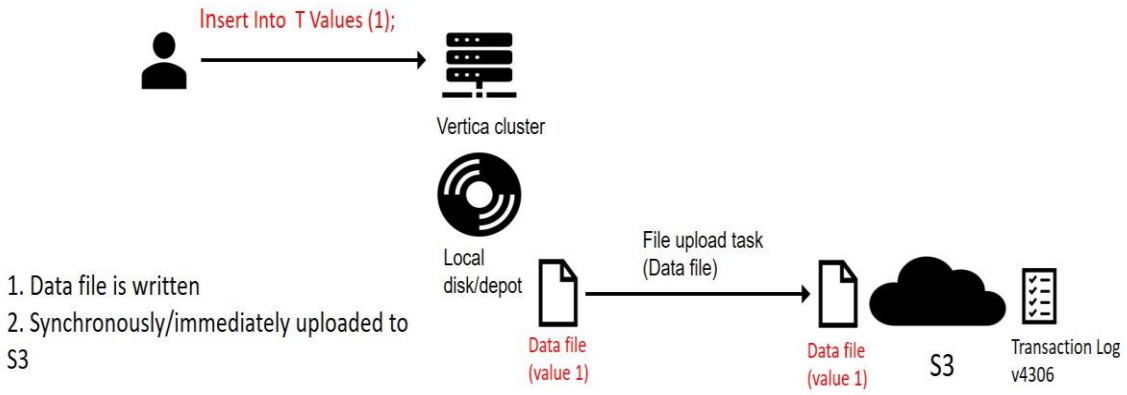
• ROS files present in depot are marked pinned, when policy is added to table and vice versa

•Subcluster level pinning policy will be added next release. In 9.3.1 release it is cluster wide

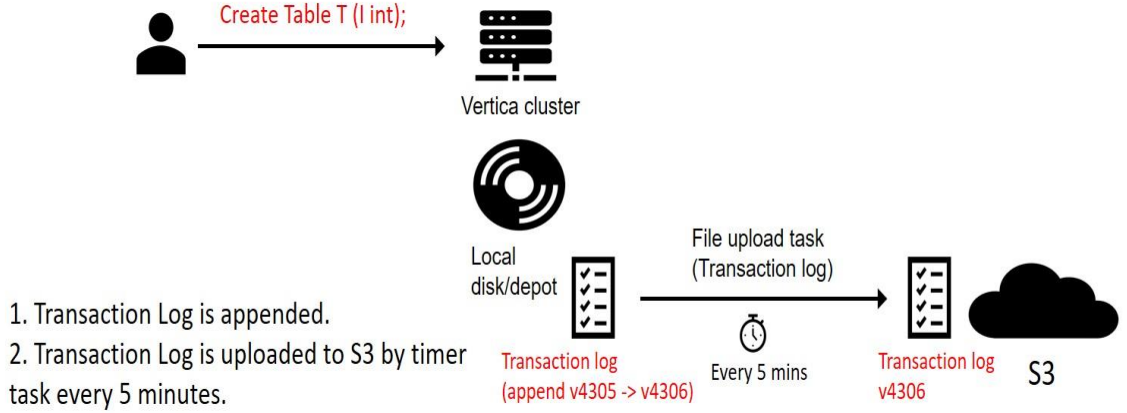


SQL in EON mode

- All data is written to S3 synchronously. *We do this to ensure durability of data in the cloud.*

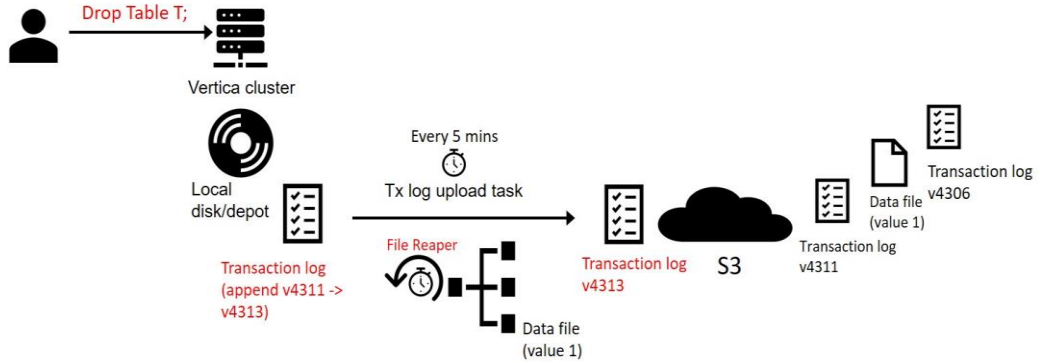


- All catalog operations and are done locally at first and distributed to all nodes synchronously . Each node asynchronously will sync its catalog/txn log to S3 periodically. *We do this because S3 is immutable and does not support file append operations.* Hence, we must batch Catalog and Transaction Log syncs (typically every 5 min).



SQL in EON mode

- All file deletes (**DROP/TRUNCATES**) are done asynchronously (lazily). *We do this to enhance performance.* The SQL can run fast(er) and deletes can be batched for efficiency and reduced api costs (we delete 1000 files at a time, with a 2 hour delay)



1. Transaction log is appended.
2. Transaction log is uploaded to S3 by Timer task.
3. Data file is added to the Reaper Queue for subsequent/asynchronous delete.

- UPDATE is effectively an INSERT+DELETE
- Tuple Mover is effectively an INSERT+DELETE
- EON mode does not have WOS, so we never lose data on single process/node crash

- If you terminate the entire cluster abruptly
  - You can lose (last) 5 min of comitted data
  - Files exist on S3 that should have been deleted

Reduce number of ROS files

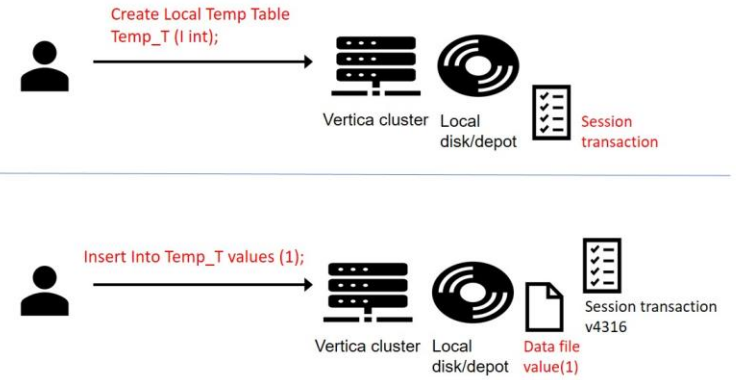
- Reading and writing to S3 is **slow** and **costly**.
- To avoid lots of round trips and too many files; load in large batches as much as possible.

Don't read and write from S3 if you can avoid it

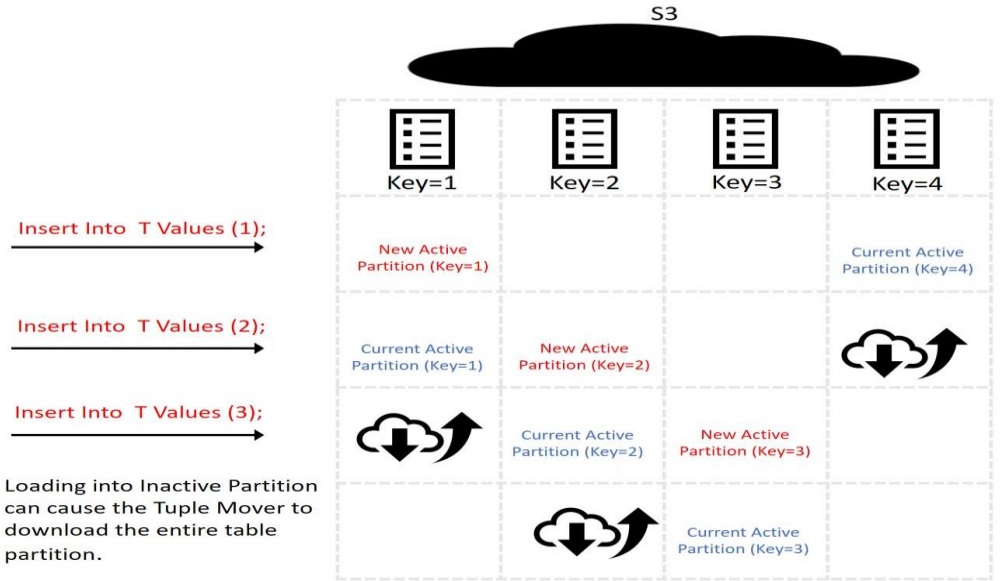
- Use Local Temporary Tables (LTT) for intermediate/non-persistent data (staging tables).

- Partition your tables correctly. Vertica merges inactive partitions immediately. This forces a download and upload of the inactive partition.
  - Avoid loading into inactive partitions.
  - Set active partition count on a tables correctly

```
ALTER TABLE public.store_orders SET ACTIVEPARTITIONCOUNT 4;
```



- Creating local temporary table updates the session transaction and is not uploaded to S3.
- Inserting into local temporary table does not create a file(s) on S3.
- There is no upload or download to S3 for operations on a local temporary table.

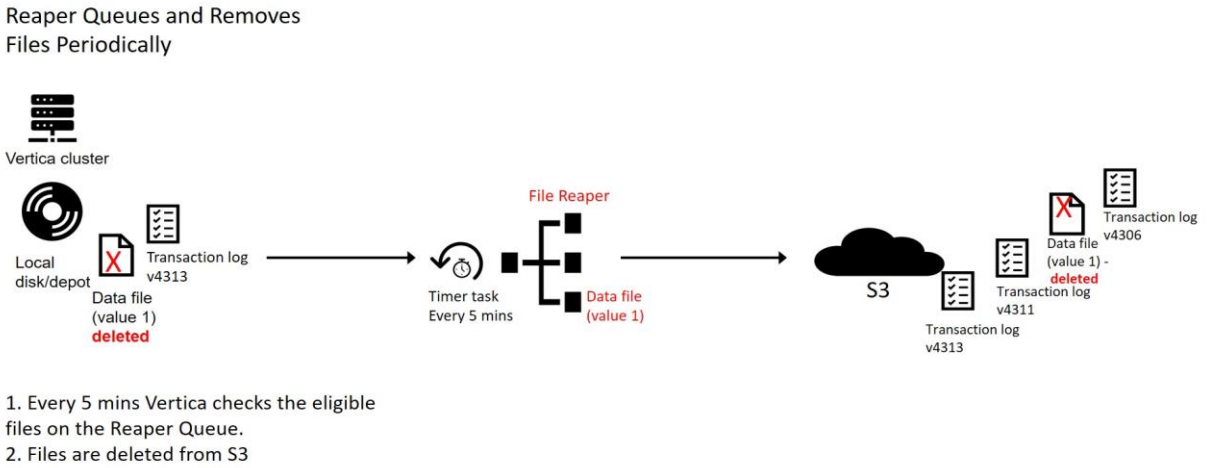


Loading into Inactive Partition can cause the Tuple Mover to download the entire table partition.

Deleting files and reaper queues

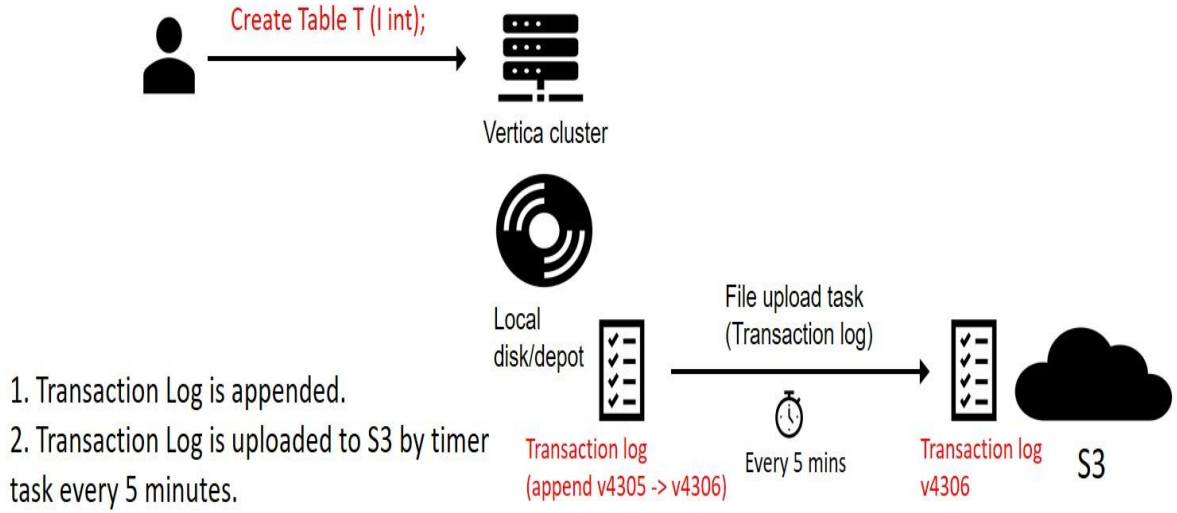
- Depending on the velocity of your ETL the number of deleted files can be very large
- Deletes incur api access costs. Use local temporary tables where possible.

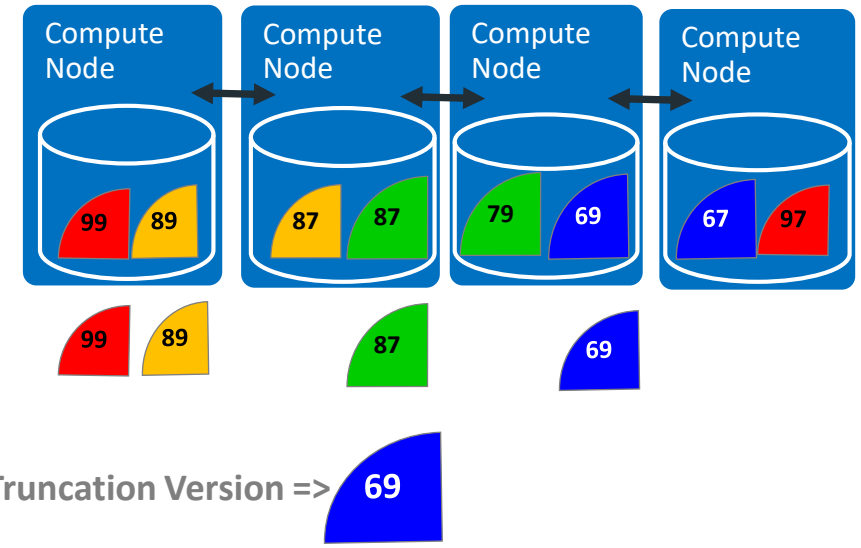
- In case of abrupt processing or instance termination you may need to remove deleted files.
  - `select clean_communal_storage();`
  - `select * from dc_communal_cleanup_records;`
- reaper queue tracks the files eligible for deletion
  - `select * from vs_reaper_queue;`
  - `select flush_reaper_queue();`
  - `select get_config_parameters('S3DeleteBatchSize'); --1000`
  - `select get_config_parameters('DelayForDeletes'); --2hrs`
  - `select get_config_parameters('ReaperCleanUpTimeoutAtShutdown'); --5 min`
  - `select get_config_parameters('FileDeletionServiceInterval'); --1 min`



Catalog sync in EON mode

- Every node syncs its catalog independently every 5 min.
- Catalog sync state
  - `select * from vs_catalog_sync_state;`
  - `select * from vs_catalog_truncation_status;`
  - `select sync_catalog();`
  - `select get_config_parameters('CatalogSyncInterval');`
- Catalog truncation version is the version that the cluster will revive to, in the event of abrupt termination.
  - `S3://<bucket>/metadata/<dbname>/cluster_config.json`
- The primary subcluster is responsible for sync-ing the catalog to S3
- data files and catalog snapshots before truncation version are eligible to be reaped/deleted.
- On shutdown we sync the catalog. When you terminate a cluster, ensure catalog has synced before you kill the instance or processes.



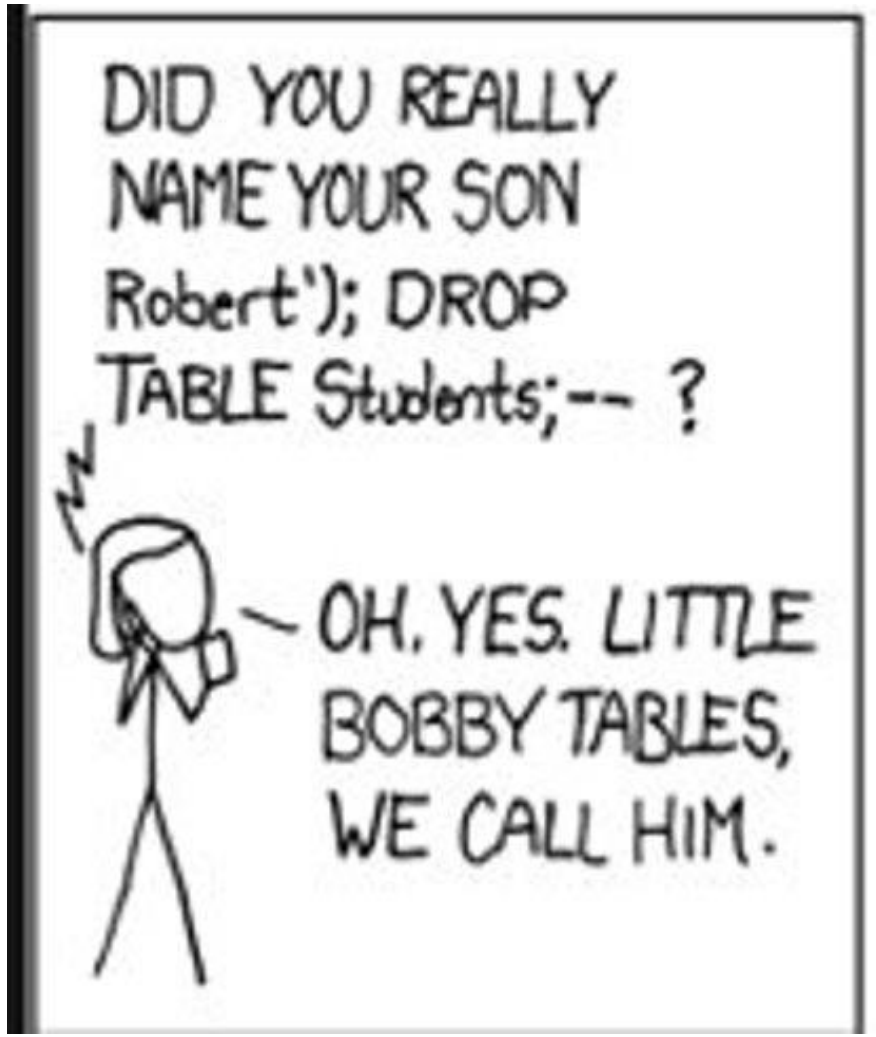


Backups

- You are on S3, do you really need to backup data? maybe
  - Designed for 99.9% availability over a given year (~ 8 hrs)
  - AWS will make commercially reasonable effort
  - And give you service credits
- S3 does not protect against accidental "DROP TABLE;"

Why not! - storage is cheap

- Replicate to a cluster in a different region (DR).
- Backup to a different region




VERTICA
Q&A
www.vertica.com
MICRO FOCUS

Vertica Academy

Your source for new comprehensive Vertica on-demand training

- Self-Paced Learning
- Online Training
- Knowledge Check
- Certificate of Completion & Certifications

Sign up for free today at **academy.vertica.com**

VERTICA

Vertica Forums

Vertica Engineers are available to answer your questions and moderate discussions right now

For more information visit **forum.vertica.com**