DESDM Databases

2014 Spring Collaboration Meeting

Todd Tomashek, NCSA

Database Cluster

- Eleven compute nodes
  - Single entry point (“service”) for client connections
  - Each node runs an “instance” of the database with shared storage
- 116 cores
- 135 drives, most 2TB; mirrored for redundancy
- Hosts three DES databases, plus a few others

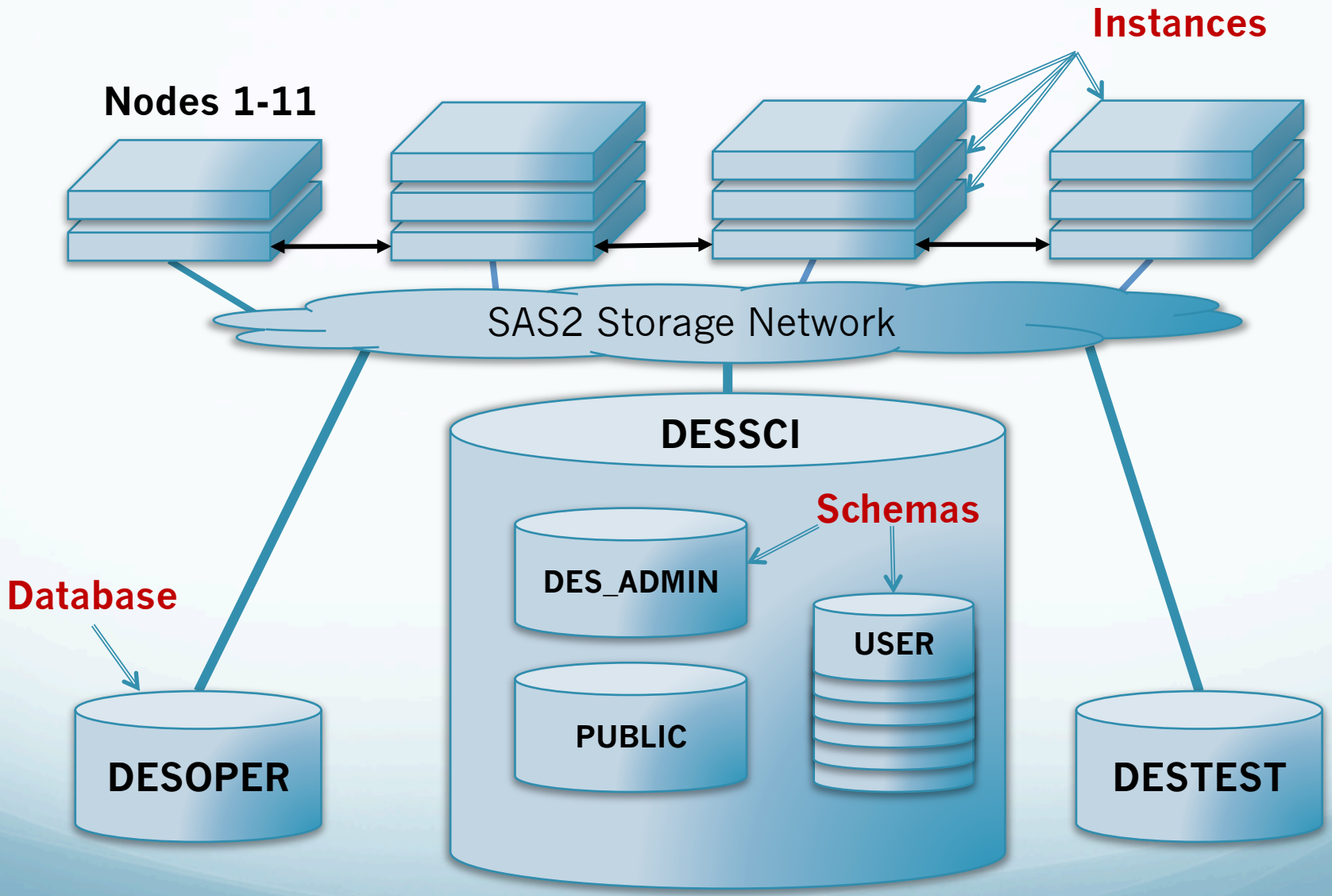
DES Databases

- DESOPER
  - Primary operations database
  - Optimized for data loading and operational use
- DESSCI
  - Primary user database
  - Optimized for data query
- DESTEST
  - Development and Test database
  - Accounts granted based on need

Database Schemas

- A “schema” in Oracle is a namespace
  - Tables, Views, Synonyms, and other objects must have unique names within a schema
- Every database user has a personal schema that is private by default. The schema name is the user name
  - References to “MyDB” are to this user schema
  - Quotas are enforced. Email “help” if you need more space
- A user owns all objects in his/her schema
- Production schema is DES_ADMIN
- PUBLIC is a globally-accessible schema
- Name resolution is login schema then PUBLIC schema

Overview


Instances
Nodes 1-11
SAS2 Storage Network
DESSCI
Schemas
DES_ADMIN
USER
PUBLIC
Database
DESOPER
DESTEST

Tables of Interest

- File Metadata
  - Location
    - Exposure
    - Image
    - Catalog
    - Coadd
- Objects
  - objects_current (desoper)
  - coadd_objects (desoper)
  - standard_stars_all (desoper)
  - sva1_objects (dessci)
  - sva1_coadd_objects (dessci)
  - USNOB_CAT1, UCAC2 (dessci)
  - APASS_DR7 (desoper)
- Processing
  - RUN
  - BLOCK
  - RUN_DATA_STATE
  - RUNTAG
  - EXPTAG
- Other Data
  - MOLYGON (mangle)
  - COADDTILE
  - COADD_SRC
  - coadd_objects_uniquetile
  - SN* (supernova tables)

Metadata Tables

- DES_TABLES (view based on all_tables)
- DES_INDEXES (view based on all_indexes, all_ind_columns)
  - ...where table_name='ALL_CAPS_TABLE_NAME'
- DES_SYNONYMS
  - "...where db_link is not null" to find linked objects
- ALL_TAB_COLUMNS
- ALL_IND_COLUMNS
- ALL_PART_KEY_COLUMNS
  - Use this to see which columns are partitioning keys

Synonyms and Database Links

- Most tables have synonyms in PUBLIC schema, which allows users to reference tables without the schema (so “select * from run” instead of “select * from des_admin.run”)
  - Name resolution looks in current schema first, then PUBLIC if not found
- Databases are “linked” so that users can perform queries against one database while connected to another.
  - For example, one could query the RUN table while connected to DESSCI
  - Use @<database> table suffix to look across a link
    - Select count(*) from image@DESOPER; while connected to dessci
- Some synonyms use a link to make a remote table appear local (so “select * from run” will work on dessci)
  - Create public synonym run for des_admin.run@desoper;

Fast Queries

- Query Optimizer
  - Powerful, but imperfect
  - DES_DB_JOB_PKG.des_explain_plan(querystring, ID_string)
    - ID_string is an arbitrary ID you want to assign to the plan. Typically your username plus a numeric suffix like “TOMASHEK_012345”
  - JOB_SUBMIT_EXPLAIN_PLAN
    - select * from job_submit_explain_plan where statement_id=‘TOMASHEK_012345’
- Selectivity is the key
  - Partitions (pruning)
  - Indexes (local)
    - Btrees (_BTX, _IDX)
      - Good for lookup-style operations with small number of predicates
    - Bitmap (_BMX)
      - Excellent for low-cardinality dimensions with multiple predicates
- Use DES_INDEXES and ALL_PART_KEY_COLUMNS
- Do not try to join multiple large tables across a DBlink

Object Query Details

- Y1P1_COADD_OBJECTS
  - Partitioned on RA, subpartitioned on RUN (48)
    - Queries that filter on RA and RUN will be able to ignore a significant portion of the table up front
  - Bitmap indexes on
    - ra, dec, tilename, catalogid_[girzY], imageid_*, class_star_*, spread_model_*, mag_model_*, mag_auto_*, mag_psf_*
  - ROWIDs can be found by ANDing bitmaps
- Views such as Y1P1_COADD_SPTE map fields to RUNs, thereby using the RUN partitioning to prune

Working in your Schema

- create table tablename ()
- create index indexname on tablename (column1,...)
  - BTree index by default
  - Use “create unique index” if unique
  - Use “create bitmap index” for bitmap index (no updates!)
- create table tablename as select ID, RA, DEC from objects
  - Very efficient way to create subset of data for further analysis
- To allow others to see your objects:
  - grant select,insert,update,delete on tablename to username
    - use “DES_READER” to give everyone read access
    - users will need to preface table name with your schema (your username):
      - select * from buckley.cool_new_catalog

Broadly Needed?

- If you create a dataset that will be of broadly and generally useful, email “help” to request that it be promoted to a common schema
  - I can then assist with partitioning, indexing, etc.
  - A PUBLIC synonym can be made so it is no longer necessary for users to preface with your schema

Loading Data

- Oracle's bulk data loading tool is SQL*Loader
- SQL*Loader can bypass logging and write directly to data files: orders of magnitude faster
- SQL*Loader is freely available with the full Oracle Client, which is available for many platforms (executable is sqlldr)
- Fairly simple to load ascii text files. Can also load binary data

Resources

- Clients
  - SQL Developer (free; http://www.oracle.com/technetwork/developer-tools/sql-developer/overview/index.html)
  - Toad for Oracle (very good, but fairly exspensive)
  - TrivialAccess (https://desweb.cosmology.illinois.edu/confluence/display/PUB/Downloading+and+Installing+the+DESDM+Client)
- Online documentation is comprehensive
  - Look for Oracle 11.2

Questions