Dataset Bookkeeping Service

Anzar Afaq & Vijay Sekhri
(For DBS Team)

CD/CMS - FermiLab

Outline

- DBS role in CMS
- DBS services
- DBS Query Language (QL)
- DBS CLI
- Data Discovery web interface
- Advance Features
  - DBS APIs
  - Analysis datasets
  - DBS MART

DBS

- DBS stands for Data Bookkeeping System
- It is authoritative source of information about data produced and stored in CMS
- It is meta-data service, therefore it does not hold data itself, only information about the produced data, such as dataset names, size, creation date, etc.
- DBS keeps its data in Databases. We use ORACLE and MySQL DBMS back-ends. Data is recorded/accessed through various interfaces.
- DBS has an ever growing number of deployments, some are centrally managed at CERN, some are at other T1/T2s.
- There is a single “SACRED” instance of DBS that’s called DBS-GLOBAL that contains “Blessed Datasets”.
- Read/Write access is controlled via Grid Certs, ReadOnly access is open to public.

Basic Concepts in DBS

- Output of data producing jobs are “Files” which have “Logical File Names” (LFNs) and Physical File Names (PFNs).
- Files are produced using a certain “release” of CMSSW, using a certain “config” file or set of “Parameter Sets”.
- Collection of files are logically grouped into File “Blocks”
- Set of files created using a specific “release”, creates a logical grouping called “Dataset”.
- A dataset name is combination of THREE parts
  - Primary Stream
  - Name given to that particular processing (Processed)
  - Type of process (Tier)
  - Represented as DatasetPath or simply “Path”
    - /primary/processed/tier(s)
    - /CSA07Muon/CMSSW_1_6_7-CSA07-Tier0-A1-Chowder/RECO
- Files live at Sites (storage), DBS only knows where it lives…

Interacting with DBS

- Users interact with DBS via several interfaces
  - Via DBS API (Written in Python, used by other tools like CRAB, ProdAgent, TOAST etc.). API also used by expert users.
  - Via DBS CLI (To perform data searches and other operations)
  - Via web interface
- The most convenient way of interaction with DBS search is to use DBS Query Language (DBS QL)
  - DBS QL can be used via DBS CLI or DBS Data Discovery Web Interface

DBS Query Language (QL)

<Vijay>

https://twiki.cern.ch/twiki/bin/view/CMS/QL

DBS CLI interface

- Easy to use unix-like CLI
- Part of CMSSW distribution
- A variety of data discovery operations supported
- Supports full syntax of Query Language (QL)
- Allow creation of Analysis Datasets (ADS)
- Save your queries locally (DBS-MART) for reuse and creation of ADS/PADS

Note: Python version and compatibility with libraries in CMSSW issues are being resolved

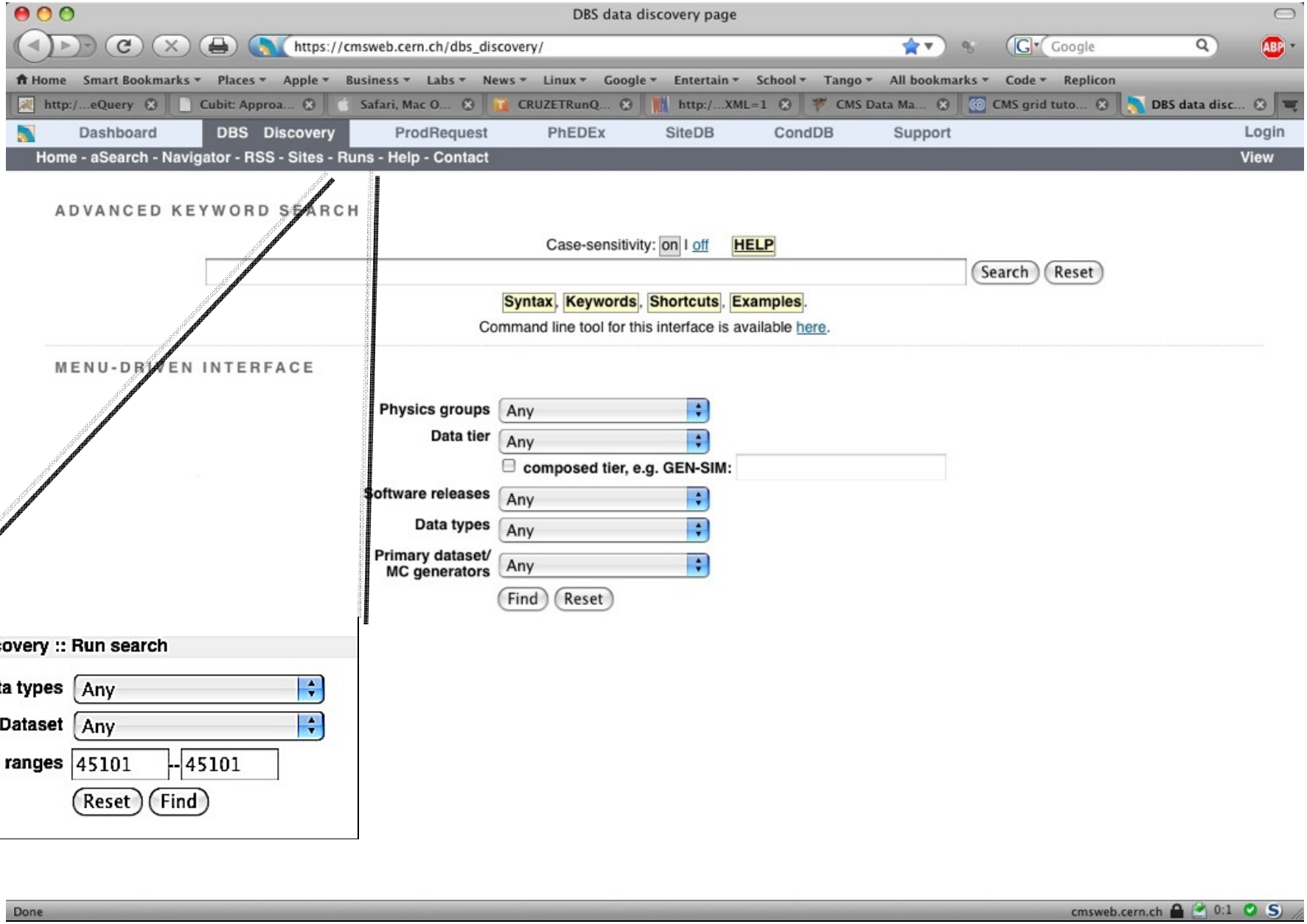
https://twiki.cern.ch/twiki/bin/view/CMS/DBSTutorial06202008

Data Discovery Web Page

- https://cmsweb.cern.ch/dbs_discovery/
- Various Data Discovery interfaces, aSearch, Navigator, Run,Site-searchers described at this URL
- https://twiki.cern.ch/twiki/bin/view/CMS/DataDiscoveryInterface

NOTE: I have some notes in slides below, but you can go through them yourself.

Data Discovery web interface(s)



Grid vs List view

Case-sensitivity: on | off

find dataset where release=CMSSW_2_0_0_pre5 Search

DBS discovery :: Adv. search :: Results **Physicist**

Found 8 processed dataset. Show all View results: grid | list mode Sort by creation date desc | asc

PATH	CREATED (dd/mm/yy)	SIZE	BLOCKS	FILES	EVENTS	SITES	CRAB	LINKS
/RelValSingleMuMinusPt10/CMSSW_2_0_0_pre5-RelVal-1205444401/RECO	13/02/08	2.0GB	1	cff, txt, (2)	4500	2	cfg	More info
/RelValSingleMuPlusPt1000/CMSSW_2_0_0_pre5-RelVal-1205444401/RECO	13/02/08	1.9GB	1	cff, txt, (2)	3900	2	cfg	More info
/RelValSingleMuMinusPt100/CMSSW_2_0_0_pre5-RelVal-1205444401/RECO	13/02/08	2.0GB	1	cff, txt, (2)	4500	2	cfg	More info
/RelValSingleMuPlusPt1/CMSSW_2_0_0_pre5-RelVal-1205444401/RECO	13/02/08	1.9GB	1	cff, txt, (2)	4500	2	cfg	More info
/RelValMinbias/CMSSW_2_0_0_pre5-RelVal-1205444401/RECO	13/02/08	4.0GB	1	cff, txt, (4)	5000	2	cfg	More info
/RelValSingleMuMinusPt1/CMSSW_2_0_0_pre5-RelVal-1205444401/RECO	13/02/08	1.9GB	1	cff, txt, (2)	4500	2	cfg	More info
/RelValSingleMuPlusPt10/CMSSW_2_0_0_pre5-RelVal-1205444401/RECO	13/02/08	2.0GB	1	cff, txt, (2)	4500	2	cfg	More info
/RelValSingleMuPlusPt100/CMSSW_2_0_0_pre5-RelVal-1205444401/RECO	13/02/08	2.0GB	1	cff, txt, (2)	4500	3	cfg	More info

Number of processed datasets per page 10 Result page: 1

find dataset where release=CMSSW_2_0_0_pre5 Search

DBS discovery :: Adv. search :: Results **Physicist**

Found 8 processed dataset. Show all View results: grid | list mode Sort by creation date desc | asc

/RelValSingleMuMinusPt10/CMSSW_2_0_0_pre5-RelVal-1205444401/RECO
Created 13 Mar 2008 22:47:28 GMT, contains 4500 events, 2 files, 1 block(s), 2.0GB, located at 2 sites (show, hide), LFNs: cff, plain
Block info , Run info , Conf. files , Parents , Children , Description , PhEDEx , Create ADS , ADS , crab.cfg , ProdRequest

/RelValSingleMuPlusPt1000/CMSSW_2_0_0_pre5-RelVal-1205444401/RECO
Created 13 Mar 2008 22:43:19 GMT, contains 3900 events, 2 files, 1 block(s), 1.9GB, located at 2 sites (show, hide), LFNs: cff, plain
Block info , Run info , Conf. files , Parents , Children , Description , PhEDEx , Create ADS , ADS , crab.cfg , ProdRequest

Sorting on Data Discovery

Found 9767 processed dataset. Show all (can be time consuming) View results: grid | list mode Sort by PATH desc | asc

PATH	CREATED (dd/mm/yy)	CREATOR	SIZE	BLOCKS	FILES	EVENTS	SITES	CRAB	∫L	LINKS
/zz3j-alpgen/CMSSW_1_6_7-HLT-1205907476/GEN-SIM-DIGI-RECO	19/02/08	ceballos Guillelmo Gomez Ceballos	15.1GB	1	cff, txt, (6)	5455	1	cfg	N/A	More info
/zz3j-alpgen/CMSSW_1_6_7-CSA07-1205907722/RECO	20/02/08	ceballos Guillelmo Gomez Ceballos	20.6GB	1	cff, txt, (6)	5455	1	cfg	N/A	More info
/zz3j-alpgen/CMSSW_1_6_7-CSA07-1205736930/GEN-SIM-DIGI-RAW	17/02/08	ceballos Guillelmo Gomez Ceballos	12.1GB	1	cff, txt, (6)	5455	1	cfg	N/A	More info
/zz3j-alpgen/CMSSW_1_4_9-CSA07-4131/GEN-SIM	10/02/08	Jose-Hernandez	5.4GB	1	cff, txt, (5)	5455	1	cfg	N/A	More info

Found 9767 processed dataset. Show all (can be time consuming) View results: grid | list mode desc | asc

- ✓ Sort by PATH
- Sort by CREATIONDATE
- Sort by CREATEDBY
- Sort by TOTALSIZE
- Sort by NUMBEROFBLOCKS
- Sort by NUMBEROFFILES
- Sort by NUMBEROFEVENTS
- Sort by NUMBEROFSITES

PATH	CREATED (dd/mm/yy)	CREATOR	SIZE	BLOCKS	FILES	EVENT	LINKS
/zz3j-alpgen/CMSSW_1_6_7-HLT-1205907476/GEN-SIM-DIGI-RECO	19/02/08	ceballos Guillelmo Gomez Ceballos	15.1GB	1	cff, txt, (6)	545	e info
/zz3j-alpgen/CMSSW_1_6_7-CSA07-1205907722/RECO	20/02/08	ceballos Guillelmo Gomez Ceballos	20.6GB	1	cff, txt, (6)	5455	1 cfg N/A More info
/zz3j-alpgen/CMSSW_1_6_7-CSA07-1205736930/GEN-SIM-DIGI-RAW	17/02/08	ceballos Guillelmo Gomez Ceballos	12.1GB	1	cff, txt, (6)	5455	1 cfg N/A More info

PATH	CREATED (dd/mm/yy)	CREATOR	SIZE	BLOCKS	FILES	EVENTS	SITES	CRAB	∫L	LINKS
/CSA07AllEvents/CMSSW_1_6_7-CSA07-Tier0-A1-Gumbo/RECO	08/10/07		195.3TB	1090	cff, txt, (73639)	110038983	6	cfg	N/A	More info
/CSA07AllEvents/CMSSW_1_6_7-CSA07-Tier0-A2-Stew/RECO	20/09/07		159.9TB	880	cff, txt, (79074)	78611634	8	cfg	N/A	More info
/CSA07AllEvents/CMSSW_1_6_7-CSA07-Tier0-A3-Chowder/RECO	29/10/07		140.7TB	766	cff, txt, (38758)	63139589	3	cfg	N/A	More info
/CSA07AllEvents/CMSSW_1_6_7-CSA07-Gumbo-B1-PDAllEvents-ReReco-100pb/RECO	29/11/07		115.1TB	496	cff, txt, (49521)	65037607	5	cfg	N/A	More info

More info tips

Step 1: Find file with run 34850

Case-sensitivity: on | off

find file where run=34850 Search Reset

DBS discovery :: Adv. search :: Results **Physicist**

Found 2 logical file name. Show all View results: grid | list mode Sort by LOGICALFILENAME desc | asc

LOGICALFILENAME	CREATED (dd/mm/yy)	CREATOR	CHECKSUM	EVENTS	FILESIZE	TYPE	STATUS	LINKS
/store/data/CalPrivateGlobal/A/000/034/850/CalPrivateGlobal.00034850.0001.A.storageManager.0.0000.dat	12/02/08	tier0	0	4	218679	STREAMER	VALID	More Info
/store/data/CalPrivateGlobal/A/000/000/000/RAW/0000/A05C6C32-75D9-DC11-BEF6-001617C3B652.root	14/02/08	tier0	4188673584	4	230773	EDM		

Number of logical file names per page 10 Result page: 1

- ✓ More Info
- Find run
- Find physics group name
- Find analisis dataset path
- Find storage element
- Find processed dataset
- Find analisis dataset version
- Find data tier
- Find analisis dataset name
- Find luminosity section
- Find software release
- Find primary dataset
- Find processed dataset
- Find block

Step 2: Now lookup dataset for one of the file

Case-sensitivity: on | off

find dataset where file=/store/data/CalPrivateGlobal/A/000/000/000/RAW/0000/A05C6C32-75D9-D(Search Reset

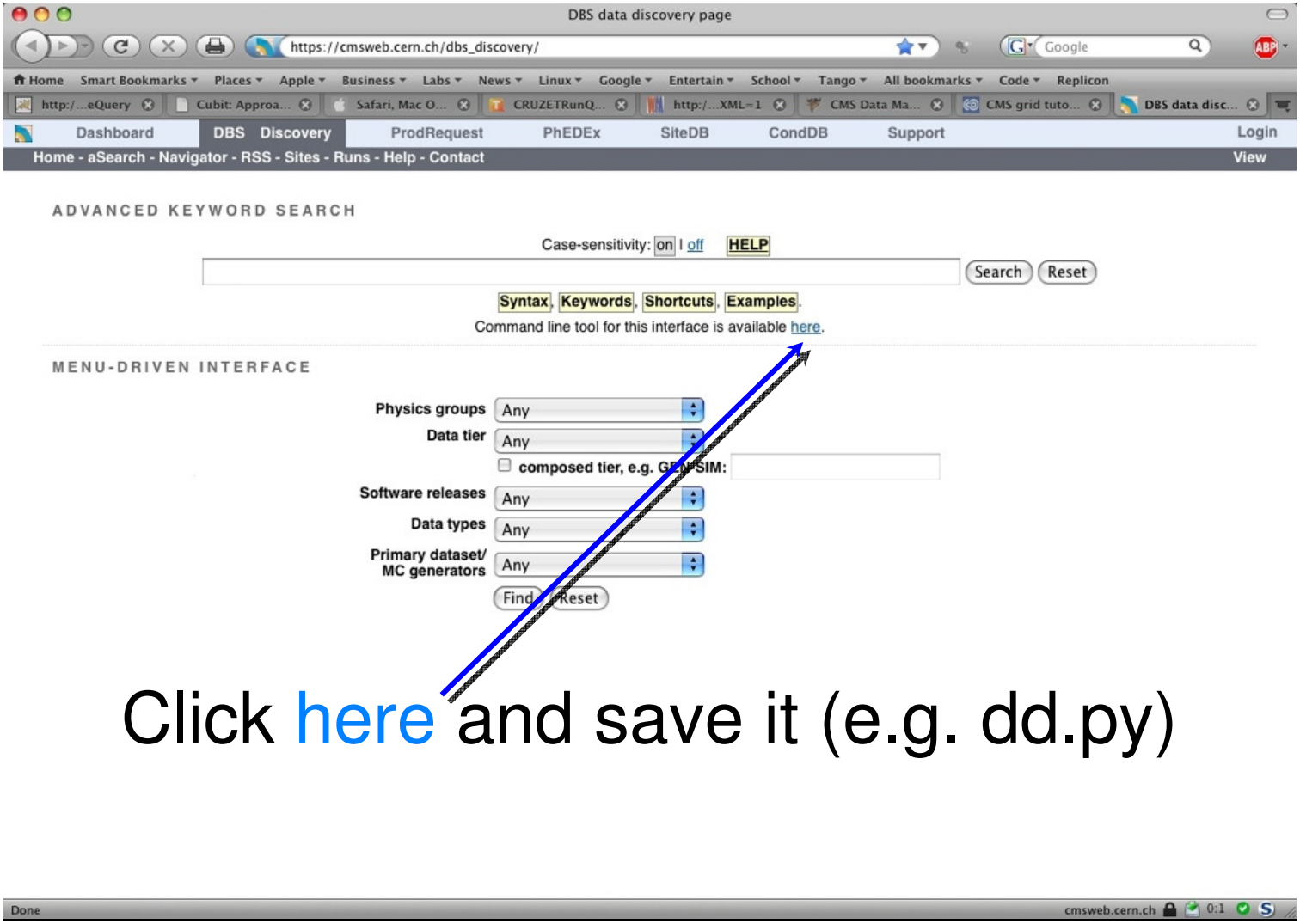
DBS discovery :: Adv. search :: Results **Physicist**

Found 1 processed dataset. Show all View results: grid | list mode Sort by PATH desc | asc

PATH	CREATED (dd/mm/yy)	CREATOR	SIZE	BLOCKS	FILES	EVENTS	SITES	CRAB	∫L	LINKS
/CalPrivateGlobal-A/Online-CMSSW_1_7_5/RAW	06/01/08	tier0	28.5GB	39	cff, txt, (672)	5431134	3	cfg	N/A	More info

Number of processed datasets per page 10 Result page: 1

aSearch Data Discovery CLI



Click here and save it (e.g. dd.py)

DBS APIs

- DBS has a python API which can be used to interact with DBS programmatically.
- DBS APIs is used by CRAB, ProdAgent, etc., but you can use them on your own.
- There are two types of DBS APIs
  - list API, e.g. listPrimaryDataset, in order to get information from DBS
  - admin API, e.g. insertPrimaryDataset, to administer data in DBS
- DBS API docs
  - https://twiki.cern.ch/twiki/bin/view/CMS/DBSApiDocumentation
  - https://twiki.cern.ch/twiki/bin/view/CMS/DBS_Client_API_Doc
- See <path>/DBS/Clients/Python/DBSAPI/UserExamples

Using DBS API

```
#!/usr/bin/env python

import sys, optparsefrom DBSAPI.dbsException import *from
DBSAPI.dbsOptions import DbsOptionParserfrom DBSAPI.dbsApi import
DbsApi
optManager  = DbsOptionParser()(opts,args) = optManager.getOpt()

# init DBS API
api = DbsApi(opts.__dict__)

# find and print all primary datasets
print "Primary Datasets"for primary in api.listPrimaryDatasets('*'):
     print "  %s"%primary
```

Thanks

- cms-dbs-support@cern.ch
- https://twiki.cern.ch/twiki/bin/view/CMS/DBS-TDR