



DBS/DLS Data Management and Discovery

Lee Lueking

3 December, 2006

Asia and EU-Grid Workshop
1-4 December, 2006



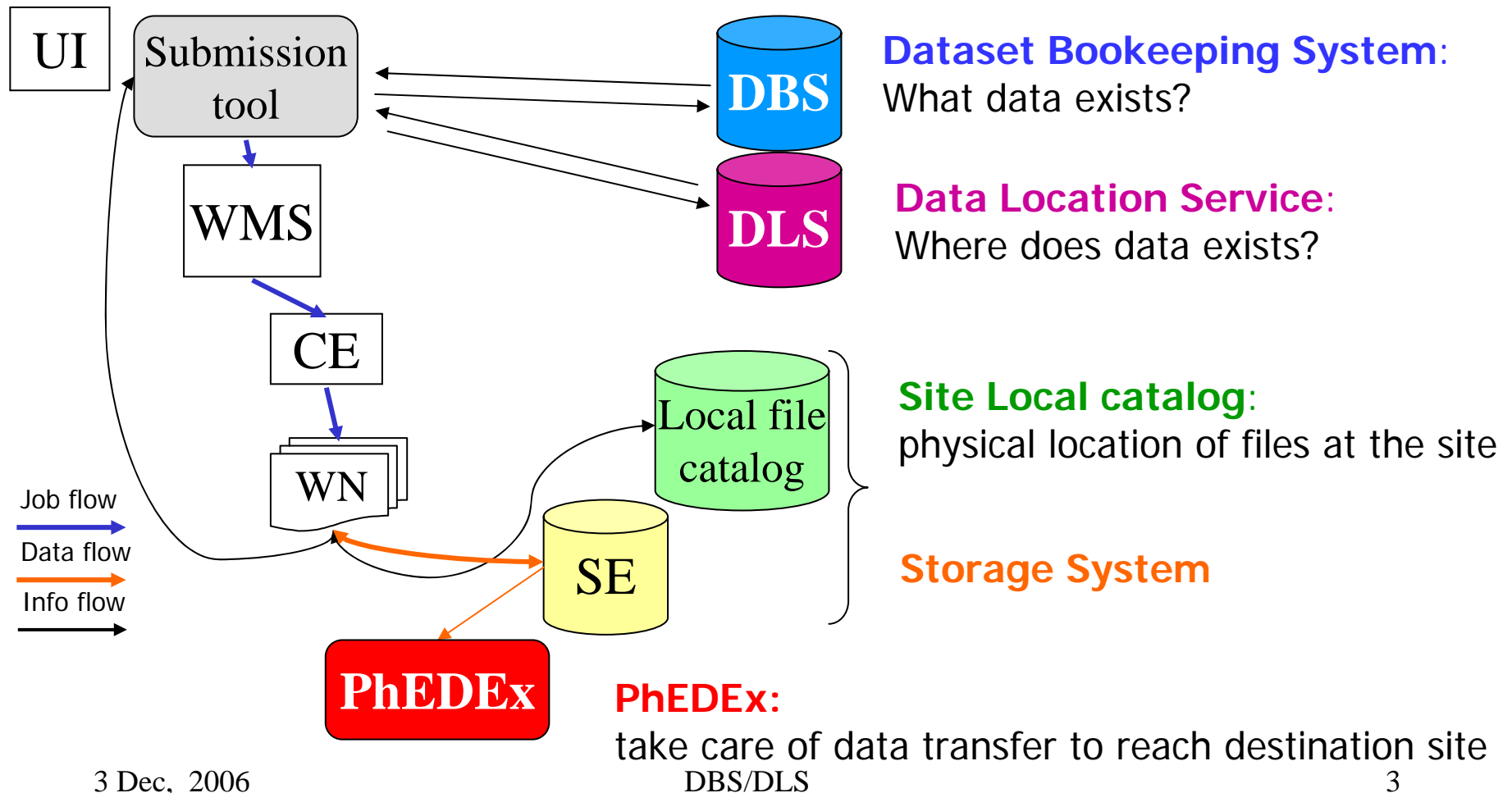
Contents

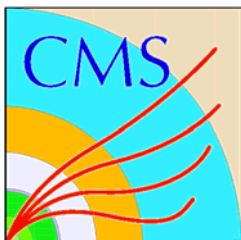
- Overview of DBS/DLS
- CMS Data Management Concepts
- DBS/DLS Data Discovery



Data processing workflow

- Data Management System allow to discover, access and transfer event data in a distributed computing environment





Dataset Bookkeeping System (DBS)

- DBS provides the means to define, discover and use CMS event data.

Data definition:

- Dataset specification (content and associated metadata)
- Track data provenance

Data discovery:

- What data exists
- Dataset organization in terms of files/fileblocks
- Site independent information



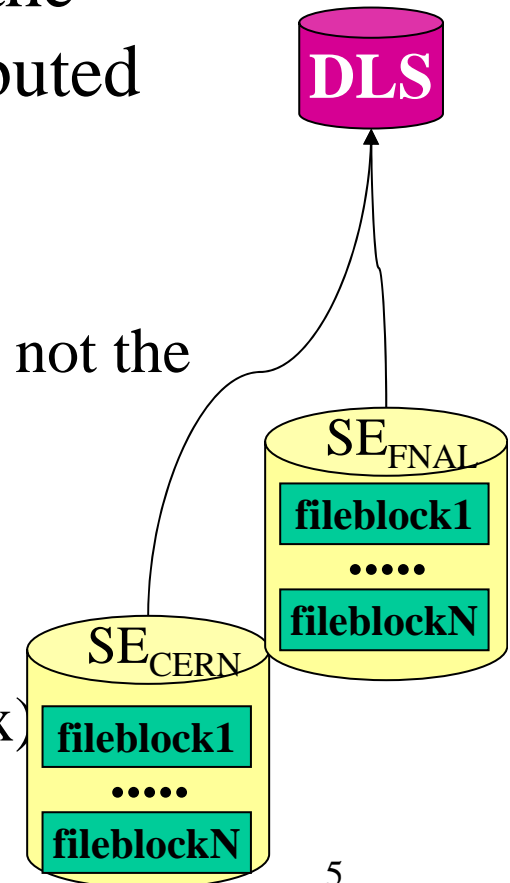
Use:

- Distributed analysis tool (CRAB)
- MC Production system
- Data distribution tool (PhEDEx)
- User data discovery



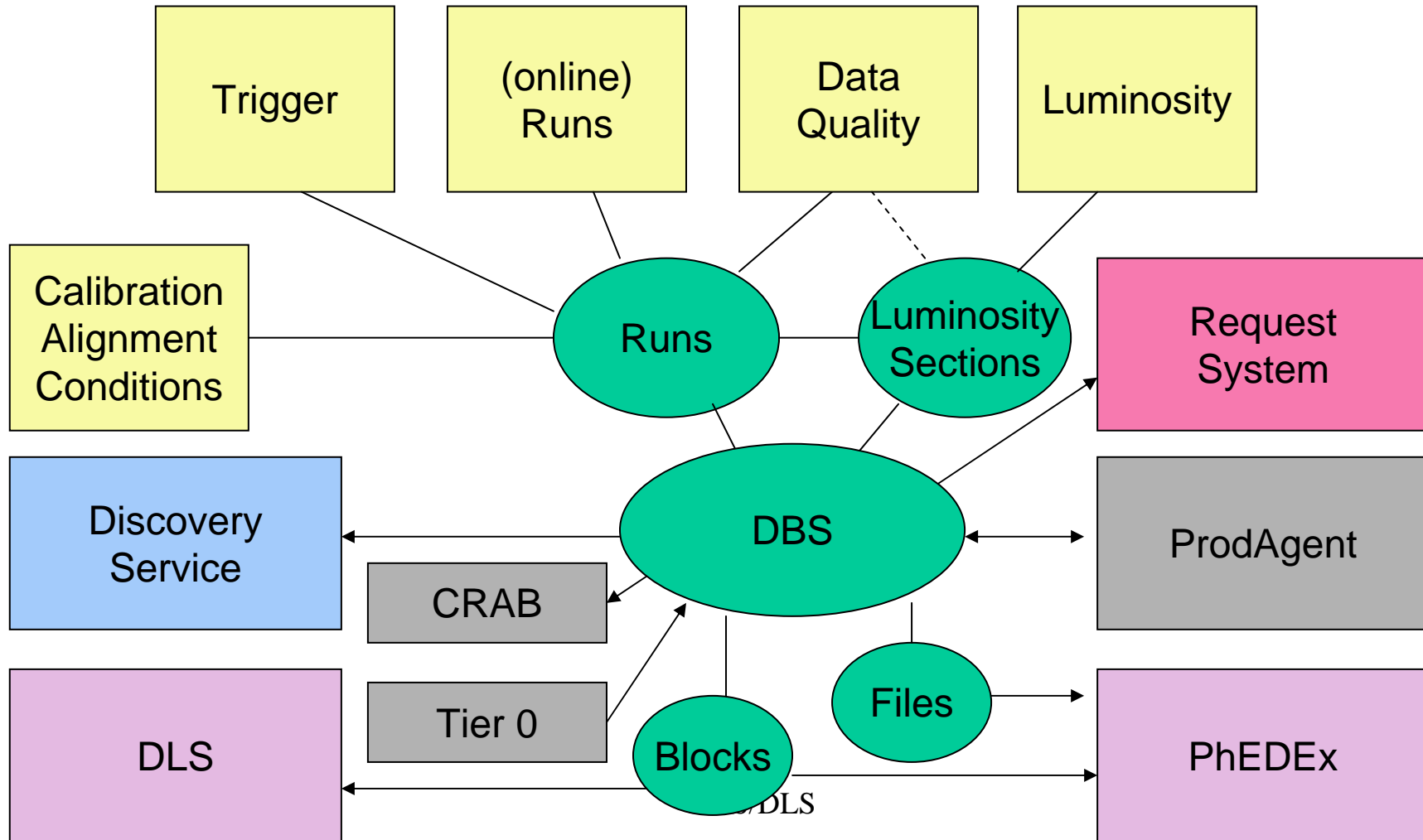
Data Location Service role

- The Data Location Service (DLS) provides the means to locate replicas of data in the distributed computing system
 - Maps file-blocks to storage elements (SE's)
 - Provide only names of sites hosting the data and not the physical location of constituent files at the sites
- Interactions with DLS
 - Insert file-blocks produced at a site (Production)
 - Insert file-blocks upon data replication (PhEDEx)
 - Query to locate file-blocks (CRAB, Production)





Relationships to DBS





Concepts: Dataset

- A set of data representing a coherent sample for analysis
 - **Primary Dataset:** determined by HLT event classification or MC production parameters
 - **Processed Dataset:** a slice of a primary dataset with a consistent processing history. Note: *May include multiple copies of some events with slight differences in processing.*
 - **Analysis Dataset:** a snapshot of a subsets of processed dataset representing a coherent sample for physics analysis
- CMSSW can produce merged data that do not conform to these classifications. We call this “*fruit salad*” and data of this nature is also accommodated.



Concepts: Lumi Sections, Data Tier

- Luminosity Section
 - A period of approximately constant instantaneous luminosity
 - Unit of accounting for integrated luminosity
 - Production data files will contain whole luminosity sections
- Data Tier
 - A set of objects grouped together in output files
 - Defined by the release configuration files
 - *Data Tier definitions may change slightly with major releases*



DBS/DLS Data Discovery

Valentin Kuznetsov
Cornell University




DBS/DLS Discovery Intro

- Set of tools which provide CLI and Web interfaces
- Combines information from DBS and DLS to find out your data
- Provides three avenues to data discovery:
 - ‘navigator menu’ approach
 - ‘keyword’ search
 - ‘site’ search



Navigator Menu

 **DBS/DLS DATA DISCOVERY PAGE**

Navigator	Navigator menu
Keyword search	DBS instances <input type="text" value="MCGlobal/Writer"/>
Site	Tier sites <input type="text" value="All"/>
Datasets	Application <input type="text" value="/CMSSW_1_0_6/Skimming/cmsRun"/>
Summary	Primary dataset <input type="text" value="CSA06-103-os-EWKSoup0-0"/>
History	Data tier <input type="text" value="All"/>
About...	<input type="button" value="Find"/>
Hide panel	DBS glossary

Drop down menus adjust hierarchically and quickly help you to find your data



Keyword Search

DBS/DLS DATA DISCOVERY PAGE

Navigator

- Keyword search**
- Site
- Datasets
- Summary
- History
- About...
- Hide panel

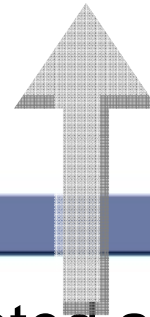
Data keyword search

The search is case insensitive and the following special symbols are supported: '(', ')', 'and', 'or' and 'not'.

You may use boolean expressions, e.g., (word1 or (word3 and word4) and not word2)

Any keywords:

Advanced search



Search keywords are evaluated as normal python expression and they are case insensitive



Site Search

DBS/DLS DATA DISCOVERY PAGE

Navigator

- Keyword search
- Site**
- Datasets
- Summary
- History
- About...
- Hide panel

Site search

Use this form to show detailed information about particular site.

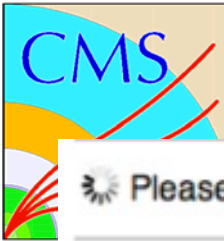
NOTE: the DLS queries may take a lot of time, since they go through LFC.

Choose DBS instance

Please select a site:

Description

Mostly for site managers, who wants to know which data resides on their site



Results

Please wait, while we retrieve your data

Results Parents Validation Parameter Set Release Specs

Processed datasets (plain view):

[/CSA06-103-os-EWKSoup0-0/RECO SIM/CMSSW_1_0_4-hg_Higgs_mc2gamma_Filter-1161045561](#)

contains 2797 events, 18 files, 1.4GB.

Show: **Blocks Summary Both**

row	Location	Events	Files	size
1	castor.sc.grid.sinica.edu.tw	1709	11	890.5MB
2	sc.cr.cnaif.infn.it	1088	7	572.4MB

Show: **Blocks Summary Both**

row	Location	Events	Files	status	size	LFN list
2_1	sc.cr.cnaif.infn.it	1088	7	OPEN	572.4MB	cff, plain /CSA06-103-os-EWKSoup0-0/RECO SIM/CMSSW_1_0_4-hg_Higgs_mc2gamma_Filter-1161045561-4e7d-950e-79295...
1_1	castor.sc.grid.sinica.edu.tw	1709	11	OPEN	890.5MB	cff, plain /CSA06-103-os-EWKSoup0-0/RECO SIM/CMSSW_1_0_4-hg_Higgs_mc2gamma_Filter-1161045561-4a6a-9662-1a4c8...



CLI interface

```
shell# cvs co COMP/{DBS,DLS}; cd COMP/DLS/Client; make; cd </path>/COMP/DBS/Web/DataDiscovery
shell# . scripts/setup.sh
shell# ./DBSHelper.py --help
usage: DBSHelper.py [options]
```

options:

```
-h, --help          show this help message and exit
--quiet            be quiet and don't print exceptions
--dict=DICTIONARY use to generate JavaScript dictionary, pass Global/All
--primaryDataset=PRIMD
                    specify primary dataset, e.g.
                    --primaryDataset=CSA06-081-os-minbias
--dataTier=DT      specify Data Tier within dataset, e.g. --dataTier=RECO
--app=APP          specify application keys (version,family,exe), e.g.
                    --app=CMSSW_0_8_1,Merged,cmsRun
--dbsInst=DBSINST specify DBS instance to use, e.g.
                    --dbsInst=MCLocal_1/Writer
--showProcDatasets be quiet and show only processed datasets
--site=SITE        specify DLS site you're interesting, e.g.
                    --site=fnal.gov
--search=SEARCH    specify any keywords to search your data, e.g.
                    --search='CMSSW_0_8_1 and Merged'
-v, --verbose      be verbose
shell# ./DBSHelper.py --dbsInst=MCLocal_2/Writer --app=CMSSW_1_0_4,Merged,cmsRun --primaryDataset=CSA06-103-os-
ZMuMu0-0
```

3 Dec, 2006

DBS/DLS

15



Lucene extension

- Lucene is a high-performance, full-featured text search engine
- Plan to use it for indexing DBS information, e.g. CMSSW configuration parameter sets for advanced searches
 - search with relational qualifiers for key-value pairs
 - ‘google-like’ searches give rating for each found hit
 - work in progress to provide an in-memory and file index



Summary

- DBS is the catalog of all existing CMS data and information about how it was produced.
- DLS tells us where the data is physically located.
- A DBS/DLS Discovery system provides a convenient way to find the data you are interested in, and know where it is located.
- Note: New features and improvements are constantly being added. Your feedback to us is very useful.



Finish