



Science & Technology
Facilities Council

Ingesting Data into ICAT

Kevin Phipps (RAL/STFC)

CLF and ISIS Service Manager
& ICAT Developer

October 2016

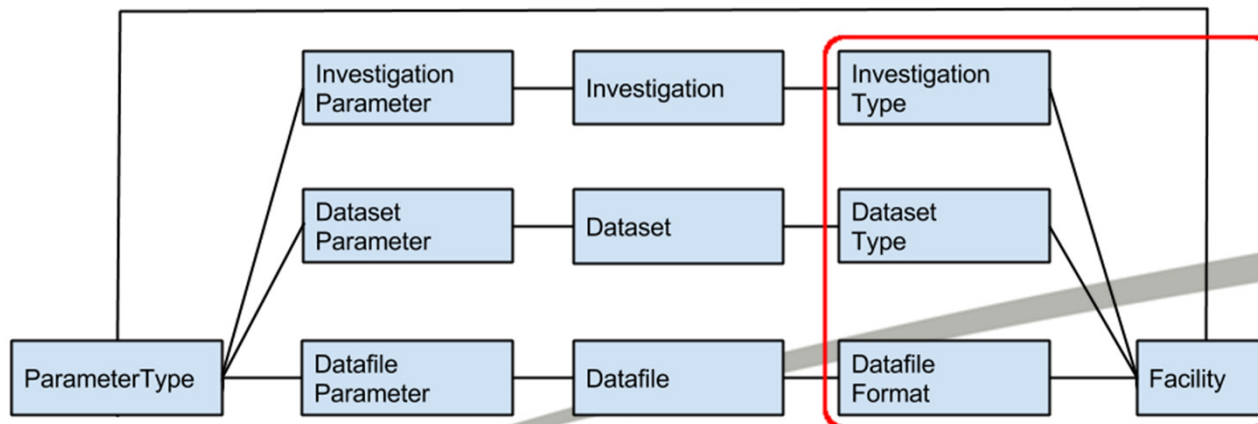
Overview

- Preparing the ICAT
- Setting up Investigations and Users
- Discovering new data
- Making data searchable
- The IDS – ingesting with and without
- Performance and scalability
- ISIS ingestion – a closer look



Preparing the ICAT

- Functional ingest account - CRU on most entities
- Create Facility, InvestigationTypes, DatasetTypes and DatafileFormats
- Create Instruments (for linking via InstrumentScientist and InvestigationInstrument)
- All just need a name (DatafileFormat also version)
- Code this to make it repeatable



Investigations and Users

- Set up in advance of experiment
- Export data from User Office system
- Nightly/weekly job
- Create Investigations (name, visitId, title)
- Link Investigation to Instrument(s)
- Create User (name = ldap/rqw38472, fullName and email useful)
- Link Investigation to User and assign a role (for use in Rules)

Week beginning	Gemini	
	Gemini	ATA2
06-Jan-15	System startup	
12-Jan-15		
19-Jan-15		
26-Jan-15	Ma	
02-Feb-15	13210052	
09-Feb-15		
16-Feb-15	Maintenance	
23-Feb-15		
02-Mar-15	McKenna	
09-Mar-15	13210040	
16-Mar-15		
23-Mar-15		
30-Mar-15		
06-Apr-15	System Access	
13-Apr-15		
20-Apr-15		
27-Apr-15		
04-May-15	Maintenance	
11-May-15		
18-May-15	Jaroszynski	
25-May-15	15110013	
01-Jun-15		
08-Jun-15		Hooker
15-Jun-15		15110002
22-Jun-15		
29-Jun-15		
06-Jul-15	Open Week	
13-Jul-15	Laser	
20-Jul-15	Quantel Service	
27-Jul-15		
03-Aug-15		Hooker
10-Aug-15	Najmudin + Lopes	15110002
17-Aug-15	15110009 + 1511008	
24-Aug-15		
31-Aug-15		
07-Sep-15		
14-Sep-15		
21-Sep-15	Maintenance	
28-Sep-15		
05-Oct-15		
12-Oct-15		
19-Oct-15	Zepf	
26-Oct-15	15110010	

Discovering new data

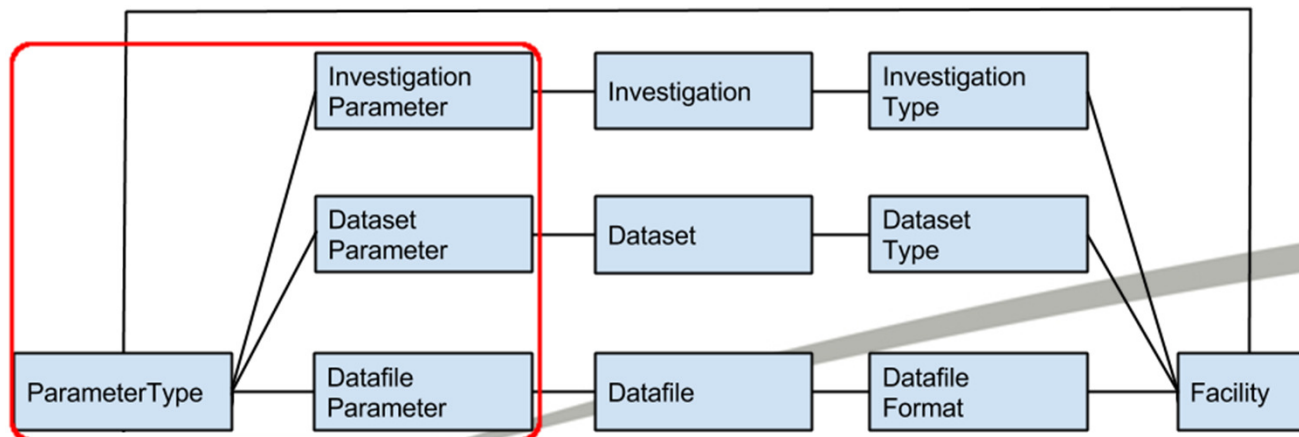
- Methods in use:
 - single input directory
 - filesystem watcher
 - trigger/drop files
- Which Investigation?
 - use the file name or path
 - metadata in the file or drop file
 - file timestamp and Instrument
- What is a Dataset?
 - shot, run, scan etc.

```
20150329 GS 00114835 S_LEG2_GREEN_NF.dat
20150329 GS 00114835 S_LEG2_GREEN_NF.mdt.xml
20150329 GS 00114835 S_LEG2_GREEN_NF.png
20150329 GS 00114835 S_PUMP_TIMING_TRACE.dat
20150329 GS 00114835 S_RED_TIMING_TRACE.dat
20150329 GS 00114835 S_SCINTILLATOR_TRACE.dat
```



Making data searchable

- Use Parameters (Investigation, Dataset and Datafile) to store metadata
- Name-value pairs of type NUMERIC, STRING, DATE_AND_TIME
- Nothing in a Datafile is searchable unless you create DatafileParameters



The IDS – ingesting with or without

- With:
 - for 2 level storage
 - if the data needs moving anyway
- Without:
 - large disk storage accessible from instruments
 - just create a Datafile entry in ICAT
 - add any metadata as DatafileParameters

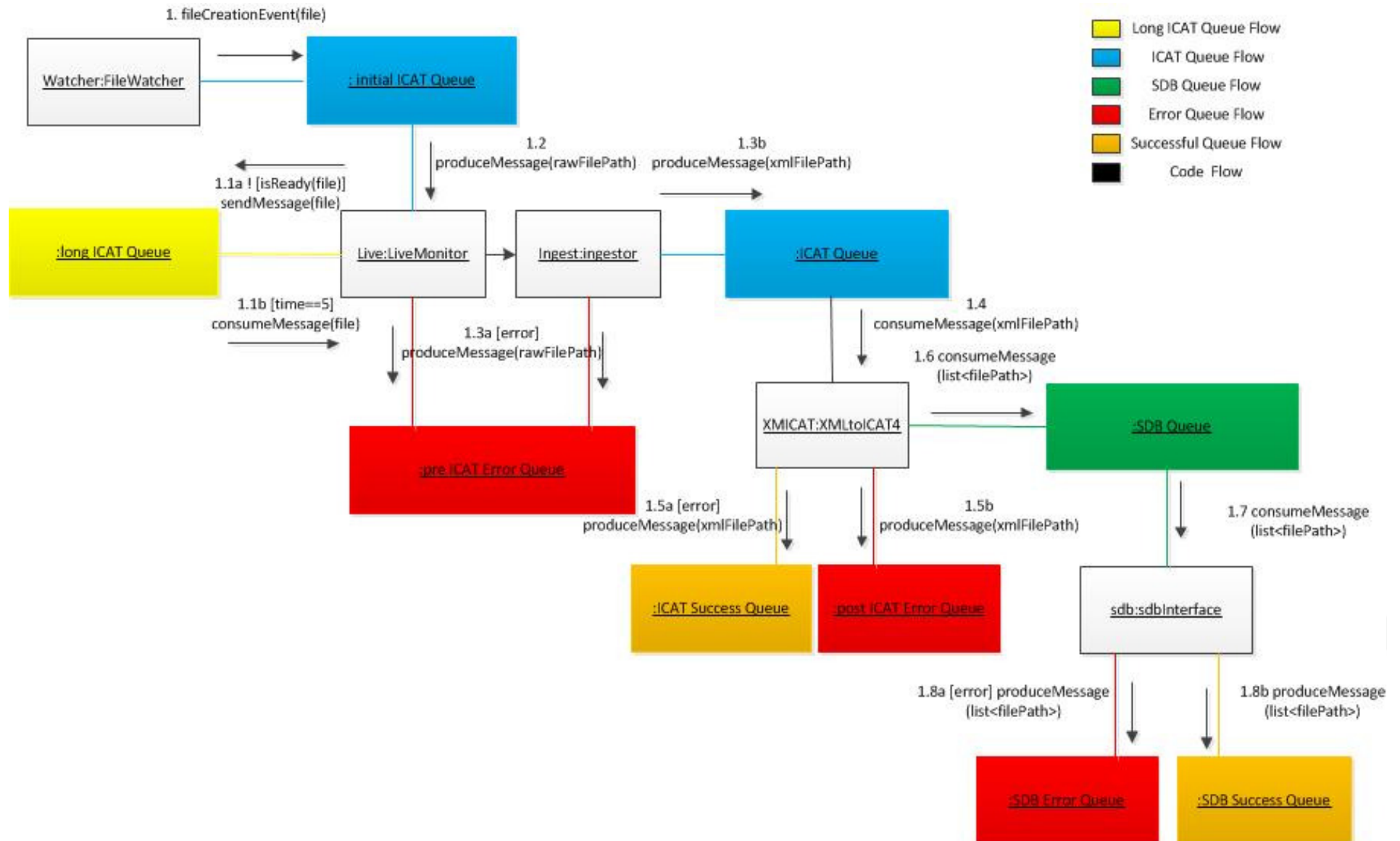


Performance and scalability

- Cache entities used frequently – DatasetTypes, DatafileFormats, ParameterTypes
- Cache recently used Investigations and Datasets
- Make it multi-threaded – one thread for queueing plus multiple ‘worker’ threads
- Design around an enterprise queueing system and run multiple ingestion program instances



ISIS ingestion – a closer look



Conclusions

- Import Investigation and User information from a User Office
- Decide how searchable data needs to be and use Parameters accordingly
- Size of available file stores and/or budget will determine the IDS setup
- Design for performance and scalability





Science & Technology
Facilities Council

Thanks for listening.

Questions ?