

# Large Synoptic Survey Telescope (LSST) Data Management

# Technical items to honor a tech great

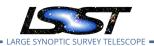
William O'Mullane

**DMTN-130** 

Latest Revision: 2019-09-11

### **Abstract**

We have been asked to consider naming some part of the technical system in honor of Jim Gray. This document is intended to give some background and options for that.



# **Change Record**

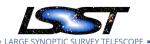
Version	Date	Description	Owner name
1	YYYY-MM-	Unreleased.	William O'Mullane
	DD		

Document source location: https://github.com/lsst-dm/dmtn-130



# **Contents**

1	Introduction	1
2	Things to name	1
	2.1 The Archive	2
	2.2 The Data Facility	2
	2.3 Qserv	2
	2.4 Butler - LSST Database	2
	2.5 Prompt Processing	3
	2.6 Science Platform	3
	2.7 imgServ	3
Α	References	3
В	Acronyms	3



# Technical items to honor a tech great

#### 1 Introduction

Tony Tyson has been contacted by Bill Gates with the idea to have some part of the DM technical stack named after Jim Gray. Gray while at IBM was one of the founders of relational databases, at Microsoft research he gave us Skyserver and helped make CasJobs work. As DM project manager and someone who worked with Jim on SDSS I think this is an excellent idea. There was mention of some donation for this too but that is really icing - the fact the Gates would like to honor Jim Gray in some way through LSST is fitting and good. This document gives some ideas on the topic. We need a list of potential candidates for Tony to discuss with Gates.

We are totally unclear on how much the agencies may care about any of these. Something to discuss here may be hosting "the archive" on Azure for free or highly discounted.

## 2 Things to name

We could look at our oft maligned product tree in DM for this. Under software we have:

- · Batch Production
- · Database Back Bone
- LSST Science Platform
- Prompt Processing
- Science Pipelines
- Quality Control
- Supporting including Qserv, Butler, task, ADQL and imgServ.

Of the above not all are useful - QC, Batch, DBB, are not very visible. ADQL seems small though relevant.



#### 2.1 The Archive

Not mentioned in the product list is the archive, this is somewhat related to Section 2.4. The archive is the collection od images and catalogs from LSST. It will always exist even after LSST. It is accessed through the Science Platform (Section 2.6) and Qserv (Section 2.3) is part of that. Any of the components may change but the archive would always be there.

This is probably the least contention item in DM for this purpose. It is also highly visible and NSF may care if we start using the term Jim Gray LSST Archive or such.

#### 2.2 The Data Facility

In operations the NCSA end of things will be labeled the LSST Data Facility and might provide an opportunity. It is a physical location and so could have a plaque. Physical locations and plaques will almost certainly draw the interest/involvement of NSF which may complicate things.

#### 2.3 Qserv

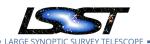
In terms of appropriateness, Gray being one of the SQL founders, a database would be very appropriate, a relational system like Qserv would be the top thing. <sup>1</sup> One worry may be the lifetime of such a product. According to Tony there is no indication of a need to be Microsoft1t based. One wonders though if a good path forward for Qserv might indeed be a collaboration with Microsoft to make a Qserv on Sqlserver, something which would make MyDB easier and could actually lead to a long term product like the Jim Gray Petascale DB, or perhaps Grayscale DB.

We are unsure of the status of Graywulf from JHU which was a DB system named in honor of Jim Gray. Qserv is post Jim Gray era - and they are a little hesitant on this option.

#### 2.4 Butler - LSST Database

We do not really think of the Data Release as a database but one could consider the data system underlying the science platform as a database and name it. This would give longevity as it will always exist even if the technology changes. The front end manifestation of this is the

<sup>&</sup>lt;sup>1</sup>Jim's Turing Prize citation was: "For seminal contributions to database and transaction processing research and technical leadership in system implementation"



Butler, which contains the relational registry of all image metadata. I imagine it might amuse Jim if in our code we had jim\_gray.get(...) and jim\_gray.put(...), some might find it disrespectful. This would require some code change but if there is a large donation it may be doable.

#### 2.5 Prompt Processing

There are two things here which one could potentially name: the alert stream and the prompt products database. Both will be long lived in the project. The alert stream is of course one of th highest profile parts of LSST. On May imagine most people would still call it the alert stream but it it could be branded and referred to officially as the Jim Gray Alert Stream or such.

#### 2.6 Science Platform

The science platform will definitely be long lived - even if the technology changes the name will stick, So one could consider this an viable option. One may in this case at least want to consider an Azure deployment, they were at least considering supporting K8s (check). Jim was not a big proponent of open software and this is a large open software project (true for all DM) - so there is some reluctance the use the science platform for this opportunity.

#### 2.7 imgServ

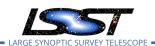
The image service could be an option. It will be long lived, given Jim's association with Sky-server and Terraserver this would be appropriate. It is an underlying service and may not be very visible though.

#### A References

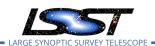
#### References

## **B** Acronyms

Acronym	Description
ADQL	Astronomical Data Query Language



Alert  Batch Pro-	A packet of information for each source detected with signal-to-noise ratio > 5 in a difference image during Prompt Processing, containing measurement and characterization parameters based on the past 12 months of LSST observations plus small cutouts of the single-visit, template, and difference images, distributed via the internet  Computational processing that is executed as inputs become available, in a
duction	distributed way across multiple enclaves when needed, while tracking sta-
duction	tus and outputs. Examples of Batch Production include offline processing
	for Prompt data products, calibration products, template images, and Spe-
	cial Programs data products. Prioritization protocols for the various types
	of batch production are given in LDM-148
Butler	A middleware component for persisting and retrieving image datasets (raw
	or processed), calibration reference data, and catalogs
DB	DataBase
DBB	Data Back Bone
DM	Data Management
DMLT	DM Leadership Team
DMTN	DM Technical Note
Data Release	The approximately annual reprocessing of all LSST data, and the installa-
	tion of the resulting data products in the LSST Data Access Centers, which
	marks the start of the two-year proprietary period
IBM	International Business Machines
LSST	Large Synoptic Survey Telescope
Prompt Pro-	The processing that occurs at the Archive Center on the nightly stream of
cessing	raw images coming from the telescope, including Difference Imaging Anal-
	ysis, Alert Production, and the Moving Object Processing System. This pro-
	cessing generates Prompt Data Products
QC	Quality Control
Qserv	Proprietary Database built by SLAC for LSST
Quality Con-	Services and processes which are aimed at measuring and monitoring a
trol	system to verify and characterize its performance (as in LDM-522). Qual-
	ity Control systems run autonomously, only notifying people when an
	anomaly has been detected. See also Quality Assurance
SDSS	Sloan Digital Sky Survey
SQL	Structured Query Language



Science	The library of software components and the algorithms and processing	
Pipelines	pipelines assembled from them that are being developed by DM to gen-	
	erate science-ready data products from LSST images. The Pipelines may	
	be executed at scale as part of LSST Prompt or Data Release processing,	
	or pieces of them may be used in a standalone mode or executed through	
	the LSST Science Platform. The Science Pipelines are one component of	
	the LSST Software Stack	
Science Plat-	A set of integrated web applications and services deployed at the LSST Data	
form	Access Centers (DACs) through which the scientific community will access,	
	visualize, and perform next-to-the-data analysis of the LSST data products	
background	In an image, the background consists of contributions from the sky (e.g.,	
	clouds or scattered moonlight), and from the telescope and camera optics,	
	which must be distinguished from the astrophysical background. The sky	
	and instrumental backgrounds are characterized and removed by the LSST	
	processing software using a low-order spatial function whose coefficients	
	are recorded in the image metadata	
metadata	General term for data about data, e.g., attributes of astronomical objects	
	(e.g. images, sources, astroObjects, etc.) that are characteristics of the	
	objects themselves, and facilitate the organization, preservation, and query	
	of data sets. (E.g., a FITS header contains metadata)	
stack	a grouping, usually in layers (hence stack), of software packages and ser-	
	vices to achieve a common goal. Often providing a higher level set of end	
	user oriented services and tools	