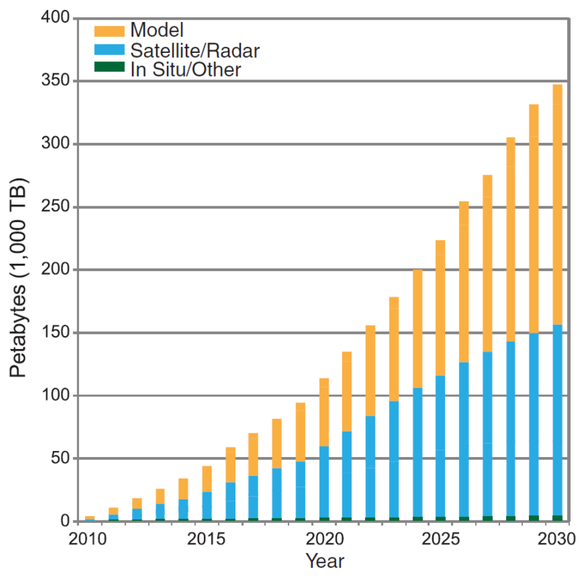
Designing Scientific Databases for High Time-Resolution Astronomical Data

Paul Zervos *– Supervisors: Dr. Tara Murphy and Prof. Alan Fekete*

School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

1. INTRODUCTION

- Science is undergoing a massive data explosion in terms of quantity and transmission speed.
- The datasets which scientists are dealing with are becoming much more complex.



The amount of total global climate data [1]
Increases from 5 PB in 2010 to 347 PB in 2030

RESEARCH QUESTIONS

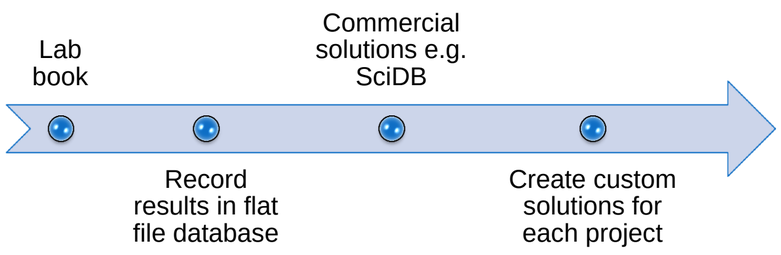
How do scientists use databases?

What are the scientists requirements for their database?

Does the evidence agree with their perceptions?

History of How Scientists Store Results

- Scientists have progressed from using lab books to store their results to custom created databases due to the quantity and specificity of the data they are dealing with.
- In 2001 "the average Wal-Mart had a better database than the average astronomer[2]", highlighting science had not kept up with commercial products.



Australian Square Kilometer Array Pathfinder (ASKAP) Survey for Variables and Slow Transients (VAST)

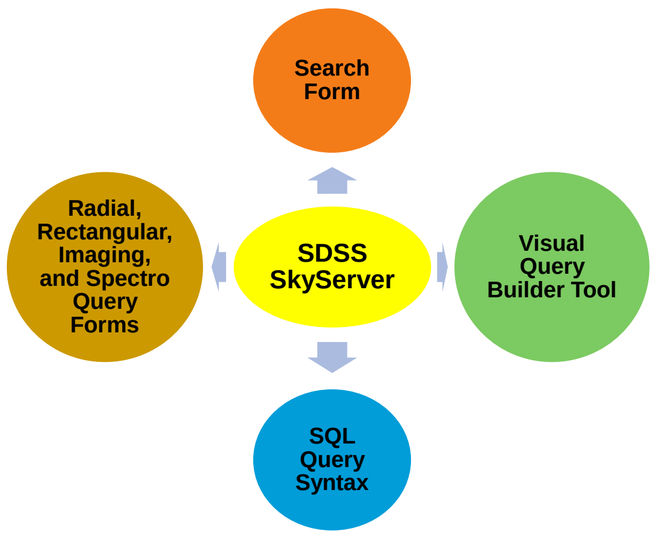
- A next generation telescope hoping to make advances in science such as learning about galaxies and stars that change rapidly.
- VAST is a project which will study radio transients and variable phenomena like supernova explosions.

CHALLENGE: Scientists require a time variable phenomena database which is capable of ingesting 2.72 GB/s of data.

2. SLOAN DIGITAL SKY SERVER

- I analysed the log files from the SDSS which is the world's first large online astronomical database, because the potential users of the VAST project find it difficult to predict how they will be using the database.
- The aim of the analysis was to ascertain how scientists actually interact with a large scale database.

Four Ways Which the Users Can Query the Database

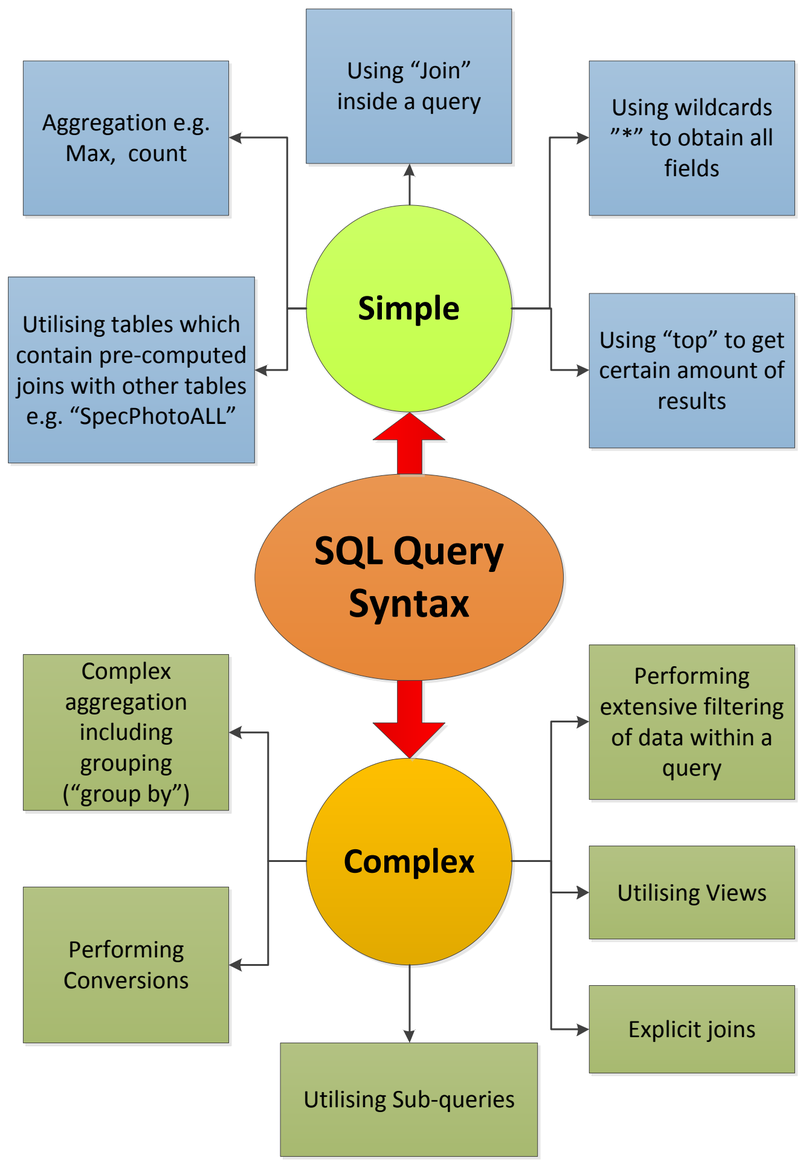


Top Three Queries Performed

Query	Number
SELECT top 1 p.objID, p.run, p.rerun, p.camcol, p.field, p.obj, p.type, p.ra, p.dec, p.u, p.g, p.r, p.i, p.z, p.Err_u, p.Err_g, p.Err_r, p.Err_i, p.Err_z, FROM fGetNearbyObjEq(195,2.5,0.5) n, PhotoPrimary p, WHERE n.objID=p.objID	179,635
select name, type from DBObjects where type ='U' and name NOT IN ('LoadEvents', 'QueryResults') order by name	9,030
select name, type from DBObjects where name like 'Constants' or name like 'Defs' order by name	2,464

Categorisation of Written Queries

- I performed an automatic analysis of 2 million log entries using Python programs, and manual analysis by hand of 5,000 queries which were written by the users.
- I categorised the queries written by the users as being simple or complex according to the following scheme I developed:



3. ASKAP VAST

- To establish the requirements for this database I surveyed 13 potential users and asked questions about their likely use of the system and what information they would like to retrieve.
- For the open-ended questions qualitative analysis was performed to extract and categorise the participants responses.
- The tables below highlight the responses of question 23 from the survey. The letter after the number 23 is a unique identifier for each different concept.
- I classified every new concept and placed them under a heading in the table with the corresponding participant listed below it which are numbered and colour coded.

Typical Queries Users Will Perform

23A	23B	23C	23D	23E	23F
Queries displaying all objects	Return sources with certain variability	Queries involving searching for light curves	Displaying "interesting" sources	Query to determine particular sources e.g. if it is a supernova	Finding all objects with some given conditions
1	1	1	11	13	1
2	6	8			8
4	12	9			10
6		10			
8		13			
9					
13					

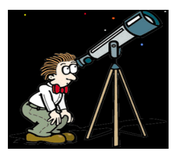
23G	23H	23I	23J	23K
Finding objects similar to a given object from another telescope	Queries that will be run e.g. every day to display new results	Joining multiple tables	Involving mathematical calculations	Do not know yet
2	3	1	9	7
12		4		
		12		

Types of ASKAP VAST Users

- From the analysis of the results I categorised the participants into three types of users which are shown below:

Astronomer / Researcher Who is a SQL / Programming Guru

Typical Astronomer / Researcher



Astronomer / Researcher Who is a Novice SQL Writer

4. DISCUSSION

- A portion of the participants from the survey state they will write and execute complex queries to search the database.
- However from the analysis of the query logs over 95% of the queries were performed by using the search forms. For the queries which were written by the users, most were considered to be simple.

5. MY CONTRIBUTIONS

- I established the requirements for the ASKAP VAST project and developed use cases to illustrate these requirements
- I also provided an analysis of *how* scientists use and interact with scientific databases.
- I have synthesised the above results and they will now form the foundation for database developers to structure and create the online database for the VAST project.

References

1. Lee, K., et al., 2005. Managing the ct data explosion: initial experiences of archiving volumetric datasets in a mini-pacs. *Journal of Digital Imaging*, 18(3):188–195.
2. Thakar, A.R., *The Sloan digital sky survey: Drinking from the fire hose.* Computing in Science & Engineering, 2008. **10**(1): p. 9-12.