A Database with Enterprise Application for Mining Astronomical Data Obtained by MOA

By

Huawei Xu

A thesis submitted in partial fulfillment of the requirements for the degree of the Master of Information Science in Computer Science

Massey University at Albany,
Auckland, New Zealand
cyxhw@hotmail.com
February 2007

Abstract

The MOA (Microlensing Observations in Astrophysics) Project is one of a new generation of modern astronomy endeavours that generates huge volumes of data. These have enormous scientific data mining potential. However, it is common for astronomers to deal with millions and even billions of records. The challenge of how to manage these large data sets is an important case for researchers. A good database management system is vital for the research. With the modern observation equipments used, MOA suffers from the growing volume of the data and a database management solution is needed. This study analyzed the modern technology for database and enterprise application. After analysing the data mining requirements of MOA, a prototype data management system based on MVC pattern was developed. Furthermore, the application supports sharing MOA findings and scientific data on the Internet. It was tested on a 7GB subset of achieved MOA data set. After testing, it was found that the application could query data in an efficient time and support data mining.

Table of Contents:

1.0 Introduction/ MOA project

On a clear night, light has travelled through space for about millions of years and then reaches the earth. It quickly disappears but its trace as photons is captured by astronomical telescope and is recorded as the digital images. Then astronomers digitize the images and start their research on the mystical universe.

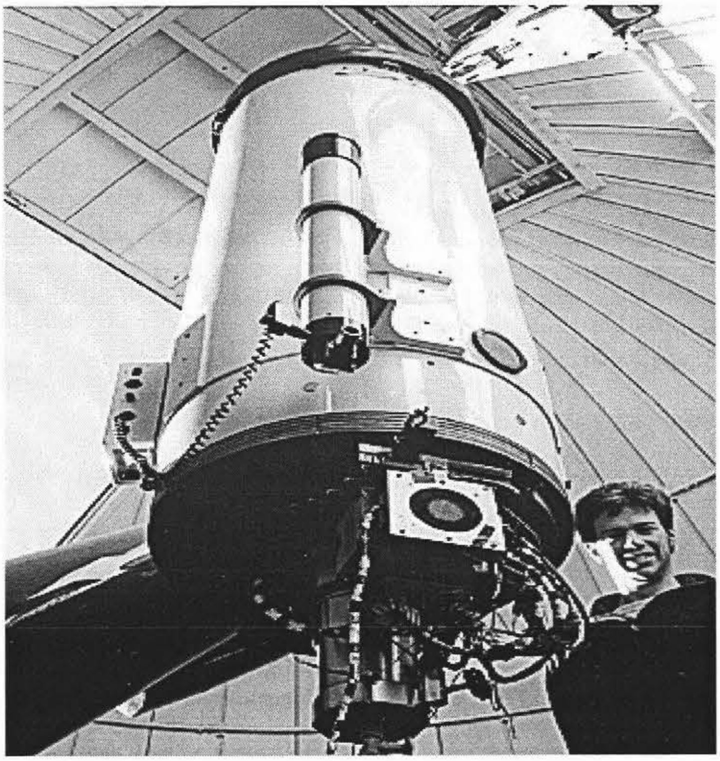
The combination of modern computer technology, wide-field CCD detectors and telescope technology have opened up new opportunities for astronomy research. As a result, astronomical research groups have used these advanced technologies in their study. For example, as the most capable Infrared imaging survey instrument, WFCAM is used by the UK Infrared Telescope on Mauna Kea (The Royal Observatory Edinburgh, 2005). This instrument is built by a cryogenic camera with four state-of-the-art detectors (four Rockwell 2048x2048 detectors), associated electronics and computing. Furthermore, its large optics, including a new f/9 secondary and a complete auto guiding system, supports to take perfect photos in a short time and its software is able to process the real-time pipeline data in time.

At the same time, in order to improve the survey quality, scientists have been built more and more professional telescopes. For instance, SkyMapper telescope is dedicated to scanning the night time sky. It is a 1.35m survey telescope with an 8-sq degree field of view. And it has an integrated 16kx16k mosaic CCD and will be on operation in 2007. SkyMapper telescope will be able to scan the night time skies quicker and deeper than before (RSAA, 2006). As a space-based telescope, the famous Hubble Space Telescope (HST) is 2.4m reflecting telescope in orbit around the Earth. HST was designed as a long term observatory and the instruments include three cameras, two spectrographs, and fine guidance sensors (Space Telescope Science Institute, 2006). Because of HST's location above the Earth's atmosphere, this science instruments can produce high resolution images of astronomical objects than the ground-based telescope. In addition, advanced ground-based telescopes also are commonly used for observation as well. Large Synoptic Survey Telescope (LSST) is a planned wide-field "survey" reflecting telescope and it will be used in 2013. The LSST has an unbelievable size-8m and a wide field view for exposure. Its 3.2 billion-pixel camera can take a 15-second exposure in every 20 seconds. This camera is expected to take over 200,000 pictures (1.28 pet bytes uncompressed) per year (LSST, 2006). In contrast, Visible and Infrared Survey Telescope (VISTA) will be available early than LSST; it will be in field in 2007. It will be a 4m class wide field survey telescope with a near infrared camera containing 67 million 0.34 arcsec pixels and available broad band filter. VISTA will produce about 315 GB on a typical night (Emerson, 2006).

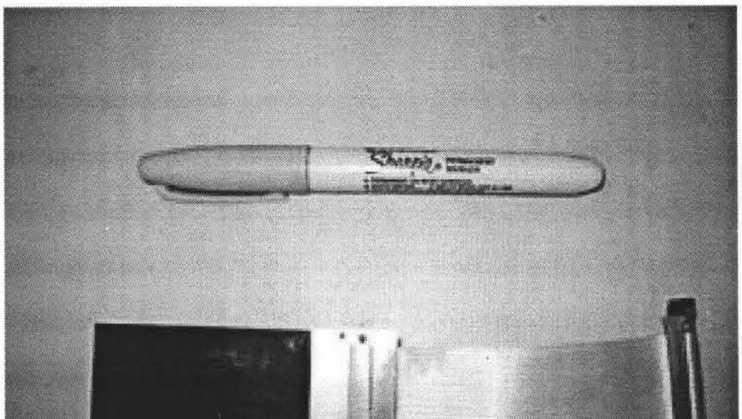
However, with the use of new modern equipment, the volume of data is large and is being accumulated at higher and higher rates. Managing and effectively data mining the enormous telescope output data is the most technical and difficult part of the projects. Therefore, astronomers need to discover a suitable database management system for them.

1.1 MOA Project and Equipments

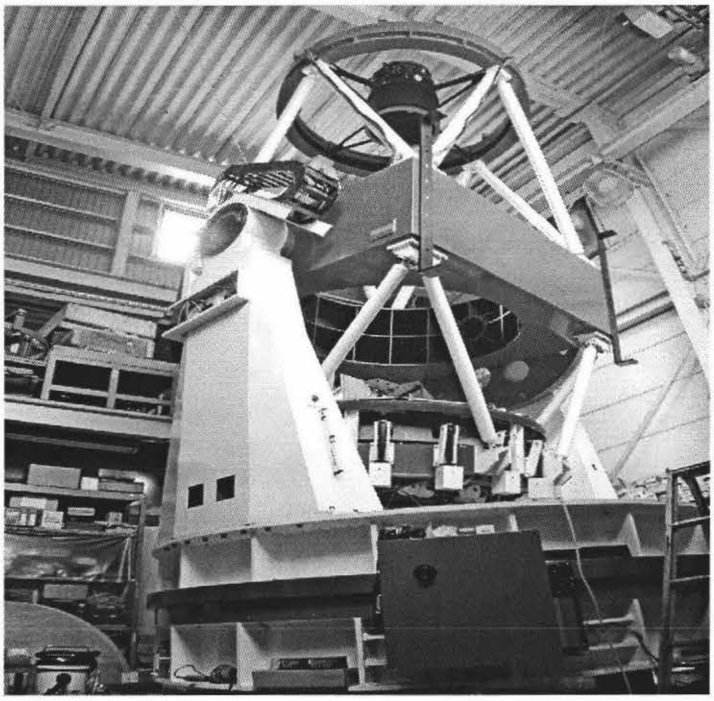
Microlensing Observations in Astrophysics (MOA) is a joint Japanese/New Zealand collaboration designed to perform large-scale astronomical photometry to detecting interesting astrophysical phenomena. MOA is one of a new generation of astronomical research projects. It started in 1995; it makes "observations and measurements on dark matter, extra-solar planets and stellar atmospheres using the gravitational microlensing technique" (MOA, 2006). Between 1998 and 2005, MOA used a 0.6-metre Boller and Chivens telescope for the observation. Picture 1 shows the 0.6m telescope with mosaic CCD camera called MOACam2 and it has 4096x6144-pixels. MOACam2 has three 2048x4069-pixel site CCD chips and the Picture 2 shows one of them. In 2002, MOA received a new 1.8-metre-diameter telescope from the Ministry of Education, Culture, Science & Technology of Japan. It is the largest telescope in New Zealand at the moment; Picture 3 shows it. During 2005, the new MOA-II telescope (Picture 4) equipped with new mosaic CCD camera called MOACam3 (Picture 5) launched was commissioned. The MOACam3 is an 80Mpixels CCD camera, which has ten CCD chips, and each chip is 2048x4096 pixels.



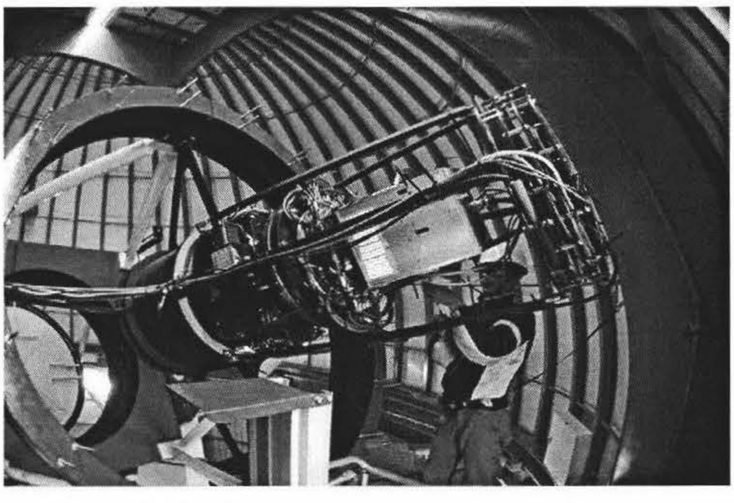
Picture 1: The 0.6 m telescope with MOACam2 attached



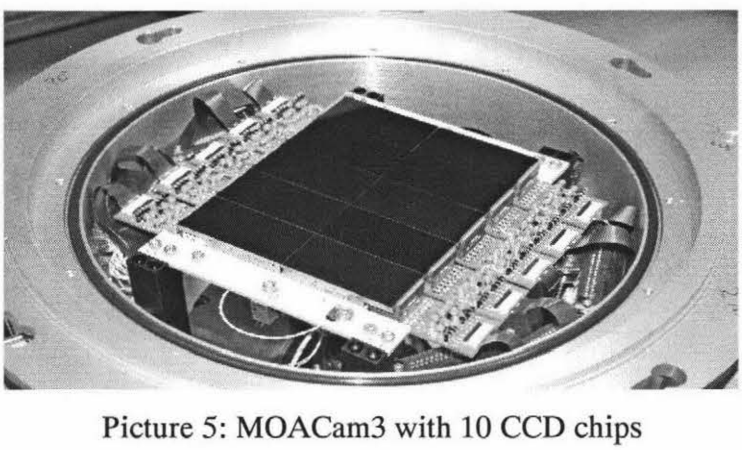
Picture 2: One of CCD chip used in MOACam2



Picture 3: The new 1.8m MOA-II telescope



Picture 4: MOACam3 attached to MOA-II telescope



Picture 5: MOACam3 with 10 CCD chips

Obtained Data

Table 1 shows the volume of data MOA obtained among different phases. From Table 1 we can see, there were about 500 GB data obtained by MOA-I telescope with MOACam1 between 1995 and 1997. In contrast, for the observations that were carried out by the new MOACam2 from 1997 to 2004, the accumulated data increased to 3TB. Since 2005 the data volume has increased dramatically. When the MOA-II telescope with new camera was launched, MOA obtains more than 3TB data each year. With the growth of the data size, how to manage these data is a challenge for MOA.

Table 1: Timeline of MOA data

Time	Telescope	CCD Camera	Number of CCD	Dimensions of each CCD	Data produced
1995-1997	0.6m	MOACam1	9	1000 x 1018	About 500GB
1997-2004	0.6m	MOACam2	3	2048 x 4096	About 3 TB
2005-now	1.8m	MOACam3	10	2048 x 4096	3 TB per year

At the moment, MOA uses a set of Python scripts to analyze the data. Roughly, they use three steps to do the research. Firstly, MOA members transform imaging data into measurement that records position, fluxes and other information of the observed objects. During the transformation, the data sizes are reduced and the useful information is retained. Secondly, MOA stores the measurement into data files so the objects can be studied. Finally, MOA members extract data knowledge from the data files based on association rules, this process is called data mining. Therefore, database system is fundamental facility for astronomy research. Matthew (2002) stated

> "Astronomers can use computers programs to find new phenomena, relationships and useful knowledge about the universe and ultimately reduce the gap between data captured and analysis".

In short, a good database management system can support researchers take advantages of the database to perform data mining.

1.2 Description of MOA data

In its present operation, the data acquired by MOA are saved in flat files. These data include three types.

- **Imaging data.** George (1997) stated "An Image is an optical representation of an object produced by light rays from the object being refracted or reflected by a lens or mirror." In short, an image is taken by a telescope for stars or galaxies to record what they look like at that moment. These imaging data are useful for a number of research areas, such as detecting asteroids and new earth objects. Hundreds of thousands of individual images (one example is shown in Figure 1) have been accumulated by MOA. Images obtained by MOA telescope are output in Flexible Image Transport System (FITS) format. FITS format files are able to show astronomical images in multiple dimensions. Normally, the FITS format data always comes with "metadata". Metadata is "data about data" and describes image information. It interprets the basic image data information, such as the number of dimensions in the image and the image size. Because imaging data is very important for the research, these should be saved in the database.

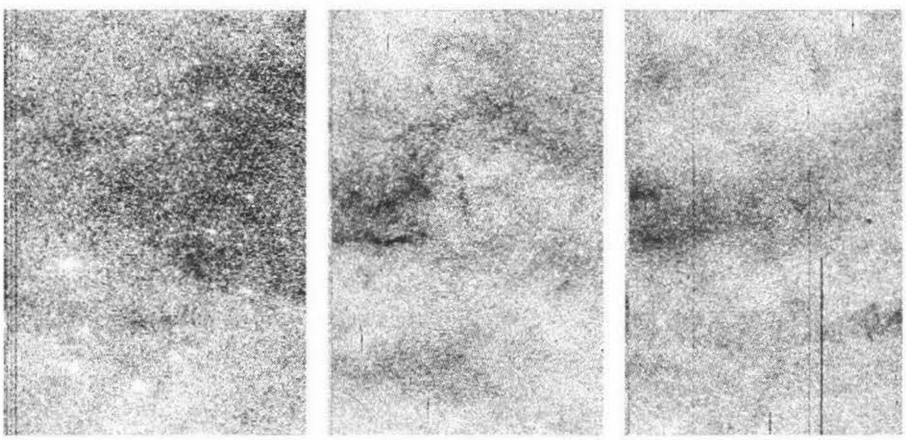


Figure 1: Three images from each CCD of MOACam2

- **Time series photometry data.** These data describe time series intensity measurement data for each star within the field. In the sky, some stars have almost constant luminosity; these kinds of stars are known as constant stars. In contrast, many stars undergo significant variations in brightness over time, and these are called as variable stars. The study of variable stars is of enormous interest to astronomers. American Association of Variable Star Observers (AAVSO) (2006) stated researching on variable stars is important because it provides information about stellar properties, such as mass, radius, luminosity, temperature, internal and external structure, composition, and evolution. Since MOA members started observation, they have been tracking about 5 millions variable stars. Through studying a long-time behavior of these stars, MOA member can analyze the variable stars behavior. In addition, the variable stars come with a wide range of types and classificatrion. For instance, there are intrinsic variable stars, wherein "variability is caused by physical changes such as pulsation and eruption"(AAVSO) and extrinsic variable star, wherein "variability is caused by eclipse of one star by another or by effects of stellar rotation"(AAVSO). Figure 3 shows an example of just three types of variability.

A light curve is a graph which shows the light intensity of a star over a period. Based on these light curves, astronomers can calculate light curve period, amplitude and magnitude. Then astronomers can use this information to schedule the observation time of the certain stars. Moreover, they can use this information to investigate astronomy science. Therefore, the database needs to record these intensity measurement data.

- **Calibration data.** These data support information on converting object positions from the images to standard astronomical coordinate systems. MOA uses astronomical coordinate system in 2000 to record each object's location. But the objects' position is shown as reference position on the captured image. Therefore, astronomers need to translate the objects position from image's reference value to coordinate system value through photometry and instrument. As a result, the database needs to store both reference position and coordinate position for each observed object in order to support their study.

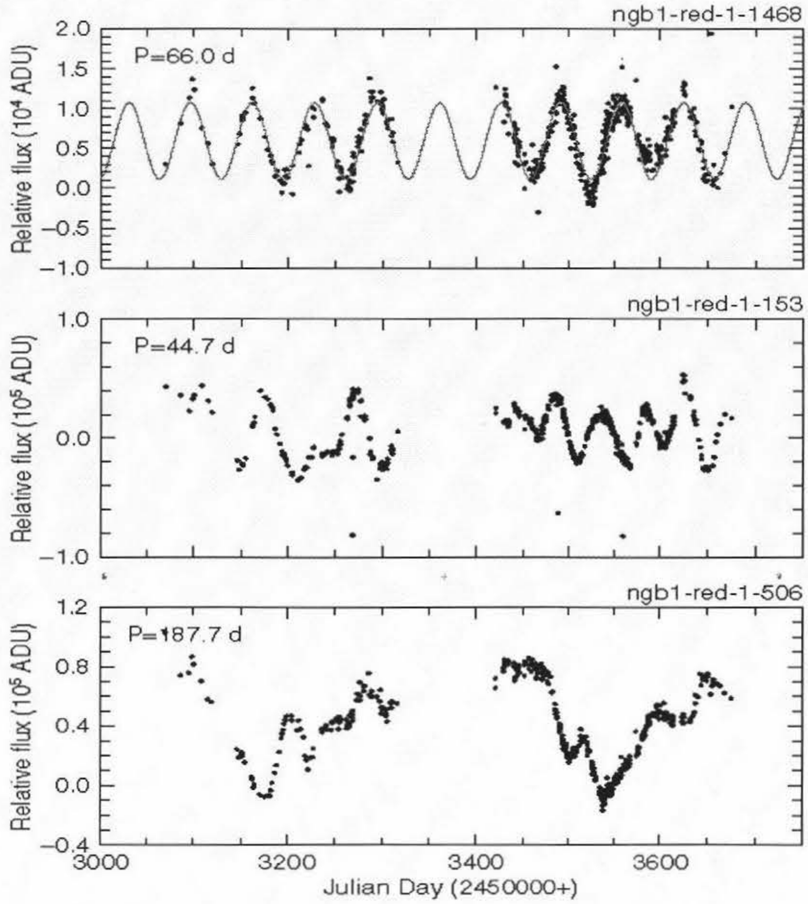


Figure 2: Light curve examples

(a) Regular pulsating variable star

(b) Semi-regular variable star

(c) Irregular variable star

1.3 Current representation of MOA data

Under the current situations, MOA has no real database management system in place although they have made some studies. They use a highly specialized collection of text files and programming scripts to manage the data. However, this is not flexible and is not extensible. Furthermore, some data saved in the flat text files are repeated and it is hard to cross-reference data among different text files. With this state of affairs, it becomes a rather cumbersome operation for any other interested parties to carry out any data mining investigations.

1.4 Overview of this project

Under current situation, MOA needs a new database with database management system to store and effectively manage their data. The database should save astronomy data, support data mining and can be extended easily in the future. In addition, in order to query data easily, MOA needs a good interface to manage the data. The interface can implement users' query on the database and also deliver the query results back to the user. Moreover, the interface can link with research tools such as GNUPLOT. For example, the interface can call the application GNUPLOT with inputting data. So the interface is able to make some of the research operations become pipeline. Then data mining operations will become easier, thereby improving the scientific productivity of the accumulated data.

This project analyzed the modern technology for database and enterprise application. Then designed a suitable application for MOA to manage their tremendous growth data. The prototype supports and benefits astronomical data mining. Due to security consideration, the main aim of enterprise application is to read astronomical data from database and to share them on the Internet. Another stand-alone Java program, it is offline from the Internet, is used for MOA to update and to modify the database (data ingestion). In addition, MOA likes to have a more powerful website to introduce their findings and discuss astronomy events with other on the Internet. It is desirable to have a functional web based interface to the data management system.