University of North Carolina Asheville
Journal of Undergraduate Research
Asheville, North Carolina
May 2020

Quasar Research Database: Architecture to Store and Access Processed SDSS DR 16 Data

Tiffany Shreves
Computer Science
University of North Carolina
Asheville, North Carolina 28804 USA

Faculty Advisors: Britt Lundgren, Lee Johnson, Marietta Cameron

Abstract

The Physics and Astronomy department at the University of North Carolina at Asheville (UNCA) is currently conducting research to glean valuable information about the cosmic web, interstellar medium, and evolution of cosmological structure using spectra from distant quasar objects observed by the Sloan Digital Sky Survey (SDSS).[1,7,9] An existing processing pipeline was developed in C++ and Python by researchers at UNCA, the University of Chicago, and Yale University to identify systems of metal-enriched gas located along the line of sight between the Earth and the observed quasars, essentially taking a 'core-sample' of the Universe.[2] In this project the team analyzed, updated, and optimized this pipeline for a more recent SDSS data release, Data Release 16 (DR 16). The second phase of the project produced a relational MySQL database to store this processed data, and a web application to serve as a user interface for researchers to access the data. The processed data from the SDSS DR 16 has been stored in a MySQL database hosted on a locally accessible Apache server. This database is serviced by a web application which provides a variety of tools to search and access pipeline output and general information for each SDSS quasar observation. The newly processed data, MySQL database, and web application user interface will facilitate research at the University of North Carolina at Asheville with the potential to expand the scope to researchers across different projects and institutions. These tools will support a variety of research topics which will yield valuable insights about the interstellar medium, formation of galaxies, and evolution of large-scale cosmological structure.

1. Introduction

The UNCA Physics and Astronomy department is conducting research to glean valuable information about the cosmic web, interstellar medium, and evolution of cosmic structures using spectra from objects called quasars, which are some of the brightest and therefore most distant observable objects from Earth.[2] Current interstellar medium research at UNCA is primarily done using data from a smaller set of quasar objects from earlier data releases of the Sloan Digital Sky Survey (SDSS); however, this survey has been continually accruing observational data, and recent data releases have included data from more than 500,000 objects, many of which have not been manually examined by researchers in detail.[1,2]

The documentation for SDSS DR 7 describes a preprocessing pipeline which is run prior to public data releases of the eBOSS survey, producing critical data about observed objects such as a quasar catalog, 1D calibrated spectra, and object redshifts, using two independent software packages, spectro1d and specBS.[6] In SDSS DR 7, the amount of manual visual inspections required for preprocessing was reduced, and these visual inspections were further cut down in later data releases.[2,4,6] The SDSS DR 7 identified 105,783 objects in the catalog, which is an increase of 28,354 quasars from the number of quasars in the fourth data release.[6] In the late 1990s, the Sloan Digital Sky Survey experimentally developed a data storage model translating a Objectivity/DB schema into a SQL Schema which required few changes and provided some significant benefits including easier and more intuitive ways to group related objects for a more scientifically useful query system, which influenced the database architecture methodologies

adopted during the design phase of this project.[6] The quasar catalogs produced during the preprocessing stages of SDSS DR 14 and DR 16 contained information about the quasar objects identified in the data set, which was critical for the data acquisition and data processing phases of the project.[1,4,6]

This project utilized and improved an existing processing pipeline written in C++ and Python that was written by researchers and faculty at Yale University, the University of Chicago, and the University of North Carolina at Asheville over the past decade. This pipeline takes a set of fits files containing quasar spectra from the Sloan Digital Sky Survey and identifies the shape of the quasar continuum emission. The pipeline divides this line from the initial spectrum, resulting in a spectrum of small bumps and wiggles. The pipeline then identifies characteristic absorption lines in this normalized spectrum that often indicate bodies of metal-enriched gas. It determines the redshift of these lines from laboratory conditions and looks for other indicative absorption lines at the same redshift to evaluate a confidence grading for the presence of bodies of metal-enriched gas, often from intervening galaxies, which are located along the line of sight between the Earth and each quasar.[2] The pipeline then marks these systems of absorption lines, determines their confidence, and plots the original spectrum and the normalized spectrum with overlaid information about identified systems. The pipeline outputs these two plots as well as several data files with information about the systems. The results can then be used to create 3-dimensional maps of the gas and galaxies in the cosmic web and to study how the composition of the universe has evolved over billions of years.[2,8]

The large scale of the most recent data release exacerbates existing issues with the pipeline and requires optimization for processing data from SDSS DR 16. Since future data releases will likely continue to grow in scale, it was necessary to update and develop a codebase that will continue to sustainably support expansion with larger data releases over the next few years. Early stages of this project determined that the speed of the current pipeline should be improved to make processing feasible for the scale of SDSS DR 16 and future data releases.[8] There are also some issues with unpredictable pipeline faults and crashes at runtime that become a significant obstacle to processing at this scale, and which needed to be investigated and addressed.[8] Early stages of the project aimed to continue optimization efforts by further developing a supervised machine learning algorithm that refines the output and removes erroneous line identifications to improve the quality of pipeline output and decrease the responsibility of scientific validation placed on researchers,[2] although this goal was extended into a future research goal due to complications in project timing and resource allocation with the transition to remote coursework in the last few months of the Spring 2020 semester.

This project optimized the processing pipeline to run on the largest ever collection of quasar spectra from the SDSS DR 16.[1] It then designed a database to house this processed data, and implemented a web application to allow researchers to query this database and access the processed data for astronomy research.

2. Methodology

This project was divided into three primary phases: pipeline analysis and improvement, database architecture, and user interface design. All three phases of the project focused on creating a product that could be easily used for years after the conclusion of project development, and which will be easily extensible for improvements and expansion with future, larger-scale SDSS data releases. The methodology incorporated elements of the Agile development methodology and utilized weekly stakeholder meetings as an opportunity to update project specifications and methods and as a motivator to continually complete usable minimum features and add value to the product at each stage of development. Weekly meetings with the primary stakeholder in the Physics and Astronomy department helped to track the progress of the project and conduct scientific validation and confirm that the product met stakeholder requirements at every interval of minimum viable functionality.

Prior research on viable and affordable alternatives to proprietary architectures for storing SDSS and other large scale astronomy survey data suggests that relational SQL database designs are well-suited to scale with the growing needs of massive astronomy data projects.[3,7] This project utilized a relational MySQL database with a database model based on the most useful information stored in the existing processed DR 7 database. This DR 7 database is accessible by a password-protected website which served as the base model for the updated DR 16 web-based user interface. Iterations of the new user interface were developed to provide at minimum the functionality of the extant DR 7 data access website.

This project included an opportunity to travel to Princeton University to consult with one of the primary architects of the DR 7 database and data access website, Dr. Yusra AlSayyad. This opportunity to meet with Dr. AlSayyad provided valuable insight into the changing needs of the project since the creation of the DR 7 architecture, ways to approach persistent faults in the legacy code of the processing pipeline, and features that would be valuable in future iterations of the user interface to improve the user experience for researchers.

3. Application

3.1 Application of Processing Pipeline Improvements

The first step to update the processing pipeline was to adapt it to run with an input of SDSS DR 16 data. Several fields changed in the SDSS data model between the seventh and sixteenth data releases, and these fields had to be updated so that the processing pipeline would recognize the data in the newer files. A legacy Segmentation Fault in the underlying C++ code of the processing pipeline also posed a significant obstacle to large-scale processing and would unpredictably halt processing batches after between a few hundred and a few thousand files, requiring constant supervision to completely process larger batches of files. By supplementing garbage collection in the wrapping python code, this Segmentation Fault issue was eliminated from single-thread processing. The fault reappeared during parallelization; however, it occurred significantly less frequently, allowing processing to occur on much larger scales without constant supervision. Ongoing research on possible origination points for the Segmentation Fault has produced several potential comprehensive solutions that are being considered for future implementation to further improve the processing pipeline and to continue to reduce the time and resources needed for researcher supervision during later SDSS data releases.

3.2 Application of Database Architecture

The database architecture was designed to mirror the most useful components of the existing processed DR 7 database. The database was implemented as a relational MySQL database with three linked quasar tables and one table dedicated to user login information. The ‘quasar’ table consists of top-level quasar information drawn from an output file from the processing pipeline and from the quasar catalog produced by Pâris et al.[4] Each ‘quasar’ can have at most one ‘absorption_list’. This ‘absorption_list’ table contains information about an identified system of metal-enriched gas, and can have one or more ‘absorption_line’. Each ‘absorption_line’ has specific information about the absorption line identified by the processing pipeline, along with the confidence rating that the pipeline and machine learning model component have assigned the line. This information about the ‘absorption_list’ and its ‘absorption_line’s have been drawn directly from the output of the processing pipeline.

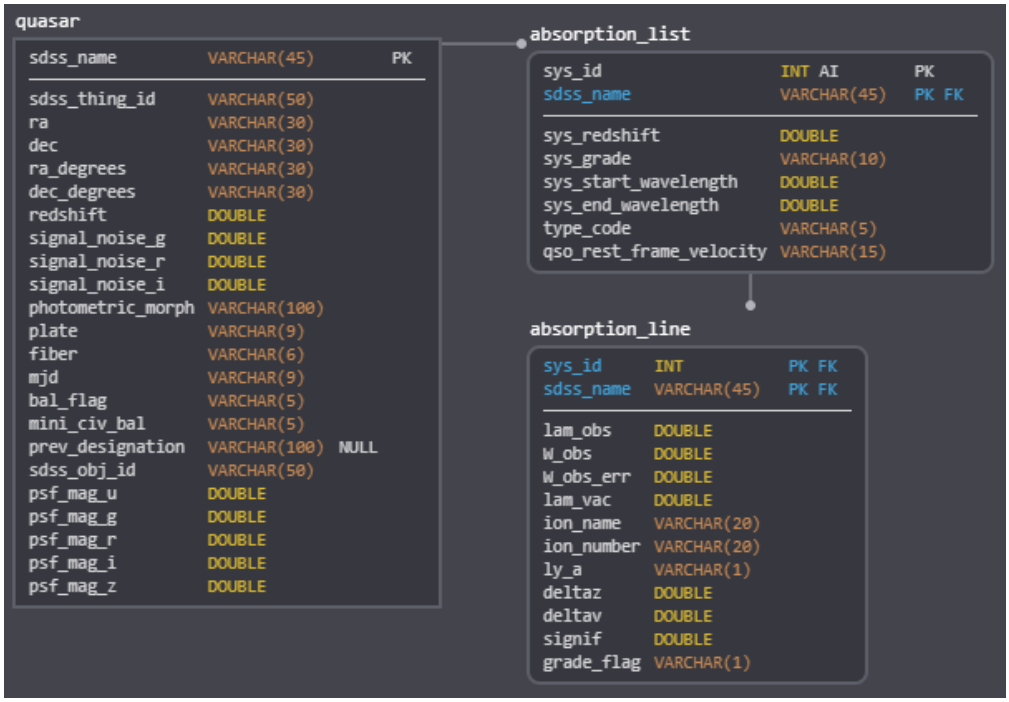

Figure 1: Relational database model diagram for processed SDSS DR 16 quasar data; implemented as a MySQL database

3.3 Application of User Interface Design

The user interface was developed by iterating on functionality present in the extant SDSS DR 7 website which currently provides access to earlier processed quasar data from earlier data releases. This password protected website allows processing pipeline developers and authorized researchers to access processed data from SDSS DR 8 to DR 16. Users can search through observations by plate and fiber number and get information about each observation such as an observed object's redshift, right ascension, declination, and SDSS identifiers. The profile page for a given object also includes a list of the observed systems of metal-enriched gas and the confidence levels for each system. The profile allows users to view and save high quality plots of the original spectrum for the object as well as the normalized spectrum with color-coded absorption lines for each system. The user also has the ability to open and save four files containing the initial data from the object and some other intermediate and formatted data output by the processing pipeline. Three buttons on the profile page allow the user to navigate to outside resources such as the NASA/IPAC Extragalactic Database (NED) and the SDSS Sky Server containing additional information about the observed object and the galaxies within a designated radius surrounding it. An additional feature allows users to search for an object and surrounding objects within a specified radius by right ascension and declination coordinates, providing an alternative to browsing by plate and fiber.

4. Data

Large scale processing of SDSS quasar data provided a useful tool to identify some anomalous observations in the most recent data release. Two files raised similar errors during processing which suggested that the objects may have been misidentified as quasars during the SDSS preprocessing stages and these objects have been flagged for future investigation. These errors highlighted several unhandled edge cases in the pipeline code that have been updated and supplemented with additional error handling and documentation to make similar objects easier to identify in the future. A series of files produced another error across a range of plates from plate 7082 to plate 7516 that revealed a potential issue with fibers 526, 527, and 528. These files have likewise been flagged for future investigation, and illustrate how the processing pipeline can provide unique opportunities to identify and explore undetected issues from observational or pre-processing stages.

5. Conclusion

The primary objective of this project has been to produce a database of processed quasar data that Physics and Astronomy researchers can use to investigate fascinating questions about the evolution of cosmological structure and the composition of the Universe. All development efforts focused on implementing data management architecture and supplementing the processing pipeline codebase so that both will continue to be useful and easy to adapt to meet the changing needs of researchers for years after the conclusion of this project. These new research tools have been designed with extensibility in mind to hopefully accommodate future SDSS data releases of increasing scale over the next few years.

6. Acknowledgements

The author would like to sincerely thank Dr. Britt Lundgren for her constant support, knowledge, and encouragement. Special thanks also goes out to Professor Lee Johnson for his direction and guidance, and to Dr. Marietta Cameron for her supervision, advice, and assistance throughout this project. Also special thanks to the UNC Asheville Undergraduate Research Program for their support and funding for research and travels, and to Dr. Yusra AlSayyad and the other architects of the processing pipeline and DR 7 website that made this project possible.

7. References

1. Abolfathi, Bela et al., “The Fourteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the extended Baryon Oscillation Spectroscopic Survey and from the second phase of the Apache Point Observatory Galactic Evolution Experiment”. 2018, ApJS, 235, 42. https://iopscience.iop.org/article/10.3847/1538-4365/aa9e8a/pdf
2. Blanton, Michael, et al., “Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe”. 2017, The Astronomical Journal, Volume 154, https://iopscience.iop.org/article/10.3847/1538-3881/aa7567/pdf
3. Lamb, Evelyn. “How to Share the Data from LSST.” *Symmetry Magazine*, Fermilab/SLAC, 7 Nov. 2019, www.symmetrymagazine.org/article/how-to-share-the-data-from-lsst.
4. Pâris, Isabelle et al. “The Sloan Digital Sky Survey Quasar Catalog: Fourteenth Data Release” Astronomy and Astrophysics, Volume 613 A51 (2018). https://www.aanda.org/articles/aa/abs/2018/05/aa32445-17/aa32445-17.html
5. Ross, Nicholas, et al., “THE SDSS-III BARYON OSCILLATION SPECTROSCOPIC SURVEY: QUASAR TARGET SELECTION FOR DATA RELEASE NINE”. 2012, The Astrophysical Journal Supplement Series, Volume 199. https://iopscience.iop.org/article/10.1088/0067-0049/199/1/3/pdf
6. Schneider, Donald et al., “The Sloan Digital Sky Survey Quasar Catalog. V. Seventh Data Release”. The Astronomical Journal, Volume 139, Issue 6, article id. 2360 (2010). https://arxiv.org/pdf/1004.1167.pdf
7. Smee, Stephene et al., “The Multi-object, Fiber-fed Spectrographs for the Sloan Digital Sky Survey and the Baryon Oscillation Spectroscopic Survey” The Astronomical Journal, Volume 146, Issue 2, article id. 32, 40 pp. (2013). https://arxiv.org/abs/1208.2233
8. Szalay, Alexander et al., “Designing and Mining Multi-Terabyte Astronomy Archives: The Sloan Digital Sky Survey”. SIGMOD '00 Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. p. 451-462. https://jimgray.azurewebsites.net/papers/MS_TR_99_30_Sloan_Digital_Sky_Survey.pdf
9. York, Donald et al., “The Sloan Digital Sky Survey: Technical Summary” The Astronomical Journal, Volume 120, Issue 3, pp. 1579-1587. (2000). https://arxiv.org/abs/astro-ph/0006396