Sloan Digital Sky Survey Analysis and Data Engineering

Joeri Hermans
joeri.hermans@doct.ulg.ac.be

November 21, 2017

1 Introduction

The Sloan Digital Sky Survey[1] or SDSS, is a major multi-spectral imaging and spectroscopic redshift survey using a dedicated 2.5 meter wide-angle optical telescope at Apache Point Observatory in New Mexico, United States. The Sloan Digital Sky Survey has created the most detailed three-dimensional maps of the Universe ever made, with deep multi-color images of one third of the sky, and spectra for more than three million astronomical objects. The SDSS began regular survey operations in 2000, after a decade of design and construction. It has progressed through several phases, SDSS-I (2000-2005), SDSS-II (2005-2008), SDSS-III (2008-2014), and SDSS-IV (2014-). Each of these phases has involved multiple surveys with interlocking science goals. The three surveys that comprise SDSS-IV are eBOSS, APOGEE-2, and MaNGA.

In this project, you will be working with data from the eBOSS (Extended Baryon Oscillation Spectroscopic Survey) experiment. eBOSS precisely measures the expansion history of the Universe throughout eighty percent of cosmic history, back to when the Universe was less than three billion years old, and improve constraints on the nature of dark energy. “Dark energy” refers to the observed phenomenon that the expansion of the Universe is currently accelerating, which is one of the most mysterious experimental results in modern physics.

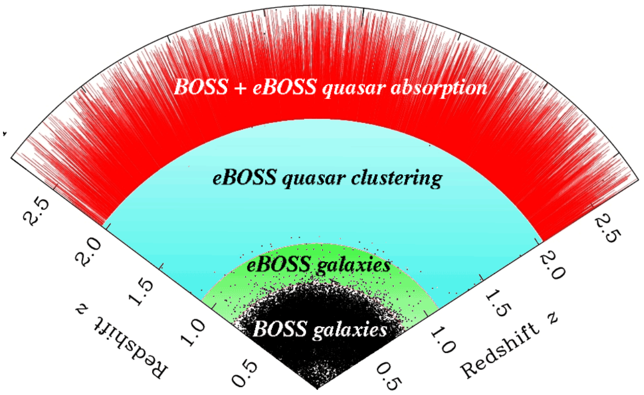

Figure 1: Converage of the eBOSS experiment up to redshift $(z) = 3$.

In physical terms, redshift, described in Equation 1, happens when light or other electromagnetic radiation from an object is increased in wavelength, or shifted to the red end of the spectrum. In general, whether or not the radiation is within the visible spectrum, "redder" means an increase in wavelength – equivalent to a lower frequency and a lower photon energy, in accordance with, respectively, the wave and quantum theories of light. Some redshifts are an example of the Doppler effect, familiar in the change of apparent pitches of sirens and frequency of the sound waves emitted by speeding vehicles. A redshift occurs whenever a light source moves away from an observer.

$$z = \left(\frac{\lambda_{\text{observed}}}{\lambda_{\text{rest}}}\right) - 1 \tag{1}$$

In astronomy, redshift can be utilized to measure the *accelerating* expansion of the universe. This is exactly one of the key questions posed by the eBOSS survey. In princple, eBOSS measures this by identifying the wavelengths of emission and absorption lines, and then comparing with the known spectra in a vacuum for those elements and thereby obtaining an average redshift for all spectra using Equation 1.

[1] https://www.sdss.org

Furthermore, these emmision lines and their accompanying fluxes (number of photons that caused a certain amount of electrons to move) can be utilized whether the observed instance is a *star* (subclasses of the spectral types can also be identified[2]), *galaxy* or a *quasar*, as shown in Figure 2 and Figure 3.

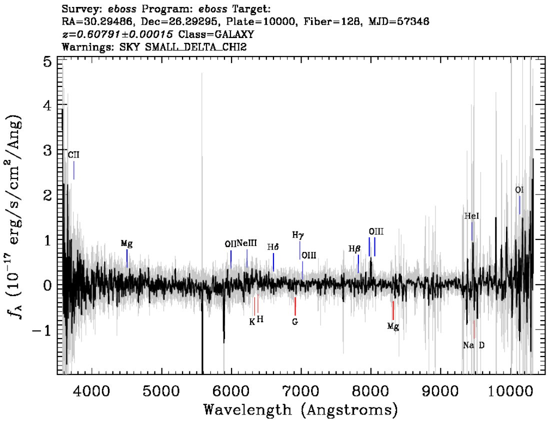

Figure 2: Spectra with accompanying fluxes of a galaxy with identified absorption lines.

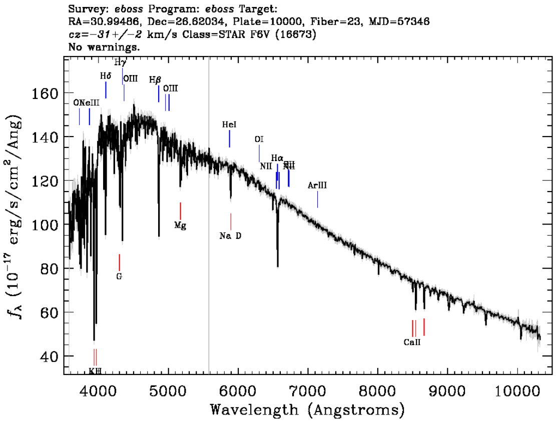

Figure 3: Spectra with accompanying of a star with identified absorption lines.

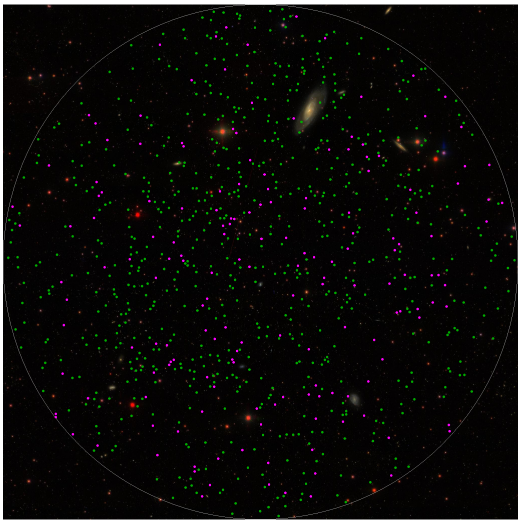

Figure 4: SDSS plate and target selection (green for galaxies and purple for quasars).

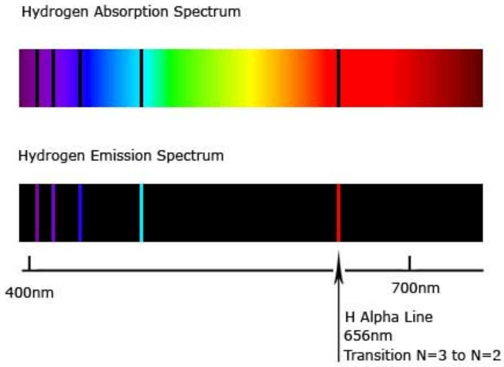

Figure 5: Balmer Series of the Hydrogen. Shows absorption and emission lines from orbital 2 to 3 and higher.

The data collection procedure is done by drilling several holes in an aluminimum plate which are positioned in a particular way to obtain individual fluxes of target objects, as shown in Figure 4. This approach allows the eBOSS survey to observe multiple targets at once by attaching an optical *fiber* to every drilled hole, causing the individual observations to be independent of each other. Furthermore, this approach allows SDSS to filter out nearby galaxies which are known to be *blueshifted* (negative redshift, i.e., objects which are blueshifted have a component which move towards us), and obtain the individual background flux for every observed object.

[2] en.wikipedia.org/wiki/Stellar_classification#Spectral_types

2 Data Products

The data of the eBOSS survey that I will provide you with is only a partial subset of the complete eBOSS survey data. For every plate and MJD (Modified Julian Data), SDSS provides several processed data products:

- A collection of *spZall* data files.
- A collection of *spZbest* data files.
- A collection of *spZline* data files.
- A single *platelist* file describes all plates and their data quality.

The most relevant data product to this project, *spZall*, contains the summary statistics of all fibers on a single plate, including the class and subclass of the object that is being observed. The remaining data files, *spZbest* and *spZline*, hold the data on the best fits of the observed data, and data on the fluxes of the received wavelengths respectively. All data in stored in the FITS format, which is a common format in astronomy. Intuitively, the format can be described as a collection of tables, and images. For instance, you could produce a FITS file which stores an image of the object you are observing, and also a table which describes every wavelength of the light that has been received.

3 Deliverables & Logistics

To complete this project succesfully, I would like you to solve and provide proof for the following questions. These questions have to be solved with Apache Spark. The proof for these questions have to be provided in a *reproducable* Jupyter Notebook with a full description what you are doing (using Markdown cells). This will suffice as your report. The project has to executed in groups with a maximum number of two persons per group. Groups need to be registered on the submission site of Montefiore. The deadline of this project is 22 December 2017 at midnight.

1. *Looking at galaxies, is the expansion of the Universe uniform across all regions of the sky? Meaning, is the redshift of the galaxies about equal across the sky?*
2. *Is the expansion of the Universe accelerating?*
3. *What is the average velocity of the galaxies which are redshifted?*
4. *What is the average velocity of the quasars which are redshifted?*
5. *Are there galaxies with a relatively small flux which are blueshifted?.*
6. *What is the distribution of the spectral type of all observed stars?*

To solve these questions, you will be reading and parsing the data products described in Section 2. However, in order to do this in an efficient manner, you most likely have to come up with an intermediate representation constructed from the original data products to perform efficient IO, i.e., it does not really make sense that you have to parse the original data structure every time when you wish to extract some valuable information. Before obtaining such a representation, you also have to figure out a way to efficiently parse the original data products (in FITS). To summarize, you need to provide *code to transform the original data into some other representation* in a way that it efficiently can be read into an Spark RDD or a Spark DataFrame.

Finally, to improve the SDSS query engine, I would like you to come up with a method which is able to efficiently generate plots such as in Figure 2, and Figure 3, for a specific region of the sky, i.e., given a specific range of *right ascension* and *declination.* This method will be utilized to evaluate the efficiency of your data representation (so optimize your schema for these type of queries).

4 Bonus

A bonus point (1/20) is awarded to the group which is able to provide the fastest average random query time over the complete eBOSS dataset (171 GB). The query that will be ran to evaluate the performance of your data architecture, will be concerned with selecting objects in a particular section in the sky. Meaning given a minimum RADEC (Right Ascenssion and Declination), and a maximum RADEC, return the list of objects which are within this coordinate system. Your result might have practical implications as well, since the current SDSS query tool is *very* slow.

5 Cluster Access

For this project, you will be able to deploy your code on a multi-node Spark cluster. The connection details, and examples of how to submit a job to the cluster, will be released in the following days in this document. An e-mail will be sent the cluster configuration has been finalized.

6 Assistance

After every lecture, the assistent will be present to help you with questions surrounding this project.

7 Data Access

To prevent you from downloading all the required SDSS data, a stick will be available at the start and end of every lecture session. This USB stick contains only a specific subset of the complete eBOSS survey dataset. However, it can be used to provide proof for the questions posed above.