CS543: Massive Data Storage and Retrieval

Group 1

Name:	Kanya Kreprasertkul	Harshini Bonam	Yifan Liao
NetID:	kk1003	sdb202	yl1463

Milky Way Bulge PSF Photometry Visualization

Abstract

We established a multilevel interactive map to demonstrate over 613 million stars in the near-IR region of the milky-way bulge area. On our map selected information is shown which is regarded to its spatial position, brightness, and spectrum. We are able to process these huge amounts of data in a relatively short time using distributed computation in Spark Streaming ecosystem and graph waves theory. Graph waves theory is an algorithm that separates graphs into subgraphs efficiently introduced by James Abello and Daniel Nakhimovich.

Introduction

Our primary goal is to establish an interactive interface to demonstrate our information of stars. To be more specific, for each tile, we show the particular information about the star, such as the total number of stars, minimum brightness, maximum brightness and average brightness of the tile. Every star is considered as connected to the neighboring stars if the astronomical distance between the two stars is less than a certain threshold *0.0001 radians*. With the help of this map, astronomers are able to grow up with a better idea about which planet may have certain metal according to the spectrum and brightness shown in. In addition, graph waves theory was applied to our data, which is an algorithm that separates graphs into subgraphs efficiently.

Dataset Description

Our dataset is based on a catalog released by the European Southern Observatory (ESO) with a total size of about **100 GB**. It contains 20 columns and over 600 million records. Each record represents a star respectively, that results from performing the precise PSF-fitting photometry on its J, K band images. Those **613,723,417 stars** are divided into **196 tiles** which cover the Galactic bulge by their location in space. In this project, 7 columns will be mainly focused on, i.e. source ID, 4 columns of its spatial position and 2 columns of its brightness.

CS543: Massive Data Storage and Retrieval

Group 1

Name:	Kanya Kreprasertkul	Harshini Bonam	Yifan Liao
NetID:	kk1003	sdb202	yl1463

Brief Implementation Details:

1. Downloaded the FITS dataset using the download script provided by ESO facility.
2. Read the FITS in Spark Streams in batches of 10000, and performed EDA.
3. Perform K-means clustering on these batches and get the cluster centers down to ~60Million.
4. Perform Graph city algorithm on these 60Million cluster centers where each cluster center is a graph node connected to another node if the distance between them is < 0.0001 radians in angular distance.

$$\text{Distance} = \sqrt{\left[(\alpha_1 - \alpha_2)\left(\cos\left(\frac{\delta_1 + \delta_2}{2}\right)\right)\right]^2 + (\delta_1 - \delta_2)^2}$$

References:

- ESO milky way bulge PSF photometry dataset https://www.eso.org/qi/catalog/show/272
- Dataset Description https://www.eso.org/rm/api/v1/public/releaseDescriptions/140
- Graph waves algorithm http://ceur-ws.org/Vol-2578/BigVis7.pdf
- ESO VVV exprtiment https://arxiv.org/pdf/1902.01695.pdf
- The astropy library https://www.astropy.org/
- VISTA Variables in the Via Lactea (VVV): The public ESO near-IR variability survey of the Milky Way https://arxiv.org/pdf/0912.1056.pdf
- The VVV Phase 3 Data release VIRCAM https://www.eso.org/sci/publications/announcements/sciann17222.html, https://www.eso.org/sci/observing/phase3/data_releases/vvv_dr3.pdf
- Distance Between Two Stars http://spiff.rit.edu/classes/phys373/lectures/astrom/astrom.html