Fast distributed cross-matching of big astronomical data and parameter estimation of a moving point source model

Petar Zecevic, mag.ing.rac
(petar.zecevic@fer.hr)
mentors: prof. dr. sc. Sven Lončarić, prof. dr. sc. Mario Jurić
University of Zagreb Faculty of Electrical Engineering and Computing, University of Washington

1. Introduction

The amount of astronomical data available for analysis is growing at an ever increasing rate. These datasets are delivered through astronomical data archives and are typically analyzed by extracting a subset of the data and downloading it on the researcher's machine.
The future will pose a challenge for the continuation of this model due to the increase in data volumes (LSST project: 1000s of observations of ~20 billion objects; ~10 PB) and changes in the nature of scientific investigations that are becoming increasingly important:

- Exploratory data analysis examining whole datasets
- Time-series analyses
- Positional cross-matching of measurements from multiple catalogs

2. Problem Description

The ideal astronomical data analysis tool:

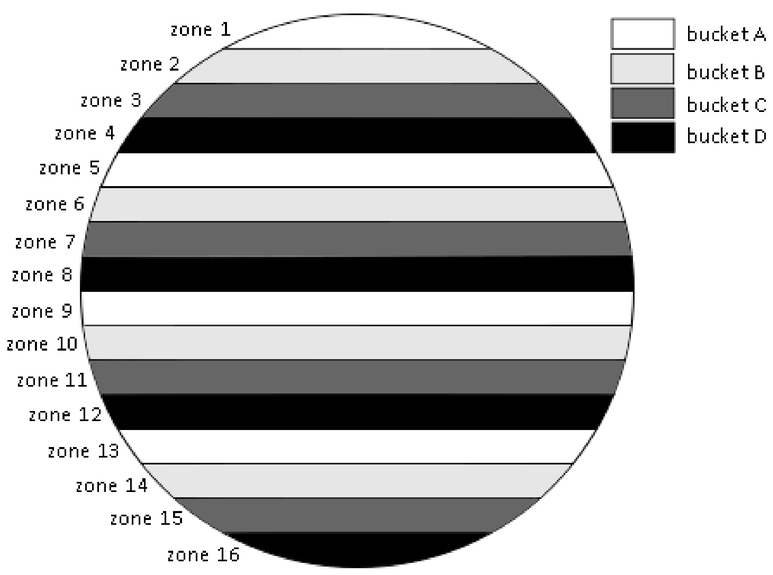
- Implements efficient cross-matching
- Based on industry standards
- Provides simple but powerful astronomical API extensions
- Easy to use on premises or in the cloud

3. Methodology

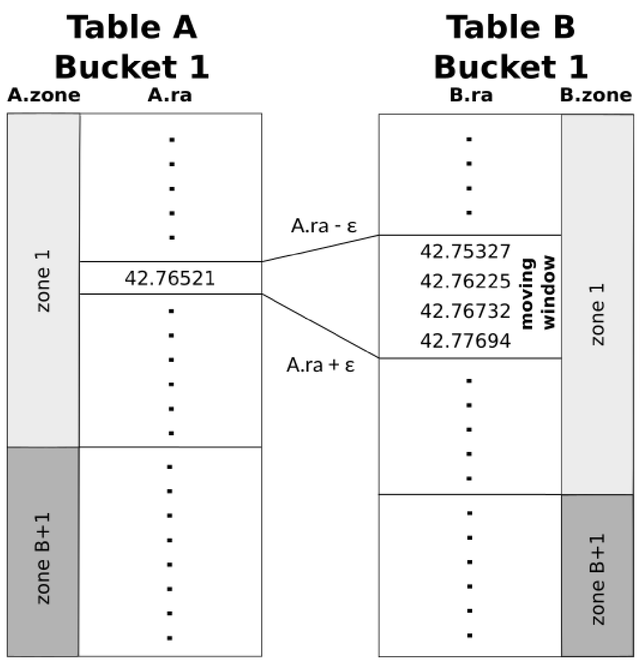
Positional cross-matching of catalogs L and R:

$$\{(l, r) \mid (l, r) \in L \times R, \mathrm{dist}(l, r) \leqslant \varepsilon\}$$

Our **Distributed Zones Algorithm** consists of a data (sky) partitioning (bucketing) scheme based on zones:



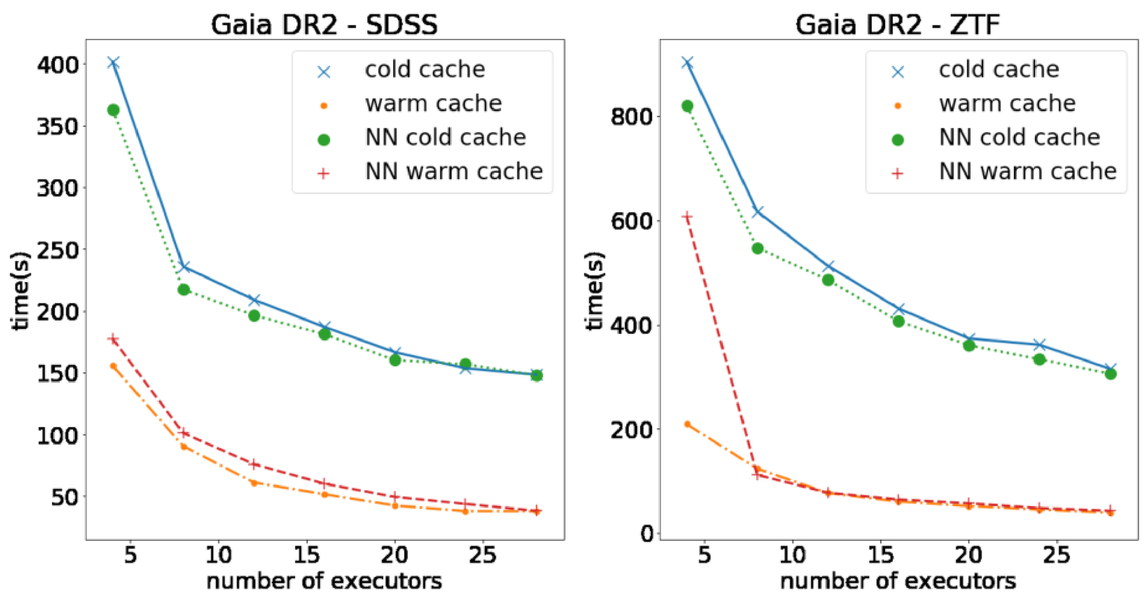
And an "epsilon join"-like optimization of range queries:




```
SELECT * FROM GAIA g JOIN SDSS s
ON g.zone = s.zone
AND g.ra BETWEEN s.ra - e AND s.ra + e
AND distance(g.ra, g.dec, s.ra, s.dec) <=e
```

4. Results

Performance tests of cross-matching Gaia DR2 (1.7 bn rows), SDSS (0.8 bn rows) and ZTF (3.1 bn rows) catalogs in scenarios with file system buffers empty or full (we used Linux OS-level caching, not Spark caching), when returning all matches or just the first nearest neighbor ("NN" results):



Data Partitioning: Size of the partitioned catalogs (in GB) and time needed to partition the data (in minutes) depending on the number of zones used

Catalog	5400 z.		10800 z.		21600 z.	
	Size	Time	Size	Time	Size	Time
SDSS	66	12	71	12	82	12
Gaia	430	89	464	86	532	150
Allwise	352	120	384	119	444	133
ZTF	1124	547	1169	545	1334	523

Cross-matching duration (in seconds) depending on the number of zones and whether cold or warm OS cache was used for each catalog combination

Catalogs	5400 z.		10800 z.		21600 z.	
	Warm	Cold	Warm	Cold	Warm	Cold
G—A	32	207	31	226	36	240
G—S	33	128	37	148	36	151
Z—A	47	260	38	296	39	283
Z—S	48	209	47	239	49	227
G—Z	37	271	39	315	44	326
A—S	27	114	29	122	29	130

5. Conclusion

The research project resulted in the system called AXS (Astronomy eXtensions for Spark). AXS enables scalable and efficient querying, cross-matching, and analysis of astronomical data sets. It is built on top of Apache Spark, with minimal extensions to the core.

Note: the second part of the theme (parameter estimation of a moving point source model) is still a work in progress.

6. Project Acknowledgement

The project was supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS).

DATACROSS