thesis_text.2

by Alexandar Mechev

General metrics

8,050	1,157	65	4 min 37 sec	8 min 54 sec
characters	words	sentences	reading time	speaking time

Score

99

This text scores better than 99% of all texts checked by Grammarly

Writing Issues

2 Issues left

Critical

2 Advanced

Plagiarism

This text seems 100% original. Grammarly found no matching text on the Internet or in ProQuest's databases.

Writing Issues

2	**Clarity**	
1	Wordy sentences	
1	Passive voice misuse	

Unique Words

Measures vocabulary diversity by calculating the percentage of words used only once in your document

37%

unique words

Rare Words

Measures depth of vocabulary by identifying words that are not among the 5,000 most common English words.

40%

rare words

Word Length

Measures average word length

5.7

characters per word

Sentence Length

Measures average sentence length

17.8

words per sentence

thesis_text.2

Conclusion

As astronomical observatories collect ever-growing data-sets, the processing challenges for these data will continue increasing. Large scale surveys expected to produce petabytes of data can no longer be processed on a single machine or small dedicated clusters at scientific institutions. Large scale distributed processing is needed to serve the scientific requirements of these survey projects.

CERN's World-Wide computing grid provides sufficient resources for such projects; however, its focus is on distributed Monte-Carlo simulations. The design choices made to tackle these computations also present some design constraints for radio astronomical processing. Namely, porting complex workflows to a grid-like environment requires a framework to distribute and monitor jobs. Additionally, a workflow orchestration software is needed to schedule and automate processing.

Summary of Thesis Achievements

This work focuses on the software built to accelerate, parallelize, and automate LOFAR processing, as well as the insights obtained into large scale processing of LOFAR data. To date, we have helped process an unprecedented eight

petabytes of data for the LOFAR Two-Meter Sky Survey (LoTSS), data which has led to more than 30 publications. We describe a generic platform for scaling astronomical processing across multiple clusters, focused on the application of bulk LOFAR processing.

We have built software that can encapsulate LOFAR processing steps and distribute them across a heterogeneous infrastructure. Our tools have been used by several scientists, implementing multiple complex pipelines, processing a total of 8 petabytes of data.

We implemented a modern monitoring suite along our processing to track the performance of individual pipeline steps.

Answers to Research Questions

[4em]8emResearch Question 1: How can we use a distributed shared infrastructure for efficient LOFAR data processing?

In Chapters 2 and 3, we detail our success with massively distributed processing of LOFAR data. We describe the underlying platform, inherited from the High Energy Physics community and the modifications to these tools that were required to host sophisticated processing software. We detail these modifications and discuss the resulting increase in throughput. Finally, we make estimations on the processing time saved by parallelizing LOFAR data processing. The work described in these chapters is essential to producing

scientific data sets at a high cadence, particularly considering the high data rates produced by LOFAR.

[4em]8emResearch Question 2: How can we build software to effortlessly accelerate complex pipelines for Radio Astronomy?

Chapters 2 and 3 detail the advances in parallelizing complex scientific pipelines on a distributed shared infrastructures. We integrate a mature workflow orchestration package with distributed LOFAR processing. We discuss the need for this orchestration, as well as the abilities to support additional complex pipelines. As an example application, we build a Continuous Integration pipeline tasked with verifying and validating the initial steps of LOFAR processing.

[4em]8emResearch Question 3: Can we automatically collect performance information during massively distributed processing and predict run times for future data sets?

Chapters 2 and 3 describe a performance monitoring suite for LOFAR data and our scalability model for LOFAR processing. When running massively distributed processing, scientists are unable to monitor the performance of the underlying software. Collecting these statistics is necessary for understanding processing inefficiencies and suggest ways to accelerate data processing.

Performance data can also be used to understand the effect of processing parameters on the resource usage of complex pipelines. We study this in detail, building a model that can be used to understand the scalability of multiple processing steps. This model shows the limitations on scientific parameters imposed by limited processing resources as well as suggestions on decreasing processing time without sacrificing scientific data quality.

Limitations

Using the software described, the LOFAR Surveys team was able to process several petabytes of archived data and produce scientific quality images. Despite the successes of the project, several issues occasionally impede data processing and prevent rapid deployment of software pipelines.

The primary issue is difficulty deploying processing software on the restricted computational clusters at Forschungszentrum Julich, one of the three LOFAR archive locations. As this cluster neither supports docker, singularity, nor CVMFS, deploying new software is difficult and time-consuming. Additionally, orchestrating jobs at this site requires additional integration with our tools.

Furthermore, our tools do not include data quality checks and automatic reprocessing, meaning some user overhead is needed to check the quality of each data-set and to debug common processing errors. As the software and scientific pipelines mature, we expect this issue to be slowly resolved.

Finally, our current software distribution does not give semantic versions to software images, nor is there a way to store these images or cite them in

related papers. Implementing these features will not only make data processing easily reproducible but also make it possible to recognize the effort put into building and distributing software images.

Future Work

This work focuses on the first stages of LOFAR data processing because of the substantial gains possible by parallelization. We take in mind the complexity of our processing workflows, the full range of scientific pipelines and the heterogeneous nature of the underlying infrastructure. Because of these factors, a wide range of astronomical pipelines can use the software presented in this work. Moreover, we can incorporate processing hosted at scientific institutions and cloud providers to scale scientific processing horizontally. One application for large scale distributed processing is the Square kilometer Array.

The Square Kilometer Array, (SKA) is a planned aperture synthesis radio telescope expected to have a total collecting area of one square kilometer. It is expected to produce more than 160 TB per day johnston2017taming, data which needs to be promptly processed. Scaling our tools to SKA-size processing requires a federation of clusters able to handle a high throughput workload. Nevertheless, as the SKA data processing will use different software tools than LOFAR, further study is needed on the optimal processing strategy for each of the many SKA science projects.

Current and future astronomical surveys provide infrastructure for scientists to efficiently process, reprocess archival data and import it into an interactive

environment. In this environment, an astronomer can study the data in detail using their tools or software packages provided to them. Our software can be included in such a science portal to make LOFAR processing easy, fast, and accessible.

In recent years, academia has begun focusing on ease of access and reproducibility of science. Science done with cutting edge instruments, such as LOFAR, tends to be time-consuming to reproduce. Barriers such as setting up a working software environment and downloading massive data-sets prevent scientists from quickly and easily reproducing the results of their peers. These barriers make it difficult to verify the accuracy of new[1] discoveries and need to be overcome[2] in order to make astronomy more honest and transparent. While our tools make it easy to do massive data processing, they can be further integrated into a science portal, where users can reproduce a publication's results using the exact software, environment and input data. Integrating our tools with this portal is crucial to making Radio Astronomy both more accessible and more authoritative.

tocchapterAcknowledgments

tocchapterBibliography

1.	~~new~~	Wordy Sentences	Clarity
2.	*be overcome*	Passive Voice Misuse	Clarity