HPC and Distributed Computing for Students in Science and Non-Science Programs

Suzanne K. McIntosh
Cloudera Inc. and New York University
New York, N.Y. U.S.A.
mcintosh@cs.nyu.edu

***Abstract*—The unprecedented growth in applications that leverage high performance computing (HPC) platforms ranging from supercomputers and grid computing systems, to Hadoop clusters, has created demand for graduates with skills in parallel programming, data science, and big data engineering.**

Traditionally, HPC courses have been part of the Computer Science curriculum. As HPC applications find their way into finance, medicine, life sciences, advertising, and other industries, an HPC curriculum to simultaneously address the needs of Computer Science students and students studying business, life sciences, or mathematics is required.

Keywords—distributed computing, education, curriculum, Hadoop, parallel programming

I. Introduction

In developing a High Performance Computing (HPC) curriculum for students from various disciplines, one challenge was bridging the programming languages gap. In particular, Computer Science and Engineering students typically learn C, C++, Java, or Python, all of which are supported by Hadoop's compute component, MapReduce. However, students in other disciplines may program in higher level languages such as SQL or R, for example.

Fortunately, Hadoop provides several programming language options to bridge the programming gap and make HPC available to a wider user base.

II. Instructional Approach

The resulting curriculum focuses on teaching parallel programming concepts using Hadoop, a distributed computing system that has become widely adopted by a variety of industries in part due to its low cost of adoption. Hadoop's compute component, MapReduce, is characterized as a SIMD HPC system. SIMD is an acronym for 'single instruction, multiple data' processing where a different dataset is processed by each of the nodes in the distributed computing system, but all nodes run the same instructions. Other HPC systems support both SIMD and MIMD processing (multiple instruction, multiple data).

To complete early labs, students use either a Hadoop cluster or they use a cost-free, fully configured Hadoop virtual machine (VM) available from a number of Hadoop distributors [1]. The VM is sufficient for completing the early labs because the data sources are small. These labs support introductory lessons on the Hadoop architecture, core Hadoop, and the Hadoop ecosystem of tools.

Core Hadoop is comprised of a distributed storage component, HDFS (which is inspired by the Google File System [4]), and a distributed compute component, MapReduce [3]. Students use the main course textbook [5] and required papers, e.g. [3][4], to supplement their learning.

In addition to core Hadoop, the course covers Hadoop ecosystem tools including the Hive [7] and Pig [6] high level programming languages. Hive is a SQL-like language, similar to the SQL used in programming database systems. Pig provides a data flow language with simple commands for transforming a data set. Both of these languages facilitate programming tasks, particularly for users who are less comfortable programming in languages supported at the MapReduce level (e.g. Java, C++, Python, etc.).

III. Analytics Projects

In the second half of the course, students form project teams to develop self-designed analytics projects using the tools learned in the first half of the course. Students identify and obtain multiple large data sources in support of their projects, and datasets are transferred to Hadoop where the analytic executes.

Students apply at least two Hadoop technologies of their choosing in the development of their projects. This flexibility allows students with SQL skills the option to use Hive or Impala [9] in developing their analytics. The Pig Latin language is yet another high level language option for programming a Hadoop solution. Other HPC systems do not typically provide such a high level programming abstraction. It is by virtue of these abstractions that HPC programming is approachable by non-Computer Science majors. The advantage here is that those with domain knowledge can interact directly with HPC systems to solve problems and explore data.

The team-based approach brings together students from various disciplines thereby collecting complementary skills and domain knowledge into each project team. Not only are

the programming skills of the team enhanced through a diverse set of members, but this in fact models real-world teams, and exposes students to the challenges of working together to achieve a common goal.

To complete the analytics projects, teams either use the university Hadoop cluster, or they use public cloud resources. Public cloud options include Amazon Elastic MapReduce (EMR), or installation of Hadoop on VMs in Amazon Elastic Compute Cloud (EC2). Hadoop can be installed manually, or through the use of automated methods such as [11] and [10].

IV. Outcomes

Approximately one hundred twenty graduate students have completed this course first offered in the Spring 2013 semester. Another one hundred students will complete the course in this semester alone, owing to the ever-growing demand for Hadoop skills in a variety of industries.

The efficacy, quality, and timeliness of topics covered are tracked using anonymous course evaluations completed by students midway through the semester and again at the end of the semester. Course content is continually adjusted to incorporate student feedback and to track with the rapidly evolving Hadoop landscape. Ultimately, the goal of the course is to ensure students learn those skills which are key differentiators in the prevailing job market.

Finally, the course requires every student to develop an analytic of their choosing. Students are encouraged to form teams with a maximum of four members per team. Experience has shown that it is feasible to define projects of sufficient coding workload for teams of this size.

Students have developed analytics in the domains listed below. One project, "Detecting Movement Paths and Patterns Using Wireless Access Point Logs", was awarded an NYU Courant Innovation Fellowship in 2014.

Business, Finance, Legal Projects

- News- and Social Media-based Stock Performance Prediction
- Predictions for Success of Startup Investments
- U.S. Patents Analytic
- Local Real Estate Value Prediction

Health and Safety

- Breast Cancer Staging and Prediction
- Healthcare Relationship to Employment in the U.S.
- Public Safety and Mapping Analytic

Sports and Entertainment

- Prediction of NBA Game Pace, Dynamically Ranking Soccer Players
- Movie, Song, and Book Recommendation
- Song Popularity Based on Listening Patterns

Miscellaneous

- Traffic Monitoring/Impact of Weather on Road Travel
- A Smarter Transportation Plan for New York City
- Drought and Earthquake Prediction
- Effect of Title and Posting Time on Popularity of Social Media Posts
- Resume Matching for Job Seekers
- Twitter Feed Sentiment Analysis
- Understanding Hierarchies in Technical Conferences

V. Future Work

The Hadoop landscape is continuously and rapidly changing. In order to keep pace with the technology, this course is updated frequently. Alternately, one may use a Hadoop curriculum maintained by a Hadoop distributor [8].

The next major course upgrade includes the addition of Spark, a language that introduces the concept of resilient distributed datasets, or RDDs, which can be persisted in memory to allow for highly performant data processing [2].

VI. Conclusion

In the past, access to a supercomputer was a prerequisite for learning about HPC. In recent years, we have discovered Hadoop clusters as low-cost platforms for learning HPC. We can build a Hadoop cluster using commodity, off-the-shelf hardware, rather than purchasing specialized and very expensive supercomputers. Supercomputers are not only expensive to acquire, they are expensive to operate and to maintain.

With access to a Hadoop cluster, we are now in a position to leverage this low-cost platform to teach HPC to students who would otherwise have no opportunity to learn about HPC.

VII. Acknowledgements

We thank Amazon for supporting student projects through educational grants. We also thank Pennsylvania State University's College of Information Sciences and Technology for providing access to the CiteSeerx data. Finally, thanks to the domain experts for supporting the development of student analytics projects.

VIII. References

[1] http://www.cloudera.com/content/cloudera-content/cloudera-docs/DemoVMs/Cloudera-QuickStart-VM/cloudera_quickstart_vm.html

[2] M. Zaharia, M. Chowdhury, M. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. 2nd Usenix Conference on Hot Topics in Cloud Computing, 2010.

[3] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In Proceedings of the 6th Symposium on Operating Systems Design and Implementation, OSDI'04, 2004.

[4] S. Ghemawat, H. Gobioff, and S. Leung. The Google File System. 19th ACM Symposium on Operating Systems Principles, 2003.

[5] T. White. Hadoop: The Definitive Guide, 3rd Ed.. Yahoo Press, 2010.

[6] A. Gates, O. Natkovich, S. Chopra, P. Kamath, S. M. Narayanamurthy, C. Olston, B. Reed, S. Srinivasan, and U. Srivastava. Building a High-Level Dataflow System on top of Map-Reduce: The Pig Experience. Proceedings of the VLDB Endowment, 2009.

[7] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Antony, H. Liu, P. Wyckoff, and R. Murthy. Hive - A Warehousing Solution Over a Map-Reduce Framework. Proceedings of the VLDB Endowment, 2009.

[8] http://www.cloudera.com/content/dev-center/en/home/academic-partnership.html

[9] Impala. http://www.impala.io/

[10] http://cloudera.com/content/cloudera/en/products-and-services/director.html

[11] Apache Whirr. https://whirr.apache.org/