Apache Crunch

SCOTT MCCLARY[1,*]

[1] School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
[*] Corresponding authors: scmcclar@indiana.edu

paper-002, April 12, 2017

Apache Crunch is an Application Programming Interface (i.e. API) designed for the Java programming language. This software is built on top of Apache Hadoop as well as Apache Spark and simplifies the process of developing MapReduce pipelines. Apache Crunch abstracts away the explicit need to manage MapReduce jobs. This defining characteristic alleviates much of the steep learning curve inherently within developing scalable applications that utilize a MapReduce type approach. Therefore, developers using Apache Crunch are able to streamline the process of converting Big Data solutions into runnable code. As a result, this Java API is leveraged in industry and academia to develop efficient, scalable and maintainable codebases for Big Data solutions.

Keywords: Big-Data, Cloud, Hadoop, MapReduce

https://github.com/cloudmesh/sp17-i524/blob/master/paper2/S17-IO-3011/report.pdf

1. INTRODUCTION

Apache Crunch is an open source Java API that is "built for pipelining MapReduce programs which are simple and efficient;" more specifically, Crunch allows developers to write, test and run MapReduce pipelines with minimal upfront investment [1, 2]. The minimal upfront investment of this API lowers the barrier for entry for Big Data developers.

Apache Crunch's purpose is to make "writing, testing, and running MapReduce pipelines easy, efficient, and even fun" [2, 3]. This open source Java API provides a "small set of simple primitive operations and lightweight user-defined functions that can be combined to create complex, multi-stage pipelines" [3]. Apache Crunch abstracts away much of the complexity from the user by compiling "the pipeline into a sequence of MapReduce jobs and manages their execution" [2, 3].

1.1. History

Josh Wills at Cloudera was the major contributor/developer to the Crunch project in 2011 [4, 5]. The original version of this software was based on Google's FlumeJava library [4–6]. From 2011 until May 2012 (i.e. version 0.2.4), the Apache Crunch project was open sourced at GitHub [4, 7]. After May 2012, the original Crunch source code was donated to Apache by Cloudera and shortly after "the Apache Board of Directors established the Apache Crunch project in February 2013 as a new top level project" [4]. Since February 2013, the Apache Crunch project continues to be used, maintained and improved in an open source fashion by the software's user and developer community. The user community has grown to include large reputable companies such as Spotify and Cerner [8, 9].

1.2. Advantages

As Hadoop continues to grow in popularity, the variation of data (i.e. satellite images, time series data, audio files, and seismograms) that is stored in HDFS grows as well [3, 10]. Many of these data "formats are not a natural fit for the data schemas imposed by Pig and Hive;" therefore, "large, custom libraries of user-defined functions in Pig or Hive" or "MapReduces in Java" have to be written, which significantly "drain on developer productivity" [2, 3, 11, 12]. The Crunch API provides an alternative solution, which does not inhibit developer productivity. Apache crunch integrates seamlessly into Java and therefore, allows developers full access to Java to write functions. Thus, Apache Crunch is "especially useful when processing data that does not fit naturally into relational model, such as time series, serialized object formats like protocol buffers or Avro records, and HBase rows and columns" [13–15].

1.3. API

Apache Crunch is a Java API that is used "for tasks like joining and data aggregation that are tedious to implement on plain MapReduce" [13]. The Apache Software Foundation provides thorough documentation of the API for Apache Crunch and even provides useful examples of how to explicitly leverage this API from a Java application [13].

1.3.1. Shell Access

For users of the Scala programming language, there is the "Scrunch API, which is built on top of the Java APIs and in-

cludes a REPL (read-eval-print loop) for creating MapReduce pipelines" [13].

2. LICENSING

The Apache Software Foundation, which includes the software tool named Apache Crunch, is licensed under the Apache License, Version 2.0 [16].

2.1. Source Code

Apache Crunch leverages Git for version control, which allows the user and developer communities to contribute freely to this open source project [7, 17].

3. ARCHITECTURE & ECOSYSTEM

In the simplest of terms, Apache Crunch runs on top of Hadoop MapReduce and Apache Spark [13]. Therefore, Apache Crunch abstracts away the need for the programmer to explicitly manage the MapReduce jobs through a Java API. However, the Apache Crunch's place within the software stack (i.e. on top of Hadoop MapReduce and Apache Spark) indicates its reliance on the MapReduce software subsystem. Given Apache Crunch's dependence on Hadoop MapReduce and Apache Spark, this API provides the ability for developers use the Java programming language to efficiently and effectively leverage MapReduce style processing to solve their complicated and complex Big Data problems.

4. USE CASES

Apache Crunch has its applicability in the Cloud Computing and Big Data industry, as shown in the following section. The widespread usage of Java, Apache Hadoop and Apache Spark in Cloud Computing help promote Apache Crunch in industry and academia alike.

4.1. Use Cases for Big Data

The Apache Hadoop ecosystem indicates that Apache Crunch is built on top of Hadoop MapReduce and Apache Spark, which both go hand in hand in solving many complicated and challenging Big Data problems. The following sections demonstrate Apache Crunch's applicability in Big Data problems. Furthermore, these use cases explains that the software facilitates the rapid and clean development of the respective Big Data solutions. The benefits realized at companies such as Cerner and Spotify are due in part to Apache Crunch's well-defined applicability in the Big Data space.

4.2. Cerner

Cerner, "an American supplier of health information technology (HIT) solutions, services, devices and hardware" [18], employs Apache Crunch to solve many of their Big Data problems [9]. Cerner decided to use Apache Crunch since it interestingly solves what they refer to as "a people problem" [9]. As a company, they have noticed that Apache Crunch diminishes a potential steep learning curve for new employees and/or teams to leverage Big Data technologies in their projects.

Cerner definitively believes that Apache Crunch stands above the other "options available for processing pipelines including Hive, Pig, and Cascading" since the Apache Crunch API allows their employees to straightforwardly code solutions to Big Data problems [9, 11, 12, 19]. The diminished learning curve as a result of using Apache Crunch allows Cerner to focus their time, effort and money on performance tuning and/or algorithm adjustments rather than wasting a significant amount the developers time simply translating a Big Data problem into runnable and efficient MapReduce code [9].

4.3. Spotify

Spotify, the popular "music, podcast, and video streaming service" [20], leverages Apache Crunch to process the many terabytes of data generated every day by their large user community [8]. Spotify has been using Apache Hadoop since 2009 and have spent significant effort since then to develop tools that make it simple for the Spotify "developers and analysts to write data processing jobs using the MapReduce approach in Python" [2, 8].

However, in 2013 Spotify came to the realization that this approach wasn't performing well enough so they decided to start using Java and Apache Crunch to solve their Big Data problems [8]. This transition to Apache Crunch resulted in higher performance, higher-level abstractions (e.g. filters, joins and aggregations), pluggable execution engines (e.g. MapReduce and Apache Spark) and added simple powerful testing (e.g. fast in-memory unit tests) [8]. Apache Crunch has given Spotify a significant enhancement for both their "developer productivity and execution performance on Hadoop" [8].

5. EDUCATIONAL MATERIAL

Apache Crunch makes the process of developing applications that leverage MapReduce and Apache Spark easier; therefore, the learning curve is much less significant in relation to developing applications that directly interact with MapReduce and Apache Spark. The Apache Software Foundation provides a lot of useful documentation. For instance, there is API documentation [21] as well as getting started information [22], a user guide [23] and even source code installation information [17]. If this is not enough, complete and extensive third-party code examples explain how to develop "hello world" applications that use Apache Crunch [24].

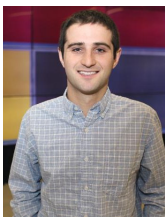
6. CONCLUSION

In general, Apache Crunch simplifies the process of writing and maintaining large-scale parallel codes by abstracting away the need to manage MapReduce jobs. This abstraction diminishes the inherent learning curve in solving Big Data problems and therefore allows developers to focus their time and effort in developing the general concept of their solution rather than in the detailed process of writing their code. The aforementioned benefits of Apache Crunch are proven by its widespread use in industry (e.g. Spotify and Cerner) and in academia. This software tool helps diminish the gap between domain scientists solving Big Data problems and the potentially complicated Computer Science tools/mechanisms provided to the Cloud Computing/Big Data community.

ACKNOWLEDGEMENTS

The authors would like to thank the School of Informatics and Computing for providing the Big Data Software and Projects (INFO-I524) course [25]. This paper would not have been possible without the technical support & edification from Gregor von Laszewski and his distinguished colleagues.

AUTHOR BIOGRAPHIES



Scott McClary received his BSc (Computer Science) and Minor (Mathematics) in May 2016 from Indiana University and will receive his MSc (Computer Science) in May 2017 from Indiana University. His research interests are within scientific application performance analysis on large-scale HPC systems. He will begin working as a Software Engineer with General Electric Digital in San Ramon, CA in July 2017.

WORK BREAKDOWN

The work on this project was distributed as follows between the authors:

Scott McClary. He completed all of the work for this paper including researching and testing Apache Airavata as well as composing this technology paper.

REFERENCES

[1] Edupristine, "Hadoop ecosystem and its components," Web Page, apr 2015, accessed: 2017-3-26. [Online]. Available: http://www.edupristine.com/blog/hadoop-ecosystem-and-components

[2] I. Wikimedia Foundation, "MapReduce - Wikipedia," Web Page, apr 2017, accessed: 2017-4-9. [Online]. Available: https://en.wikipedia.org/wiki/MapReduce

[3] J. Wills, "Introducing crunch: Easy mapreduce pipelines for apache hadoop," Blog, oct 2011, accessed: 2017-3-26. [Online]. Available: http://blog.cloudera.com/blog/2011/10/introducing-crunch/

[4] The Apache Software Foundation, "Apache Crunch - About," Web Page, 2013, accessed: 2017-3-26. [Online]. Available: https://crunch.apache.org/about.html

[5] Cloudera, Inc., "Big data | Machine Learning | Analytics | Cloudera," Web Page, 2017, accessed: 2017-4-09. [Online]. Available: https://www.cloudera.com

[6] C. Chambers, A. Raniwala, F. Perry, S. Adams, R. R. Robert R. Henry, R. Bradshaw, and N. Weizenbaum, "FlumeJava: Easy, Efficient Data-Parallel Pipelines," in *2010 ACM SIGPLAN Conference on Programming Language Design and Implementation*, ser. PLDI '10. Toronto, Ontario, Canada: ACM, 2010, pp. 363–375. [Online]. Available: http://doi.acm.org/10.1145/2609441.2609638

[7] GitHub, Inc., "GitHub," Web Page, 2017, accessed: 2017-4-09. [Online]. Available: https://github.com

[8] J. Kestelyn, "Data processing with apache crunch at spotify," Blog, feb 2015, accessed: 2017-3-26. [Online]. Available: http://blog.cloudera.com/blog/2015/02/data-processing-with-apache-crunch-at-spotify/

[9] M. Whitacre, "Scaling people with apache crunch," Blog, may 2014, accessed: 2017-3-26. [Online]. Available: http://engineering.cerner.com/blog/scaling-people-with-apache-crunch/

[10] The Apache Software Foundation, "Welcome to Apache Hadoop!" Web Page, mar 2017, accessed: 2017-4-9. [Online]. Available: http://hadoop.apache.org

[11] ——, "Welcome to Apache Pig!" Web Page, jun 2016, accessed: 2017-4-9. [Online]. Available: https://pig.apache.org

[12] ——, "Apache Hive TM," Web Page, 2014, accessed: 2017-4-9. [Online]. Available: https://hive.apache.org

[13] ——, "Apache Crunch Simple and Efficient MapReduce Pipelines," Web Page, 2013, accessed: 2017-3-26. [Online]. Available: https://crunch.apache.org

[14] ——, "Welcome to Apache Avro!" Web Page, may 2016, accessed: 2017-4-9. [Online]. Available: https://avro.apache.org

[15] ——, "Apache HBase – Apache HBase Home," Web Page, apr 2017, accessed: 2017-4-9. [Online]. Available: https://hbase.apache.org

[16] ——, "Apache license, version 2.0," Web Page, jan 2004, accessed: 2017-3-26. [Online]. Available: http://apache.org/licenses/LICENSE-2.0.html

[17] ——, "Getting the source code," Web Page, 2013, accessed: 2017-3-26. [Online]. Available: https://crunch.apache.org/source-repository.html

[18] I. Wikimedia Foundation, "Cerner - Wikipedia," Web Page, mar 2017, accessed: 2017-3-26. [Online]. Available: https://en.wikipedia.org/wiki/Cerner

[19] Xplenty Ltd, "Cascading | Application Platform for Enterprise Big Data," Web Page, 2016, accessed: 2017-4-9. [Online]. Available: http://www.cascading.org

[20] I. Wikimedia Foundation, "Spotify - Wikipedia," Web Page, mar 2017, accessed: 2017-3-26. [Online]. Available: https://en.wikipedia.org/wiki/Spotify

[21] The Apache Software Foundation, "Apache crunch 0.15.0 api," Web Page, 2017, accessed: 2017-3-26. [Online]. Available: https://crunch.apache.org/apidocs/0.15.0/

[22] ——, "Apache Crunch - Getting Started," Web Page, 2013, accessed: 2017-3-26. [Online]. Available: https://crunch.apache.org/getting-started.html

[23] ——, "Apache Crunch - Apache Crunch User Guide," Web Page, 2013, accessed: 2017-3-26. [Online]. Available: https://crunch.apache.org/user-guide.html

[24] N. Asokan, "Learn Apache Crunch," Blog, Mar 2015, accessed: 2017-3-26. [Online]. Available: http://crunch-tutor.blogspot.com

[25] Gregor von Laszewski and Badi Abdul-Wahid, "Big Data Classes," Web Page, Indiana University, Jan. 2017. [Online]. Available: https://cloudmesh.github.io/classes/