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 9, 2017

Apache Crunch is a Java API that simplifies the process of developing MapReduce pipelines. This library is built on top of Apache Hadoop and Apache Spark and is therefore used in industry to develop efficient, scalable and maintainable codebases for Big Data solutions. The main benefit of Apache Crunch is that the explicit need to manage MapReduce jobs has been abstracted away. Thus, Apache Crunch alleviates much of the steep learning curve inherently within developing scalable applications that utilize a MapReduce type approach.

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" [1]. More specifically, Crunch allows developers to write, test and run MapReduce pipelines with minimal upfront investment [1]. As explained in Section 1.2, this Java API was developed by Josh Wills at Cloudera and is based on Google's FlumeJava library [2, 3].

Apache Crunch "aims to make writing, testing, and running MapReduce pipelines easy, efficient, and even fun" [4]. 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" [4]. 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" [4].

1.1. 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 [4]. 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" [4]. The Crunch API provides an alternative solution, which does not inhibit developer productivity. Apache crunch integrates seamlessly into Java and therefore, allow 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" [5].

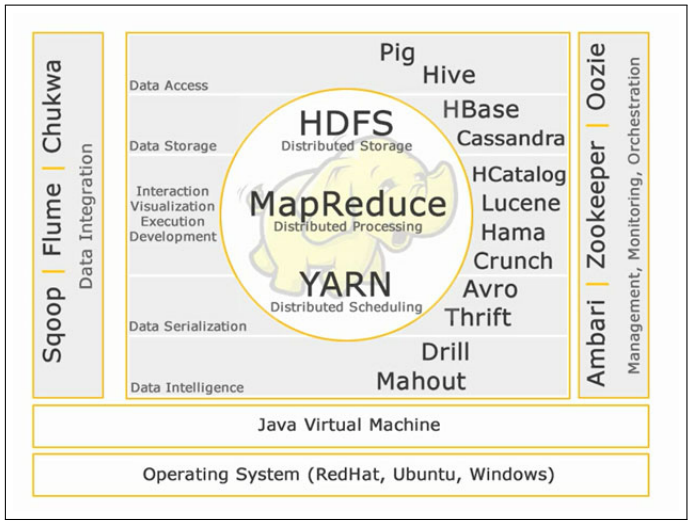


Fig. 1. The image above depicts the Apache Hadoop Ecosystem, including where Apache Crunch lies within the software stack [1].

1.2. About & History

The major contributor to the "initial source code of the Apache Crunch project [was] ... Josh Wills at Cloudera in 2011" [3]. Up until May 2012 (i.e. version 0.2.4), the Apache Crunch project was open sourced at GitHub [3]. After May 2012, "Cloudera donated the source code to Apache and the project entered

the Apache Incubator, ... [a]fter 9 months at the Incubator, ... the Apache Board of Directors established the Apache Crunch project in February 2013 as a new top level project" [3]. Since February 2013, the Apache Crunch project continues to be used, maintained and improved in an open source fashion, as explained in Section 2.1.

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" [5]. Section 5 explains that the Apache Software Foundation provides thorough documentation of the API and provides examples of how to explicitly leverage this API from a Java application.

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 includes a REPL (read-eval-print loop) for creating MapReduce pipelines" [5].

2. LICENSING

The Apache Software Foundation, which includes Apache Crunch, is licensed under the Apache License, Version 2.0 [6].

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].

3. ARCHITECTURE & ECOSYSTEM

Figure 1 and Figure 2 provide a complete graphical representation of the MapReduce ecosystem and explicitly indicates Apache Crunch's place within the software stack. In the simplest of terms, Apache Crunch runs on top of Hadoop MapReduce and Apache Spark [5].

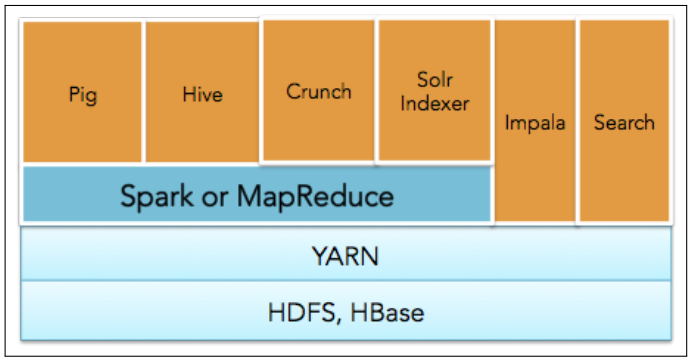


Fig. 2. The image above depicts the Apache Hadoop software stack, including Apache Crunch [8].

4. USE CASES

Apache Crunch has its applicability in the Big Data industry, as shown in Section 4.1. The widespread usage of Apache Hadoop and Apache Spark help to promote Apache Crunch in industry and academia alike.

4.1. Use Cases for Big Data

As explained in Section 3, 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.

4.2. Cerner

Cerner, "an American supplier of health information technology (HIT) solutions, services, devices and hardware" [9], employs Apache Crunch to solve many of their Big Data problems [10]. Cerner chose to use Apache Crunch since it interestingly solves what they refer to as "a people problem" [10]. As a company, they have noticed that Apache Crunch diminishes a potential steep learning curve for new employees and/or teams to use Big Data technologies. Specifically, Apache Crunch stands above the other "options available for processing pipelines including Hive, Pig, and Cascading" since Apache Crunch allows Cerner employees to "easily ... translate how [they] ... described a problem into concepts [that they] ... can code" [10]. 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 spending a significant amount time simply translating the problem into runnable code [10].

4.3. Spotify

Spotify, the popular "music, podcast, and video streaming service" [11], leverages Apache Crunch to process the many terabytes of data generated every day by their large user community [12]. Spotify has been using Hadoop since 2009 and have spent significant effort since then to develop tools that make it "easy for [their] developers and analysts to write data processing jobs using the MapReduce approach in Python" [12]. 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 [12]. 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) [12]. Apache Crunch has given Spotify a "huge boost for both [their] ... developer productivity and execution performance on Hadoop" [12].

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 [13] as well as getting started information [14], a user guide [15] and even source code installation information [7]. If this is not enough, complete and extensive third-party code examples explain how to develop "hello world" applications that use Apache Crunch [16].

6. CONCLUSION

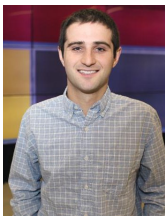
In general, Apache Crunch simplifies the process of writing and maintaining large-scale parallel codes by abstracting away the need to manage MapReduce jobs. This abstraction diminishes

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, shown in Section 4.

ACKNOWLEDGEMENTS

The authors would like to thank the School of Informatics and Computing for providing the Big Data Software and Projects (INFO-I524) course [17]. 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: https://crunch.apache.org/source-repository.html

[2] 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

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

[4] 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/

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

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

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

[8] J. Jairam Ranganathan, "Apache Spark in the Apache Hadoop Ecosystem," Blog, Sep 2014, accessed: 2017-3-26. [Online]. Available: https://vision.cloudera.com/apache-spark-in-the-apache-hadoop-ecosystem/

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

[10] 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/

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

[12] 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/

[13] 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/

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

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

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

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