Viability of map-reduce algorithms for the measurement of Higgs boson properties with the ATLAS detector at the LHC

Sheena O'Connell, School of Physics, University of the Witwatersrand, Johannesburg 2050, South Africa

Introduction

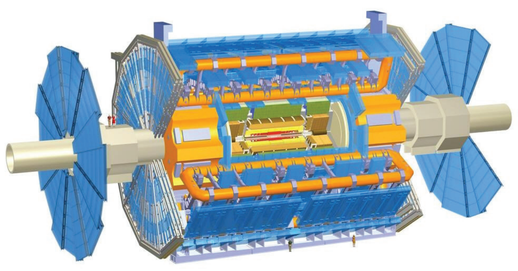
ATLAS is an experiment at the Large Hadron Collider at CERN that uses particle physics in order to search for new behaviours in the head-on collisions of protons of extraordinarily high energy. Following CERN's discovery of the Higgs boson, shifted interest in Higgs decay analysis at ATLAS has resulted in a many-fold increase in the volume of data required for computations, thus introducing the requirement more data transport over a computing cluster's internal network. This effects overall process efficiency as data transport is an expensive operation. Hadoop, an industry standard software platform for the analysis of big data, is designed to address many of the efficiency concerns faced in typical HEP analysis.

Aim

To use Hadoop to parallise ROOT based HEP analysis such that the solution is:

- comparable in efficiency to a ROOT + PROOF solution
- useful specifically for the analysis of higgs to two photon decay events

CURRENT METHODOLOGY



low lever triggering and selection removes

erronous and uninteresting data

GRID

tiered storage

ROOT basesed analysis

made parallel using PROOF

PREVIOUS APPROACHES

HEPOOP

AVRO

Z-Lib compressed ROOT file data

Pig

Jython

Possibly 3-10x slower than ROOT for certain tasks. not confirmed.

Python

Probably v.2.5.2

Numpy + Scikit Learn

Integrated nicely with a variety of industry standard technologies BUT no ROOT based implementation.

Performance "significantly inferior" to ROOT + PROOF based analysis especially when dealing with large nested structures

RootOnHadoop

HDFS

Standard ROOT files
Chunking disabled

List of file paths

Copy file to local storage
or
access file using fuse -20%-30% delay

Delay for disk access and takes up space

Java mapper (sFilePath)

Output saved as file

Standard ROOT analysis code

Java reducer (sFilePath)

Output path

~24% slower than ROOT + PROOF

Hadoop Streaming

GRID

Standard ROOT files stored as normal

List of file paths

ROOT map code

Mapper (sFilePath)

ROOT reducer

Reducer

Output path

~21% slower than ROOT + PROOF

PROPOSED SOLUTION

HBASE

ROOT files optimized for column store

Java mapper

Java reducer

ROOT wrapper

Makes use of alternative constructors instead of reading straight from files.

Slightly refactored ROOT analysis code

- OR -

HDFS

Simplified ROOT files (no error checking etc)

Java mapper

Java reducer

Custom loading methods to make sure no splits occur

ROOT

Receives unbroken bytestream for each file

- ⊕ Can tie into Pig for more complex algorithms
- ⊕ Column store unlocks lazy or sparce loading (later)
- ⊖ Makes the use of metadata difficult: It must be re-implemented for future use cases

- ⊕ Use almost standard constructors
- ⊕ Keeps ROOT's metadata capabilities
- ⊖ Lazy/Sparce loading probably not possible

Conclusion

Attempts to integrate Hadoop and ROOT have been made in the past. These integrations were successful in that they could run analysis code, they were however inefficient as compared to ROOT+PROOF based analysis. A new approach to the integration of the two technologies was described. This new approach side-steps many of the sources of inefficiency found in previous solutions but is yet to be fully tested.

Literature cited

- Gumpert C, Moneta L, Cranmer K, Kreiss S and Verkerke W Software for statistical data analysis used in Higgs searches
- Rene Brun and Fons Rademakers, Sep. 1996, ROOT - An Object Oriented Data Analysis Framework, Proceedings AIHENP'96 Workshop, Lausanne, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81-86
- White T, O'Reilly Media Inc 2012, Hadoop - The Definitive Guide, Third Edition
- ATLAS Experiment official website URL http://atlas.ch/, last accessed 20 May 2015
- Bhimji W, Bristow T, Washbrook A HEPDOOP: High-Energy Physics Analysis using Hadoop
- Russo S, Pinamonti M, Cobal M Running a typical ROOT HEP analysis on Hadoop MapReduce
- Lehrack S, Duckeck G, Ebke J Evaluation of Apache Hadoop for parallel data analysis with ROOT
- Press G The Hadoop Bubble Quivers As Hortonworks Misses, Posted on March 10, 2015 URL http://whatsthebigdata.com/2015/03/10/the-hadoop-bubble-quivers-as-hortonworks-misses/, Originallyposted on Forbes.com
- Hadoop: Industry Solutions URL http://hortonworks.com/industry/, last accessed 20 May 2015
- root-6.02.08 source code