A Study in Hadoop Streaming with Matlab for NMR data processing

Kalpa Gunaratna [1], Paul Anderson [2], Ajith Ranabahu [1]
&
Amit Sheth [1]

[1] **Kno.e.sis** - Ohio Center of Excellence in Knowledge-Enabled Computing
Wright State University, Dayton, Ohio
[2]**Air Force Research Laboratory**, Biosciences & Protection Division
Wright-Patterson AFB, Dayton, Ohio

12/01/2010

Outline

- Introduction
- Background
- Design
- Implementation
  - Baseline correction
  - Hadoop streaming
- Results & Discussion
- Conclusion

Introduction

- Biologists confront with huge amount of data (NMR spectrometers, etc).
- Have to undergo numerical processing like baseline correction, normalization, etc. even before doing anything useful.
- Important observation in a Biologists' context,
  - Even though increase in distributed computing tools they avoid using them much.
  - User friendly and domain specific tools are preferred over their lack of performance.
- *Best of both worlds for biologists….*
- Matlab code is run on Hadoop.

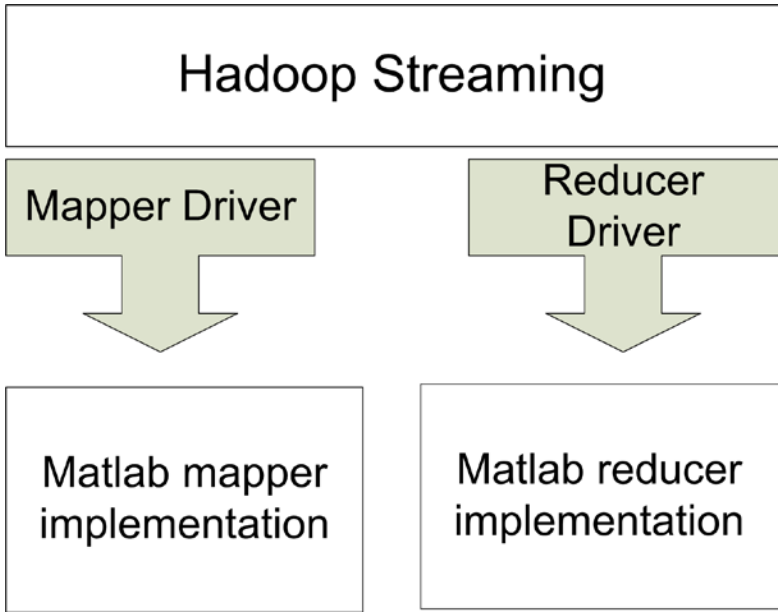
Background

- NMR (Nuclear Magnetic Resonance) data analysis normally consists of Giga bytes of data files.
  - A typical 1H NMR or C13 spectrum contain thousands of resonances.
- Metabolomics
  - Assess end product unlike proteomics and genomics.
  - NMR spectroscopy of biofluids is an effective method for identifying variations in states.

Background cont.

- Baseline distortion
  - Arise from hardware and processing sources.
  - Can lead to incorrect metabolites quantification which leads to spurious scientific conclusions.

Design



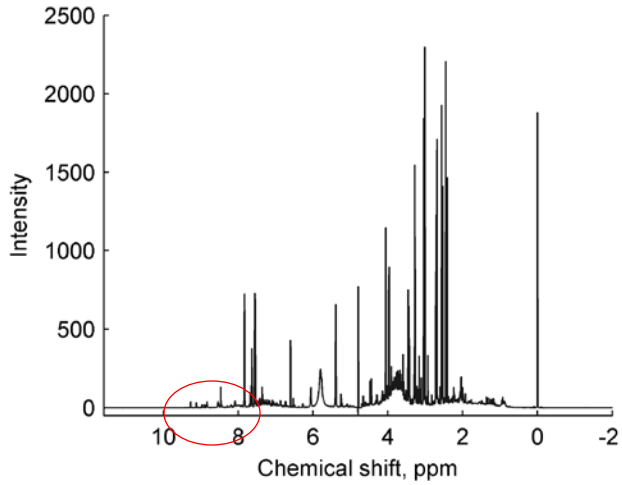
- Hadoop streaming is used with C++ driver applications.

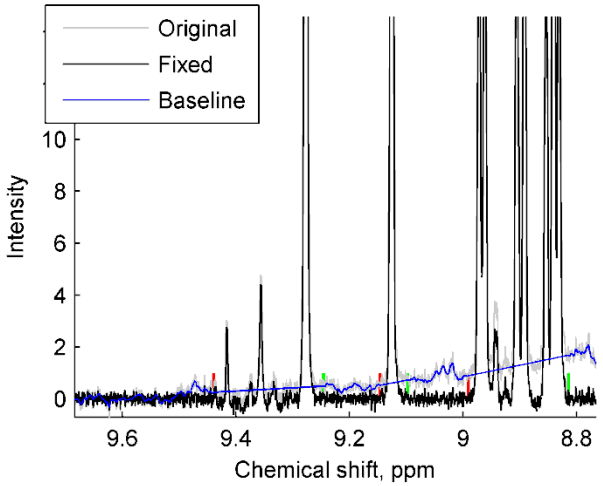
Design cont.

- Driver applications are used to read data from the source and call Matlab functions.
- The driver application is responsible for calling relevant Matlab code segments for computations.

Implementation

- Baseline correction





Implementation cont.

- Baseline correction
  - Whittaker Smoother algorithm is used.
  - The algorithm is written completely in Matlab.

Implementation cont.

- NMR Data Streaming
  - Driver application is written in C++.
  - Matlab code is compiled with C++ to create a shared library.
  - Driver acts as an interface for mapper in Hadoop and calls Matlab function.
- NMR spectra consist of columns and hence it is inverted to a row oriented file (Hadoop reads line by line).
- Our original Matlab baseline correction desktop code version is trivially changed here.

Implementation cont.

- Driver creates a relevant Matlab object for a column and passes to the Matlab function.
- For this specific example, a reducer is not necessary since each spectrum is restricted to a single row.
  - If spread across rows, reducers may be needed to format the output.

Implementation cont.

- Technical issues
  - Matlab seemed to have problems with reading directly from Hadoop streaming (need of driver application).
  - Matlab instances need to be available in nodes.

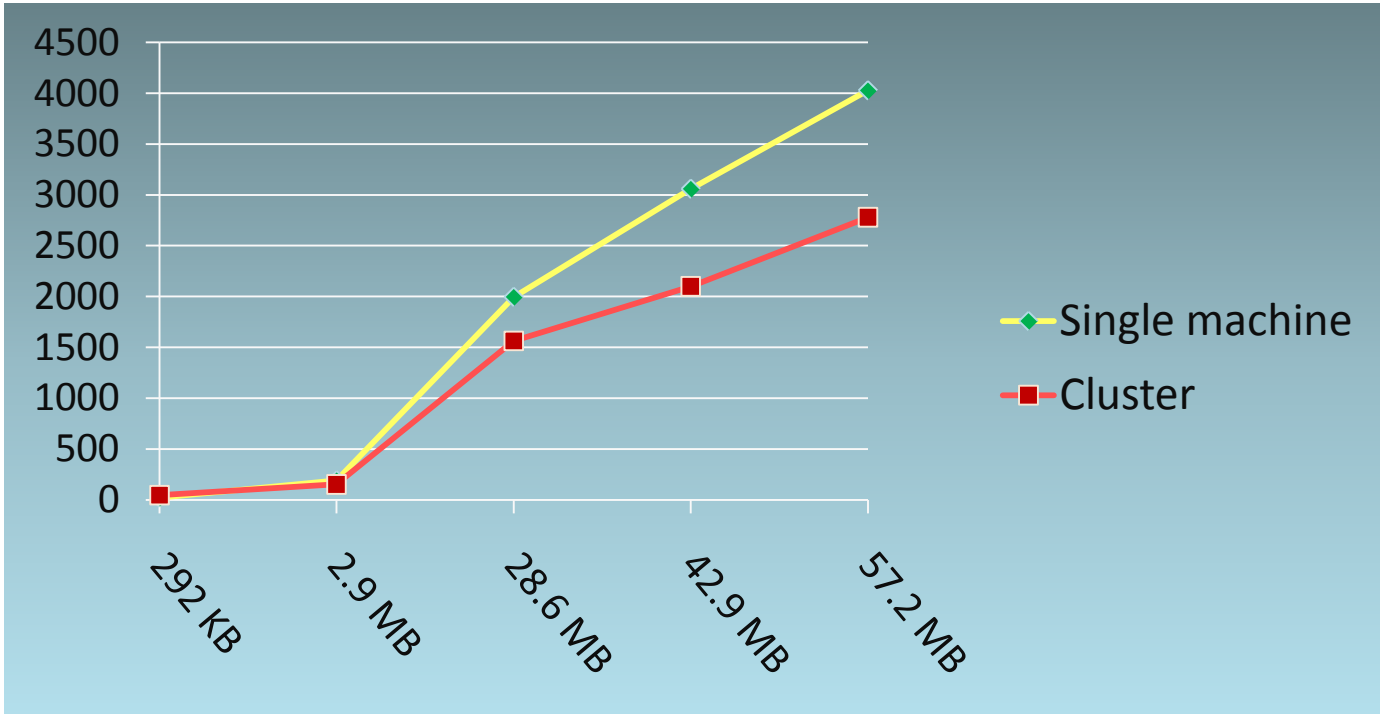
Results & Discussion

Size	Single machine(sec)	Cluster (sec)
292 KB (1 spectrum)	22	46
2.9 MB (10 spectra)	192	152
28.6 MB (100 spectra)	1996	1563
42.9 MB (150 spectra)	3059	2100
57.2 MB (200 spectra)	4027	2780

Cluster – 16 nodes of Quad core AMD Opteron with 16 GB of RAM

Single machine – 3 GHz dual core CPU with 4 GB of RAM

Results cont.



Results cont.

- Advantages of using Matlab on Hadoop.
  1. Scientists are relieved from learning new technologies having sharp learning curves (sometimes scripting languages are even incompatible with requirements of biologists).
  2. Non distributed code implementations which are readily available could be used in cloud environment without significant change.
- No need of paradigm shift. Code adoption cost is often expensive and repetitive.
- Facilitates **rapid testing** and **prototyping** where necessary.

Conclusion

- Cloud computing would not be feasible for scientists if they have to deviate from their routine practices significantly.
- Hence Hadoop streaming allows to use existing Matlab programs in Hadoop clusters.
- Our experiment reflects that using Matlab in Hadoop is feasible and could be extended for various requirements.

Questions

Thank You!

http://knoesis.org