Getting Code Near the Data: A Study of Generating Customized Data Intensive Scientific Workflows with Domain Specific Language

Ashwin Manjunatha[1], Ajith Ranabahu[1], Paul Anderson[2] and Amit Sheth[1]

[1]Ohio Center of Excellence in Knowledge Enabled Computing (Kno.e.sis)
[2]Air Force Research Laboratory, Biosciences & Protection Division
[1]{ashwin, ajith, amit}@knoesis.org / [2]paul.anderson2@wpafb.af.mil

1. Introduction

Metabolomics

The term metabolomics is defined as a comprehensive analysis in which metabolites of a biological system are identified and quantified. Any technique that can quantify metabolites can be used for metabolomics, but there are two primary techniques seen in the literature: nuclear magnetic resonance (NMR) and mass spectrometry with a prior online separation step such as high performance liquid chromatography (HPLC) or gas chromatography (GC). While neither technique is strictly superior, each technique has its own advantages and disadvantages. Existing applications include the identification of biomarkers associated with responses to toxin and pathophysiologic changes, sample classification based on the type of toxic exposure, large scale human studies, clinical diagnosis, and the study of genetic disorders.

Toxicology

Toxicology is the branch of pharmacology that deals with poisons and their effects on plant, animal and human life. Metabolic profiling especially of urine or blood plasma samples can be used to detect the physiological changes caused by toxic insult of a chemical.

NMR

Nuclear magnetic resonance (NMR) spectroscopy is an experimental technique that exploits the properties of an atom's nucleus. It can be used to obtain information about the concentration and structure of molecules. NMR studies magnetic nuclei by applying a static magnetic field followed by applying a second oscillating magnetic field. Specifically, only nuclei with an odd number of protons or neutrons can be measured using NMR; however, the two most common atoms studied are ^{1}H And ^{13}C.

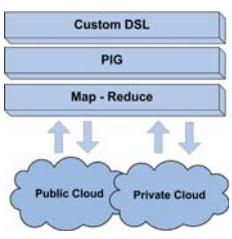
NMR Spectrometer

2. What is the problem ?

- Metabolomics generates large data sets
  - Hard to move around for processing
  - May not work well for the case of Service Oriented Workflow
- Processing is computationally intensive
  - Cannot be processed in a single machine
- Some institutions have regulations on data
  - E.g. Air Force Research Lab have strict regulations on data usage

Need a better way to take code near the data,
Not data near the code !

3. Using a DSL for mini Workflows



Architecture of the current prototype

Use a Domain Specific Language (DSL) to create a mini Workflow

- The DSL can be transformed into a Map-Reduce (Hadoop) program.
- Flexible enough to be compiled into other types of executable programs (E.g. Matlab program)
- Reduces the *number of hops* to create the Workflow. No need to do iterative programming with developers. The biologist can create the program by him/herself.

4. Example Workflow

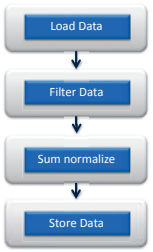
```
# load data
originaldata =
load_data_from_csv(rawdatafile)

#filter out a range
filtered = range_filter({:min=> 20,:max
=>50},originaldata)

# sum normalize
normalized = sum_normalize(filtered)

# write the file
store (normalized_datafile,normalized)
```

Sample DSL

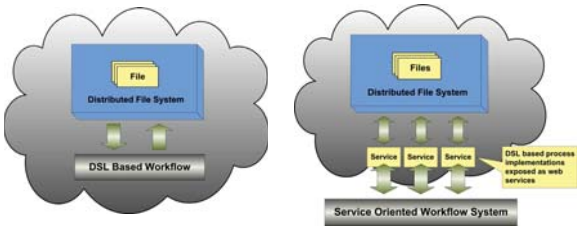


Workflow

```
A = LOAD '$filename' USING PigStorage(',')
      AS (colnum:int, value:double);
B = GROUP A BY colnum;
C = FOREACH B GENERATE group, SUM(A.value);
D = COGROUP A by colnum inner,  C by $0 inner;
F = FOREACH D GENERATE group, FLATTEN (A),FLATTEN (C);
G = FOREACH F GENERATE $0,($2/$4)*100;
STORE G INTO '$filename_processed' USING PigStorage (',');
```

PIG script for Sum Normalization

5. Usage Patterns



The DSL can be used in different ways

1. Can be used as a standalone workflow creator
2. Can be used to implement individual services for a service oriented workflow

6. Discussion

Convenience
Custom DSL is comprehensive to a biologist rather than a generic solution like PIGLatin.

Tools
Sophisticated tooling is required to enable the biologist to conveniently build the workflows.

Provenance and Metadata
DSL makes it easier to integrate provenance and other metadata.

Efficiency
The efficiency may be reduced but, it can be easily compensated by a cheaper improvement of computing power.

7. What is the bottom line for the Biologist ?

Faster processing and result generation
The backend services can run much faster than any single computer and hence provides the results faster.

Ability to handle very large datasets
Clouds are capable of handling much larger datasets than any single computer.

Works well inside regulated environments
Data can easily be processed in house without violating institutional regulations.

Reduced *hops* makes for quick turnaround
Most tasks can be programmed by the biologist without the need for skilled developers.