Big Data Management with Myria

Brandon Haynes and Magdalena Balazinska

Department of Computer Science & Engineering
University of Washington
http://myria.cs.washington.edu
Myria Team
Magda Balazinska, Bill Howe, and Dan Suciu (faculty)
Software Engineer: Tobin Baker
Grad Students
  Shumo Chu
  Eric Gribkoff
  Brandon Haynes
  Jeremy Hyrkas
  Paris Koutris
  Brendan Lee
  Brandon Myers
  Ryan Maas
  Dominik Moritz
  Laurel Orr
  Jennifer Ortiz
  Jingjing Wang
Ugrad Students
  Yuqing Guo
  Dan Radion

Alumnae/Alumni: Victor Almeida, Lee Lee Choo, Dan Halperin, Vaspol Ruamviboonsuk, Emad Soroush, Mayukha Vadari, Andrew Whitaker, Shengliang Xu
Acknowledgments

The Myria Team

Our science collaborators

Our sponsors

• National Science Foundation, Moore & Sloan Foundations, Washington Research Foundation, eScience Institute, ISTC Big Data, Petrobras, and EMC
Overview of the Myria Stack
Stack for big data management and analytics

- A new big data mgmt & analytics **system**
  - Available open source
  - Runs in shared-nothing clusters (Amazon EC2)
  - Also runs in an HPC cluster at MIT
  - Think of it as Hive/Hadoop but faster
  - Think of it as Spark but faster
  - This of it as SQL Server or PostgreSQL but more scalable

- An **operational service** deployed at UW

- Developed by the UW database group and eScience
Myria Big Data Management Service

Myria is a cloud service: Just open browser and go!

```plaintext
1  good_op_vct = scan(armbrustlab:sealow:good_op_vct_v4);
2  def avg_sd(x): [avg(float(x)), stdev(float(x))];
3  beads = select * from good_op_vct where pop = "beads";
4  bead_stats = select avg_sd(fsc_small) as [fsc_avg, fsc_sd],
5      avg_sd(chl_small) as [chl_avg, chl_sd],
6      avg_sd(pe) as [pe_avg, pe_sd],
7      Cruise from beads;
8  store(bead_stats,
9      armbrustlab:sealow:bead_stats_v4_bycruise_untrans);
```

Visualization of the logical and optimized physical query plan.

Relational algebra converted and optimized into a Myria Physical Plan.
Myria Demo on Amazon

http://demo.myria.cs.washington.edu
Myria Is a Cloud Service

- Browser
- MyMergerTree
- Python/iPython

RACO
- Web UI
  - Language Parser
  - Query Optimizer
- Google App Engine

MyriaX
- Parallel Query Execution
Myria Supports Different Back-Ends

**RAGO Goals:** Optimizing middleware for big data systems – write once, run efficiently anywhere: Hadoop, Spark, MyriaX, straight C, RDBMS, PGAS, MPI, ...
Myria is a parallel data management system.

MyriaX

REST Server

Coordinator

Catalog

Worker

Catalog

Worker

Catalog

Worker

Catalog

Shared-nothing cluster

Primary data store:

Can ingest data from:

S3

HDFS

HDFS

HDFS

RDBMS

RDBMS

RDBMS

JSON query plans & other instructions

Magdalena Balazinska - University of Washington
Performance Example

Input data: Twitter Graph 1.5 billion edges and 41 million vertices

![Performance Example Graph]
Example Myria Applications

Environmental Flow Cytometry

Galaxy Simulations

FSC

Ps3.FCS...subset

 depicts Prochlorococcus

FSC

Small Stuff

692-40 little stuff

Phytoplankton

Retail Analytics

Bibliometrics
Myria Demo

http://demo.myria.cs.washington.edu

- Demonstrating Web interface to Myria
  - Analyzing pre-loaded data
  - Analyzing data stored in S3
- Demonstrating Python interface to Myria
  - Basic Python and iPython
- Demonstrating spinning up your own cluster
  - Demo with Amazon EC2 but private cluster also OK
Myria Documentation

Docs: http://myria.cs.washington.edu/docs/index.html

Issues: Please post on github
https://github.com/uwescience/myria-stack

Users mailing list: myria-users@cs.washington.edu
To subscribe:
https://mailman.cs.washington.edu/mailman/listinfo/myria-users