A Flexible and Robust Workflow Analytics Framework for Cloud-based Data Intensive Computing

**PI:** Li Xiaorong (IHPC)

**Co-PI(s):** Qin Zheng (IHPC), Zhu Yongqing (DSI), Bharadwaj Veeravalli (NUS), Terence Hung (IHPC)
Where big data comes from?

- Sensors, mobile apps, etc
- Scientific modeling simulations
- Internet

[Images of sensors, mobile devices, and scientific modeling simulations]
Paradigm Shift in Computing

Azure Services Platform

Rackspace Hosting

NetApp

Eucalyptus Systems

Elastra

Salesforce.com

Cloudera

Joyent

Hosted Solutions

EDBT 2011 Tutorial
Big data challenges

Volume

Variety

Big Data

Velocity
Flexible Workflow for Cloud Data Analytic Services

Background & Motivation

- Structured and unstructured data
- Automatic sys. for big data processing

Methodology

- Analytic workflow automation & elastic resource allocation and optimization

Web-based tools for workflow management in Clouds

Process workflow tasks automatically

Integrative Climate Data Management
Scalable and adaptive workflow management in Clouds

- Manage **big, distributed** nature of data for data analytics workflow applications in Hybrid Clouds.

- Automate the workflow process for data analytics with multi-objective optimisation: **budget + performance**

- Dynamic resource allocation for **adaptive** services

*Successfully employed for weather data processing, traffic analysis, DNA synthesis, optimisation problem, ...*

**Achievement:**
- 2 Best Paper Awards (DISIO’11, TGCC’11)
- 1 Best Prototype Award (SC-BDA-2012)
- Top-tier conf papers (ICSOC’11, CloudCom’12, etc.)
- IEEE SCALE 2013 (1st prize)
- 1 TD submitted.
Auto-Scaling

Scale on the fly and real-time monitoring

1. Select the scaling methods

2. List VMs (VMs can be added or released dynamically during the processing)

3. Monitor workflow execution and generate the reports
# Fault Tolerant

## Statistics tables for execution

<table>
<thead>
<tr>
<th>Node Name</th>
<th>Done Jobs</th>
<th>Failed Jobs</th>
<th>Submit Time</th>
<th>Done Time</th>
<th>Time Taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>1</td>
<td>0</td>
<td>6:0:11</td>
<td>6:0:21</td>
<td>10</td>
</tr>
<tr>
<td>A15</td>
<td>1</td>
<td>0</td>
<td>6:0:57</td>
<td>6:1:49</td>
<td>52</td>
</tr>
<tr>
<td>A7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A5</td>
<td>1</td>
<td>7</td>
<td>6:5:14</td>
<td>6:5:18</td>
<td>6:5:18</td>
</tr>
<tr>
<td>A14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Chart

```
10.217.168.226: A24, A8, C2
10.217.168.227: A3, A4, A9, C3
10.217.168.235: A23
10.217.168.236: A7, A17, A12, A22, End
10.217.168.239: A2
10.217.168.240: A10, A6, A21
10.217.168.245: A7, A16, A19
10.217.168.250: Start, A15, A11, A19, A25, A5, A14, C1
10.217.168.251: A18, A16
175.41.155.66: A5, A14, Cost: 137
46.137.196.61: A11, A5, Cost: 137
```
Security-conscious data flow management

- Configuration and auto mapping

(a) Configuration of the security levels.

(b) Secured resources

Other methods: authorization, authentication, encryptions, etc
Self-learning and iterative optimization

Adaptive workflow engine for processing DAG and non-DAG workflows for big data processing

Interactive cost & performance Optimization in public, private, or hybrid-clouds
Various workflows types, MapReduce-enabled
Workflow-enable big data analytic framework

**Data**
- Climate: Temperature, humidity, rain force, etc. (Source: NEA)
- Geo-Population
- Life cycle of Aedes mosquito
- Dengue records (Source: MOH)

**Analytic Models**
- Evolutionary data analytics
  - Best resource allocation for parallel data processing!
  - Best workflow for data analytics!
  - Best sources for relevant data!

**Data Processing**
- Workflow Mgt
- Resource Mgt

**Visualization**
- A graphical user interface

**New!**
CASE STUDY AND DEMOS
Traffic patterns analysis

Analytics workflow

- Traffic News (as text)
- Text Parser and Extractor
- Spatial Information Elicitation Engine
- Spatial Relational Table

- Road Network Data
- Rainfall Data (as image)
- Geo Referencing Engine
- Vector on Raster Operation Engine
- Spatially-harmonized Data

- Road Network Data
- Spatially-harmonized Data
- Spatial Harmonization and Data Mash-up
- Harmonized and Mashed Data

- Road Network Data
- Spatially-harmonized Data
- Vector on Raster Operation Engine
- Geo Referencing Engine
- Image Aggregator
- Traffic Data (as images)
Climate-associated Dengue Fever Prediction in Clouds

- Predict the dengue fever by analyzing the temperature, rain forces, population density, etc.
- Visual map of high risk zones by spatial-temporal analysis
- Validate the analysis results with the real dengue cases

Example
The area with red color is highly likely to have more cases than those with dark blue.
Scalable Complex System for MRT Disruption

Scenario

7:00PM

No disruption
Typical day

with disruption
Disruption from: 6PM to 7PM

90 MRT stations
3000 trains
2 - 6.9 million agents
1000 AWS VM instances
Hybrid clouds
Set-up Simulation Scenario

- Set-up RTS Network Structure
- Set-up All Possible OD-Pairs and Routes
- Prepare Simulation

Run Simulation

- Create Station Agents
- Create Train Agents
- Create Commuter Agents

For $t$ in (Simulation time)

- Check All Commuter at time $t$
- Check all Station at time $t$
- Check all Train at line $l$ on time $t$

- Update Commuter
- Update Station
- Update Train
IEEE Scale Challenge Competition 2013 (CCGrid 2013)

- Manage **big, distributed** nature of data for data analytics workflow applications in Hybrid Clouds.
- Automate the **workflow process** for data analytics with multi-objective optimisation: **performance** and **budget**
- Dynamic resource allocation for **adaptive** services

Cloud-based satellite image processing for rainfall analysis

Achievement & Deliverables

- Prototype
  - 1 prototype of web-based cloud data workflow management system with monitoring, processing and monitoring has been implemented;
  - A new application on the big-data analysis for rainfall analysis of Singapore island via satellite images.
  - Presentation and demonstration in various events such as A*STAR Scientific conference 2012, Keynote paper in The 3rd international conference on emerging applications of information technology (EAIT 2012), ICPADS SC-BDA2012.
Parallel Spare Allocation & Optimisation

Workflow for Parallel Spare Allocation & Optimisation

Application
• Spares allocation optimisation across multiple airports (EADS)

Challenges
• Heuristics-based → many runs
• Memory intensive
• Quick turnaround needed

Improvement
• Initially: 60 parts in 1 week
• From 60 parts in 3 hours to 1066 parts across 59 airports in less than 1 hour

Chee Shin Yeo, Elaine Wong, Yong Siang Foo, “Handling Large Datasets in Parallel Metaheuristics: A Spares Management and Optimization Case Study”, in Proc. 25th AINA 2011
Extensibility: Compatible with other analytics tools

- Compatible with various analytic tools, such as GeoServer, Ncview, Rapidminer, RapidAnalytics, etc. to support data visualisation, data mining, and collaborative data analytics

Lu Sifei, et.al "A Framework for Cloud-Based Large-Scale Data Analytics and Visualization: Case Study on Multiscale Climate Data," HPCCloud Workshop, CloudCom2011.

THANK YOU