Towards an Optimized Big Data Processing System

The Doctoral Symposium of the IEEE/ACM CCGrid 2013
Delft, The Netherlands

Bogdan Ghiț, Alexandru Iosup, and Dick Epema

Parallel and Distributed Systems Group
Delft University of Technology
Delft, The Netherlands
PhD at TU Delft

Candidate: Bogdan Ghiț
Group: Parallel and Distributed Systems
Supervisors: Dick Epema, Alexandru Iosup
Research Topic: Resource management in grids and clouds
Start date: 24 October 2011
Finish date: 24 October 2015
“The Data Deluge”

“According to one estimate mankind created 150 exabytes (billion gigabytes) of data in 2005. This year, it will create 1,200 exabytes.”

The data is difficult to store, even harder to analyze it.
Data Sources

- **Computer Science**

- **LinkedIn**
  - Daily batch processing for "People you may know" recommendations

- **Facebook**
  - 30 PB by March 2011 = 3,000 x Library of Congress
MapReduce and Beyond

- Master-slave model
- MR-cluster
  - Stack of frameworks for large-scale data processing

- Multiple users vs. Isolation
  - MR-clusters on-demand
  - Isolation w.r.t. performance, data, failure, and versioning

- Data volume vs. Limited resources
  - Use resources from multiple clusters
  - Dynamically change the size

- Performance vs. Fairness
  - Capacity-based model
  - Capability-based model
Road Map

- **PhD start-up**: Nov. 2011
  - MR-clusters on-demand
  - Dynamic resizing of an MR-cluster
  - MTAGS 2012 Best Paper Award

- **Nov. 2012**: I. System prototype
  - Workload analysis
  - IEEE Big Data 2013

- **Nov. 2013**: II. Benchmarking Tool for MR
  - Fairness between multiple MR-clusters

- **Nov. 2014**: III. Performance model for MR
  - Scheduling decisions based on the data and the workload
  - Identify the performance boundaries

- **Nov. 2015**: IV. Optimize the resource management for MR
  - Applicability to other frameworks and infrastructures

- **PhD thesis**: Nov. 2015
Dynamic MapReduce Clusters

- Complex resource management
  - Single / multiple physical clusters
  - Placement and scheduling policies
  - Change resource allocations at runtime
  - Data management issues

- MR-cluster structure: *data replication vs. data locality*

  **Core nodes**
  - Execute tasks and store data locally
  - Replication required when removed

  **Transient nodes**
  - Execute tasks, do not store data
  - Data transfers to read/write data
Resizing Mechanism

Question: Given an MR-cluster, how can you tell if it is overloaded or underloaded?

- Monitor the MR cluster utilization: \( F_{\text{min}} \leq \frac{\# \text{tasks}}{\# \text{slots}} \leq F_{\text{max}} \)

- Grow-Shrink Policy (GSP) – with transient nodes
  - Size of grow and shrink steps: \( S_{\text{grow}} \) and \( S_{\text{shrink}} \)
  - Baseline policies: grow with core nodes (GGDP) or grow with transient nodes (GGP)

Timeline

\( F_{\text{min}} \leq \frac{\# \text{tasks}}{\# \text{slots}} \leq F_{\text{max}} \)
System Prototype

Koala Grid Scheduler
• Enables processor and data co-allocation
• Implements placement and scheduling policies
• Application types: cycle-scavenging, workflows, OpenMPI

Koala and MapReduce
• Developed an MR-Runner module to schedule MR jobs
• Provides isolated MR-clusters on a per-user basis
• Koala mechanism for resizing the MR-clusters
• MR jobs submissions transparent to Koala

DAS-4 Infrastructure
• Real-world experiments on a multicluster system
• 6 clusters, over 1600 cores, 150 machines, 180 TB, 1-10 Gbit/s
Transient Nodes

- Wordcount scales better than Sort on transient nodes

50 GB data set
Resizing Performance

- Resizing bounds
  \[ F_{\text{min}} = 0.25 \]
  \[ F_{\text{max}} = 1.25 \]

- Resizing steps
  - GSP
    \[ S_{\text{grow}} = 5 \]
    \[ S_{\text{shrink}} = 2 \]
  - GG(D)P
    \[ S_{\text{grow}} = 2 \]

50 MR jobs
1...50 GB
Road Map

- **PhD start-up**: Nov. 2011
- **MR-clusters on-demand**: Nov. 2012
- **Dynamic resizing of an MR-cluster**: Nov. 2012
- **MTAGS 2012 Best Paper Award**:

**I. System prototype**
- **IEEE Big Data 2013**

**II. Benchmarking Tool for MR**
- **Workload analysis**

**III. Performance model for MR**
- **Fairness between multiple MR-clusters**

**IV. Optimize the resource management for MR**
- **Scheduling decisions based on the data and the workload**
- **Identify the performance boundaries**
- **Extended system**
- **Applicability to other frameworks and infrastructures**

**PhD thesis**: Nov. 2015
Workload Analysis

**Question:** Which are the major MapReduce use cases?

- **Google, Facebook, Yahoo!, Cloudera, Microsoft**
  - Findings from 12 published production traces
  - Our analysis of other 4 production traces

- **Complex Workload**
  - Large variations in job submissions rates
  - 90% of the jobs in all traces process and generate less than 1 GB, and complete in under 1 minute
  - For large jobs, variations in job sizes vs. job durations
  - Our PDS group analyzes 15 TB of BitTorrent logs with MapReduce
Benchmarking Tool

• Real-world applications
  ➢ Text processing, web searching, machine learning

• Trace-based workloads
  ➢ Analysis and modeling of traces from production clusters

• BTWorld use case
  ➢ Complex MR workflow
  ➢ 14 Pig queries / 33 MR jobs
  ➢ Aggregations, selections, joins projections

![Makespan for different data sets](image)

Makespan [h]

- 5GB
- 10GB
- 50GB
- 100GB
Road Map

I. System prototype
- PhD start-up
- MTAGS 2012
  - Best Paper Award
- Dynamic resizing of an MR-cluster

II. Benchmarking Tool for MR
- Nov. 2011
- IEEE Big Data 2013
- Workload analysis

III. Performance model for MR
- Nov. 2012
- Nov. 2013
- Scheduling decisions based on the data and the workload
- Identify the performance boundaries

IV. Optimize the resource management for MR
- Nov. 2014
- Nov. 2015
- Extended system
- Applicability to other frameworks and infrastructures

PhD thesis
**Fair-Sharing Across Multiple Users**

*Question: Given multiple MR-clusters, how can you tell if one is working better than another?*

- Schedule and provision **concurrent** MR-clusters
- Differentiate users and converge to a division of resources such that they get **similar performance**

- **Weighted proportional allocations:**
  - Take **snapshots** in time of the queue sizes
  - Maintain a **history** of finished jobs
Provisioning Policies

Question: Can we obtain better performance with the dynamic MR-clusters?

• Data is hard to move
  ➢ Aprox. 3 h to transfer 1 TB between HDFS and the local storage (900 Mbps write speed)
  ➢ Removing a node with 100 GB makes ~ 6 failed jobs (1 Gbit/s Ethernet, avg. map task duration – 24 s, most jobs have less than 150 tasks)

• Explore a large space of policies:
  ➢ Policies for establishing the weights (fair-shares)
  ➢ Policies for growing (core or transient nodes, single or multiple clusters)
  ➢ Policies for shrinking (preemptive or non-preemptive)
Performance Model

*Question:* Which are the performance boundaries of the MR processing system?

- Analytical and statistical methods
- Metrics:
  - Fairness – users get similar performance
  - Elasticity – dynamic MR clusters
  - Performance isolation – multiple MR clusters
  - Velocity of data processing
  - Adaptivity to data explosion

http://research.spec.org/
Road Map

PhD start-up
Nov. 2011

I. System prototype
Nov. 2012

II. Benchmarking Tool for MR
Nov. 2013

III. Performance model for MR

IV. Optimize the resource management for MR
Nov. 2014

PhD thesis
Nov. 2015

- Dynamic resizing of an MR-cluster
- Workload analysis
- Fairness between multiple MR-clusters
- Identify the performance boundaries
- Scheduling decisions based on the data and the workload
- Applicability to other frameworks and infrastructures

- MTAGS 2012 Best Paper Award
- IEEE Big Data 2013
Optimize the MapReduce System

**Question:** Are the results obtained so far relevant for the large domain of data-processing systems?

- Provisioning policies with different optimization targets
- Incorporate knowledge about the workloads in scheduling and provisioning decisions
- Release the extended system with the full functionalities
- Investigate the applicability to other programming models and infrastructures
More Information

• Team: D. Epema, A. Iosup, M. Capotă, T. Hegeman, N. Yigitbasi, L. Fei,…

• PDS publication database
  ➢ www.pds.ewi.tudelft.nl/research-publications/publications

• Home pages
  ➢ www.pds.ewi.tudelft.nl/ghit
  ➢ www.pds.ewi.tudelft.nl/epema
  ➢ www.pds.ewi.tudelft.nl/~iosup

• Web sites:
  ➢ KOALA: www.st.ewi.tudelft.nl/koala