Towards Quantitative Analysis of Data Intensive Computing

A Case Study of Hadoop

Peng Wang, Dan Meng, Zhaoxia Han, Xu Liu*
Institute of Computing Technology, Chinese Academy of Sciences
*Rice University
Motivation

• Hadoop widely used in massive data processing

• Performance Diagnosis of MapReduce Applications:

➢ How can we evaluate MapReduce clusters?
Approach to Profiling

• Artemis [WASL ’08]
  Performance Debugging [ICDCS ’10]
    – Approach: Visualization

• Our Work: Quantitative approach
Quantitative Metrics

- **Data Locality Ratio**

- **Load Balance Coefficient**

- **Access Balance Coefficient**
  1. `SELECT COUNT(*) FROM student GROUP BY school`
  2. `SELECT AVG(age) FROM student GROUP BY school`

\[
DLR = \frac{\sum_{i=1}^{n} l_i}{\sum_{i=1}^{n} t_i}
\]

\[
GC = \frac{2\sum_{i=1}^{n} ix_i}{n\sum_{i=1}^{n} x_i} - \frac{n + 1}{n}
\]

\[
LBC = 1 - GC
\]

\[
ABC = -\sum_{i=1}^{n} \frac{f_i \log_2 f_i}{\log_2 n}
\]
Evaluation

• 270-node Hadoop production cluster at Tencent

• Application: Web Click Stream Analysis
  – More than 4 TB per day

• Prototype
  – Python Language
  – Parsing Hadoop Logs, analysis, and visualization
Results
Thanks!

Q&A