CIS 602: Provenance & Scientific Data Management

MapReduce Provenance and Provenance Analytics

Dr. David Koop
Reminders

• Start work on course projects now!

• Reading Responses:
  - Your goal is to **think** about the material the paper is presenting
  - **Understand** the ideas and show your understanding in the critique and questions
Today’s Class

- Two different topics:
  - MapReduce provenance
  - Provenance analytics

- MapReduce provenance
  - coarse versus fine-grained provenance
  - tracking dependencies
  - provenance storage

- Provenance analytics
  - Analyze provenance data for insight
  - What can be determined from provenance data?
Provenance for Generalized Map and Reduce Workflows

R. Ikeda, H. Park, and J. Widom

Presented by: Kevin Lydon
Related Work

• “Provenance for MapReduce-based Data-Intensive Workflows”, D. Crawl, J. Wang, and I. Altintas
• “Putting lipstick on pig: enabling database-style workflow provenance”, Y. Amsterdamer, S. B. Davidson, D. Deutch, T. Milo, J. Stoyanovich, and V. Tannen
MapReduce Provenance in Kepler

4. APPLICATIONS

4.1 Word Count

Word Count is the canonical example for the MapReduce programming model [5]. This application counts the number of text occurrences for each word read in Map, and outputs the total count. To achieve parallel processing, the input directory contains three files, and the Hadoop job produces a list of key-value pairs of each word and count. The reference database is required for each workflow execution.

![Provenance graph of Word Count workflow](image)

### Diagram Details
- **Map**: Reads input files and processes key-value pairs.
- **Reduce**: Aggregates results and outputs final counts.
- **SDF Director**: Coordinates dataflow and handles dependencies.

4.2 BLAST

BLAST [2] discovers the similarities between two biological sequences and is one of the most widely used algorithms in bioinformatics. Executing BLAST can be a data-intensive process since the query or reference data can have thousands of sequences.

![BLAST workflow](image)

### Diagram Details
- **Map**: Processes query or reference data.
- **Reduce**: Aggregates results and outputs final BLAST scores.
- **SDF Director**: Coordinates dataflow and handles dependencies.

[D. Crawl et al.]

In this section we present two applications that were built in our architecture. The Kepler Query API provides an interface to retrieve provenance information from workflows, which can be used to improve users' understanding of their data-intensive applications. Applications using the provenance data model used to store data. The Hadoop job is shown as a demonstration of how to query provenance data. The Kepler Query API can be used to query provenance data independent of the provenance model used to store data. This can be found in the visualization.
5.2 Provenance Queries

The Input Data row shows only the amount of data read as tokens in the Map tasks. These values do not include all the data read by workflow processes, e.g., the BLAST program reads the 20 GB reference database. The amount of captured provenance information for BLAST is roughly five times larger than the input for Word Count tasks. We were unable to get an accurate measure of the indexes’ communication times between NDB Data Servers takes the majority of the time. In the experiments with two slave nodes, MySQL Server must retrieve the data from different data servers. However, in the experiments with four and eight slave nodes, MySQL Server can answer any query without contacting other nodes since each node has the entire set of data. However, in the experiments with four and eight slave nodes, the query times appear to slightly decrease as the number of slave nodes increases. However, the query times for the Word Count workflow are the smallest for two nodes and greatest for four nodes. Since the amount of data retrieved during the BLAST program. This query takes less time than finding all the processes that ran the BLAST program. For these artifacts.

Table 1: Provenance information for each input data configuration.

<table>
<thead>
<tr>
<th>Input Data (MB)</th>
<th>Provenance Data (MB)</th>
<th>Artifacts</th>
<th>Processes</th>
<th>Dependencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>509</td>
<td>66,572</td>
<td>49,931</td>
<td>5,518,978</td>
</tr>
<tr>
<td>200</td>
<td>1,001</td>
<td>85,068</td>
<td>63,803</td>
<td>10,907,935</td>
</tr>
<tr>
<td>300</td>
<td>1,494</td>
<td>95,740</td>
<td>71,807</td>
<td>16,317,464</td>
</tr>
<tr>
<td>3.2</td>
<td></td>
<td>2.248</td>
<td>1.608</td>
<td>4.493</td>
</tr>
</tbody>
</table>

Table 1: Provenance information for each input data configuration.
Lipstick System

(a) Legend

(b) Coarse-grained provenance

(c) Fine-grained provenance

[Y. Amsterdamer et al.]
Lipstick System

- Based on Pig Latin language for Pig Hadoop system
- Use a graph with the tokens and semi-ring operations to communicate provenance
- plus is alternative use
- dot is joint use
- delta indicates duplicate elimination
MapReduce Provenance

- MapReduce poses significant challenges for storage, timing if we want to store all relationships between all intermediate inputs and outputs
- Lots of provenance data generated for “simple” MapReduce tasks
- Easy to describe the general map and reduce rules but tracking all of the provenance requires significant storage
- Would provenance storage/compression help here?
- Do advantages of storing full provenance outweigh the costs of doing so?
Examining Statistics of Workflow Evolution
Provenance: A First Study

L. Lins, D. Koop, E.W. Anderson, S. P. Callahan, E. Santos,
C. E. Scheidegger, J. Freire, and C. T. Silva

Presented by: Sivaraj Srinivas Busayavalasa
Provenance Analytics

• Event at Provenance Week 2014 on Provenance Analytics
  - Algorithms for provenance analysis and transformation
  - Machine learning techniques applied to provenance
  - Reasoning systems for provenance
  - Implementation, scalability, and performance of provenance analysis
  - Predictive models based on provenance
  - Applications of provenance analysis/analytics
  - Provenance based quality analysis, trust rating, reputation
  - Online and offline use of provenance
  - Provenance for auditing and accountability
  - Privacy issues pertaining to provenance
  - Reasoning with incomplete or uncertain provenance
  - Usability of provenance
  - Industrial experience with provenance
Provenance Analytics

  - CollabMap: online crowdsourcing mapping application
  - users rate each others’ contributions
  - determine trust for particular contributions based on ratings, extrapolate trust in new contributions based on provenance
  - also shows interest in a specific issue/region, interactions between authors
Provenance Analytics

• “Experiences in using provenance to optimize the parallel execution of scientific workflows steered by users”, M. Mattoso, J. Dias, F. Costa, D. de Oliveira, E. Ogasawara
  - Tune parallel execution engines based on provenance of past executions
  - Computational steering informed by provenance
  - Examples:
    • Lanczos algorithm convergence
    • Modify long-running workflows while running
    • Uncertainty Quantification
Provenance Analytics

- “Application of PROV Model for Modeling a VM Overload Mitigating Strategy: Task Eviction”, A. Albatli, L. Lau, J. Xu
  - Keep track of cloud execution provenance (resource usage, etc.) and use this to inform eviction strategies

- “Provenance for Online Decision Making”, A. S. Keshavarz, T. D. Huynh, and L. Moreau
  - Use provenance to make decisions
Provenance Analytics

- Data analytics is a big concern in academic, industry, and government
- Provenance is data about how data is generated and used
- Analyzing provenance of data should also inform our understanding of the data itself
Reminders

• Start work on course projects now!
• Reading Responses:
  - Your goal is to think about the material the paper is presenting
  - Understand the ideas and show your understanding in the critique and questions