Graphulo: Graph Processing for Accumulo Databases

Dylan Hutchison[1,2,3]   Jeremy Kepner[1,2,4]   Vijay Gadepally[1,2]

[1]MIT Lincoln Laboratory   [2]MIT Computer Science & AI Laboratory   [3]University of Washington   [4]MIT Mathematics Department

Apache Accumulo

- Distributed key/value store in Hadoop HDFS Java ecosystem
- Best for large (TBs to PBs), de-normalized tables (NoSQL)
- Cell-level visibility access control; row store by default
- Performance record at 115M entries inserted/second[3]
- Implements stored procedures via a Java iterator framework

Graphulo

- Library for Accumulo to orchestrate server-side graph processing[2]
- Implements GraphBLAS.org standard sparse matrix operations
- Graphulo's clients use its core matrix math primitives to build algorithms
- Open Source: http://graphulo.mit.edu

> Graphulo enables diverse **graph algorithms** atop the **GraphBLAS** primitives on many **graph schemas** in **Accumulo**

Graphulo Stored Procedure Calls

↓ : "is specialization of"
parent
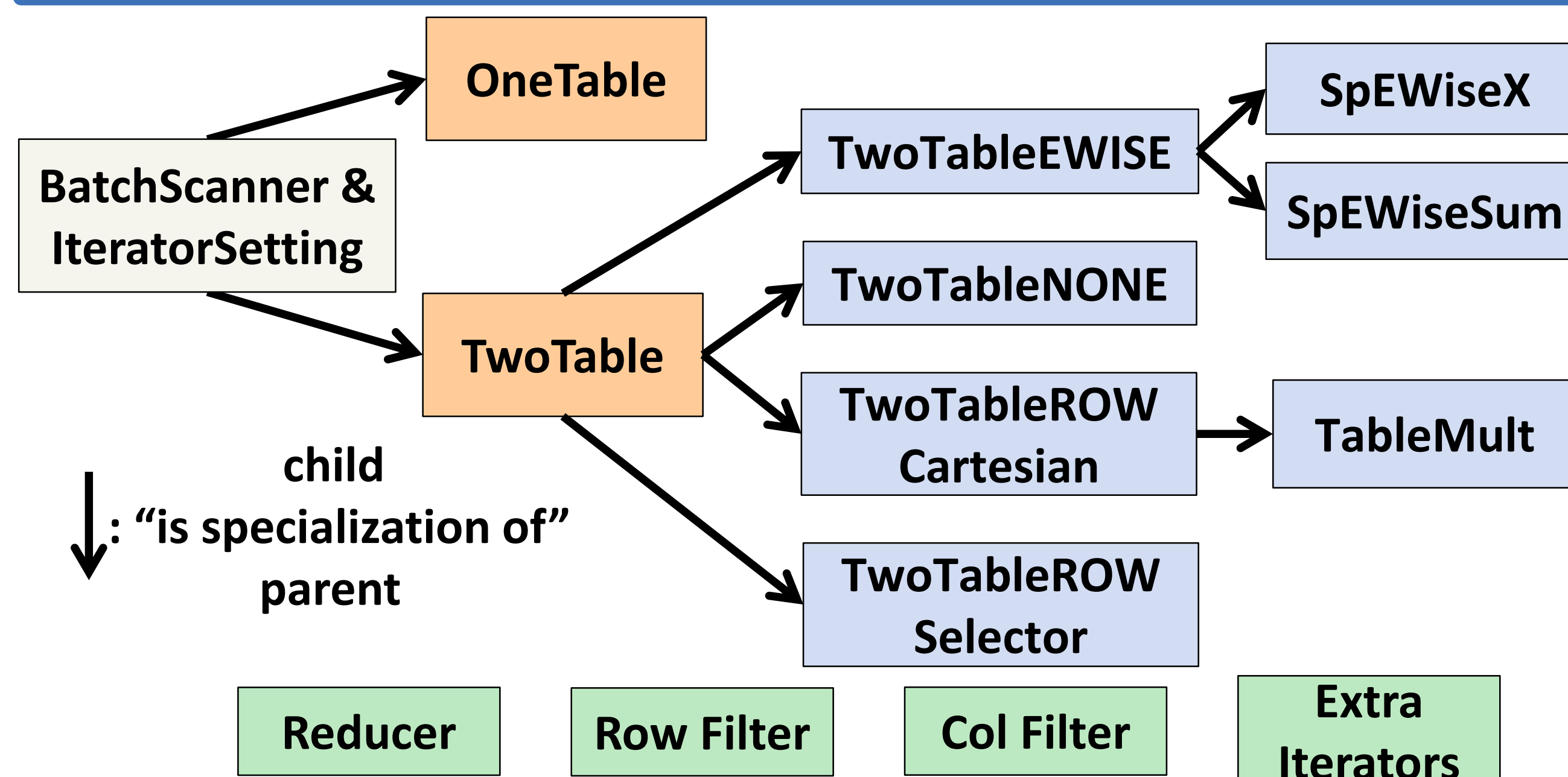
Performance

- Single-node matrix multiply experiment: Graphulo vs. multiply outside Accumulo in Matlab using D4M
- Graphulo outer product alg. scales with Accumulo[1]
- Matlab network transfer cost is bottleneck
- Future work is multi-node scale testing Graphulo

References

1. D. Hutchison, J. Kepner, V. Gadepally, and A. Fuchs, "Graphulo implementation of server-side sparse matrix multiply in the Accumulo database," in IEEE HPEC, 2015.
2. V. Gadepally, J. Bolewski, D. Hook, D. Hutchison, B. Miller, and J. Kepner, "Graphulo: Linear algebra graph kernels for NoSQL databases," in IEEE IPDPSW, May 2015.
3. J. Kepner, W. Arcand, D. Bestor, B. Bergeron, C. Byun, V. Gadepally, M. Hubbell et al., "Achieving 100,000,000 database inserts per second using Accumulo and D4M," in IEEE HPEC, 2014.