From Graphs to Tables the Design of Scalable Systems for Graph Analytics

Joseph E. Gonzalez
UC Berkeley, AMPLab
jegonzal@eecs.berkeley.edu

ABSTRACT

From social networks to language modeling, the growing scale and importance of graph data has driven the development of numerous new graph-parallel systems (e.g., Giraph and GraphLab). By restricting the types of computation that can be expressed and by introducing new techniques to partition and distribute graphs, these systems can efficiently execute sophisticated graph algorithms orders of magnitude faster than more general data-parallel systems.

However, the same restrictions that enable graph-parallel systems to achieve substantial performance gains also limit their ability to express many of the important stages in a typical graph-analytics pipeline. Moreover, while graph-parallel systems are optimized for iterative diffusion algorithms like PageRank they are not well suited for more basic tasks like constructing the graph, modifying its structure, or expressing computation that spans multiple graphs. While existing systems address specific stages of a typical graph-analytics pipeline, they do not address the entire pipeline, forcing the user to deal with multiple systems, complex and brittle file interfaces, and inefficient data-movement and duplication.

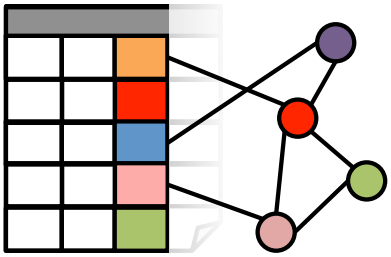
To fill the need for a holistic approach to graph-analytics we introduce GraphX, which unifies graph-parallel and data-parallel computation under a single API and system. GraphX recasts advances in graph-processing in the context of relational algebra and distributed join optimization enabling more general data-parallel systems to process graphs efficiently. We evaluate the GraphX system on several real-world tasks and show that its end-to-end performance can exceed that of specialized systems.

This talk describes recent work at the UC Berkeley AMPLab in collaboration with Reynold S. Xin, Ankur Dave, Daniel Crankshaw, Michael J. Franklin, and Ion Stoica.

Categories and Subject Descriptors

H.2.4 [**Information Systems**]: DATABASE MANAGEMENT

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04. .



1. CONTENT

In this talk, we review the graph-parallel abstraction and describe how it can be used to express important machine learning and graph analytics algorithms like PageRank and latent factor models. We describe how systems like GraphLab and Pregel exploit restrictions in the graph-parallel abstraction along with advances in distributed graph representation to efficiently execute iterative graph algorithms orders of magnitude faster than more general data-parallel systems. We then discuss the limits of the graph-parallel abstraction in the context of an end-to-end graph-analytics pipeline.

Motivated by these challenges we introduce GraphX and describe how it unifies graph-parallel and data-parallel computation. We begin by reinterpreting vertex-partitioning and the graph-parallel API through the lens of distributed joins processing and show how this view enables GraphX to achieve performance comparable to specialized graph processing systems while exposing a more flexible API. Moreover, we show how a simple set of GraphX operators can be used to express graph-parallel computation and how, by applying a collection of query optimizations derived from our work on graph-parallel systems, we can execute entire graph-analytics pipelines efficiently in a more general data-parallel distributed fault-tolerant system. Finally, we present the results of our performance analysis of the GraphX system when compared against existing tools for graph analytics.

Short Biography

Joseph Gonzalez is cofounder of GraphLab Inc. and a postdoc in the AMPLab at UC Berkeley. Joseph received his PhD from the Machine Learning Department at Carnegie Mellon University where he worked with Carlos Guestrin on parallel algorithms and abstractions for scalable probabilistic machine learning. Joseph is a recipient of the AT&T Labs Graduate Fellowship and the NSF Graduate Research Fellowship.