GraphX: A Resilient Distributed Graph System on Spark

Kartikeya Upasani (kuu2101)

Paper Review, 28 November 2016

1 Motivation

Several computational algorithms involve graphs which are difficult to perform in parallel on a multi-core or distributed architecture. However, performing them in parallel is important for several common applications such as social-networks and targeted advertising. Moreover, existing techniques do not offer parallelism in graph construction and transformation, and cannot be coupled with each other easily.

2 Goal

Developing a interactive fault-torelant graph engine on top of Spark (Resilient Distributed Graphs), and implements popular operations like those of PowerGraph and Pregel.

3 Key Idea

Partitioning graph using vertex-cut and building on top of Spark's RDDs to provide common graph operations.

4 Approach

GraphX extends Spark's RDDs to RDGs by allowing certain immutable transformations such as $vertices()$, $edges()$, $filterVertices(predicate)$, etc. that return a new graph. Such graph-parallel computation adopts a vertex centric view of data computation, and requires each vertex or edge to be processed in the context of its neighborhood. Vertex-cuts are used over edge-cuts for dividing the graph for parallel computation as they have been proven to be efficient in terms of storage and communication. This is achieved by using three unordered horizontally partitioned tables implemented as Spark RDDs- EdgeTable, VertexDataTable, and VertexMap. Pregel and Powergraph features are implemented using this architecture.

5 Results

GraphX was compared against Apache Mahout and PowerGraph in experiments on the PageRank algorithm. It proved to be 8x faster than Mahout but 7x slower than PowerGraph, which is a dedicated graphing engine.

6 Conclusion

GraphX is a fast general purpose distributed computing engine that allows effortless implementation of common PowerGraph and Pregel functions, as an advantage over other general purpose engines like Mahout, or over Spark that does not have these features.