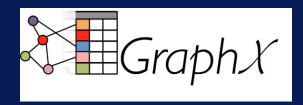
Spark GraphX



After this video you will be able to...

- Describe what GraphX is
- Explain how Vertices and Edges are stored
- Describe how Pregel works at a high level



MLlib G

GraphX

Spark GraphX

Spark Core

GraphX is Apache Spark's API for graphs and graph-parallel computation.

GraphX uses a property graph model.

Both Nodes and Edges can have attributes and values

Properties → Tables

Vertex Table

Node properties

Edge Table

Edge properties

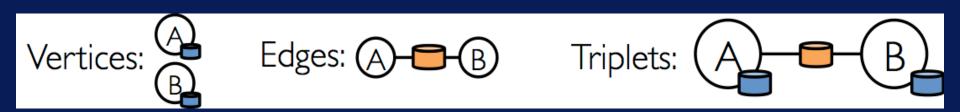
GraphX uses special RDDs

VertexRDD[A] extends RDD[(VertexID, A)]

EdgeRDD[ED, VD] extends RDD[Edge[ED]]

Triplets

The triplet view logically joins the vertex and edge properties.



<u>http://spark.apache.org/docs/latest/img/triplet.png</u>

Pregel API

Bulk-synchronous parallel messaging mechanism

Constrained to the topology of the graph

GraphX

Graph Parallel Computations

Special RDDs for storing Vertex and Edge information

Pregel operator works in a series of super steps