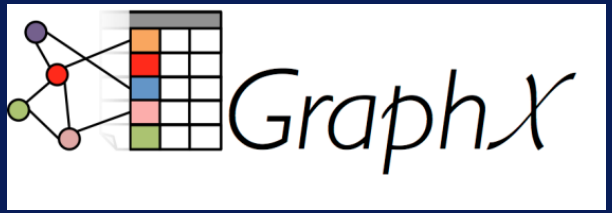
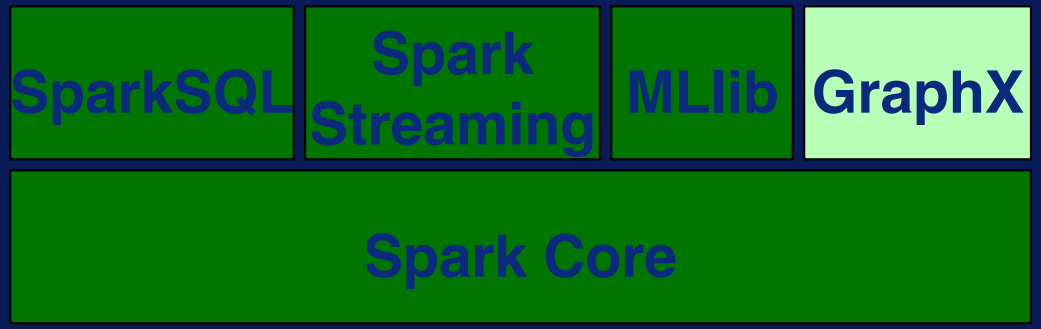
Spark GraphX



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



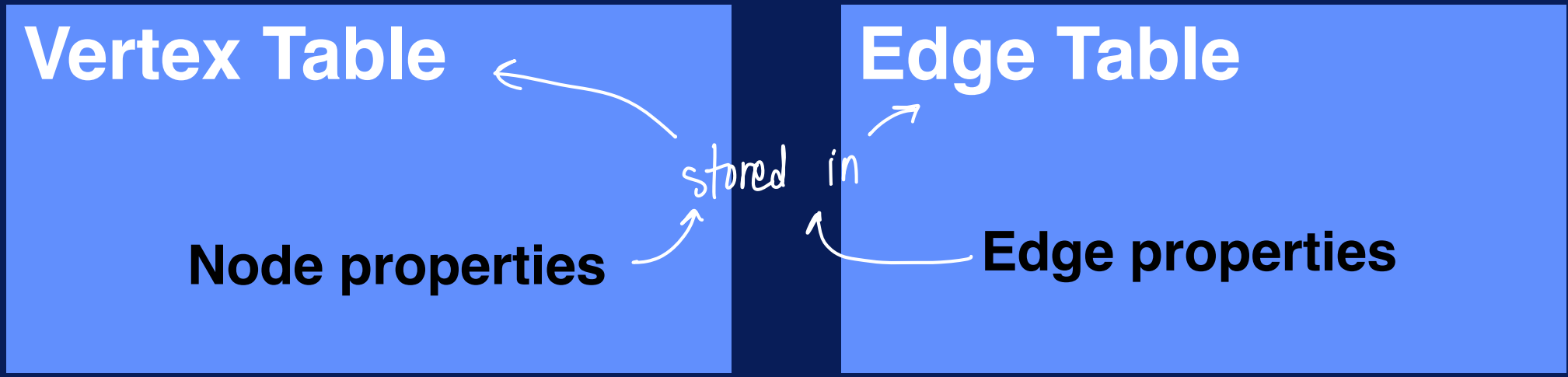
Spark GraphX

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



connectivity information (wich edge connects which nodes) is stored separately from the node and edge properties

GraphX uses special RDDs

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

— represents a set of vertices all of which have an attribute called A

* defined to be unique by design

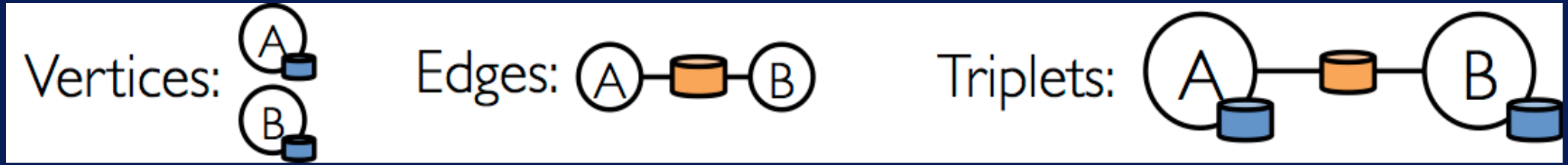
EdgeRDD[ED, VD] extends RDD[Edge[ED]]

— extends this basic edge storing by the edges in columnar format on each partition for performance

* object with a source vertex and destination vertex and edge attribute

+ Triplets (in addition to vertex and edge views)

The triplet view logically joins the vertex and edge properties.



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

Pregel API

- operator that can execute operators from pregel library for graph analytics
- executes in a series of super steps which defines a messaging protocol for vertices

Bulk-synchronous parallel messaging mechanism

Constrained to the topology of the graph

GraphX

spark can be used for:

Graph Parallel Computations

graphx uses

Special RDDs for storing Vertex and Edge information

Pregel operator works in a series of super steps