Lesson 8

Apache® Spark™ GraphX

.

IBM System G

- Offers a set of Big Data tools for graph computations. G (stands for graph, which may be a property graph, Bayesian network graph, or cognitive network graph)
- Graph may be static or dynamic, small or large, topological or semantic.

IBM System G

- G has library functions for graph analytics
- G applications include creating and analyzing the database, visualization and middleware for graph

Graph Analytics

- Compute degree centralities, degree distribution, separation of degrees, betweenness centralities, closeness centralities, neighbourhoods, PageRank, shortest path, Breadth First Search, minimum spanning tree (forest), connected components, spectral clustering and cluster coefficient.

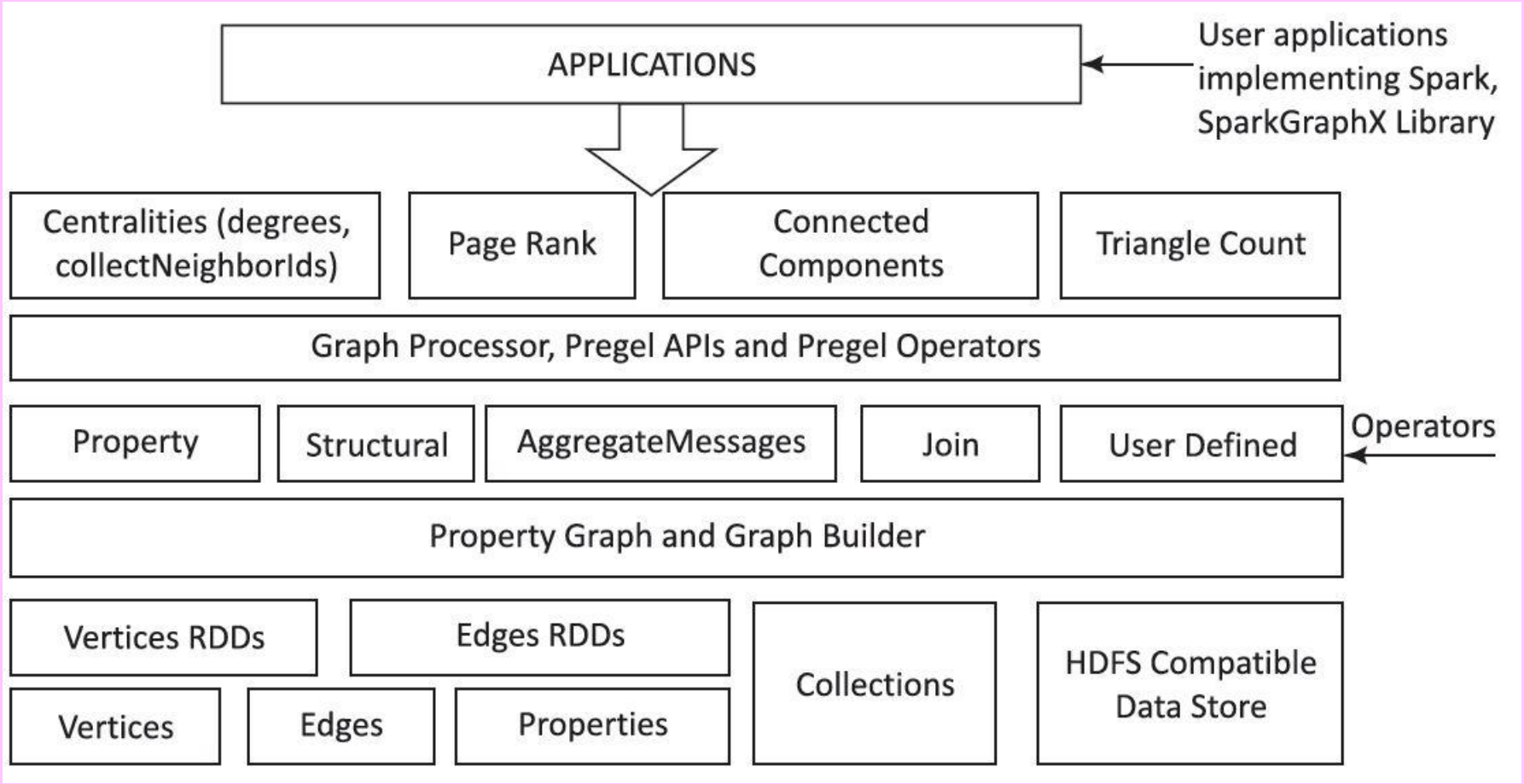
Apache Spark GraphX

- An open source software that provisions a number of functions and operators for graph stored in HDFS environment
- Apache Spark refers to a multi-component platform for Big Data computing that uses data store at a HDFS file system, HDFS compatible data sources, such as HBase, Cassandra, Ceph or S3. .

Spark GraphX

- Provides a set of fundamental operators such as subgraph, joinVertices and aggregateMessages
- Provides for computations using the property graphs
- Identifier in GraphX is 16-bit long unique-key. Edges have vertexIDs for corresponding source-destination paths.

Figure 8.8 GraphX Architecture



GraphX Operators

1. **Aggregation operator—** used in computing the shortest path to a source, smallest reachable vertex id, connected components and PageRank
2. Pregel operator—executes in a series of supersteps in which vertices receive the sum of their inbound messages from the previous superstep

Pregel Operator

- GraphX Pregel-API provides computation of messages in parallel as a function of the edge triplet. A message for computation provide accesses to both attributes of source and destination vertices.
- Pregel operator in GraphX is a bulk-synchronous parallel-messaging abstraction

GraphX Operators

3. ConnectedComponents algorithm— labels each connected component of the graph with the ID. GraphX contains an implementation of the algorithm for the ConnectedComponentsObject.
4. PageRank─ The function measures the importance of each vertex in a graph.

graph.connectedComponents().vertices

1. Degree Computation Objects: Functions graph.inDegrees.reduce(max); graph. outDegrees.reduce(max); graph.degrees.reduce(max); for
- analyzing the degree distribution also at the vertices.

graph.connectedComponents().vertices

2. Collection neighbour Ids and neighbours Operators: The functions collectNeighborIds (ed geDirection: EdgeDirection); collectNeighbors (edgeDirection: EdgeDirection)

graph.connectedComponents().vertices

3. Triangle Count Algorithm—Determines the number of triangles passing through each vertex. The count is a measure of clustering.

Property Operators

- Property Operators in Class Graph [VD, ED] are mapVertices [VD2](), mapEdges [ED2](), mapTriplets [ED2]()}. These functions yield a new Graph.

Structural operators in Class Graph [VD, ED]

- Use subgraph(), mask() and groupEdges(). Function reverse( ) returns a new graph with all directions reversed

Join operators

- Joins the data from external collections (RDDs) with the graphs
- Merges extra properties with an existing graph
- Main join functions: joinVertices and outerJoinVertices().

Page Rank Analytics

- GraphX provides static and dynamic implementations of PageRank as methods of the PageRank object: ranksByUsername =users.join(ranks).map {case (id, (username, rank)) => (username, rank)}

Summary

We learnt:

- Spark GraphX
- Pregel Operator
- Connected Components
- Aggregation Functions
- Join, Property and Structural operators
- Page Rank Analysis

End of Lesson 8 on

Apache® Spark™ GraphX