Efficient Application of Complex Graph Analytics on Very Large Real World RDF Datasets

Zhe Wu, Jay Banerjee

{alan.wu, jayanta.banerjee}@oracle.com

Oracle USA

RDF [1] Graph modeling is a foundational technology in the whole semantic web (SW) technology stack. Since its debut in 2004, RDF graph has enjoyed many applications in the enterprise domain. Examples of these applications include, but certainly not limited to, integration and federated query of heterogeneous data sources, flexible and extensible representation of enterprise knowledge base, ad-hoc query and navigation on top of schema-less graph model of enterprise data, social network representation and link analysis, and metadata processing in the context of master data management (MDM). In the past decade, many mature open source and commercial RDF platforms and solutions [6] have been developed to store and index RDF graph data (triples and quads), edit and manage OWL [2] ontologies, perform logical inference, execute pattern matching and graph navigation (SPARQL [3]), visualize RDF graph data and OWL ontologies, and link data in RDF format and also other data types including relational (RDB2RDF [4,5]). As a graph modeling language, RDF provides great flexibility for enterprise applications and it adds precision, through the use of URI and formal semantics, to enterprise data. SPARQL query and OWL inference have been two key functions for semantic web applications. A somewhat less obvious application of RDF is that such a graph model is also a great candidate for graph analytics.

Large-scale graph analytics [7-10] (page ranking, community detection, etc) for enterprise applications have many challenges, including efficient and scalable graph data management, high-performance implementation of graph algorithms, user- and operation-friendly management interfaces, and tight integration with high quality tools. To effectively address such challenges, we propose a comprehensive graph analytics architecture built upon the Oracle platform. Key components of

this platform include Oracle Database 12c, SQL-based graph analytics, parallel in-memory graph processing engine, and the RDF Semantic Graph capabilities. We show why complex graph-based analytics matter for large enterprise-scale RDF datasets, and we share our experiences in implementing several graph analytical functions, such as page ranking, clustering, path analysis, and so on. We apply and evaluate our implementation on several large-scale real world RDF graph datasets, including graphs from social networks, social media domain and the linked data domain. We make best practice recommendations based on our experiences.

References

[1] http://www.w3.org/TR/rdf11-concepts/

[2] http://www.w3.org/TR/owl2-syntax/

[3] http://www.w3.org/TR/rdf-sparql-query/

[4] http://www.w3.org/TR/rdb-direct-mapping/

[5] http://www.w3.org/TR/r2rml/

[6] http://www.w3.org/wiki/LargeTripleStores

[7] http://graphlab.org/projects/index.html

[8] http://www.oracle.com/technetwork/oracle-labs/parallel-graph-analytics/overview/index.html

[9] http://spark.apache.org/graphx/

[10] http://giraph.apache.org/