Parallel Processing of RDF Data in a Distributed Environment

Tigran Shahinyan

Institute of Informatics and Automation Problems
Yerevan, Armenia
e-mail: tigran.shahinyan@gmail.com

ABSTRACT
Efficient processing of huge amounts of data has become one of major challenges during recent years. Two of the main approaches are parallel relational databases and non-relational data processing systems. Another ecosystem of promising technologies is Semantic Web, which provides theoretic and technological basis for distributed knowledge representation and reasoning. The paper represents an approach of using MapReduce paradigm for processing of RDF data.

Keywords
Distributed computing, MapReduce, Semantic Web

1. INTRODUCTION
Huge amounts of data and the necessity of its efficient processing using the computing power of the Internet, computer grids and cloud system caused the emergence of non-relational data processing paradigms and technologies. MapReduce is a paradigm introduced by Google and implemented by several vendors [1]. One of the most widespread implementations of the MapReduce paradigm is the open source Apache Hadoop [2].

2. DISTRIBUTED DATA PROCESSING APPROACHES
There are two main approaches for distributed data storage and processing. Parallel relational database systems seek to improve performance through parallelization of various operations, such as loading data, building indices and evaluating queries [3].
Parallel databases are based on one of the following architectures:

- Shared memory architecture, where multiple processors share the main memory space, as well as mass storage (e.g. hard disk drives).
- Shared disk architecture, where each node has its own main memory, but all nodes share mass storage, usually a storage area network. In practice, each node usually also has multiple processors.
- Shared nothing architecture, where each node has its own mass storage as well as main memory [4].

Parallel relational database systems show high performance especially in homogeneous environments [5].
The most important advantages of MapReduce are fault tolerance and the ability to operate in heterogeneous environments [6].

3. DATA REPRESENTATION IN SEMANTIC WEB
The Resource Description Format (RDF) and OWL (Web Ontology Language) are used to represent information resources modeled as a directed labeled graphs, where edges represent the named link between two resources, represented by the graph nodes [7].
RDF uses URI references to identify resources and properties.
RDF graphs can be read and written by using the Jena software package, and queried using the SPARQL query language [8].
Semantic web is penetrating many areas of information processing and exchange activity. An example is UniProt, and effort to create a comprehensive catalog of protein data in RDF [9].

4. SEMANTIC WEB DATA PROCESSING
We have used Hadoop as an implementation of MapReduce to process protein data from UniProt catalog. The data was stored in HDFS distributed file system in line-based NTriples. For processing the data we have used Jena with its Elephas library, which provides Hadoop InputFormat and OutputFormat implementations for RDF. It covers all RDF serializations that Jena supports and extensions by custom formats.
Elephas splits and parallelizes processing of input where the RDF serialization allows it.
We have used and extended various reusable basic Mapper and Reducer implementations covering the following common tasks: counting, filtering, grouping, splitting, transformation.

REFERENCES
[1] Jeffrey Dean, Sanjay Ghemawat, MapReduce: A Flexible Data Processing Tool, Communications of the ACM, Volume 53 Issue 1, January 2010
[2] Apache Hadoop, https://hadoop.apache.org/
[3] T. Shahinyan, A comparison of Different Approaches of Distributed Data Processing, *CSIT 2013*, Yerevan, Armenia
[4] David DeWitt, Jim Gray, Parallel database systems: the future of high performance database systems, Communications of the ACM, Volume 35 Issue 6, June 1992
[5] A. Pavlo, A. Rasin, S. Madden, M. Stonebraker, D. DeWitt, E. Paulson, L. Shrinivas, and D. J. Abadi. A Comparison of Approaches to Large Scale Data Analysis. In Proc. Of SIGMOD, 2009.
[6] A. Abouzeid, K. Bajda-Pawlikowski, D. Abadi,A. Silberschatz, A. Rasin, HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads, VLDB 2009
[7]W3C Semantic Web Acivity, http://www.w3.org/2001/sw/
[8] Michael Grobe, RDF, Jena, SparQL and the "Semantic Web" SIGUCCS'09, October 11–14, 2009, St. Louis, Missouri, USA.
[9] Universal Protein Resource. http://www.uniprot.org/.