# Performance Analysis of Semantic Web Environment Based on MapReduce

**Vrushali T. Lanjewar** **Dr. V. M. Thakare**

***Abstract*** **— The Semantic Web is a framework that allows sharing and reusability of data throughout the group of interacting people, enterprise, and application. With growth of semantic data, forthcoming surge of ontology arises many challenges in performing efficient reasoning. This paper discusses the issues related to process large ontology to improve performance of semantic data and use of MapReduce programming model in enhancing performance of web semantic environment. This paper proposes performance evaluation method on large scale ontologies by using MapReduce, which realizes runtime searching, reasoning for knowledge base. Finally, a prototype system is proposed on a Hadoop framework and discusses expected results to validate the usability and effectiveness of the proposed approach.**

***Key Words*** **— HDFS, MapReduce, ontology reasoning, RDF, Semantic Web**

## I. INTRODUCTION

Semantic Web is a network, which contains some parts of the document and the document describes the significant relationship between things, including the semantic information in order to facilitate the automatic processing machine. The main goal of semantic web is to extend the principles of web from the document to data. In order to achieve the goal described above, the most important is to be able to define and describe the relations among data (resources) on the Web. Semantic web applications found in wide range such as knowledge representation and analysis, data integration, cataloging services, improving search algorithms.

Resource Description Framework (RDF) is the fundamental building block of semantic web that defines relations. RDF is used to describe relationship between subject and object. RDF is basic representation of ontology. RDF is XML based representation for metadata defined by W3C. RDF is a tool which gives finer more detailed classification and characterization of those relationships as well as resources being characterized. The basic RDF data model is a triple, and the basic object types are the following: resources, properties, statements, which are also known as declaration. Schema definition language (RDFS) defines a new vocabulary that includes typing, inheritance of classes, properties. SPARQL is a syntactically-SQL-like language for querying RDF graphs via pattern matching. The language features includes basic conjunctive patterns, value filters, optional patterns, and pattern disjunction.

Hadoop framework is an open source Java implementation of MapReduce that allows for the distributed processing of large datasets across clusters of computers via simple programming models. It can scale up from single server to thousands of machines, each offering local computation and storage, and manages execution details such as data transfer, job scheduling, and error management. MapReduce is a massively scalable, parallel processing framework that works with HDFS. With MapReduce and Hadoop, computation is executed at the location of the data, rather than moving data to the compute location; data storage and computation coexist on the same physical nodes in the cluster.

The Hadoop Distributed File System (HDFS) is a distributed file system that is designed for the Hadoop MapReduce framework and has master-slave architecture. Its cluster consists of a single NameNode and a master server that manages the namespace of file system and regulates the access of clients. In addition, there are a number of DataNodes (usually one by node) that manages the storage nodes.

## II. BACKGROUND

There are various parameters that can be modified in semantic web considering various factors such as complex query processing time, performing efficient reasoning, semantic inference performance on the basis of accuracy and consistency in heterogeneous concepts sets used in different information systems. Due to the deluge of large semantic data, fast growth of ontology challenges occurs in performing scalable reasoning. Thus distributed methods of reasoning are required to improve scalability and performance of inferences. Most of developed methods are tested and evaluated on hundreds of documented not used in real applications such as web so an semantic annotation method are developed.

## III. PREVIOUS WORK DONE

Bo Liu et al. [1], presented a method to speed up the updating process with newly-arrived data and fulfill the requirements of end-users for online queries. Incremental Distributed Inference Method (IDIM) is based on MapReduce and Hadoop, which can well influence the old and new data to minimize the updating time and reduce the reasoning time when facing big RDF datasets to speed up the updating process with newly-arrived data and fulfill the requirements of end-users for online queries. A representation method TIF/EAT to support incremental inference over large-scale RDF datasets which can efficiently reduce the

storage requirement and simplify the reasoning process. A real-world application on healthcare domain is presented. IDIM performs reasoning on this medical ontology. Meanwhile, users can execute their query more efficiently without computing and searching over the entire RDF closure used in the prior work.

Yujiang Liu et al. [2], Algorithm is to study the use of RDF as the description data to set out rules and to learn. How to use the time during the excavation of the real knowledge to discover is more interesting rules, which can show the use of ILP methods into the feasibility on mining the Semantic Web. It is an important guiding role on the full sense future semantic Web mining.

Maria Jose Ibanez et al. [3], proposed the scalability of the algorithm for computing the RG has been evaluated. The critical part of the algorithm is the implementation of the function checking the equivalence of two states (net markings). The computing time is considerably high. Main aim was to evaluate the feasibility of computing the RG of an input system (clearly, a much more efficient version of the prototype tool need to be implemented in future). This paper presents U-RDF-PN systems as a high-level formalism for the modeling of SBPs with a well-defined semantics, which allow the set of system states and state transitions to be generated and used as the inputs of the model checker. In order to prove the feasibility of the proposed approach, a fully functional tool for model checking has been implemented.

Jingzhi Guo et al. [4], have been discussed a semantic inference problem that requires reasoning between heterogeneous e-marketplace activities. It replaces domain-wide ontology by ConexNet concepts, which are collaboratively created between heterogeneous domains. It has introduced a concept separation strategy to separate an activity into concept denotation, concept connotation, and concept implementation. With this separation, any heterogeneous activity is interoperable utilizing ConexNet, which is related to the work, researched in maintaining semantic consistency between heterogeneous concepts. To implement this strategy, a RuleXPM schema has been designed for governing the message handling using defeasible logic, SWRL, and ConexNet concept and a semantic inference engine has been developed for deriving a next activity for the intended recipient of EMpNet.

Jianling Sun et al [5]; Instead of implementing another distributed system adopted the well-known open source project: HBase. Also, proposed a MapReduce query processing algorithm against the storage schema. They analyze the common problems for processing SPARQL queries with MapReduce and propose our solution to bypass the limitations.

## IV. EXISTING METHODOLOGY

1. Semantic web architecture provides solution for various problems of existing search methodology. An incremental and high performance inference engine provides scalability and efficient reasoning. A representation method TIF/EAT to support incremental inference over large-scale RDF datasets which can efficiently reduce the storage requirement and simplify the reasoning process [1].

2. Search algorithms for resource discovery is proposed in semantic web usage in data mining. It can reduce the computational algorithms and improve the accuracy, which can take advantage of RDF data clustering to learn [2].

3. SBP analysis is an emerging research field. A fully functional tool for model checking is implemented in semantic business process. As the experimental results proved, the real bottleneck corresponds to the computation of the RG. Semantic inference problem between heterogeneous e-marketplaces handled to achieve maximum accuracy providing semantic consistency by using semantic inference engine [3].

4. The RuleXPM method, targets at achieving 100% accuracy for semantic interpretation across domains/ contexts. To implement this strategy, a RuleXPM schema has been designed for governing the message handling using defeasible logic, SWRL, and ConexNet concept and a semantic inference engine has been developed for deriving a next activity for the intended recipient of EMpNet. In this engine, a generic RIA has been introduced, which guarantees the correct semantic inference. The correctness of the approach is demonstrated in a prototype where experiments are made to test the performance [4].

5. Another approach proposes a semantic environment, combining semantic resources, machine learning and Hadoop tuning concepts. The approach is based on the Hadoop Ontology and subdivided into three main modules: preprocessing, the parameter tuning generator and search. Being a semantic environment, it is proposed that all knowledge representation generated by collaborative tools and by the insertion of external knowledge should be organized in ontologies[5].

## V. ANALYSIS AND DISCUSSION

### *A. Analysis of Existing methods:*

1. IDIM to deal with large-scale incremental RDF datasets to our best knowledge and IDIM uses a different representation method.

2. Resource discovery is a search algorithm to classify the learning resources and data pre-processing.

3. High-level model checking as an analysis tool; Linear Temporal Logic based model checking used to analyze logs in SBP. The implementation of model checker is based on an adaptation of the labeling algorithm proposed in as the inputs are an RDF-KS and an RDF CTL formula.

4. Semantic Inference Engine can work in EMpNet for heterogeneous e-marketplace activity inference. In this model, the execution of concept match in each recipient system is strictly sequential starting from the concept that is going to match.

5. MapReduce join algorithm for SPARQL, BGP processing with evaluation results is proposed. RDF storage schema is presented on HBase. Jena inference engine is used to obtain inference results from generated LUBM datasets.

*B. Attribute and Parameter Considered:*

To evaluate the performance of semantic web environments various parameters are considered are discussed below:

In order to show the performance of IDIM method, the reasoning time is recorded when each part is input one-by-one. For IDIM, the output triples are the ones in TIF and EAT, and the time for generating TIF/EAT is recorded. The use of Semantic Web Mining on the Web ontology will greatly improve the results of Web mining as well as the efficiency. High level Model checking, performance is evaluated through the influence of the number of RDF triples on the time for verifying a formula. In this engine, a generic RIA has been introduced, which guarantees the correct semantic inference. The correctness of the approach is demonstrated in a prototype where experiments are made to test the performance. Semantic inference engine provides concept separation strategy to solve the heterogeneous activity inference problem, which is extremely useful for heterogeneous business process integration and interoperation. MapReduce join algorithm for SPARQL**,** supports for parallel processing. Large volume RDF dataset usually has too many literals, which may lead to an extremely large mapping table. Thus, scalability through HBase and MapReduce on RDF store is achieved.

*C. Effect of Outcome of Various Attributes and Parameter:*

1. Hadoop configurations are identical to that in IDIM. Two methods run on each dataset three times and then calculate the number of the output triples and the time needed for the reasoning. Hadoop cluster with eight computing nodes is also configured for distributed computing.

2. From the analysis point of view; the limitation on the Web environment of the existing data mining method is studied by recovery algorithm.

3. In high level model checking technique parameter improved are computation time and query processing time. On the other hand, the efficiency of the prototype must be improved in order the tool could deal with more complex systems.

4. RuleXPM method, suggested in this paper, targets at achieving 100% accuracy for semantic interpretation across domains/contexts.

5. MapReduce join algorithm for SPARQL, reduced 110 cost. For a particular triple pattern, data can be loaded from a single row of the corresponding table. HBase stores data of the same.

*D. Comparison and Drawback:*

1. In order to show the performance of the method, comparing IDIM with WebPIE, this is the state of the art for RDF reasoning. Further, to compare the performance when the input data are incremental, the whole dataset is divided into four parts and input them into the system gradually.
2. RDFMS hierarchical clustering method based on semantic distance data is proposed with the analysis of semantic level on RDFMS resource description.
3. Comparing with other research approaches, high level model checking process show control flow structure of ordinary petri nets by showing how complex structures are supported. RDF checker model studies execution correctness and behavioral properties of business process (considering data flow aspects).
4. In semantic inference engine, the accuracy problem of ontology interpretation across domains/contexts is actually not solved. The accurate interpretation is still a problem for using heterogeneous ontologies, an e- marketplace. The drawback is its higher complexity of the design of the inference engine and the implementation of rules.
5. For large-scale RDF data based on HBase have been studied. The main disadvantage in the method is that it requires more storage space. There are several copies data which locate in different tables. But reduce execution steps during query processing.

## VI. PROPOSED METHODOLOGY

We proposed an ontology-based semantic inference using MapReduce to improve performance of information extraction in semantic web. An information extraction system is used to extract annotating documents for recognizing mapping of words to existing domain concepts. An incremental reasoning method is used on large growing ontology based on MapReduce architecture. It deals with the reasoning problem in case of large ontology. As MapReduce is build on the idea to use large scale data such as cluster, cloud environment so it is suitable for large RDF datasets in which relationship between new arrived data and existing data are considered. Recompilation will take place when every time new data arrives in RDF dataset, to avoid such time consumption incremental reasoning is proposed.

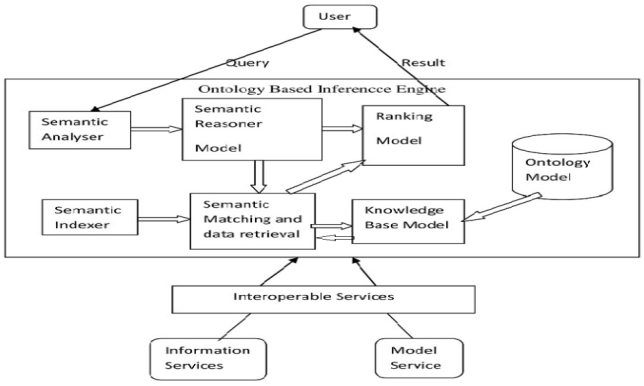


**Fig:** Block diagram of Ontology Based Semantic Inference Engine

Query submits and get results are transparently occurred when user accesses a particular web service.

The fig. shows the basic flow of inputted data to obtain the knowledge. This engine includes and ontology supported knowledge based model, semantic web enabled resources, semantic reasoner model, syntax analyser and ranking model based on matrix can be developed. The syntax analyzer will be able to parse user queries and reformulate them as semantic queries to the reasoned. The semantic reasoner will support detailed inference based on knowledge base models. The results from the reasoner will be mapped to different instances of data and services.

Ontology and semantic annotations work can be explained in incremental reasoning in number of steps:

1. When User logs in particular sites with web browser [Web browser with annotations].Html page and Http Request occurs.
2. Annotations (PDF/XML) from user is stored and stored/retrieved though Postgre/SQL form annotations project.
3. Annotations then inputted to desired application.
4. User query can be handled with external ontology services is provided by ONKI service. Ontology concepts are result of query.
5. Ontology is used in information extraction which gives concepts and entities when input is given as a document.

## VII. Expected Result

An inference engine can be use to improve the scalability as well as efficient reasoning. MapReduce can handle large collection data by handling load balancing problem and limit data exchange by dynamically scheduling jobs on computing nodes. Transfer and assertion model will reduce storage and also reduces computation time. RDF closure is not recomputed so reasoning time will significantly reduce.

## Conclusion

With the reasoning on large scale data is a major challenge which increases complexity of tasks. So the ontology-based semantic inference can be implemented to achieve better performance. This paper presents an approach to achieve high performance and scalability of ontology reasoning in semantic web. The large scale extraction of data is achieved and query processing time can be reduced with the use of inference engine. The goal is to provide an environment for improving parameters of the Hadoop configuration based on semantic knowledge through ontologies. Thus, users would get a better performance. Semantic Web technologies can be used in a variety of application areas for example; healthcare and life sciences, business process management, expert systems, e-marketplace, Web services composition and cloud system management.

## Future Scope

The ontology-based semantic approach is used for Hadoop MapReduce environment for configuration parameters associated with workloads and their specific characteristics semantically modeled.

## References

[1] Bo Liu; Keman Huang; Jianqiang Li; MengChu Zhou, "An Incremental and Distributed Inference Method for Large-Scale Ontologies Based on MapReduce Paradigm," *Cybernetics, IEEE Transactions on* , vol.45, no.1, pp.53,64, Jan. 2015

[2] Yujiang Liu, "Study of Semantic Web Usage Mining," *Communication Systems and Network Technologies (CSNT), 2013 International Conference on*, vol., no., pp.678,680, 6-8 April 2013

[3] Ibanez, M.J.; Fabra, J.; Alvarez, P.; Ezpeleta, J., "Model Checking Analysis of Semantically Annotated Business Processes," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* , vol.42, no.4, pp.854,867, July 2012

[4] Jingzhi Guo; Lida Xu; Zhiguo Gong; Chin-Pang Che; Chaudhry, S.S., "Semantic Inference on Heterogeneous E-Marketplace Activities," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* , vol.42, no.2, pp.316,330, March 2012

[5] Jianling Sun; Qiang Jin, "Scalable RDF store based on HBase and MapReduce," *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*, vol.1, no., pp.V1-633,V1-636, 20-22 Aug. 2010

## Author's Profile

**Vrushali Tarachand Lanjewar**
received B.E degree in Computer Science from RTM Nagpur University and currently pursuing M.E degree in Computer Science and Information Technology from Department of Computer Science, S.G.B.Amravati University, Amravati. Her research interests include use of MapReduce of hadoop in real life applications.