Dealing with Ontological Queries

1. Example Dataset
   - POS67
   - RecommendedName
   - Type

2. Example Graph Pattern
   - structuredName
   - Protein
   - Amyloid beta A4 protein
   - 865M triples
   - Loading and inference: 290h
   - 64 nodes with 4GB memory
   - AlternativeName

3. Inference: compute the full closure of RDF dataset with rules.
   - 1) Ontology
   - RecommendedName
   - subPropertyOf
   - structuredName
   - 58:54
   - In the cloud (e.g. Amazon EC2), time is money!!
   - 37:18
   - With NTGA, MR workflow

Existing Techniques

- **Challenge:** Current dominant approach - Materialize then Query
- Static Materialization of RDF graph is expensive!

BigOWL3 3.1*

- Commercial RDF database engine
- Reasoning against 865M explicit triples: 1h
- Loading and inference: 290h on a single node.

- In the cloud (e.g. Amazon EC2), time is money!!
  - Virtuoso: 61 hours + 2 hrs to run queries vs. 8 hrs to run queries in situ.

- Techniques better suited for private clouds and private data (not data transfer)
  - There is room for improvement on 8hrs.

Query Rewriting: Expand into UCQs

- Computes only the closure of schema with restricted RDFS rules.*
- Re-write a graph pattern into multiple graph patterns involving with schema elements .
- Groups all the expanded queries (EQs) using unions.

- Processing grouped queries requires n × K MR jobs. (e.g., 60 MR jobs for LUBM Q5.)

1) Example Graph Pattern
2) Expanded Patterns

**Possible Optimization: MQO and NTGA**

1. Factor out Common Triples using Multi-Query Optimization**
   1) Expanded Patterns
   2) A single pattern with optional clauses

2. Processing Optional Graph Patterns with NTGA***
   - Comparison between Relational-style approach and NTGA-based approach.

   Consider triple relation T and query Q, with two star-patterns:

   \[ \sigma_{\text{T}}(s) = (\text{vendor}, \text{price}, \text{country}) \]

3. MR Workflow using Optional Patterns: 4 MR jobs (3 MR jobs and 1 Map Only job)

With NTGA, MR workflow for the example query: 1 MR job

**RAPID+ Architecture/ Preliminary Evaluation**

**RAPID+**: an extended Apache Pig system that integrates the NTGA data model and algebra.

1. Preliminary performance evaluation of Union, LOJ-Union (Optional), and NTGA approaches.
   (the 231GB dataset of LUBM dataset on a 5-node cluster)

2. Process Unigrep in Bio2RDF with 30 Nodes
   - In LOJ-Union, the 2nd MR job took more time than expected due to the joins with high degree multi-value property.

References:
**Ravindra Li, Anupa Sen, Anirudh Srivastava, Yogeshwar Dutt and Kripa V. Scalable batch query optimization for BigRDF. In Proc. SIGMOD ’12.