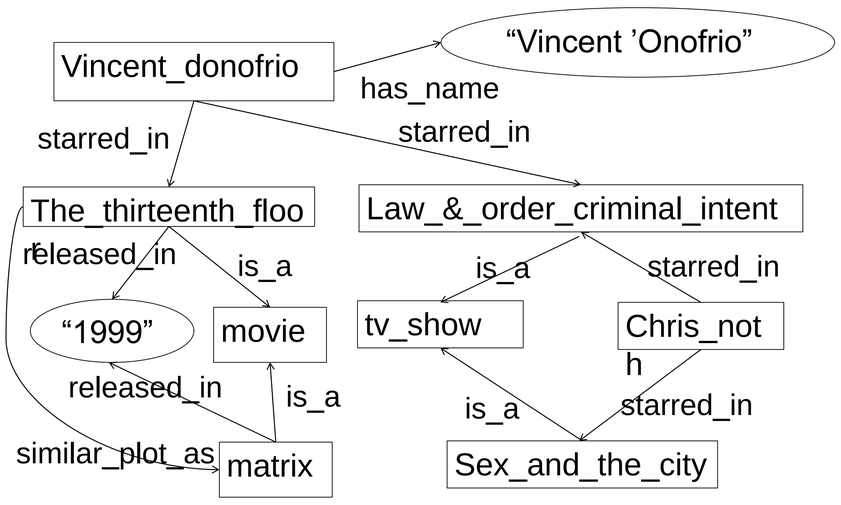
A Two-Tiered Index Architecture for Scalable RDF Processing

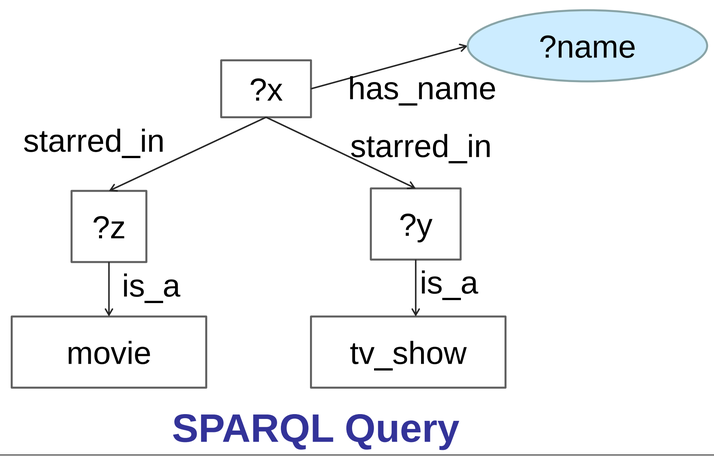
Sairam Gurajada *and* Martin Theobald {gurajada,mtb} @ mpi-inf.mpg.de

Introduction

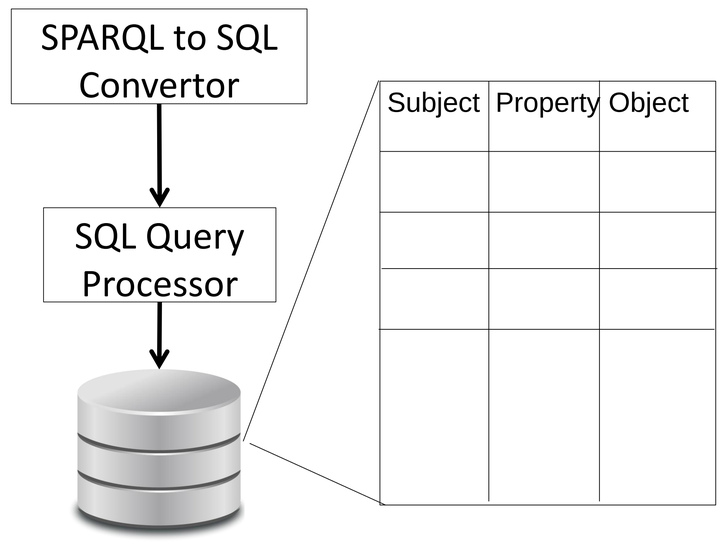


RDF (Graph structured knowledge base)

Who acted in a "tv-show" as well as in a "movie"?
Vincent Donofrio is one answer



SPARQL Query



RDF Triple Store – Relational Model

Motivation

How to scale RDF System?

- Over 30 Billion triples in the linked data cloud
- **Distributed Approach**

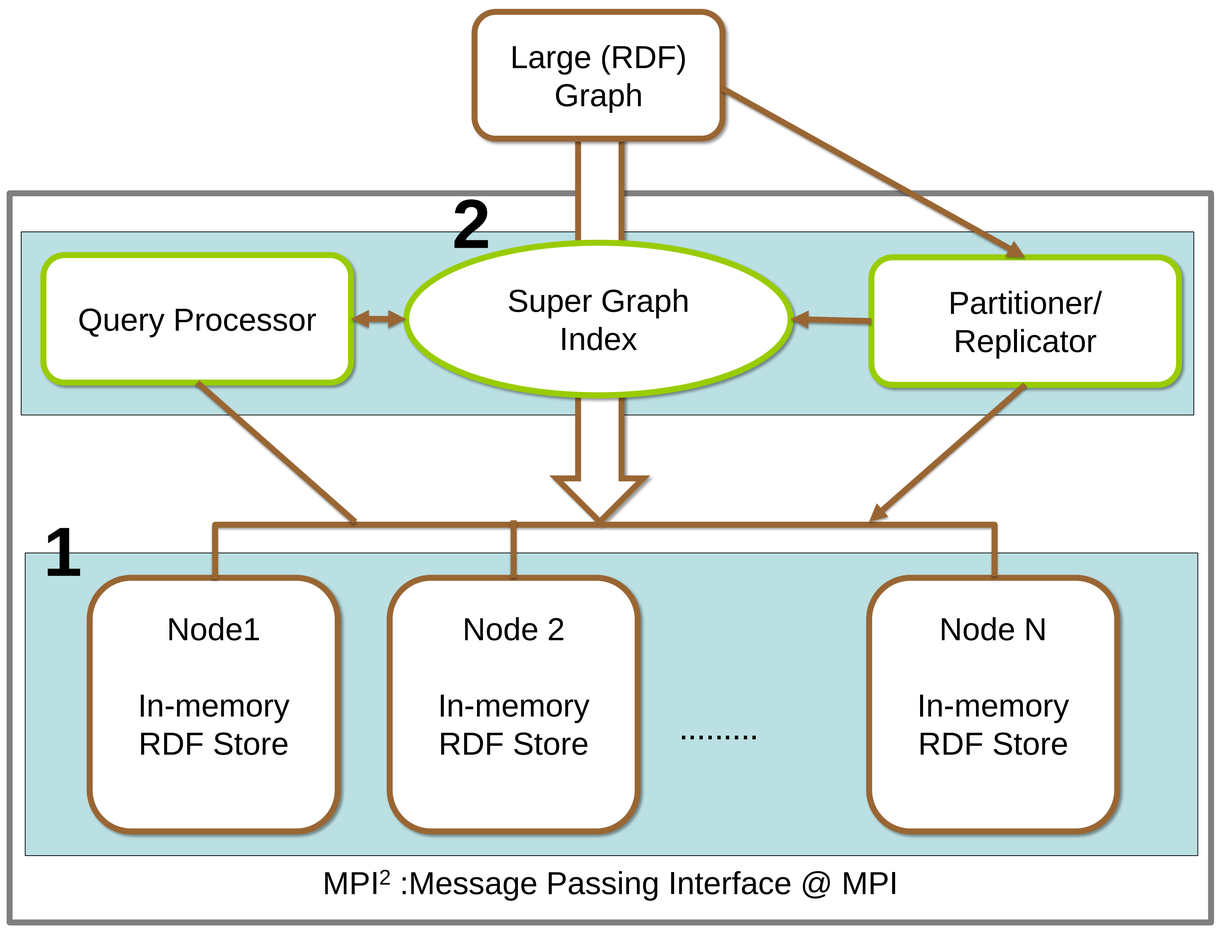
Challenges

- Minimize inter-node communication by effective ***partitioning*** and ***replication*** approaches
- Parallel query processing and effecient load balancing

Existing Approaches

- Distribute the triples by applying ***hashing*** on Subject or Object
  Inefficient to answer path queries, but Efficient load balancing
- Use ***graph clustering algorithms*** on large RDF graph and distribute the triples with 1 or 2-hop replication
  Poor load balancing, but efficient in processing path queries?

Architecture



Problem Statement

- A Partitioner P partitions the RDF graph G into (G_1, G_2, G_3... G_k)
- A query Q which requires multiple (m per say) partitions to answer is split in to (Q_1, Q_2,... Q_m)
- Let Ri be the set of results returned by executing Q_i ***independently*** on partition G_i .
- During join operation ,
  - R_i results are shiped to partition R_j ($R_i < R_j$) and joined to form result set R_{ij} ($<< R_i$ and $<< R_j$)
  - *Tuples Communicated:* R_i

- **How to optimize the number of tuples communicated in a join operation? which requries filtering out some of the tuples in Ri that do not participate in the join operation**
  Our Idea is to have a "Two-Tiered Index archiecture"
  Tier 1 : *Super graph index* contains the summary of original graph
  Tier 2 : *Original graph index*

- Queries are first posed to Super graph index which directs the search over regular graph index at cluster nodes

Current work

- How to build Super Graph (Summary graph) from RDF graph?
- Indexing approaches for Super Graph and regular RDF graph
- Can the state of the art single site RDF systems can be used locally at each slave?
- How to design a replicator for two-tiered index architecture
- .....

References:

1) Jiewen Huang, Daniel J. Abadi, Kun Ren: Scalable SPARQL Querying of Large RDF Graphs. PVLDB 4(11): 1123-1134 (2011)
2) Thomas Neumann, Gerhard Weikum: The RDF-3X engine for scalable management of RDF data. VLDB J. 19(1): 91-113 (2010)
3) Lei Zou et al. gStore: answering SPARQL queries via subgraph matching, VLDB 2011
4)Sairam Gurajada, P. Sreenivas Kumar: On-line index maintenance using horizontal partitioning. CIKM 2009: 435-444

max planck institut informatik

UNIVERSITÄT DES SAARLANDES

International Max Planck Research School for Computer Science