Ge Song, Frédéric Magoulès (Supervisor), Fabrice Huet (Co-Supervisor)
Lab MICS
CentraleSupélec
Université Paris-Saclay

Parallel and Continuous Join Processing for Data Stream

Thèse pour l'obtention du grade de Docteur

Table of Contents

Introduction

Big Data Everywhere

- Google: 24 PB / day
- Facebook: 10 millions photos + 3 billion “likes” / day
- Youtube: 800 million visitors / month
- Twitter: Doubling its size every year

Issues

- The most significant issue comes from the size of Big Data.
- The flip side of size is speed.
- The cost of network communication in transferring data.
- The dynamics of data.

Dynamic Data Stream

Persistent Static Relations $\Rightarrow$ Transient Dynamic Data Streams

Batch-oriented data processing $\Rightarrow$ Real-time stream processing

$\Downarrow$

Architecture Level: possible to add or remove computational nodes based on the current load

Application Level: able to withdraw old results and take new coming data into account

Objective: parallel and continuous processing for Join operation

Join: a popular and often used operation in the big data area.

- Data parallelism $\Rightarrow$ Data Driven Join $\Rightarrow$ kNN
- Task parallelism $\Rightarrow$ Query Driven Join $\Rightarrow$ Semantic Join on RDF data

Part I: Data Driven Stream Join (kNN)

Outline

- Related Work
- Parallel Workflow
- Theoretical Analysis
- Continuous kNN
- Experiment Result
- Conclusion

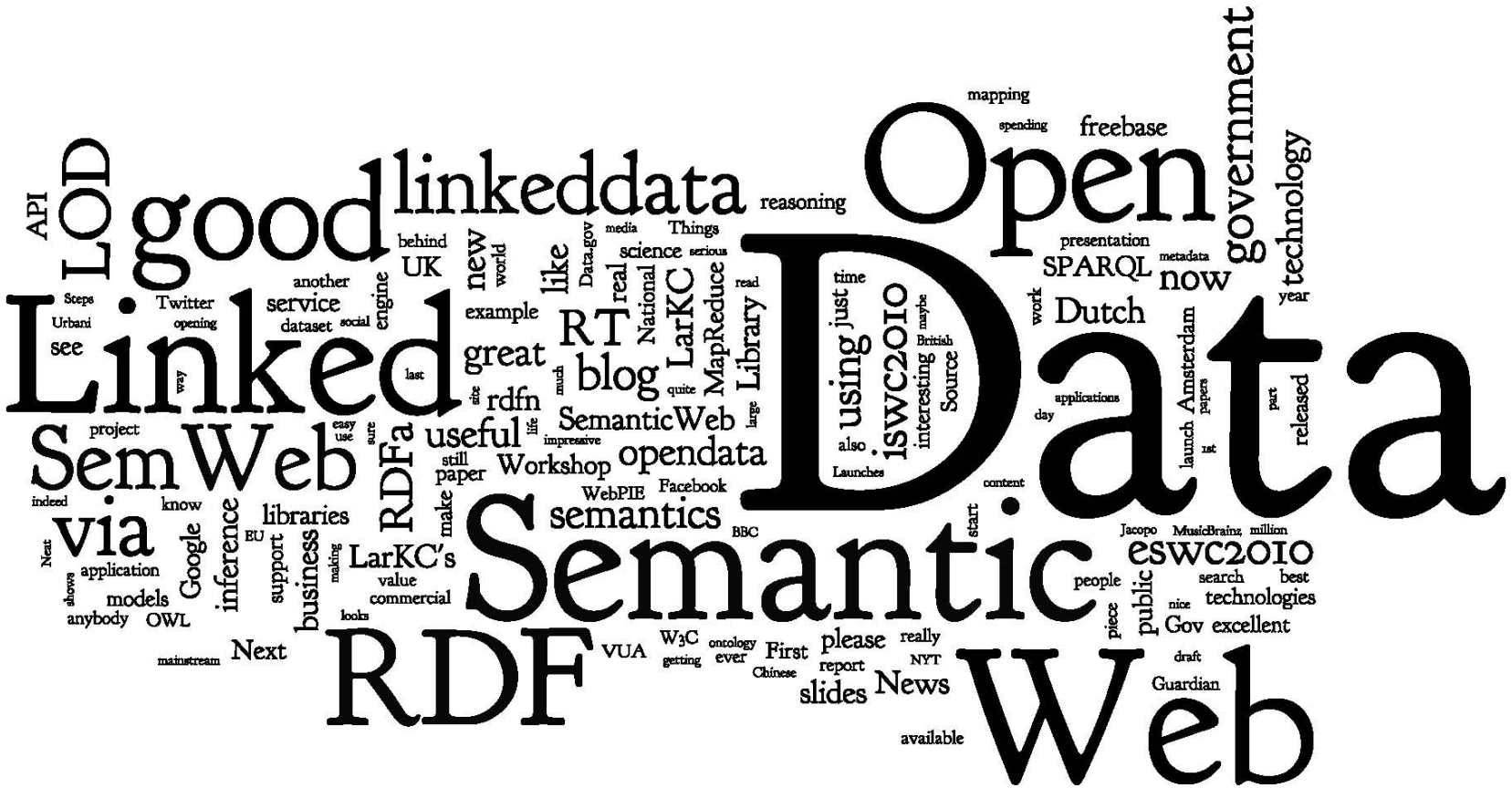
Outline

- Related Work
- Parallel Workflow
- Theoretical Analysis
- Continuous kNN
- Experiment Result
- Conclusion

Definition: kNN

Given a set of query points $\mathbb{R}$ and a set of reference points $\mathbb{S}$, a $\mathbb{k}$ **nearest neighbor join** is an operation which, for each point in $\mathbb{R}$, discovers the $\mathbb{k}$ nearest neighbors in $\mathbb{S}$.

Part II: Query Driven Stream Join (RDF)



Outline

- Related Work
- Query Decomposition and Distribution
- Data Partition and Assignment
- Parallel and Distributed Query Plan
- Continuous Join
- Analysis
- Implementation
- Experiment Result
- Conclusion

Outline

- Related Work
- Query Decomposition and Distribution
- Data Partition and Assignment
- Parallel and Distributed Query Planner
- Continuous Join
- Analysis
- Implementation
- Experiment Result
- Conclusion

Thank You!