A novel MapReduce-based approach for distributed frequent subgraph mining

Sabeur Aridhi, Laurent d'Orazio, Mondher Maddouri and Engelbert Mephu Nguifo

July 04, 2014

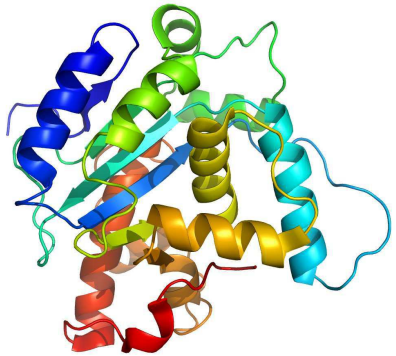
Context and motivations

Application domains

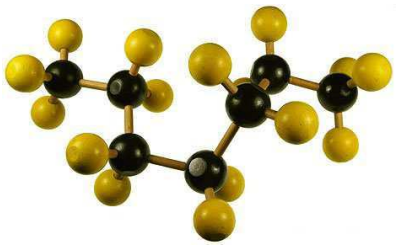
- Computer networks,
- Social networks,
- Bioinformatics,
- Chemoinformatics.

Graph representation

- Data modeling.
- Identifying relationship patterns and rules.



Protein structure



Chemical compound

Social network

Context and motivations

Mining graph data

- Graph mining aims to find patterns, hidden relations and behaviors in data.

Context and motivations

Mining graph data

- Graph mining aims to find patterns, hidden relations and behaviors in data.

Mining graph goals

- Computing graph properties:
  - Density, diameter, radius, ...
- Mining substructures from graph databases.
  - Substructures: paths, trees, subgraphs.
  - Frequent Subgraph Mining (FSM) task.

Context and motivations

Availability of graph data

- Exponential growth in both size and number of graphs in databases.

Context and motivations

Availability of graph data

- Exponential growth in both size and number of graphs in databases.
- Availability of graph data sources:
  - The protein data bank (PDB) contains 95280 of protein 3D structures.
  - Facebook loads 60 terabytes of new data every day **[Thusoo 2010]**.
  - Google processes 20 petabytes of data per day **[Dean 2008]**.

Context and motivations

Availability of graph data

- Exponential growth in both size and number of graphs in databases.
- Availability of graph data sources:
  - The protein data bank (PDB) contains 95280 of protein 3D structures.
  - Facebook loads 60 terabytes of new data every day **[Thusoo 2010]**.
  - Google processes 20 petabytes of data per day **[Dean 2008]**.
- 3Vs of Big Data (Volume, Velocity and Variety).
- Availability of cloud computing environments.

Context and motivations

In this work

- We are interested to FSM from graph databases.

Context and motivations

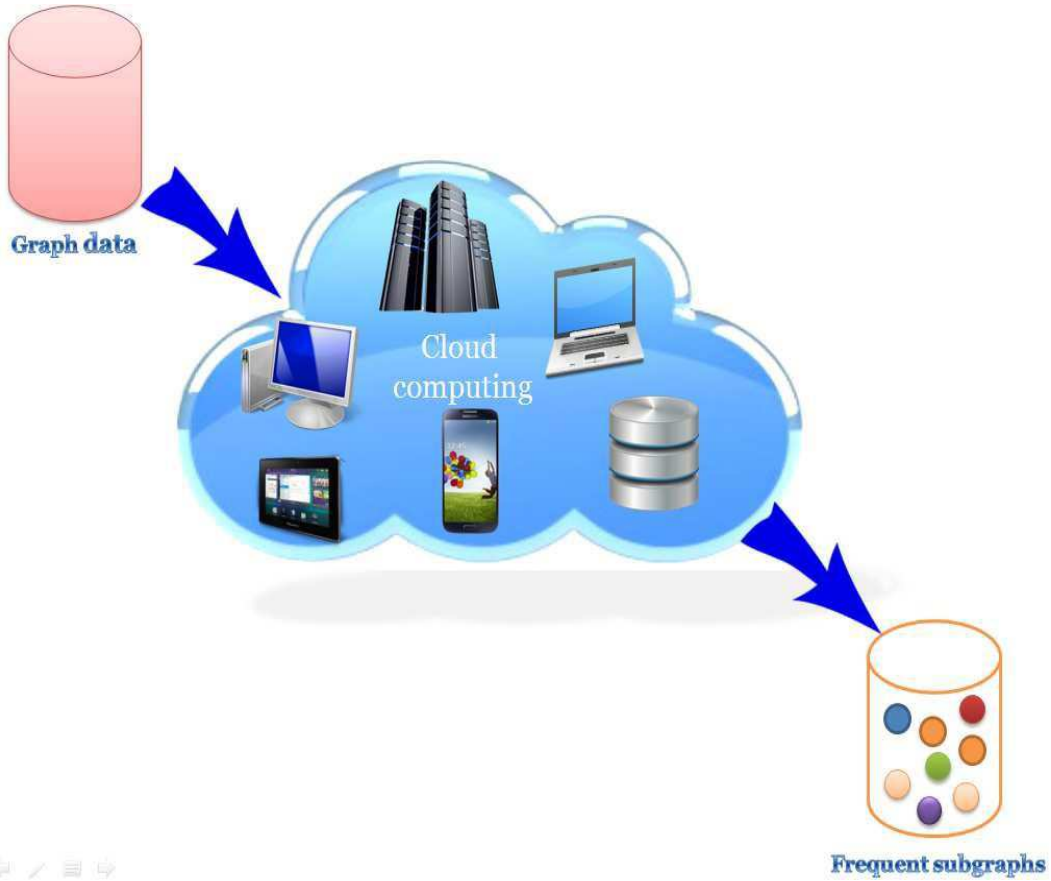
In this work

- We are interested to FSM from graph databases.

Frequent subgraph mining algorithms

- Various approaches of FSM.
- Existing approaches are mainly:
  - Tested on centralized computing systems.
  - Evaluated on relatively small databases.
- Few works for FSM in the cloud.

Goals



Questions

- Distributed FSM from large graph database.
- Data/computation distribution.
- Tuning cloud parameters.

Outline

Outline

Outline

Background

Graph

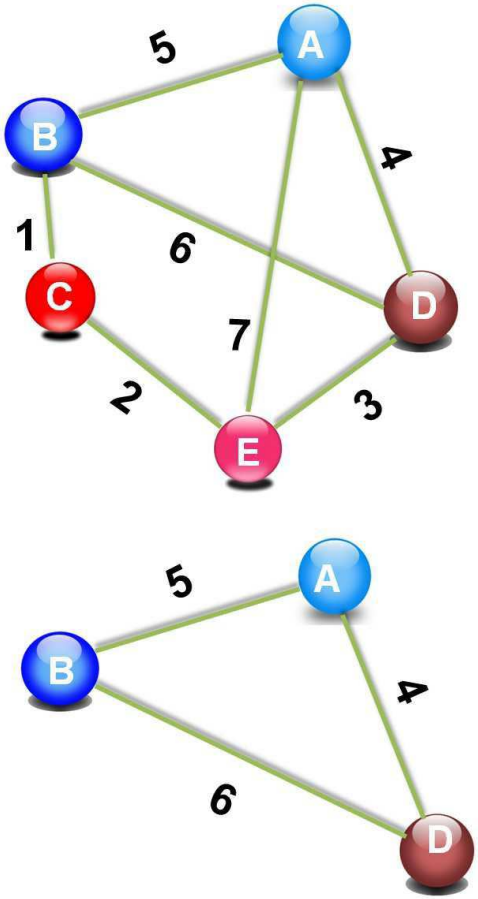
A graph is denoted as $G = (V, E)$ where V is a set of nodes and E is a set of edges.

Subgraph

A graph $G' = (V', E')$ is a subgraph of another graph $G = (V, E)$ iff: $V' \subseteq V$, and $E' \subseteq E \cap (V' \times V')$.

Density

The density of a graph $G = (V, E)$ is calculated by $density(G) = \frac{2 \cdot |E|}{(|V| \cdot (|V| - 1))}$.



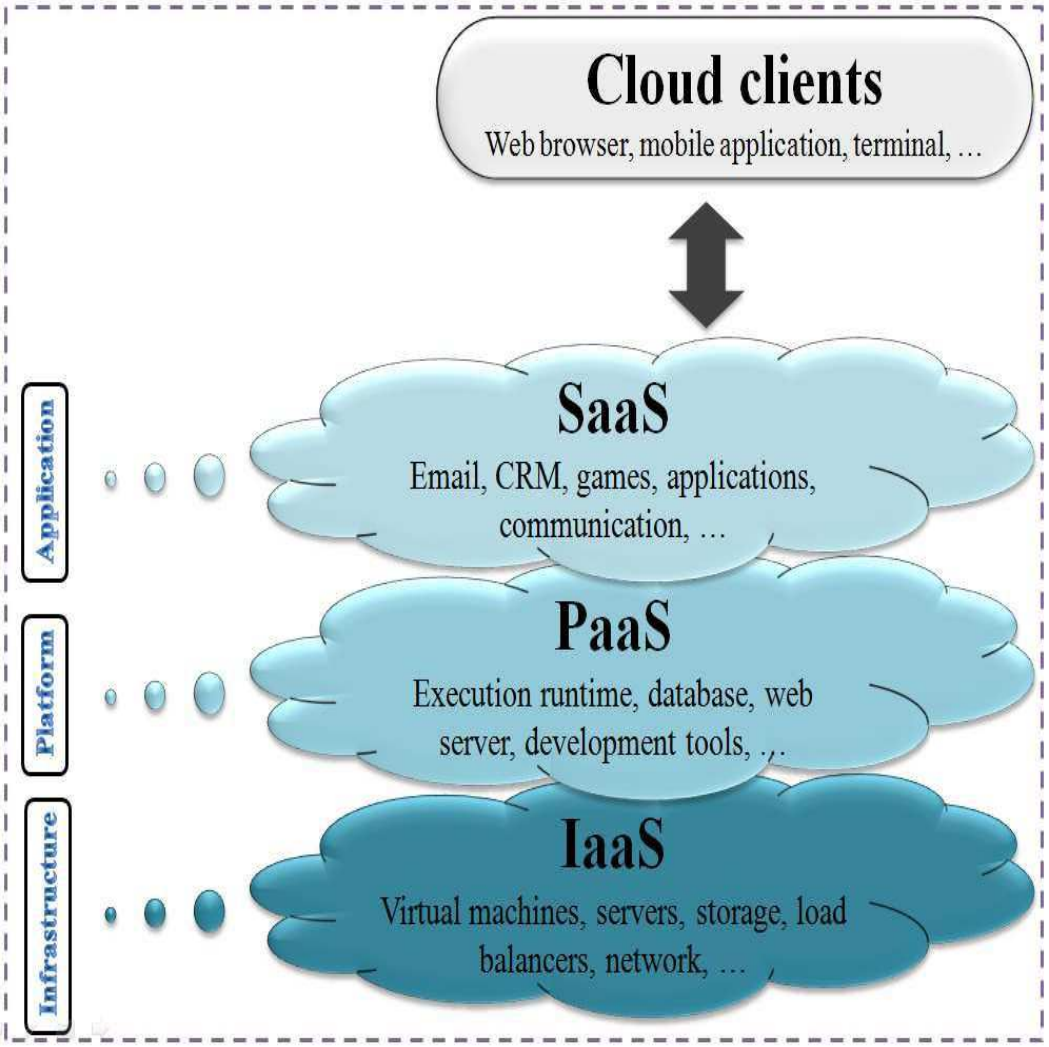
Outline

Background

Cloud computing

- Large number of computers that are connected via Internet.
- Applications delivered as services.
- Hardware and system software delivered as services.
- Pay as you go.
- Cloud services can be rapidly and elastically provisioned.

Background



Service models

- Software as a Service (SaaS).
- Platform as a Service (PaaS),
- Infrastructure as a Service (IaaS),

Outline

Background

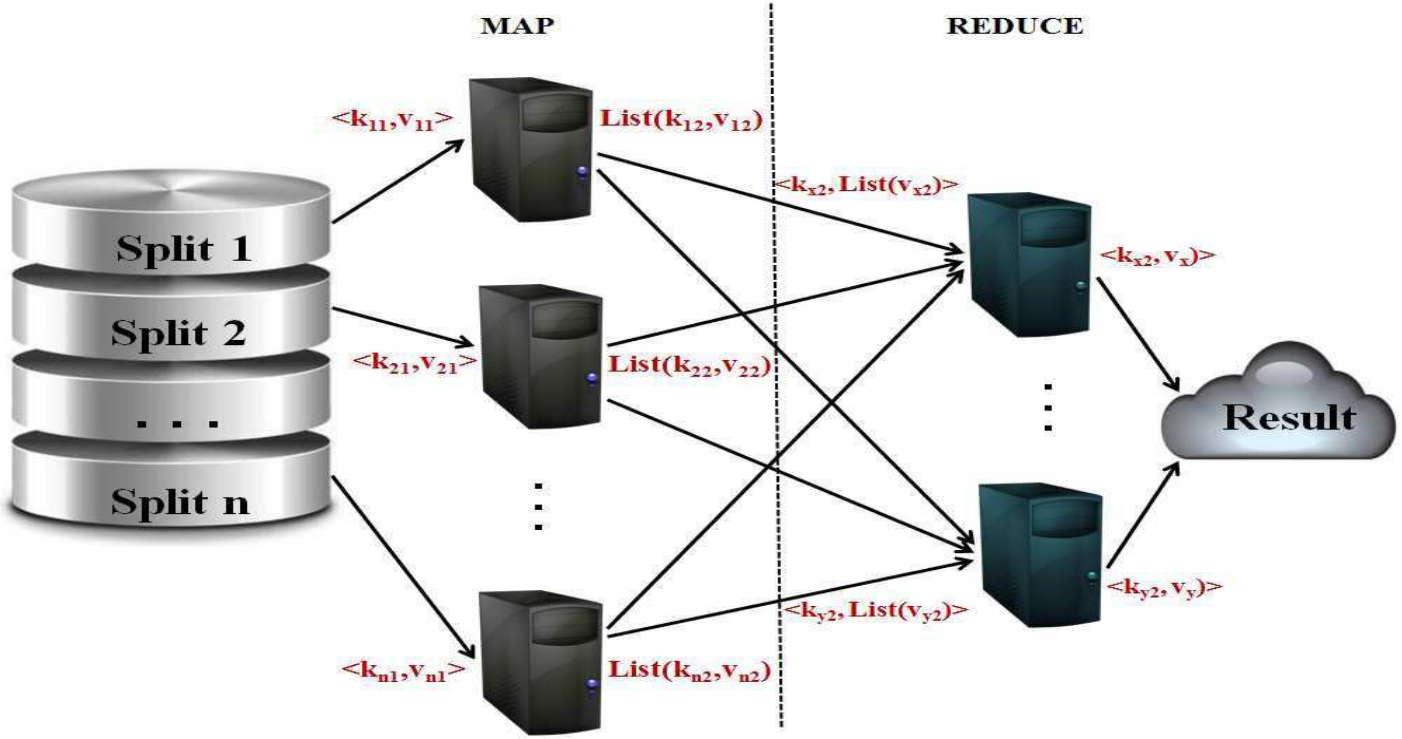
MapReduce framework

- A framework for processing huge datasets.
- Large number of computers and task/node failures.

Background

MapReduce framework

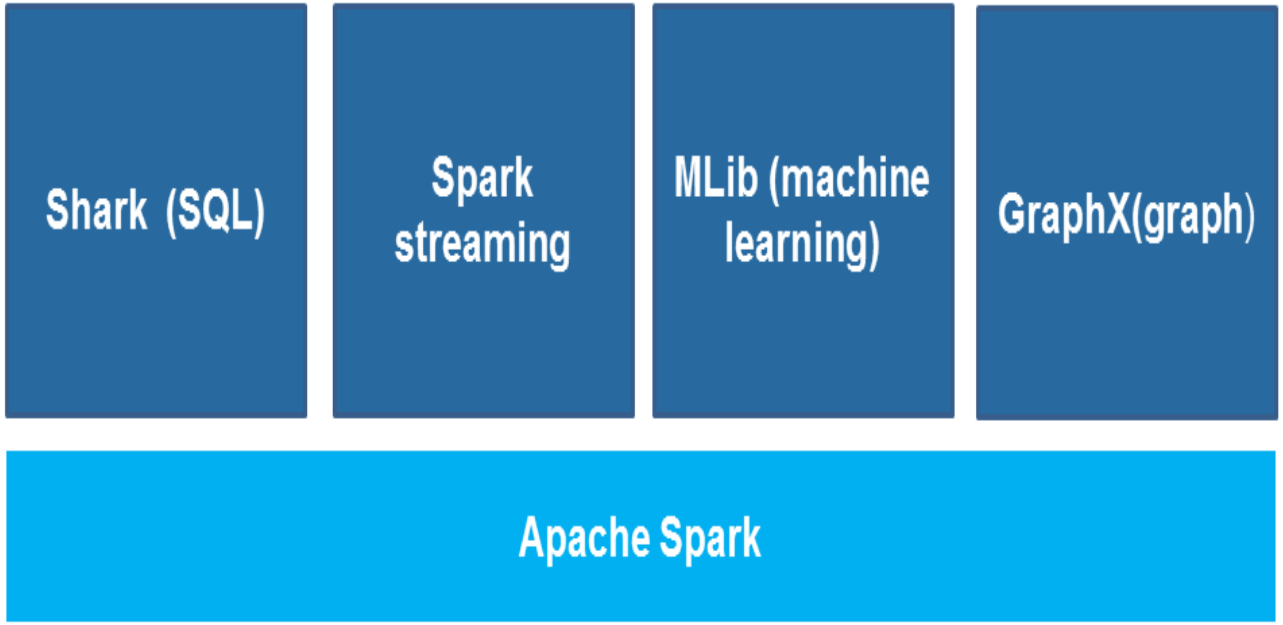
- A framework for processing huge datasets.
- Large number of computers and task/node failures.



Background

SPARK framework

- A general engine for large-scale data processing.
- Combine SQL, streaming, and complex analytics.
- It offers several high-level operators that make it easy to build parallel applications.



Background

SPARK framework: GraphX

- A part of the Apache Spark project.
- API for graphs and graph-parallel computation.
- PageRank.
- Connected components.
- Label propagation.
- Strongly connected components.
- Triangle count.