Big Graph Mining : Frameworks and Techniques

Sabeur Aridhi and **Engelbert Mephu Nguifo**

EU COST BigSkyEarth Workshop
Sopron (Hungary), February 23-24, 2017

Big Data Research 6 (2016) 1–10

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Big Graph Mining: Frameworks and Techniques

Sabeur Aridhi [a,*], Engelbert Mephu Nguifo [b,c]

[a] Aalto University, School of Science, P.O. Box 12200, FI-00076, Finland
[b] Clermont University, Blaise Pascal University, LIMOS, BP 10448, F-63000 Clermont-Ferrand, France
[c] CNRS, UMR 6158, LIMOS, F-63173 Aubiere, France

ARTICLE INFO

Article history:
Received 6 January 2016
Received in revised form 12 June 2016
Accepted 24 July 2016
Available online 24 August 2016

Keywords:
Big graphs, data mining
Pattern mining
Graph processing frameworks

ABSTRACT

Big graph mining is an important research area and it has attracted considerable attention. It allows to process, analyze, and extract meaningful information from large amounts of graph data. Big graph mining has been highly motivated not only by the tremendously increasing size of graphs but also by its huge number of applications. Such applications include bioinformatics, chemoinformatics and social networks. One of the most challenging tasks in big graph mining is pattern mining in big graphs. This task consists on using data mining algorithms to discover interesting, unexpected and useful patterns in large amounts of graph data. It aims also to provide deeper understanding of graph data. In this context, several graph processing frameworks and scaling data mining/pattern mining techniques have been proposed to deal with very big graphs. This paper gives an overview of existing data mining and graph processing frameworks that deal with very big graphs. Then it presents a survey of current researches in the field of data mining/pattern mining in big graphs and discusses the main research issues related to this field. It also gives a categorization of both distributed data mining and machine learning techniques, graph processing frameworks and large scale pattern mining approaches.

Motivation : Research topic

Our research team : **Data, Services and Interoperability**

Two main directions for Big Data :

- Data Management
  - Query optimisation approaches
    - Integrated systems
    - Hybrid storage systems
  - Indexing technique for biological data

- Pattern mining – Machine learning
  - Rule mining
  - Graph mining
    - Cloud computing : Partitioning approach
    - Evolving graphs
  - Preferences in data mining
  - Missing data in Spatial Datawarehouse
  - Recommender systems

Motivation : Multidisciplinarity

Theme DSI --- some Projects :

LabEx IMobS^3 2012-2022

- "Innovative Mobility: Smart and Sustainable Solutions" Track #2

Investissement d'avenir BreedWheat 2011-2020

- Wheat, Biology, Bioinformatics, Genetics, Genomics, Ecophysiology, high throughput phenotyping and genotyping, …

CNRS Mastodons PETASKY 2012-2015

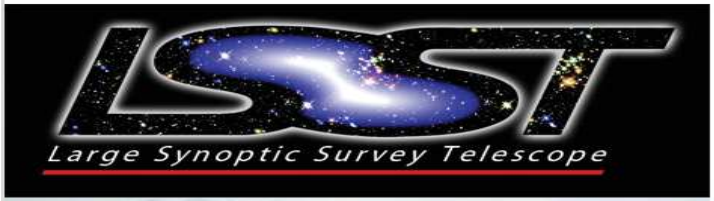
- Management and exploration of massive scientific data from astronomy observations
- EU COST Project BigSkyEarth 2015-2018

French-Brazil CNRS-INRIA-FUNCAP Project LSTG 2016-2018

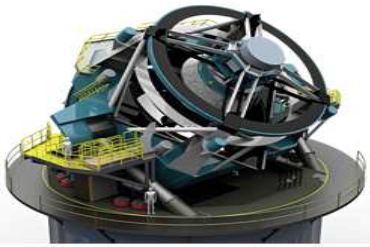
- UCA / USVQ - Universitade Federal do Ceara (Fortaleza)
- *Large-Scale Time Dependent Graphs*

Motivation : Mastodons project Petasky

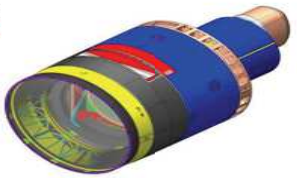
- Partners : LIMOS, LIRIS, LABRI, LIF, LIRMM, PRISM, ..., LPC, LAL, APC, LAM, ... IN2P3
- Study of challenges related to astronomical data
  - Data Storage and Indexing - Query Optimisation
  - Photometric redshifts reconstruction – Pattern mining



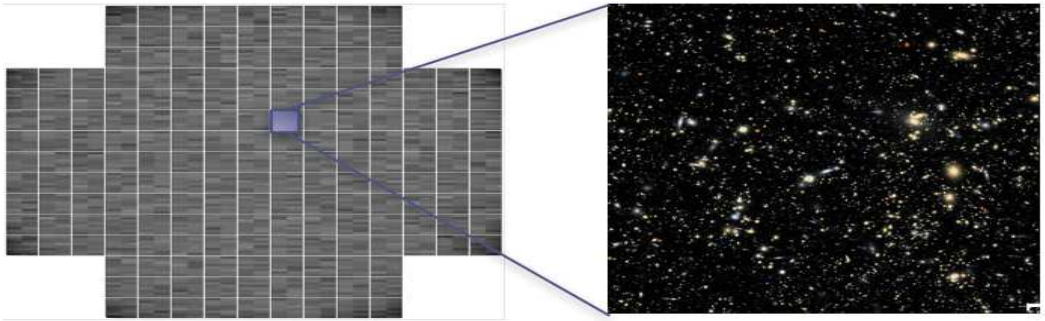
2020



Camera : 198 CCD (16 Mpix)
→ 3,2 G pixels !
~ 6 Gbyte / 17 secondes
→ **15 TB / night**

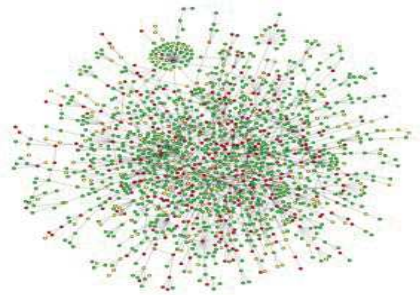


10 years : 60 Pbytes of data

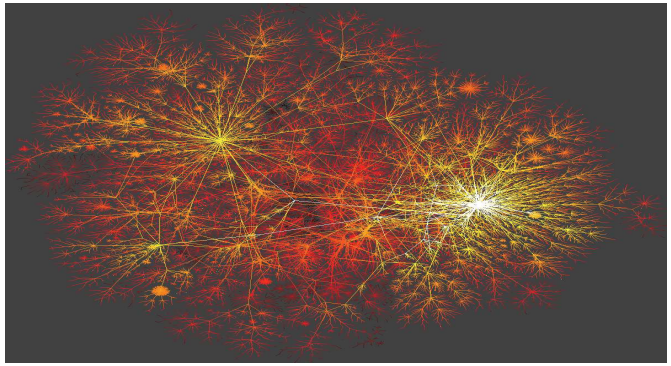


Motivation : data structure

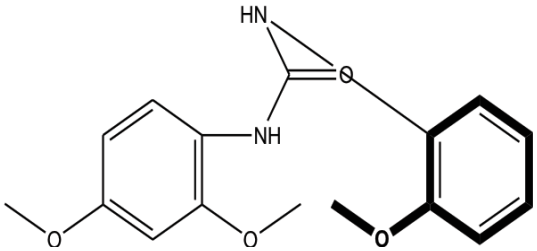
Graphs are very useful for modeling variety of entities and their relations



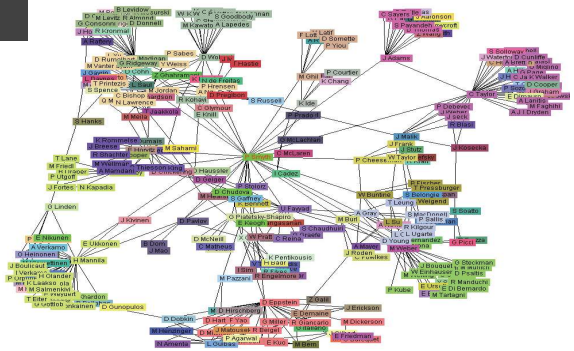
Yeast protein interaction network



Internet



Chemical compounds



Co-author network

Motivation : Large Graphs in the world !

- Social graphs : Facebook, Twitter, Google+, LinkedIn, etc…

  Facebook [2012] : a billion users (nodes) and more than 140 billion friendship relationships (edges).

- Endorsement graphs : web link graph, paper citation graph, etc…

  #web pages = 30 billions or more, and #devices = probably more than a billion.

- Location graphs : map, power grid, phone network, etc…
- Co-occurrence graphs : term-document bipartite, click-through bipartite, etc…

  MEDLINE adds from 1 to 140 publications per day

Motivation : EU COST Project BigSkyEarth → Prospective

where are Graphs in Earth or Sky ?

- C. Schwarz's Talk on Earth observation :
  - Image analysis in the feature domain
    - Extract features, categories, and objects (Features extractor)
  - Image analysis in the stacked data domain (linked data)
  - Image analysis by visualization
    - high level descriptors to identify relationships and unexpected events

- D. Vinkovic'Talk on « Forest Data project »
  - Graph modeling : a tree is node, and relationship can be neighborhood, …

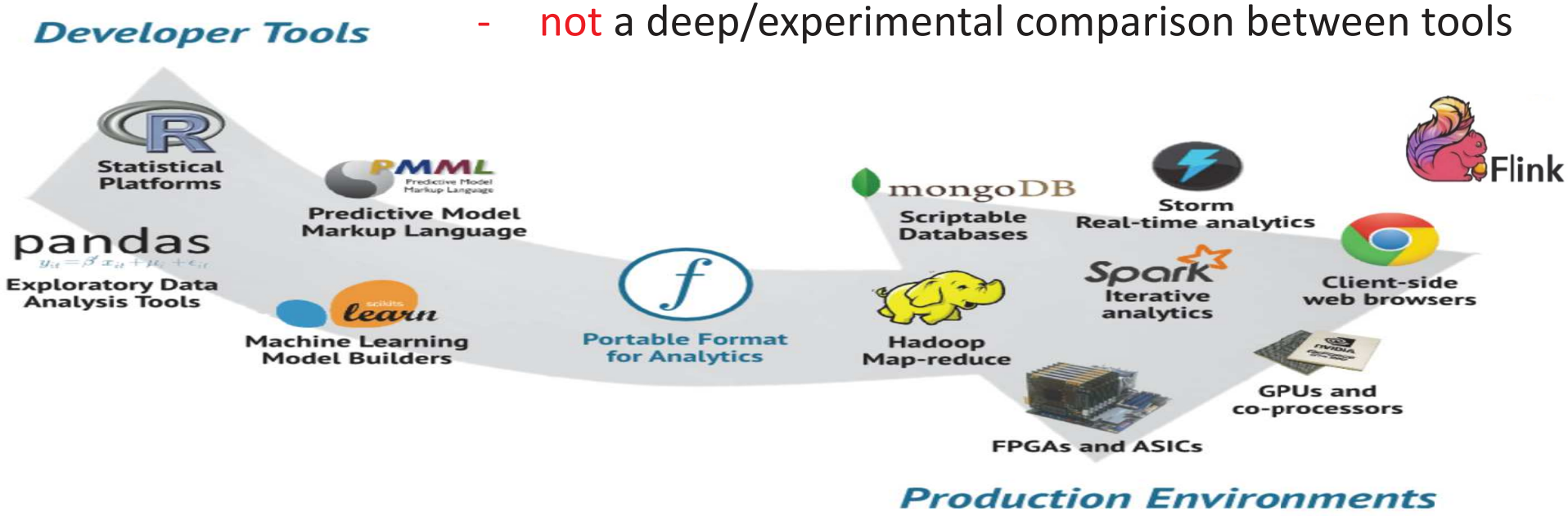
- P. Baumann's Talk on Datacubes : Rasdaman (multiple array)
  - ? Sparsity, empty cells in multidimensionnality

Motivation : Graphs are everywhere !

We need Big Graph Analytics

This talk is :

- a survey of *main* frameworks and techniques
- probably not exhaustive
- not a deep/experimental comparison between tools



Outline

- Graphs and graph mining
- Big data analytics
- Big Graph frameworks
- Some contributions
- Conclusion

Graph

Graph

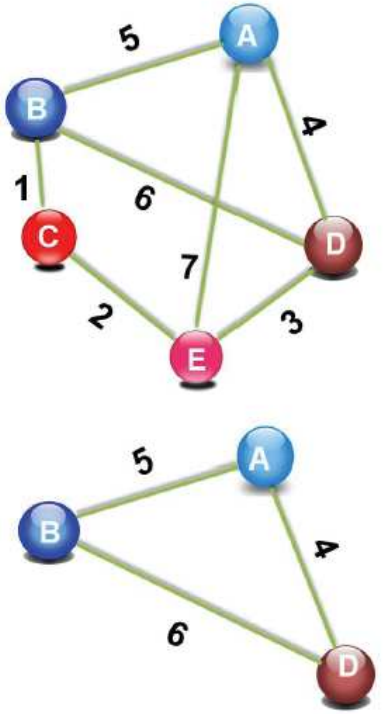
A graph is denoted as $G = (V, E)$ where V is a set of nodes and E is a set of edges.

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 $density(G) = \frac{2 \cdot |E|}{(|V| \cdot (|V| - 1))}$.



Graph processing/mining

Mining graph data

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

Mining graph goals

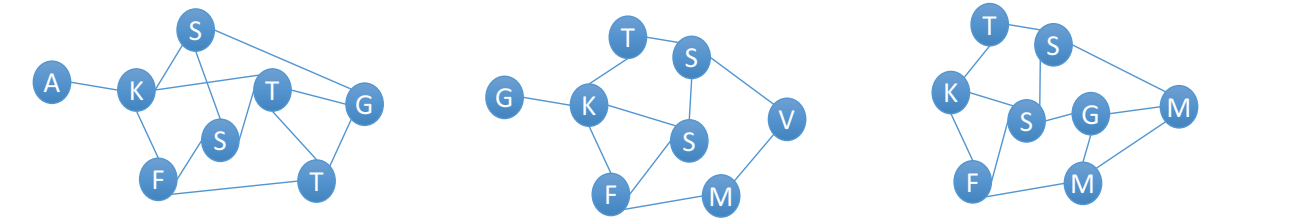
- Computing graph properties:
  - Density, diameter, radius, ...
- Mining substructures from graph databases.
  - Substructures: paths, trees, subgraphs.
  - Frequent Subgraph Mining (FSM) task.

Frequent subgraph discovery

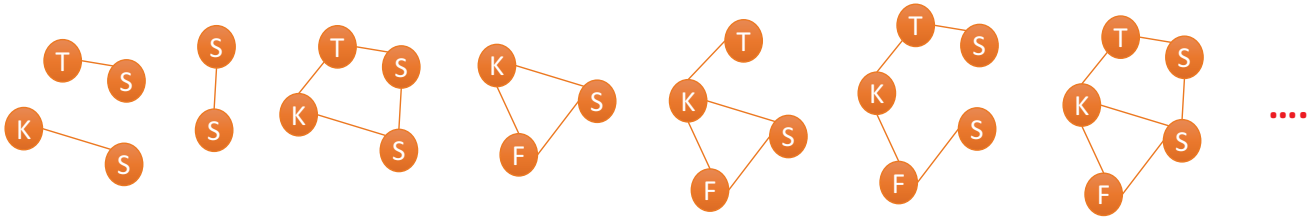
Goal:

- Finding subgraphs that occur in graph data, giving a minimum support

Example:



Graph database



Frequent Subgraphs
(minimum support = 3)

Outline

- Graphs and graph mining
- Big data frameworks/analytics
- Big graph frameworks/analytics
- Some contributions
- Conclusion

Big Data Framework

MapReduce :

•2004 : Google

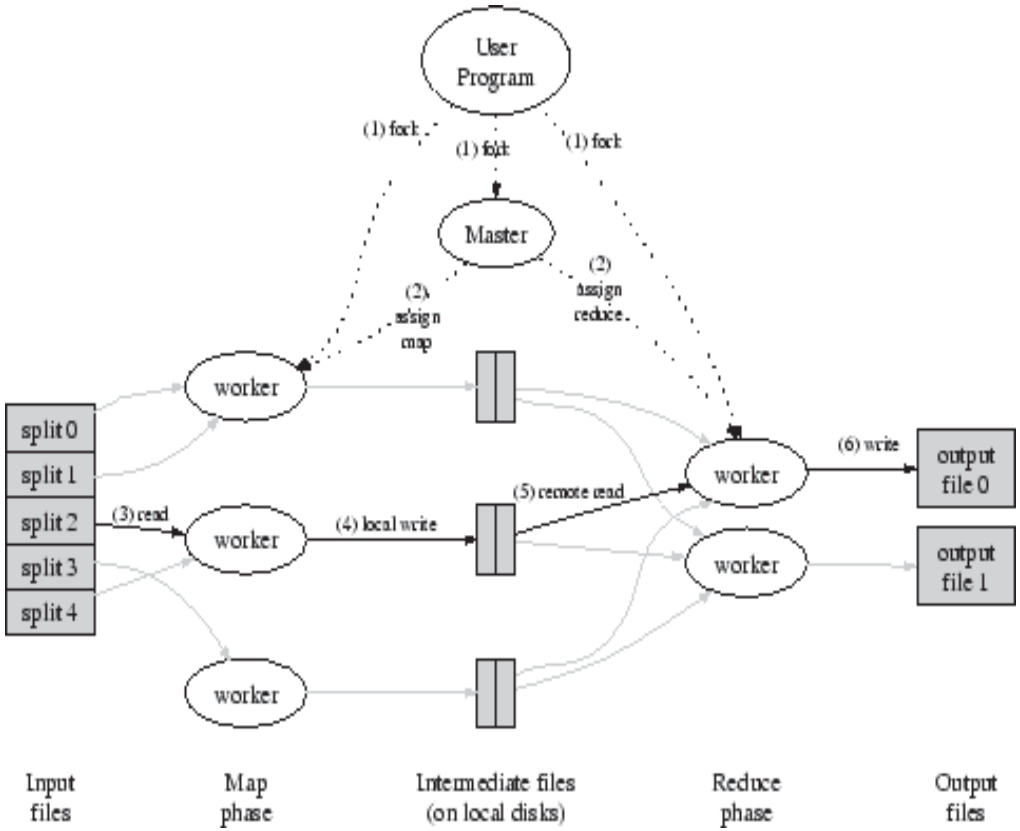
•A programming model and not an algorithm

•Executed in a distributed environment

•Several implementations such as **Hadoop**

•Based on two functions (Map and Reduce)

•Based on key-value format

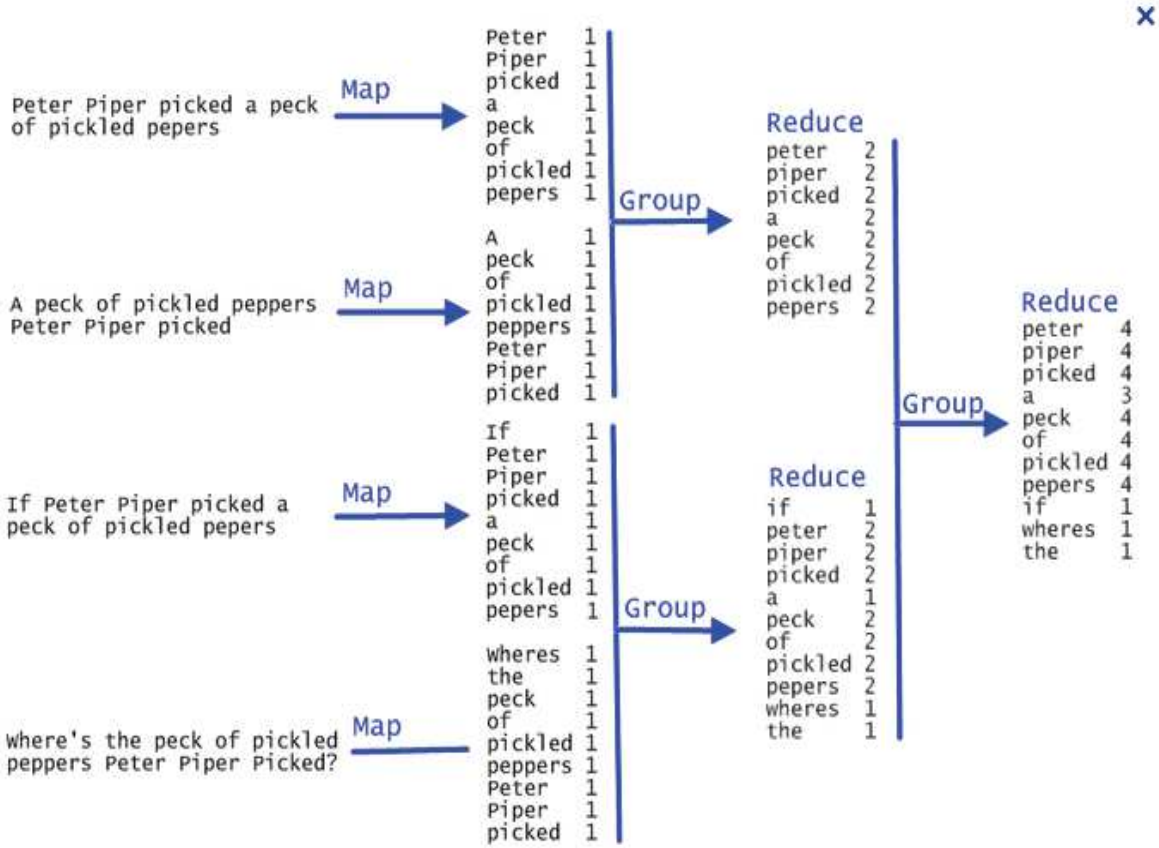


Big Data Frameworks

MapReduce :

Example : mothergooseclub.com

- Peter Piper picked a peck of pickled pepers
- A peck of pickled pepers Peter Piper picked
- If Peter Piper picked a peck of pickled pepers
- Where's the peck of pickled pepers Peter Piper Picked ?



Big Data Frameworks

MapReduce / Hadoop : Google 2004

Strength : General-purpose framework, powerful and simple, open-source
Limitation : General-purpose framework → e.g. : Less efficient on relational data

Some **solutions** : Problem and data-specific
- *Pig Latin* : Yahoo (2008) – Iterative data processing
- *Hive* : Facebook (2009) – SQL-like workload
- *Pregel* : Google (2010) – Graph processing. →. *GraphLab* (2011), *PowerGraph* (2012)
  - → *Giraph* : Facebook (2012) --- Iterative graph processing (Pregel open source)
  - → *Trinity* : Microsoft (2013) --- **started 2010**
- *SpatialHadoop* : U. Minnesota (2013) – Geospatial data processing

Alternatives :
- *Dryad* : Microsoft (2007) – https://www.microsoft.com/en-us/research/project/dryad/ started in **dec. 2004**
  - Quite expressive
  - Subsumes other computation frameworks, such as Google's map-reduce, or the relational algebra
  - Commercial (not open-source)
- …

Big Data Analytics

more ….

- **Spark** (Zaharia *etal*., USENIX 2010) https://spark.apache.org/
  - a powerful general purpose processing framework that provides an ease of use tool for **efficient analytics** of heterogeneous data
  - Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk
  - Spark runs on Hadoop, Mesos, standalone, or in the cloud.
  - It can access diverse data sources including HDFS, Cassandra, HBase, and S3.

- **Flink** https://flink.apache.org/
  - **Stratosphere** (Alexandrov *etal*., VLDB J. 2014) funded by German Research Foundation (DFG)
  - an open source framework for processing data in both real time mode and batch mode

- **Storm** (Toshniwal etal, SIGMOD 2014) http://storm.apache.org/ Twitter
  - an open source framework for processing large structured and unstructured data in real time

- **H2O** (Alagiannis etal., SIGMOD 2014),
  - a system that brings database-like interactiveness to Hadoop
  - develop an analytical interface for cloud computing, providing users with tools for data analysis

Comparisons

- https://fr.slideshare.net/sbaltagi/flink-vs-spark

- http://www.metistream.com/comparing-hadoop-mapreduce-spark-flink-storm/

Big Data ML-DM

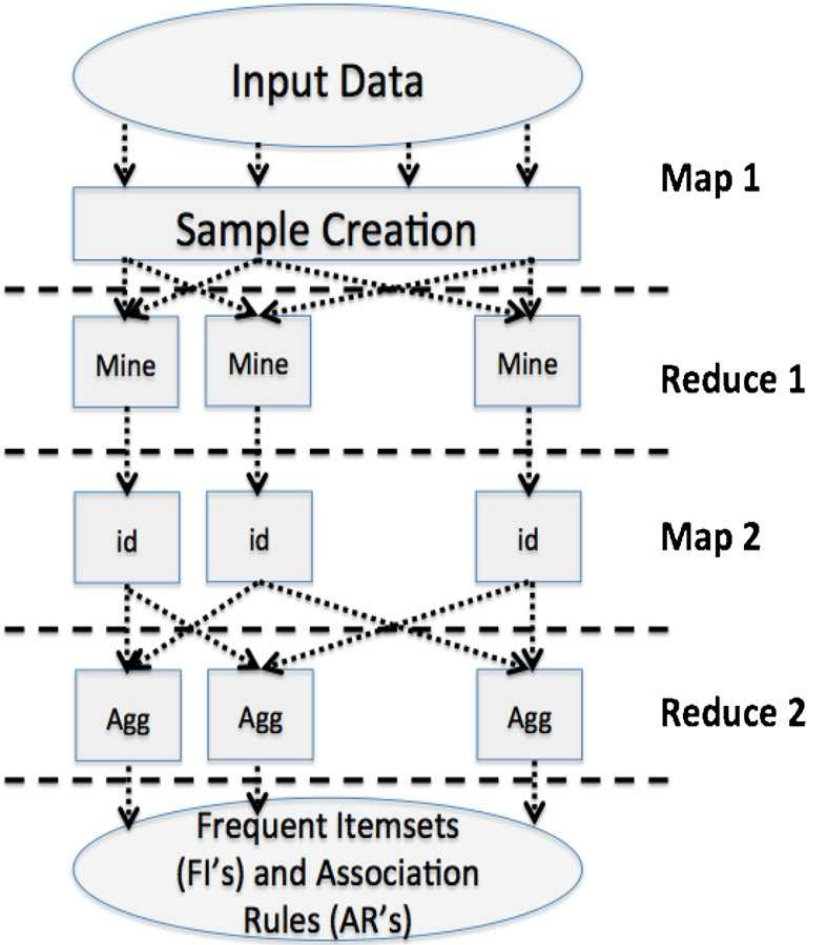
Table 1.
Overview of data mining and machine learning techniques.

Approach	Programming language	Programming model
NIMBLE	JAVA	MapReduce, OpenCL, MPI
Mahout	JAVA	MapReduce
SystemML	JAVA and DML	MapReduce
PARMA	JAVA	MapReduce
MLlib	JAVA and Scala	Spark
FlinkML	JAVA and Scala	MapReduce

Spark

PARMA [Riondato *etal*, CIKM 2012]

- Tool for mining frequent itemsets and association rules
- Run in a distributed mode
- Based on MapReduce



Big Data ML-DM

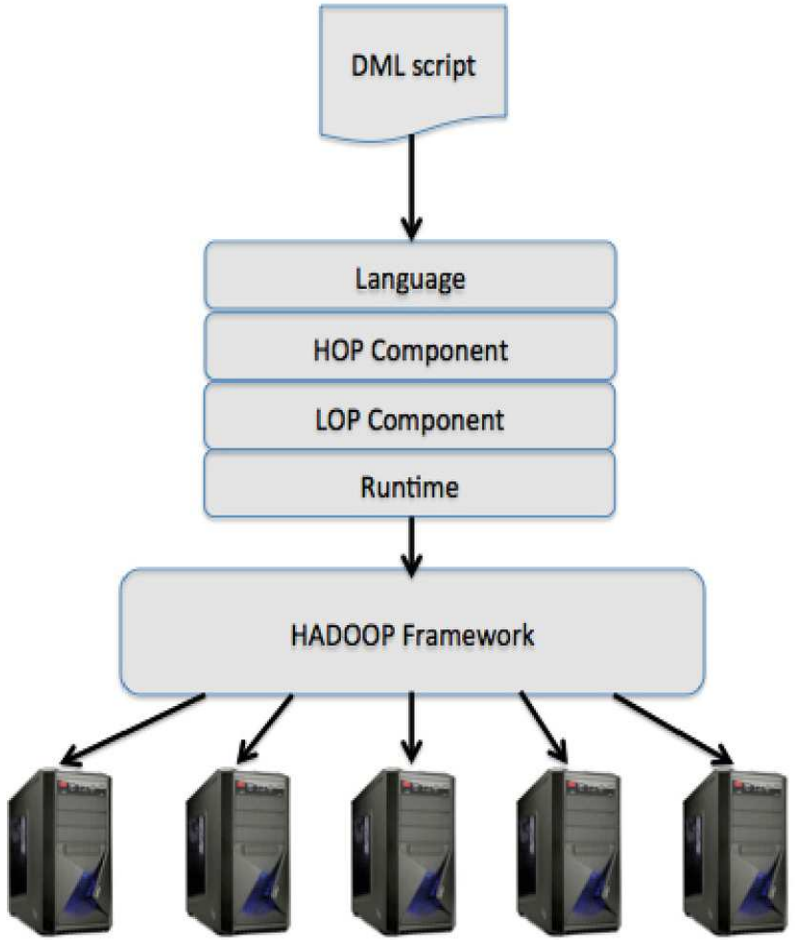
SystemML [Ghoting *etal*, ICDE 2011]

- Distributed framework
- Provides Machine learning solutions
- Run on top of Hadoop
- Based on a Declarative Machine

Learning Language

- High-Level Operator Component
- Low-Level Operator Component

A version on Spark [VLDB 2016]



Big Data Analytics

Mahout

http://mahout.apache.org/

- Set of scalable machine learning algorithms
- Based on Hadoop MapReduce
- Examples:
  - Recommender systems
  - Distributed Principal Components Analysis (PCA) techniques for dimensionality reduction
  - Clustering algorithms

Outline

- Graphs and graph mining
- Big data frameworks/analytics
- Big Graph frameworks/analytics
- Some contributions
- Conclusion

Graph Processing Frameworks

Table 2.
Overview of graph processing frameworks.

Framework	Asynchronous execution	Resources	Programming model
PEGASUS	No	Distributed system	Matrix operations
Pregel	No	Distributed system	Vertex-centric
Blogel	No	Distributed system	Graph-centric
GraphX	No	Distributed system	Edge-centric
GraphLab	Yes	Parallel systems	Vertex-centric

also : Trinity (Microsoft), Giraph (Facebook), …

Graph Processing Frameworks

Pregel

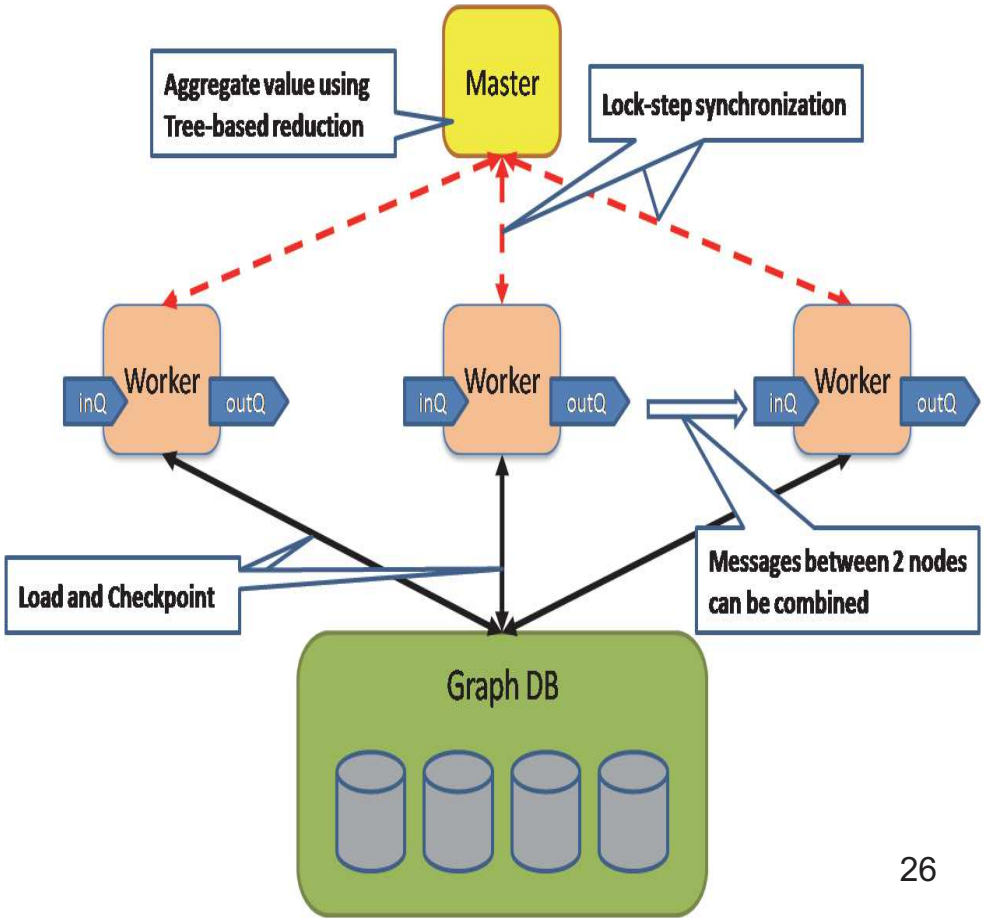
•Scalable and fault tolerant

•Vertex-oriented computing

•Flexible API to use in various graphs processing problems

•Compatible with the Hadoop ecosystem and master/slaves architecture.

•Based on the communications and the exchange of messages between the vertices.



Graph Processing Frameworks

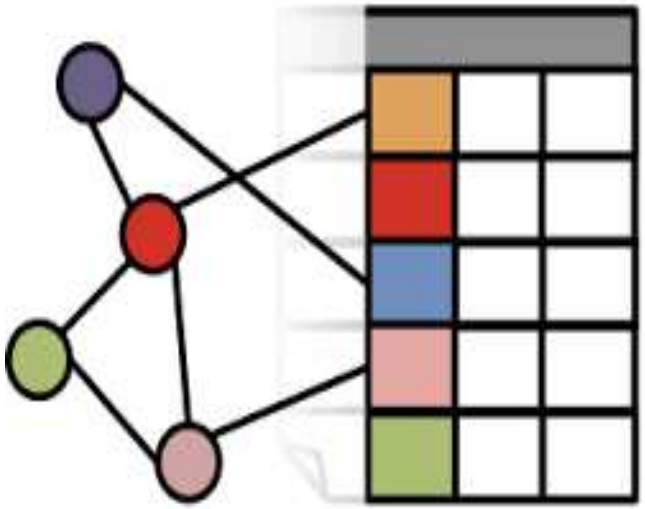
Running a Pregel program

1. Send the graph to master node.
2. Load the graph in the master (in GFS or BigTable).
3. Assign sub-graph to workers candidates.
4. Each worker follows the execution of their partitions (communication, exchange of messages, ..)
5. The master follows the states of the workers to get the final results.
6. Send the final results to user

Graph Processing Frameworks

GraphX

- Distributed and In-memory processing framework
- Sub project of Spark
- Compatible with many systems (HDFS, Apache Spark)
- Allows persistence to the RAM
- Uses the Resilient Distributed Datasets (RDD) concept
- Uses the Data Frame in the Spark project



Graph Processing Frameworks

GraphX