Hands-on GraphLab

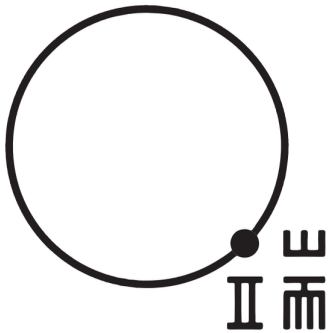
(Hands-on Massive Data Processing Platform Series)

Pili Hu

Initium Media 端傳媒

June 26, 2015

bit.ly/hkosc15-graphlab



Preparation & Prerequisites

Prerequisites: (for hands-on)

- Python required
- Python pandas is a plus
- Python networkx is a plus

Preparation:

- GraphLab https://dato.com/download/
- IPython Notebook:
  - https://github.com/initiummedia/hkosc2015-workshop

Sample Commands

virtualenv venv

source venv/bin/activate

pip install --upgrade --no-cache-dir https://get.dato.com/GraphLab-Create/1.4.1/e-hkosc15@hupili.net/64CA-557D-92A5-F7D8-91A9-CC38-C9C3-BBB9/GraphLab-Create-License.tar.gz

pip install -r requirements.txt

ipython notebook

(better to apply your own GraphLab Create trial license)

This Workshop (Series) Is About

- Hands-on three massive data processing platforms:
  - Hadoop
  - Spark
  - GraphLab
- Get the basic programming concept of the framework
- Get a feel of command-line/ shell of the framework

This Workshop (Series) Is NOT About

- How to install/ configure a cluster
- Rigorous performance evaluation
- Mathematical principle behind the frameworks
- Architecture of the platform from an implementation perspective

Expected Take-aways

- Demythify “Big Data Platforms”
- Benefit of framework:

  Dealing with small == dealing with big

Show-off to your friends:

Yeah, I got my hands-on XXX!

Choices of Platforms

- Hadoop: 1st widely adopted platform by industry; popularised MapReduce
- Spark: A lot optimisation over Hadoop to reach hundreds times acceleration; current de facto standard
- GraphLab: Cutting edge framework to implement Machine Learning algorithms; New programming concept -- Vertex Program
- ~~Storm: widely adopted Streaming platform~~

Agenda of the GraphLab Hands-on Session

- Overview
- Some examples
  - Recommender
  - Pagerank
- Hands-on time

Mis-conception

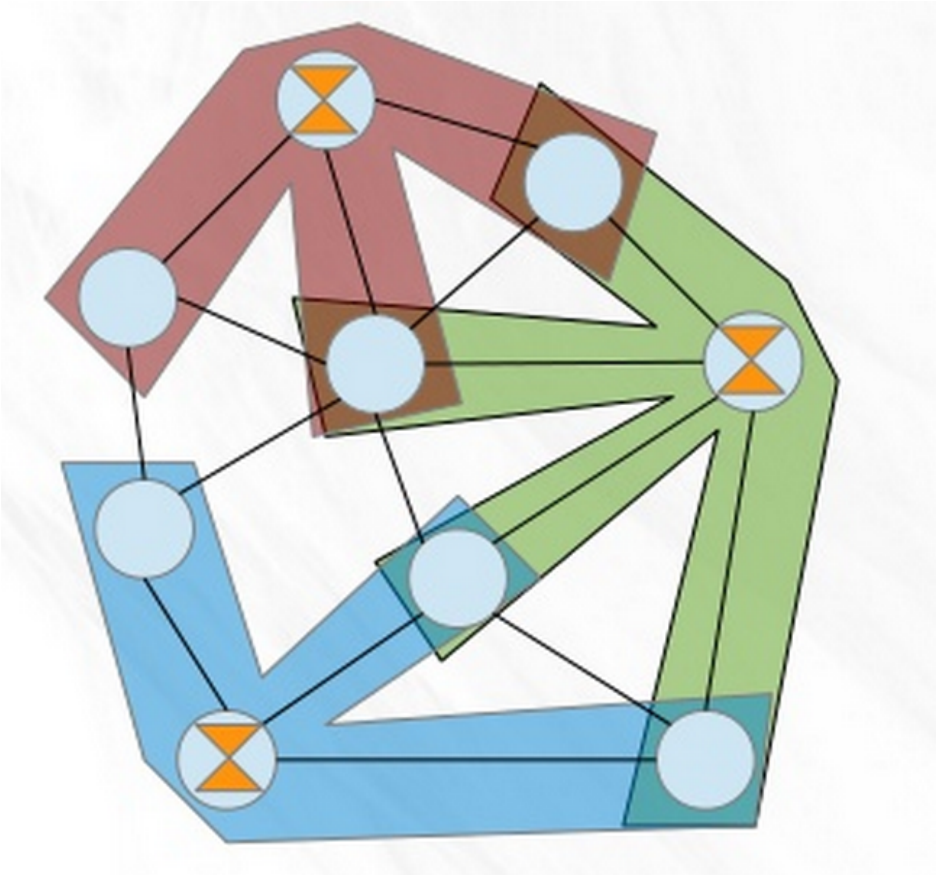
Wrong: GraphLab is to process graph

(it can; but not designed for...)

Correct: GraphLab models computation flow over a graph structure

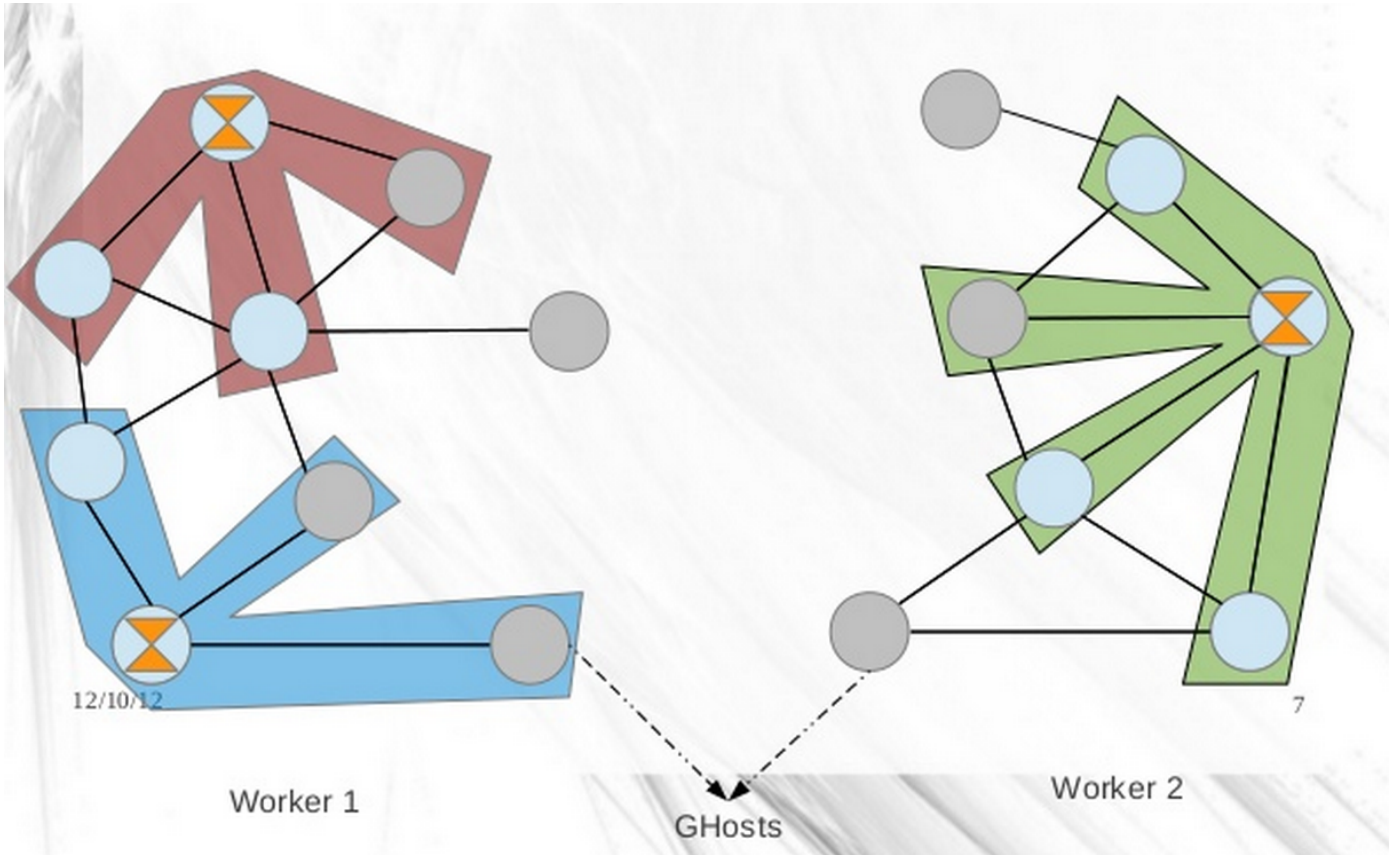
(your problem might not look like a graph but computation flow might be abstracted as a graph -- Recommender System)

Computation Flow on Graph



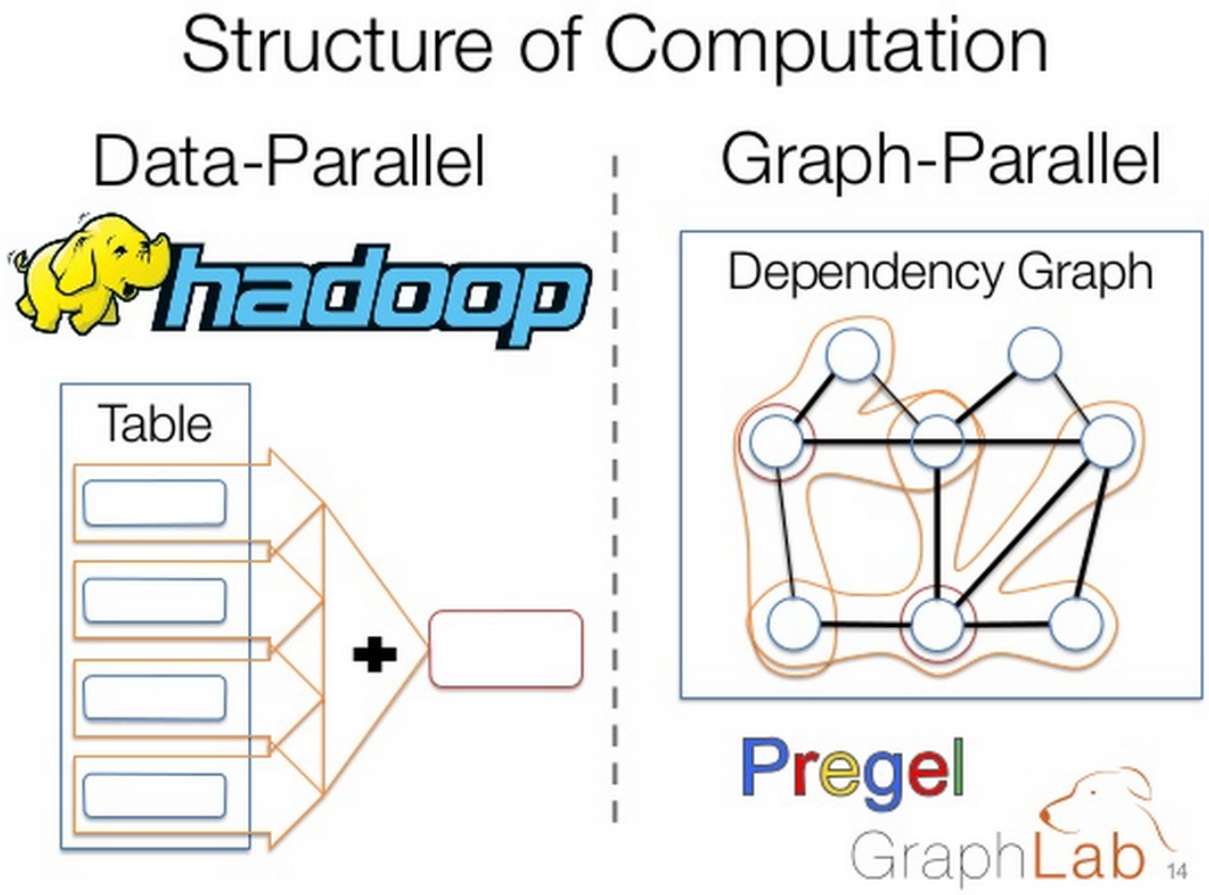
http://www.slideshare.net/Zuhairkhayyat/graphlab-under-thehood

Computation Flow on Graph



http://www.slideshare.net/Zuhairkhayyat/graphlab-under-thehood

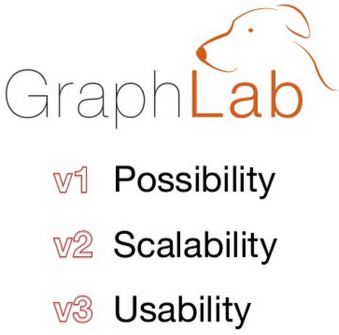
Comparison to Hadoop

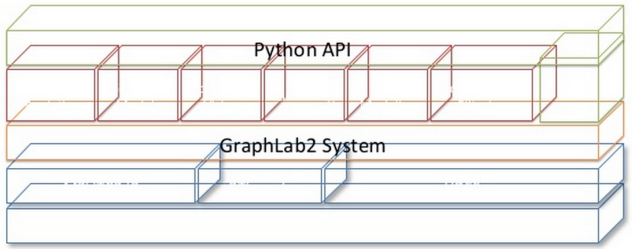


http://www.slideshare.net/SessionsEvents/joey-gonzalez-graph-lab-m-lconf-2013

GraphLab

- GraphLab 1.0:
  - Initial model
  - C++ API
- GraphLab 2.0 (PowerGraph):
  - Deal with skewed data
  - GAS: Gather -- Apply -- Scatter
- GraphChi:
  - Disk based version
- GraphLab Create:
  - Rich ML libs; Python integration
  - A startup effort





Data Structures

- graphlab.SArray → pandas.Series
- graphlab.SFrame → pandas.DataFrame
- graphlab.SGraph

S: Server-side

Algorithms on Graphs

Most are already in GraphLab Create Lib

- **Collaborative Filtering**
  - Alternating Least Squares
  - Stochastic Gradient Descent
  - Tensor Factorization
  - SVD
- **Structured Prediction**
  - Loopy Belief Propagation
  - Max-Product Linear Programs
  - Gibbs Sampling
- **Semi-supervised ML**
  - Graph SSL
  - CoEM
- **Graph Analytics**
  - PageRank
  - Shortest Path
  - Triangle-Counting
  - Graph Coloring
  - K-core Decomposition
  - Personalized PageRank
- **Classification**
  - Neural Networks
  - Lasso

...

http://www.slideshare.net/SessionsEvents/joey-gonzalez-graph-lab-m-lconf-2013

General Usage

- Represent your data in SXXX structure
- Pick algorithm and run it

(where is graph¿)

(well, you don’t see the computation layer, ...)

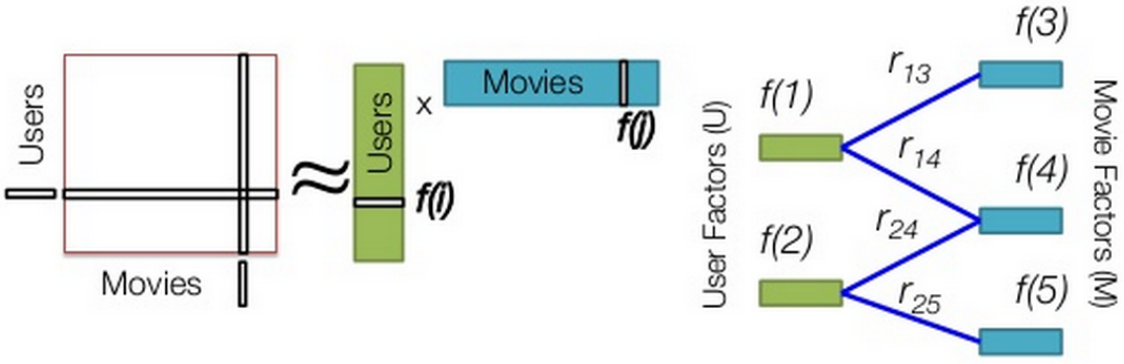
A practical view of GraphLab:

Ready for production/ easy to use Machine learning toolkits

Example: Recommender

Recommending Products

Low-Rank Matrix Factorization:



Iterate:

$$f[i] = \arg\min_{w \in \mathbb{R}^d} \sum_{j \in \mathrm{Nbrs}(i)} \left(r_{ij} - w^T f[j]\right)^2 + \lambda ||w||_2^2$$

8

http://www.slideshare.net/SessionsEvents/joey-gonzalez-graph-lab-m-lconf-2013

Example: Recommender

One example input:

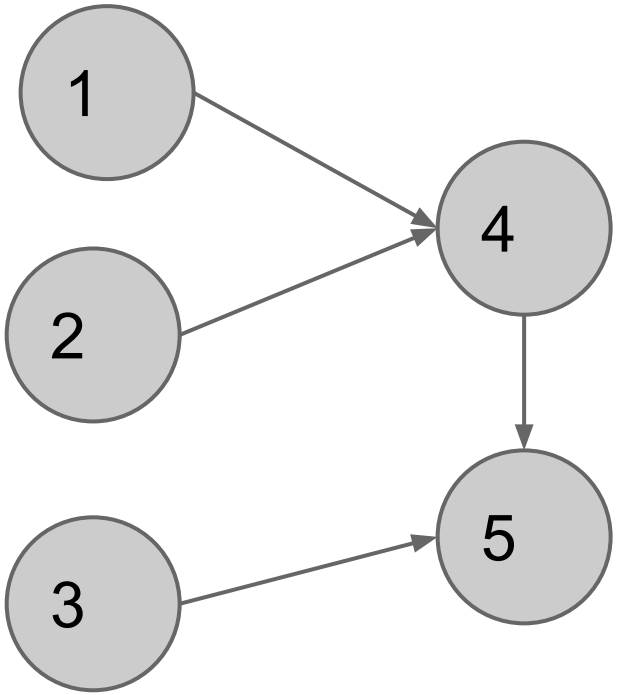
	score				
item_id	**a**	**b**	**c**	**d**	**e**
user_id					
1	1	5	NaN	1	NaN
2	2	5	NaN	NaN	2
3	NaN	1	2	5	5
4	4	NaN	2	5	NaN

Example: Recommender

One example result:

	score				
item_id	**a**	**b**	**c**	**d**	**e**
user_id					
1	1.000000	5.000000	-0.896185	1.000000	1.035765
2	2.000000	5.000000	-0.213599	1.425343	2.000000
3	3.471785	1.000000	2.000000	5.000000	5.000000
4	4.000000	0.199973	2.000000	5.000000	3.736209

Example: Pagerank



Futher

- GraphLab C++ API:
  - 2014 offering of ENGG4030 @ CUHK
  - http://project.hupili.net/engg4030/t11-graphlab/
  - Write Vertex Program
  - Write GAS model
  - Different implementations of PageRank is provided in that tutorial
- Introduction to Graph Analysis in Python
  - April 22, 2015 @ General Assembly
  - https://drive.google.com/folderview?id=0B8i0lKkzNhjsaFo2bkdtam52NGs&usp=sharing

Q/A & Hands-on

Contact me:

http://hupili.net

bit.ly/hkosc15-graphlab

WE ARE HIRING