# GraphX

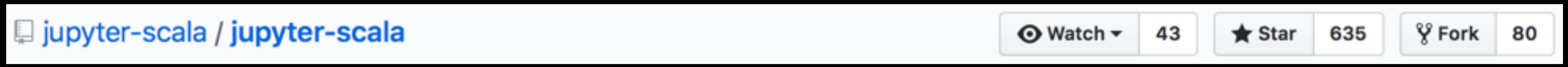
Amber, 2017.10.13

# Agenda

1. Spark + Scala + Jupyter
2. Spark GraphX
3. Hands On

# Spark + Scala + Jupyter

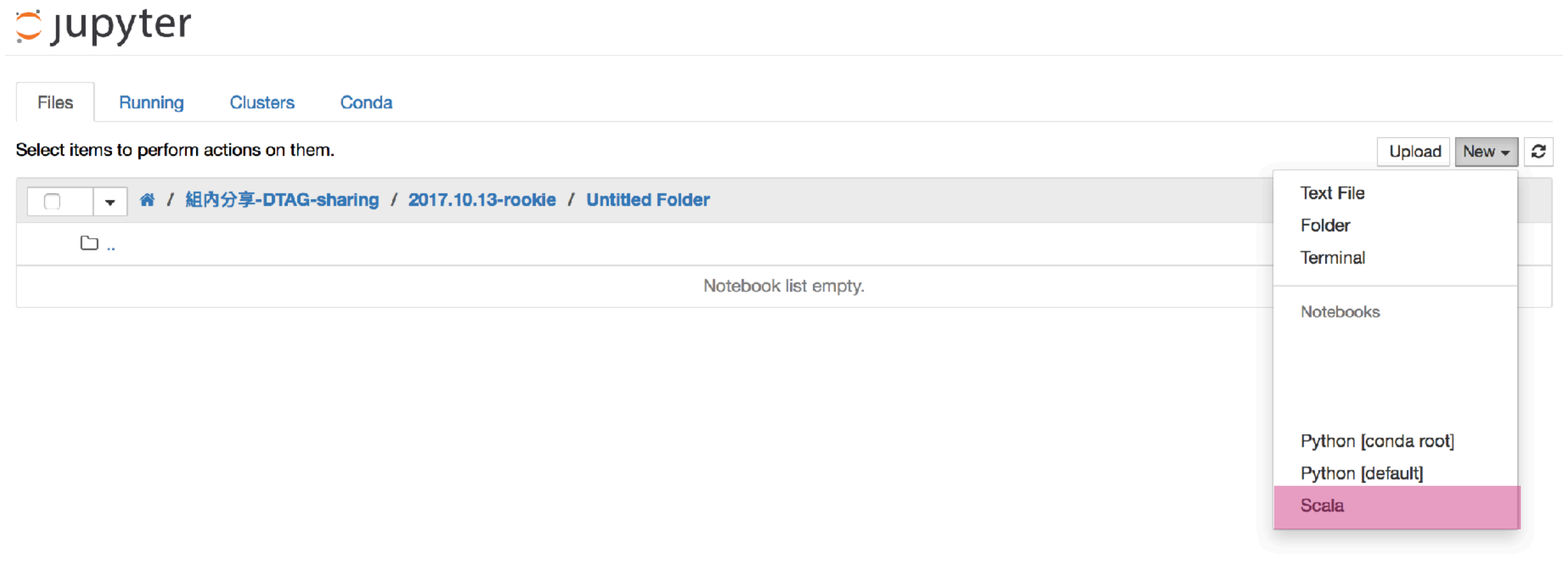
```
$ git clone https://github.com/jupyter-scala/jupyter-scala.git
$ cd jupyter-scala/
$ sh jupyter-scala

Run jupyter console with this kernel with
  jupyter console --kernel scala
Use this kernel from Jupyter notebook, running
  jupyter notebook
and selecting the "Scala" kernel.

$ jupyter kernelspec list

Available kernels:
  python2  /Users/…/anaconda2/lib/python2.7/site-packages/ipykernel/resources
  scala    /Users/…/Library/Jupyter/kernels/scala
```



jupyter-scala / jupyter-scala

# Spark + Scala + Jupyter

```
$  jupyter notebook
```



Select items to perform actions on them.

Notebook list empty.

線上管理頁面：http://localhost:4040/jobs/

# Spark GraphX

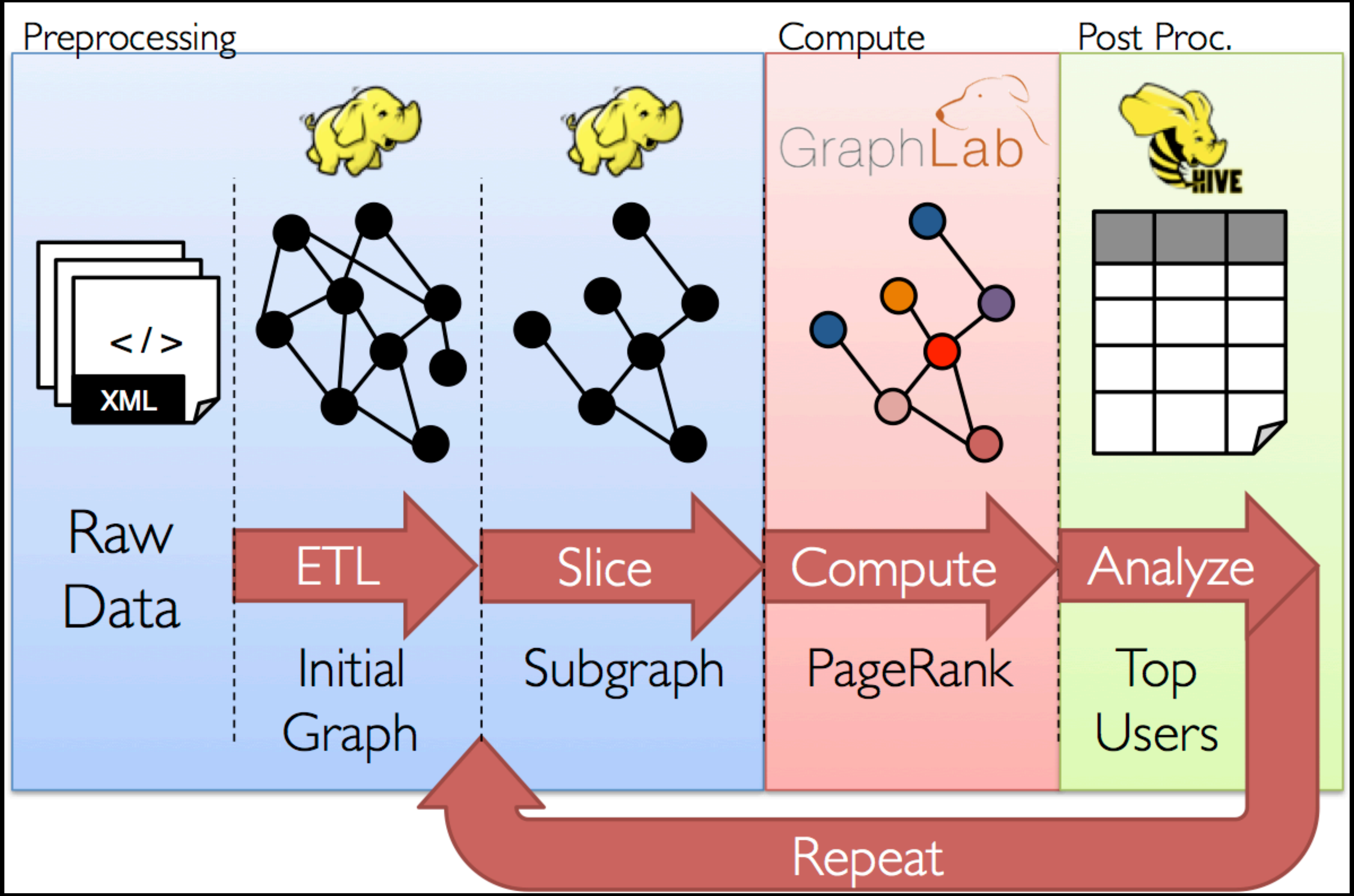
# Spark GraphX 介紹

GraphX 透過引入 Resilient Distributed Property Graph：一種帶有頂點和邊屬性的有向多重圖，來擴展Spark RDD。

為了支援圖形的運算，GraphX 擁有基本運算子和 Pregel API 的優化。

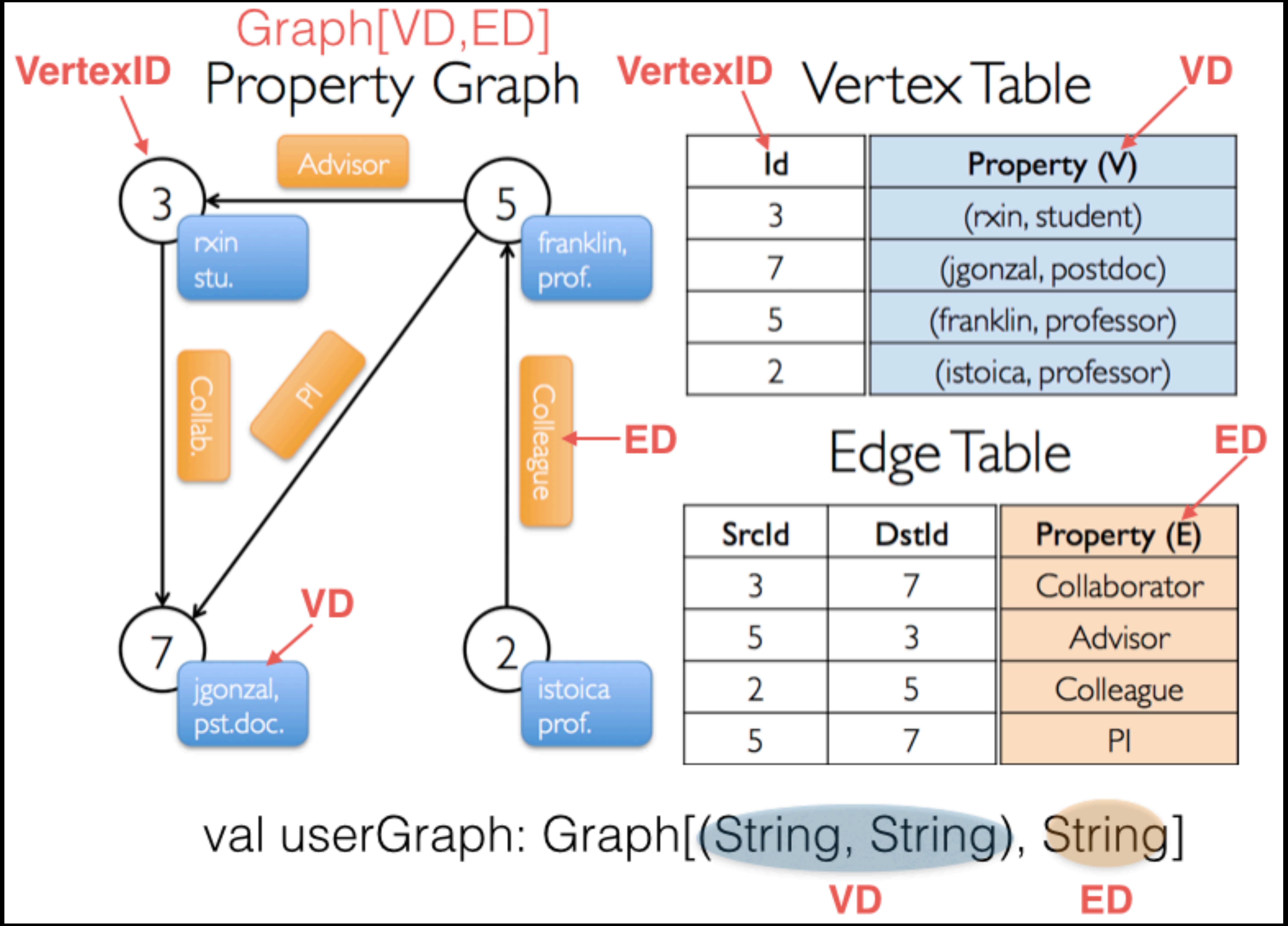
- 例如：subGraph、joinVertices、aggregateMessages

# Spark GraphX 介紹



GraphX允許使用者將資料視為一個圖形和集合（例如：RDDs），而不需要任何的資料搬移和複製。

# Spark GraphX 資料結構



| Id | Property (V) |
|---|---|
| 3 | (rxin, student) |
| 7 | (jgonzal, postdoc) |
| 5 | (franklin, professor) |
| 2 | (istoica, professor) |

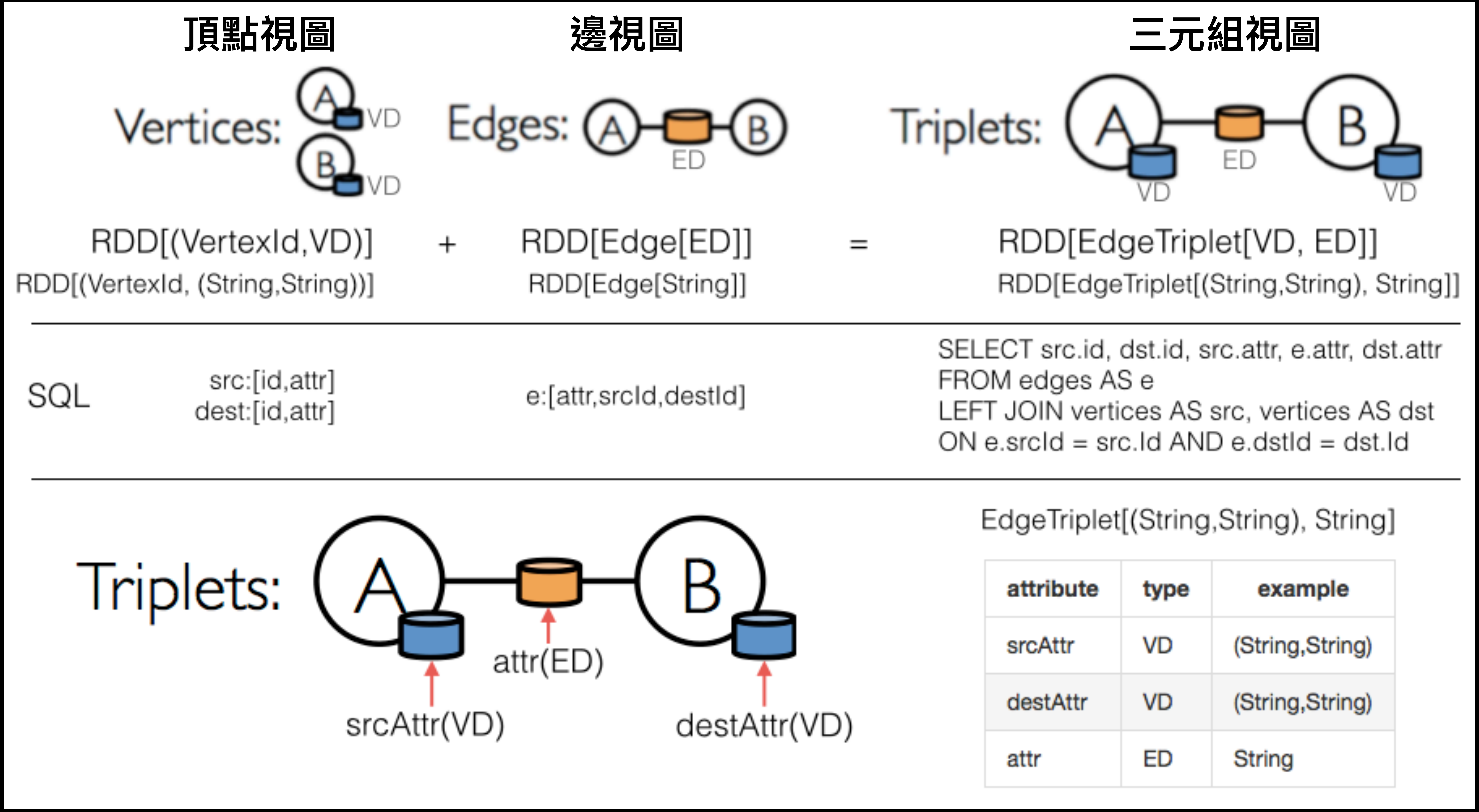
| SrcId | DstId | Property (E) |
|---|---|---|
| 3 | 7 | Collaborator |
| 5 | 3 | Advisor |
| 2 | 5 | Colleague |
| 5 | 7 | PI |

val userGraph: Graph[(String, String), String]

**VertexRDD[VD] = RDD[(VertexId,VD)]**
VertexRDD[(String,String)] = RDD[(VertexId, (String,String))]

**EdgeRDD[ED] = RDD[Edge[ED]]**
EdgeRDD[String] = RDD[Edge[String]]

# Spark GraphX 資料結構



頂點視圖　　邊視圖　　三元組視圖

RDD[(VertexId,VD)] + RDD[Edge[ED]] = RDD[EdgeTriplet[VD, ED]]

RDD[(VertexId, (String,String))]　　RDD[Edge[String]]　　RDD[EdgeTriplet[(String,String), String]]

SQL

src:[id,attr]
dest:[id,attr]

e:[attr,srcId,destId]

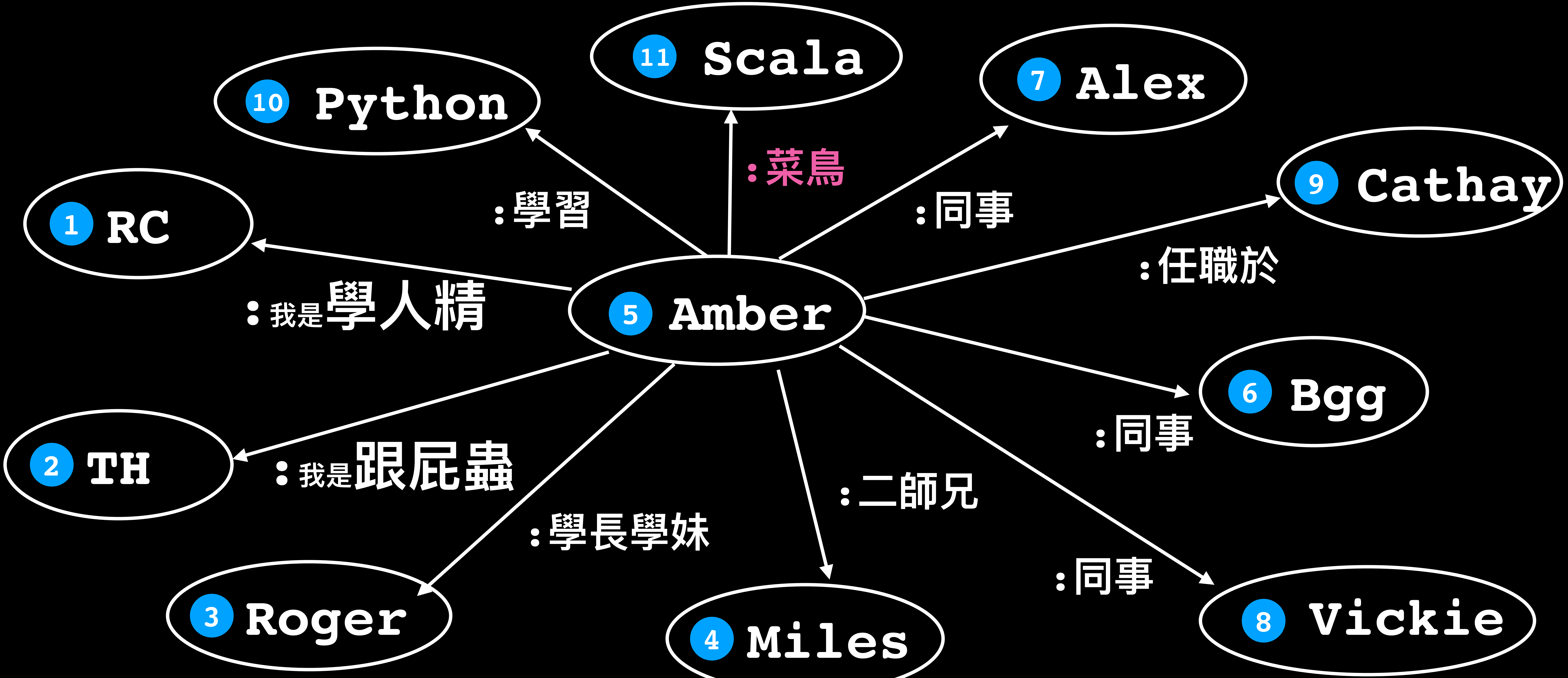
```
SELECT src.id, dst.id, src.attr, e.attr, dst.attr
FROM edges AS e
LEFT JOIN vertices AS src, vertices AS dst
ON e.srcId = src.Id AND e.dstId = dst.Id
```

EdgeTriplet[(String,String), String]

| attribute | type | example |
|---|---|---|
| srcAttr | VD | (String,String) |
| destAttr | VD | (String,String) |
| attr | ED | String |

# 先來一點參與感

# About Amber



11 Scala

10 Python

7 Alex

:菜鳥

9 Cathay

1 RC

:學習

:同事

:任職於

:我是學人精

5 Amber

6 Bgg

:同事

2 TH

:我是跟屁蟲

:二師兄

:學長學妹

:同事

3 Roger

4 Miles

8 Vickie

# 下集待續

# Agenda

1. Pregel API
2. Graph Algorithm
   - PageRank
   - Connected Components
3. Case Study