

Spark GraphFrames

Anurag Nagar, Ph.D.

Big Data Class

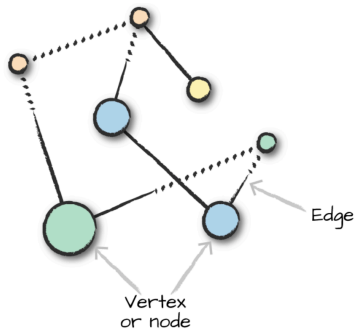
Outline

- 1 Background
 - Installation
- 2 Example 1
- 3 Example 2

Background

GraphFrames

- graph processing library for GraphX
- graph library based on DataFrames
- uniform API for data and graph processing
- easier way to create graphs from tabular data



Installation on Databricks

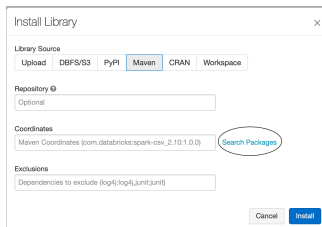
Steps

1. Start Databricks cluster
2. Go to clusters links and hover your mouse over the cluster
3. Click “Libraries” link
and then on the next page, click “Install New” button.

A rectangular button with a light gray background and rounded corners. It contains a small icon of a document with a plus sign, followed by the text "Install New".

Installation on Databricks

4. On the popup window, go to the Maven tab and then click on Search package
5. Search for “GraphFrame”
Spark package and choose the release that matches your cluster configuration.
6. Finally click on Install and your cluster should contain the required jar dependency.



Installation on a SBT project

Install the appropriate version according to your Spark and Scala versions from

<https://mvnrepository.com/artifact/graphframes/graphframes>

or

<https://spark-packages.org/package/graphframes/graphframes>

Outline

- 1 Background
 - Installation
- 2 Example 1
- 3 Example 2

Example

Let's create a DataFrame of nodes

```
import org.graphframes.GraphFrame
val v = sqlContext.createDataFrame(List(
  ("a", "Alice", 34),
  ("b", "Bob", 36),
  ("c", "Charlie", 30),
  ("d", "David", 29),
  ("e", "Esther", 32),
  ("f", "Fanny", 36),
  ("g", "Gabby", 60)
)).toDF("id", "name", "age")
```


Example

Let's create a DataFrame of edges (relationships)

```
val e = sqlContext.createDataFrame(List(  
  ("a", "b", "friend"),  
  ("b", "c", "follow"),  
  ("c", "b", "follow"),  
  ("f", "c", "follow"),  
  ("e", "f", "follow"),  
  ("e", "d", "friend"),  
  ("d", "a", "friend"),  
  ("a", "e", "friend")  
)).toDF("src", "dst", "relationship")
```

Example

Finally, let's create the graph and run some queries.

```
val g = GraphFrame(v, e)

// indegrees of each node
display(g.inDegrees)

// youngest person
val youngest = g.vertices.groupBy().min("age")
display(youngest)
```

Example

Go through the complete example at:

https://graphframes.github.io/graphframes/docs/_site/user-guide.html

Outline

- 1 Background
 - Installation
- 2 Example 1
- 3 Example 2

Example 2

We will work through further examples from Chapter 30 of [Spark: The Definitive Guide](#)

<https://github.com/databricks/Spark-The-Definitive-Guide>