From Graph to Knowledge Graph

Mining Large-scale Heterogeneous Networks Using Spark

KDD 2019 Anchorage

Aug 8th, 2019

Microsoft Research – Microsoft Academic Graph team

Iris SHEN, Charles HUANG, Chieh-Han WU, Anshul KANAKIA

Module I: Welcome and Intro

9:30 am - 10:30 am

WHAT & WHY

- What would be covered
  - Graph and Knowledge Graph basics
  - Microsoft Academic Graph (MAG) case study
- Who are the targeted audiences
  - Interested in Graph and Knowledge Graph
  - Industrial practitioner: data scientist / data analyst / applied researcher
- Who are we
  - Microsoft Academic Team
  - We build the MAG (from scratch) since 2014
  - Researcher / Data scientist / Data Pipeline Architect

Tutorial Overview

Morning (9:30am - 12:00pm) ***Graph***	*Presenters:*
• *Module 1*: Welcome and intro (environment setup + dataset)	*Iris / Charles*
• *Module 2*: Graph basics	*Iris / Chieh-Han*
• *Module 3*: Graph representation learning	*Iris / Charles*

----12:00 – 1:00pm ---- Lunch Break ----

Afternoon (1:00pm – 3:30pm) ***Knowledge Graph (KG)***	
• *Module 4*: KG fundamentals and construction	*Iris / Chieh-Han*
• *Module 5*: KG inference and application	*Iris / Anshul*
• *Module 6*: Summary and looking forward	*Iris*

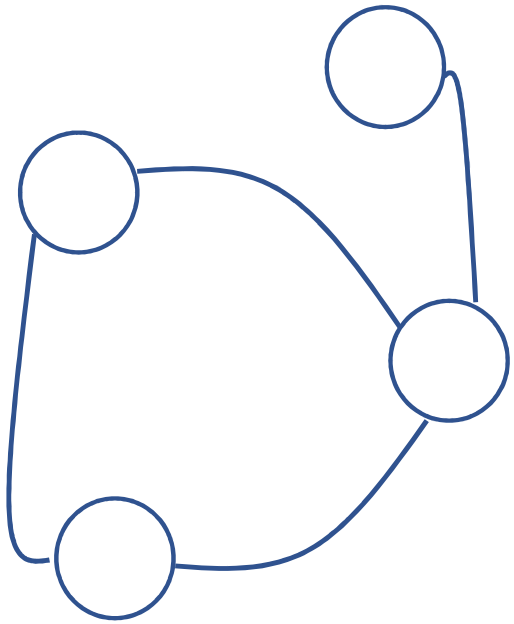
More comprehensive algorithms and theories, see our edX course:
https://www.edx.org/course/from-graph-to-knowledge-graph-algorithms-and-applications

Module I Overview

- Basics
- Environment Setup
- Labs – Dataset understanding

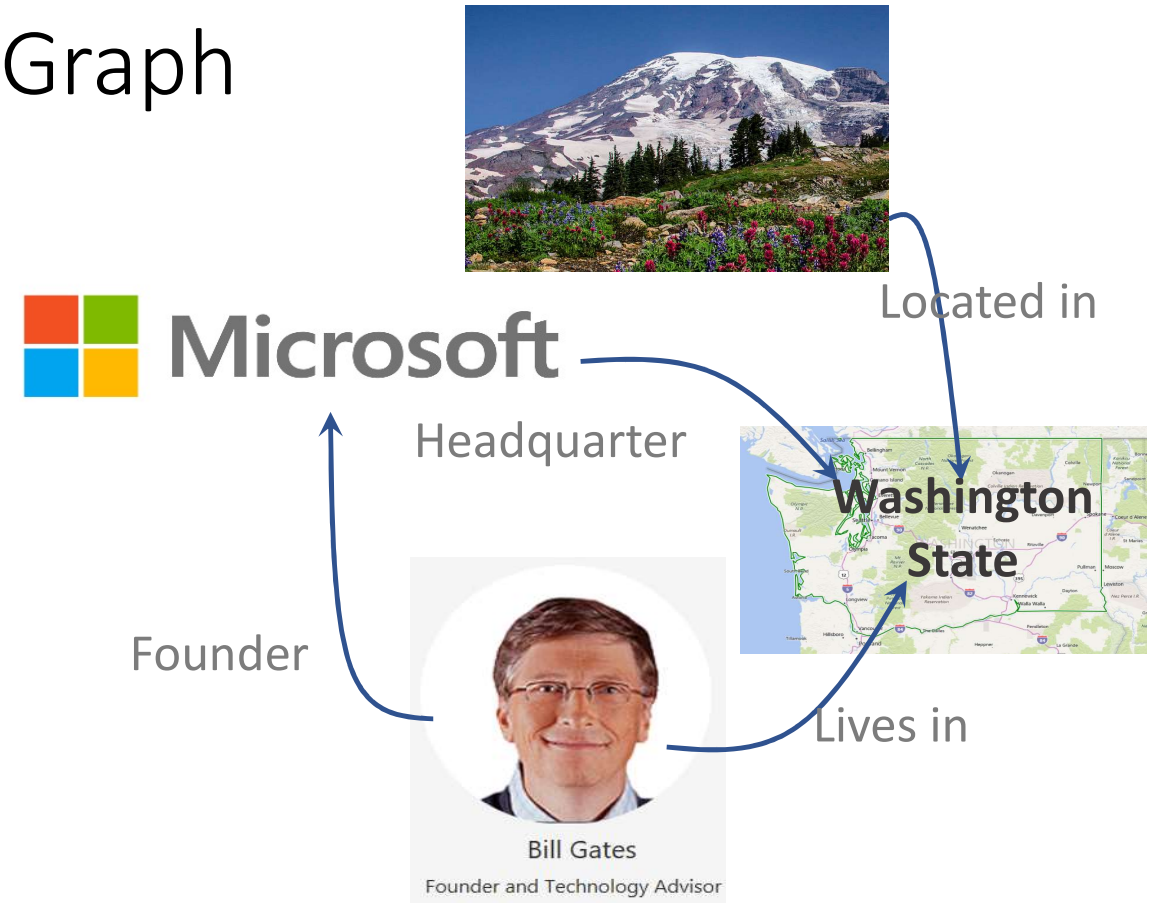
Basics – Graph

- Node
- Edge
- Structure



Basics – Knowledge Graph

- Node - entity
- Edge - relation
- Structure - semantics



Basics – Spark and Databricks

- Apache Spark
  - An open-source distributed general-purpose cluster-computing framework
  - Provides an interface for programming entire clusters with implicit data parallelism and fault tolerance
- Databricks
  - A company founded by the original creators of Apache Spark
  - Develops a web-based platform for working with Spark, that provides automated cluster management and IPython-style notebooks
- Azure Databricks
  - Fast, easy, and collaborative Apache Spark–based analytics service

Getting Microsoft Academic Graph (Full Graph)

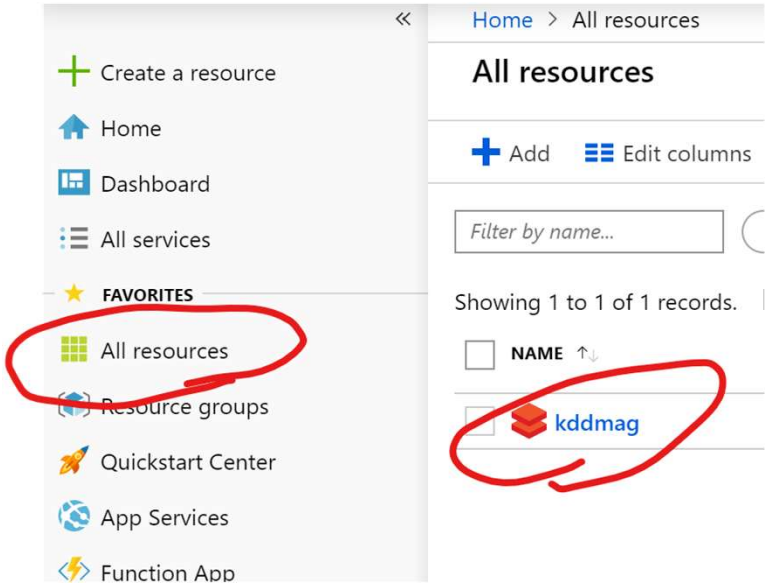
- Microsoft Academic Graph Documentation (aka.ms/mag)
- Get Microsoft Academic Graph
  - Create Azure subscription
  - Create Azure storage account
  - Follow instruction to submit request to get MAG
- Set up Azure Databricks
  - Creates Azure Databricks service
  - Create Spark Cluster

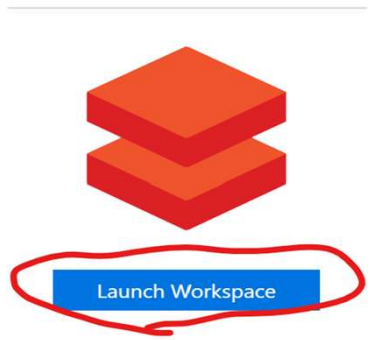
Environment Setup (1)

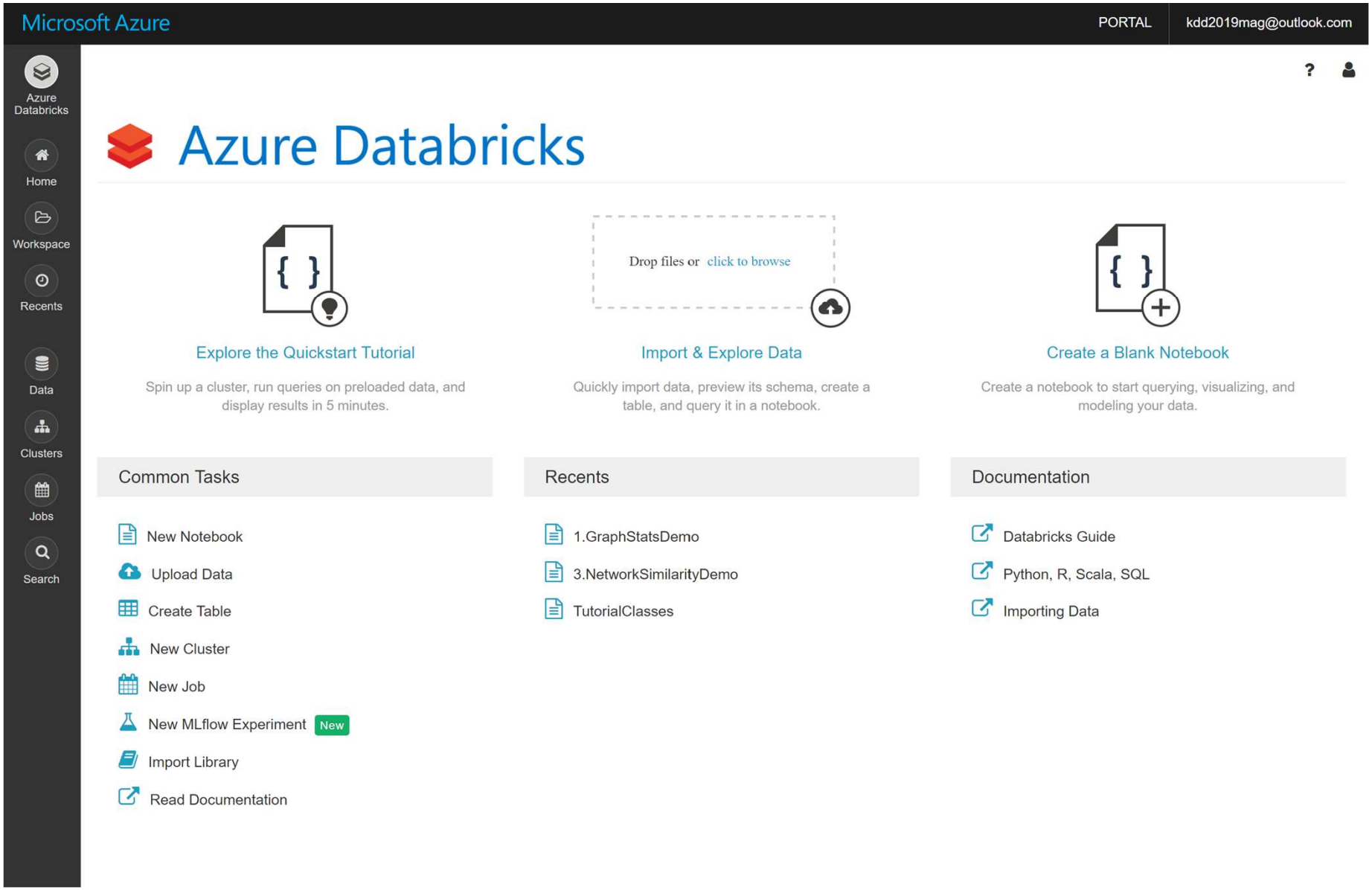
- Tutorial sub-graph
  - No need to submit request
  - Only available during KDD tutorial
- Sign in at the table near entrance for accessing tutorial Databricks
- Go to Azure portal https://portal.azure.com/ and login using the same email

Environment Setup (2)

- Launch Databricks workspace



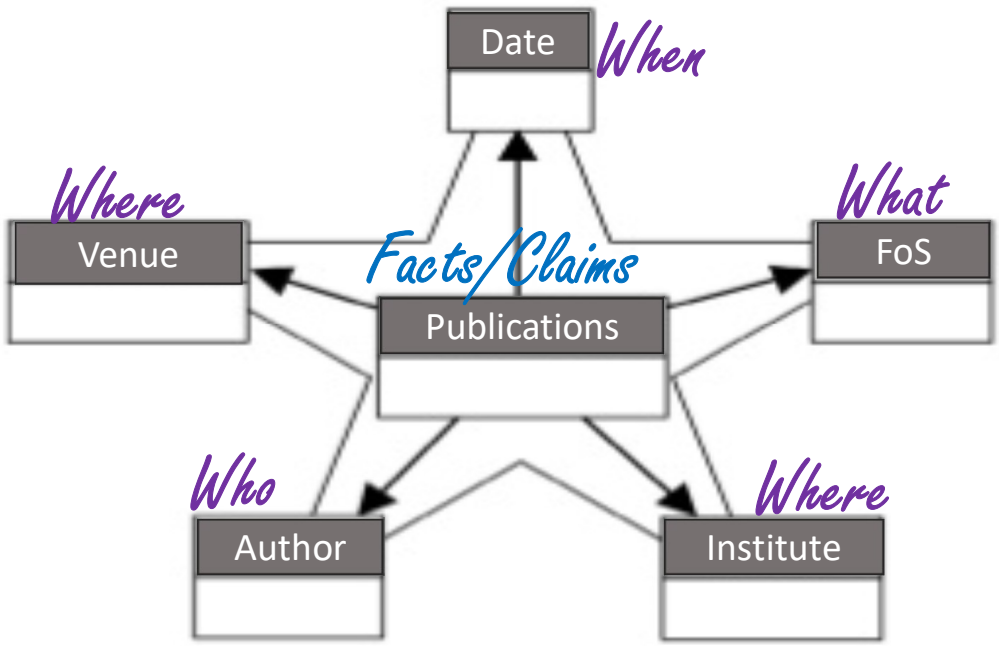


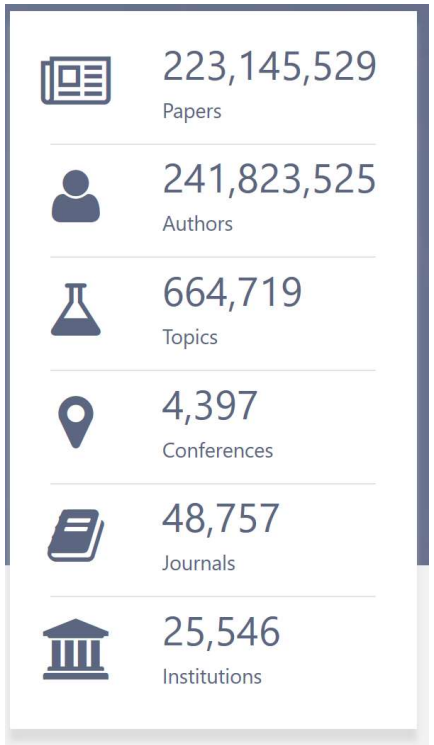

Microsoft Azure
PORTAL
kdd2019mag@outlook.com
Azure Databricks
Home
Workspace
Recents
Data
Clusters
Jobs
Search
Azure Databricks
Explore the Quickstart Tutorial
Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.
Drop files or click to browse
Import & Explore Data
Quickly import data, preview its schema, create a table, and query it in a notebook.
Create a Blank Notebook
Create a notebook to start querying, visualizing, and modeling your data.
Common Tasks
New Notebook
Upload Data
Create Table
New Cluster
New Job
New MLflow Experiment
New
Import Library
Read Documentation
Recents
1.GraphStatsDemo
3.NetworkSimilarityDemo
TutorialClasses
Documentation
Databricks Guide
Python, R, Scala, SQL
Importing Data

Lab 1: Get MAG & Basic Stats

- MAG Schema
- Understand MAG entities and relations
- Top CS conference sub-graph

MAG Schema





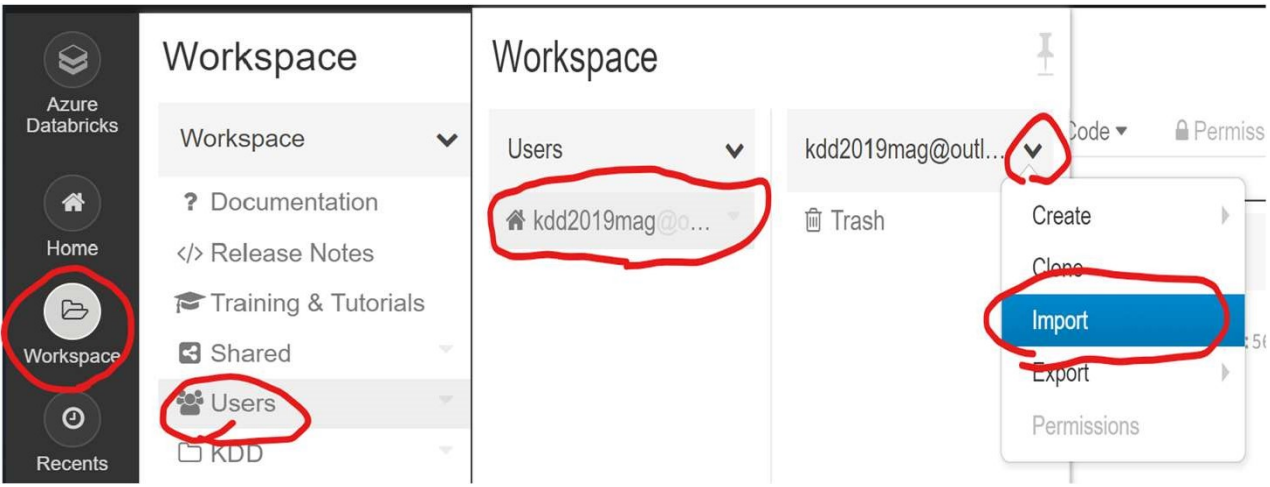
Full schema documentation: https://docs.microsoft.com/en-us/academic-services/graph/reference-data-schema

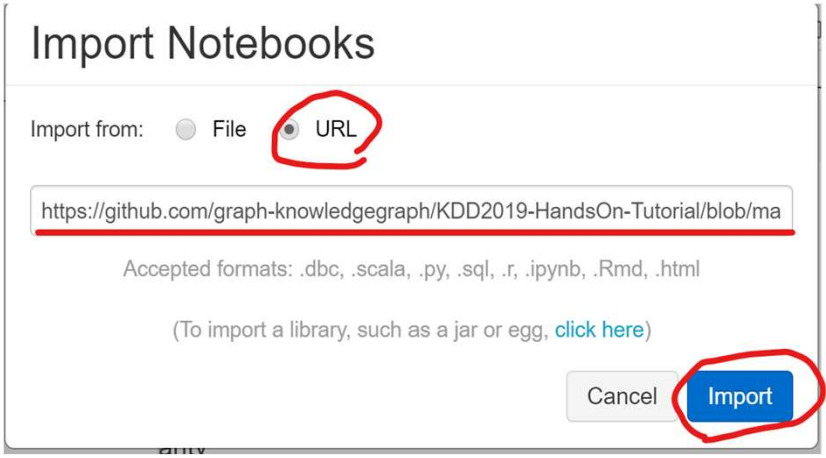
MAG CS Top conferences subgraph

- 103 selected top tier CS conferences
- Direct linked Papers / Authors / Affiliations
- All fields-of-study and the taxonomy

Lab 1: Get MAG & Basic Stats (1)

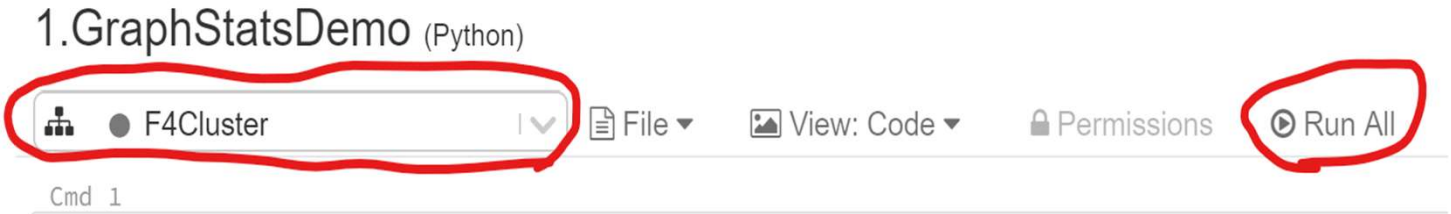
- GitHub Repository
  https://github.com/graph-knowledgegraph/KDD2019-HandsOn-Tutorial
- Import Databricks notebooks
  - TutorialClasses.py (https://github.com/graph-knowledgegraph/KDD2019-HandsOn-Tutorial/blob/master/Module_I/TutorialClasses.py)
  - 1.GraphStatsDemo.py (https://github.com/graph-knowledgegraph/KDD2019-HandsOn-Tutorial/blob/master/Module_I/1.GraphStatsDemo.py)





Lab 1: Get MAG & Basic Stats (2)

- Run 1.GraphStatsDemo




```
Table              |     Count
-------------------+----------
Papers             |    728201
Authors            |    810779
FieldsOfStudy      |    664301
ConferenceSeries   |       103
Journals           |     48753
Affiliations       |      8596
```