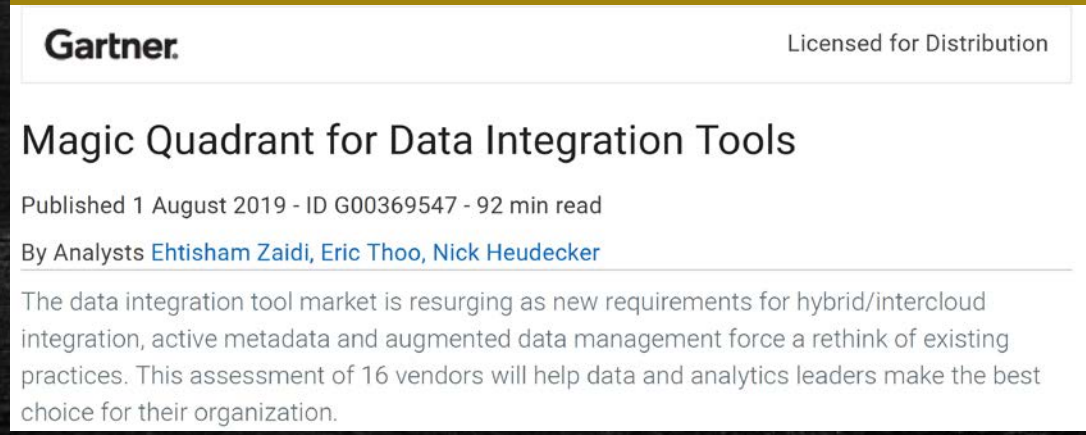
DSE 203 (Fall 2020)

Class Logistics, Assignment and Project

Why This Course?

- By 2021, more than 80% of organizations will ***use more than one data delivery style*** to execute their data integration use cases.
- By 2022, organizations utilizing active metadata to ***dynamically connect, optimize and automate data integration processes*** will reduce time to data delivery by 30%.
- ***By 2022, manual data integration tasks (including recognition of performance and optimization issues across multiple environments) will be reduced by 45% through the addition of ML and automated service-level management.***
- By 2023, improved location-agnostic semantics in data integration tools will reduce design, deployment and administrative costs by 40%.

Gartner, officially known as Gartner, Inc. is a global research and advisory firm providing information, advice, and tools for businesses in IT, finance, HR, customer service and support, legal and compliance, marketing, sales, and supply chain functions.


Gartner

Licensed for Distribution

Magic Quadrant for Data Integration Tools

Published 1 August 2019 - ID G00369547 - 92 min read

By Analysts Ehtisham Zaidi, Eric Thoo, Nick Heudecker

The data integration tool market is resurging as new requirements for hybrid/intercloud integration, active metadata and augmented data management force a rethink of existing practices. This assessment of 16 vendors will help data and analytics leaders make the best choice for their organization.

Instructors

- Instructor:
  - Amarnath Gupta
  - Office: SDSC E-312 (on Zoom)
  - Email: a1gupta@ucsd.edu
- TAs
  - Beidan Huang b5huang@eng.ucsd.edu
  - Pushpak Gautam p1gautam@ucsd.edu
- Canvas
  - https://canvas.ucsd.edu/courses/18944
- Prulu
  - https://prulu.com/class/5489/dse-203-section-14088/

Syllabus

- What is data integration?
- The problem of value matching
- The problem of entity matching
- The problem of cross-model data integration
- Data Integration architectures and data warehousing
- Modern trends in data integration
  - Data lakes
  - Knowledge-based integration
  - Polystores

Several of these problems need a combination of data management and machine learning solutions

Evaluation: Assignments

- No mid-term and final exam
- 3 Programming Assignments (20 X 3 = 60 points)
- Assignment 1
  - Using data sets given to you and a library of different distance functions
    - Find the best value matching method
    - Implement a similarity join method
  - Assignment 2
    - Given a data set and an entity resolution library
    - Solve the same problem using an active learning library
  - Assignment 3
    - Given a JSON data set and a relational data set
    - Implement a workflow that will create a combined JSON data

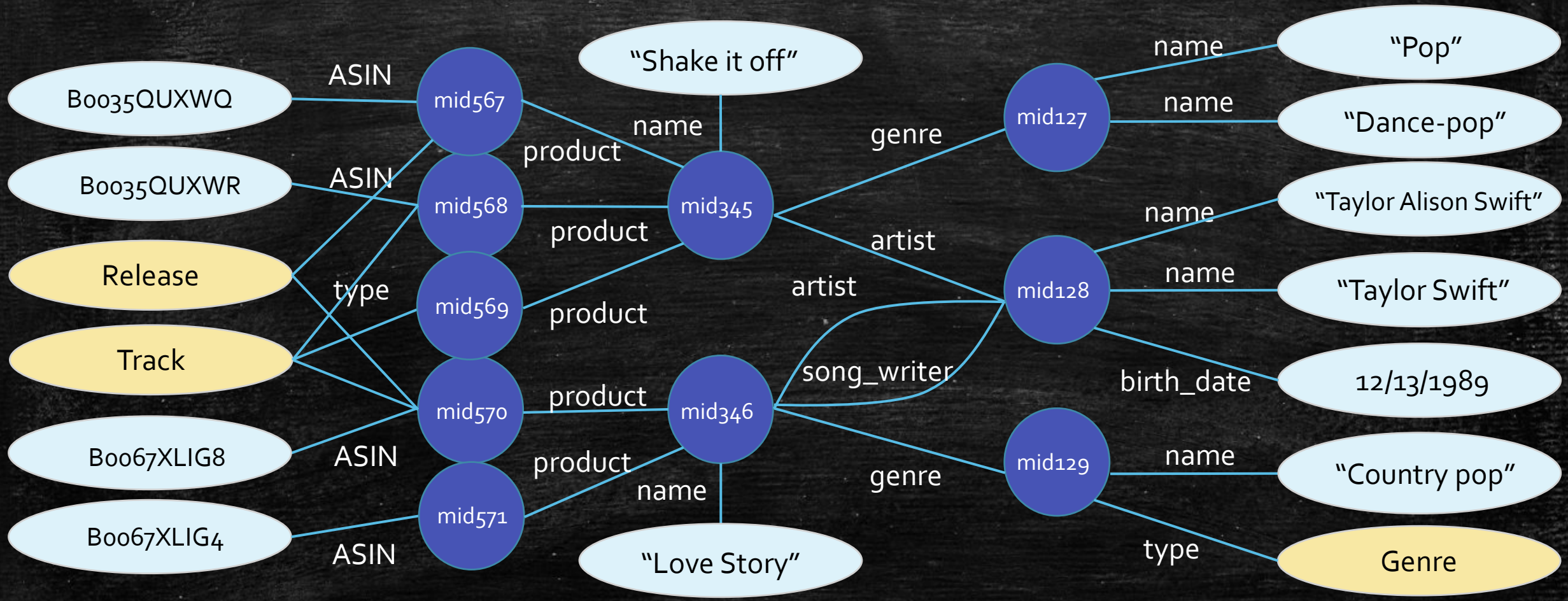
Evaluation: Project

- Goal:
  - Integrating structured, semistructured and unstructured data (40 points)
- Groups of 3 (9 groups)
- Each group will
  - Get 3 different data sets and a taxonomy from instructors
  - Create a knowledge graph
  - Load the graph in Neo4J
  - Answer a set of queries on the knowledge graph
  - Some of the queries would have analytical functions
- Deliverables
  - Project presentation with a functional demo
  - The steps of integration as Jupyter Notebook
  - The final Neo4J graph as a zipped file

Last Year's Project

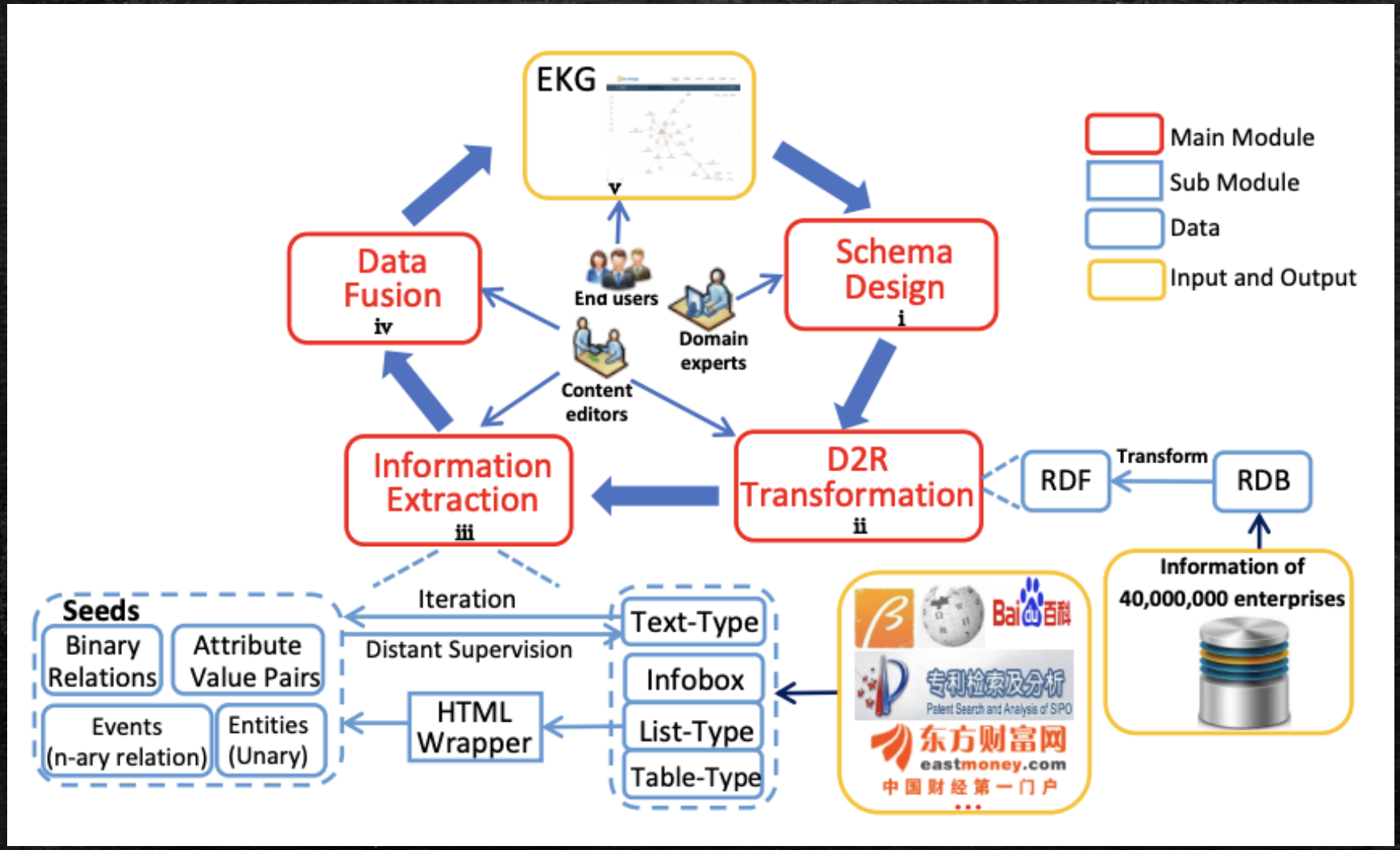
- Each group chose a company from Wikipedia
- For each company they chose three collaborator/partner/subsidiary company
- Using Wikipedia, they mined text to gather information from them
- With this information and information other sources they produced, they created a knowledge graph to answer queries like
  - Find pairs of companies that compete in some areas and cooperate in other areas. Find these areas.
  - Which companies have acquired new companies to start a new product or service line?

Product Graph Example for 2 Songs (Amazon.com)



Luna Dong, Amazon

Building a Knowledge Graph



Read This

https://towardsdatascience.com/a-practical-guide-to-build-an-enterprise-knowledge-graph-for-investment-analysis-3a15363098b7

Your Tasks

- Form your project group and inform the TAs within 1 week
- Get into a groupwise meeting with me to (Week 2)
  - Get your data sets/ pointer to data sets
  - Define the knowledge graph specs for your design
- Every week
  - Some groups meet with me/TAs to discuss progress/difficulties

This Class Project is a precursor for your Capstone Project

This is a Practical "Design-A-Solution" Class

- This class has a "real world" nature
  - We mix theory and practice
  - It combines techniques learned in other classes
  - We don't use "toy data"
- For each topic we discuss, you will find
  - There are many existing algorithms and techniques
  - None of them fits all applications
  - There is often no concrete formula by which you will get the "correct" solution
  - You have to experiment with multiple techniques
  - You may have to combine multiple techniques to get good results

For each class, you should be able to add one or more lines to your CV