American Colleges and Economic Localities

Joe Downs and Reid Kay

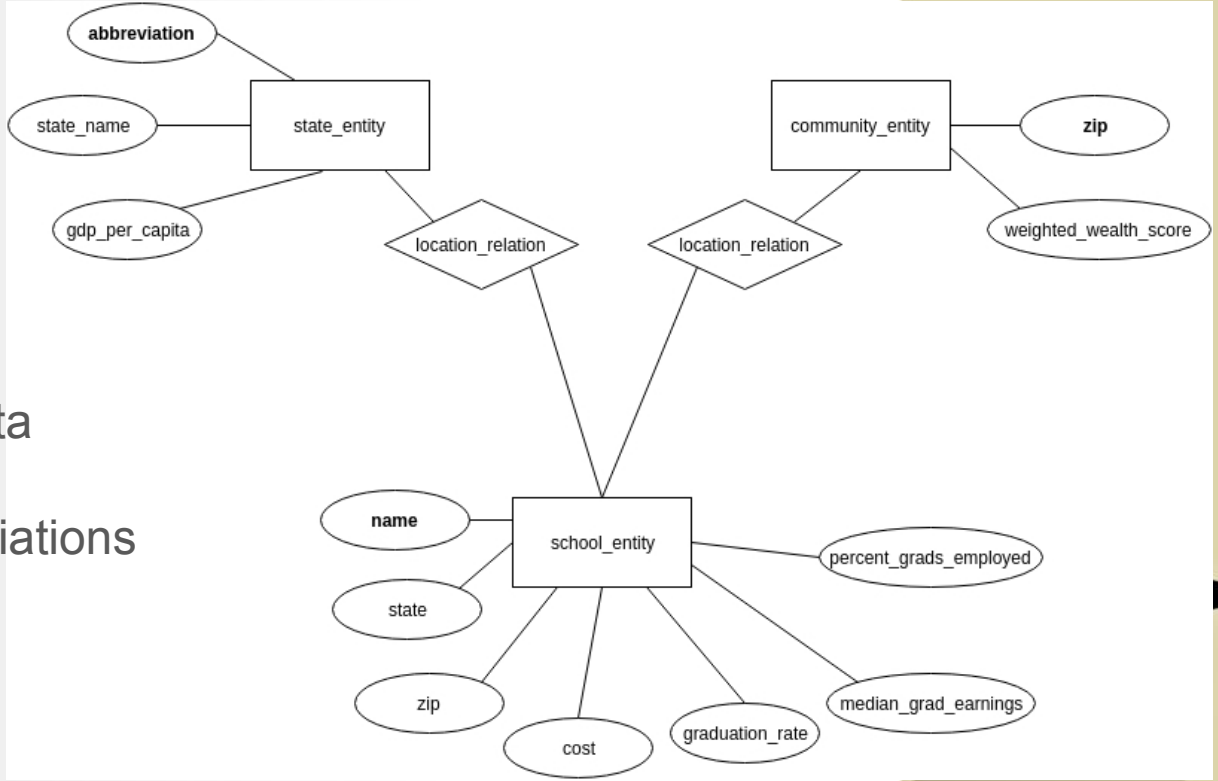
Problem Overview

- Colleges costs vary widely
- College outcomes vary widely:
  - Earnings after graduation
  - Graduation rate
  - Employment after graduation
- Economic factors vary across the countries
- Objective:
  - Rank schools based on outcomes
  - Rank schools based on outcomes per cost
  - Identify relationships between outcomes and economic localities

Data Sources & Entity Relationships

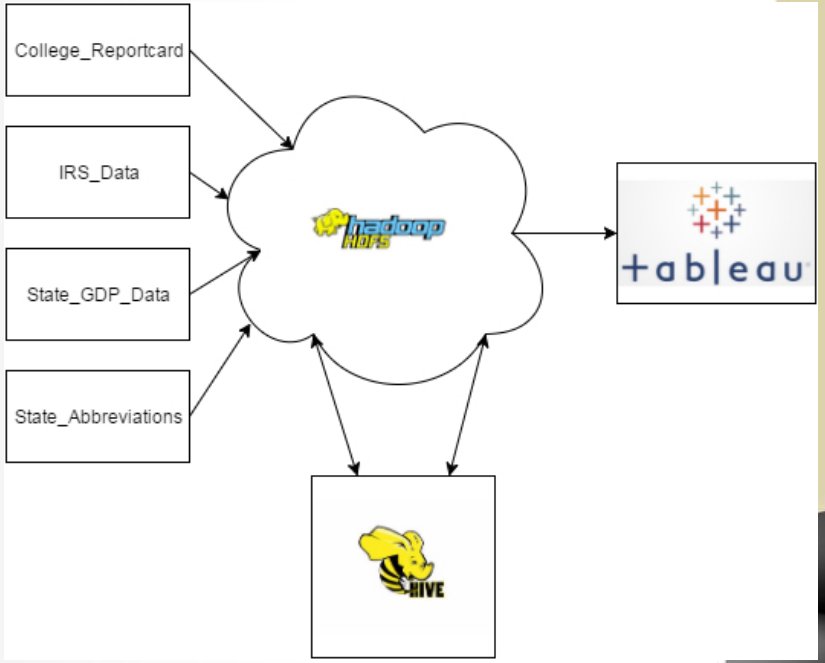
Data Sources:

- College Scorecard
- IRS Zipcode Data
- BEA State GDP Data
- State Name/Abbreviations



Architecture and Technical Implementation Details

- HDFS data-lake
- Hosted on AWS
- Hive ETL Engine
- Tableau Presentation

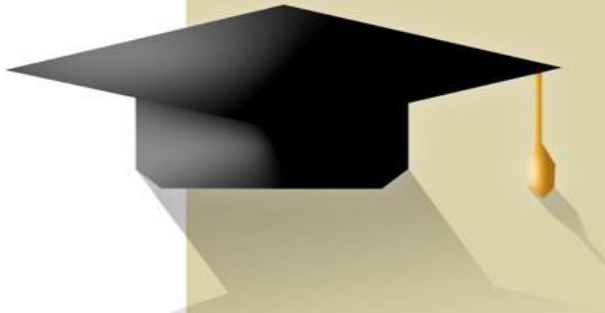


Technical Challenges

- Establishing workflow
  - Share results and processing
  - No single, stable server
- Data Ingestion
  - Automation was tricky
  - Some HTML forms were hard to send through cURL
  - Initial data set had thousands of columns, required script to ingest
- ETL
  - Good ER diagram tricky to develop
  - Some zip codes span states, which required altering the diagram

Algorithms

- Mostly leveraged existing technologies
- Hive
  - Joins
  - Ranking
  - Correlation
- Tableau
  - Geographic association
  - Filtering and sorting
  - Simple aggregations



Results

- Graduation correlation switches sign when factoring in cost
  - Student incentive?
  - School incentive?

	Without Cost	With Cost
Educational Metrics Correlate with Region Wealth?	Positive correlation for all metrics	Low-to-negative correlation for most metrics
Best Institution Types	Universities and Nursing Programs	Trade Schools and Community Colleges

Limitations

- Not all data automatically ingested
- Summarizing numerous data for visual display
  - Mean
  - Median
- Low data velocity and old batches
- Correlational study cannot identify causality