BENCHMARK BIG DATA SYSTEMS ON COMPLEX ANALYTIC QUERIES

Shumo Chu, Edward Wu and Xiaoyi Zhang

WHY

- Big data systems can handle the big volume of data if the query is not complex.
    - How do they perform against complex queries?

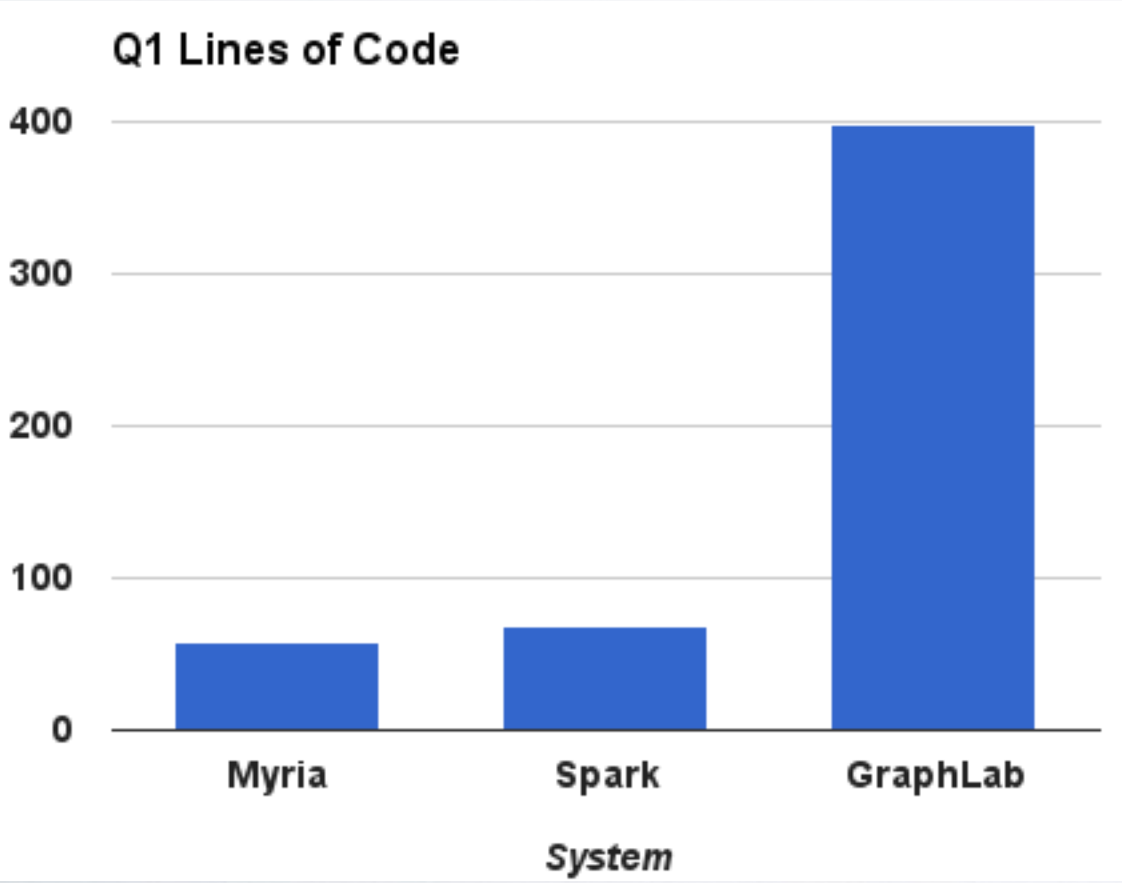
Myria Spark GraphLab

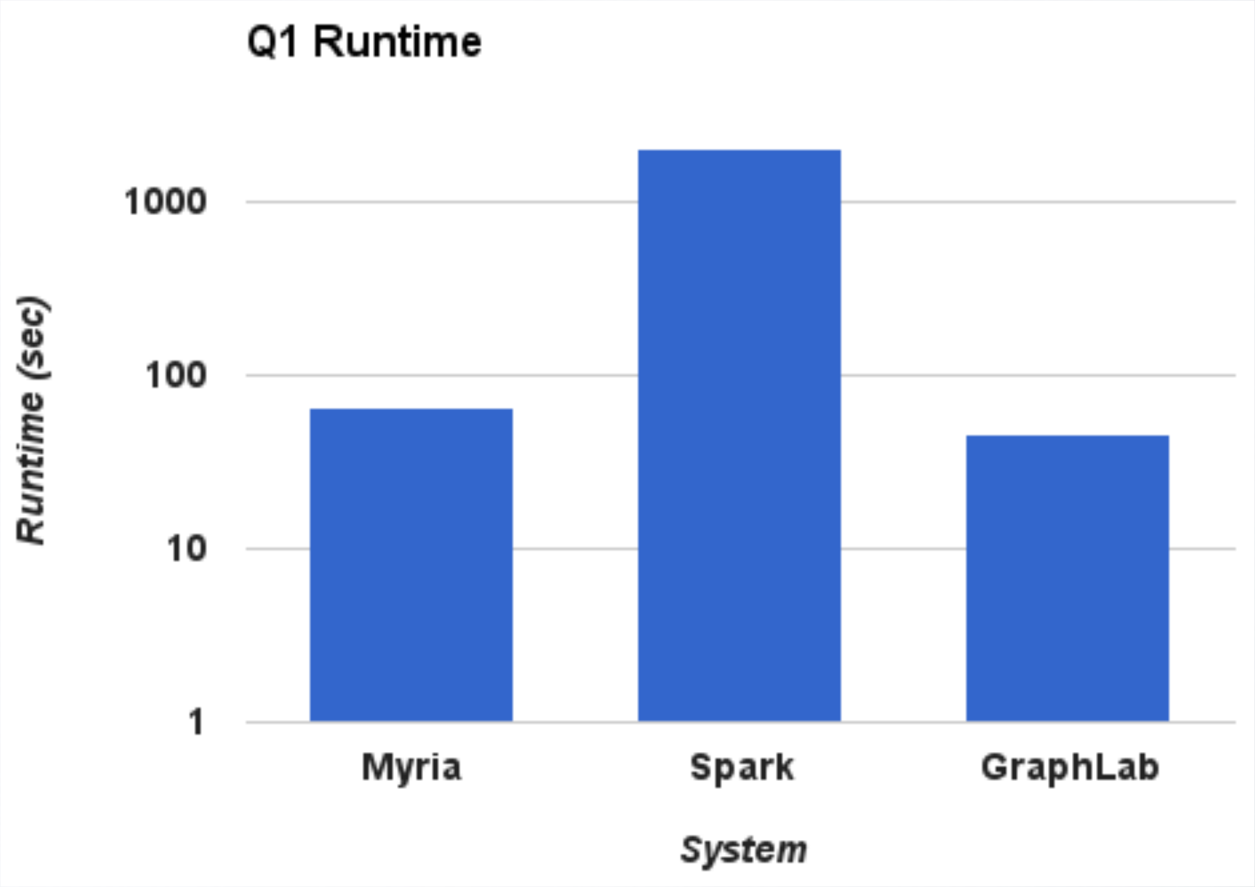
WHAT&HOW TO COMPARE

- Define "complex" query
  - Iterative
  - Aggregation and Filtering
  - Multiple data sources
- Evaluation Metrics
  - Lines of code
  - Runtime

BENCHMARK QUERIES

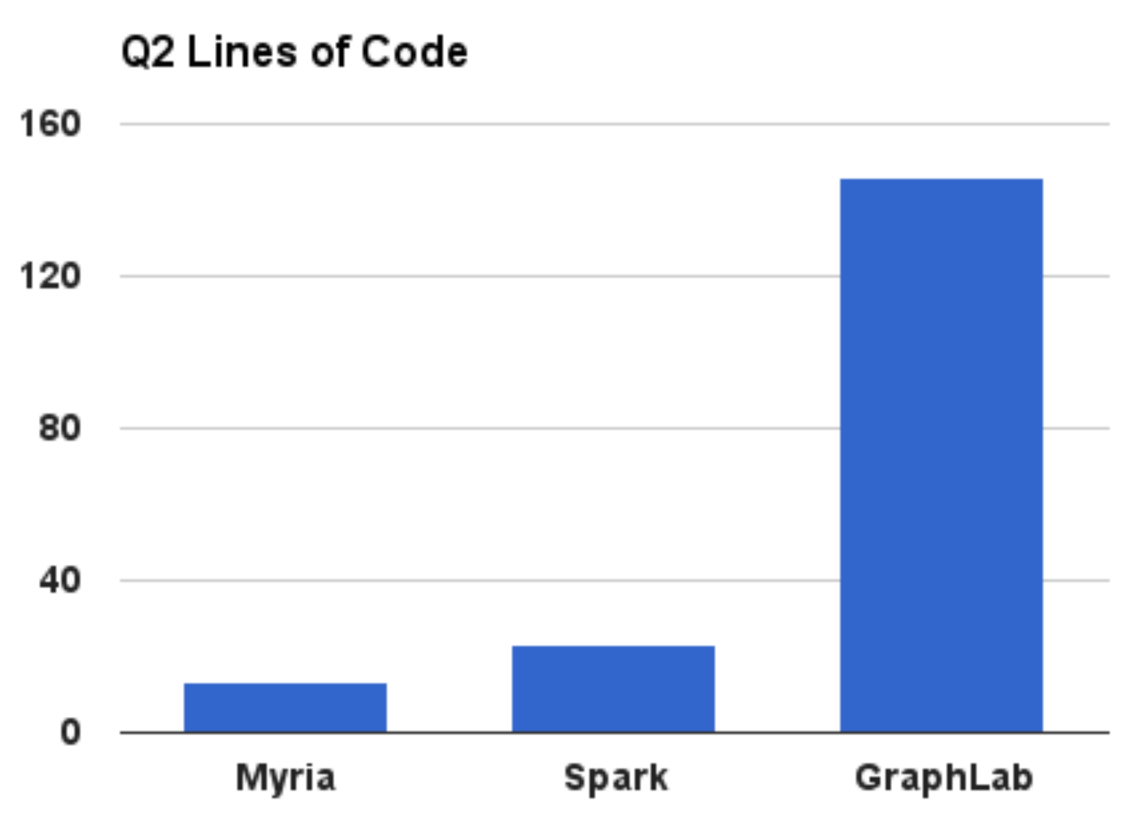
- Query 1: Compute the least common ancestor (LCA) of two academic papers

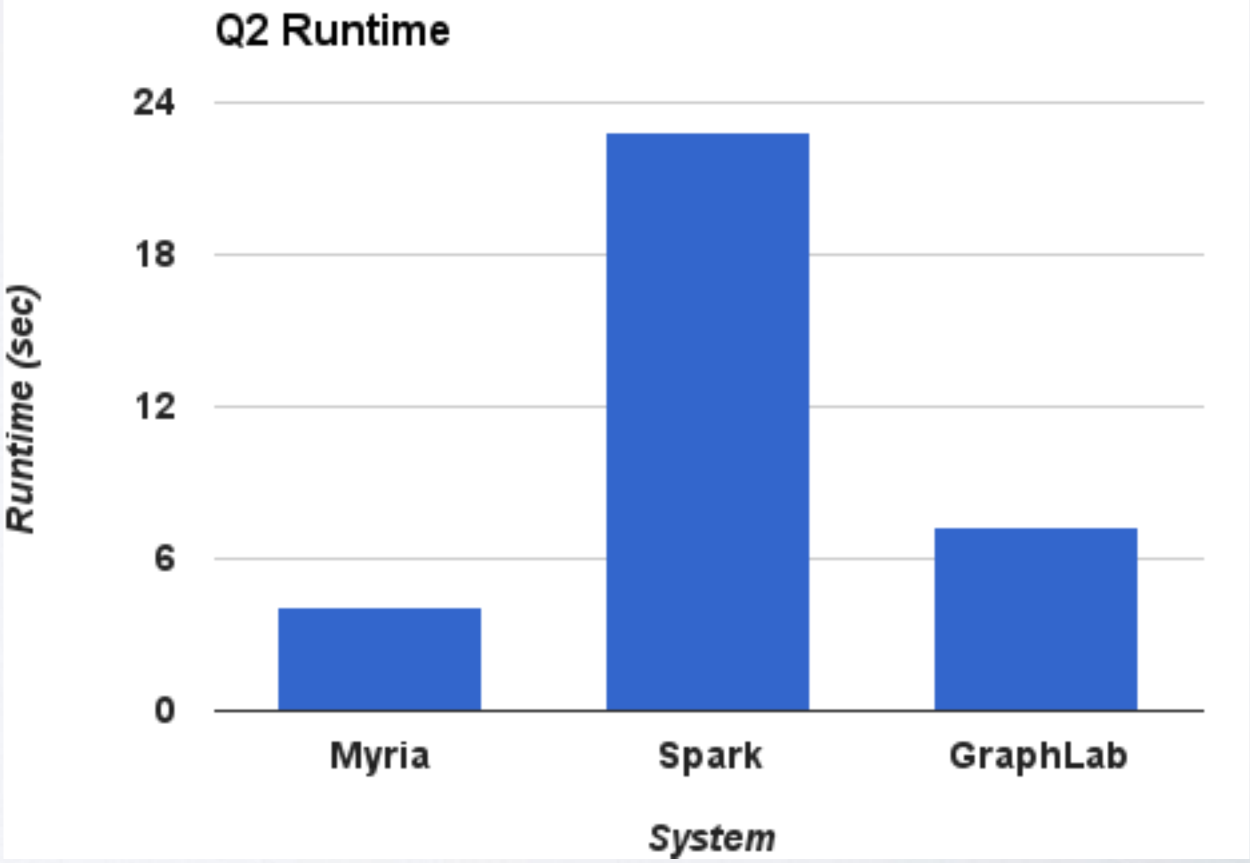




BENCHMARK QUERIES

- Query 2: Compute the k-core (or k-degenerate graph) of an undirected graph





BENCHMARK QUERIES

- Query 3: Compute the merger tree, a hierarchical assembly of galaxies, by tracking the merging of small galaxies

COMPARISON

	Myria	Spark	GraphLab
Pros	Great runtime; Less line of code	Well matured system and eco-system	
Cons	Data ingestion process	Performance; No automatic control of parallelism of RDD	