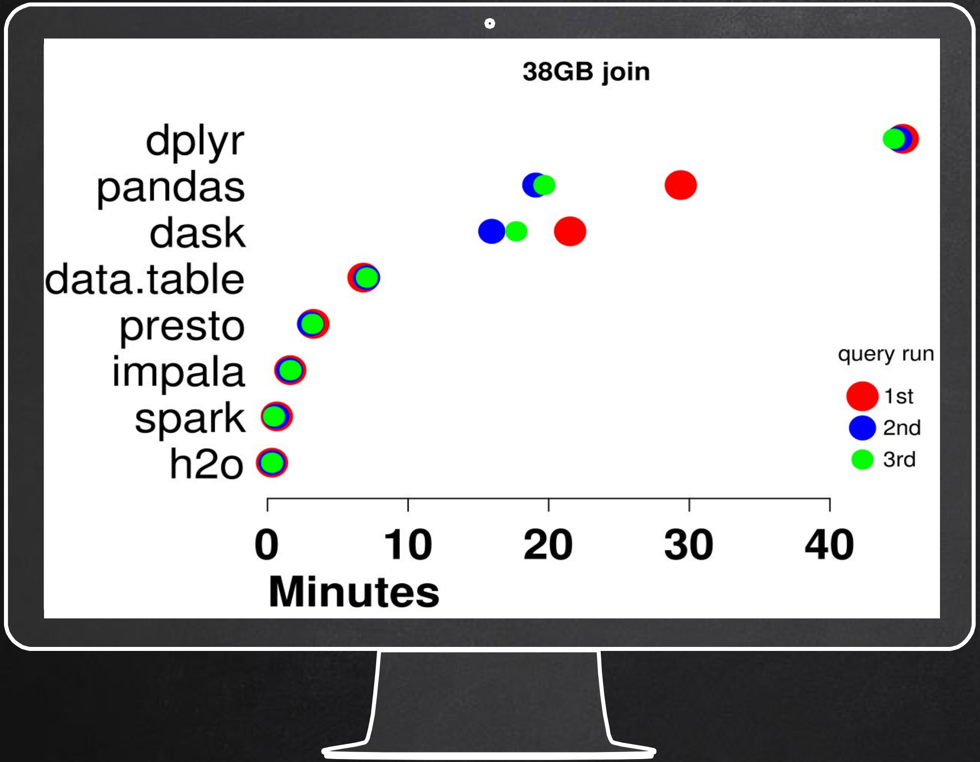
Data Munging

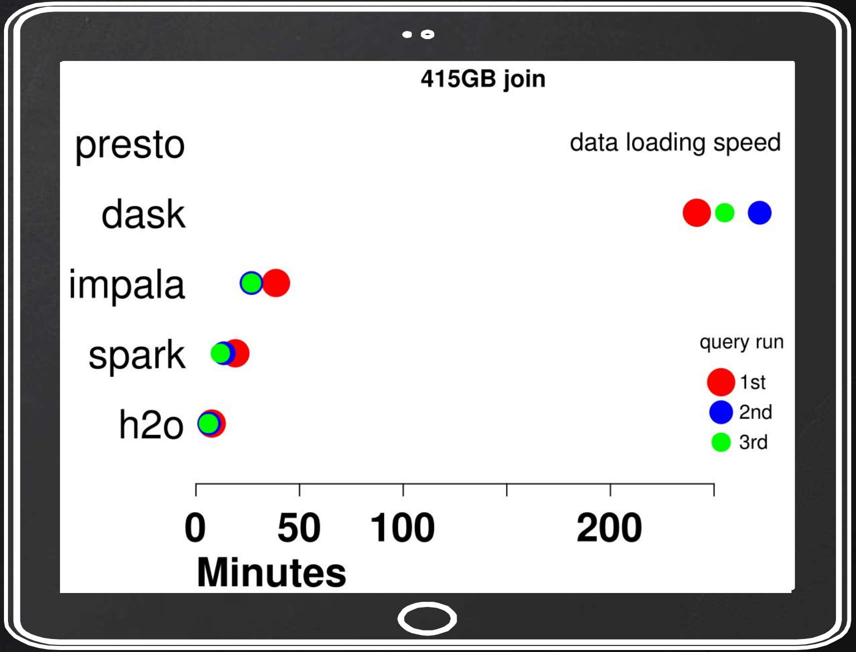
- Processing of raw data into another form
- Exploratory data analysis
- Feature engineering
- Data cleaning
- Preparing data for machine learning

Data Munging

Task	Operation
calculated columns	select/update
data type conversion	select/update
cleaning (trim, substring)	select/update
remapping, lookup	join
summaries	group by
filtering	special case of join
top by group	group by + sort
order data	sort

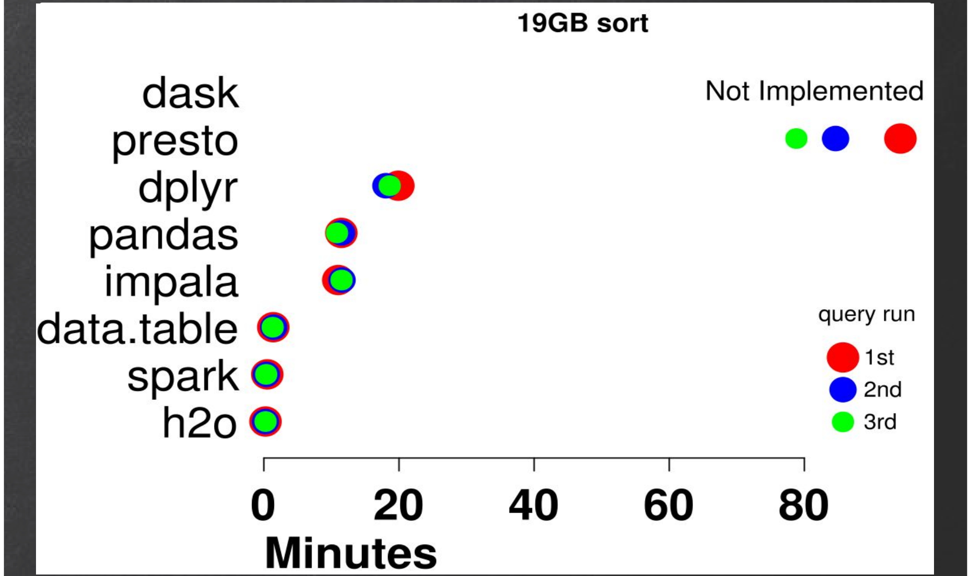
High cardinality big join

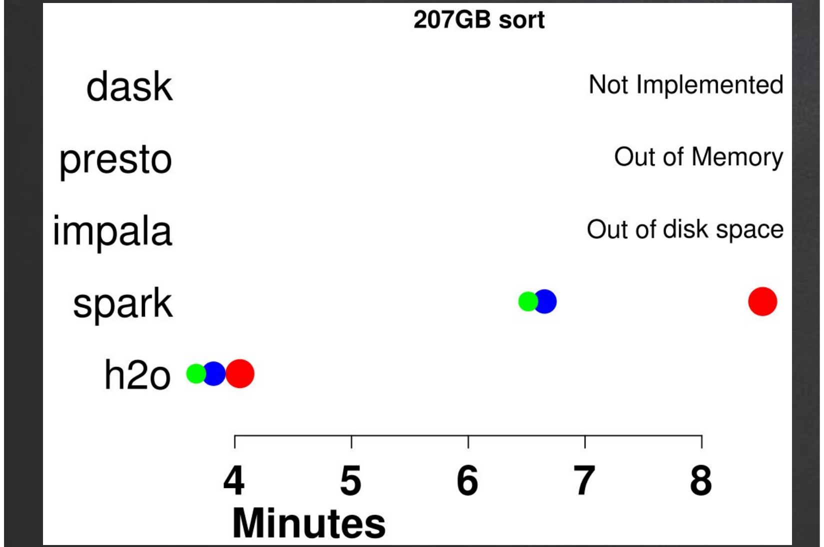




- Two integer columns in each table
- N rows x N rows join
- High cardinality join column

Big sort





- Two integer columns
- High cardinality sort column

Benchmark

- Cluster of 10 physical machines
- 32 CPU per node
- 200 GB memory per node

Open and reproducible benchmark:

https://github.com/h2oai/db-benchmark

Jan Gorecki github.com/jangorecki