CSE P 590

Building Data Analysis Pipelines

Fall 2024

Big data

Today

- More on big data processing
- Wrap up and open discussion
- In-class 5

More on big data processing

Recap: Big data

Characteristics

- *Volume*: data sets are (too) big → distributed analysis
- *Variety*: data formats: structured, semi-structured, unstructured
- *Velocity*: data changes rapidly → real-time analysis

- *Variability*: meaning of data changes
- *Veracity*: noisy data (tradeoff between noisy and useful)
- *Value*: informed decision making (value of collected data)

Recap: Big data challenges

Compute bound

- Compute-intensive simulations
- Real-time processing
- High-volume data

Memory bound

- Data exceeds memory on a single machine

I/O bound

- Different data sources (structured, semi-structured, unstructured)
- Data sharing among different processes

Recap: sparklyr (a Spark DSL)

Distributed data storage

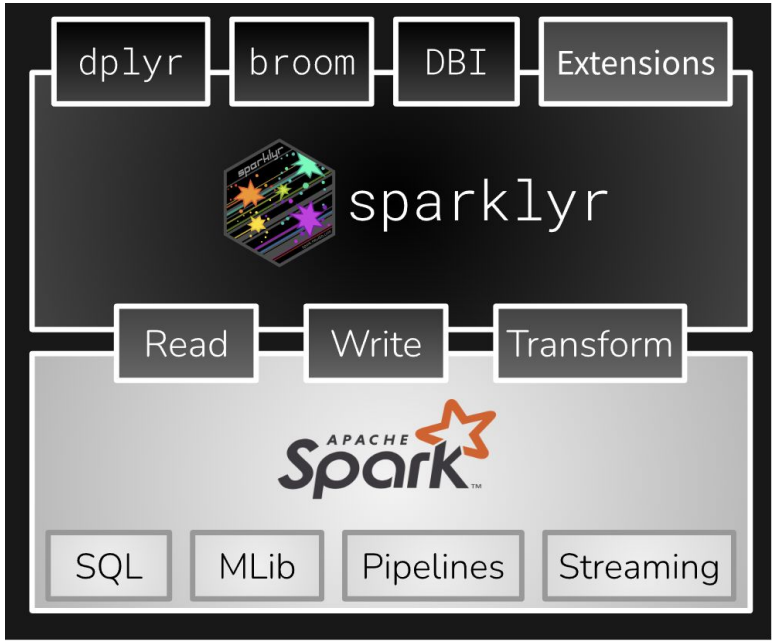
- A dplyr backend
- Supports SQL (like dbplyr)
- Lazy evaluation

Distributed data processing

- Support for many data formats

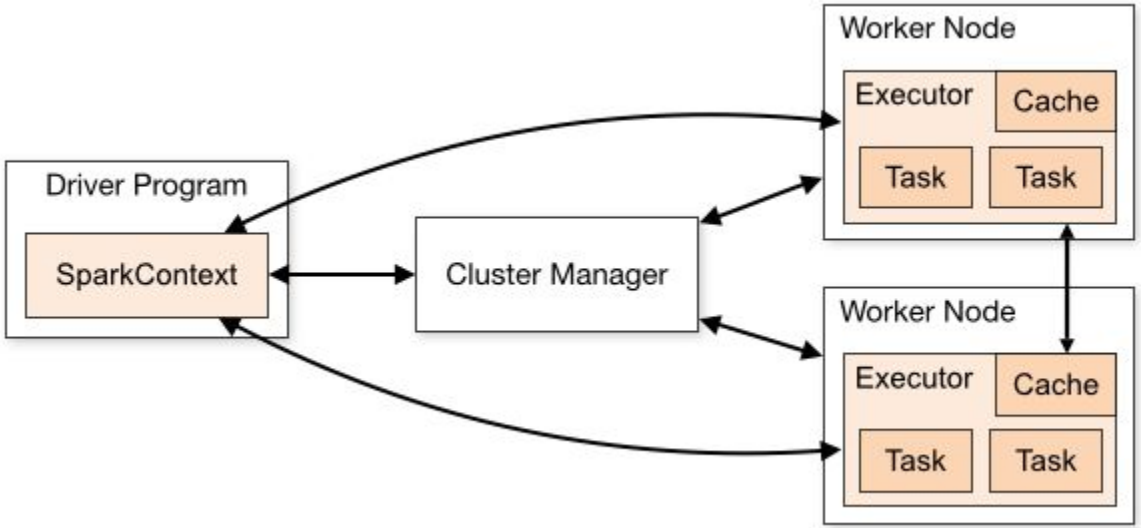
Modeling/ML

- Support for many common model types



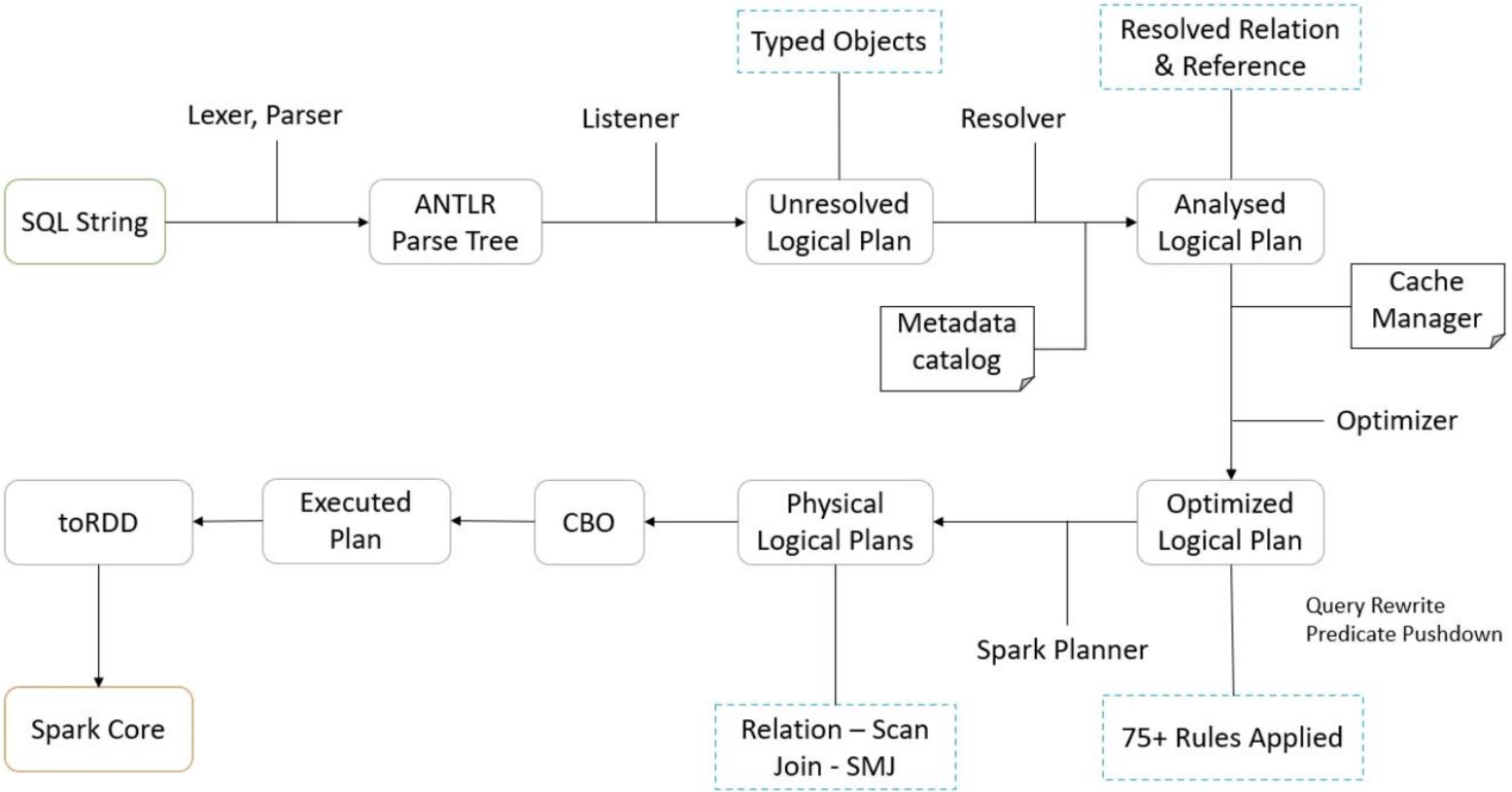
https://spark.posit.co

Spark: architecture



https://spark.apache.org/docs/3.5.3/cluster-overview.html

Spark: query optimization



Dplyr backends: Spark vs. DB

- Sparklyr and dbplyr as dplyr backends
- Automatic translation from dplyr verbs to SQL queries
- Lazy evaluation to minimize network traffic

> (If your data fits in memory there is no advantage to putting it in a database: it will only be slower and more frustrating.)

> Most of the time you don't need to know anything about SQL, and you can continue to use the dplyr verbs that you're already familiar with:

> However, in the long-run, I highly recommend you at least learn the basics of SQL. It's a valuable skill for any data scientist, and it will help you debug problems if you run into problems with dplyr's automatic translation.

https://dbplyr.tidyverse.org/articles/dbplyr.html
https://spark.posit.co/guides/dplyr.html

Live demo: Spark vs. DB

Open discussion

Why R for this course?

1. Leveling the playing field (Python knowledge distribution is bimodal)
2. (My opinion) dplyr, ggplot2, etc. are superior: great for data wrangling/viz.
3. Focusing on (non-ML) analysis concepts
4. Many specialized data science packages in R

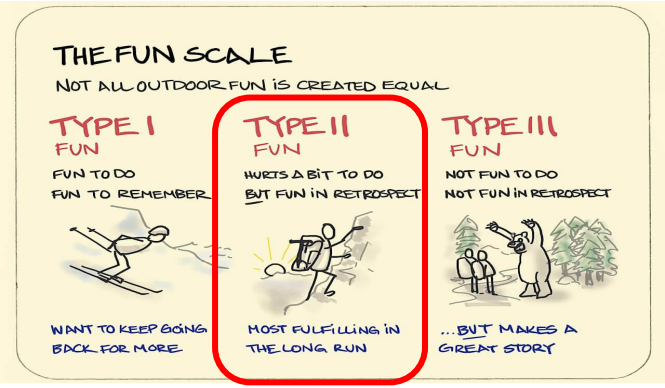
Pick the right tool/ecosystem for the job!

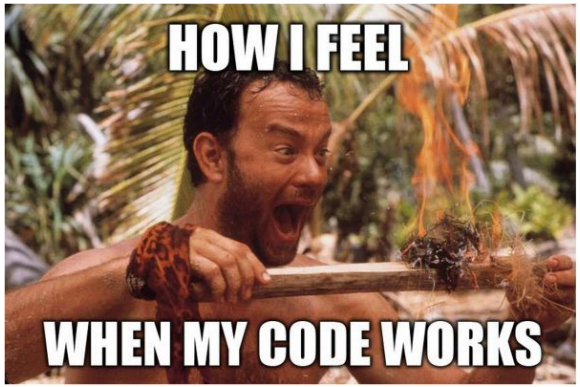
Open discussion

Example topics

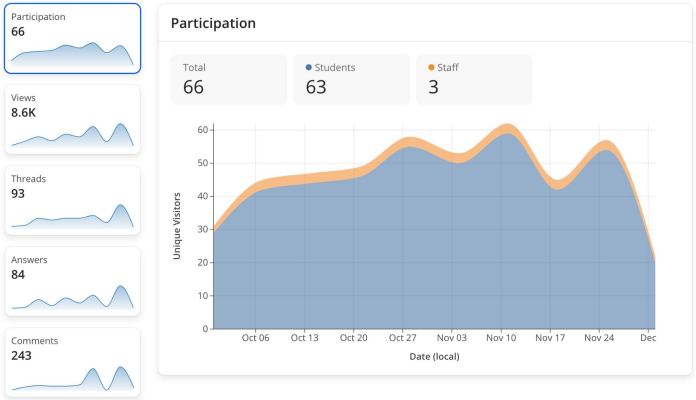
- Highlights and frustrations
- Open questions and confusions
- Experiences with AI-assisted coding
- R vs. Python (your conclusions)

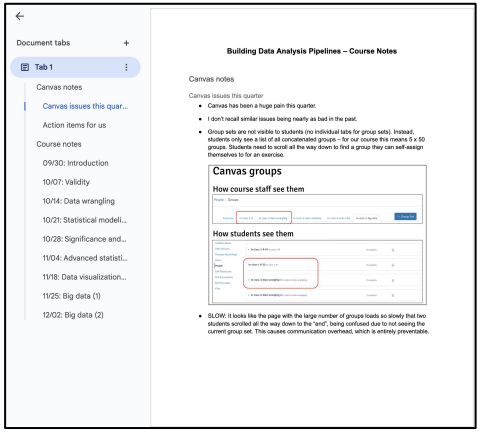
10 weeks flew by!

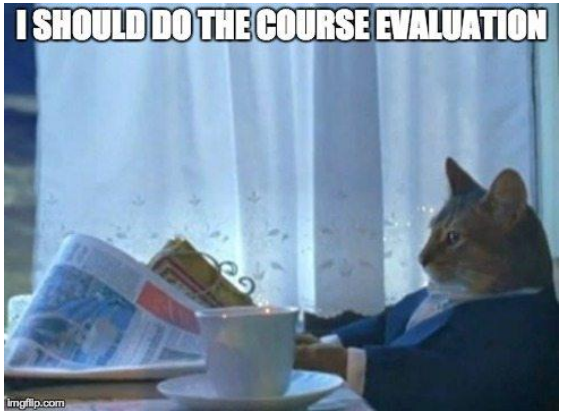




Thanks for a great quarter!







https://uw.iasystem.org/survey/297879

In-class 5