PySpark Data Manipulation

Big Data & Predictive Analysis Lanjut

Arif Laksito, M. Kom

Data Manipulation

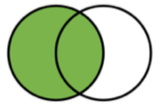
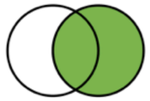
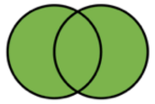
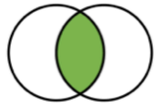
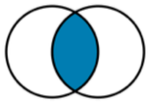
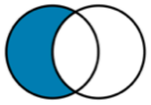
As we have seen in previous meeting, working with PySpark data is as familiar as working with Pandas Dataframe, or SQL.

.

Data Manipulation

- Using our understanding of Dataframes and SQL, there won't be much to talk about in theories.
- Let's now just try to refresh our mind regarding all these keywords:
    1. Select
    2. Filtering
    3. Aggregation
    4. Grouping
    5. Joining

Types of Join in PySpark

1. **Left** Join 
2. **Right** Join 
3. **Full** Join 
4. **Inner** Join 
5. **Semi** Join 
6. **Anti** Join 

Let's code

PySpark Data Manipulation