HW 01 Beam on Spark

Jiho Choi
(jihochoi@snu.ac.kr)

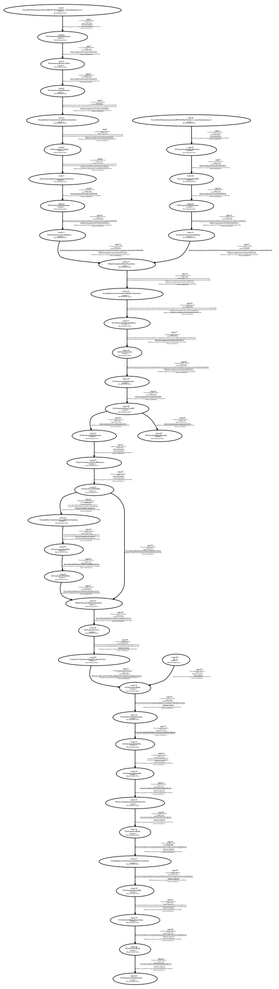
Nemo DAG and

Although Nemo optimizes (Fold, …) the DAG using compiler optimizer, I couldn't find much differences in the sense of vertices and edges. However, I believe with different SQL statements and setting, the topologies may have changed.

Execution Plan

Starting from empty logical plan, the execution plan fills and groups the topologies with whom are not dependent to each other.

Original Nemo IR (Intermediate Representation)



Querying without parallel processing

- SELECT rank_num, country FROM RANKING WHEN rank_date = '2018-06-07'
- SELECT rank_num, RANKING.COUNTRY, PLAYER.height, PLAYER.weight
  FROM PLAYER GROUP BY country
- SELECT rank_num, RANKING.COUNTRY, PLAYER.height, PLAYER.weight,
  BMI(PLAYER.height, PLAYER.weight) FROM RANKING
  INNER JOIN PLAYER ON RANKING.country = PLAYER.country

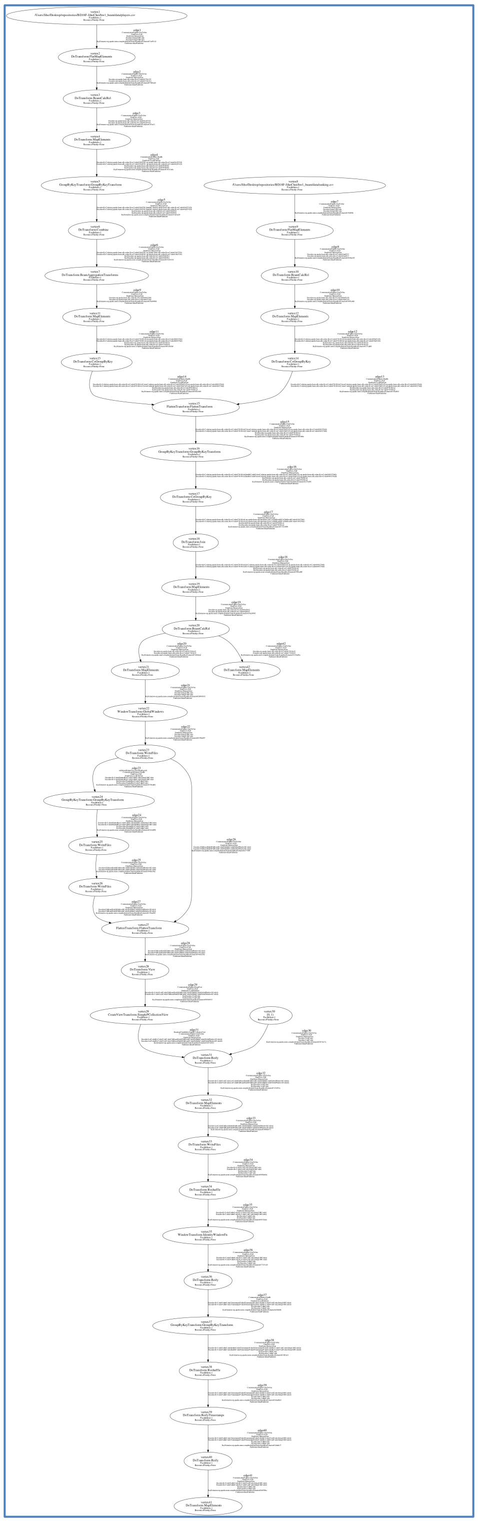
Data flow (with distributed systems) → requires a lot more

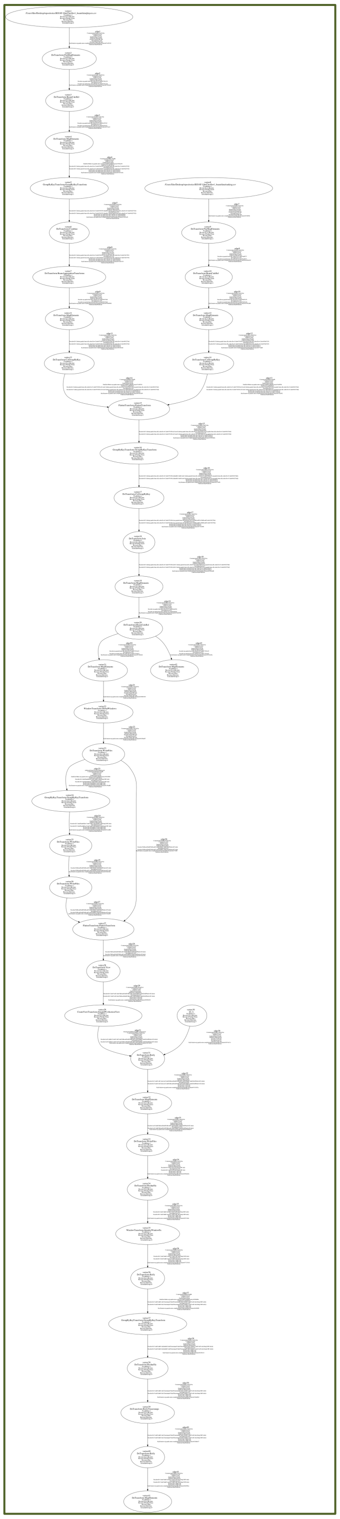
- PIPELINE_OPTION, PIPELINE
- PCOLLECTION
- PTRANSFORMATION

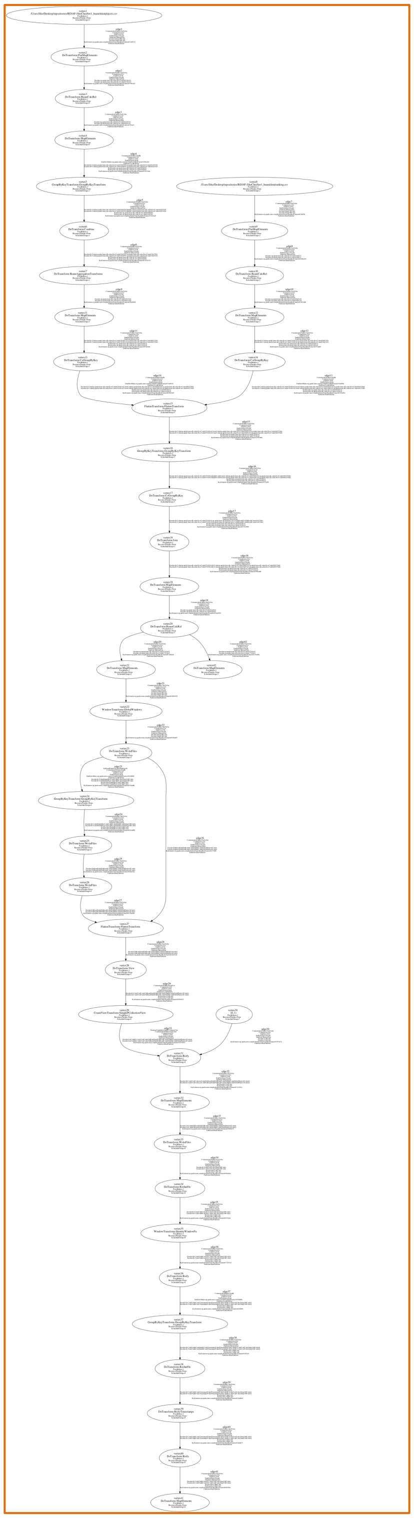
References

- Nemo Compiler Optimizer
  https://github.com/apache/incubator-nemo

IR 0 → IR After Default Policy, IR After Compression Pass







Plan –1 Logical → Plan 0 Submitted → Plan 1 Periodic → Plan 2 Periodic → Plan 3 Final