The IFs of Hive

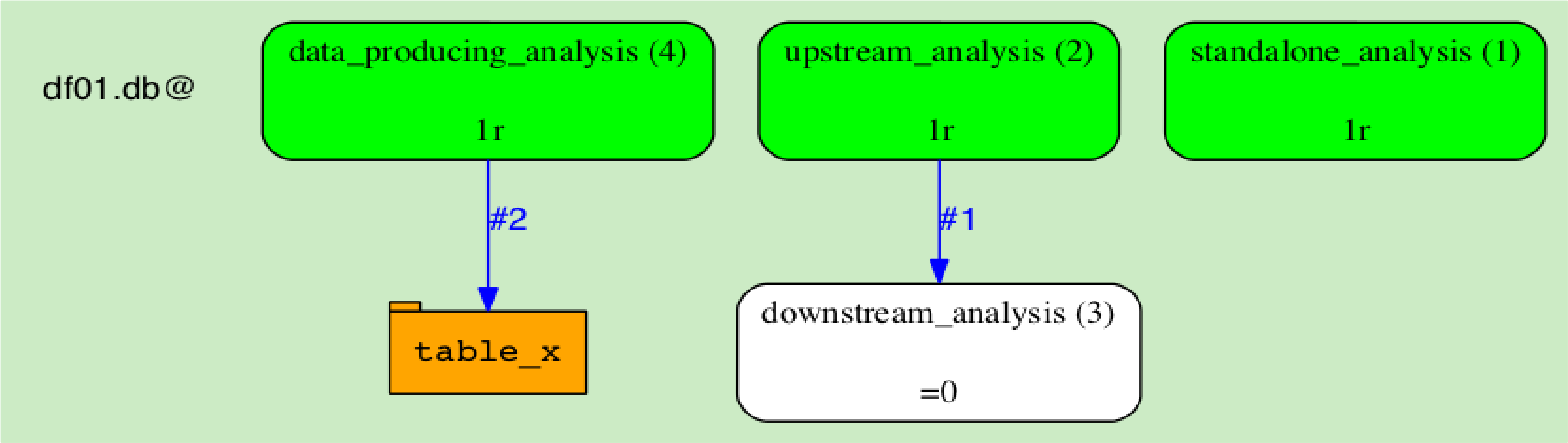
Leo Gordon

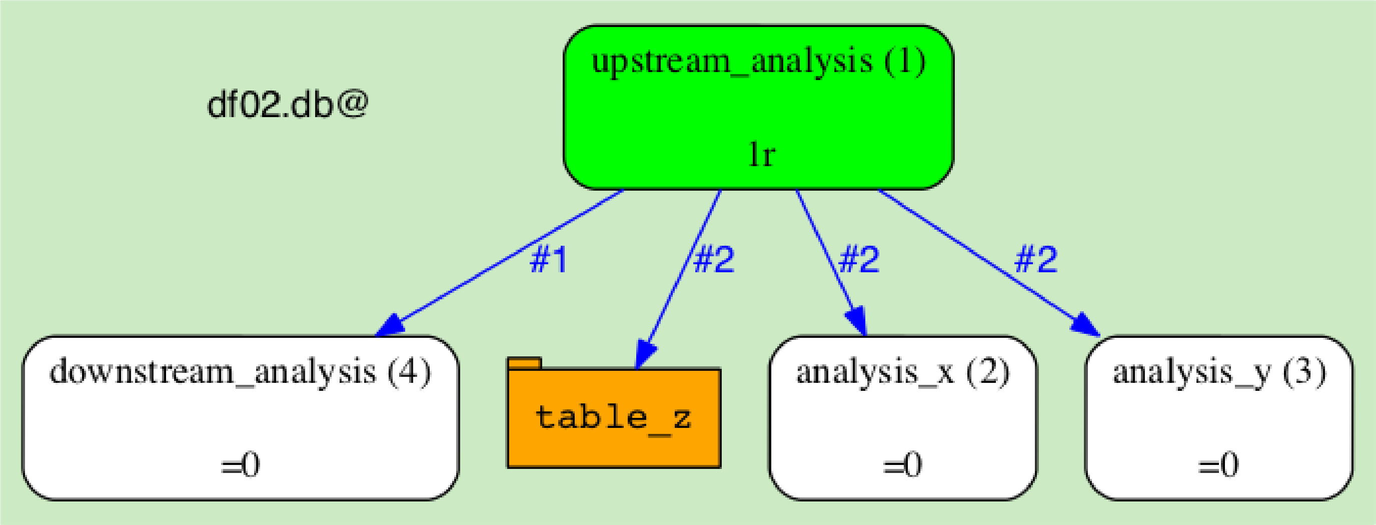
A3-145
ehive-users@ebi.ac.uk

Flow in eHive

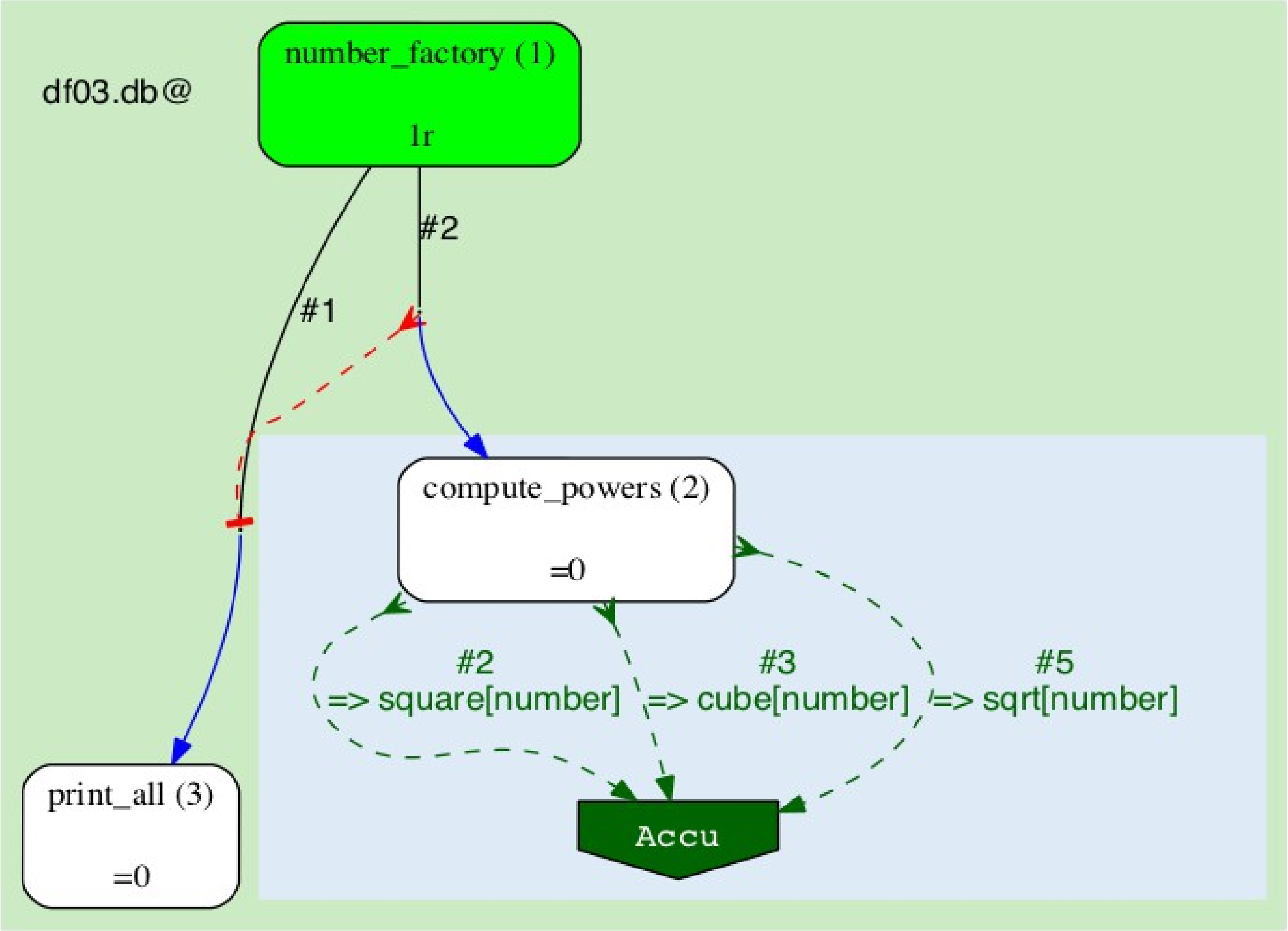
- **eHive is our main tool for building powerful and complex pipelines**
- **We use stratification to minimize the development effort:**
  - *Runnables* are simple & generic building blocks;
    they package the code
  - *PipeConfigs* have enough complexity to parametrize and link the above;
    they describe pipeline flow diagrams
- **In eHive dataflow and control flow tend to converge together, based on a simple messaging protocol:**
  - *Runnables* emit messages (send hashes into numbered channels)
  - *PipeConfigs* decide where they should go. *It is the only linking mechanism.*

Flow in eHive – simple examples



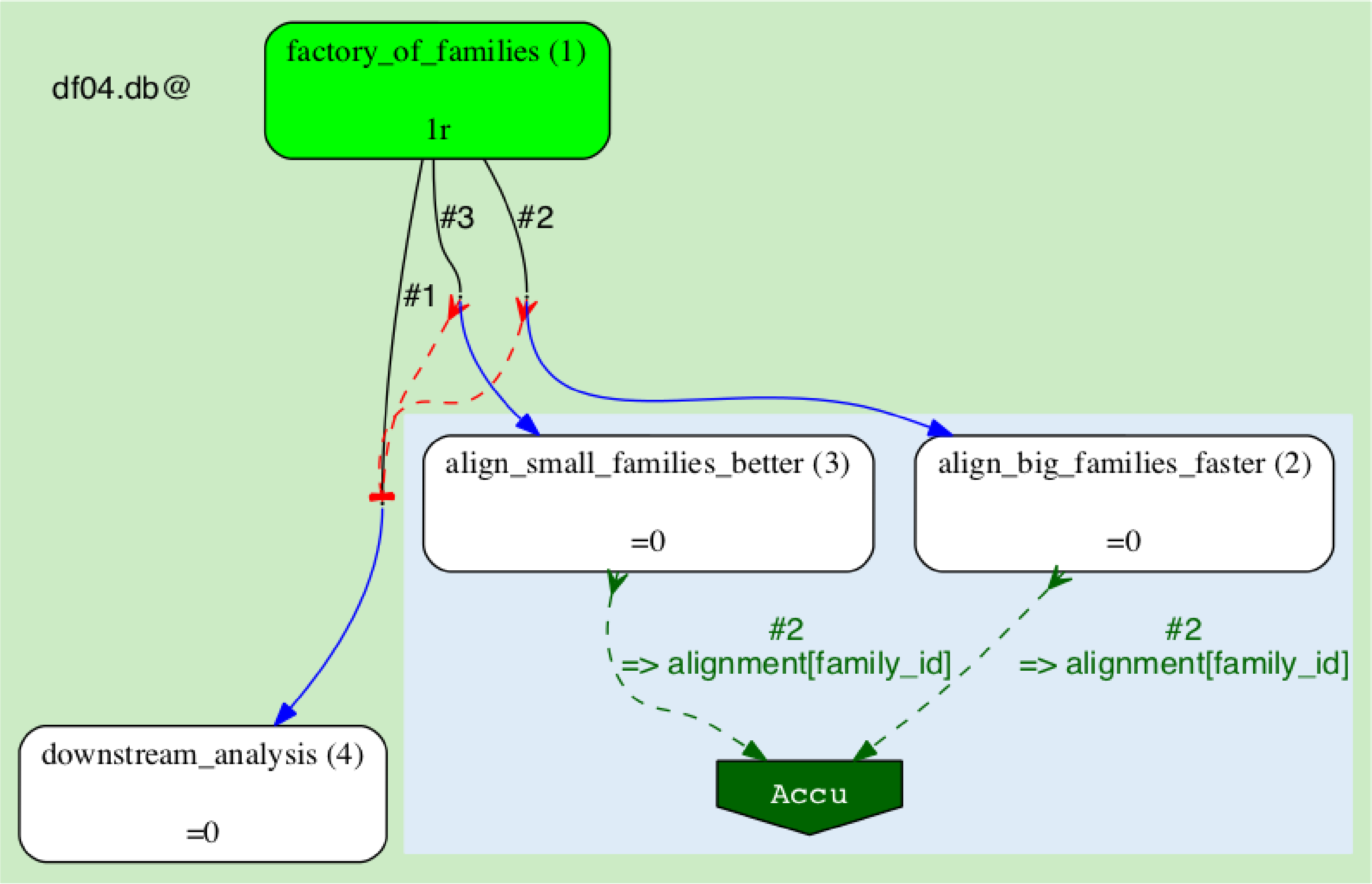


Flow in eHive – *Semaphores* and *Accus*



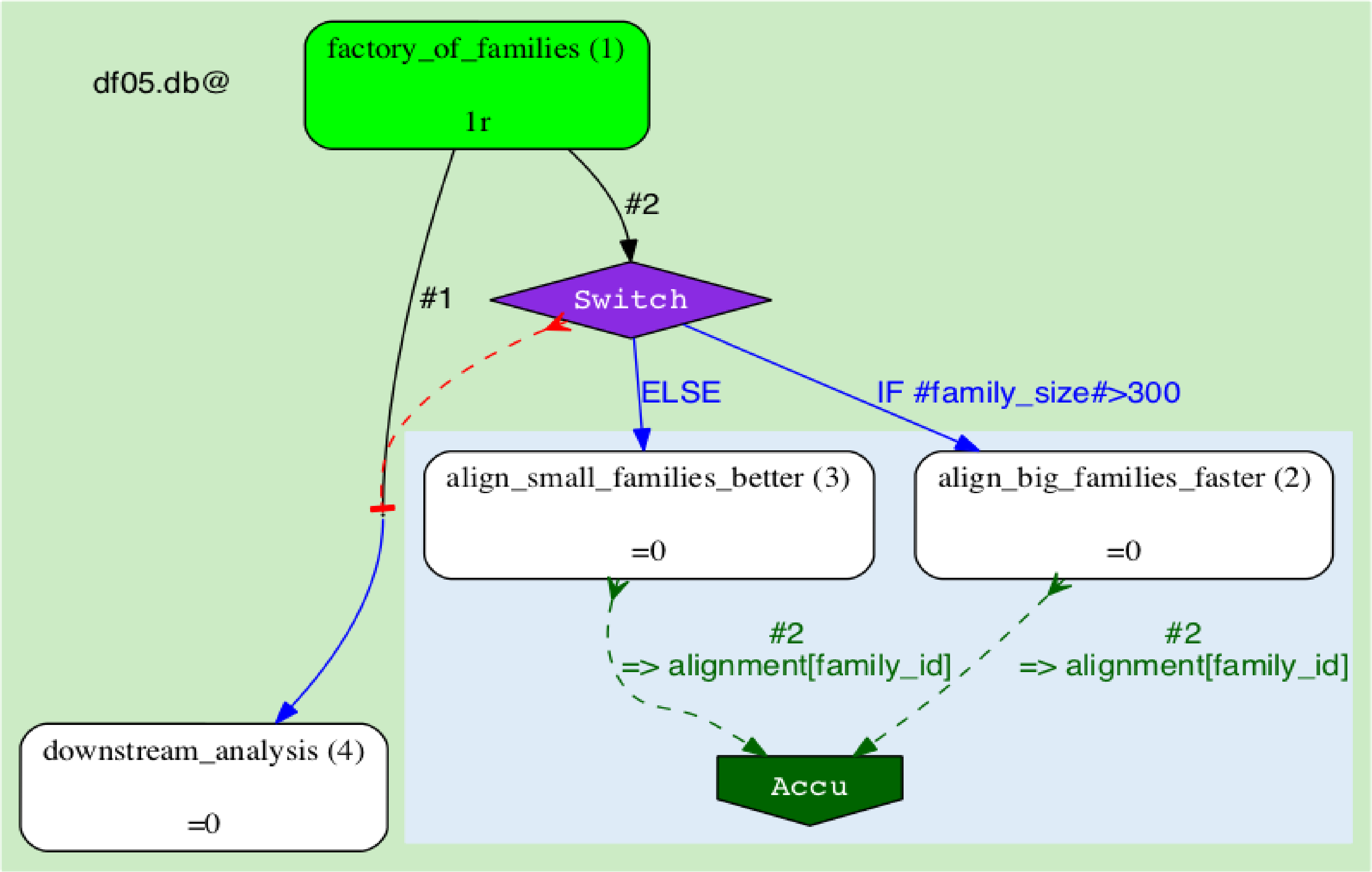
- eHive's way to express fanning out, parallel processing and merging back
- In semaphore context the concepts of control flow and dataflow converge
- An extra type of dataflow targets in semaphore fan context: accumulators

Old way: choice built into *Runnables*



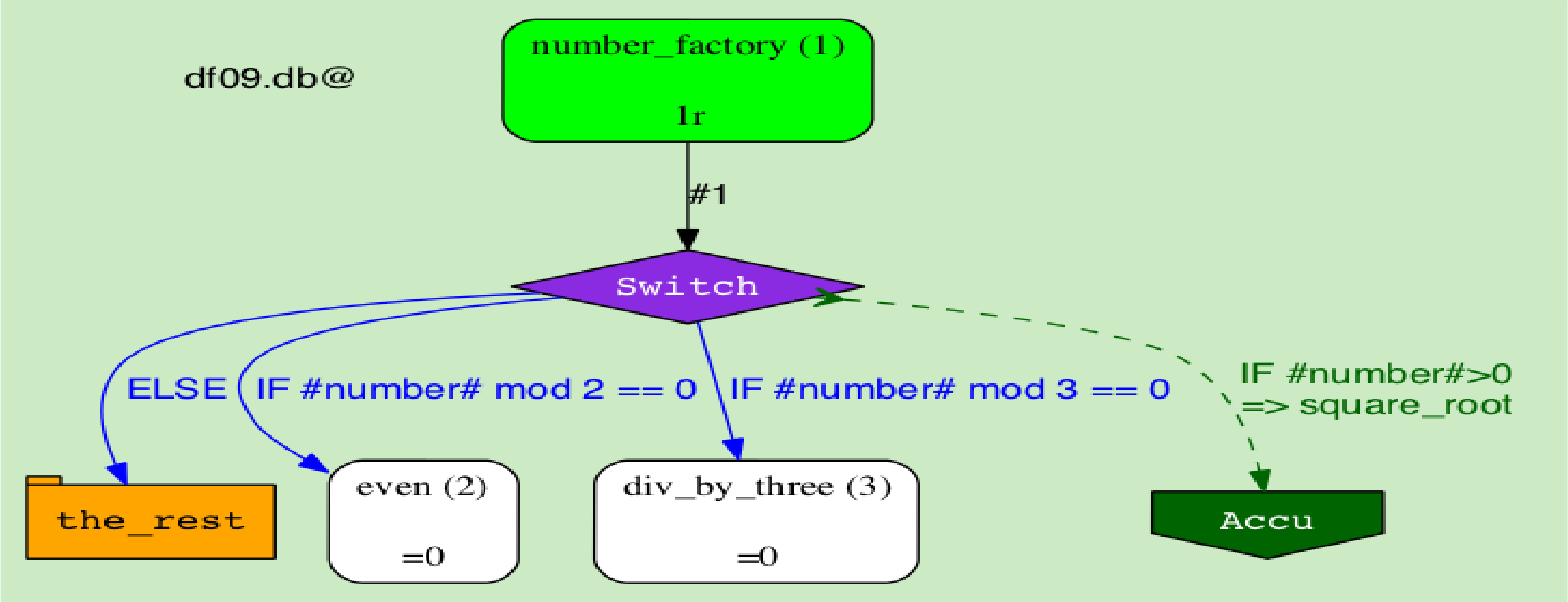
- Solution where Runnable defines the selection criterion is inflexible (even when parametric).
- It is also invisible to the pipeline engineer or anyone looking at the diagram.

New way: choice on *PipeConfig* level



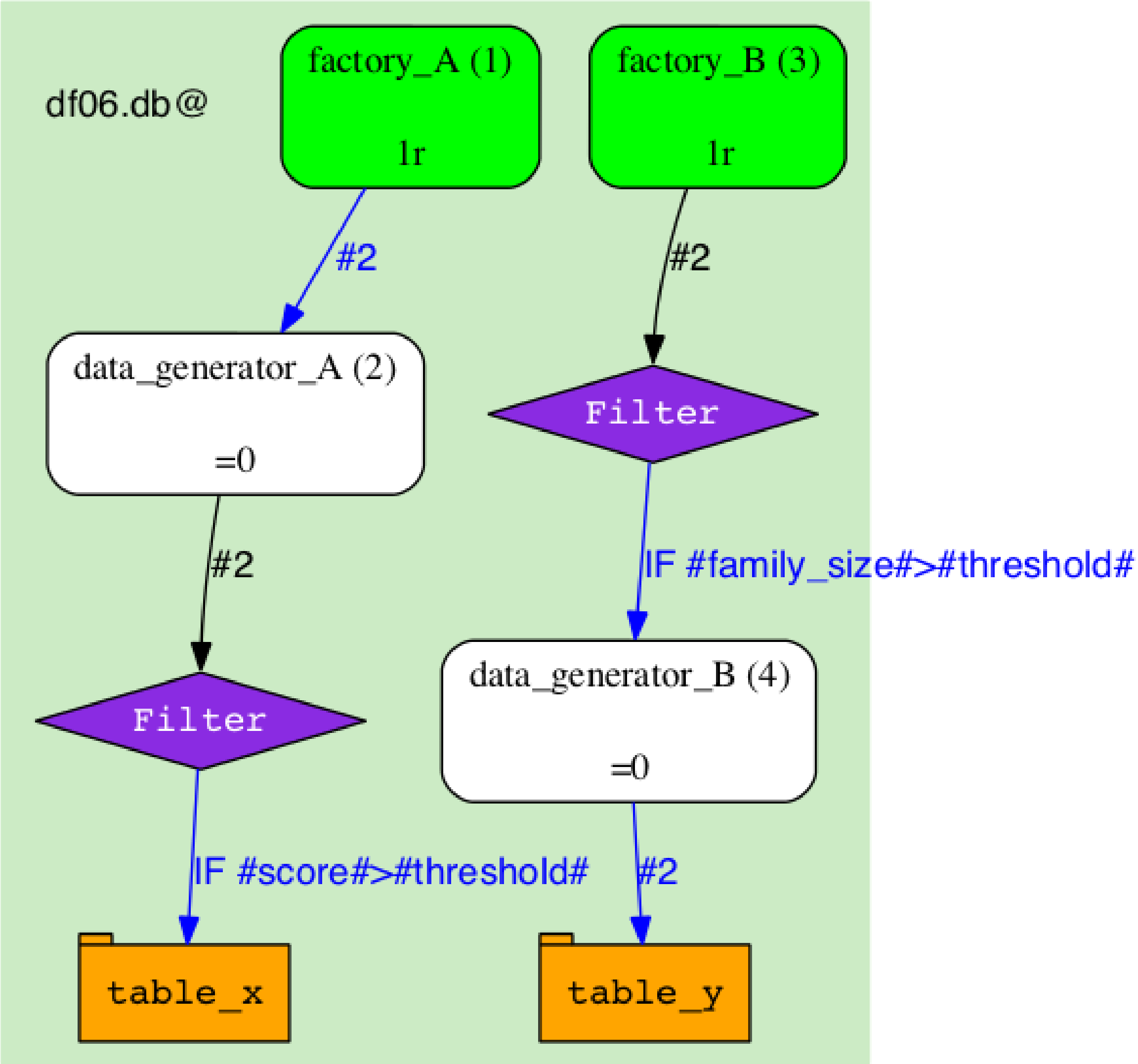
- We can leave the Runnable alone, treat it as a “black box”.
- The condition can be used before any type of target (filter or sort data, create jobs, or both)

Conditional flow: pros and cons



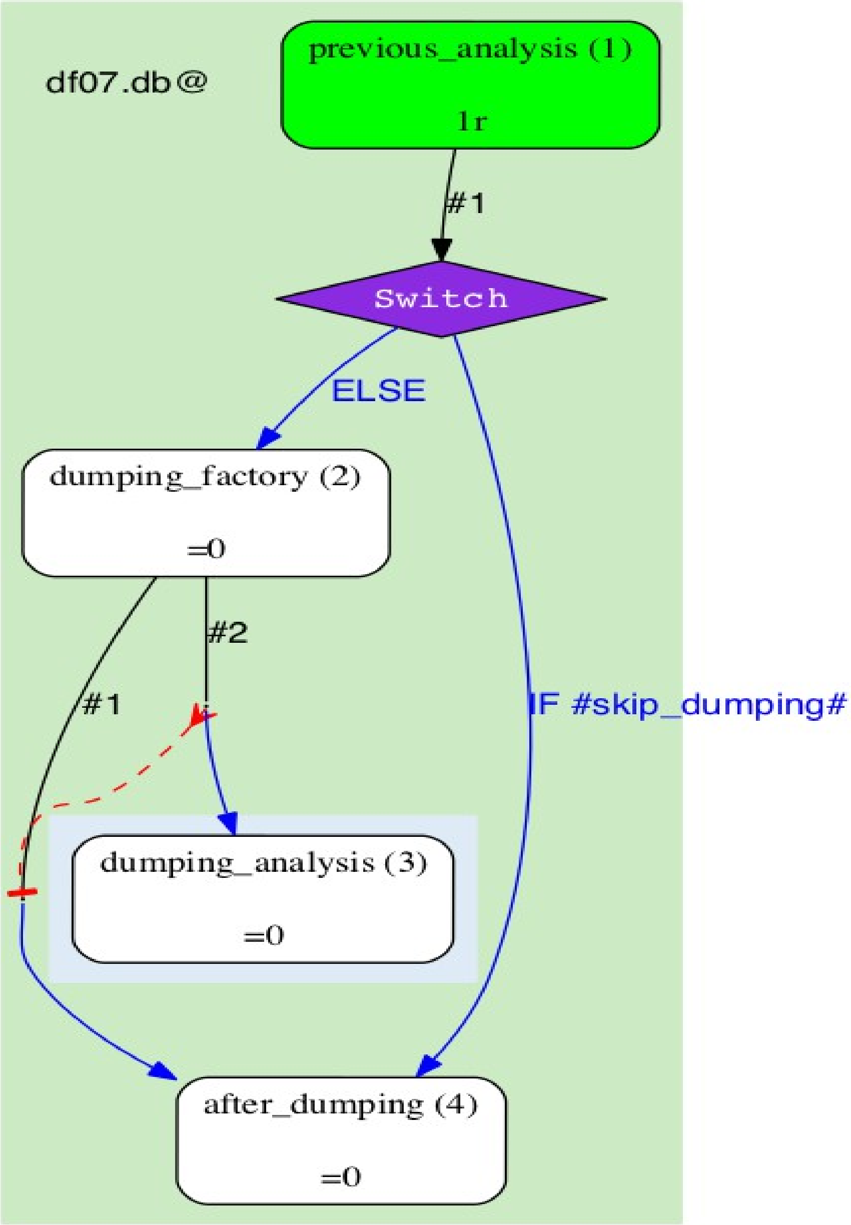
- **Pros**:
  - Conditions are not mutually exclusive and are computed without an order (it's a “parallel switch”)
  - They share a common “ELSE” branch
  - A condition can be used before any type of target: filter or sort data, create jobs, accumulators, or any mixture
  - Diagrams become more readable
- **Cons**:
  - Conditions are not mutually exclusive and are computed without an order (they cannot be chained or nested)

Usage of conditional flow: filtering

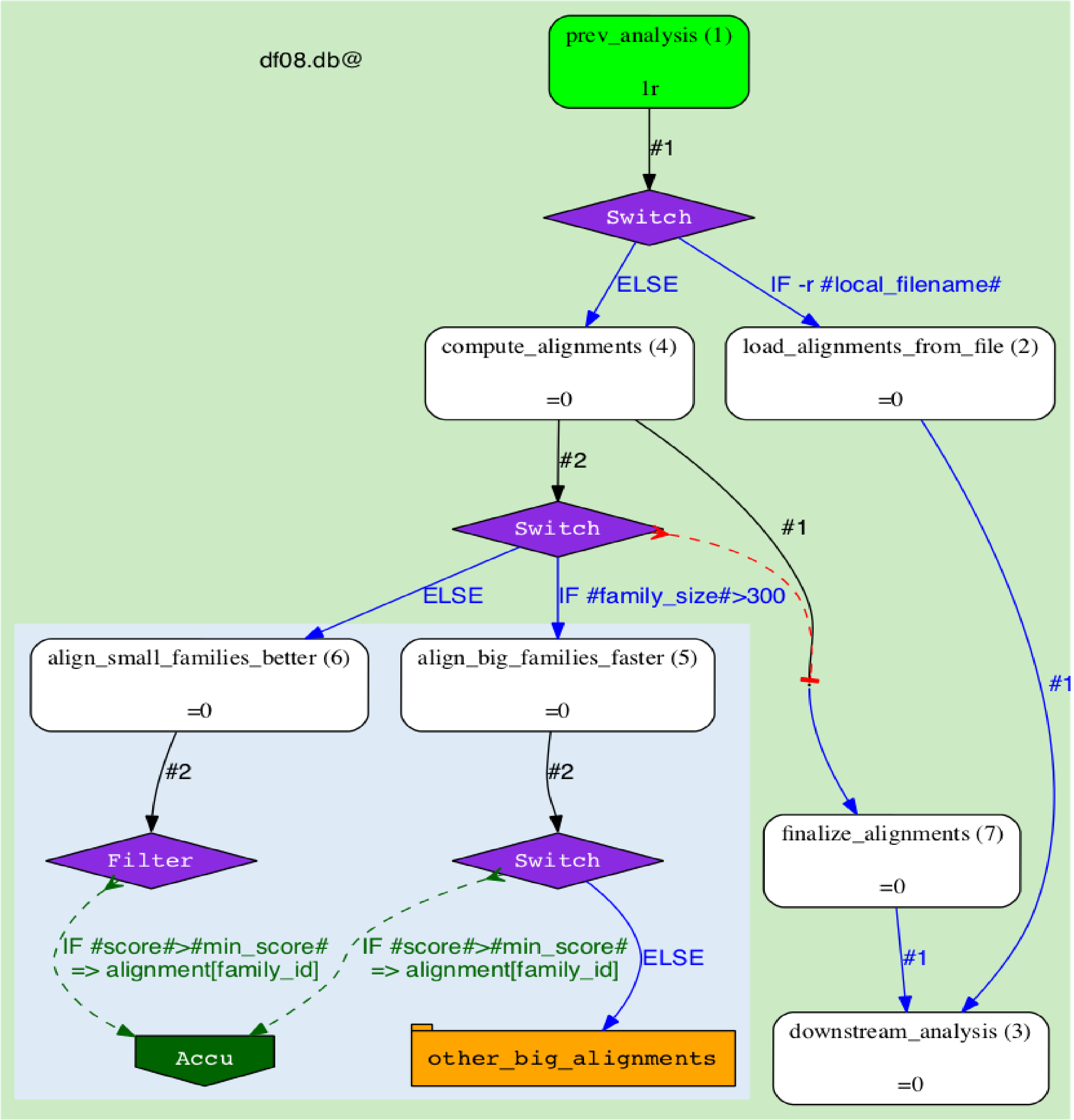


- **Filtering individual items**:
  - data before storing it in a table
  - jobs before creating them in an analysis

Usage of conditional flow: bypass



- **Skipping whole parts of the pipeline**:
  - statically (via pipeline_wide_parameters)
  - or dynamically (by passing the parameter in)


df08.db@
prev_analysis (1)
1r
#1
Switch
ELSE
IF -r #local_filename#
compute_alignments (4)
=0
load_alignments_from_file (2)
=0
#2
#1
Switch
ELSE
IF #family_size#>300
#1
align_small_families_better (6)
=0
align_big_families_faster (5)
=0
#2
#2
Filter
Switch
IF #score#>#min_score#
=> alignment[family_id]
IF #score#>#min_score#
=> alignment[family_id]
ELSE
finalize_alignments (7)
=0
#1
Accu
other_big_alignments
downstream_analysis (3)
=0

Availability

- Conditional flow can be tried on *master* branch
- Officially in from *version/2.4*

Questions?

Acknowledgements

The Entire Ensembl Team, esp. Matthieu Muffato

Fiona Cunningham[1], M. Ridwan Amode[1], Daniel Barrell[1,2], Kathryn Beal[1], Konstantinos Billis[1], Simon Brent[2], Denise Carvalho-Silva[1], Peter Clapham[2], Guy Coates[2], Stephen Fitzgerald[1], Laurent Gil[1], Carlos García Girón[1], Leo Gordon[1], Thibaut Hourlier[1], Sarah E. Hunt[1], Sophie H. Janacek[1], Nathan Johnson[1], Thomas Juettemann[1], Andreas K. Kähäri[2], Stephen Keenan[1], Fergal J. Martin[1], Thomas Maurel[1], William McLaren[1], Daniel N. Murphy[1,2], Rishi Nag[1], Bert Overduin[1], Anne Parker[1], Mateus Patricio[1], Emily Perry[1], Miguel Pignatelli[1], Harpreet Singh Riat[1], Daniel Sheppard[1], Kieron Taylor[1], Anja Thormann[1], Alessandro Vullo[1], Steven P. Wilder[1], Amonida Zadissa[1], Bronwen L. Aken[1], Ewan Birney[1], Jennifer Harrow[2], Rhoda Kinsella[1], Matthieu Muffato[1], Magali Ruffier[1], Stephen M.J. Searle[2], Giulietta Spudich[1], Stephen J. Trevanion[1], Andy Yates[1], Daniel R. Zerbino[1] and Paul Flicek[1,2,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Funding

wellcome trust

National Human Genome Research Institute

CTTV Centre for Therapeutic Target Validation

BLUEPRINT epigenome

EMBL

BBSRC bioscience for the future

BioMedBridges

MedBioinformatics

Co-funded by the European Union