Description of workflow

Figure 1 and 2 are the abstract workflow structure diagrams of the workflows. The ellipse represents the job in the workflow, and the rectangle refers to the input and output data. Arrows represent the dependencies between jobs. As for the job naming rules in the figures, the job name is preceded by the “_”, the transformation catalog adds descriptions (including job script) to the job based on the job name, and the ID of the job after “_”, the workflow description file identifies each job by the ID. Jobs can have the same name,and the jobs with the same name will have the same job descriptions in the transformation catalog, but the same workflow description file can not have the same ID jobs, because this will make it impossible to distinguish each job form workflow.

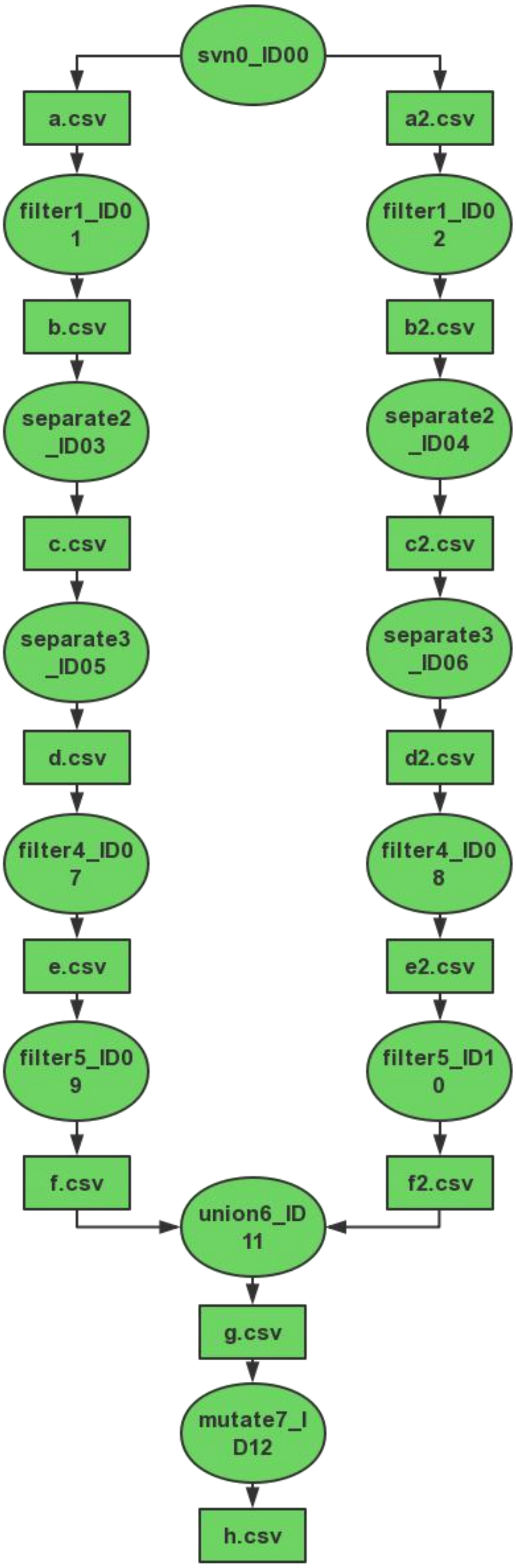

Figure 1 Construction of traffic data wrangling workflow A

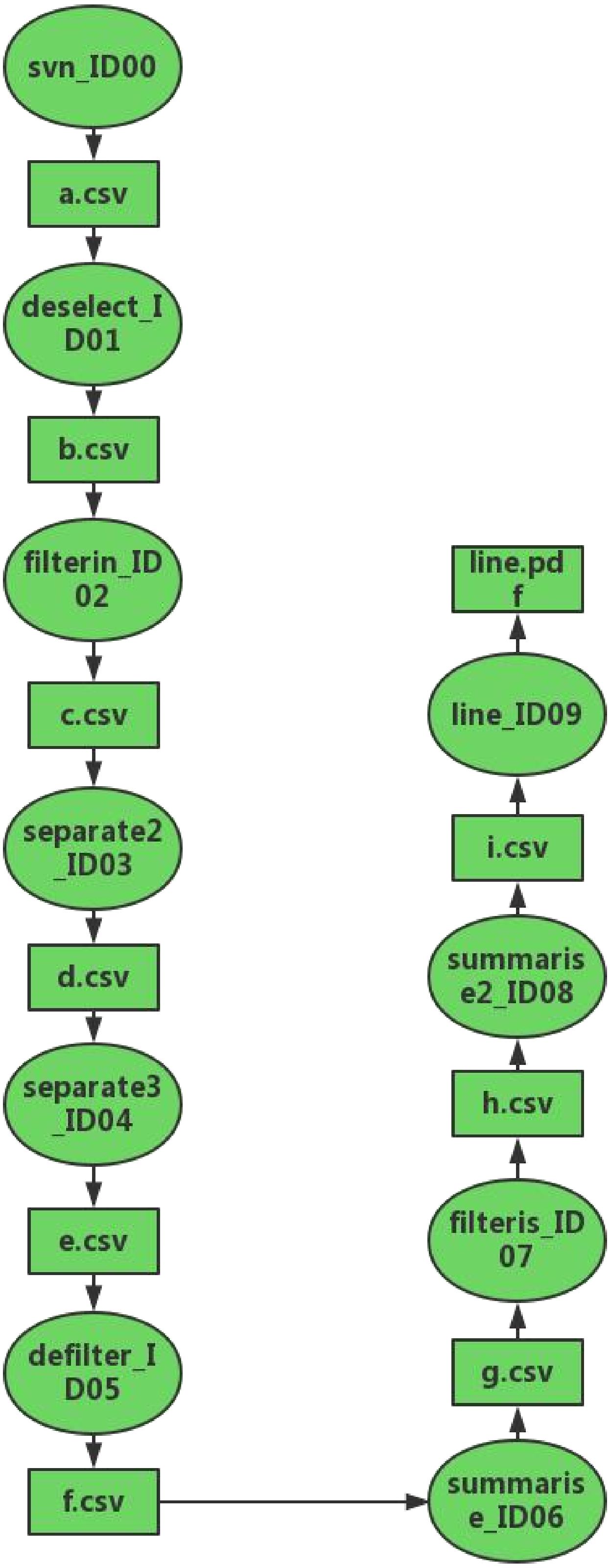

Figure 2 Construction of traffic data wrangling workflow B

Workflow A has 13 jobs, svn0_ID00 gets the traffic data sets a.csv and a2.csv from the data source, filter1_ID01 filters the “DirectionDescription” data attribute in the a.csv, filters out the data whose value is North, and uses them as the output data set b.csv. Then, separate2_ID03 splits the “sdata” attribute in b.csv into “data” and “time” attributes, and generates data set c.csv. After that, the “data” attribute in c.csv is divided into “year”, “mouth” and “day” by separate3_ID05, and the data set d.csv is generated. Then filter4_ID07 and filter5_ID09 filter the “day” attribute and “time” attribute separately, traffic data information from 17:00 to 18:00 every Friday is obtained and output as f.csv. In the same way, f2.csv is obtained through filter_ID02, separate2_ID04, separate3_ID06, filter4_ID08 and filter_ID10. The next step is to use union6_ID11 to integrate f.csv and f2.csv into a data set g.csv. Finally, the final data set h.csv is generated by computing the JT value (avgspeed/3.01) for each row of data through mutate7_ID12.

Workflow B has 10 jobs. First, the traffic data set a.csv is retrieved from the data source via svn_ID00, then the useless data attributes are deleted using deselect_ID01, after that, similar to the previous workflow, the north workday data is filtered through filterin_ID02, separate2_ID03, separate3_ID04 and defilter_ID05 and final output is f.csv. Then, the total volumes per hour of every working day are counted by summarise_ID06 and written into g.csv. Afterwards, filteris_ID07 filters out all the Monday data as h.csv and sends it to summarise2_ID08, summarise2_ID08 counts the cumulative volumes per hour of all Mondays as i.csv. Finally, line_ID09 draws i.csv into a line chart and outputs it into PDF format.

Figure 3 and 4 are relationship diagrams of three workflows jobs, jobs scripts and called web services. The green ellipse is job of the workflow, the yellow ellipses are job scripts, the white ellipses are traffic data wrangling web services (operator).

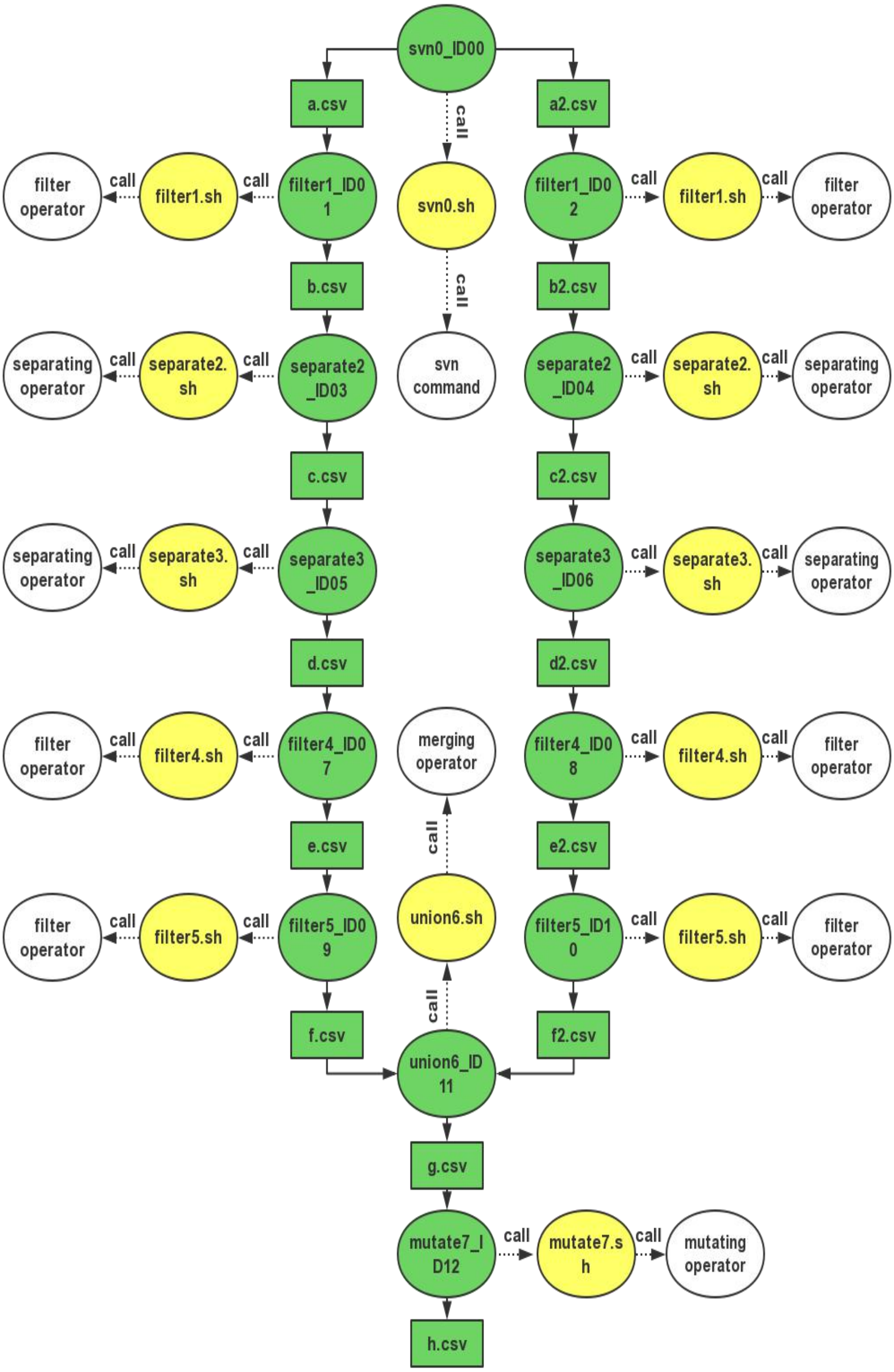

Figure 3 Call graph of workflow A

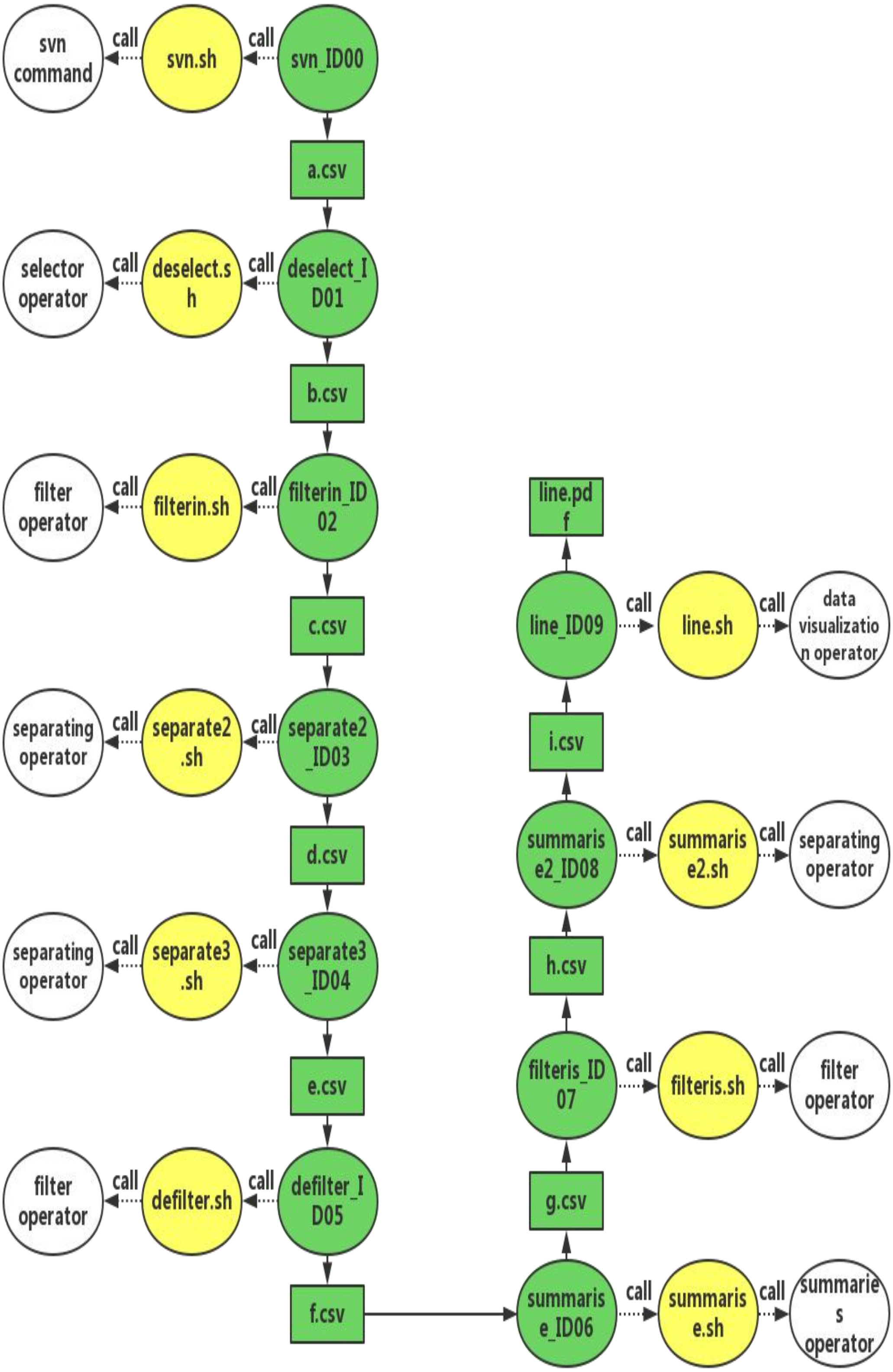

Figure 4 Call graph of workflow B