Dominique W. Chandra D.
6162101104
Mata Kuliah Rekayasa Data

Tugas 3

Dalam tugas ini, diberikan sebuah data mentah bernama student.csv dan data tersebut akan ditransformasi sehingga data tersebut siap digunakan. Namun, sebelum memulai membersihkan data, akan dilakukan beberapa eksplorasi data sederhana untuk lebih memahami data. Beberapa hal yang ditemukan setelah melakukan eksplorasi data, antara lain:

- Identifikasi nilai yang hilang: dalam dataset students.csv, terdapat 7 kolom, yaitu nama, umur, jenis kelamin, nilai tes masuk, persentase sekolah, kota, dan status penerimaan. Dari ketujuh kolom tersebut, semuanya memiliki nilai yang hilang, dengan rincian:
  - Nama: 10 data hilang
  - Umur: 10 data hilang
  - Jenis kelamin: 10 data hilang
  - Nilai tes masuk: 11 data hilang
  - Persentase sekolah: 11 data hilang
  - Kota: 10 data hilang
  - Status penerimaan: 10 data hilang
- Identifikasi data duplikat: terdapat beberapa data duplikat, contohnya seperti yang ada di cuplikan di bawah ini

Name	Age	Gender	Admission Test Score	High School Percentage	City	Admission Status
Ayesha	24	Male	94	98.43	Multan	Rejected
Ayesha	24	Male	94	98.43	Multan	Rejected

Namun, keanehan terlihat pada gambar di bawah ini

Name	Age	Gender	Admission Test Score	High School Percentage	City	Admission Status
Hassan	24	Female	79	75.67		
Hassan	24		79	75.67	Karachi	Accepted

Dapat dilihat bahwa nama, usia, nilai tes masuk, dan status penerimaan sama persis namun terdapat data yang hilang di keduanya. Hal ini bisa dimanfaatkan untuk mengisi kolom satu sama lain.

- Identifikasi data salah: menurut pengamatan penulis, terdapat beberapa data yang kemungkinan besar salah, contohnya seperti pada gambar di bawah

Name	Age	Gender	Admission Test Score	High School Percentage	City	Admission Status
Aliya	-1	Male	101	54.59		
Shehroz	-1	Female	61	69.48	Quetta	Rejected
	-1	Male	66	79.07	Rawalpinc	Rejected
Zunaira	-1	Female	84	58.77	Quetta	Rejected
Shoaib	-1	Male	91	80.12	Quetta	Accepted

Seperti yang dapat dilihat, bahwa terdapat data siswa yang menyatakan bahwa umur dari siswa tersebut adalah -1. Hal ini tidak masuk akal karena tidak ada umur yang mulai dari bilangan negatif. Oleh karena itu, nilai ini tentu harus diganti. Selain itu, kesalahan juga terdapat di kolom nilai tes masuk.

Name	Age	Gender	Admission Test Score	High School Percentage	City	Admission Status
Aliya	-1	Male	101	54.59		
Rehan	19	Female	-5	61.91	Quetta	Rejected
Umar	22	Male	150	77.69		Rejected

Perhatikan bahwa dari gambar di atas, terdapat nilai tes masuk yang lebih dari 100 dan kurang dari 0. Hal ini cukup aneh karena dengan asumsi bahwa nilai tes masuk hanya boleh diberi nilai dalam rentang 0-100 maka tentu ketiga data tersebut salah dan harus diperbaiki.

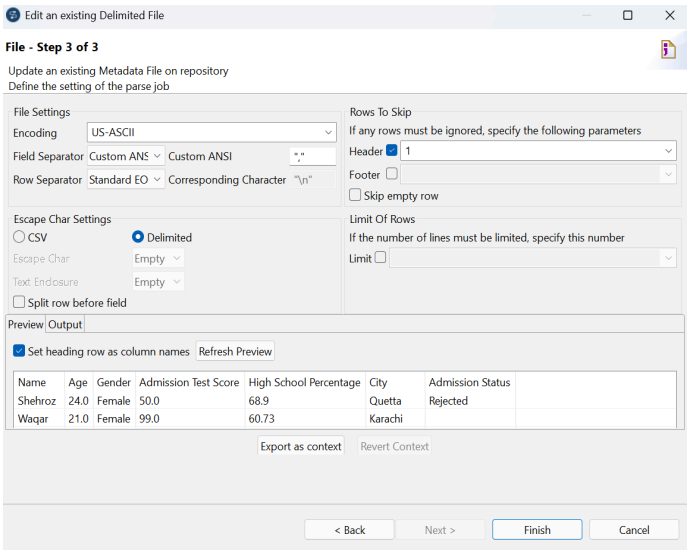
Name	Age	Gender	Admission Test Score	High School Percentage	City	Admission Status
Shafiq	17	Male	78	-10	Quetta	Rejected
Maryam	19	Female	74	110.5	Lahore	Accepted

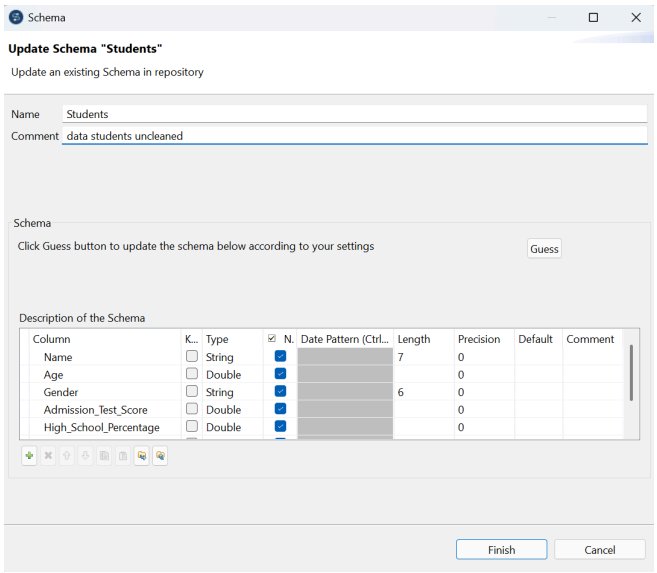
Hal yang sama juga terjadi pada kolom persentase sekolah. Dengan asumsi bahwa persentase suatu sekolah hanya boleh dinilai dengan nilai dalam rentang 0-100, maka kedua data tersebut aneh karena memiliki persentase sekolah di bawah 0 dan di atas 100.

- Terkait format dan tipe data: sudah konsisten dengan kolom yang bersesuaian, seperti nama, jenis kelamin, kota, dan status penerimaan memiliki tipe data string (artinya tidak ada kolom yang terisi dengan numerik). Untuk kolom umur, nilai tes masuk, dan persentase sekolah juga sudah sesuai terisi dengan data numerik (artinya tidak ada kolom yang terisi dengan string).

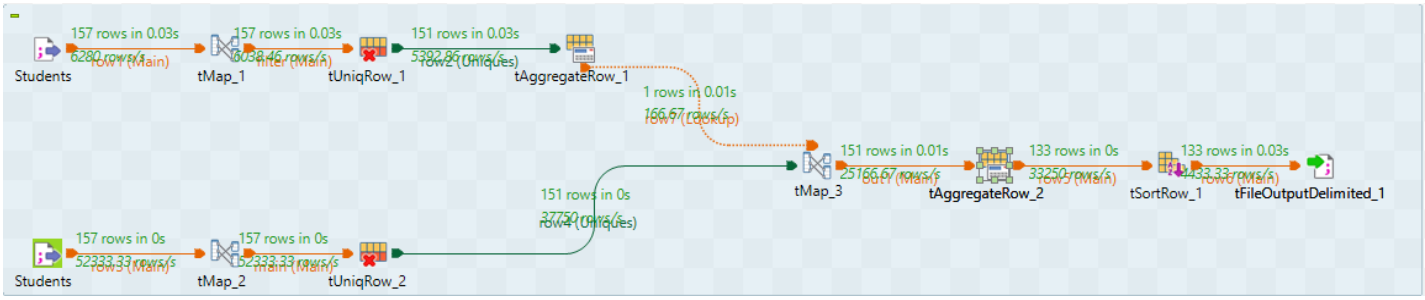
Setelah melakukan eksplorasi data sederhana, selanjutnya adalah mengekstrak data ke dalam *software* Talend. Ketika mengekstrak data ke dalam Talend, penulis menggunakan metadata agar lebih memudahkan proses ekstraksi ke dalam Talend dan menyesuaikan tipe datanya menjadi:

- *Name*: string
- *Age*: *double*
- *Gender*: string
- *Admission test score*: *double*
- *Highschool percentage*: *double*
- *City*: string
- *Admission status*: string.

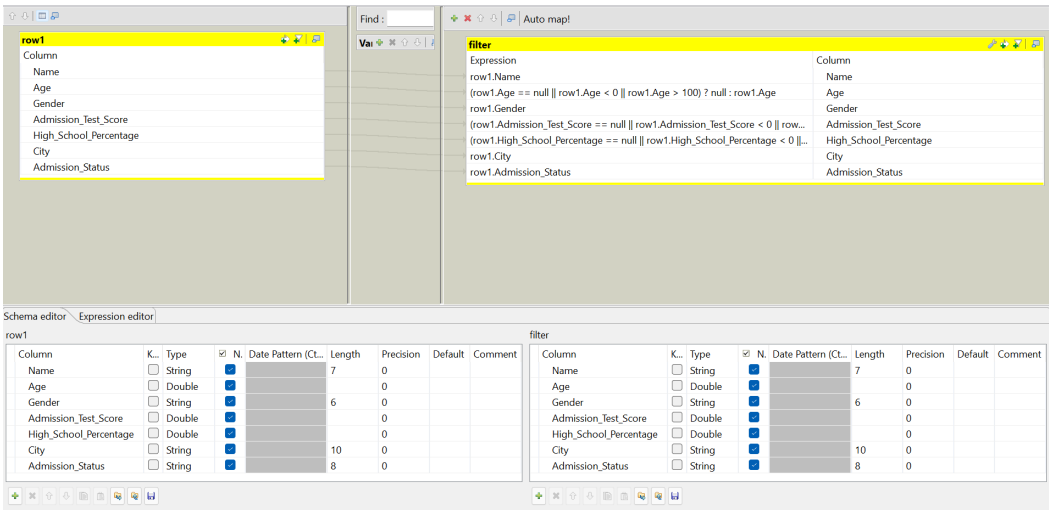




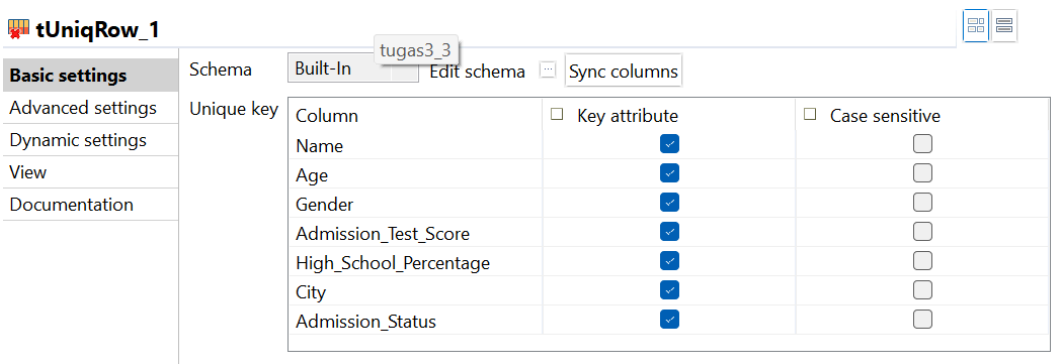
Keseluruhan proses pembersihan data terbagi menjadi 2 alur. Alur pertama digunakan untuk agregasi dan alur kedua merupakan duplikat dari alur pertama namun tanpa agregasi. Cuplikan mengenai alur dapat dilihat pada gambar di bawah ini



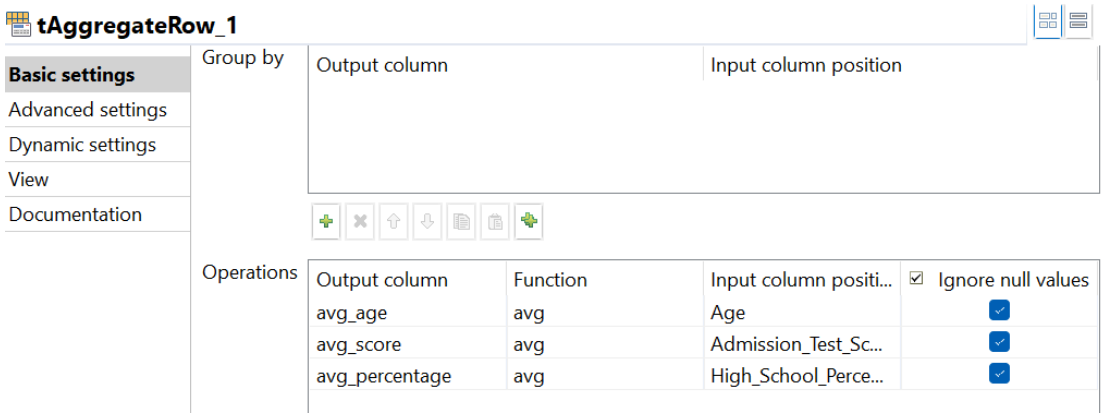
Akan dibahas alur mengenai pertama terlebih dahulu. Sebelum membersihkan data, untuk memastikan bahwa data hilang pada kolom numerik (age, admission test score, dan highschool percentage) dianggap sebagai null dan mengubah data numerik dengan nilai negatif dan lebih dari 100, data dihubungkan dengan tMap dan pada kolom age, admission test score, dan highschool percentage diberikan kode sebagai berikut:
(row1.Age == null || row1.Age < 0 || row1.Age > 100) ? null : row1.Age



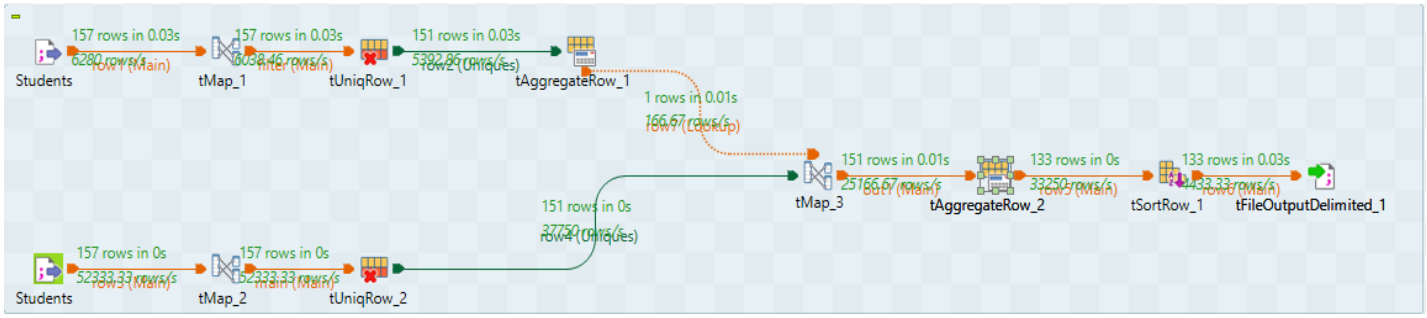
Sehingga melalui proses ini, setiap data kosong dan data yang memiliki nilai negatif serta lebih dari 100 dianggap sebagai null. Setelah itu, baru akan ditangani tentang data duplikat terlebih dahulu dengan menggunakan tUniqRow. Pengaturan dalam tUniqRow dapat dilihat pada gambar di bawah ini



Dengan menceklis setiap kolom, hal ini memastikan bahwa setiap data yang memiliki informasi sama persis dibuang salah satunya. Setelah itu, data akan diolah menggunakan tAggregatedRow. Agregasi dilakukan karena ingin menghitung nilai rata-rata dari setiap data numerik untuk mengisi data yang hilang dan data yang salah. Metrik rata-rata dipilih karena terbilang cukup aman (artinya sebaran data dapat menjadi normal dan mengurangi resiko outlier) digunakan sebagai inputasi. Cuplikan mengenai pengaturan dalam tAggregatedRow dapat dilihat pada gambar berikut



Terdapat 3 kolom baru yang ditambahkan, yaitu avg_age untuk menyimpan nilai rata-rata dari kolom age, avg_score untuk menyimpan nilai rata-rata dari kolom admission test score, dan avg_percentage untuk menyimpan nilai rata-rata dari kolom highschool percentage. Checkbox ignore nul values diaktifkan agar Talend menghiraukan nilai kosong yang ada di kolom age, admission test score, dan highschool percentage. Nantinya tAggregateRow akan dihubungkan dengan tMap kedua untuk proses inputasi.



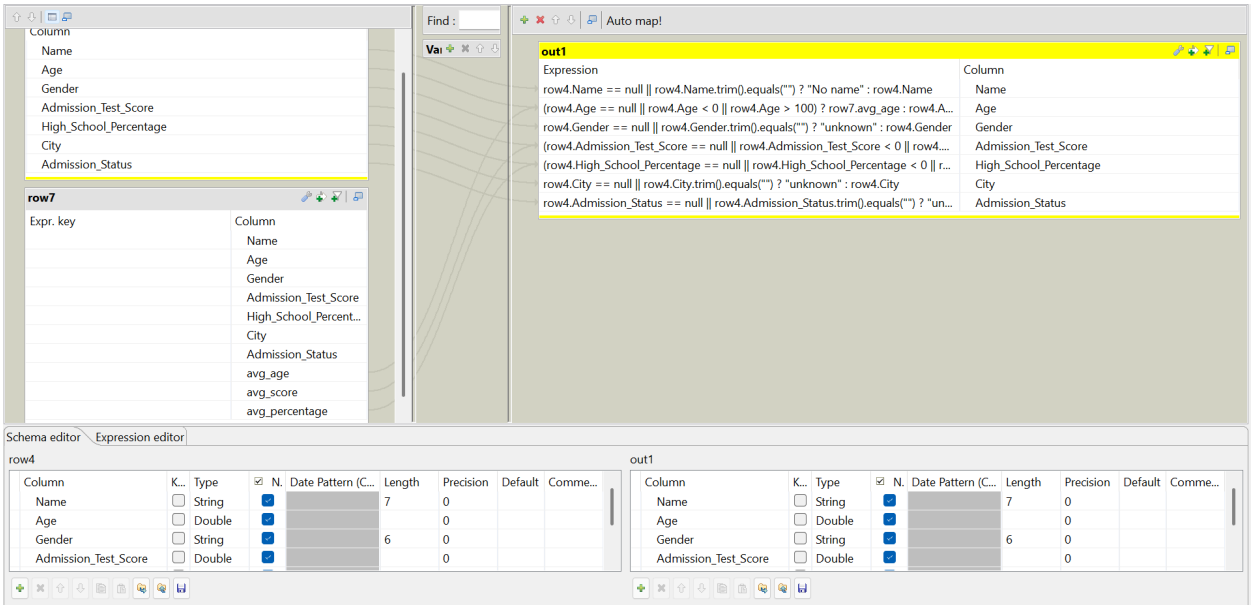
Perhatikan kembali gambar alur di atas, jika dilihat setelah tAggregateRow hanya mengeluarkan keluaran berupa 1 baris. Satu baris tersebut hanya memuat informasi avg_age, avg_score, dan avg_percentage. Hal ini lah yang membuat penulis memutuskan untuk mebuat alur kedua. Jika dilihat kembali, alur kedua hampir sama dengan alur pertama namun tanpa proses agregasi dan langsung menghubungkannya dengan tMap_3. Di dalam tMap_3 terjadi beberapa proses seperti:

- Mengisi data yang hilang pada kolom nama menjadi “No name”, penulis memutuskan untuk tidak membuang baris yang tidak memiliki nama karena terdapat kemungkinan data dapat digunakan sebagai bahan untuk pengolahan data lebih lanjut seperti

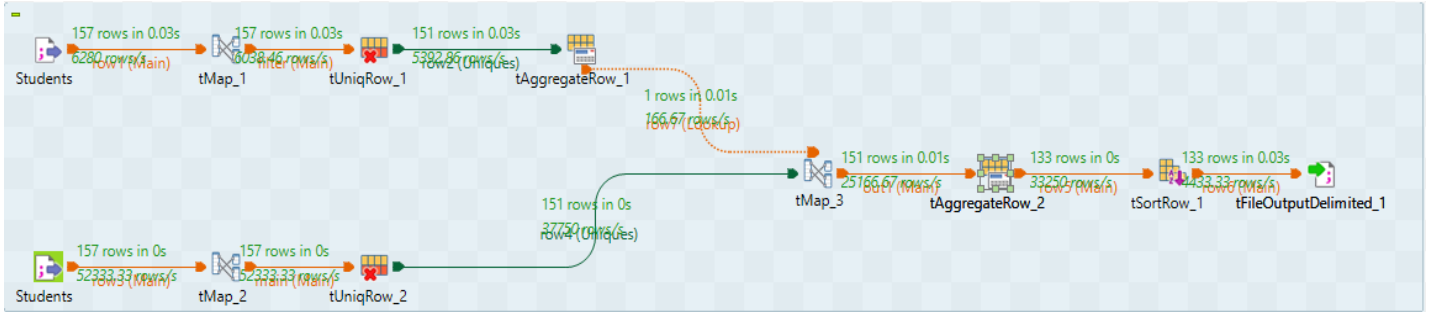
membuat model machine learning dan sejenisnya. Kode yang digunakan: row4.Name == null || row4.Name.trim().equals("") ? "No name" : row4.Name

- Mengisi data hilang pada kolom gender, city, dan admission status menjadi “none”. Seperti namanya, “unknown” artinya tidak diketahui. Dengan asumsi bahwa, penginput data (murid) lupa atau memang sengaja tidak mengisi informasi mengenai gender dan city. Selain itu, data yang hilang di kolom admission status diberi nilai “unknown” karena terdapat kemungkinan bahwa murid tersebut memang status penerimaannya belum diketahui. Kode yang digunakan: row4.Gender == null || row4.Gender.trim().equals("") ? "none" : row4.Gender
  row4.Admission_Status == null || row4.Admission_Status.trim().equals("") ? "unknown" : row4.Admission_Status
- Mengisi data yang hilang pada kolom age, admission test score, dan highschool percentage dengan rata-rata yang sudah diihitung saat proses agregasi di tAggregateRow. Kode yang digunakan: (row4.Age == null || row4.Age < 0 || row4.Age > 100) ? row7.avg_age : row4.Age

Cuplikan mengenai ketiga proses tersebut dapat dilihat pada gambar di bawah ini



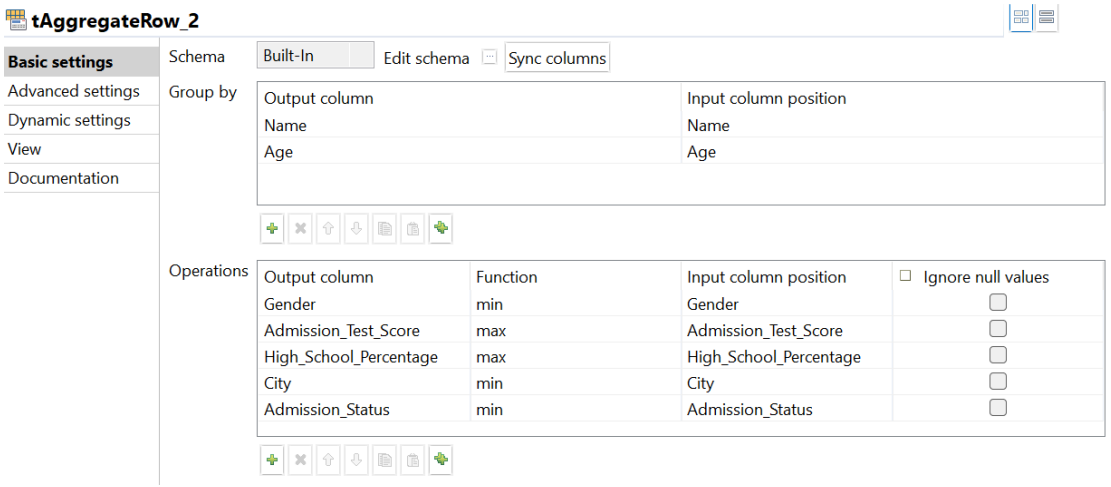
Kembali dengan alur dalam Talend



Setelah melalui proses di tMap_3, proses dilanjutkan dengan tAggrageteRow_2. Agregasi ini dilakukan untuk mengatasi permasalahan yang sudah disinggung di awal mengenai keanehan dalam data duplikat, seperti pada gambar di bawah ini

Name	Age	Gender	Admission Test Score	High School Percentage	City	Admission Status
Hassan	24	Female	79	75.67		
Hassan	24		79	75.67	Karachi	Accepted

Jika dilihat sekilas, tentu bentuk data seperti gambar memang duplikat namun informasi mengenai murid bernama “Hassan” terbagi menjadi dua kolom. Oleh karena itu dalam tAggregateRow_2 diatur seperti pada gambar berikut



Penulis melakukan Group by pada kolom name dan age, lalu melakukan operasi Max pada kolom admission test score dan highschool percentage karena ingin diambil nilai terbesarnya sedangkan untuk kolom gender, city, dan admission status menggunakan operasi Min karena urutan alfabet dalam Talend disesuaikan dengan urutan ASCII, di mana huruf kapital diurutkan lebih dulu daripada huruf kecil sehingga dalam urutan ASCII, huruf kapital memiliki nomor urut lebih kecil dibandingkan dengan huruf biasa/kecil. Sehingga data milik “Hassan” menjadi seperti berikut

Hassan	24	Female	79	75.67	Karachi	Accepted

Setelahnya, baru menggunakan tSortRow untuk mengurutkan data berdasarkan kolom name dan diurutkan mulai dari alfabet A (ascending). Hasil akhir yang dikeluarkan terdapat 133 baris.