GGMD - Exercices applicatifs 2
Évaluation de requêtes réparties

UCBL - Département Informatique de Lyon 1 – 2023

L'objectif de ces exercices est de vous sensibiliser au calcul de coût de transfert de données d'une requête répartie.

On considère toujours la base de données *insee_deces* qui a été fragmenté et dont le schéma global est le suivant :

- personne (idp, nom, prenoms, datenaiss, #lieunaiss, datedeces, #lieudeces, age)
- region (reg, nom, #cheflieu, zone)
- departement (dep, nom, #cheflieu, #reg)
- commune (com, nom, #dep)
- mairie (#codeInsee, cp, nomOrga, nomCom, email, tel, url, adresse, latitude, longitude, dateMaj)

reposant sur des fragments définis comme suit :

$region_i = \sigma_{zone=i}(region), i \in \{1;2;3\}$
$departement_i = departement \ltimes_{reg} region_i, i \in \{1;2;3\}$
$commune_i = commune \ltimes_{dep} departement_i, i \in \{1;2;3\}$
$personne\_age = \Pi_{idp,age}(personnes)$
$personne\_info\_naiss = \Pi_{idp,nom,prenom,datenaiss,lieunaiss}(personnes)$
$personne\_info\_deces = \Pi_{idp,nom,prenom,datedeces,lieudeces}(personnes)$
$pers\_naiss_i = personne\_info\_naiss \ltimes_{lieunaiss=com} commune_i, i \in \{1;2;3\}$
$pers\_deces_i = personne\_info\_deces \ltimes_{lieudeces=com} commune_i, i \in \{1;2;3\}$
$mairie_i = mairie$

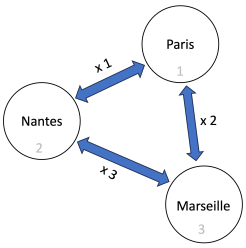
A noter, que l'on considère ici la base contenant des données 'propres' avec les contraintes d'intégrités valides.
Pour rappel, tous les fragments hébergés sur le site de Paris (P) sont indexés par 1, ceux hébergés sur le site de Nantes (N) sont indexés par 2 et ceux hébergés sur le site de Marseille (M) sont indexés par 3.
Pour les besoins de l'exercice, la table mairie n'a pas était répliquée mais migrée sur le site de Paris.

On suppose que chaque attribut vaut 0.5 Ko. (hypothèse simplificatrice pour les calculs)

On spécifie les coûts de transfert par la fonction tr(T, X, Y) qui retourne le coût pour transférer d'un ensemble de tuples T depuis le site X vers le site Y.

Pour l'exercice, on suppose que :

tr(T, P, N) = card(T) × taille(T)
tr(T, P, M) = 2 × card(T) × taille(T)
tr(T, N, M) = 3 × card(T) × taille(T)
tr(T, X, Y) = tr(T, Y, X)



A noter que des informations statistiques sur les données et les opérateurs sont précisées en annexe.

Exercice 1 : Jointures inter-site

On considère la requête Q1b émise sur le site de Nantes : "Donner les personnes (idp, nom, prénoms) nées dans la région 'Hauts-de-France' et décédées dans la région 'Occitanie'".
NB : En guise de révision, vous pourrez refaire ce TD, avec la requête Q1 pour les personnes nées en 'Auvergne-Rhône-Alpes' et décédées dans la région 'Pays de la Loire'.

Vous allez considérer deux plans d'exécution différents. Le plan P1 consiste à transférer l'ensemble des fragments sur le site de Nantes pour effectuer les calculs sur le seul site de Nantes. Le plan P2 consiste à effectuer le maximum des calculs sur les sites stockant les données.

Plan P1

- Transfert de region_1, departement_1, commune_1 et pers_naiss_1 de Paris à Nantes.
- Transfert de region_3, departement_3, commune_3 et pers_naiss_3 de Marseille à Nantes.
- Sur le site de Nantes :
  - Calcul de la sélection $\sigma_{nom='Hauts-de-France'}(\text{region}_1)$, on obtient T11.
  - Calcul de la jointure entre T11 et departement_1, on obtient T12.
  - Calcul de la jointure entre T12 et commune_1, on obtient T13.
  - Calcul de la jointure entre T13 et pers_naiss_1, on obtient T14.
  - Calcul de la projection $\Pi_{idp,nom,prenoms}(\text{T14})$, on obtient T1.
- Sur le site de Nantes :
  - Calcul de la sélection $\sigma_{nom='Occitanie'}(\text{region}_3)$, on obtient T21.
  - Calcul de la jointure entre T31 et departement_3, on obtient T32.
  - Calcul de la jointure entre T32 et commune_3, on obtient T33.
  - Calcul de la jointure entre T33 et pers_naiss_3, on obtient T34.
  - Calcul de la projection $\Pi_{idp,nom,prenoms}(\text{T34})$, on obtient T3.
- Sur le site de Nantes :
- Calcul de l'intersection entre T1 et T3, on obtient le résultat.

Plan P2

- Sur le site de Paris :
  - Calcul de la sélection $\sigma_{nom='Hauts-de-France'}(\text{region}_1)$, on obtient T11.
  - Calcul de la jointure entre T11 et departement_1, on obtient T12.
  - Calcul de la jointure entre T12 et commune_1, on obtient T13.
  - Calcul de la jointure entre T13 et pers_naiss_1, on obtient T14.
  - Calcul de la projection $\Pi_{idp,nom,prenoms}(\text{T14})$, on obtient T1.
- Transfert de T1 de Paris à Nantes.
- Sur le site de Marseille :
  - Calcul de la sélection $\sigma_{nom='Occitanie'}(\text{region}_3)$, on obtient T31.
  - Calcul de la jointure entre T31 et departement_3, on obtient T32.
  - Calcul de la jointure entre T32 et commune_3, on obtient T33.
  - Calcul de la jointure entre T33 et pers_naiss_3, on obtient T34.
  - Calcul de la projection $\Pi_{idp,nom,prenoms}(\text{T34})$, on obtient T3.
- Transfert de T3 de Marseille à Nantes.
- Sur le site de Nantes :
- Calcul de l'intersection entre T1 et T3, on obtient le résultat.

1. A partir des statistiques proposées en Annexe, calculer les coûts des transferts de données du plan P1.

2. A partir des statistiques proposées en Annexe, calculer les coûts des transferts de données du plan P2.
3. Quelle modification proposeriez-vous au plan P2 pour réduire son coût de transfert ? Proposer un plan P3 et calculer son coût.

Exercice 2 : Algorithme des semi-jointures

On considère la requête Q'1 émise depuis Paris : "Donner les personnes (idp, nom, prénoms, datenaiss, lieunaiss, datedeces, lieudeces) nées dans la région 'Hauts-de-France' et décédées dans la région 'Occitanie'"

On considère le plan d'exécution P4 suivant :

— Sur le site de Marseille :
  — Calcul de la sélection $\sigma_{nom='Occitanie'}(\text{region}_3)$, on obtient T31.
  — Calcul de la jointure entre T31 et departement_3, on obtient T32.
  — Calcul de la jointure entre T32 et commune_3, on obtient T33.
  — Calcul de la jointure entre T33 et pers_deces_3, on obtient T34.
  — Calcul de la projection $\Pi_{idp,lieudeces,datedeces}$(T34), on obtient T3.
— Transfert de T3 de Marseille à Paris.
— Sur le site de Paris :
  — Calcul de la jointure entre T1 et T3, on obtient le résultat.

1. Vérifier que le coût de transfert du plan P4 est de 6 629 577 Ko
2. Proposer P5 un plan d'exécution de Q'1 intégrant une gestion de la jointure selon l'algorithme des semi-jointures.
3. Calculer le coût de P5.

Annexe

On suppose disposer des informations suivantes :

— H1 : Card(region) = 18 ;
— H2 : Card(region_1) = 4 ;
— H3 : Card(region_2) = 9 ;
— H4 : Card(region_3) = 5 ;
— H5 : Card(departement) = 101 ;
— H6 : Card(departement_1) = 28 ;
— H7 : Card(departement_2) = 28 ;
— H8 : Card(departement_3) = 45 ;
— H9 : Card(commune) = 37 600
— H10 : Card(commune_1) = 12 849 ;
— H11 : Card(commune_2) = 8 028 ;
— H12 : Card(commune_3) = 14 123 ;
— H13 : Card(personnes) = 24 811 296
— H14 : Card(pers_naiss_1) = 8 634 189 ;
— H15 : Card(pers_naiss_2) = 5 196 351 ;
— H14 : Card(pers_naiss_3) = 7 444 568 ;
— H15 : Card(Q1) = 54 357 tuples
— H16 : Card(Q2) = 509 tuples
— H17 : Card(Q3) = 3 036 725 tuples

Statistiques sur base INSEE

zone	région	code dept	nom dept	Nb communes par dept	Nb communes par région	Nb pers. nées dans dept	Nb pers. nées dans région	Nb pers. mortes dans dept	Nb pers. mortes dans région
1	Grand Est	8	Ardennes	449	5121	154451	2342875	121493	2244779
		10	Aube	431		128841		127627	
		51	Marne	613		239082		243409	
		52	Haute-Marne	426		104996		90949	
		54	Meurthe-et-Moselle	591		319444		316826	
		55	Meuse	499		123788		85665	
		57	Moselle	725		371141		380006	
		67	Bas-Rhin	514		394780		408873	
		68	Haut-Rhin	366		272651		288539	
		88	Vosges	507		233701		181392	
	Hauts-de-France	2	Aisne	799	3788	296508	2716392	236579	2442723
		59	Nord	648		1128276		1076413	
		60	Oise	679		234596		248863	
		62	Pas-de-Calais	890		779854		610482	
		80	Somme	772		277158		270386	
	Normandie	14	Calvados	528	2652	201731	1280909	280064	1319176
		27	Eure	585		174497		184588	
		50	Manche	446		218741		201665	
		61	Orne	385		134352		127337	
		76	Seine-Maritime	708		551588		525522	
	Île-de-France	75	Paris	21	1288	1456630	2294013	828486	3332669
		77	Seine-et-Marne	507		196757		335696	
		78	Yvelines	259		179621		344292	
		91	Essonne	194		52839		309516	
		92	Hauts-de-Seine	36		140751		461266	
		93	Seine-Saint-Denis	40		113011		339794	
		94	Val-de-Marne	47		88779		420731	
		95	Val-d'Oise	184		65625		292888	

zone	région	code dept	nom dept	nb communes		Nb pers. nées dans dept	Nb pers. nées dans région	Nb pers. mortes dans dept	Nb pers. mortes dans région
2	Bourgogne-Franche-Comté	21	Côte-d'Or	698	3700	177834	1247460	217298	1260842
		25	Doubs	571		180113		195752	
		39	Jura	494		125963		107837	
		58	Nièvre	309		140526		141676	
		70	Haute-Saône	539		129060		91391	
		71	Saône-et-Loire	565		308236		274552	
		89	Yonne	423		130899		175393	
		90	Territoire de Belfort	101		54829		56943	
	Bretagne	22	Côtes-d'Armor	348	1207	337310	1495028	276793	1413708
		29	Finistère	277		483530		459555	
		35	Ille-et-Vilaine	333		320435		357897	
		56	Morbihan	249		353753		319463	
	Centre-Val de Loire	18	Cher	287	1757	163793	959279	156859	1080333
		28	Eure-et-Loir	365		142939		162274	
		36	Indre	241		148072		129626	
		37	Indre-et-Loire	272		177170		245463	
		41	Loir-et-Cher	267		135833		143429	
		45	Loiret	325		191472		242682	
	Guadeloupe	971	Guadeloupe	32	32	91658	91658	85460	85460
	Guyane	973	Guyane	22	22	13989	13989	20723	20723
	La Réunion	974	La Réunion	24	24	128711	128711	127657	127657
	Martinique	972	Martinique	34	34	96016	96016	83614	83614
	Mayotte	976	Mayotte	17	17	928	928	5315	5315
	Pays de la Loire	44	Loire-Atlantique	207	1235	366478	1163282	460085	1286782
		49	Maine-et-Loire	177		178728		258117	
		53	Mayenne	240		151399		113660	
		72	Sarthe	354		231507		230831	
		85	Vendée	257		235170		224089	

zone	région	code dept	nom dept	nb communes		Nb pers. nées dans dept	Nb pers. nées dans région	Nb pers. mortes dans dept	Nb pers. mortes dans région
3	Auvergne-Rhône-Alpes	1	Ain	393	4038	160756	2420642	173707	2799730
		3	Allier	317		194277		211010	
		7	Ardèche	335		164735		137859	
		15	Cantal	246		110983		89831	
		26	Drôme	363		137814		180764	
		38	Isère	512		257123		346578	
		42	Loire	323		323437		347349	
		43	Haute-Loire	257		147918		103129	
		63	Puy-de-Dôme	464		244751		289571	
		69	Rhône	276		422207		579873	
		73	Savoie	273		120210		141945	
		74	Haute-Savoie	279		136431		198114	
	Corse	2A	Corse-du-Sud	124	360	436	976	30707	65472
		2B	Haute-Corse	236		540		34765	
	Nouvelle-Aquitaine	16	Charente	364	4309	166576	2325198	167401	2637631
		17	Charente-Maritime	463		233064		280420	
		19	Corrèze	279		158144		145111	
		23	Creuse	256		115718		88552	
		24	Dordogne	503		211155		203372	
		33	Gironde	535		401090		588383	
		40	Landes	327		149174		157361	
		47	Lot-et-Garonne	319		114818		150016	
		64	Pyrénées-Atlantiques	546		241805		302822	
		79	Deux-Sèvres	256		157572		149311	
		86	Vienne	266		171905		188632	
		87	Haute-Vienne	195		204177		216250	
	Occitanie	9	Ariège	327	4454	82679	1715716	74982	2209859
		11	Aude	433		144749		158827	
		12	Aveyron	285		181582		144739	
		30	Gard	351		195057		256988	
		31	Haute-Garonne	586		208307		393762	
		32	Gers	461		87957		86673	
		34	Hérault	342		238137		395533	
		46	Lot	313		78445		85137	
		48	Lozère	152		57166		37371	
		65	Hautes-Pyrénées	469		96764		120179	
		66	Pyrénées-Orientales	226		114363		195398	
		81	Tarn	314		151252		168447	
		82	Tarn-et-Garonne	195		79258		91823	
	Provence-Alpes-Côte d'Azur	4	Alpes-de-Haute-Provence	198	962	46813	982036	64174	1916052
		5	Hautes-Alpes	162		51627		55272	
		6	Alpes-Maritimes	163		176369		478070	
		13	Bouches-du-Rhône	135		427835		736891	
		83	Var	153		157329		373617	
		84	Vaucluse	151		122063		208028	