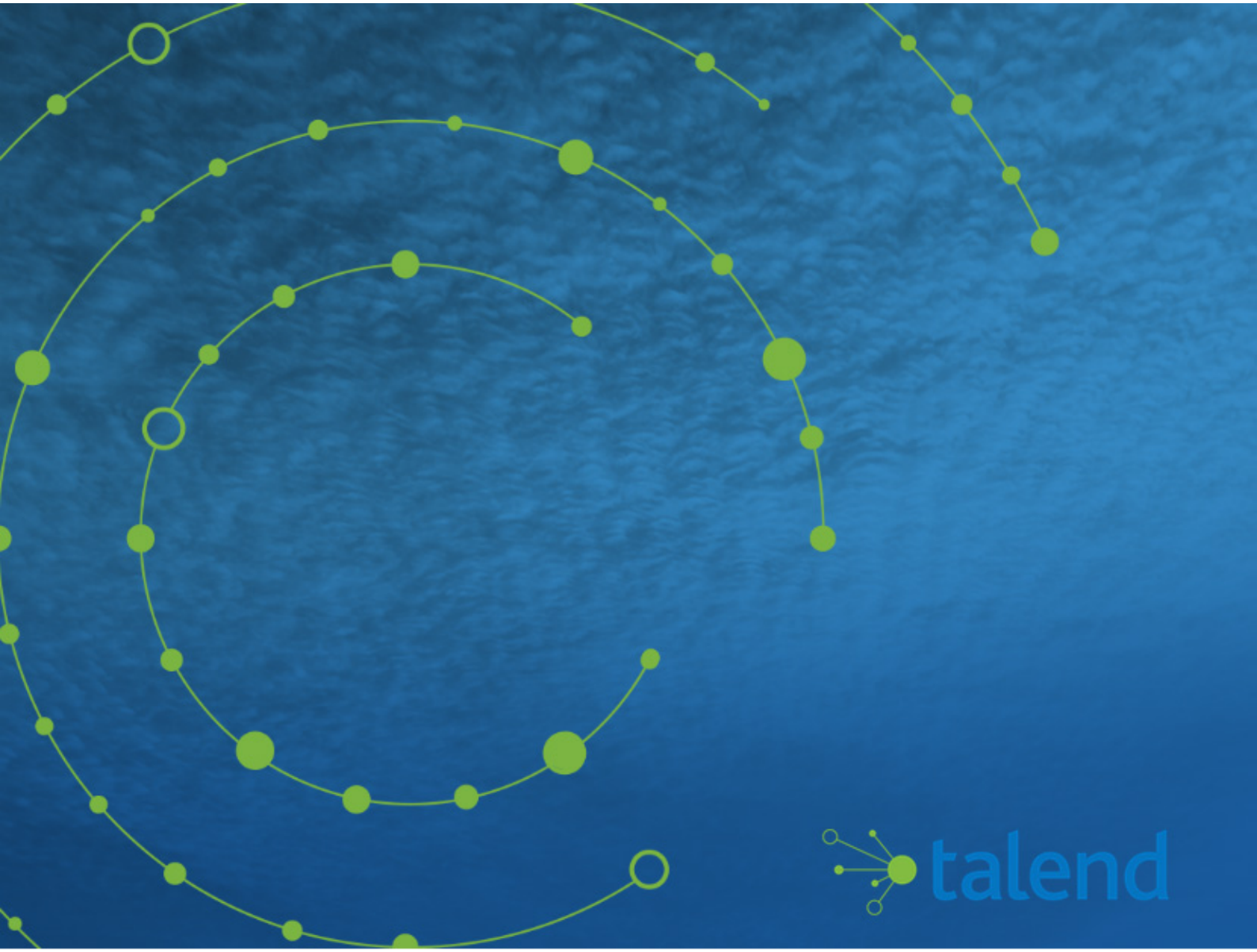


Exemples de Jobs et d'analyse de qualité de données

7.0.1

Table des matières

Copyleft

Convient à la version 7.0.1. Annule et remplace toute version antérieure de ce guide.

Date de publication : 13 avril 2018

Profiling de données clients

Incorporer des outils de qualité de données appropriés au sein de vos processus métier est vital au commencement de tout projet et durant celui-ci, afin de déterminer le type de qualité de vos données et de décider des données à résoudre.

Imaginez, par exemple, que vous démarrez une campagne dans votre service commercial et votre service marketing, ou que vous devez contacter des clients pour la facturation et le paiement et que votre source principale pour contacter les personnes concernées sont les adresses e-mail et postales. Avoir des données cohérentes et correctes est vital dans des campagnes, afin de pouvoir contacter toutes les personnes que vous souhaitez contacter.

Cette section fournit un exemple de profiling d'adresses e-mail et postales de clients des États-Unis.

Identifier des anomalies de données

La première étape dans cet exemple est le profiling de vos ressources, ici les informations de contact des clients, dans une base de données MySQL. Les résultats du profiling vous fournissent des statistiques concernant les valeurs dans chaque colonne.

Profiler les colonnes d'adresses

Utilisez votre Studio Talend pour analyser les colonnes de clients, notamment email et postal.

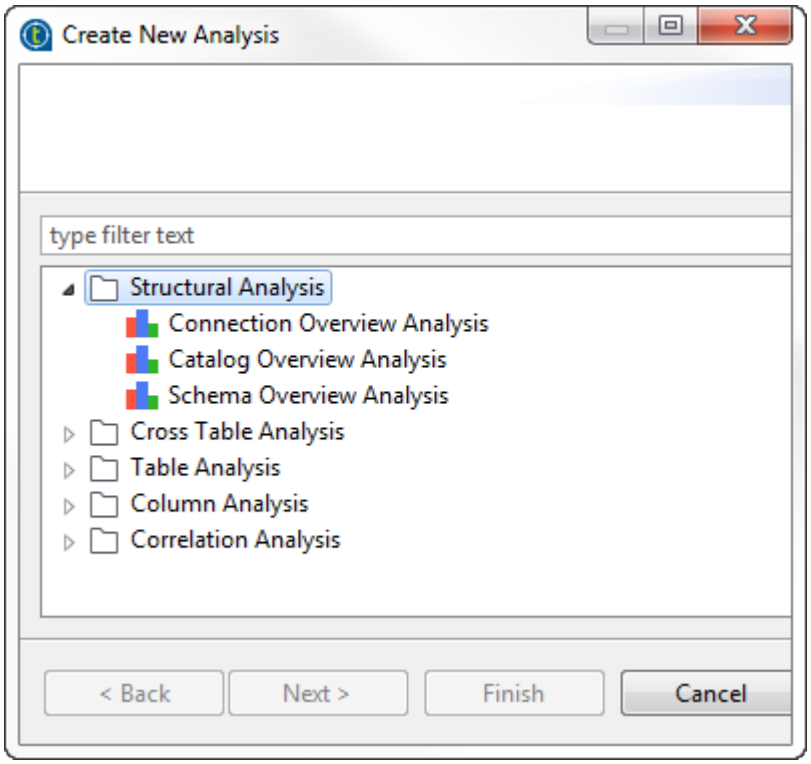
À l'aide d'indicateurs et de modèles natifs sur ces colonnes, les résultats d'analyse affichent les données d'adresses qui correspondent et ne correspondent pas, le nombre d'enregistrements les plus fréquents pour chaque modèle distinct, ainsi que le nombre de lignes, de doublons et de blancs dans chaque colonne.

Définir l'analyse de colonnes

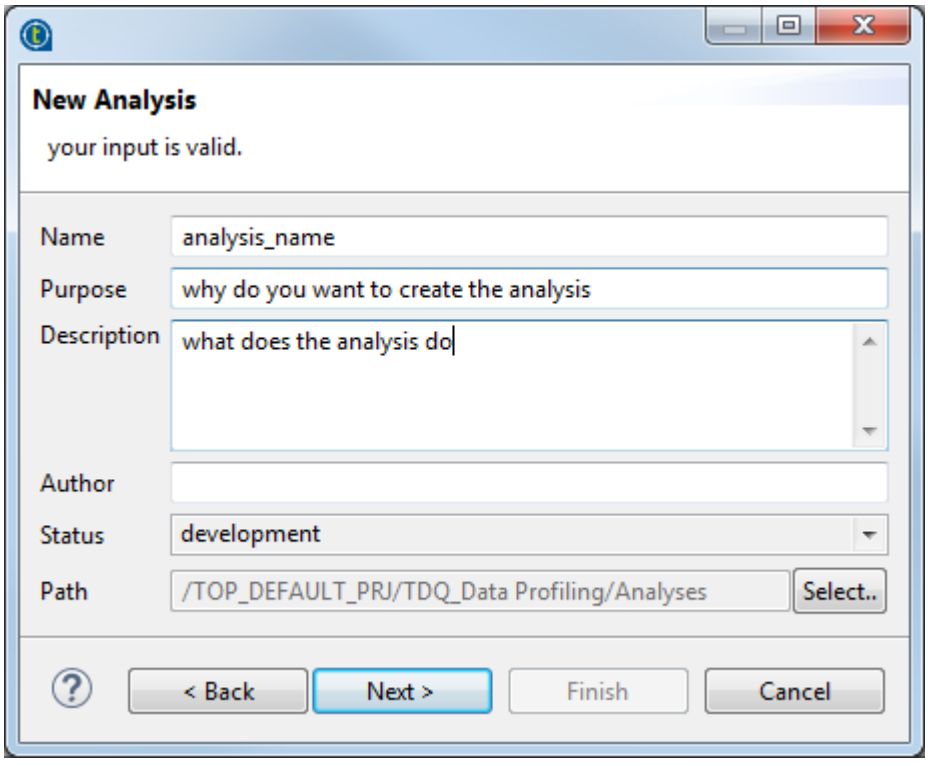
Procédure

1. Dans la vue **DQ Repository**, cliquez-droit sur le dossier **Analyses** et sélectionnez **New Analysis**.

   L'assistant **Create New Analysis** s'ouvre.



2. Commencez à saisir `Basic column analysis` dans le champ de recherche, sélectionnez **Basic Column Analysis** dans la liste et cliquez sur **Next**. Si votre Studio Talend est en français, saisissez `analyse simple de colonne`.



3. Dans le champ **Name**, saisissez un nom pour l'analyse de colonnes.

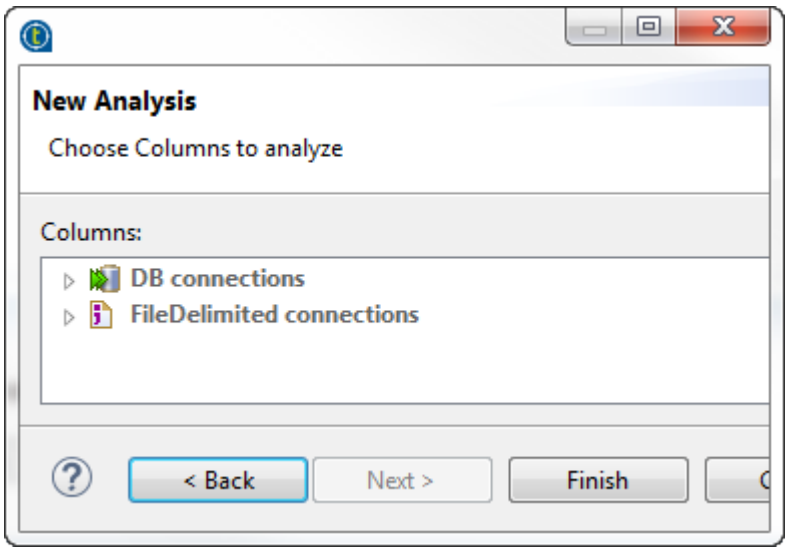
Remarque :

Il est recommandé de ne pas utiliser les caractères spéciaux suivants dans le nom de l'élément, notamment :

"~", "!", "`", "#", "^", "&", "*", "\\", "/", "?", ":", ";", "\"", ".", "(", ")", "'", "¥", "'", """, "«", "»", "<", ">".

Ces caractères seront remplacés par un "_" dans le système de fichiers. Vous risquez ainsi de créer des éléments en doublon.

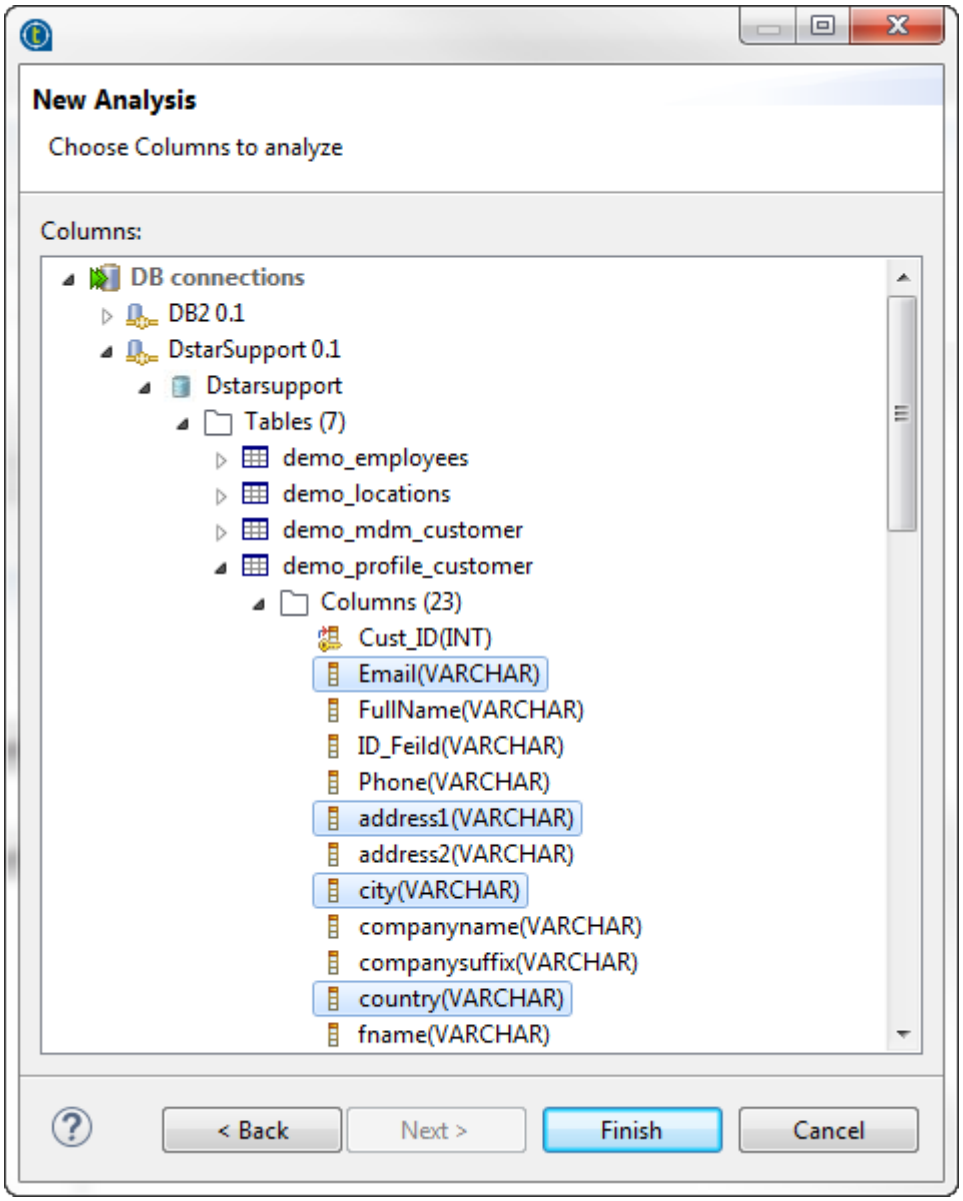
4. Configurez les métadonnées de l'analyse de colonnes (objectif, description et nom de l'auteur) dans les champs correspondants puis cliquez sur **Next**.



Sélectionner les colonnes d'adresse et configurer les données d'exemple

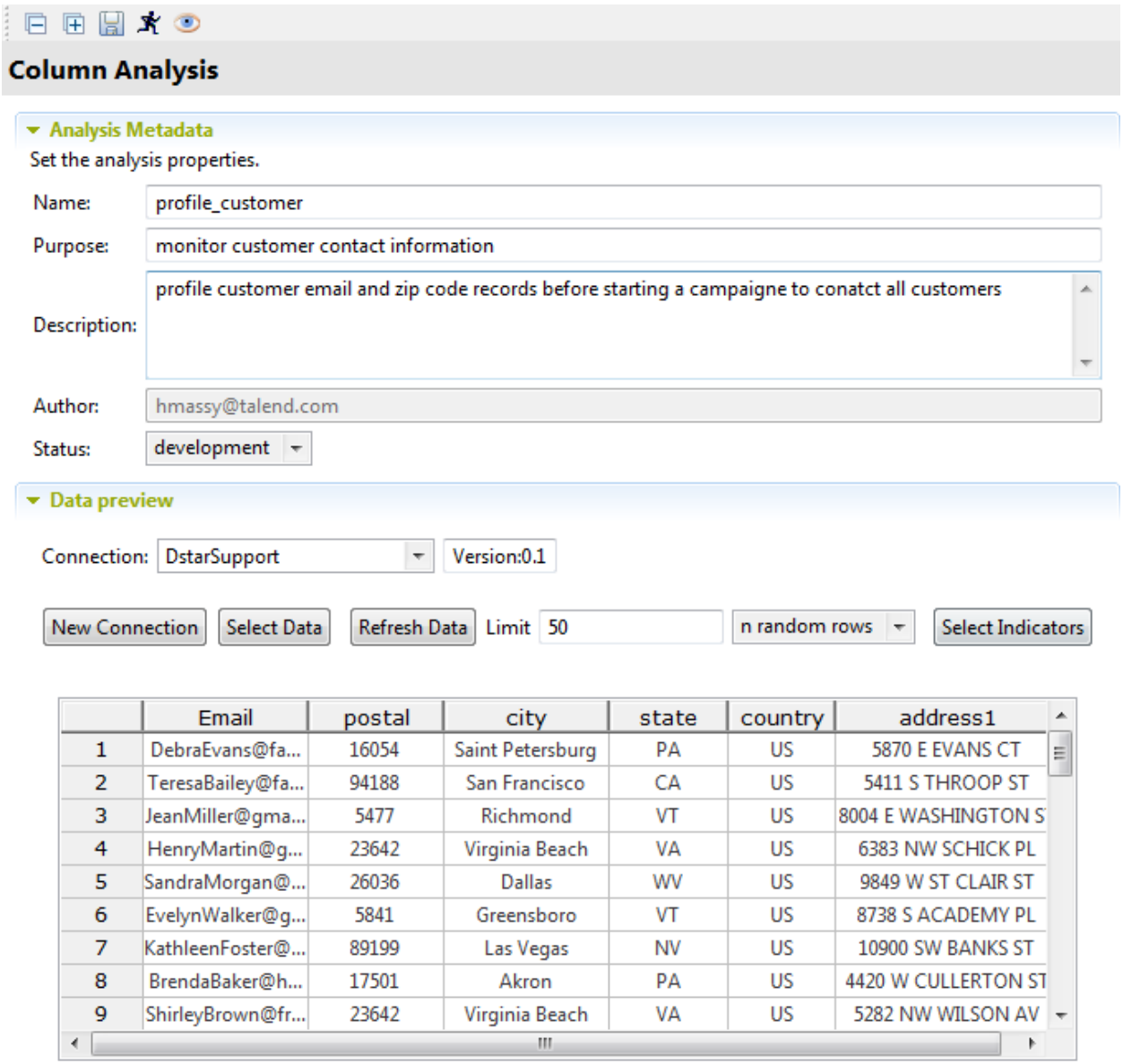
Procédure

1. Développez le nœud **DB connections** et parcourez-le jusqu'aux colonnes d'adresses que vous souhaitez analyser.



2. Sélectionnez les colonnes et cliquez sur **Finish** pour fermer l'assistant.

   Un fichier pour la nouvelle analyse de colonnes s'affiche sous le nœud **Analysis** de la vue **DQ Repository** et l'éditeur d'analyse s'ouvre sur les métadonnées de cette analyse.



3. Dans la vue **Data preview**, cliquez sur **Refresh Data**.

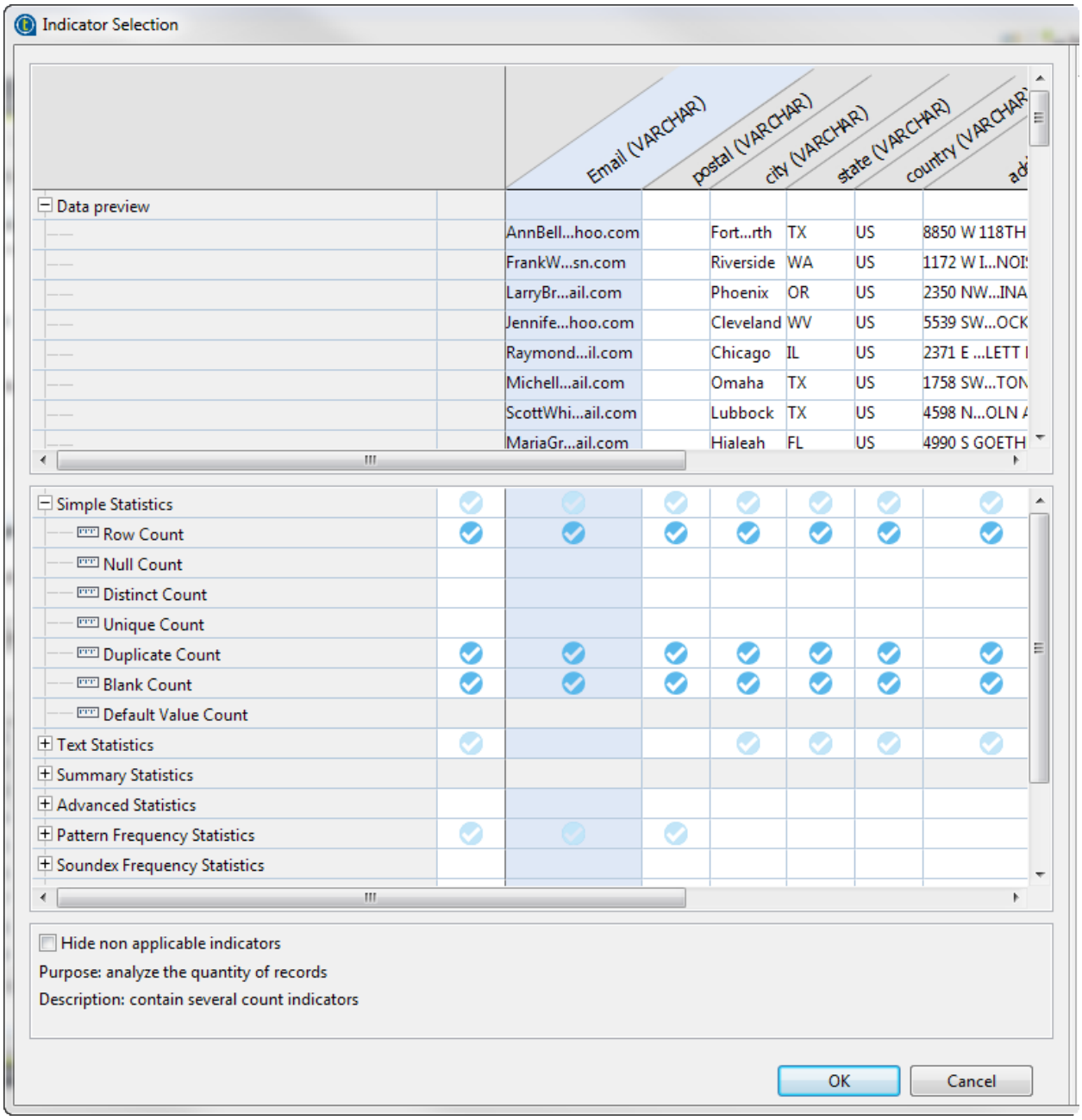
   Les données des colonnes sélectionnées sont affichées dans la table.

   Vous pouvez modifier votre source de données et les colonnes sélectionnées à l'aide des boutons **New Connection** et **Select Data**, respectivement.
4. Dans le champ **Limit**, saisissez `50` pour le nombre d'enregistrement que vous souhaitez afficher dans la table et utiliser comme données d'exemple.
5. Sélectionnez **n random rows** afin de lister `50` enregistrements aléatoires des colonnes sélectionnées.

Configurer les indicateurs système

Procédure

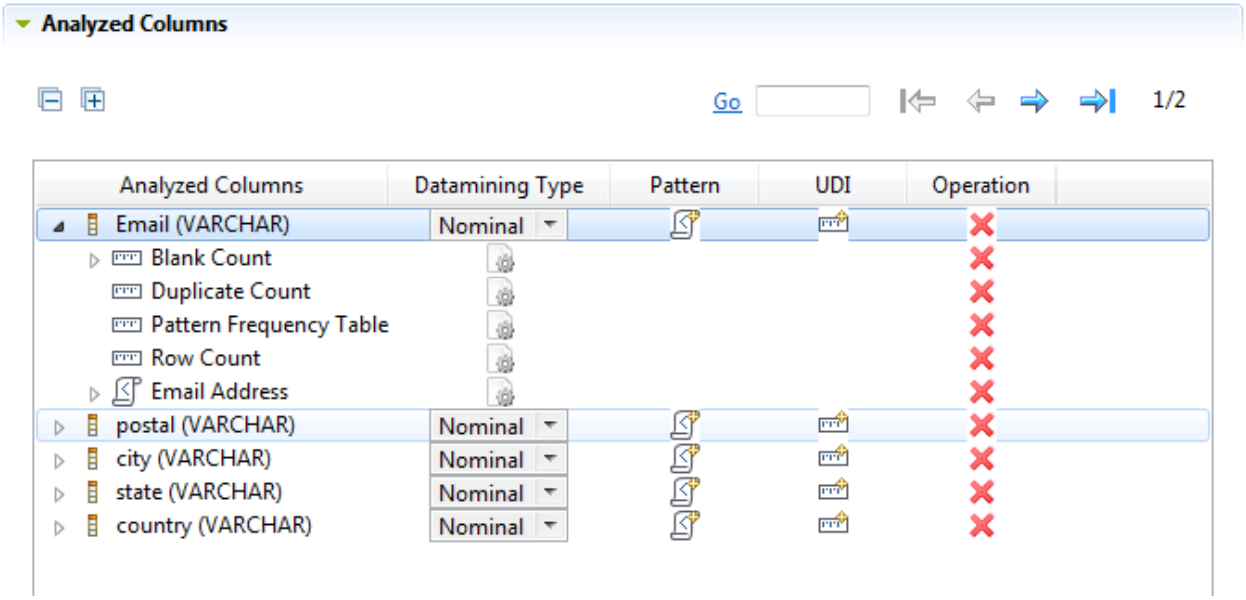
1. Dans la vue **Data preview** de l'éditeur d'analyse, cliquez sur **Select indicators** pour ouvrir la boîte de dialogue **Indicator Selection**.



2. Cliquez dans les cellules à côté des noms d'indicateurs afin de les paramétrer pour les colonnes analysées et cliquez sur **OK**.

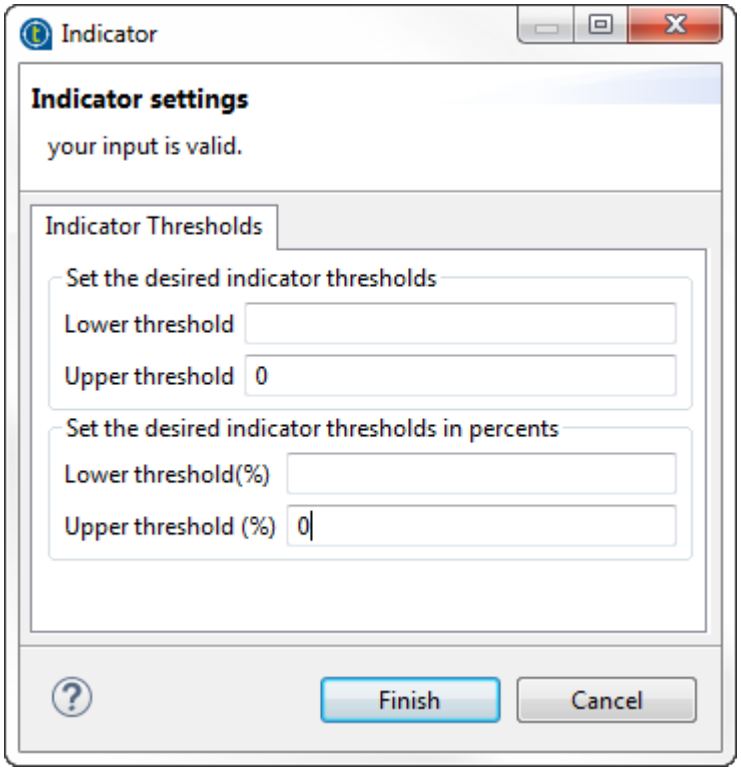
   Dans cet exemple, vous souhaitez consulter le nombre de lignes, de blancs et de doublons dans toutes les colonnes, afin de voir si les données sont cohérentes. L'indicateur **Pattern Frequency Table** est utilisé sur les colonnes email et postal afin de calculer le nombre des enregistrements les plus fréquents pour chaque modèle ou valeur distinct(e).

   Les indicateurs sont ajoutés aux colonnes dans la vue **Analyzed Columns**.



3. Cliquez sur l'icône d'option à côté de l'indicateur **Blank Count** et saisissez `0` dans le champ **Upper threshold**.

   Définir des seuils sur les indicateurs est très utile. Cela permet de marquer en rouge le nombre de valeurs nulles dans les résultats d'analyse.



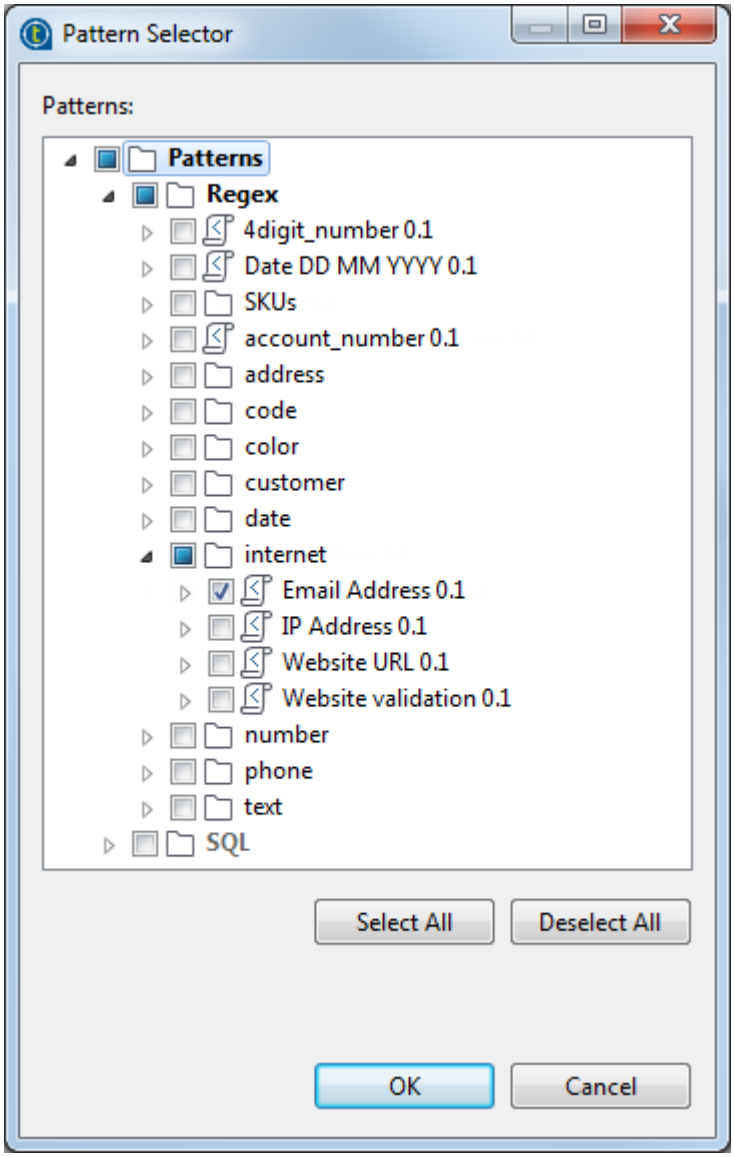
Configurer les modèles

Vous allez mettre en correspondance le contenu de la colonne email par rapport au format d'e-mail standard et le contenu de la colonne postal par rapport au format standard des code postaux des États-Unis.

Cela permet de définir le contenu, la structure et la qualité des adresses e-mail et des codes postaux, ainsi que donner un pourcentage des données correspondant aux formats standards et des données ne correspondant pas.

Procédure

1. Dans la vue **Analyzed Columns**, cliquez sur l'icône à côté de la colonne email.



2. Dans la boîte de dialogue **Pattern Selector**, développez **Regex** et parcourez l'arborescence jusqu'au nœud **Email Address**, dans le dossier **internet**, puis cliquez sur **OK**.
3. Cliquez sur l'icône d'option à côté de l'indicateur **Email Address** et saisissez `98.0` dans le champ **Lower threshold (%)**.

   Si le nombre d'enregistrements correspondant au modèle est inférieur à 98 %, ils seront marqués en rouge dans les résultats de l'analyse.
4. Répétez l'opération pour ajouter la colonne postal au modèle **US Zipcode Validation** depuis le dossier **address**.

   Pour plus d'informations concernant les types de modèles et leur utilisation lors d'analyses de données, consultez le Guide utilisateur du Studio Talend à l'adresse https://help.talend.com.

Exécuter l'analyse et afficher les résultats du profiling

Procédure

1. Sauvegardez l'analyse de colonnes dans l'éditeur d'analyse puis appuyez sur **F6** pour l'exécuter.

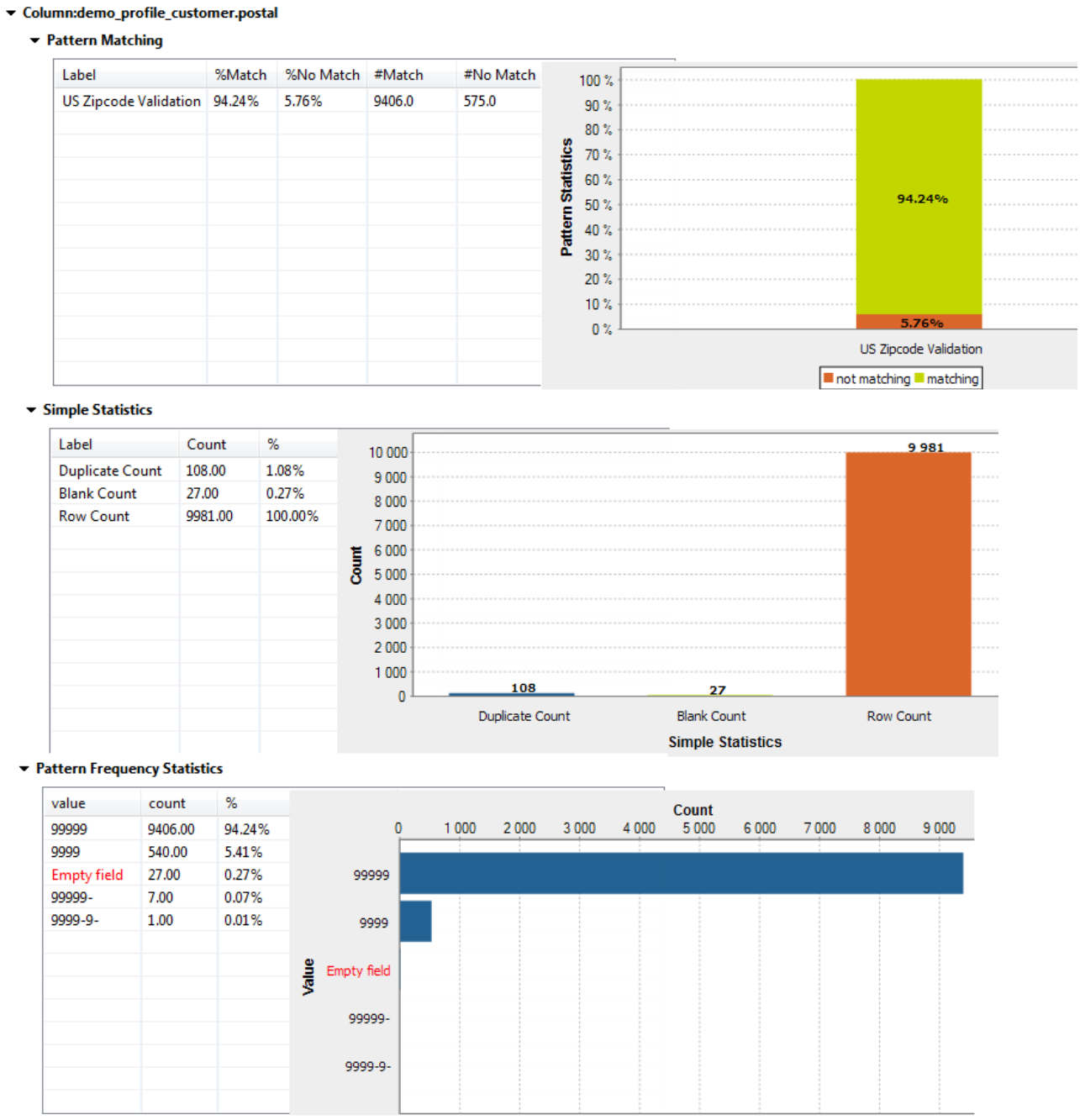
   Un groupe de diagrammes s'affiche dans le panneau **Graphics**, à droite de l'éditeur d'analyse, et montre les résultats de l'analyse de colonnes, notamment ceux de la mise en correspondance des modèles.
2. Cliquez sur l'onglet **Analysis Results** au bas de l'éditeur d'analyse pour accéder à une vue plus détaillée des résultats.

   Ces résultats affichent les graphiques générés pour les colonnes analysées, ainsi que les tables détaillant les résultats des statistiques et des mises en correspondance des modèles.

Résultats

Les résultats pour la colonne email se présentent comme suit :

Les résultats de la colonne postal se présentent comme suit :



Les ensembles de résultats pour la colonne postal donnent le nombre d'enregistrements qui correspondent et le nombre d'enregistrements qui ne correspondent pas au format standard des codes postaux des États-Unis. Les ensembles de résultats donnent également le nombre de blancs et de doublons, ainsi que le nombre d'enregistrements les plus fréquents pour chaque modèle distinct. Ces résultats montrent que les données ne sont pas vraiment cohérentes.

Un certain pourcentage des clients ne peut être contacté ni par e-mail ni par courrier. Ces résultats montrent clairement que vos données ne sont pas vraiment cohérentes et qu'il faut qu'elles soient corrigées.

Visualiser les données analysées

Après exécution de l'analyse de colonnes à l'aide du moteur SQL, dans la vue **Analysis Results** de l'éditeur d'analyse, vous pouvez cliquer-droit sur l'une des barres ou lignes des graphiques et ainsi accéder à une vue des données analysées.

Cela peut être très utile pour voir les lignes invalides, par exemple, et trouver la manière d'ajuster les données.

Procédure

1. Au bas de l'éditeur d'analyse, cliquez sur l'onglet **Analysis Results** pour ouvrir une vue détaillée des résultats d'analyse.
2. Cliquez-droit sur une ligne de données, dans les résultats des statistiques de la colonne email et sélectionnez **View rows**, par exemple.

Résultats

La perspective **Data Explorer** s'ouvre et liste les lignes invalides dans la colonne email.

Cust_ID	FullName	fname	lname	address1	city	state	postal	country	region	Phone	Email
3	Larry Bryant	Larry	Bryant	2350 NW CHINA PL	Phoenix	OR		US	USNWEST	834-367-3427	LarryBryant@gmail.com
23	Sarah Taylor	Sarah	Taylor	3284 W NEWBURG AV	Irving	TX	75063	US	USSOUTH	815-522-9116	SarahTaylor@yahoo.com
91	Alice Torres	Alice	Torres	3656 S PEARSON ST	Santa ...	CA	92799	US	USWEST	212-856-5683	AliceTorres@yahoo.com
119	Marie Murphy	Marie	Murphy	7687 S TRIPP AV	Charlo...	VT	5445	US	<null>	813-723-6681	MarieMurphy@yahoo.com
193	Kevin Garcia	Kevin	Garcia	2529 N CLARENCE AV	Oakland	TX	78951	US	USSOUTH	604-834-6474	KevinGarcia@yahoo.com
210	Scott Brooks	Scott	Brooks	10714 NE SCOTT ST	Fresno	TX	77545	US	USSOUTH	802-359-3898	ScottBrooks@yahoo.com
238	Joyce Cooper	Joyce	Cooper	10004 W PLEASANT ...	Orlando	WV	26412	US	USEAST	212-963-1775	JoyceCooper@gmail.com
298	Karen Thomas	Karen	Thomas	6651 N PONCHART...	Corpus...	TX	78480	US	USSOUTH	816-779-3247	KarenThomas@gmail.com
310	Betty Watson	Betty	Watson	7983 SW FARRAR DR	Omaha	TX	75571	US	USSOUTH	304-701-7857	BettyWatson@yahoo.com
349	Betty Bailey	Betty	Bailey	8360 NE 120TH ST	San Fra...	CA	94188	US	USWEST	622-835-3474	BettyBailey@yahoo.com
378	Nancy Bryant	Nancy	Bryant	10141 S 41ST ST	Saint P...	PA	16054	US	USEAST	623-965-5728	NancyBryant@gmail.com
379	Karen Morgan	Karen	Morgan	10166 E ABERDEEN ST	Kansas...	MO	64999	US	USMWST	303-342-8553	KarenMorgan@gmail.com
393	Nancy Hughes	Nancy	Hughes	10418 SE 45TH ST	Durham	PA	18039	US	USEAST	709-724-5816	NancyHughes@yahoo.com
394	Joyce Parker	Joyce	Parker	5123 W 32ND ST	Aurora	WV	26705	US	USEAST	508-717-4278	JoyceParker@yahoo.com
400	Henry Rogers	Henry	Rogers	6358 NE CHELTENH...	Wichita	KS	67278	US	USMWST	812-873-8641	HenryRogers@yahoo.com
430	Frank Foster	Frank	Foster	8198 W KIRKLAND AV	Phoenix	OR	97535	US	USNWEST	603-876-5331	FrankFoster@gmail.com
446	Betty Brooks	Betty	Brooks	2568 SW CABRINI ST	Cincin...	OH	45999	US	USMWST	307-392-4335	BettyBrooks@yahoo.com
453	Marie Torres	Marie	Torres	3631 SE 47TH PL	Minne...	NC	28652	US	USEAST	715-402-7674	MarieTorres@gmail.com
495	Frank Howard	Frank	Howard	4898 S FARRELL ST	Spokane	WA	99299	US	USNWEST	738-201-3245	FrankHoward@yahoo.com
496	Scott Powell	Scott	Powell	11065 W TORRENCE...	Clevela...	WV	26215	US	USEAST	218-791-8511	ScottPowell@yahoo.com
506	David Morris	David	Morris	7872 SE MAGNET AV	Akron	PA	17501	US	USEAST	734-791-2962	DavidMorris@gmail.com
539	Linda Taylor	Linda	Taylor	7498 S CASTLE ISLA...	Saint L...	OK	74866	US	USWEST	403-861-9549	LindaTaylor@gmail.com
586	Carol Harris	Carol	Harris	10191 SE EDWARD B...	Portland	TX	78374	US	USSOUTH	808-611-4555	CarolHarris@yahoo.com
647	Brian Powell	Brian	Powell	2562 NW CORBETT ...	Durham	PA	18039	US	USEAST	406-889-9681	BrianPowell@yahoo.com
650	Carol Barnes	Carol	Barnes	9642 E HICKORY AV	Detroit	TX	75436	US	USSOUTH	627-747-1229	CarolBarnes@yahoo.com