tExtractKeyValue

tExtractKeyValue properties

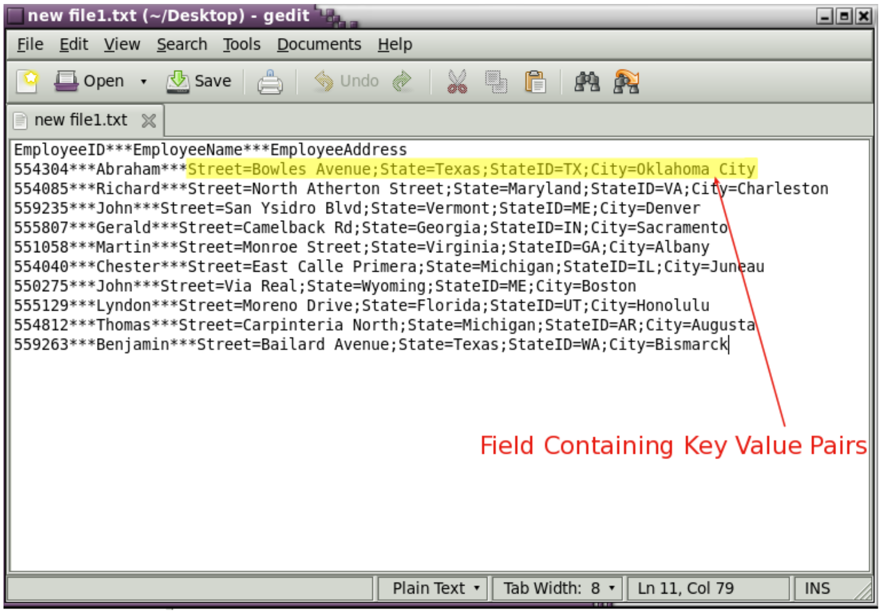
Version	2.0	
Component family	Processing/Fields	
Function	**tExtractKeyValue** generates multiple columns from a given column containing key-value pairs.	
Purpose	**tExtractKeyValue** allows you to extract “values” from “keys” in a string. The “keys” are taken as fields and “values” as its contents. (This component allows empty key-value pairs in the input).	
Basic settings	*Field*	Select an incoming field from the Field list to extract.
	Schema type and Edit Schema	A schema is a row description, i.e., it defines the number of fields that will be processed and passed on to the next component. The schema is either built-in or remote in the Repository. Click Edit Schema to make changes to the schema. Note that if you make changes, the schema automatically becomes built-in. Click Sync columns to retrieve the schema from the previous component connected to **tExtractKeyValue**.
		Built-in: The schema will be created and stored locally for this component only.
		Repository: The schema already exists and is stored in the Repository, hence can be reused in various projects and Job flowcharts.
	Field Separator(Regex)	It is used to separate key-value pairs in each row. It has to be a Regex Expression.
	Key-Value Separator(Regex)	It separates a Value from a key. It has to be a Regex Expression.
	Key Name(Regex)	**Column**: This field is automatically populated with the columns defined in the schema that you propagated.
		Key (Regex): This field should contain the “key” names as present in the input file. It has to be a Regex Expression.
		Trim Value: Select this check box to remove leading and trailing whitespaces from defined columns.
	Die on error	Select this check box to stop the execution of the Job when an error occurs. Clear the check box to skip the row on error and complete the process for error-free rows.
Advanced settings	*tStatCatcher Statistics*	Select this check box to gather the processing metadata at the Job level as well as at each component level.

Usage	This component handles flow of data therefore it requires input and output components. It allows you to extract data from a field containing key-value pairs, using a Row > Main link. Use this component to split a *Field* containing "key-value" pairs on the basis of "key". It extracts the "values" from the "keys". The "keys" are taken as fields and "values" as its contents. (This component even allows empty key-value pairs in the input). For further information, please see "Scenario 2" and "Scenario 3".
New Features	Allows the key to keep on changing for each row. Allows the key-value separator to be present in the value. Allows the key-value separator to be present in the key.
Author	Mahesh M. Pillai

Scenario 1: Extracting "values" from "keys" in a field.

The following scenario creates a three -component Job, which aims at reading a file wherein there is a field containing multiple key-value pairs .It extracts the "value" from the "keys " from the column and displays the output in the Run log console.

The contents of the Input Text File are shown below:

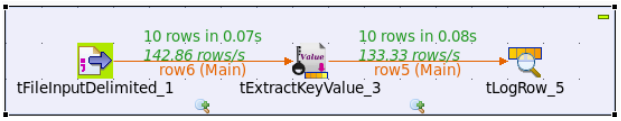


```
EmployeeID***EmployeeName***EmployeeAddress
554304***Abraham***Street=Bowles Avenue;State=Texas;StateID=TX;City=Oklahoma City
554085***Richard***Street=North Atherton Street;State=Maryland;StateID=VA;City=Charleston
559235***John***Street=San Ysidro Blvd;State=Vermont;StateID=ME;City=Denver
555807***Gerald***Street=Camelback Rd;State=Georgia;StateID=IN;City=Sacramento
551058***Martin***Street=Monroe Street;State=Virginia;StateID=GA;City=Albany
554040***Chester***Street=East Calle Primera;State=Michigan;StateID=IL;City=Juneau
550275***John***Street=Via Real;State=Wyoming;StateID=ME;City=Boston
555129***Lyndon***Street=Moreno Drive;State=Florida;StateID=UT;City=Honolulu
554812***Thomas***Street=Carpinteria North;State=Michigan;StateID=AR;City=Augusta
559263***Benjamin***Street=Bailard Avenue;State=Texas;StateID=WA;City=Bismarck
```

•The First Line is the Header.

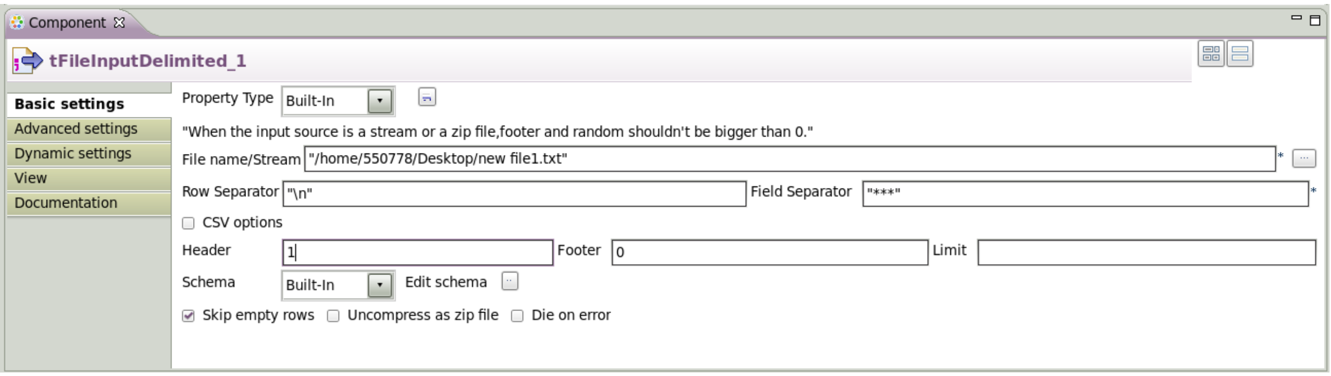
• Here the *Field* "EmployeeAddress" contains the Key-Value Pairs.

• Here the keys are "Street", "State", "StateID", and "City".



- Drop the following components from the Palette onto the design workspace: **tFileInputDelimited**, **tExtractKeyValue**, and **tLogRow**.
- Connect the three components using **Row Main** links.
- In the design workspace, select **tFileInputDelimited**.
- Click the **Component** tab to define the basic settings for **tFileInputDelimited**.
- In the **Basic settings** view, set **Property Type** to **Built-In**.
- Click the three-dot **[...]** button next to the **File Name** field to select the path to the input file.

The **File Name** field is mandatory.



Component ☒

tFileInputDelimited_1

Basic settings
Advanced settings
Dynamic settings
View
Documentation

Property Type: Built-In

"When the input source is a stream or a zip file,footer and random shouldn't be bigger than 0."

File name/Stream: "/home/550778/Desktop/new file1.txt" *

Row Separator: "\n"   Field Separator: "***" *

□ CSV options

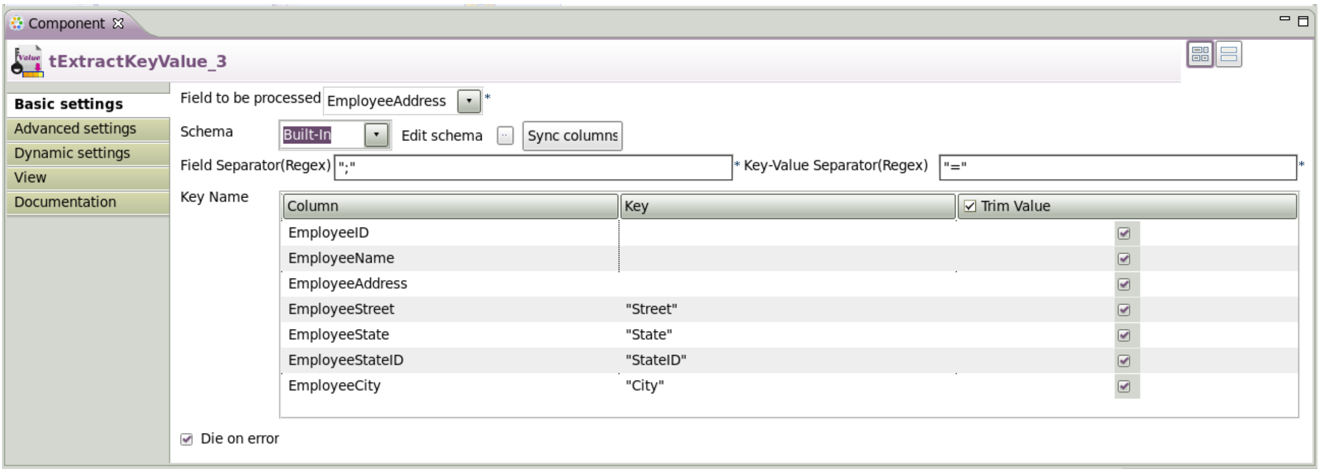
Header 1   Footer 0   Limit

Schema: Built-In   Edit schema

☑ Skip empty rows □ Uncompress as zip file □ Die on error

- The input file used in this scenario is called new *file1.txt* . It is a text file that holds three columns: *EmployeeID*, *EmployeeName*, and *EmployeeAddress*.
- Here the *Field Separator is* "***".
- Fill in all other fields as needed. In this scenario, the **header** is 1 and the **footer** is not set and there is no **limit** for the number of processed rows
- Click Edit schema to describe the data structure of this input file. In this scenario, the schema is made of the three columns, *EmployeeID*, *EmployeeName*, and *EmployeeAddress* .
- In the design workspace, select **tExtractKeyValue**.
- Click the Component tab to define the basic settings for **tExtractKeyValue**.

- From the **Field to be processed** list, select the column to split, *EmployeeAddress* in this scenario.



Component ☒

tExtractKeyValue_3

Basic settings
Advanced settings
Dynamic settings
View
Documentation

Field to be processed: EmployeeAddress *

Schema: Built-In   Edit schema   Sync columns

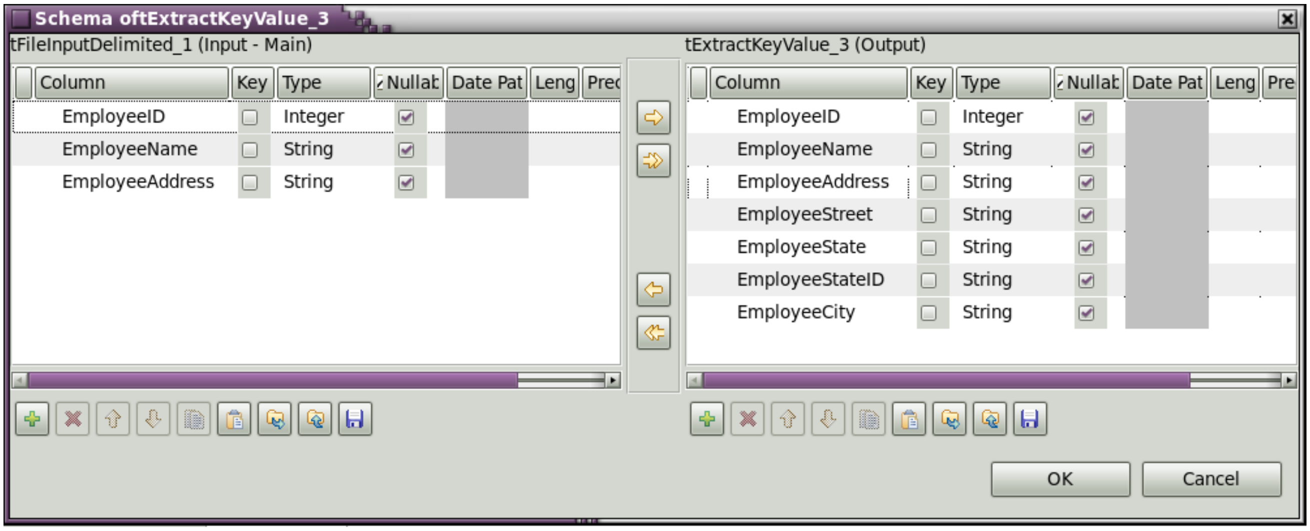
Field Separator(Regex): ";" *   Key-Value Separator(Regex): "=" *

Key Name

Column	Key	☑ Trim Value
EmployeeID		☑
EmployeeName		☑
EmployeeAddress		☑
EmployeeStreet	"Street"	☑
EmployeeState	"State"	☑
EmployeeStateID	"StateID"	☑
EmployeeCity	"City"	☑

☑ Die on error

- Set the **Schema** as either a local (Built-in) or a remotely managed (Repository) to define the data to pass on to the **tLogRow** component.
- You can load and/or edit the schema via the **Edit Schema** function.
- Click Edit schema to describe the data structure of this processing component.
  In the present scenario we have in total a maximum of 4 key-value pairs in a field.
  You may define them as shown below.



Schema oftExtractKeyValue_3

tFileInputDelimited_1 (Input - Main)

Column	Key	Type	Nullab	Date Pat	Leng	Prec
EmployeeID	□	Integer	☑			
EmployeeName	□	String	☑			
EmployeeAddress	□	String	☑			

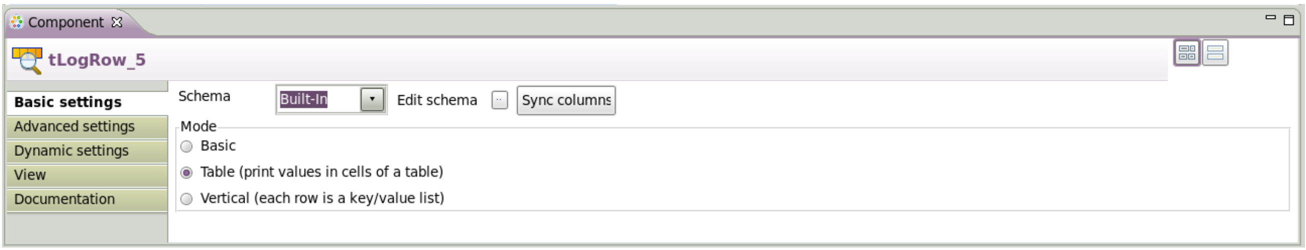
tExtractKeyValue_3 (Output)

Column	Key	Type	Nullab	Date Pat	Leng	Pre
EmployeeID	□	Integer	☑			
EmployeeName	□	String	☑			
EmployeeAddress	□	String	☑			
EmployeeStreet	□	String	☑			
EmployeeState	□	String	☑			
EmployeeStateID	□	String	☑			
EmployeeCity	□	String	☑			

- The column name can be any valid name.
- Define the **Field separator** used to delimit key-value pairs in a row (Here it is “;”). Also define the **Key-Value Separator** (Here it is “=”).
  These fields should be in Regex Patterns.
- Now the **Key Name** Field should be filled.
  Column field is automatically populated with the columns defined in the schema that you propagated.
  The **Key** field corresponds to the keys in the *Field* that we have selected.
  In the current scenario the keys are “Street”, “State”, “StateID”, and “City”.
- The **Trim Value** field can be used to remove leading and trailing whitespaces from the “values” corresponding to the fields.

•Die on Error can be set as per need. (Select this check box to stop the execution of the Job when an error occurs. Clear the check box to skip the row on error and complete the process for error-free rows.)

- Select the tLogRow and define the Field separator to use for the output display.



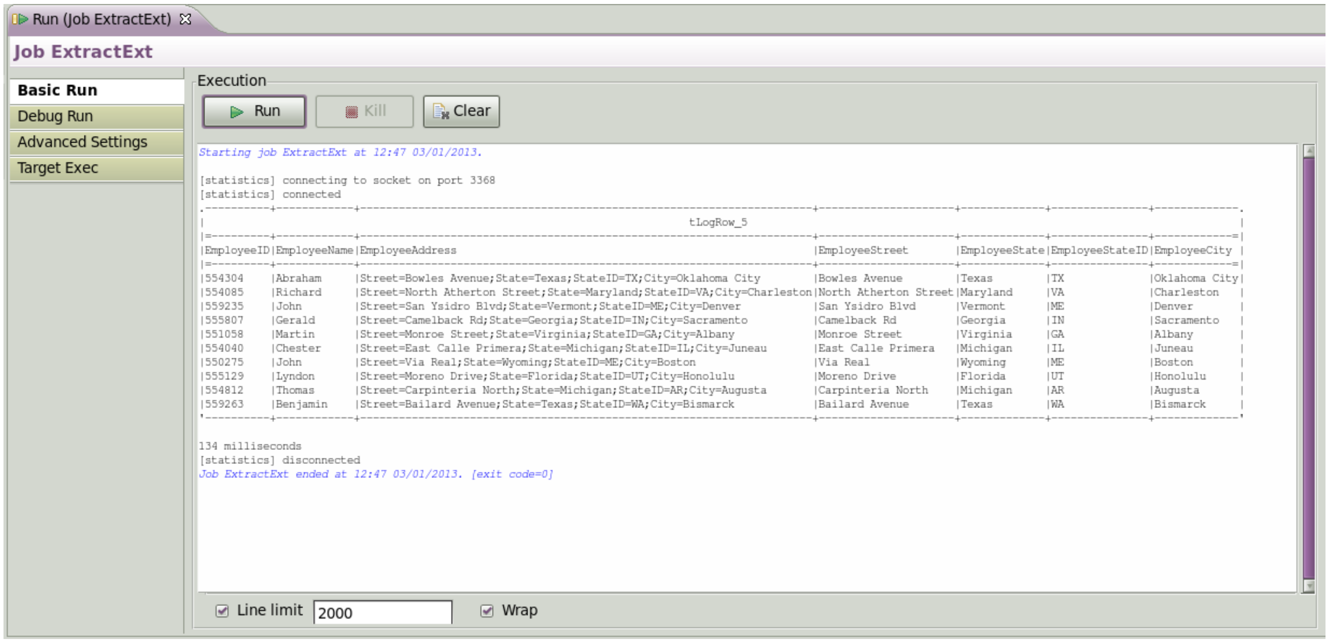
Component

tLogRow_5

Schema: Built-In Edit schema Sync columns

Mode
○ Basic
● Table (print values in cells of a table)
○ Vertical (each row is a key/value list)

- Go to **Run** tab, and click on Run to execute the Job.
  The extracted fields are displayed on the Run log as defined in both components.
  The data from the input schema will be automatically propagated into the Output by default.



```
Starting job ExtractExt at 12:47 03/01/2013.

[statistics] connecting to socket on port 3368
[statistics] connected
.----------+------------+----------------------------------------------------------------------+---------------------+-------------+---------------+-------------.
|                                                                           tLogRow_5                                                                            |
|=---------+------------+----------------------------------------------------------------------+---------------------+-------------+---------------+------------=|
|EmployeeID|EmployeeName|EmployeeAddress                                                       |EmployeeStreet       |EmployeeState|EmployeeStateID|EmployeeCity |
|=---------+------------+----------------------------------------------------------------------+---------------------+-------------+---------------+------------=|
|554304    |Abraham     |Street=Bowles Avenue;State=Texas;StateID=TX;City=Oklahoma City        |Bowles Avenue        |Texas        |TX             |Oklahoma City|
|554085    |Richard     |Street=North Atherton Street;State=Maryland;StateID=VA;City=Charleston|North Atherton Street|Maryland     |VA             |Charleston   |
|559235    |John        |Street=San Ysidro Blvd;State=Vermont;StateID=ME;City=Denver           |San Ysidro Blvd      |Vermont      |ME             |Denver       |
|555807    |Gerald      |Street=Camelback Rd;State=Georgia;StateID=IN;City=Sacramento          |Camelback Rd         |Georgia      |IN             |Sacramento   |
|551058    |Martin      |Street=Monroe Street;State=Virginia;StateID=GA;City=Albany            |Monroe Street        |Virginia     |GA             |Albany       |
|554040    |Chester     |Street=East Calle Primera;State=Michigan;StateID=IL;City=Juneau       |East Calle Primera   |Michigan     |IL             |Juneau       |
|550275    |John        |Street=Via Real;State=Wyoming;StateID=ME;City=Boston                  |Via Real             |Wyoming      |ME             |Boston       |
|555129    |Lyndon      |Street=Moreno Drive;State=Florida;StateID=UT;City=Honolulu            |Moreno Drive         |Florida      |UT             |Honolulu     |
|554812    |Thomas      |Street=Carpinteria North;State=Michigan;StateID=AR;City=Augusta       |Carpinteria North    |Michigan     |AR             |Augusta      |
|559263    |Benjamin    |Street=Bailard Avenue;State=Texas;StateID=WA;City=Bismarck            |Bailard Avenue       |Texas        |WA             |Bismarck     |
'----------+------------+----------------------------------------------------------------------+---------------------+-------------+---------------+-------------'

134 milliseconds
[statistics] disconnected
Job ExtractExt ended at 12:47 03/01/2013. [exit code=0]
```

☑ Line limit 2000 ☑ Wrap

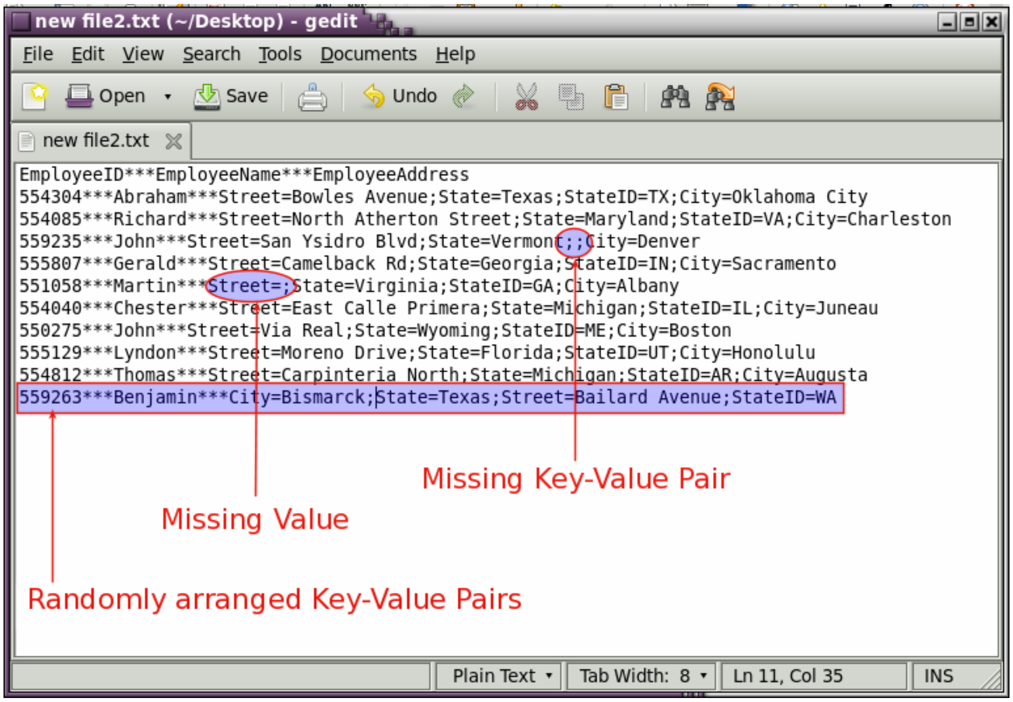
Scenario 2: Extracting “values” from “keys” in a field with
➢ Key-Value Pairs randomly distributed within the field
➢ Key-Value Pairs missing in the field

The following scenario describes the Flexibility *this component offers:*

- ➢ *It allows missing key value pairs.*
- ➢ *It allows missing values.*
- ➢ *It allows random distribution of key-value pairs within the field.*

The following scenario depicts a two-component Job, which aims at extracting the “value” from the “keys” from a field (containing multiple key-value pair) and displays the output in the Run log console.
The above mentioned (*in italics)* conditions are created in the input file.

The Input File (with Annotations) is shown below:



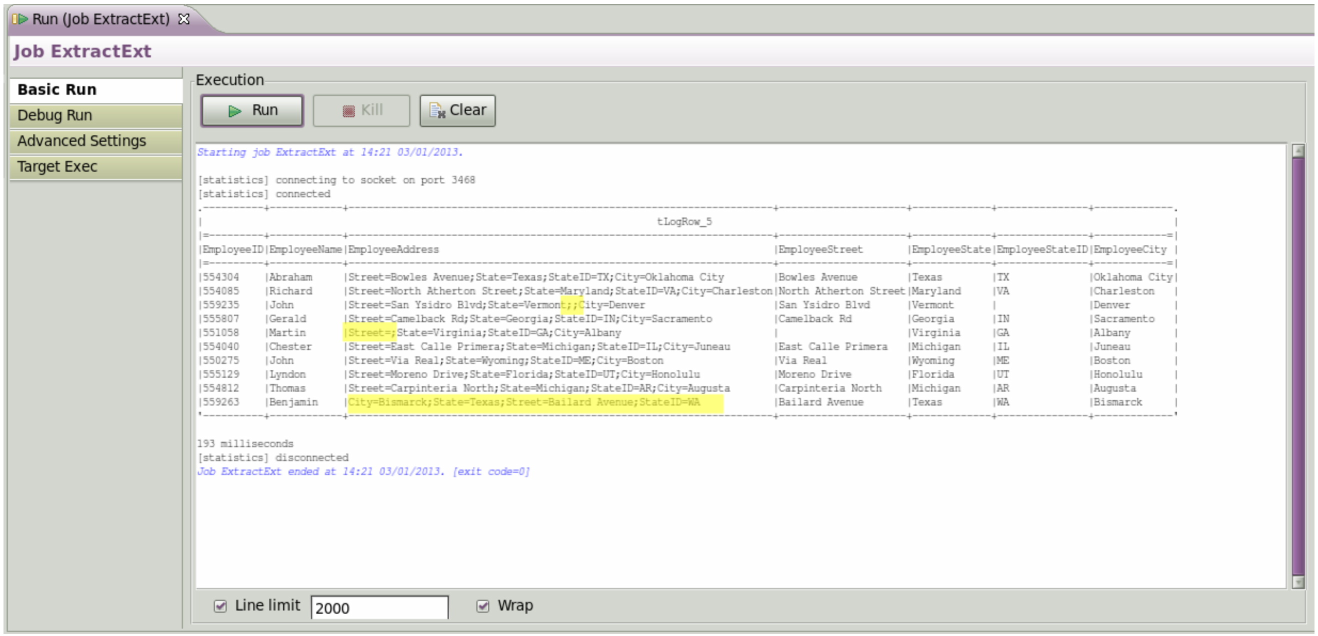
```
EmployeeID***EmployeeName***EmployeeAddress
554304***Abraham***Street=Bowles Avenue;State=Texas;StateID=TX;City=Oklahoma City
554085***Richard***Street=North Atherton Street;State=Maryland;StateID=VA;City=Charleston
559235***John***Street=San Ysidro Blvd;State=Vermont;;City=Denver
555807***Gerald***Street=Camelback Rd;State=Georgia;StateID=IN;City=Sacramento
551058***Martin***Street=;State=Virginia;StateID=GA;City=Albany
554040***Chester***Street=East Calle Primera;State=Michigan;StateID=IL;City=Juneau
550275***John***Street=Via Real;State=Wyoming;StateID=ME;City=Boston
555129***Lyndon***Street=Moreno Drive;State=Florida;StateID=UT;City=Honolulu
554812***Thomas***Street=Carpinteria North;State=Michigan;StateID=AR;City=Augusta
559263***Benjamin***City=Bismarck;State=Texas;Street=Bailard Avenue;StateID=WA
```

Missing Key-Value Pair

Missing Value

Randomly arranged Key-Value Pairs

The Job is the same as in Scenario1.
The output with annotations is shown below:



```
Starting job ExtractExt at 14:21 03/01/2013.

[statistics] connecting to socket on port 3468
[statistics] connected
.----------+------------+----------------------------------------------------------------------+---------------------+-------------+---------------+-------------.
|                                                                           tLogRow_5                                                                            |
|=---------+------------+----------------------------------------------------------------------+---------------------+-------------+---------------+------------=|
|EmployeeID|EmployeeName|EmployeeAddress                                                       |EmployeeStreet       |EmployeeState|EmployeeStateID|EmployeeCity |
|=---------+------------+----------------------------------------------------------------------+---------------------+-------------+---------------+------------=|
|554304    |Abraham     |Street=Bowles Avenue;State=Texas;StateID=TX;City=Oklahoma City        |Bowles Avenue        |Texas        |TX             |Oklahoma City|
|554085    |Richard     |Street=North Atherton Street;State=Maryland;StateID=VA;City=Charleston|North Atherton Street|Maryland     |VA             |Charleston   |
|559235    |John        |Street=San Ysidro Blvd;State=Vermont;;City=Denver                     |San Ysidro Blvd      |Vermont      |               |Denver       |
|555807    |Gerald      |Street=Camelback Rd;State=Georgia;StateID=IN;City=Sacramento          |Camelback Rd         |Georgia      |IN             |Sacramento   |
|551058    |Martin      |Street=;State=Virginia;StateID=GA;City=Albany                         |                     |Virginia     |GA             |Albany       |
|554040    |Chester     |Street=East Calle Primera;State=Michigan;StateID=IL;City=Juneau       |East Calle Primera   |Michigan     |IL             |Juneau       |
|550275    |John        |Street=Via Real;State=Wyoming;StateID=ME;City=Boston                  |Via Real             |Wyoming      |ME             |Boston       |
|555129    |Lyndon      |Street=Moreno Drive;State=Florida;StateID=UT;City=Honolulu            |Moreno Drive         |Florida      |UT             |Honolulu     |
|554812    |Thomas      |Street=Carpinteria North;State=Michigan;StateID=AR;City=Augusta       |Carpinteria North    |Michigan     |AR             |Augusta      |
|559263    |Benjamin    |City=Bismarck;State=Texas;Street=Bailard Avenue;StateID=WA            |Bailard Avenue       |Texas        |WA             |Bismarck     |
'----------+------------+----------------------------------------------------------------------+---------------------+-------------+---------------+-------------'

193 milliseconds
[statistics] disconnected
Job ExtractExt ended at 14:21 03/01/2013. [exit code=0]
```

☑ Line limit 2000 ☑ Wrap

In spite of the input being in an unordered format, the output is in the correct format.

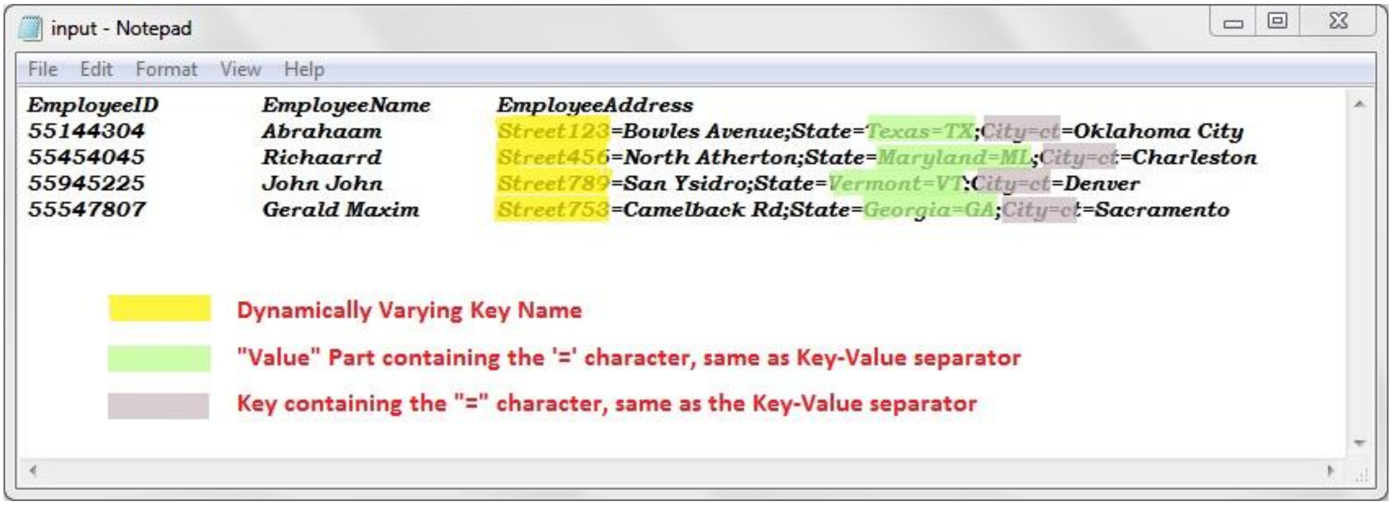
Scenario 3: Reading data from a file that contains

- ➢ **Dynamic Key Name**
- ➢ **Key-Value Separator duplicated in the "value"**
- ➢ **Key-Value Separator duplicated in the "key"**

*The following scenario describes the **latest Flexibility** this component offers:*

- ➢ *It allows the key to keep on changing for each row.*
- ➢ *It allows the key-value separator to be present in the value.*
- ➢ *It allows the key-value separator to be present in the key.*

The following scenario depicts a two-component Job, which aims at reading each row (containing multiple key-value pair) of a file, extracts the "value" from the "keys" and displays the output in the Run log console.
The above mentioned (*in italics)* conditions are created in the input file.
The Input File (with Annotations) is shown below:



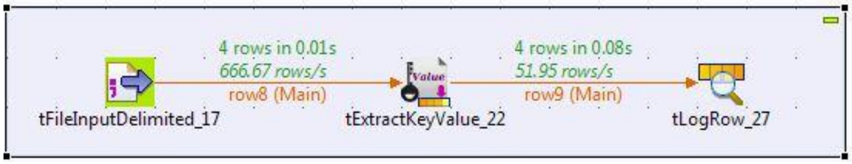
```
EmployeeID      EmployeeName     EmployeeAddress
55144304        Abrahaam         Street123=Bowles Avenue;State=Texas=TX;City=ct=Oklahoma City
55454045        Richaarrd        Street456=North Atherton;State=Maryland=ML;City=ct=Charleston
55945225        John John        Street789=San Ysidro;State=Vermont=VT;City=ct=Denver
55547807        Gerald Maxim     Street753=Camelback Rd;State=Georgia=GA;City=ct=Sacramento
```

🟨 Dynamically Varying Key Name

🟩 "Value" Part containing the '=' character, same as Key-Value separator

▨ Key containing the "=" character, same as the Key-Value separator

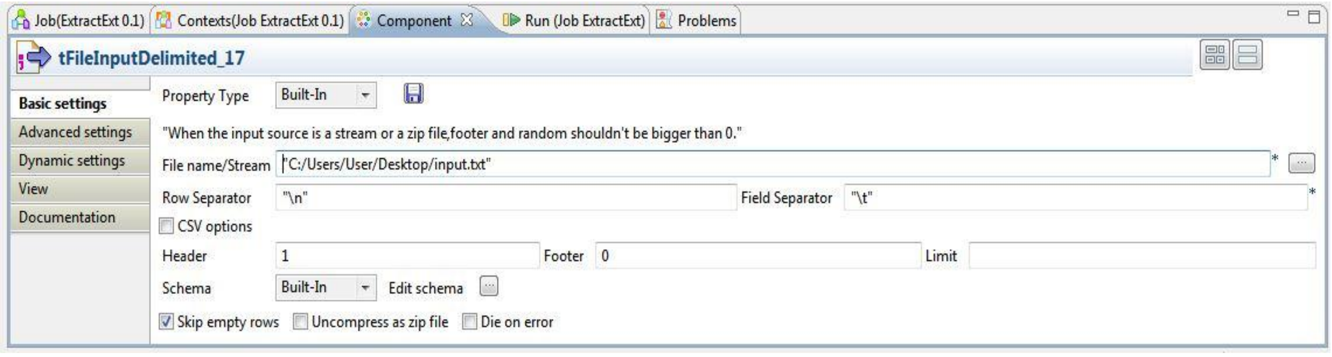
There are three tab separated fields in the above file.
The third field "EmployeeAddress" contains three Key-Value pairs.
The key-value separator is '='.
The first key-value pair contains key that keeps on changing for each row. viz., Street123, Street456, etc.
The second key-value pair has values which already contains '=' character, which is same as the key-value separator. viz., Texas=TX, Maryland=ML, etc.
The third pair has '=' in its key (City=ct) which is same as the key-value separator.
In these scenarios, the **tExtractKeyValue** component can be leveraged to filter off the value from the key.



• Drop the following components from the Palette onto the design workspace: **tFileInputDelimited**, **tExtractKeyValue**, and **tLogRow**.

- Connect the three components using **Row Main** links.
- In the design workspace, select **tFileInputDelimited**.
- Click the **Component** tab to define the basic settings for **tFileInputDelimited**.
- In the **Basic settings** view, set **Property Type** to **Built-In**.
- Click the three-dot **[...]** button next to the **File Name** field to select the path to the input file.

The **File Name** field is mandatory.



Property Type: Built-In

"When the input source is a stream or a zip file,footer and random shouldn't be bigger than 0."

File name/Stream: "C:/Users/User/Desktop/input.txt" *

Row Separator: "\n" Field Separator: "\t" *

□ CSV options

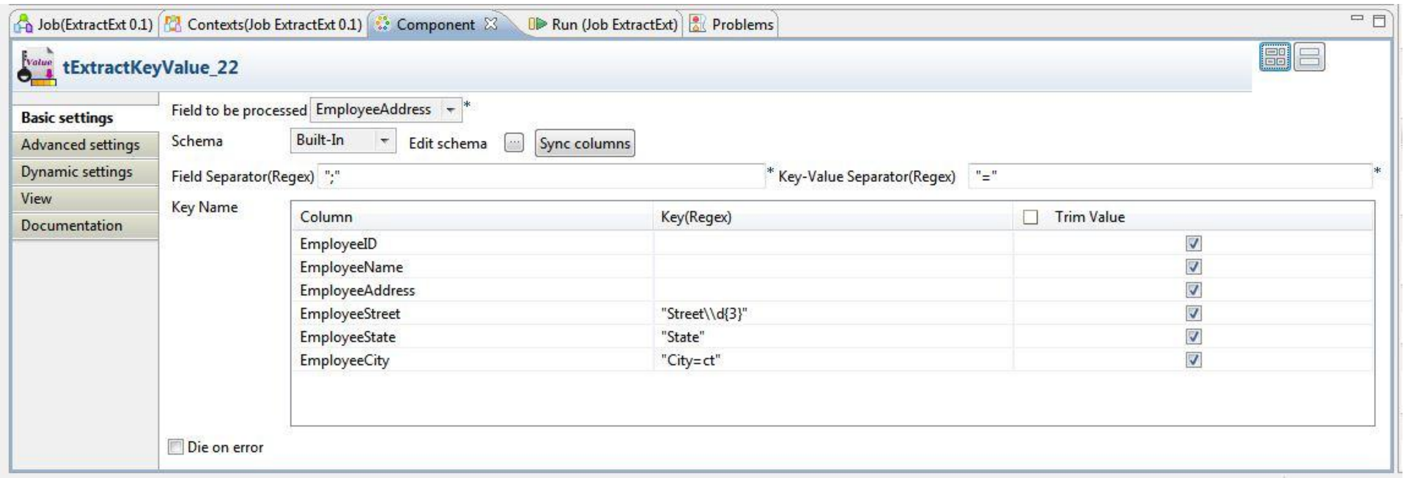
Header: 1 Footer: 0 Limit:

Schema: Built-In Edit schema

☑ Skip empty rows □ Uncompress as zip file □ Die on error

- The input file used in this scenario is called *input.txt.* It is a text file that holds three columns: *EmployeeID, EmployeeName*, and *EmployeeAddress*.
- Here the *Field Separator is* "\t".
- Fill in all other fields as needed. In this scenario, the **header** is 1 and the **footer** is not set and there is no **limit** for the number of processed rows
- Click Edit schema to describe the data structure of this input file. In this scenario, the schema is made of the three columns, *EmployeeID*, *EmployeeName*, and *EmployeeAddress* .
- In the design workspace, select **tExtractKeyValue**.
- Click the Component tab to define the basic settings for **tExtractKeyValue**.

- From the **Field to be processed** list, select the column to split, *EmployeeAddress* in this scenario.



Field to be processed: EmployeeAddress *

Schema: Built-In Edit schema

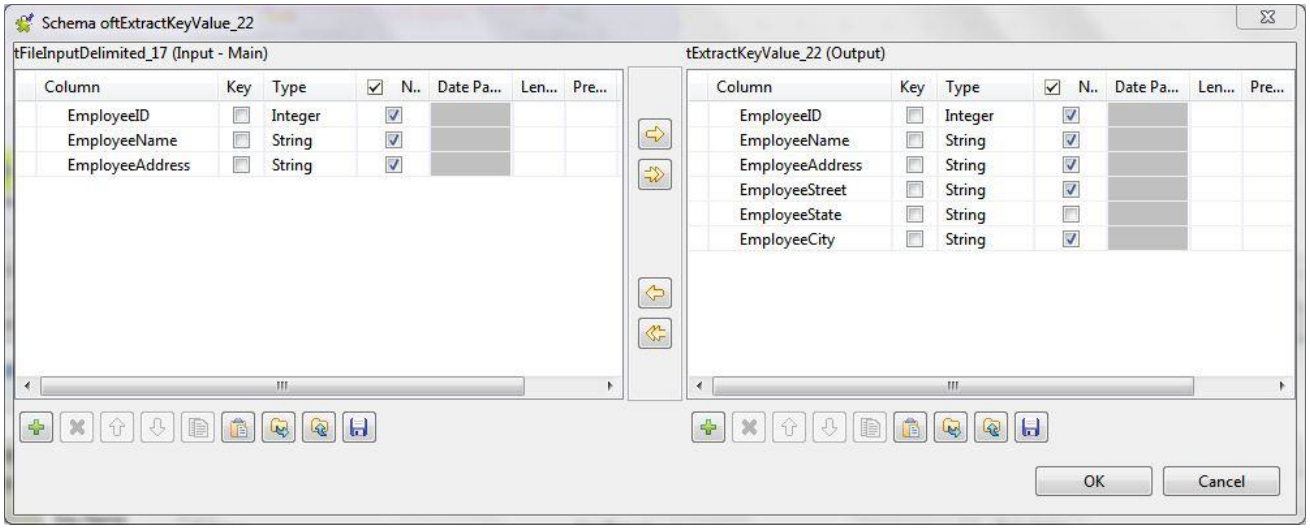
Field Separator(Regex): ";" * Key-Value Separator(Regex): "=" *

Key Name

Column	Key(Regex)	□ Trim Value
EmployeeID		☑
EmployeeName		☑
EmployeeAddress		☑
EmployeeStreet	"Street\\d{3}"	☑
EmployeeState	"State"	☑
EmployeeCity	"City=ct"	☑

□ Die on error

- Set the **Schema** as either a local (Built-in) or a remotely managed (Repository) to define the data to pass on to the **tLogRow** component.
- You can load and/or edit the schema via the **Edit Schema** function.
- Click Edit schema to describe the data structure of this processing component.
  In the present scenario we have in total 3 key-value pairs in a field.
  You may define them as shown below.



Schema oftExtractKeyValue_22

tFileInputDelimited_17 (Input - Main)

Column	Key	Type	☑ N..	Date Pa...	Len...	Pre...
EmployeeID	□	Integer	☑			
EmployeeName	□	String	☑			
EmployeeAddress	□	String	☑			

tExtractKeyValue_22 (Output)

Column	Key	Type	☑ N..	Date Pa...	Len...	Pre...
EmployeeID	□	Integer	☑			
EmployeeName	□	String	☑			
EmployeeAddress	□	String	☑			
EmployeeStreet	□	String	☑			
EmployeeState	□	String	□			
EmployeeCity	□	String	☑			

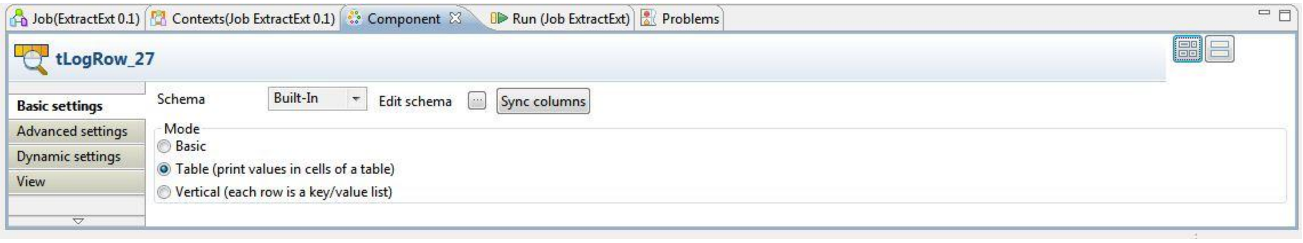
- The column name can be any valid name.
- Define the **Field separator** used to delimit key-value pairs in a row (Here it is “;”). Also define the **Key-Value Separator** (Here it is “=”).
  These fields should be in Regex Patterns.
- Now the **Key Name** Field should be filled.
  Column field is automatically populated with the columns defined in the schema that you propagated.
  The **Key** field corresponds to the keys in the *Field* that we have selected.
  The **Key** should be in Regex Patterns. This allows the user to have control over defining patterns.

  In the current scenario the key for the first field is “Street” followed by a three-digit number. So, the corresponding Regex Pattern would be "Street\\d{3}".
  The second key is "State" and the third is "City=ct".

- The **Trim Value** field can be used to remove leading and trailing whitespaces from the “values” corresponding to the fields.

•Die on Error can be set as per need. (Select this check box to stop the execution of the Job when an error occurs. Clear the check box to skip the row on error and complete the process for error-free rows.)

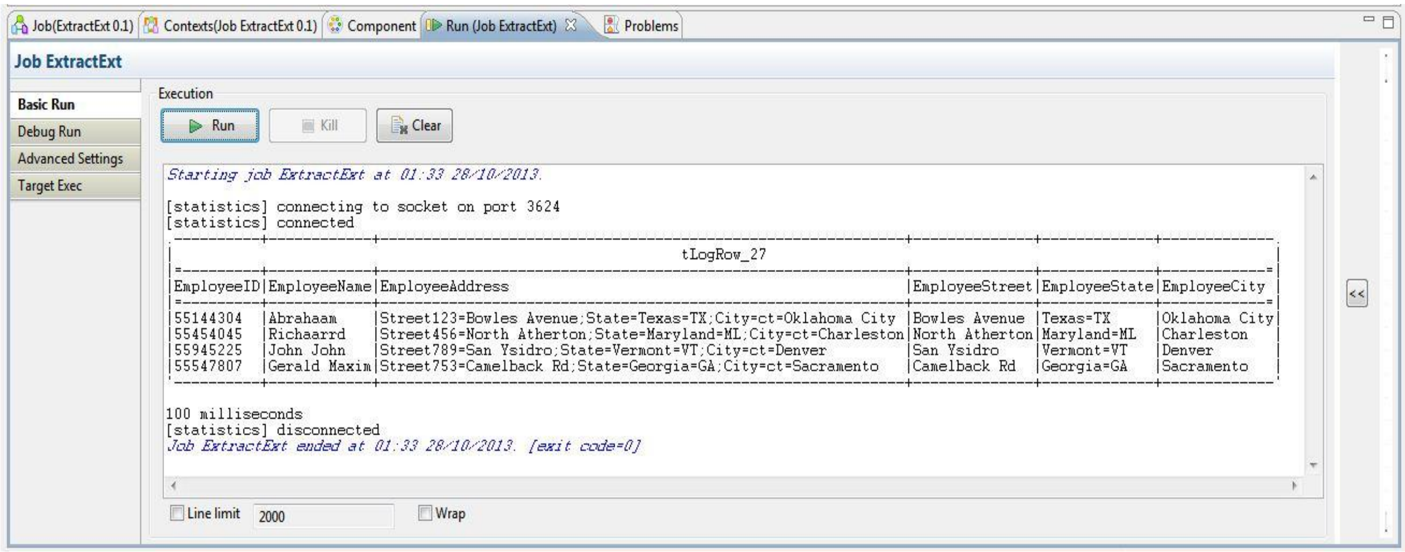
- Select the tLogRow and define the Field separator to use for the output display.



Schema Built-In

Mode
○ Basic
● Table (print values in cells of a table)
○ Vertical (each row is a key/value list)

- Go to **Run** tab, and click on Run to execute the Job.
  The extracted fields are displayed on the Run log as defined in both components.
  The data from the input schema will be automatically propagated into the Output by default.



```
Starting job ExtractExt at 01:33 28/10/2013.

[statistics] connecting to socket on port 3624
[statistics] connected
.----------+------------+-------------------------------------------------------------+--------------+-------------+-------------.
|                                                           tLogRow_27                                                           |
|=---------+------------+-------------------------------------------------------------+--------------+-------------+------------=|
|EmployeeID|EmployeeName|EmployeeAddress                                              |EmployeeStreet|EmployeeState|EmployeeCity |
|=---------+------------+-------------------------------------------------------------+--------------+-------------+------------=|
|55144304  |Abrahaam    |Street123=Bowles Avenue;State=Texas=TX;City=ct=Oklahoma City |Bowles Avenue |Texas=TX     |Oklahoma City|
|55454045  |Richaarrd   |Street456=North Atherton;State=Maryland=ML;City=ct=Charleston|North Atherton|Maryland=ML  |Charleston   |
|55945225  |John John   |Street789=San Ysidro;State=Vermont=VT;City=ct=Denver         |San Ysidro    |Vermont=VT   |Denver       |
|55547807  |Gerald Maxim|Street753=Camelback Rd;State=Georgia=GA;City=ct=Sacramento   |Camelback Rd  |Georgia=GA   |Sacramento   |
'----------+------------+-------------------------------------------------------------+--------------+-------------+-------------'

100 milliseconds
[statistics] disconnected
Job ExtractExt ended at 01:33 28/10/2013. [exit code=0]
```

□ Line limit 2000 □ Wrap

In spite of the key name being dynamic and the key/value containing the key-value separator (=), the output is obtained in the required format.