# Talend User Component tGoogleAnalyticsInput

let your work flow

http://www.cimt-ag.de

## Purpose

This component addresses the needs of gathering Google Analytics data for an large number of profiles and fine grained detail data.
The component uses the Core Reporting API 3.0 and the Authentication API OAuth 2.0 final.
To provide the ability to run in multiple iterations the component has special capabilities to avoid multiple logins through iterations. Usually automated processes should not use personal accounts. This requirement is addressed by using Service Accounts which are the only preferred way to login into Google Analytics for automated processes.
Please in case of problems check the check list at the end of this document.

## Talend-Integration

This component can be found in the palette under Business->Google
This component provides an output flow and several return values.

## Parameters

Parameters to connect to Google Analytics (setup client)

| Property | Content |
|---|---|
| Application Name | Not necessary, but recommended by Google.<br>Simple provide the name of your application gathering data. ***Required*** |
| Service Account Email | The email address of the service account. This address is created by Google within the process of creating a service account. ***Required*** |
| Key File (*.p12) | The Service Account Login works with private key file for authentication. In the process of creating a service account you download this file.<br>***Required*** |

Parameters to define the query

| Property | Content |
|---|---|
| Profile-ID | All profiles gets an unique integer identifier. This id are needed here. As profile names can be reused within other Google-Accounts and Google-Properties, a profile ID is unique over all accounts.<br>***Required!*** |
| Start Date | All queries need always a time range (only date, not time).<br>The value must be a String with the pattern “yyyy-MM-dd”. ***Required!*** |
| End Date | Time range end. If you want gather data for one date, use start date as end date.<br>The value must be a String with the pattern “yyyy-MM-dd”. ***Required!*** |
| Dimensions | Dimensions are like group clauses. The metrics values will be grouped by these dimensions.<br>See advise for notations below.<br>***Not required (since release 1.5)*** |
| Metrics | Things you want to measure. Separate all metrics with a comma. |

| | |
|---|---|
| | See advise for notations below.<br>***Required!*** |
| Filters | Contains all used filters as concatenated string.<br>See advise for notation below |
| Sorts | Contains all sort criteria as concatenated string.<br>See advise for notation below |
| Segment-ID | Segments are stored filters within Google Analytics. You need the numeric ID of the segment. |
| Sampling Level | Google Analytics can collect the result based on sampled data. This attribute tells Google Analytics which kind of sampling should be used (in case of sampling is necessary because of the amount of data).<br>These are the possible values:<br>DEFAULT: It is a balance between Speed and precision<br>FASTER: use more sampled data but the result returns faster<br>HIGHER_PRECISION: use less sampled data and it takes longer to get the result |
| Deliver Totals Data Set | The API provides a totals data set. This can be used to calculate percentage values or check results. This data set will be delivered (as first row) if option is checked or will be omitted if option is not checked. Date values (e.g. ga:date) remains empty (null) in the totals dataset. |

### Advice for notation

For dimensions, metrics, filters and sorts you have to use the notation from the Google Core API:
https://developers.google.com/analytics/devguides/reporting/core/dimsmets

Filters can be concatenated with OR or AND operator.
Separate filter expressions with a comma means OR
Separate filter expressions with a semicolon means AND

Filter comparison operators:

| Operator | Meaning |
|---|---|
| “==” | Exact match to include |
| “!=” | Exact match to exclude |
| “=~” | Regex match to include (only for strings) |
| “!~” | Regex match to exclude (only for strings) |
| “>=” | Greater or equals than (only for numbers) |
| “=@” | Contains string |
| “!@” | Does not contains string |
| “>” | Greater than (only for numbers) |
| “<=” | Lower or equals than (only for numbers) |
| “<” | Lower than (only for numbers) |

### Building Schema for Component

In the schema you need a amount of columns equals to the sum of the number of dimensions and

metrics.
Columns in the schema must start at first with dimensions (if provided) and ends with metrics.
Schema column types must match to the data types of the dimensions and metrics. The schema column names can differ from the names of dimensions and metrics. Only the order and there type are important.
Important: For date dimensions (e.g. ga:date) you must specify the date pattern as “yyyyMMdd”!

### Advanced Option Parameters

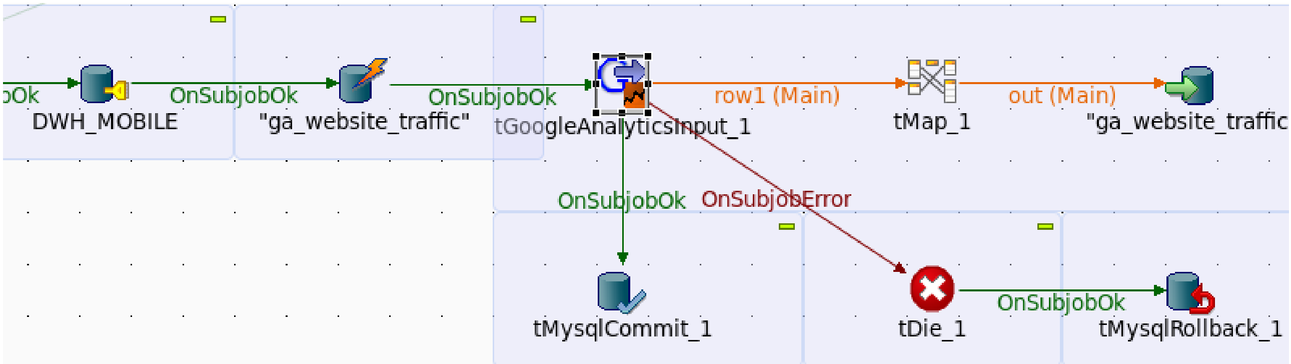
| Property | Content |
|---|---|
| Timeout in s | How long should the component wait for getting the first result and fetching all result with one internal iteration |
| Static Time Offset (to past) | Within the process of login, the component requests an access token and use the local time stamp (because these tokens will expire after a couple of seconds)<br>Google rejects all request to access tokens when the request is in the future compared to the timestamp in Google servers. If you expierence such kind of problems, this options let the requests appear to be more in the past (5-10s was recognized as good time shift) |
| Fetch Size | This is the amount of data the component fetches at once. The value is used to set the max_rows attribute. max_rows means not the absolute amount of data! The component manages setting the start index to get all data. To achieve this, the component iterates as long as the last result set are completely fetched. |
| Local Number Format | You can get numbers in various formats. Here you can define the locale in which format double or float values are should textual formatted by the API. |
| Reuse Client for Iterations | If you use this component in iterations it is strongly recommended to set this option. It saves time to authenticate unnecessary often and avoids problems because of max. amount of connects per time range. |
| Distinct Name Extension | The client will be kept with a automatically created name: Talend-Name-Component name + job name. In case this is not distinct enough, you can specify an additional extension to the name. |

### Return values

| Return value | Content |
|---|---|
| ERROR_MESSAGE | Last error message |
| NB_LINE | Number of delivered lines |
| CONTAINS_SAMPLED_DATA | True if data are sampled, means not exactly calculated. This can happen if you query to many details. |
| SAMPLE_SIZE | The amount of datasets used for the query |
| SAMPLE_SPACE | The amount of available datasets for this query |
| TOTAL_AFFECTED_ROWS | Number of rows which are collected by Google to calculate the result. |

### Scenario 1

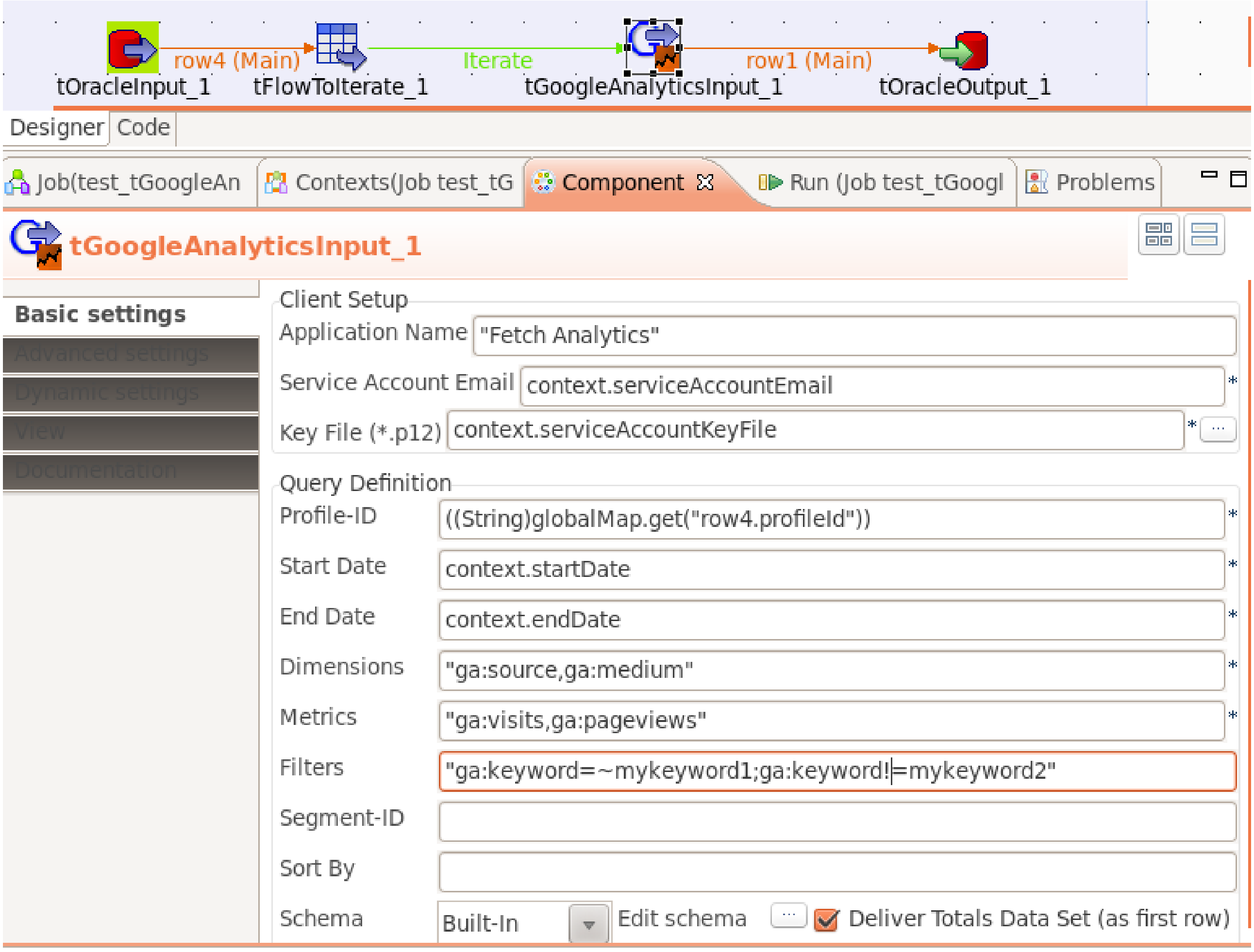
Profiles and filters are stored in a database.



In this scenario Google Analytics data are fetched in a sub job. The data will be deleted before inserting to provide restart capabilities.
This job can be used as embedded job in a surrounding job in a iteration.

### Scenario 2

Iterate through profiles stored in a database table.



tGoogleAnalyticsInput_1

Client Setup

| | |
|---|---|
| Application Name | "Fetch Analytics" |
| Service Account Email | context.serviceAccountEmail * |
| Key File (*.p12) | context.serviceAccountKeyFile * |

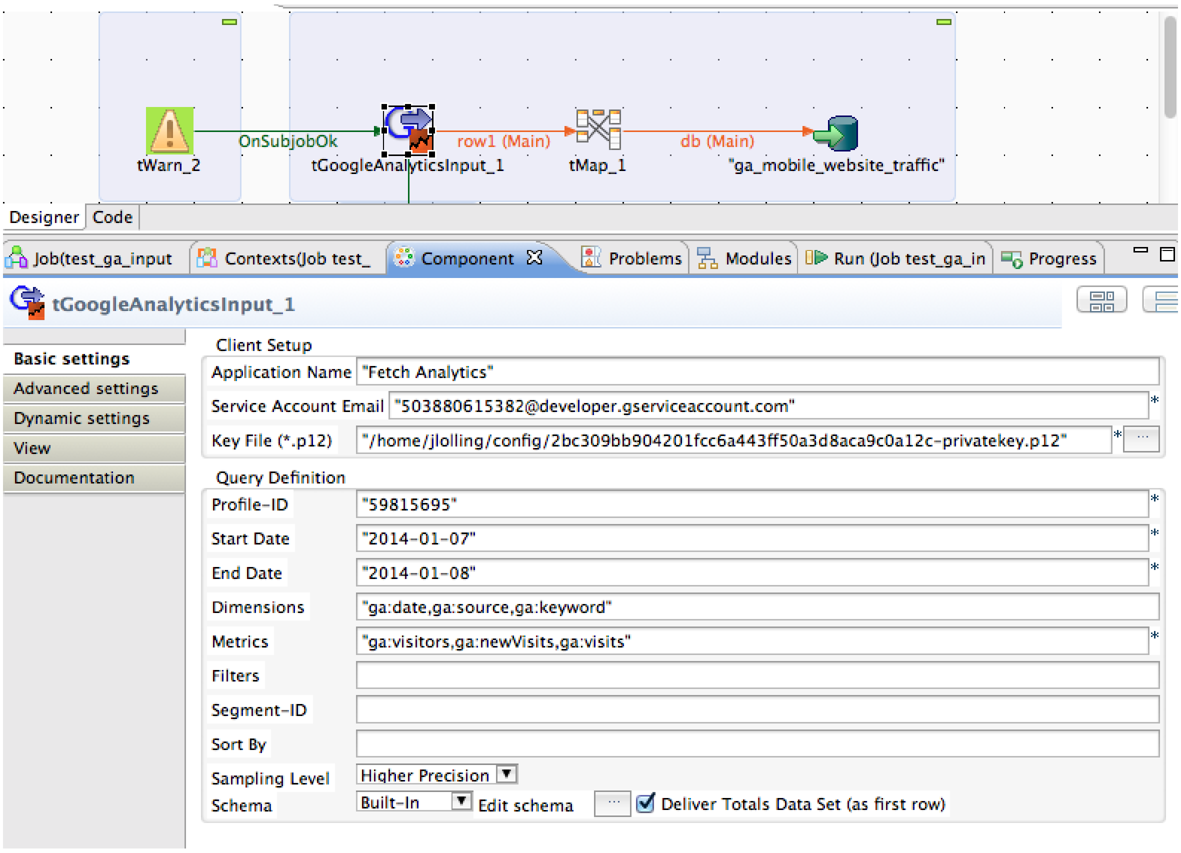
Query Definition

| | |
|---|---|
| Profile-ID | ((String)globalMap.get("row4.profileId")) * |
| Start Date | context.startDate * |
| End Date | context.endDate * |
| Dimensions | "ga:source,ga:medium" * |
| Metrics | "ga:visits,ga:pageviews" * |
| Filters | "ga:keyword=~mykeyword1;ga:keyword!=mykeyword2" |
| Segment-ID | |
| Sort By | |
| Schema | Built-In |

Edit schema ☑ Deliver Totals Data Set (as first row)

## Scenario 3

In case of the data are based on sampled data, you can control the sampling.



tGoogleAnalyticsInput_1

Client Setup

Application Name "Fetch Analytics"

Service Account Email "503880615382@developer.gserviceaccount.com" *

Key File (*.p12) "/home/jlolling/config/2bc309bb904201fcc6a443ff50a3d8aca9c0a12c-privatekey.p12" *

Query Definition

Profile-ID "59815695" *

Start Date "2014-01-07" *

End Date "2014-01-08" *

Dimensions "ga:date,ga:source,ga:keyword"

Metrics "ga:visitors,ga:newVisits,ga:visits" *

Filters

Segment-ID

Sort By

Sampling Level Higher Precision

Schema Built-In Edit schema ☑ Deliver Totals Data Set (as first row)

With the help of the new attribute Sampling Level you can control the way Google Analytics collects the necessary data. E.g. with the sampling level “Higher Precision” you can use a larger sampling size. The back draft, it will take much longer to get the result.
If there is no sampling, this will have no effect.

**<u>Checklist</u>**:

1. Is the email of the service account added to all relevant profiles?
2. Is the system time of the host running the job synchronized with a NTP server?
3. Is the Google Analytics API enable in the Google API Console?

**Tip**:
Check your query at first in the Google Analytics API Explorer to get an idea if the data works for you.