ETL development for Master Project

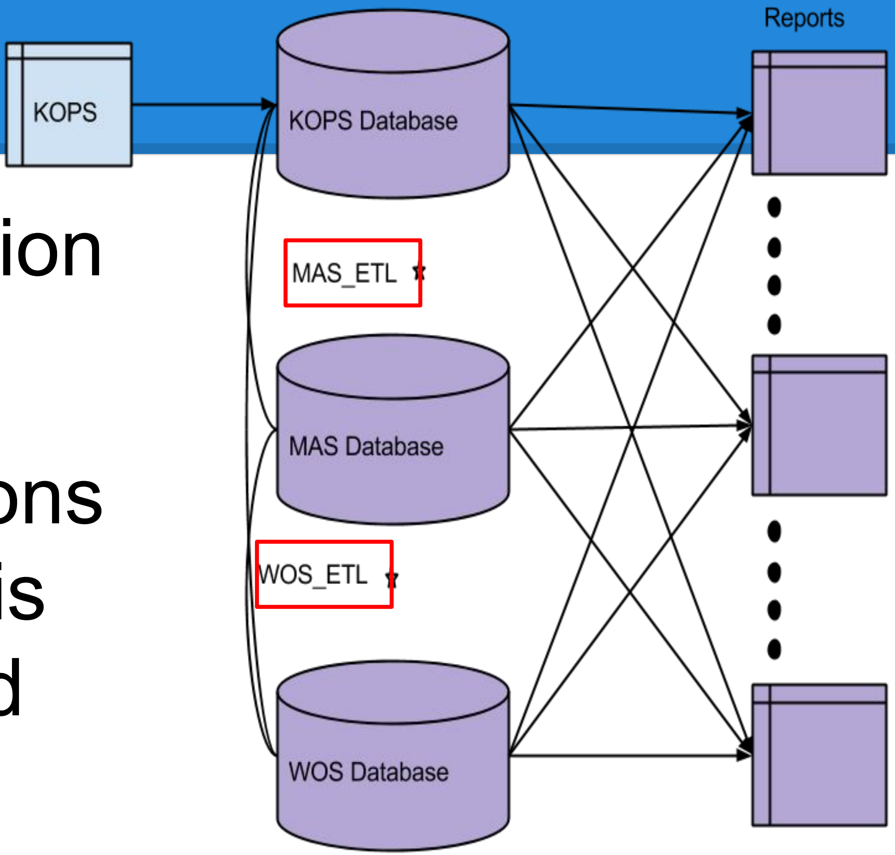
Dhiraj Eadara

Contents

1. Purpose
2. Indexing
3. Objectives
4. ETL for MAS
5. ETL for WOS

Purpose

- Gather information
  - Publications
  - Authors
- Focus on Citations
- Perform Analysis
- Many tools used

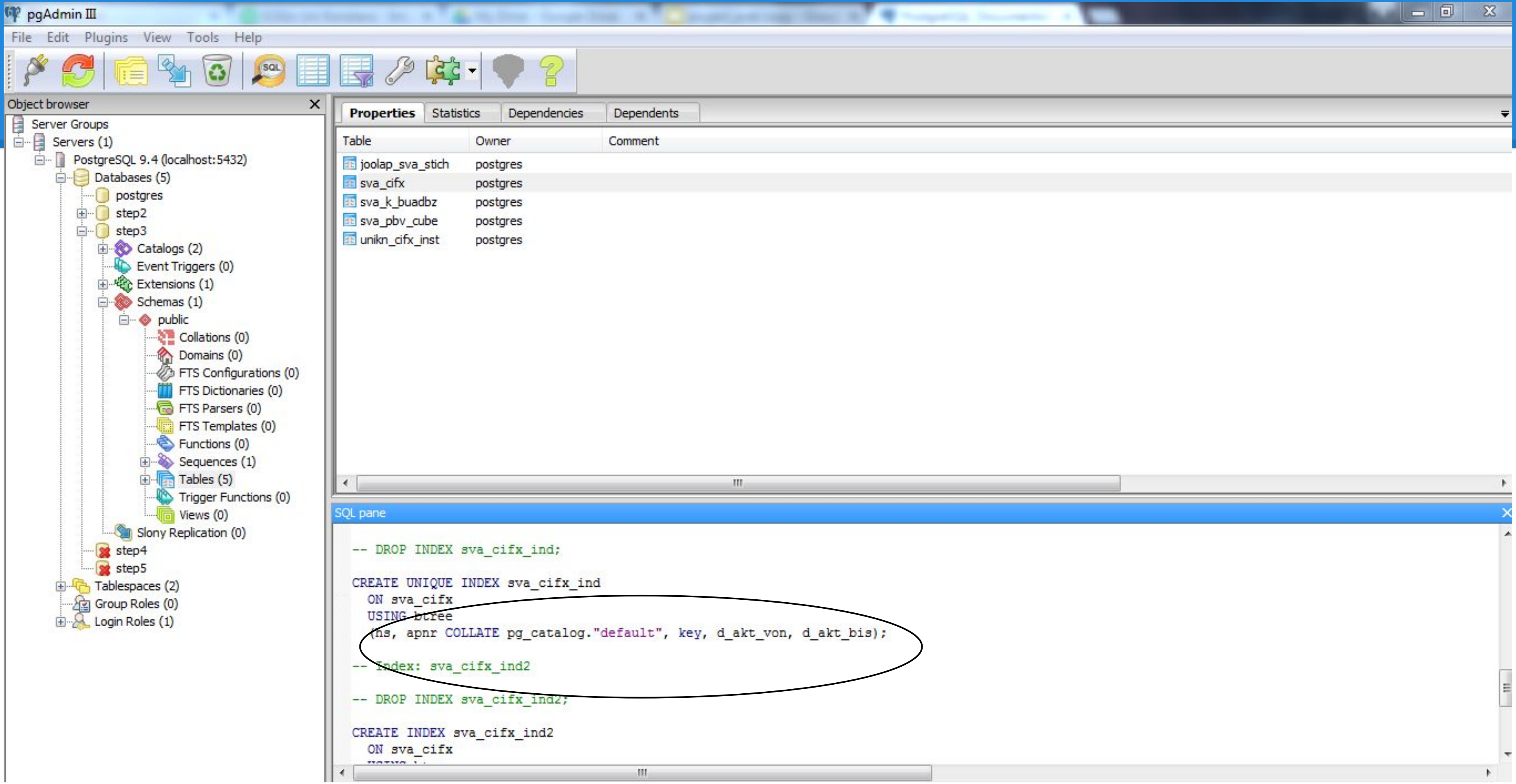


Indexing

- Different types of indexes in Postgresql
  - B-Tree
  - Hash
  - GiST
  - GIN
- Btree preferred
- Hash discouraged, others for arrays

Issues

- Some hash indexes
- No indexes on columns that matter
- Too many multi-column indexes



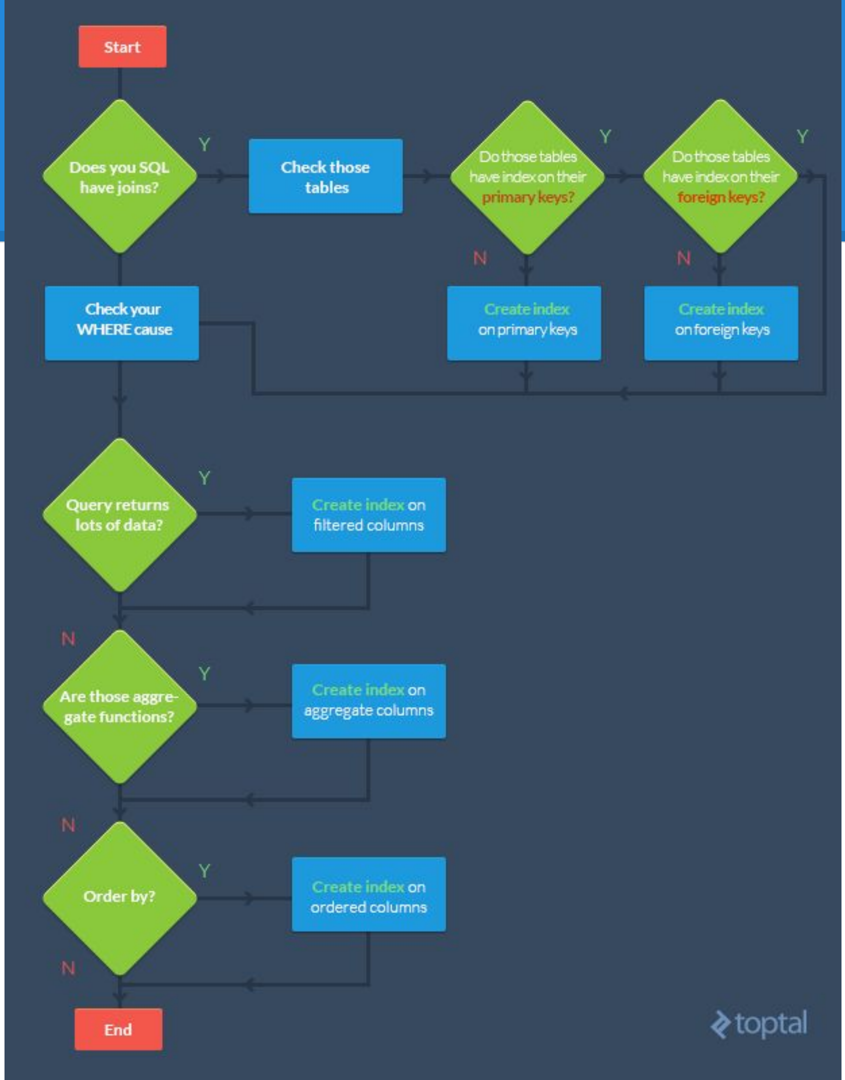
- Server Groups
  - Servers (1)
    - PostgreSQL 9.4 (localhost:5432)
      - Databases (5)
        - postgres
        - step2
        - step3
          - Catalogs (2)
          - Event Triggers (0)
          - Extensions (1)
          - Schemas (1)
            - public
              - Collations (0)
              - Domains (0)
              - FTS Configurations (0)
              - FTS Dictionaries (0)
              - FTS Parsers (0)
              - FTS Templates (0)
              - Functions (0)
              - Sequences (1)
              - Tables (5)
              - Trigger Functions (0)
              - Views (0)
          - Slony Replication (0)
        - step4
        - step5
      - Tablespaces (2)
      - Group Roles (0)
      - Login Roles (1)

Table	Owner	Comment
joolap_sva_stich	postgres	
sva_cifx	postgres	
sva_k_buadbz	postgres	
sva_pbv_cube	postgres	
unikn_cifx_inst	postgres	

```
-- DROP INDEX sva_cifx_ind;

CREATE UNIQUE INDEX sva_cifx_ind
  ON sva_cifx
  USING btree
  (ns, apnr COLLATE pg_catalog."default", key, d_akt_von, d_akt_bis);

-- Index: sva_cifx_ind2

-- DROP INDEX sva_cifx_ind2;

CREATE INDEX sva_cifx_ind2
  ON sva_cifx
```

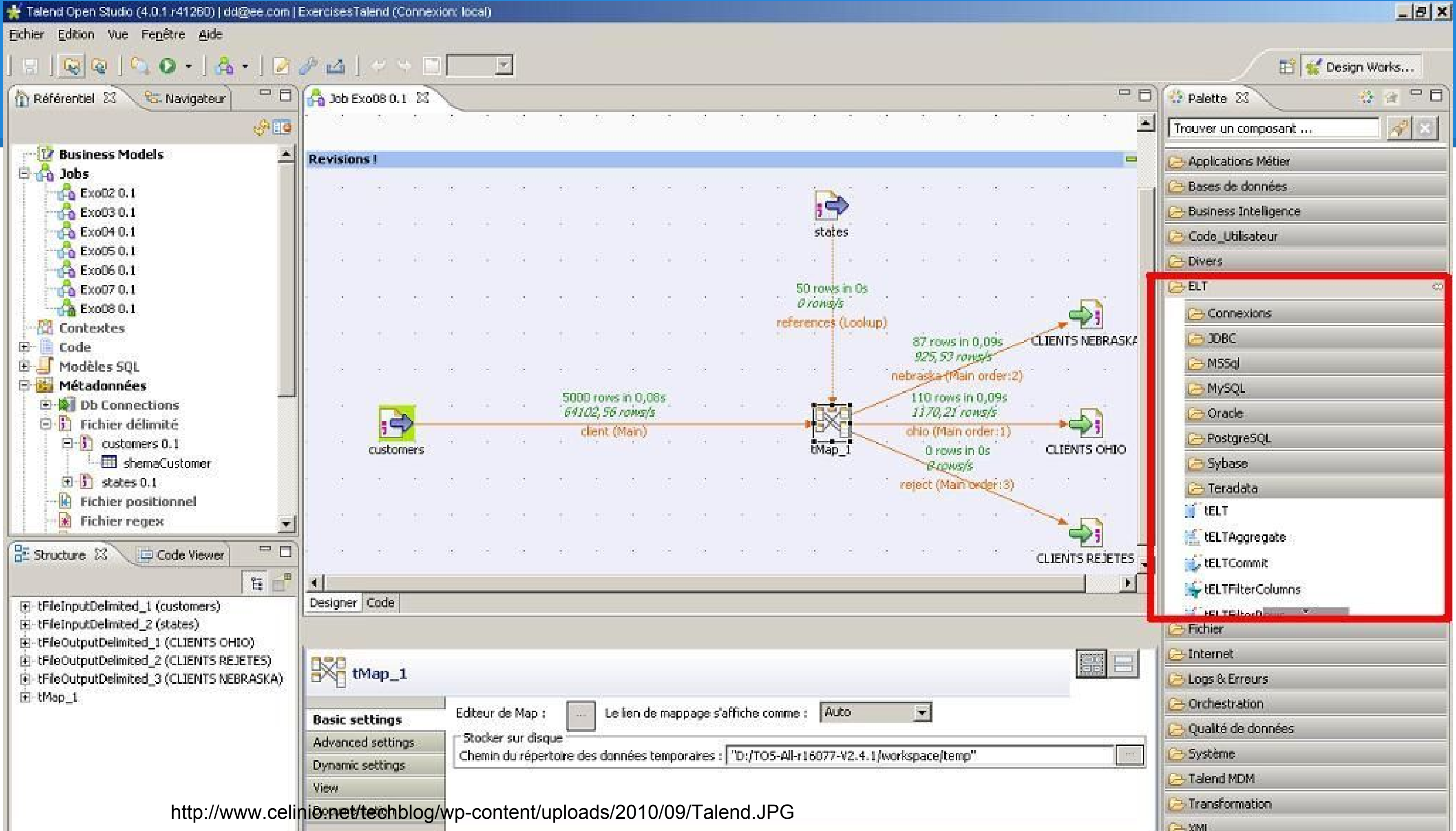
Strategy for indexing

- Check SQL query
- check which columns are used
- index them with btree
- Avoid multicolumn indexes


Start
Does you SQL have joins?
Y
Check those tables
Do those tables have index on their primary keys?
Y
Do those tables have index on their foreign keys?
Y
N
Create index on primary keys
N
Create index on foreign keys
Check your WHERE cause
Query returns lots of data?
Y
Create index on filtered columns
N
Are those aggregate functions?
Y
Create index on aggregate columns
N
Order by?
Y
Create index on ordered columns
N
End
toptal

Talend

- Swiss-army knife of ETL and Big data
- Can connect varied sources
- Offers diverse tools to manipulate data
- Popular in the corporate world
- Developed in Java



```
tFileInputDelimited_1 (customers)
tFileInputDelimited_2 (states)
tFileOutputDelimited_1 (CLIENTS OHIO)
tFileOutputDelimited_2 (CLIENTS REJETES)
tFileOutputDelimited_3 (CLIENTS NEBRASKA)
tMap_1
```

Editeur de Map : ...

Le lien de mappage s'affiche comme : Auto

Stocker sur disque

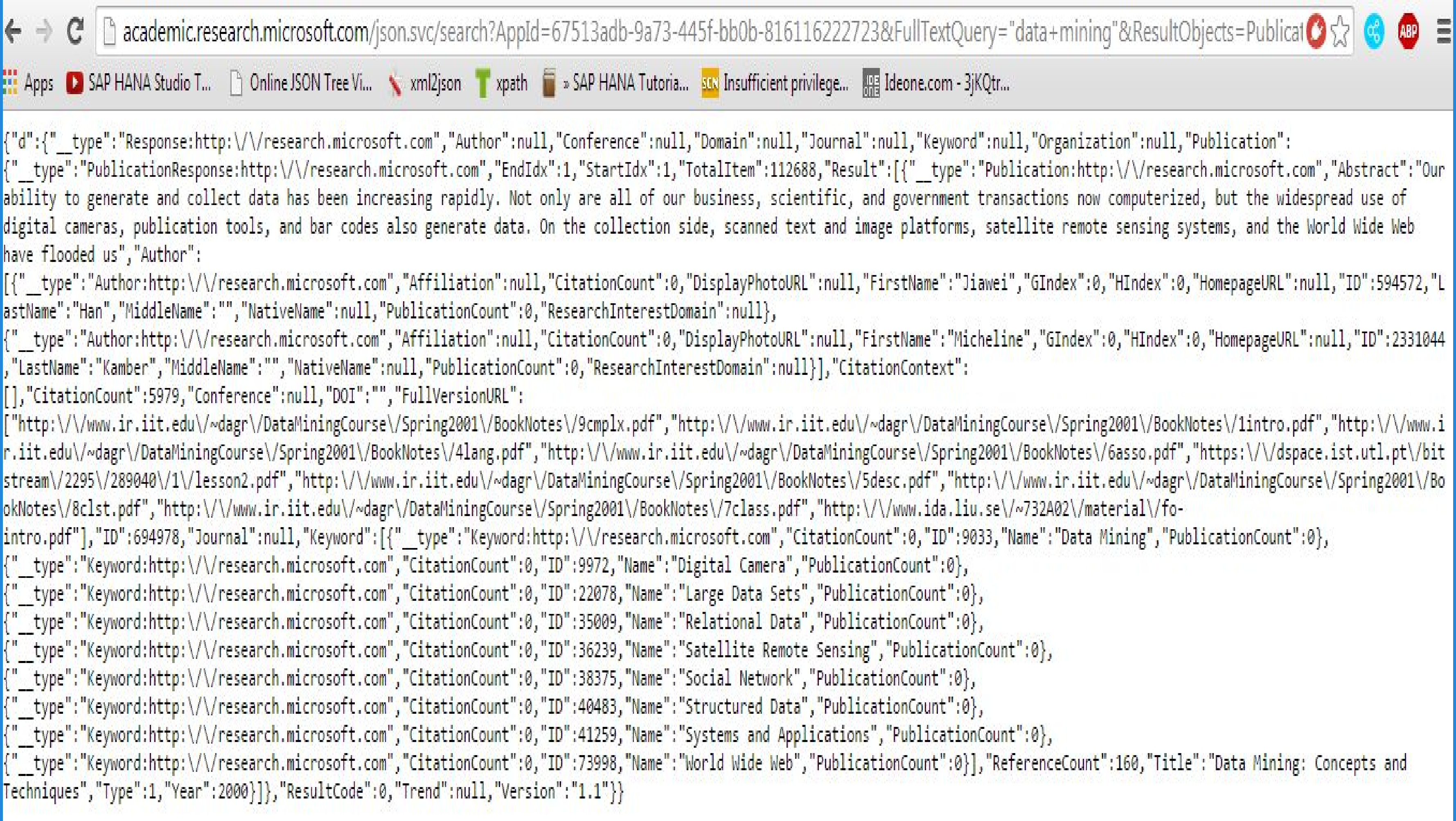
Chemin du répertoire des données temporaires : "D:/TOS-All-r16077-V2.4.1/workspace/temp"

http://www.celinio.net/techblog/wp-content/uploads/2010/09/Talend.JPG

Microsoft Academic Research

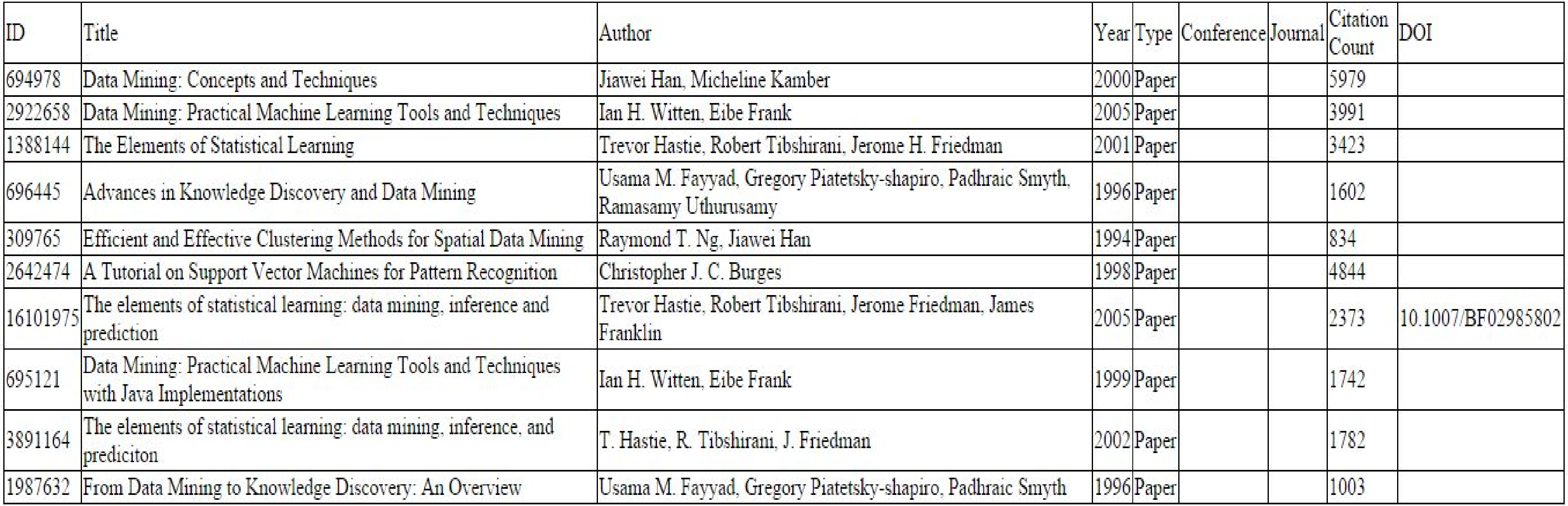
Microsoft Academic Research

- contains information in JSON format
- Information ordered per
  - **Publication**
  - **Author**
  - Conference
  - Journal
  - Organization
  - **Domain**
  - **Keyword**



academic.research.microsoft.com/json.svc/search?AppId=67513adb-9a73-445f-bb0b-816116222723&FullTextQuery="data+mining"&ResultObjects=Publicat
Apps
SAP HANA Studio T...
Online JSON Tree Vi...
xml2json
xpath
» SAP HANA Tutoria...
Insufficient privilege...
Ideone.com - 3jKQtr...

```
{"d":{"__type":"Response:http:\/\/research.microsoft.com","Author":null,"Conference":null,"Domain":null,"Journal":null,"Keyword":null,"Organization":null,"Publication":
{"__type":"PublicationResponse:http:\/\/research.microsoft.com","EndIdx":1,"StartIdx":1,"TotalItem":112688,"Result":[{"__type":"Publication:http:\/\/research.microsoft.com","Abstract":"Our
ability to generate and collect data has been increasing rapidly. Not only are all of our business, scientific, and government transactions now computerized, but the widespread use of
digital cameras, publication tools, and bar codes also generate data. On the collection side, scanned text and image platforms, satellite remote sensing systems, and the World Wide Web
have flooded us","Author":
[{"__type":"Author:http:\/\/research.microsoft.com","Affiliation":null,"CitationCount":0,"DisplayPhotoURL":null,"FirstName":"Jiawei","GIndex":0,"HIndex":0,"HomepageURL":null,"ID":594572,"L
astName":"Han","MiddleName":"","NativeName":null,"PublicationCount":0,"ResearchInterestDomain":null},
{"__type":"Author:http:\/\/research.microsoft.com","Affiliation":null,"CitationCount":0,"DisplayPhotoURL":null,"FirstName":"Micheline","GIndex":0,"HIndex":0,"HomepageURL":null,"ID":2331044
,"LastName":"Kamber","MiddleName":"","NativeName":null,"PublicationCount":0,"ResearchInterestDomain":null}],"CitationContext":
[],"CitationCount":5979,"Conference":null,"DOI":"","FullVersionURL":
["http:\/\/www.ir.iit.edu\/~dagr\/DataMiningCourse\/Spring2001\/BookNotes\/9cmplx.pdf","http:\/\/www.ir.iit.edu\/~dagr\/DataMiningCourse\/Spring2001\/BookNotes\/1intro.pdf","http:\/\/www.i
r.iit.edu\/~dagr\/DataMiningCourse\/Spring2001\/BookNotes\/4lang.pdf","http:\/\/www.ir.iit.edu\/~dagr\/DataMiningCourse\/Spring2001\/BookNotes\/6asso.pdf","https:\/\/dspace.ist.utl.pt\/bit
stream\/2295\/289040\/1\/lesson2.pdf","http:\/\/www.ir.iit.edu\/~dagr\/DataMiningCourse\/Spring2001\/BookNotes\/5desc.pdf","http:\/\/www.ir.iit.edu\/~dagr\/DataMiningCourse\/Spring2001\/Bo
okNotes\/8clst.pdf","http:\/\/www.ir.iit.edu\/~dagr\/DataMiningCourse\/Spring2001\/BookNotes\/7class.pdf","http:\/\/www.ida.liu.se\/~732A02\/material\/fo-
intro.pdf"],"ID":694978,"Journal":null,"Keyword":[{"__type":"Keyword:http:\/\/research.microsoft.com","CitationCount":0,"ID":9033,"Name":"Data Mining","PublicationCount":0},
{"__type":"Keyword:http:\/\/research.microsoft.com","CitationCount":0,"ID":9972,"Name":"Digital Camera","PublicationCount":0},
{"__type":"Keyword:http:\/\/research.microsoft.com","CitationCount":0,"ID":22078,"Name":"Large Data Sets","PublicationCount":0},
{"__type":"Keyword:http:\/\/research.microsoft.com","CitationCount":0,"ID":35009,"Name":"Relational Data","PublicationCount":0},
{"__type":"Keyword:http:\/\/research.microsoft.com","CitationCount":0,"ID":36239,"Name":"Satellite Remote Sensing","PublicationCount":0},
{"__type":"Keyword:http:\/\/research.microsoft.com","CitationCount":0,"ID":38375,"Name":"Social Network","PublicationCount":0},
{"__type":"Keyword:http:\/\/research.microsoft.com","CitationCount":0,"ID":40483,"Name":"Structured Data","PublicationCount":0},
{"__type":"Keyword:http:\/\/research.microsoft.com","CitationCount":0,"ID":41259,"Name":"Systems and Applications","PublicationCount":0},
{"__type":"Keyword:http:\/\/research.microsoft.com","CitationCount":0,"ID":73998,"Name":"World Wide Web","PublicationCount":0}],"ReferenceCount":160,"Title":"Data Mining: Concepts and
Techniques","Type":1,"Year":2000}]},"ResultCode":0,"Trend":null,"Version":"1.1"}}
```



ID	Title	Author	Year	Type	Conference	Journal	Citation Count	DOI
694978	Data Mining: Concepts and Techniques	Jiawei Han, Micheline Kamber	2000	Paper			5979	
2922658	Data Mining: Practical Machine Learning Tools and Techniques	Ian H. Witten, Eibe Frank	2005	Paper			3991	
1388144	The Elements of Statistical Learning	Trevor Hastie, Robert Tibshirani, Jerome H. Friedman	2001	Paper			3423	
696445	Advances in Knowledge Discovery and Data Mining	Usama M. Fayyad, Gregory Piatetsky-shapiro, Padhraic Smyth, Ramasamy Uthurusamy	1996	Paper			1602	
309765	Efficient and Effective Clustering Methods for Spatial Data Mining	Raymond T. Ng, Jiawei Han	1994	Paper			834	
2642474	A Tutorial on Support Vector Machines for Pattern Recognition	Christopher J. C. Burges	1998	Paper			4844	
16101975	The elements of statistical learning: data mining, inference and prediction	Trevor Hastie, Robert Tibshirani, Jerome Friedman, James Franklin	2005	Paper			2373	10.1007/BF02985802
695121	Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations	Ian H. Witten, Eibe Frank	1999	Paper			1742	
3891164	The elements of statistical learning: data mining, inference, and prediciton	T. Hastie, R. Tibshirani, J. Friedman	2002	Paper			1782	
1987632	From Data Mining to Knowledge Discovery: An Overview	Usama M. Fayyad, Gregory Piatetsky-shapiro, Padhraic Smyth	1996	Paper			1003	

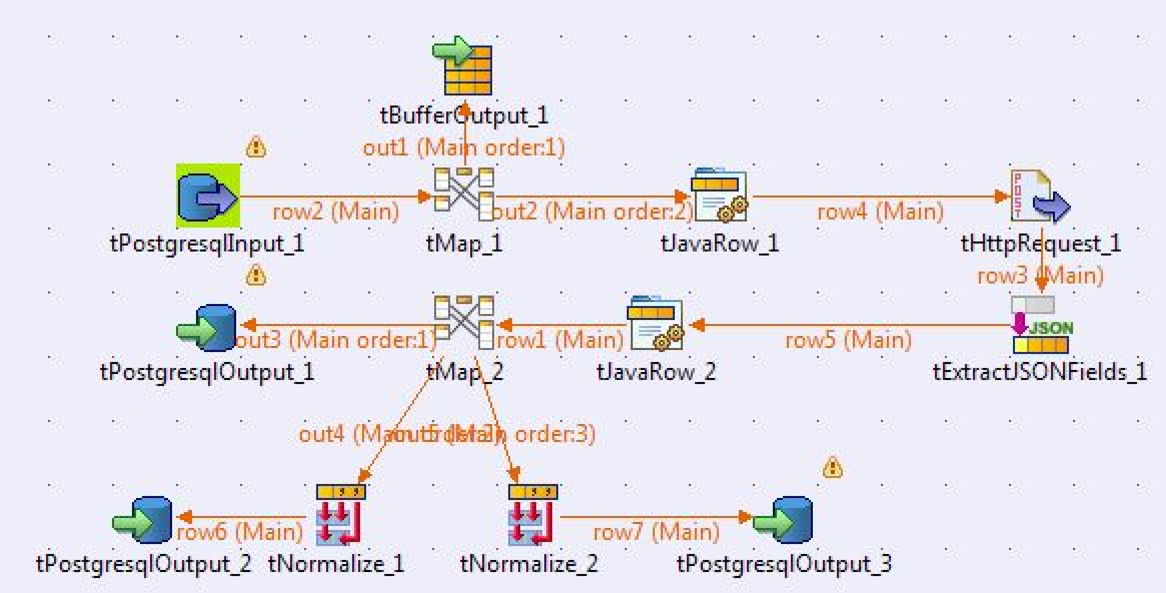
Methodology

- Search data in database in the website
- Retrieve, read and extract data in JSON
- Manipulate the data
- Transfer to database

Objectives

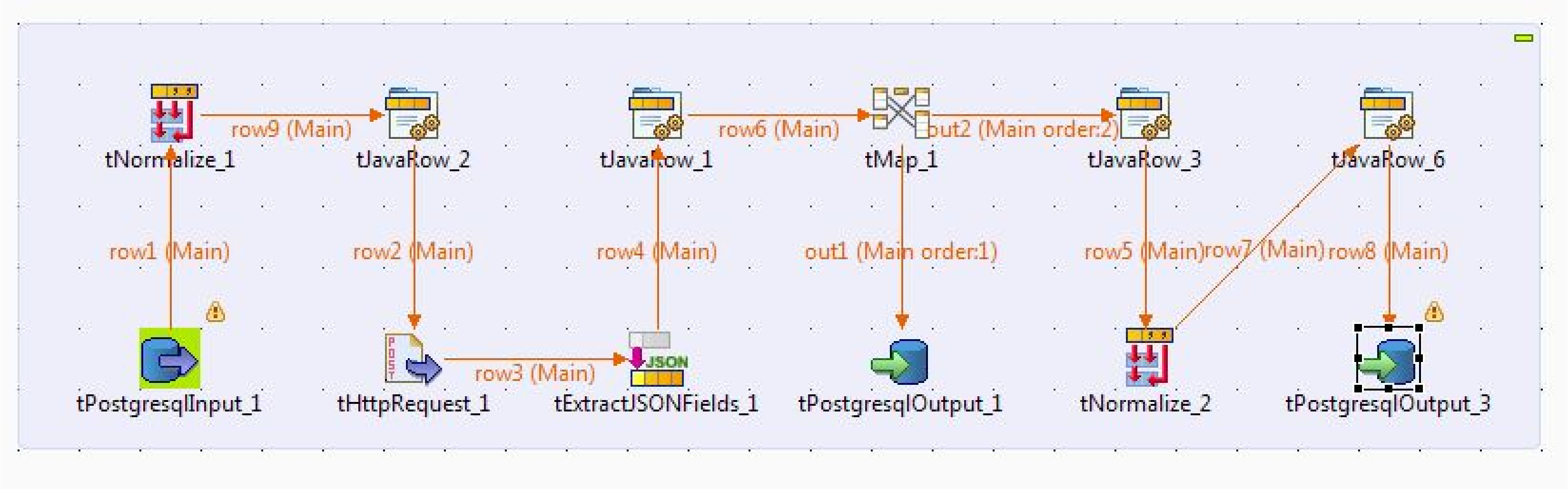
1. kops_msapi
   a. table for publications
   b. fact tables for keywords and authors
2. autoren
   a. tables for authors
   b. fact tables for research domains
3. key tables for keywords and domains
4. year_wise_citations for publication & author

kops_msapi



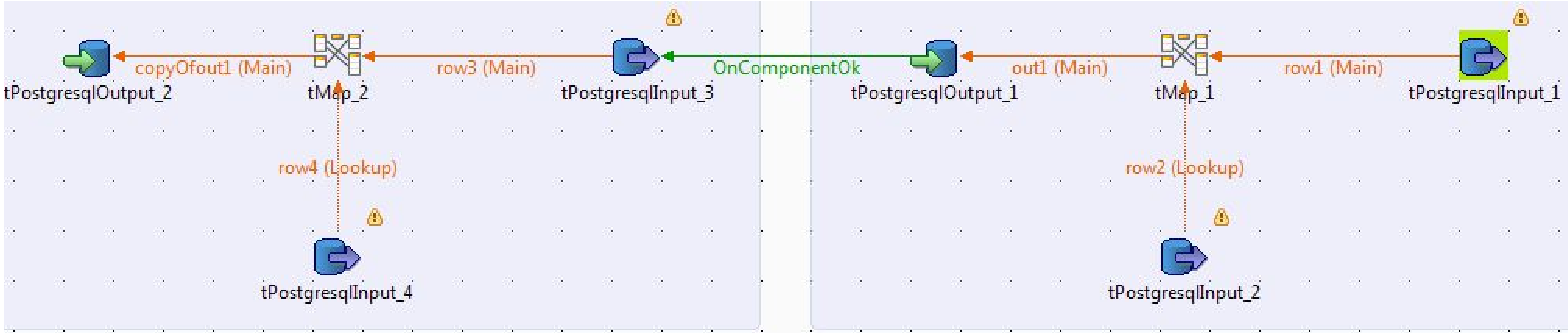
- Search for data contained in kops database
- Store it in new table by title
- Creates two fact tables w.r.t authors, keywords

autoren



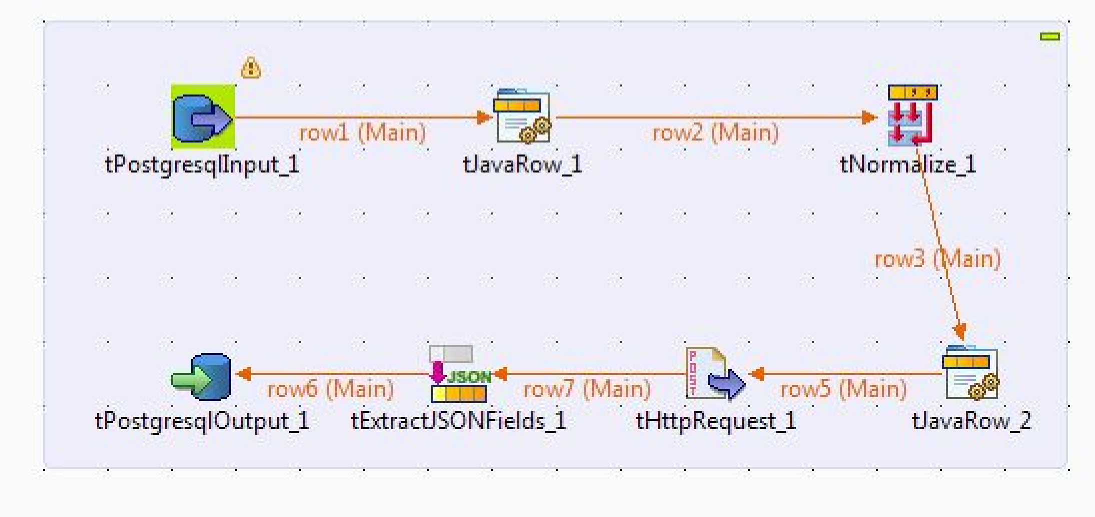
- Extract id for author & Search in database
- Store all relevant author ids in the api
- Creates a fact table for w.r.t research domains

mas_autoren_2_pop



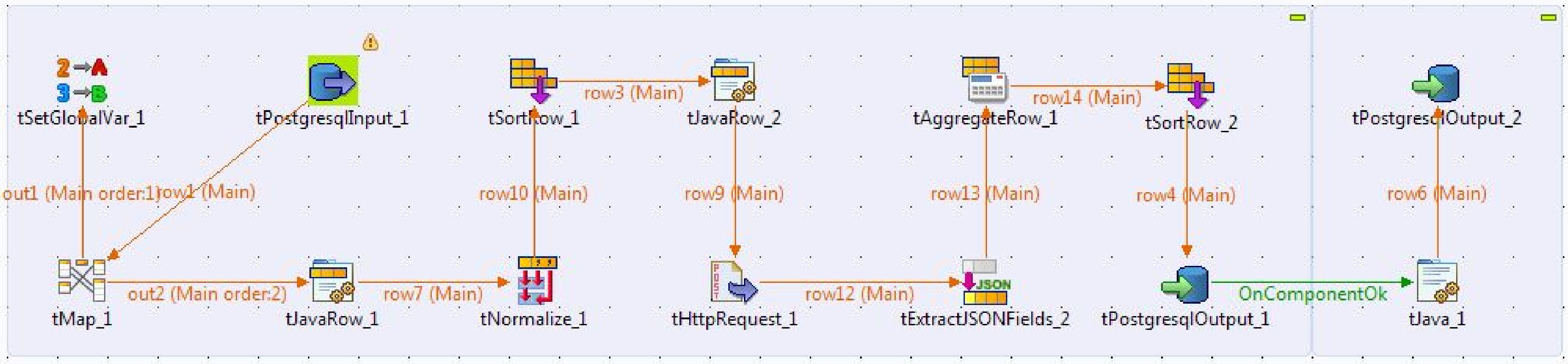
- Fill the pop_id
- Previous job leaves pop_id empty
- The key is the author’s name

keyword_details



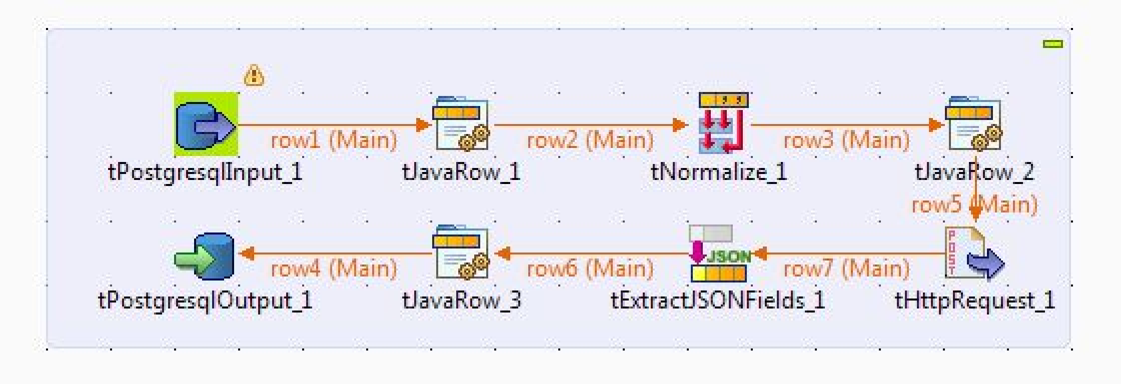
- Each keyword is then looked up with id
- All details are then stored in key table
- Publication count and citation count

year_wise_citation_publications



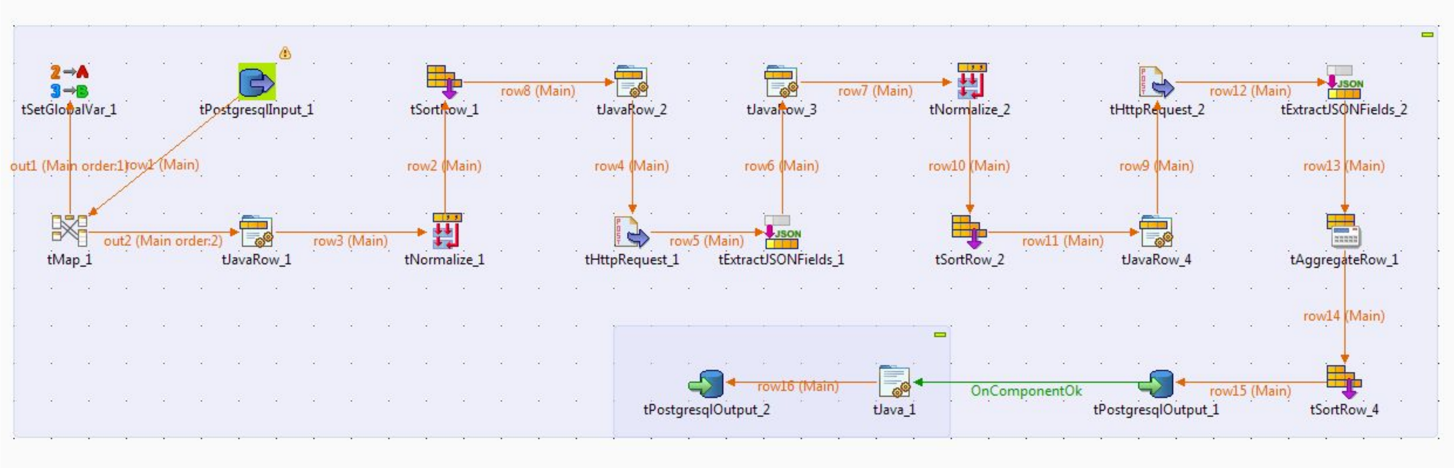
- Stores citation counts for each publication every year

domain_details



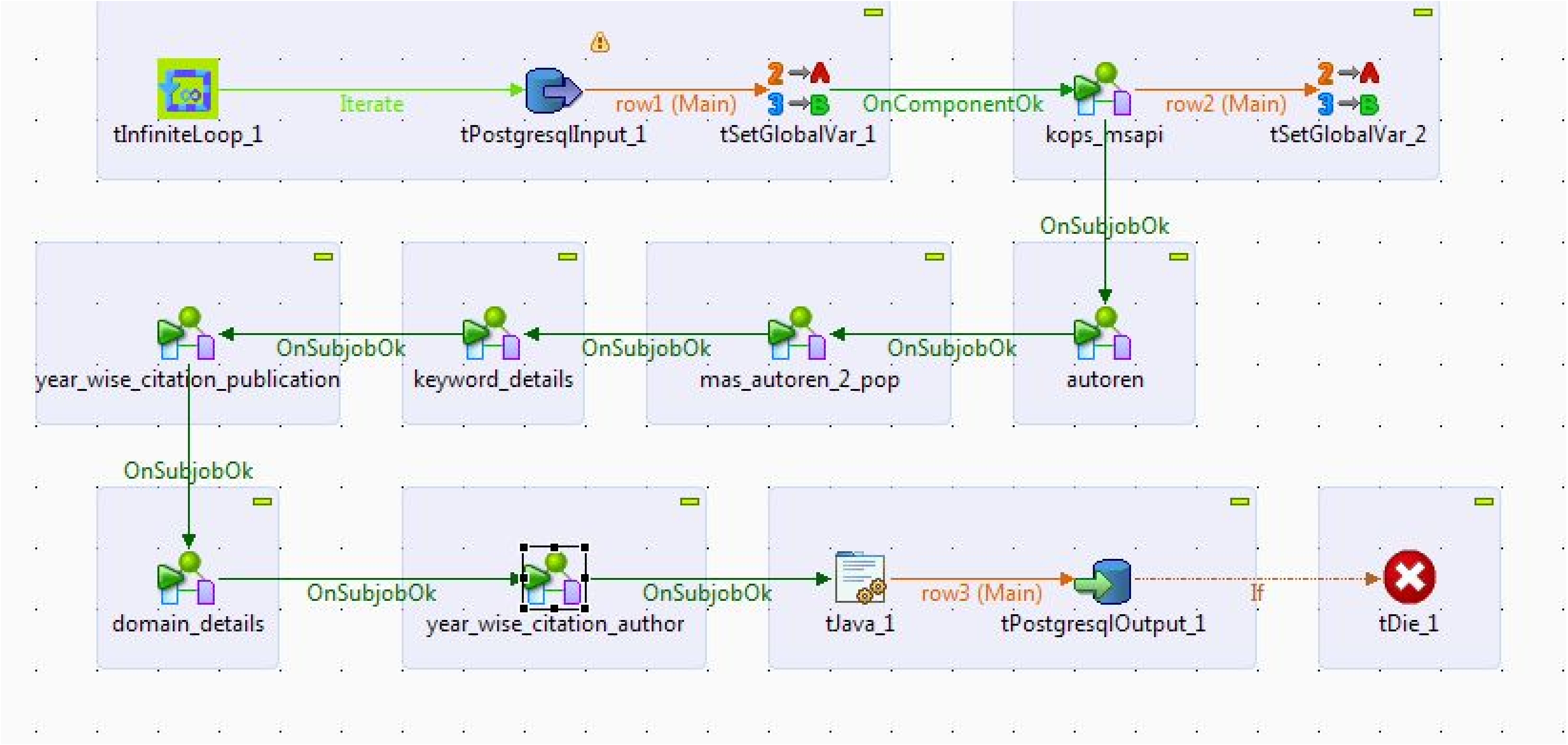
- searches by domain id
- Stores research domains, specialisation & Number of publications and citations

year_wise_citation_author



- Stores citation counts for each author every year

Overall


tInfiniteLoop_1
Iterate
tPostgresqlInput_1
row1 (Main)
tSetGlobalVar_1
OnComponentOk
kops_msapi
row2 (Main)
tSetGlobalVar_2
OnSubjobOk
year_wise_citation_publication
OnSubjobOk
keyword_details
OnSubjobOk
mas_autoren_2_pop
OnSubjobOk
autoren
OnSubjobOk
domain_details
OnSubjobOk
year_wise_citation_author
OnSubjobOk
tJava_1
row3 (Main)
tPostgresqlOutput_1
If
tDie_1

Challenges

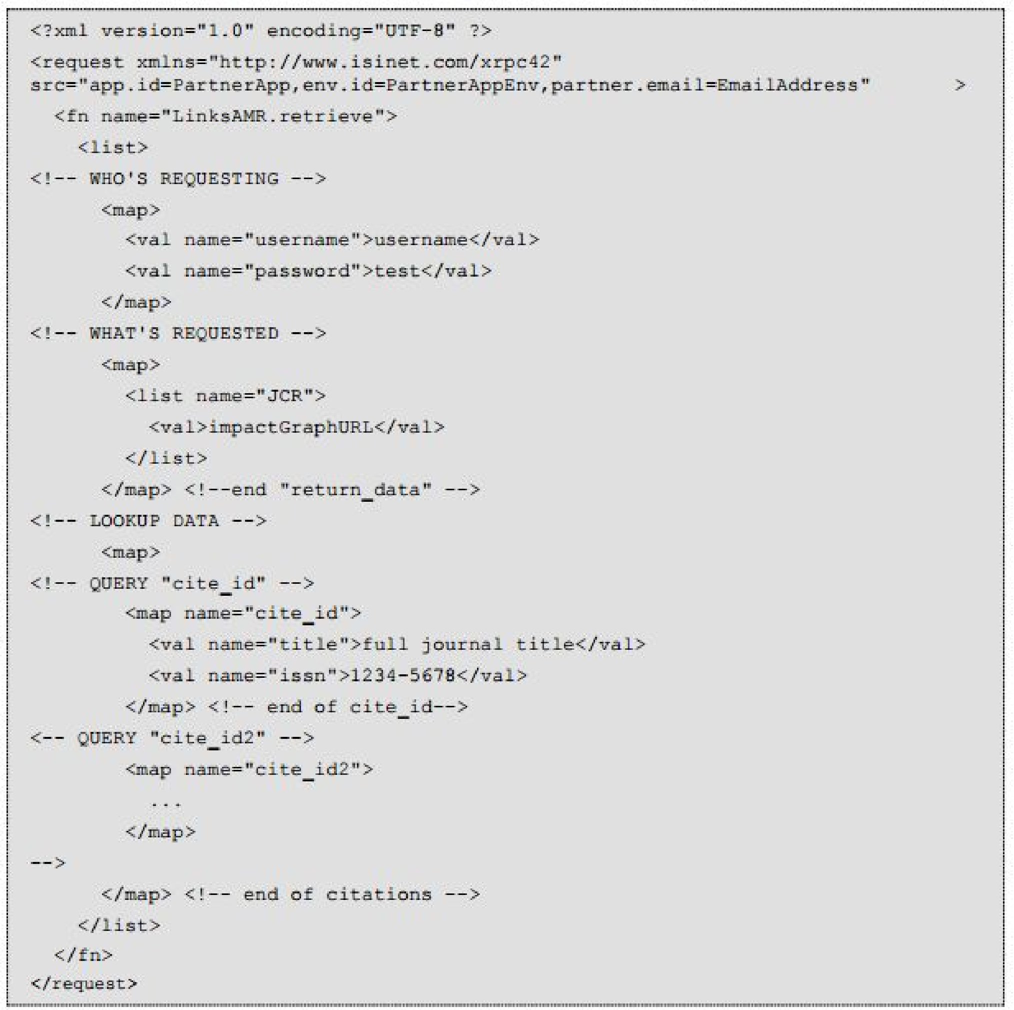
- Restrictions on flow to 2 queries per second
- Restrictions on number of entries retrieved
- Ability to start and restart any point of time

ETL Process for Web of Science of Data

Web of Science

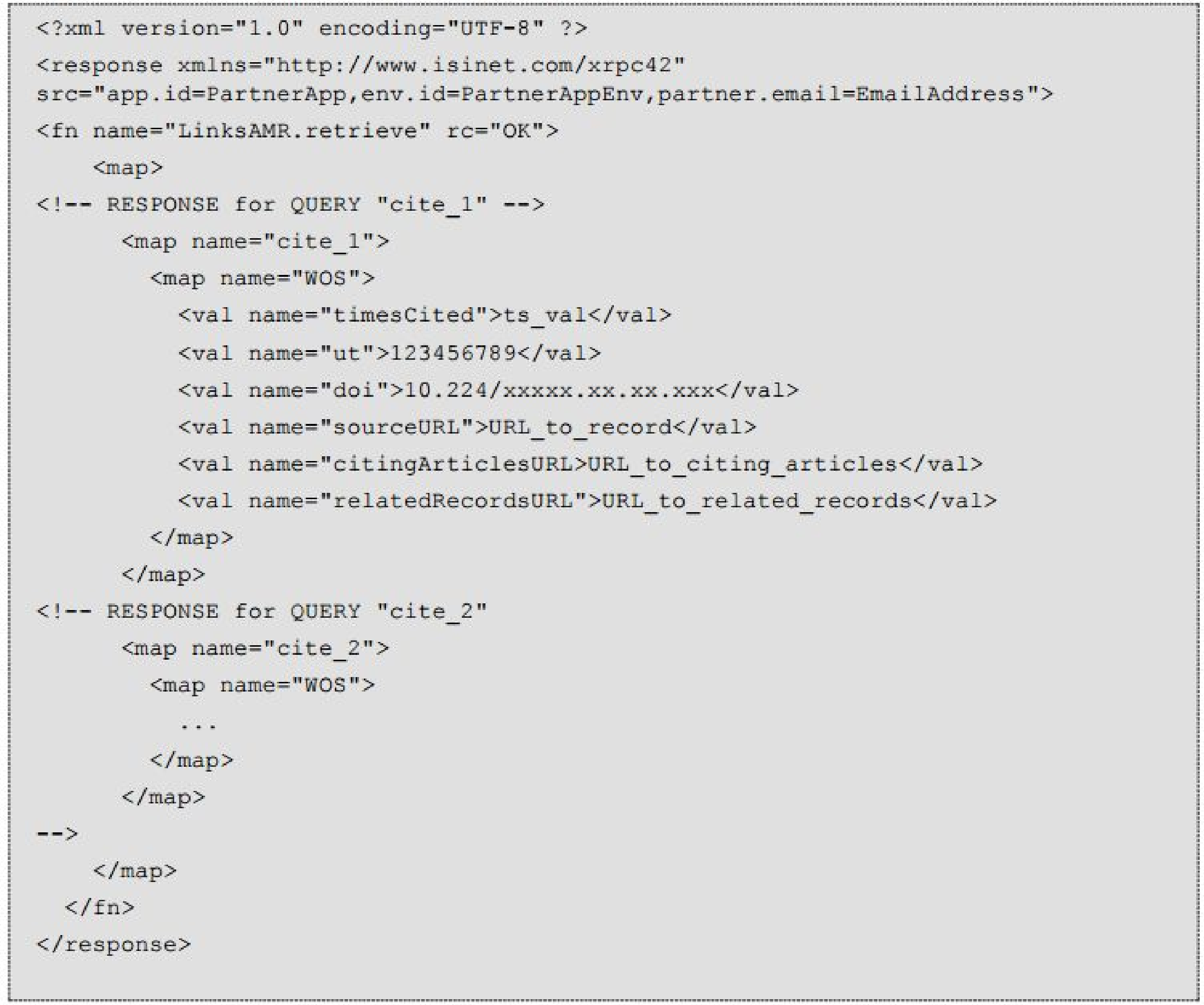
- contains information in XML format
- Information is as follows
  - doi
  - issn
  - ut
  - TimesCited
  - Title, etc.

Request



```
<?xml version="1.0" encoding="UTF-8" ?>
<request xmlns="http://www.isinet.com/xrpc42"
src="app.id=PartnerApp,env.id=PartnerAppEnv,partner.email=EmailAddress"       >
  <fn name="LinksAMR.retrieve">
    <list>
<!-- WHO'S REQUESTING -->
      <map>
        <val name="username">username</val>
        <val name="password">test</val>
      </map>
<!-- WHAT'S REQUESTED -->
      <map>
        <list name="JCR">
          <val>impactGraphURL</val>
        </list>
      </map> <!--end "return_data" -->
<!-- LOOKUP DATA -->
      <map>
<!-- QUERY "cite_id" -->
        <map name="cite_id">
          <val name="title">full journal title</val>
          <val name="issn">1234-5678</val>
        </map> <!-- end of cite_id-->
<-- QUERY "cite_id2" -->
        <map name="cite_id2">
          ...
        </map>
-->
      </map> <!-- end of citations -->
    </list>
  </fn>
</request>
```

Response

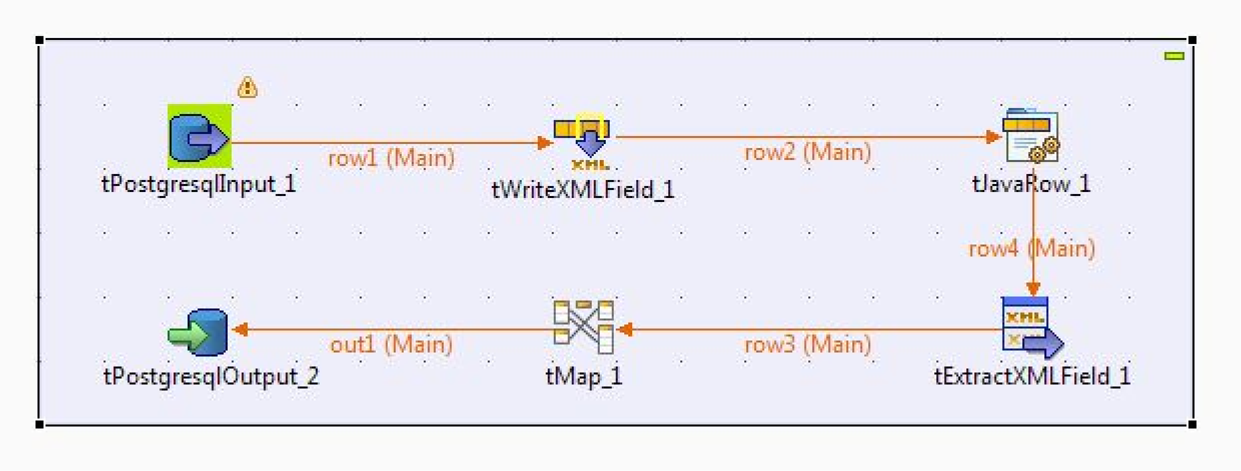


```
<?xml version="1.0" encoding="UTF-8" ?>
<response xmlns="http://www.isinet.com/xrpc42"
src="app.id=PartnerApp,env.id=PartnerAppEnv,partner.email=EmailAddress">
<fn name="LinksAMR.retrieve" rc="OK">
    <map>
<!-- RESPONSE for QUERY "cite_1" -->
      <map name="cite_1">
        <map name="WOS">
          <val name="timesCited">ts_val</val>
          <val name="ut">123456789</val>
          <val name="doi">10.224/xxxxx.xx.xx.xxx</val>
          <val name="sourceURL">URL_to_record</val>
          <val name="citingArticlesURL>URL_to_citing_articles</val>
          <val name="relatedRecordsURL">URL_to_related_records</val>
        </map>
      </map>
<!-- RESPONSE for QUERY "cite_2"
      <map name="cite_2">
        <map name="WOS">
          ...
        </map>
      </map>
-->
    </map>
  </fn>
</response>
```

Objectives

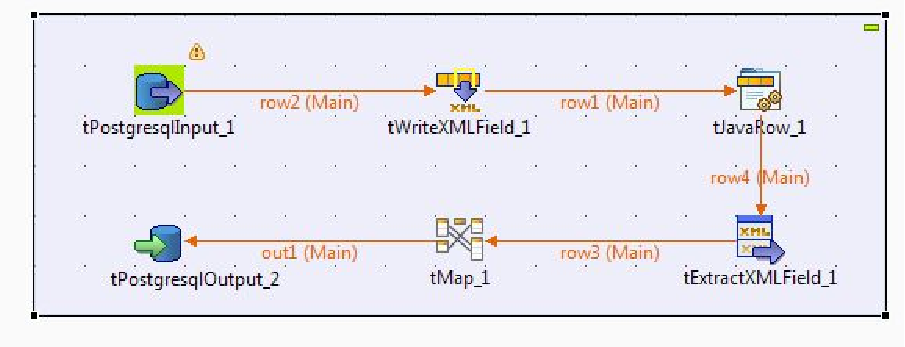
1. Create, send and parse xml responses
    a. wos_4_doi
    b. wos_xml_creator_for_authors
    c. wos_xml_creator_for_journals

wos_4_doi



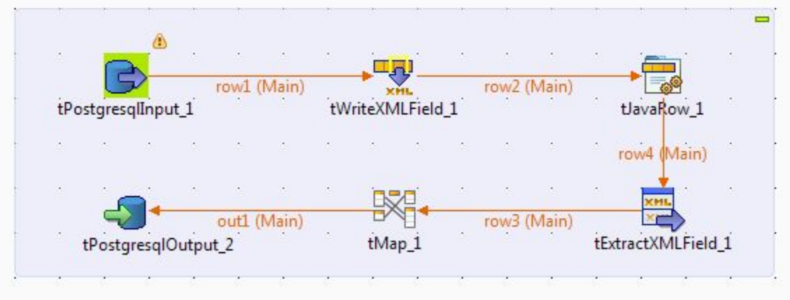
- creates & sends xml requests with doi from both mas and kops tables
- Processes responses

wos_xml_creator_for_journals



- if doi is missing use the following
    - Journal title
    - Volume
    - Issue
    - Start page or article number

wos_xml_creator_for_authors



- if doi is missing use the following
    - any author name
    - Volume
    - issn
    - Issue
    - Start page or article number

Challenges

- tHTTPRequest failed to POST requests
- Used custom code
- Derived from component and modified