

SAS® GLOBAL FORUM 2018

USERS PROGRAM

Introduction to ETL with SAS®

Presented by Vasilij Nevlev
Analytium Ltd

April 8 - 11 | Denver, CO
#SASGF

Introduction to ETL with SAS®

Presented by Vasilij Nevlev

Analytium Ltd

Why ETL is important?

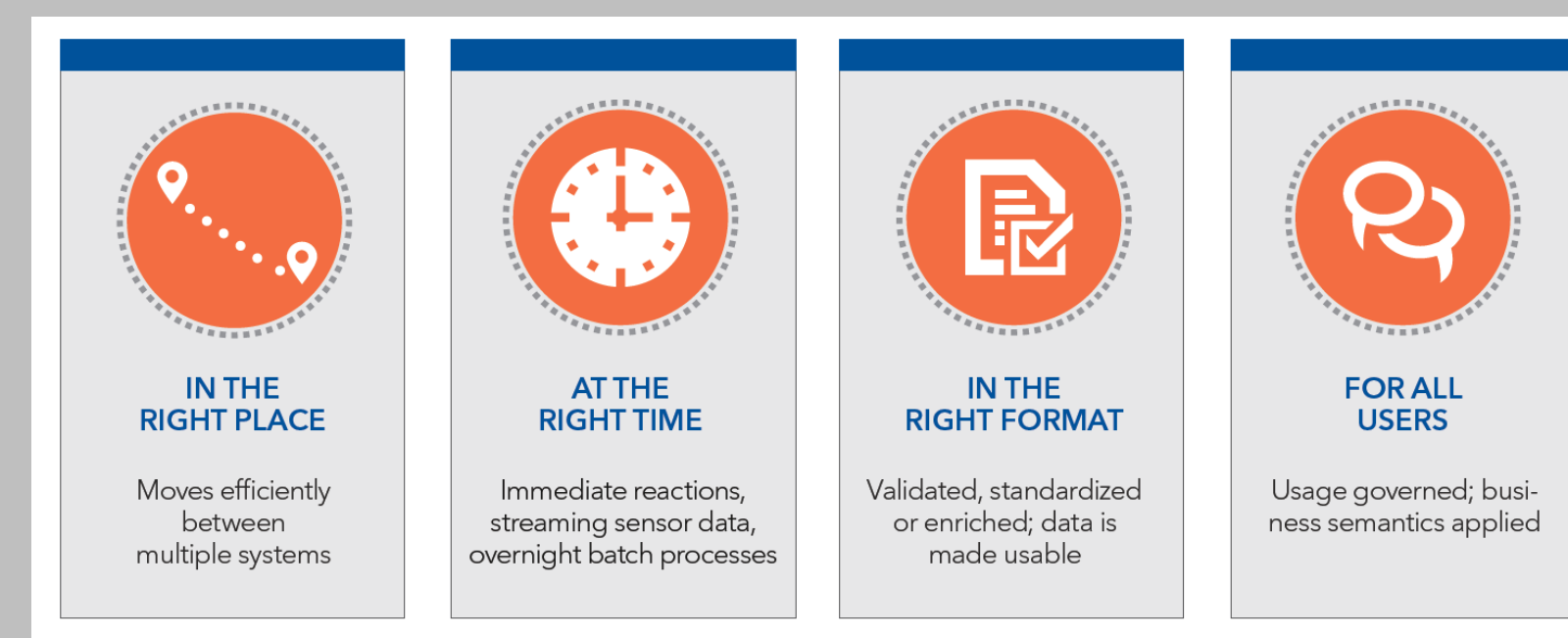
If you are here, at SAS Global Forum, you are probably involved in data management or data consumption in one or more ways.

It might sound obvious, by difficult to achieve, the data is often required to be:

- In the right place
- At the right time
- In the right format
- Available for the rights users

ETL framework existing to meet these objectives

Data Management Objectives



What is ETL?

- ETL is an acronym for “Extract, Transform, Load” which describes key stages and their order in a typical data management process.
- ETL process is often, but not always, implemented at an enterprise level as a data warehouse
- “A data warehouse is a system that extracts, cleans, conforms and delivers sources data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making”
Source: Ralph Kimball, Joe Caserta: The Data Warehouse ETL Toolkit; Wiley 2004
- The most important part to the business is “querying and analysis”
- The most complex and time consuming part is “extracts, cleans, conforms and delivers”.

Well developed ETL process helps you to:

- Access the data you need
- Improve productivity
- Reduce development and maintenance time
- Govern and secure your data
- Work faster and meet time constrains
- Eliminate overlapping and redundant tools

When there is no managed ETL

When there is no defined approached to ETL, organisations end up with creating multiples views of their data (demonstrated below). These are monolithic chunks of code that no one understands, difficult to manage and often produce unexpected results.

```
SELECT
  company_id, dbo_mag_dum_sap_sap_id, company.name, company.phone, company.street, dbo.Company.zipcode,
  company.city,
  --bestp_id_posting_p_ct_id_contact,
  p_ct_sabaction_p_sabaction,
  p_ct_academic_title_p_title, p_ct_firstname_firstname, p_ct_lastname_lastname, [email], p_ct_email_email, [email],
  p_ct_sabaction_sabaction, p_ct_academic_title_p_title, p_ct_firstname_firstname, p_ct_lastname_lastname, [email], p_ct_email_email, [email],
  --bestp_corporation_posted_source,
  bestp_id_posting, bestp_description, bestp_date_posted, sap_company_emp_resp_login, sap_company_emp_resp_firstname, sap_company_emp_resp_lastname, sap_company_emp_resp_email, sap_company_emp_resp_phone, sap_cat_value [customer category], sap_emp_reporting_line_unit

FROM dbo.Company

ORDER 2020 dbo.Postings ON dbo.Postings.ref_id_company = dbo.Company.company_id
ORDER 2020 dbo.sources ON dbo.Postings.Posted_Source = dbo.sources.code
LEFT JOIN dbo.Excluded_Words ON Postings.description LIKE '%' + Excluded_Words.Word + '%' AND Excluded_Words.country='it'
--inner 2020 dbo.posting_profession pf ON pf.ref_id_posting = Postings.ref_id_posting
--inner 2020 dbo.Profession p ON p.id_profession = pf.ref_id_profession

LEFT OUTER 2020 dbo_mag_index ON dbo_mag_index.index_id = dbo.Postings.ref_id_company AND so_country = 'it'
LEFT OUTER 2020 dbo_mag_provider_dum ON dbo_mag_provider_dum.provider_id = company.company_id AND dbo_mag_provider_dum.ref_provider = 1
LEFT 2020 dbo_mag_dum_sap ON dbo_mag_dum_sap.dum_id = dbo_mag_provider_dum.dum_id
LEFT 2020 vw_mag_dum_post_sap ON sap_dum_account_id AND sap_id IS NOT NULL
LEFT 2020 emp ON company.company_id=emp.ref_id_company
LEFT 2020 customer_sap ON customers_sap.sap_account_id = sap_id

ORDER APPLY SELECT TOP 1 id_posting IN best_p_description_corporation_data_posted_p_description_profession IN best_p_language_email ref_id_contact FROM dbo.Postings IN best_p

ORDER 2020 dbo.sources IN best_p_sources ON IN best_p_posted_source = IN best_p_sources.code
LEFT JOIN dbo.Excluded_Words ON Postings.description LIKE '%' + Excluded_Words.Word + '%' AND Excluded_Words.country='it'
ORDER 2020 dbo.posting_profession pf ON pf.ref_id_posting = IN best_p_id_posting
WHERE IN best_p_ref_id_company = dbo.Company.company_id
AND IN best_p_source_id_source IN (443)
AND IN best_p_date_posted = GETDATE(), GETDATE(), 0) - 30
AND IN best_p_date_posted = GETDATE(), GETDATE(), 0) - 30
AND IN best_p_ref_hierarchy NOT IN (1,2,4,8,16)
/*Exclude accounts if their postings only include PROFESSION equal to: 3 Architects, 13 Hotels and restaurants, tourism, reception, 18 Call center, 19 Accountancy, 20 warehouse worker, 21 Craft, manufacturing*/
ORDER BY id_posting DESC, so_preference .LIKE(email) DESC)@emp

ORDER APPLY (SELECT TOP 1 FROM dbo.Contact C JOIN WHERE C.ref_id_company = company.company_id ORDER BY C.ct_id
LEFT JOIN dbo.Contact p_ct ON bestp_ref_id_contact = p_ct_id_contact
LEFT JOIN SAP_Reporting dbo_vw_account_general_sap_company ON dbo_mag_dum_sap_sap_id = sap_company.account_id
LEFT JOIN SAP_Reporting dbo_vw_employee_sap_cat ON sap_company.emp_resp_cat = sap_cat.code
LEFT JOIN SAP_Reporting dbo_vw_employee_sap_emp ON sap_company.emp_resp_fm = sap_emp.emp1_usd

WHERE IN 3
--AND sap_company.account_id IS NULL
--AND emp.ref_id_company IS NULL
--AND (dbo.sources.country = 'de')
/* AND (and(is(sap_company.business_not_allowed, ''))= '' *)

/*-companies that posted a listings on the free governmental job board "Bundesagentur für Arbeit" the week before (Friday to Thursday previous week) */
/*Monday last week*/ AND dbo.Postings.Date_Posted = GETDATE(), GETDATE(), 0) - 30
/*Friday 00:00:00 this week*/ AND dbo.Postings.Date_Posted = GETDATE(), GETDATE(), 0) - 30

/*-country is DE (index data)*/
AND dbo.Company.country IN ('', 'DE')
/*-are not crafting (index data)*/
AND company.industry NOT IN ('084.1', '197', '198', '199.1', '199.2', '199', '199.3', '199.4', '199.5', '199.6', '199.7', '199.8', '199.9', '199.10', '199.11', '199.12', '199.13', '199.14', '199.15', '199.16', '199.17', '199.18', '199.19', '199.20', '199.21', '199.22', '199.23', '199.24', '199.25', '199.26', '199.27', '199.28', '199.29', '199.30', '199.31', '199.32', '199.33', '199.34', '199.35', '199.36', '199.37', '199.38', '199.39', '199.40', '199.41', '199.42', '199.43', '199.44', '199.45', '199.46', '199.47', '199.48', '199.49', '199.50', '199.51', '199.52', '199.53', '199.54', '199.55', '199.56', '199.57', '199.58', '199.59', '199.60', '199.61', '199.62', '199.63', '199.64', '199.65', '199.66', '199.67', '199.68', '199.69', '199.70', '199.71', '199.72', '199.73', '199.74', '199.75', '199.76', '199.77', '199.78', '199.79', '199.80', '199.81', '199.82', '199.83', '199.84', '199.85', '199.86', '199.87', '199.88', '199.89', '199.90', '199.91', '199.92', '199.93', '199.94', '199.95', '199.96', '199.97', '199.98', '199.99', '200', '201', '202', '203', '204', '205', '206', '207', '208', '209', '210', '211', '212', '213', '214', '215', '216', '217', '218', '219', '220', '221', '222', '223', '224', '225', '226', '227', '228', '229', '230', '231', '232', '233', '234', '235', '236', '237', '238', '239', '240', '241', '242', '243', '244', '245', '246', '247', '248', '249', '250', '251', '252', '253', '254', '255', '256', '257', '258', '259', '260', '261', '262', '263', '264', '265', '266', '267', '268', '269', '270', '271', '272', '273', '274', '275', '276', '277', '278', '279', '280', '281', '282', '283', '284', '285', '286', '287', '288', '289', '290', '291', '292', '293', '294', '295', '296', '297', '298', '299', '300', '301', '302', '303', '304', '305', '306', '307', '308', '309', '310', '311', '312', '313', '314', '315', '316', '317', '318', '319', '320', '321', '322', '323', '324', '325', '326', '327', '328', '329', '330', '331', '332', '333', '334', '335', '336', '337', '338', '339', '340', '341', '342', '343', '344', '345', '346', '347', '348', '349', '350', '351', '352', '353', '354', '355', '356', '357', '358', '359', '360', '361', '362', '363', '364', '365', '366', '367', '368', '369', '370', '371', '372', '373', '374', '375', '376', '377', '378', '379', '380', '381', '382', '383', '384', '385', '386', '387', '388', '389', '390', '391', '392', '393', '394', '395', '396', '397', '398', '399', '400', '401', '402', '403', '404', '405', '406', '407', '408', '409', '410', '411', '412', '413', '414', '415', '416', '417', '418', '419', '420', '421', '422', '423', '424', '425', '426', '427', '428', '429', '430', '431', '432', '433', '434', '435', '436', '437', '438', '439', '440', '441', '442', '443', '444', '445', '446', '447', '448', '449', '450', '451', '452', '453', '454', '455', '456', '457', '458', '459', '460', '461', '462', '463', '464', '465', '466', '467', '468', '469', '470', '471', '472', '473', '474', '475', '476', '477', '478', '479', '480', '481', '482', '483', '484', '485', '486', '487', '488', '489', '490', '491', '492', '493', '494', '495', '496', '497', '498', '499', '500', '501', '502', '503', '504', '505', '506', '507', '508', '509', '510', '511', '512', '513', '514', '515', '516', '517', '518', '519', '520', '521', '522', '523', '524', '525', '526', '527', '528', '529', '530', '531', '532', '533', '534', '535', '536', '537', '538', '539', '540', '541', '542', '543', '544', '545', '546', '547', '548', '549', '550', '551', '552', '553', '554', '555', '556', '557', '558', '559', '560', '561', '562', '563', '564', '565', '566', '567', '568', '569', '570', '571', '572', '573', '574', '575', '576', '577', '578', '579', '580', '581', '582', '583', '584', '585', '586', '587', '588', '589', '590', '591', '592', '593', '594', '595', '596', '597', '598', '599', '600', '601', '602', '603', '604', '605', '606', '607', '608', '609', '610', '611', '612', '613', '614', '615', '616', '617', '618', '619', '620', '621', '622', '623', '624', '625', '626', '627', '628', '629', '630', '631', '632', '633', '634', '635', '636', '637', '638', '639', '640', '641', '642', '643', '644', '645', '646', '647', '648', '649', '650', '651', '652', '653', '654', '655', '656', '657', '658', '659', '660', '661', '662', '663', '664', '665', '666', '667', '668', '669', '670', '671', '672', '673', '674', '675', '676', '677', '678', '679', '680', '681', '682', '683', '684', '685', '686', '687', '688', '689', '690', '691', '692', '693', '694', '695', '696', '697', '698', '699', '700', '701', '702', '703', '704', '705', '706', '707', '708', '709', '710', '711', '712', '713', '714', '715', '716', '717', '718', '719', '720', '721', '722', '723', '724', '725', '726', '727', '728', '729', '730', '731', '732', '733', '734', '735', '736', '737', '738', '739', '740', '741', '742', '743', '744', '745', '746', '747', '748', '749', '750', '751', '752', '753', '754', '755', '756', '757', '758', '759', '760', '761', '762', '763', '764', '765', '766', '767', '768', '769', '770', '771', '772', '773', '774', '775', '776', '777', '778', '779', '780', '781', '782', '783', '784', '785', '786', '787', '788', '789', '790', '791', '792', '793', '794', '795', '796', '797', '798', '799', '800', '801', '802', '803', '804', '805', '806', '807', '808', '809', '810', '811', '812', '813', '814', '815', '816', '817', '818', '819', '820', '821', '822', '823', '824', '825', '826', '827', '828', '829', '830', '831', '832', '833', '834', '835', '836', '837', '838', '839', '840', '841', '842', '843', '844', '845', '846', '847', '848', '849', '850', '851', '852', '853', '854', '855', '856', '857', '858', '859', '860', '861', '862', '863', '864', '865', '866', '867', '868', '869', '870', '871', '872', '873', '874', '875', '876', '877', '878', '879', '880', '881', '882', '883', '884', '885', '886', '887', '888', '889', '890', '891', '892', '893', '894', '895', '896', '897', '898', '899', '900', '901', '902', '903', '904', '905', '906', '907', '908', '909', '910', '911', '912', '913', '914', '915', '916', '917', '918', '919', '920', '921', '922', '923', '924', '925', '926', '927', '928', '929', '930', '931', '932', '933', '934', '935', '936', '937', '938', '939', '940', '941', '942', '943', '944', '945', '946', '947', '948', '949', '950', '951', '952', '953', '954', '955', '956', '957', '958', '959', '960', '961', '962', '963', '964', '965', '966', '967', '968', '969', '970', '971', '972', '973', '974', '975', '976', '977', '978', '979', '980', '981', '982', '983', '984', '985', '986', '987', '988', '989', '990', '991', '992', '993', '994', '995', '996', '997', '998', '999', '1000')
/*Exclude accounts if their postings are included in this job hierarchy list: "Internships, students", "Education", "Qualified commercial staff, blue-collar workers", "Qualified Staff", "Miscellaneous"*/
--AND p.id_profession NOT IN (1, 3, 4, 9, 11)
--AND p.id_profession NOT IN (9, 11, 18, 19, 20, 21)
/*Exclude open opportunities and customers (obsolete)*/
--AND customer_sap.sap_account_id IS NULL
--AND sap_company.emp_resp_login IN ('00000', '00000', '00000')
--AND sap_company.business_line NOT IN ('000', '000')
--AND sap_company.business_not_allowed = ''
--AND sap_company.status IS 3
--AND sap_company.protection_end_date IS NULL OR sap_company.protection_end_date = '' OR sap_company.protection_end_date <= GETDATE()
--AND sap_company.contact_permission != 'Disallowed'
```

Side note: ETL and Scale of Data Management Initiative

- ETL process is applicable at any scale
 - Enterprise Scale: Enterprise Data Warehouse
 - Business Unit Scale: ETL to deliver specific data to another operational tool
 - Developer Level: individual development tasks
- Everyone does it, even if they don't realise it
- Some do it better than the others

Introduction to ETL with SAS®

ETL Key Subsystems and Components

Presented by Vasilij Nevlev

(E)xtract data subsystems

Prepare To Start:

- Business Requirements
- Logical Data Map
- Business Terms
- Naming conventions

Judge Data:

- Data Profiling
- You need to know what data attributes you are dealing with in advance

Change Data Capture

- In many cases you need to track and version changes.

Result: Extracted Table including Format Conversion

SAS Tools:

- Profiling tools in SAS Data Quality Studio
- Profiling tools in SAS Enterprise Miner
- Profiling tools in SAS Enterprise Guide
- Full support of different databases, such as Oracle, Terradata and Netezza
- Full range of transformations aimed at data extraction
- Full support out of the box SCD1 and SCD2
- Implicit and explicit in-database processing

(L)oad: Prepare for consumption

Time variance:

- SCD Management

Bridges:

- Bridge Tables
- Multi valued dimensions
- Special dimensions

Aggregations:

- Aggregate tables (datamarts)
- OLAP Cubes

Result: Datamarts, Fact and Dimension Tables ready for consumption

SAS Tools and Techniques:

- SCD Transformations in SAS DI Studio
- SAS Scalable Performance Data Server
- SAS OLAP Cubes powered by:
 - SAS OLAP Cube Server
 - SAS OLAP Cube Studio

(T)ransform: Clean and Conform

Cleaning Machinery:

- Cleansing and data quality

Cleaning Controls:

- Error event schemas
- Audit dimension

Integration:

- Deduplication and conforming systems

Keys:

- Surrogate Key Generator

Result: Cleaned Tables and Conformed Dimensions

SAS Tools:

- Look up in memory and as a data merge in SAS DI
- Complex data quality processing rules in SAS DQ
- Exception handling and reporting in SAS DI
- Data exception tracking in SAS DI
- Out of the box, no code required, Surrogate Key Generators

(M)anage – ETL/M?

Control:

- Scheduling

Protect:

- Backup
- Recovery/Restart

Control:

- Version/Control
- Migration

Other:

- Workflow monitoring
- Performance/Scalability
- Lineage and Dependency
- Problem Escalation
- Pipeline/Parallelize
- Security
- Compliance
- Metadata Repository

SAS Tools and Techniques:

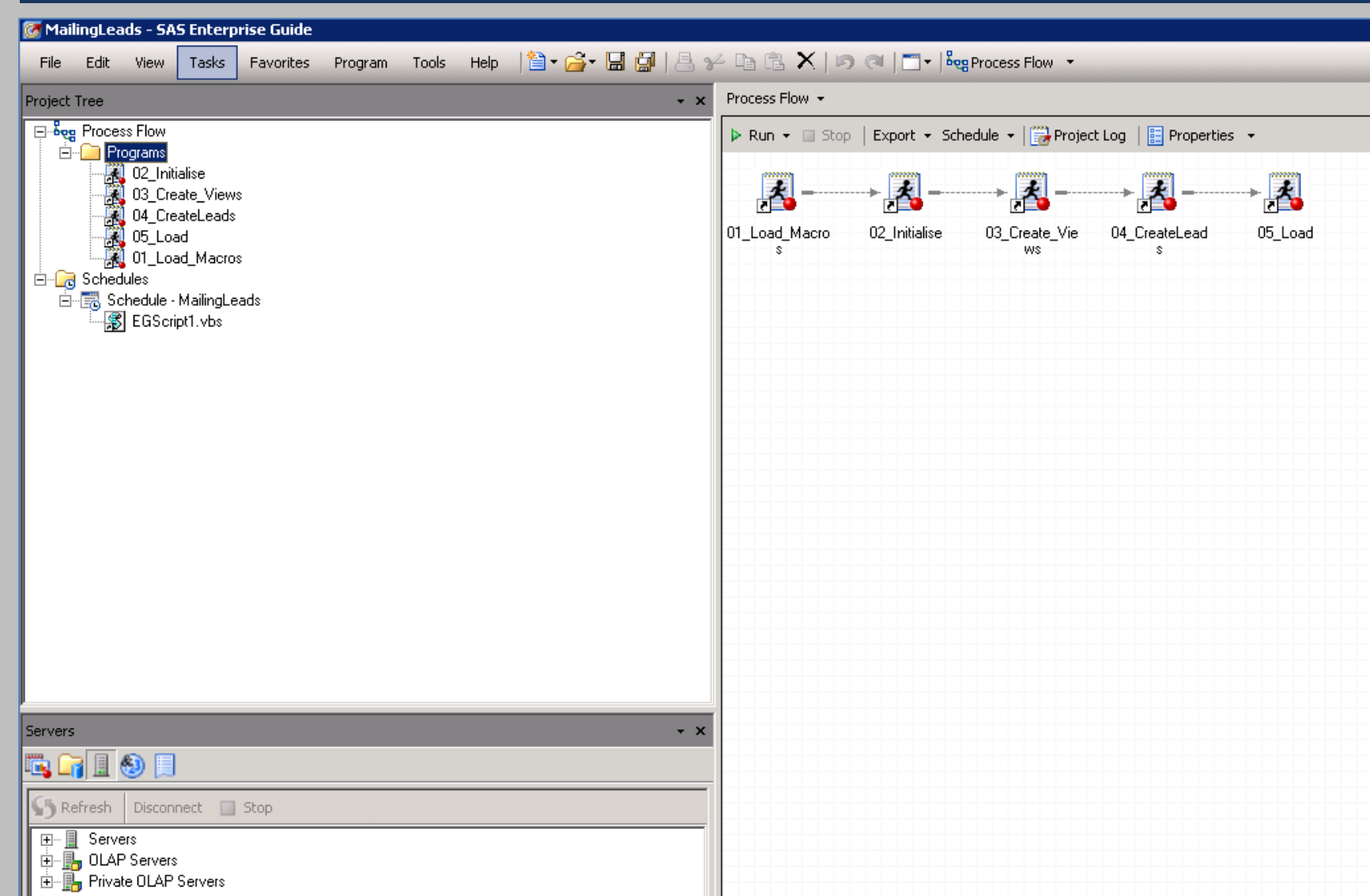
- SAS Metadata Server
- SAS Management Console
- Scheduling Plugin in SAS SMC
- Integration with IBM Suite
- Building In Process Scheduling
- Building OS Scheduling Integration
- Visual editing of flows in SAS SMC
- Visual editing of jobs in SAS DI Studio

Introduction to ETL with SAS®

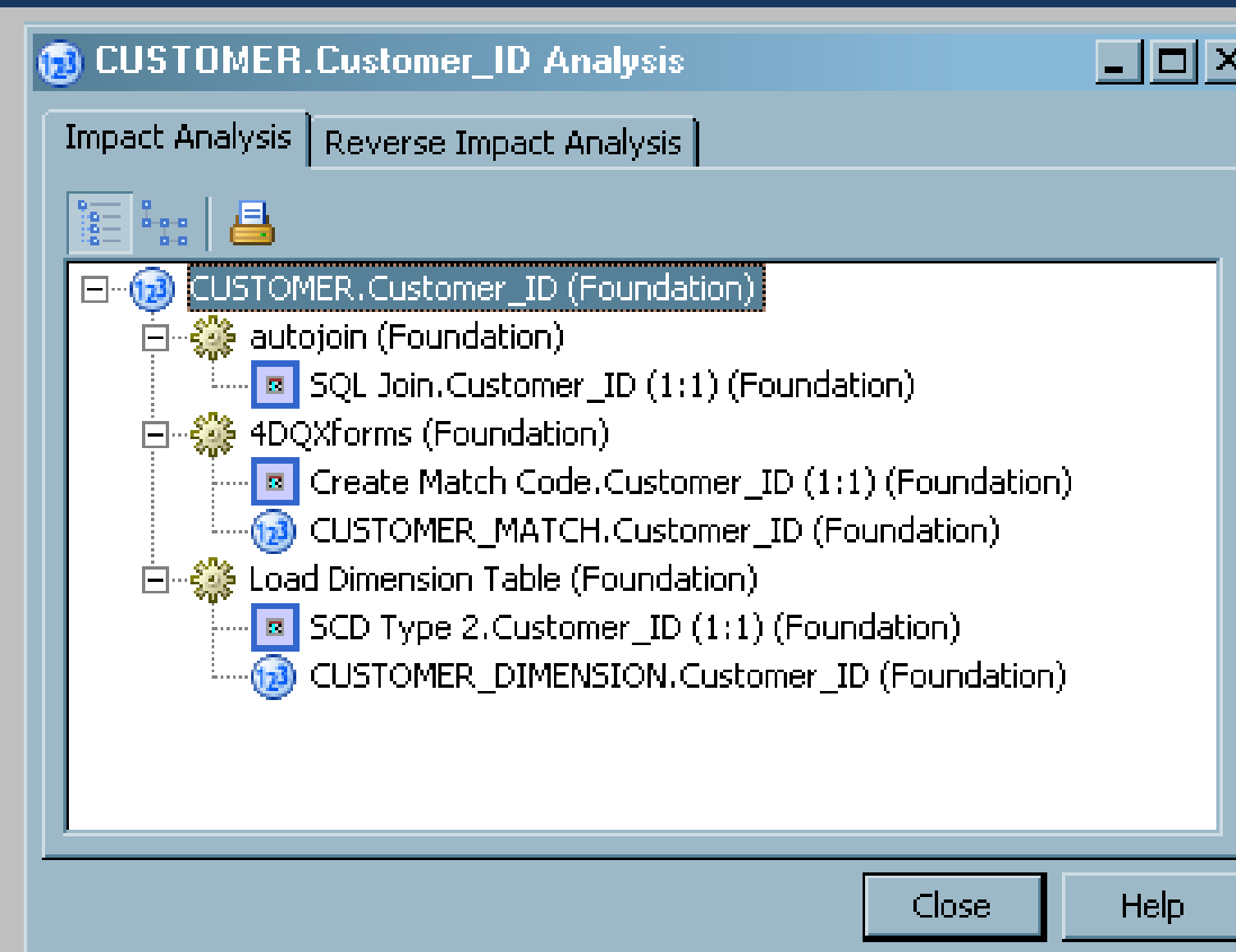
ETL Key Subsystems with SAS

Presented by Vasilij Nevlev

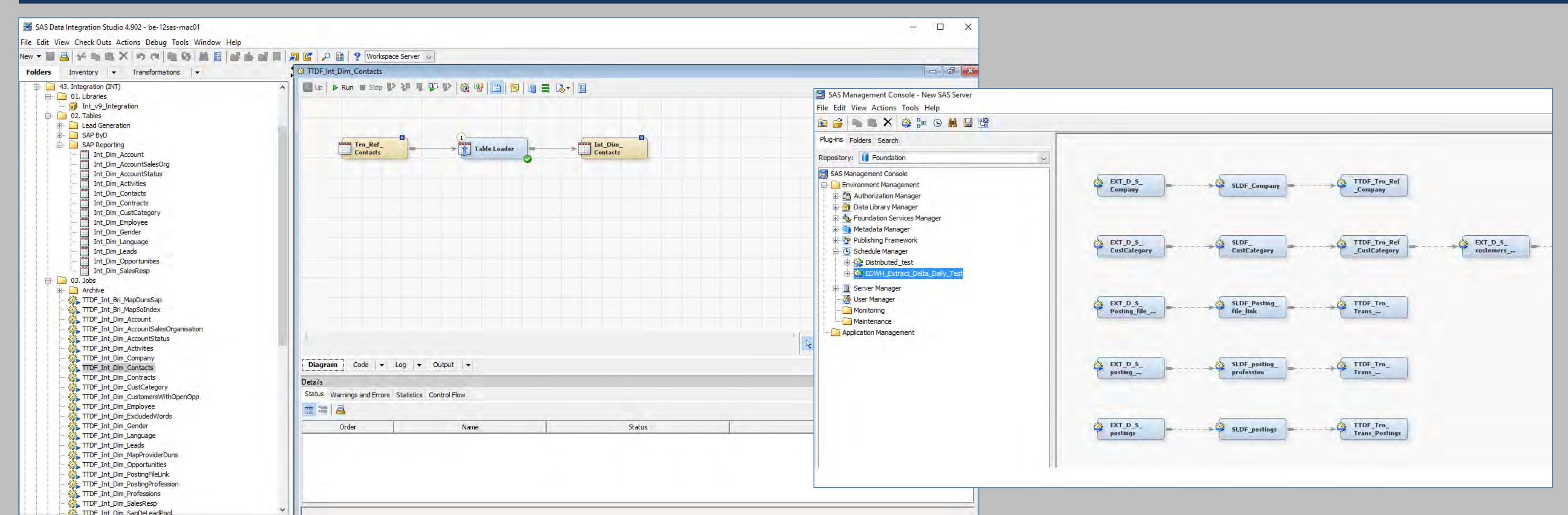
Simple ETL in SAS EG



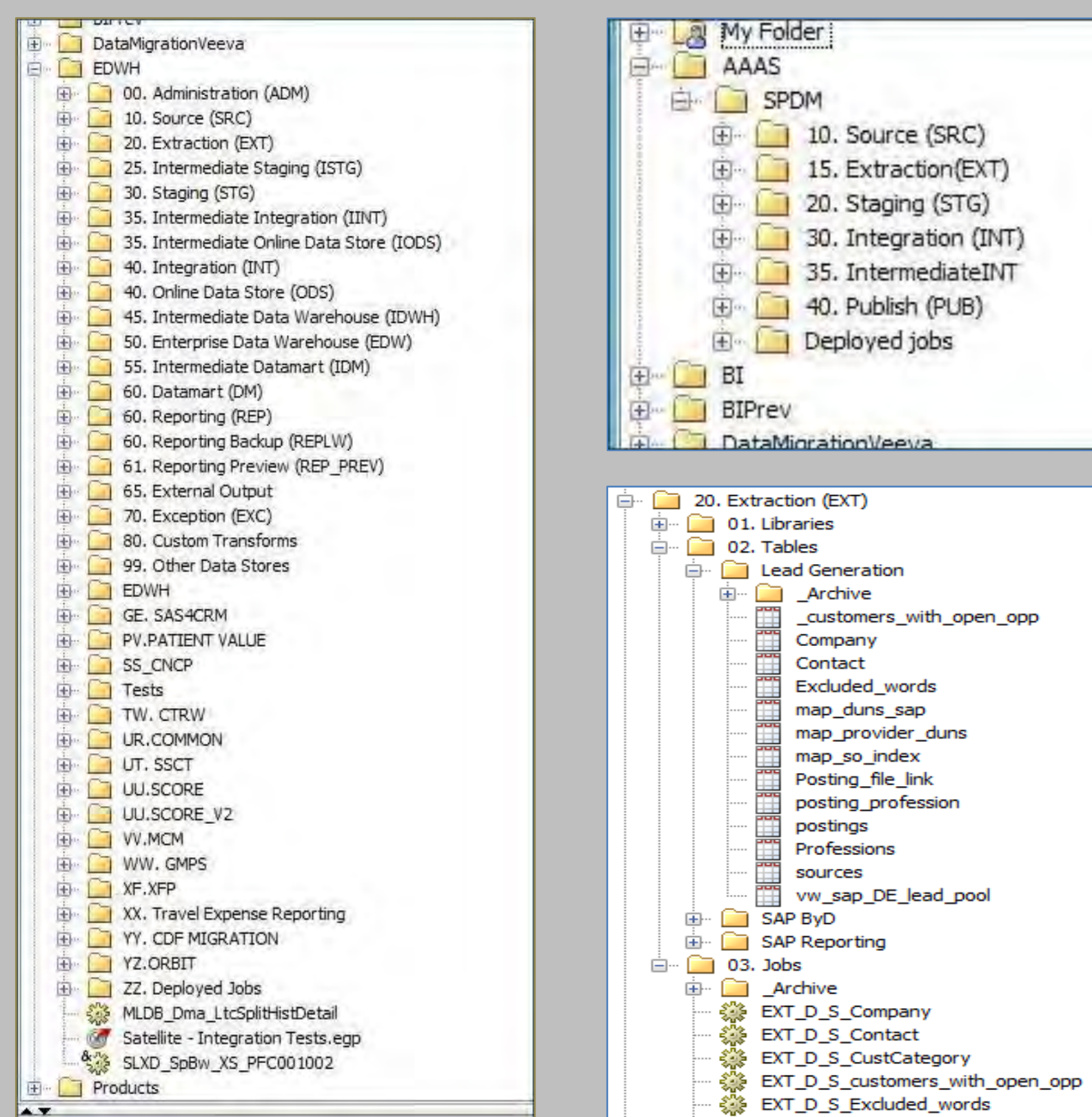
Data Lineage in SAS DI



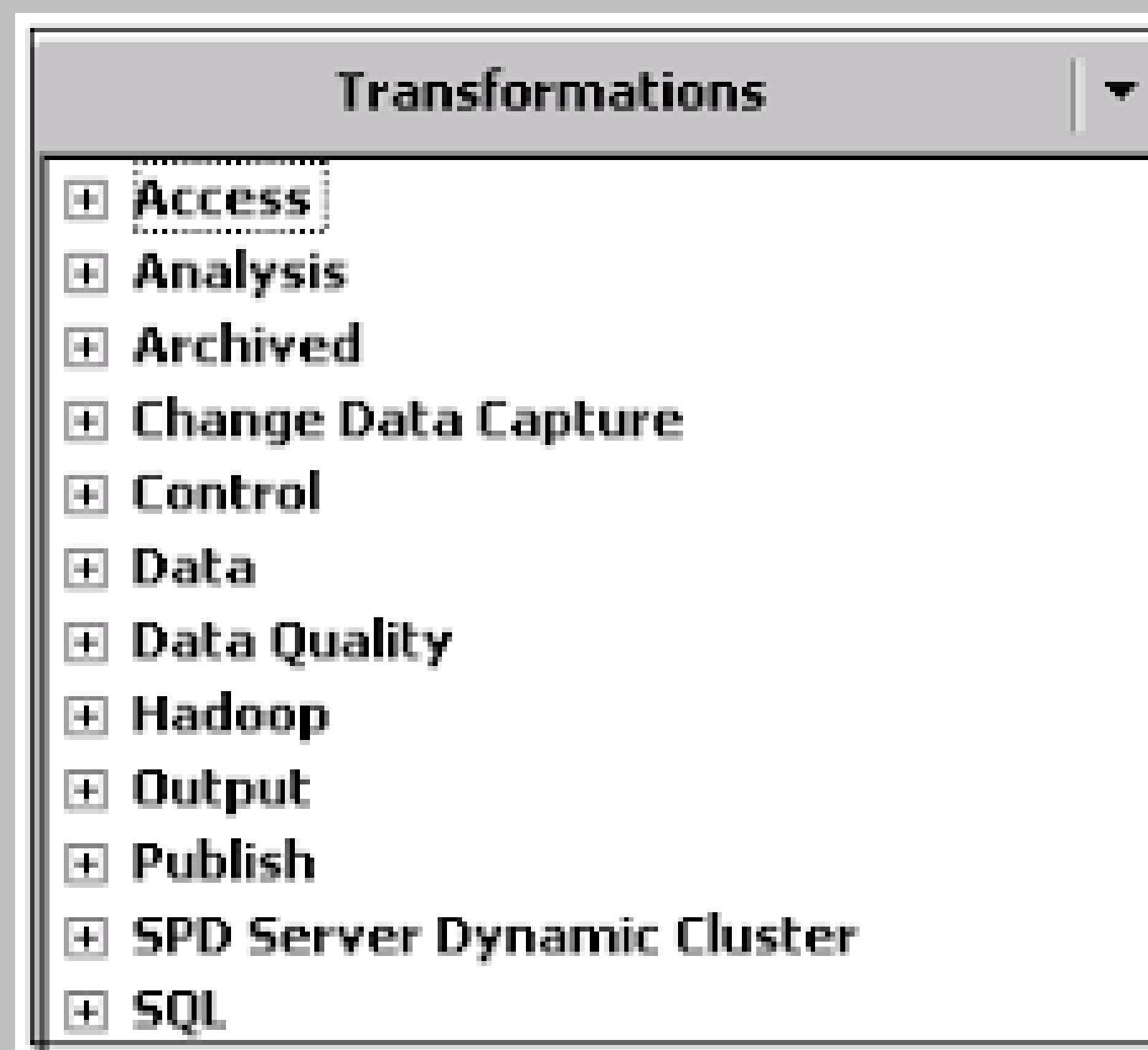
SAS DI Studio Job and SAS Management Console Flows



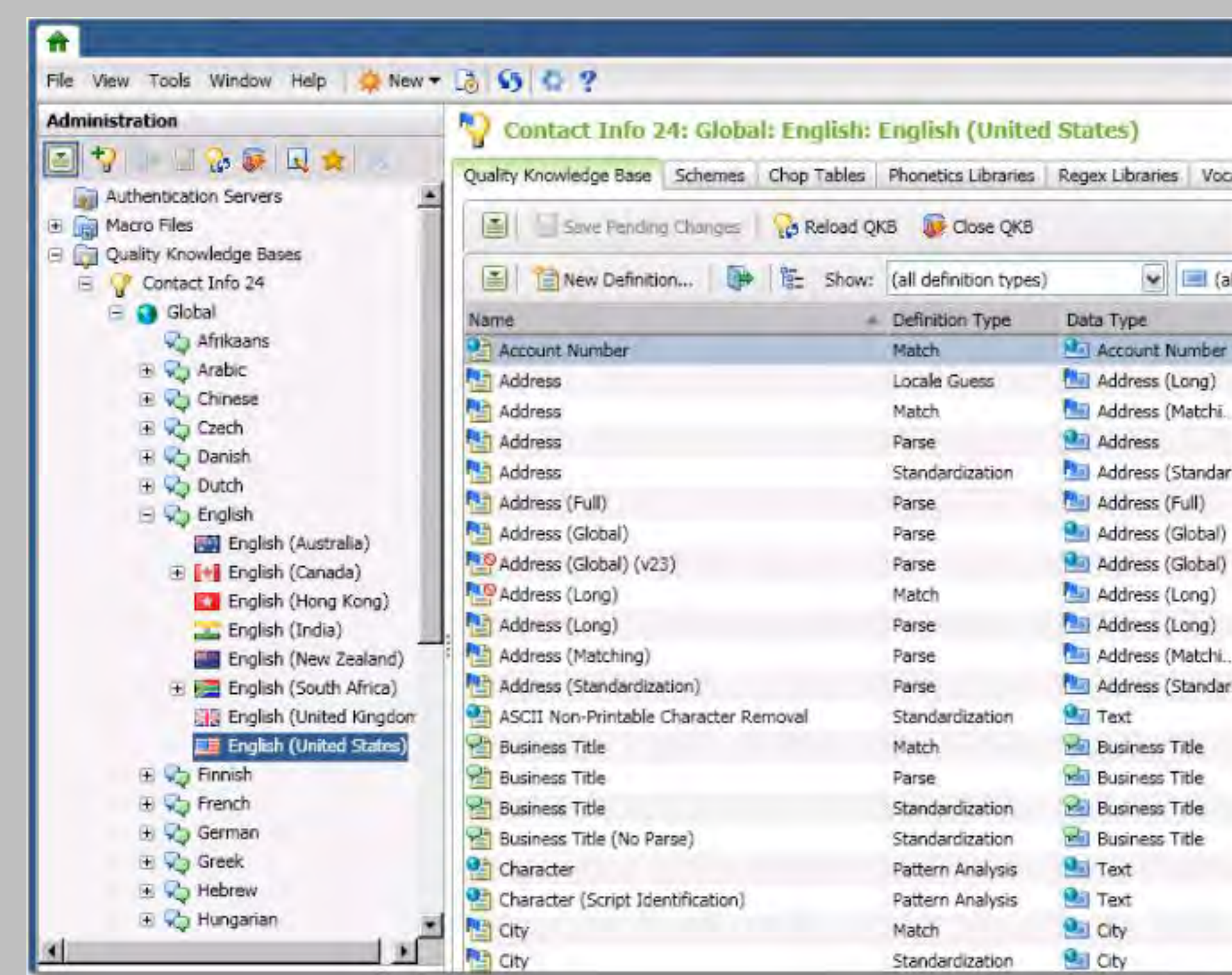
Metadata Organization in SAS



Data Transformations in SAS DI



Conclusion



SAS Offers all the tools necessary to implement all of the major ETL subsystems, from being able to profile the data sources, to organize metadata structures and orchestrate data flows according to business requirements.

Provides Knowledge Bases (shows to the left) that assesses data quality based on the region and locale. This can applied to addresses, names and lots of other areas.

SAS has the capability to almost every single ETL subsystem described earlier, so then the whole company ETL could be managed within a single ecosystem of tools and systems.



SAS[®] GLOBAL FORUM 2018

April 8 - 11 | Denver, CO
Colorado Convention Center

#SASGF