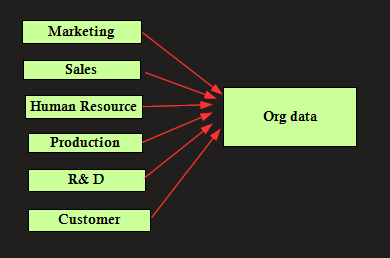
Matillion PART A

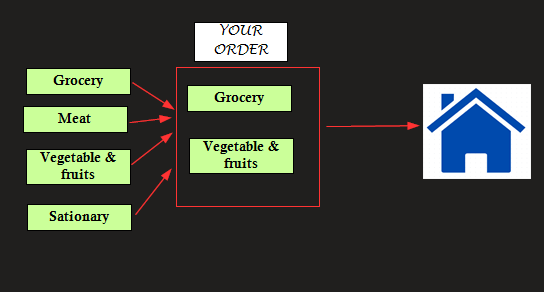
29 July 2024

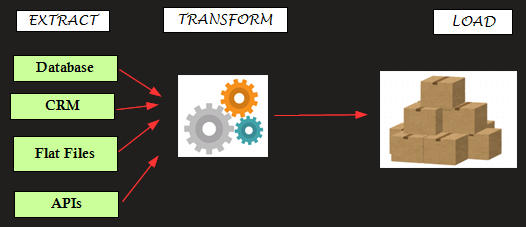
17:20

* + **Matillion YT:**
    - **Session order**

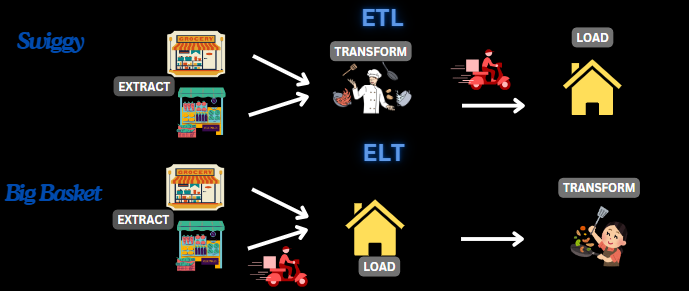
**V1**

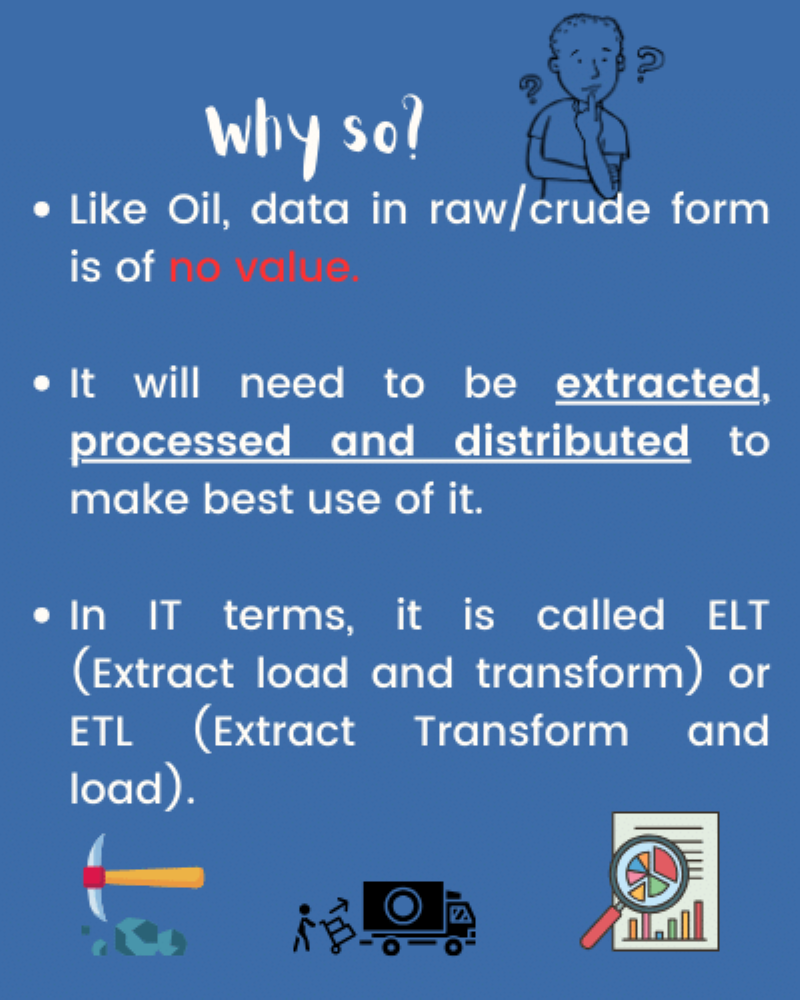
* + Official Links:
    - Website: <https://www.matillion.com/>
    - Documentation: <https://docs.matillion.com/>
    - Certification: <https://www.matillion.com/resources/certifications>
  + Intro to ETL :
    - Why data integration ?
      * We are generating lots of data everywhere. We can get valuable insights just by working on this indivisual fields. BUT we can get best out of data , if we can integrate this data into one place and analyse on it as a whole. That is where data integration comes into picture.
        + 
    - What is ETL? <https://www.matillion.com/blog/what-is-etl-the-ultimate-guide>
      * Extract , Transform , Load
      * everyday example: Swiggy or any online portal



* + In IT :
    - 
  + Have you heard of ELT?

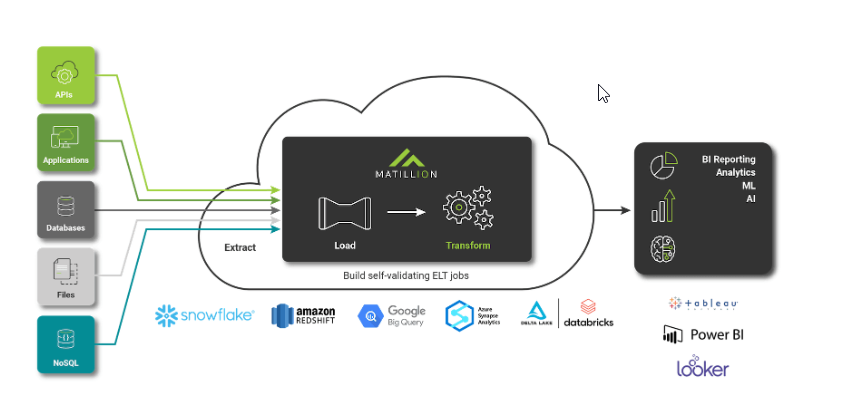
eg: buying furniture from Amazon

* + 
  + Is data the new oil? <https://www.linkedin.com/posts/renita-nazareth_why-cloud-why-matillion-activity-7151145113254031360-DYBg?utm_source=share&utm_medium=member_desktop>
    - [1\_matillion\_](https://www.canva.com/design/DAF5fxiEGak/6U3XTuJPm1vRlQLQKp3zsg/edit?utm_content=DAF5fxiEGak&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton)



**V2**

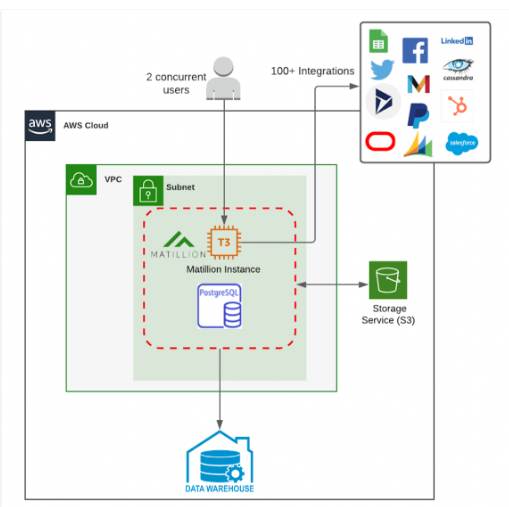
* + Introducing Matillion ETL
    - What Is Matillion and What Does It Offer?
      * Cloud based data integration and ETL tool that enables organizations to extract, load and transform data from various sources into data warehouses or cloud based data platforms.
      * What does it offer?
        + Matillion data loader
        + Matillion ETL
    - What is matillion ETL ? Features? <https://docs.matillion.com/metl/docs/1975061/>
    - Why Is Matillion a Good Choice to Consider?
      * Easy Integration with sources + easy learning curve (drag and drop) + on cloud (cost control) +data security
    - Case-study: How Can Matillion Modernize Your Business?
      * slack: <https://www.matillion.com/success-stories/slack>

* + Overview of Matillion ETL Architecture
    - 

Matillion sits on the various options like snowflake, amazon ----

extract using data loader components + load it and transform it

3 types : <https://docs.matillion.com/metl/docs/2950792/#clustered-diagram>

* + [The Single Node Architecture.](onenote:https://d.docs.live.net/c5a9c0cb524f4692/Documents/Work%20related/DA/ETL.one#1a%20Matillion&section-id={EC133B0F-98A8-4FEB-AB10-4E3CF82802E3}&page-id={DE2816D2-5146-41D7-8CBD-CF4608D2BA1A}&object-id={CB161702-56E0-4B8F-AC8A-660EEC84FE0F}&10)
    - 

* + [The Single Node & RDS Architecture.](onenote:https://d.docs.live.net/c5a9c0cb524f4692/Documents/Work%20related/DA/ETL.one#1a%20Matillion&section-id={EC133B0F-98A8-4FEB-AB10-4E3CF82802E3}&page-id={DE2816D2-5146-41D7-8CBD-CF4608D2BA1A}&object-id={CB161702-56E0-4B8F-AC8A-660EEC84FE0F}&14)
    - tCO+ 
      5-12 
      3 
      MATILLION ' 
      AWS 
      metad 
      a of 
      matilli 
      seperate_ so pts 
      lost 
      wont 

* + [The Clustered Architecture.](onenote:https://d.docs.live.net/c5a9c0cb524f4692/Documents/Work%20related/DA/ETL.one#1a%20Matillion&section-id={EC133B0F-98A8-4FEB-AB10-4E3CF82802E3}&page-id={DE2816D2-5146-41D7-8CBD-CF4608D2BA1A}&object-id={CB161702-56E0-4B8F-AC8A-660EEC84FE0F}&18)

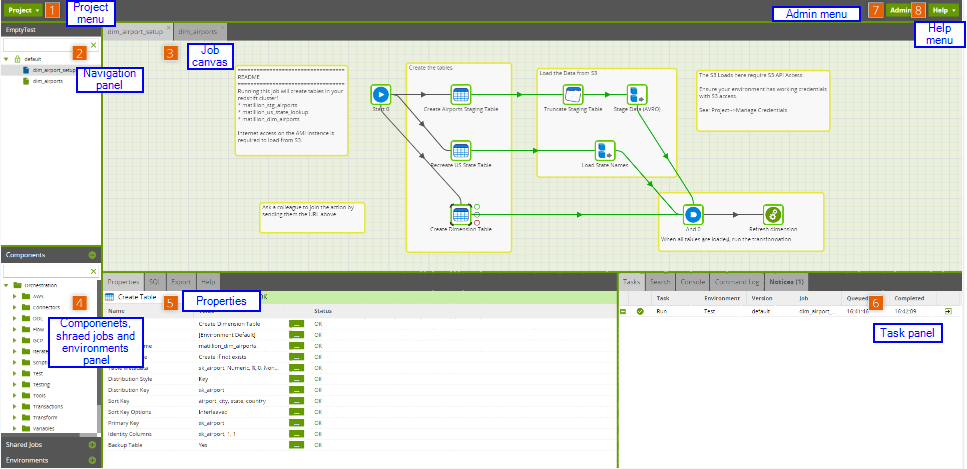
MATILLION , 
MATILLION 
AWS 
ROS 

**V3**

* + Practical - Get started with Matillion setup with AWS EC2 and snowflake
    - Launching Matillion ETL :
      * multiple ways :
        + <https://docs.matillion.com/metl/docs/launching-matillion-etl/>
        + AWS connect - discontinued
        + Matillion hub - business account
        + snowflake partner connect
      * We will use : with cloud formation template
    - Setup :
      * URL needed:
        + AWS free account <https://aws.amazon.com/free/>
        + Snowflake free tier account <https://signup.snowflake.com/>
        + Matillion hub <https://hub.matillion.com/>
      * [MASTER EXTRACT, TRANSFORM & LOAD WITH MATILLION](https://www.youtube.com/watch?v=z-JgDQPp-ro&list=PLekRk6SuZdQCQ_Uf-ojFLqUoyUPUMBbsk&index=3)
    - Accessing the Matillion ETL Client (Amazon EC2) : <https://docs.matillion.com/metl/docs/2957722/>
      * Cloud provider + Cloud data platform

V4

* + UI and basic functions: <https://docs.matillion.com/metl/docs/2694747/>
    - The majority of Matillion ETL features can be accessed using the following system features:



* + UI Intro - PART A:
    - Navigation panel
      * Each Matillion ETL project is made up of jobs. There is no limit to the number of jobs that may be contained within a project.
      * Double-clicking a job on the list will open it on the Job canvas. Alternatively, right-clicking anywhere within the panel will open a pop-up menu, which allows you to create, delete, import, and export jobs, as well as creating job folders.
      * There are two kinds of job:
        + Orchestration job: Generally loads and unloads data to tables and file storage, also known as staging the data.
        + Transformation job: Generally manipulates data in the form of tables.
    - Job canvas
      * The job canvas shows the contents of a selected job.
      * components are linked to one another in the order in which they will need to be executed. This can be done by clicking on a component icon on the job canvas to reveal the connection rings around the component. Clicking and dragging on a ring will extend a connection that can be affixed to another component.
    - Jobs: <https://docs.matillion.com/metl/docs/8347111/#overview>
      * Types in matillion:
        + orchestration Jobs handles workflow management and controls the flow. These are used for loading data from source systems into data warehouses and other logical tasks.
        + Transformation Jobs are used for data transformations and integrations
      * Components: eg: assembling a computer
        + A component is a building block that does a single task, such as loading data from Salesforce, running a Python script, performing an SQL Join or even interacting with GenAI.
        + Nodes in components:

Green: Continue if the component successfully runs

Grey: Continue flow in any event

Red: Continue if the component fails to run

Blue: Continue on to another component (regarding iterator components).

* + - Components panel:
      * This is a collapsible list of all components available to the currently selected job type.
      * Switching between an orchestration job and a transformation job will automatically update the available components
      * Each component embodies a particular function
    - Properties panel
      * Clicking a component icon on the job canvas reveals detailed information about the component in the Properties panel below.
      * Common tabs seen:
        + Properties: A unique set of user-defined properties that effect the function of the selected component—changes made to these properties will not affect the properties of any other component, even components of the same type. Read the specific component's documentation for more details.
        + Export: Export certain values from the selected component as a variable to be used elsewhere.

Variable exports :The Export tab of the Properties panel allows you to edit mappings between the runtime information that the component makes available and the variables you have already defined.

To export a value, it must be mapped to an existing variable through the component's Export tab. <https://docs.matillion.com/metl/docs/2908982/#grid-variable-export>

* + - * + Sample: Many components allow you to produce a sample output of data up to a specified number of rows—this can be used to confirm a component is set up correctly before using it in a live job.
        + Metadata: Display data for each column featured in the sample data, including column names, types, sizes, and scales. This tab is only available in transformation components.
        + Lineage: Track columns across transformations to understand how that column has been affected by components up to the point of inspection. This tab is only available in Enterprise Mode.
        + SQL: Display the SQL equivalent of the component and its current setup.
        + Help: Display a snippet of the help documentation for the selected component

Practical :

* + Create Project folder
  + Create Orchestration job
  + Bring in S3 component

V5

Practical : LOADING DATA FROM S3 TO SNOWFLAKE FROM MATILLION

* + S3 bucket and load data in S3
    - NOTE: Data created by Chatgpt.
    - Github file link
  + Create table in snowflake
    - Script :

CREATE TABLE StudentRecords (

RecordID INT PRIMARY KEY , -- Unique identifier for each record

StudentName VARCHAR(100) NOT NULL, -- Name of the student

Subject VARCHAR(100) NOT NULL, -- Subject of the exam

Marks INT NOT NULL, -- Marks obtained by the student

Year INT NOT NULL -- Year in which the marks were obtained

);

* + Create stage pointing to S3
  + Configure S3 component
  + Explore validate and execution
  + UI - PART B
    - Environments panel
      * In Matillion ETL, an environment describes a single connection to a database, on which jobs can be run. Numerous environments can be set up at once, meaning different users can use different environments on a single Matillion ETL instance.
    - Tasks panel
      * The Tasks tabs show information about tasks that have been run during the current session.
      * Tabs include:
        + Tasks: A list of up to the last 20 tasks attempted by the current user in the current project session—including queued, failed, or cancelled tasks. Read Task management for more details.
        + Search: Allows you to search the current project for matching components.
        + Console: The Matillion ETL instance console.
        + Command Log: An archive of interactions between the current browser session and the Matillion ETL server.
        + Notices: A list of notifications concerning the Matillion ETL project—including updates to Matillion ETL.
      * Task history:
        + 3 ways:

project>task history

>> job: Task history

Task tab

V6

* + DDL in matillion <https://docs.matillion.com/metl/docs/2043153/>

Practical : CREATE , DELETE , ALTER TABLE IN SNOWFLAKE FROM MATILLION

* + Common DDL
    - Create table :<https://docs.matillion.com/metl/docs/2043153/>
      * Where are we creating table? What is the target
      * Properties:
        + Create /replace -

Create - use this option if you know well in advance which tables are already there. There will be hundreds of table. I won't recommend for newbies. What will happen if there is a table - it will give an error.

Create if not exist - safest option

Replace - Very risky as you would be replacing a table. So, if you have created the table and if you know the impact , you can use it. You created but now you wanted to change the structure.

* + - * + Name - Mytable
        + Table type: <https://docs.matillion.com/metl/docs/2043153/#properties>
        + Column: - data type - what kind of value this columns as holding <https://docs.matillion.com/metl/docs/2043153/#properties>



* + - * + Collation:

A **collation** is a set of **rules** for comparing strings that belong to a character set. The impact is evident when ordering strings but not so important for storing data.

* + Delete table
  + SQL script :<https://docs.matillion.com/metl/docs/2107012/>
    - No select statements. Can use INSERT , CREATE , DELETE
  + TRUNCATE TABLE :<https://docs.matillion.com/metl/docs/2103743/>
  + Create EXTERNAL TABLE :<https://docs.matillion.com/metl/docs/2972813/#properties>

Practical :SQL SCRIPT COMPONENT

* + replace create table with SQL SCRIPT component

V7

* + variables - PART1: Types of variables? format? Order of preference? settings - Variable behaviour? data type? use case ?
    - Variables Matillion supports: Environment Variables, Automatic Variables, Job Variables, and Grid Variables.
    - Analogy:
      * Global variable and local variables
      * School Principal- env, Class Teacher - Job variable
    - referenced with the syntax: ${<variable name>}.
    - Order of preference if same name: JV > EV > AV
    - 4 types:
      * job variable - within the scope of a single job <https://docs.matillion.com/metl/docs/2943425/>
        + Will override environment variables of the same name within that specific job.
        + job variables are always included in jobs that are imported or exported and are not available for optional inclusion like environment variable
        + Setting job variable:

ways in which you can update a variable:

Iterator components: sets variable to a new value for each iteration

we need to define variables in advance and use them when configuring iterator components

when run in parallel, variable scope should be "copied" to ensure each parallel component sees its own copied copy of the variable

Python scripts -using their built-in "context" object

locations where edit job variables is seen:

existing any job -> clicking a job,>Variables>clicking Manage Job Variables

anywhere on canvas -> canvas>Manage Job Variables

Job Variables Menu

* + - * environment variable: <https://docs.matillion.com/metl/docs/2943424/>
        + environment variables can be used across Matillion ETL, in configurations and in all jobs.
        + Creating environment variables:

project > manage environment variables

* + - * + details:



Name: The name of the environment variable.

Type: Text, Numeric, or DateTime. Text is selected by default.

Behaviour: Shared or Copied. Shared is selected by default.

Default Environment Value: The default value of the variable in each environment.

Description: An environment variable can optionally be given a description

* + - * + When a job begins, all variables are initialized with their default environment value
        + Setting the variable can be done through;

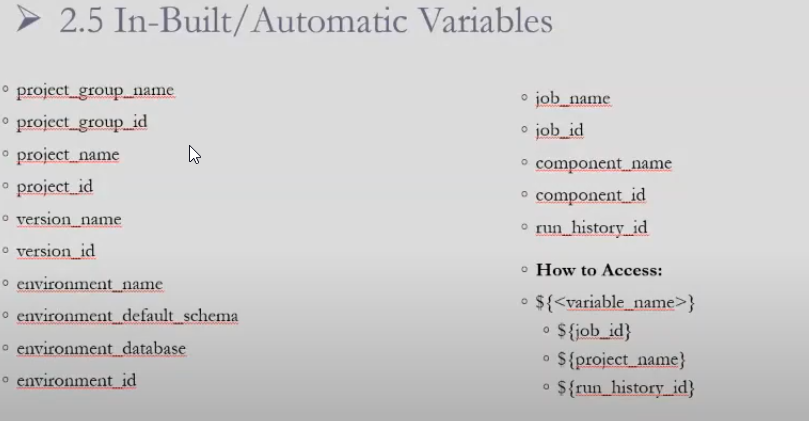
Iteration components

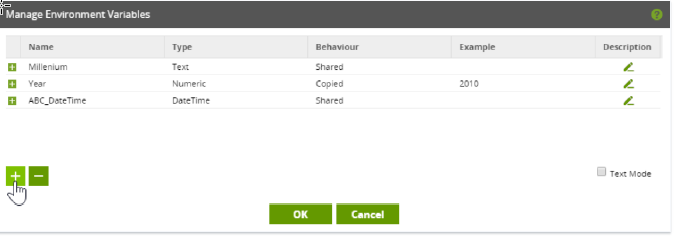
Python scripts

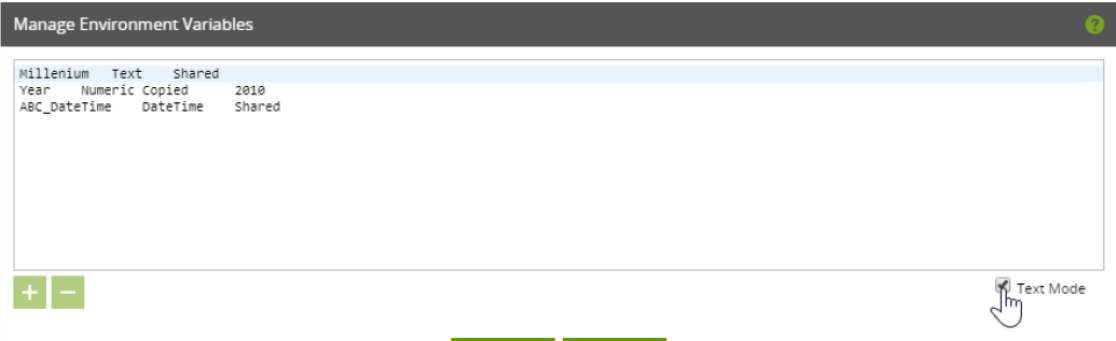
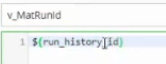
context.updateVariable("variable", "new value")

SQS runs :Job executions, triggered by SQS messages, can also set variable values using the optional 'variables' item when sending the message.

* + - * automatic variables (Environment Variables) : . <https://docs.matillion.com/metl/docs/2943424/#automatic-variables>
        + automatically available ,Already created in the Matillion account by default
        + list of vaiables: <https://docs.matillion.com/metl/docs/2943424/>



* + grid variables :is like 2D arrays . (to be covered later) [Grid variable](onenote:#1%20Matillion&section-id={C60757F1-15B3-4A3E-B1DF-BABFF853DB95}&page-id={BE0DEB51-5C91-4B5F-83C4-2668E1D93110}&object-id={C0B1FED6-AE35-437A-81C6-303D56DFD785}&80&base-path=https://d.docs.live.net/5682879ac503adc5/文件/DATALearn/Topics.one)
  + settings:
    - 

* + 2 kinds of variable behaviour:
    - Copied = local:
      * Copied variables can be updated within one branch of a job without updating variables of the same name in other branches.
      * A branch using a copied variable will begin by taking its default value, even if this variable has been updated in another, parallel branch.
    - shared = global ----------> "caring is sharing"
      * Shared variables are updated in a job-wide fashion. As such, regardless of branches, if a Shared variable is updated in one branch, all other branches will use that updated value.
  + Datatypes:
    - 
  + Text Mode:
    - On many menus, such as Manage Job Variables and Manage Environment Variables, users can switch to a Text Mode by ticking the checkbox next to Text Mode.
    - Switching to Text Mode allows users to:
      * • Edit variables as a set of space-delimited text entries.
      * • Convert all existing variable data into an appropriate format.
      * • **Copy and paste text, offering a fast way to move large numbers of variables**.
      * • Have certain types of missing data auto-completed by Matillion ETL when parsing.
      * Error check data before adding it to the project (Matillion ETL parses the data on exiting Text Mode, throwing a message if an error has been made).
    - 
  + When to use:
    - env - when used in multiple jobs - like database names , s3 bucket name
    - automatic :
      * in messaging components
      * in CDC track :
        + 
    - grid : when iterating dynamically , working will passing more than 1 value in a variable
    - all the rest of the cases - JV

V8

Practical : ENVIRONMENT VARIABLES - ACCESS BD, SCHEMA , TABLE NAME USING VARIABLES

* + Create env for Dev , stage, prod
  + create env variable for database, schema, table name and update in properties and scripts

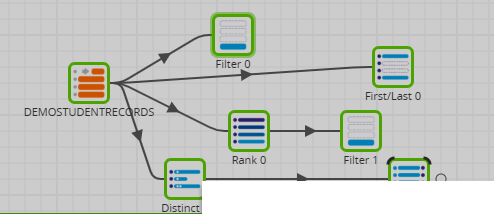
V9

Practical : Copied and shared

V10

Practical : USE TRANFORMATION COMPONENETS

* + Transformation:
    - Table input + Filter +Rank +distinct



* + Push down concept: processing is done in target cloud DW [Push down concept:](onenote:#1a%20Matillion&section-id={EC133B0F-98A8-4FEB-AB10-4E3CF82802E3}&page-id={DE2816D2-5146-41D7-8CBD-CF4608D2BA1A}&object-id={F61B45D4-6AA7-4A85-A45C-E81F5E070AA6}&4F&base-path=https://d.docs.live.net/c5a9c0cb524f4692/Documents/Work%20related/DA/ETL.one) 
    - verify in snowflake