**re: target dataset**

n general, a target consists of the set of information required to define the expected data in a dataset. Often referred to as a "schema," this target schema information can include:

* Names of columns
* Order of columns
* Column data types
* Data type format

A dataset associated with a target is expected to conform to the requirements of the schema. Where there are differences between target schema and dataset schema, a validation indicator is displayed.

**KNIGHT TRAINING --------------------------------------------------------------------------------------------------------------**

Acme Corp. receives inventory data from the local retail store on a weekly basis. This data contains information for 1000s of Acme Corp. items being sold by the retailer. As part of the inventory analysis, Acme Corp. often calculates rolling sums of key metrics such as inventory quantity, sales, shipments, etc. across monthly, quarterly, semi-annually and annual basis. The end result will be used in downstream analytics. Why calculate the rolling sum? Acme Corp. needs to know the trends over time for any specific time period.

**DEMO -------------------------------------------**

The Northwind company has asked us to help them wrangle their raw retail data for the past two years (yyyy and yyyy). They have requested the following wrangled data to be included in the deliverables:

* A cleaned, normalized, OTHER ADJECTIVES version of each raw data table, which can then be used in subsequent exploratory analyses by the home team
* Refined data tables that include inferred statistic attributes to help answer the following questions:
  + How much did each Northwind customer purchase during each quarter of yyyy and yyyy?
  + What is the % customer size of each customer for each year represented in the raw data sets?
  + What is the product velocity for each product represented in the raw data sets?

**================================**

**DEMO**

Show how histograms can be effectively used in Trifacta to quickly gain insight to a data set.

Ditto with column history.

Ditto with the profile sheet

Identify two or three selling points re what Trifacta does well

Rem to use comments to document your recipes

the histogram does...

the quality bar does...

the ribbon does...

**================================**

**Prerequisites**

NorthWind is a company that...

**Story**

**Steps**

**Raw Stage**

* Profile
  + When we first load a dataset into the app, its usually a good idea to profile it so that we can get a quick look at what might be wrong with the data as well as to provide a reference later on if we want to see how our data has changed during the wrangling process.
* Clean the data
  + fix data types per schema
    - Trifacta allows you to align your raw data to a target file. . .
    - use target schema
      * a target is the set of columns, their order, and their formats to which you are attempting to wrangle your dataset. his target can be defined through imported or created datasets and must be assigned to an existing recipe.  After it is assigned to a recipe, a target appears in the Transformer page to assist in your wrangling efforts.
* Delete unnecessary columns
  + getting rid of unnecessary columns makes it easier to work with data in the application, and it also improves job execution performance.
* Manage missing data
  + there needs to be a region associated with each order placed
  + for records with NULL regions, use lookup table -- **CREATE REGION LOOKUP**
    - if USA, use State
    - if not USA, use CITY ABBR
    - (note that Trifacta has a REST API for its Enterprise product)

**Refined Stage**

**Production Stage**

* Methodology
  + Raw Stage
  + Refined Stage
  + Production Stage
* Demo Breakdown
* Cool Trifacta Stuff
  + Profiles
    - rem: we do a Profile step on each input bc it provides a visual comparison to indicate how the file has changed during the wrangling process and also bc it provides a buffer for picking up any transformations that we may need to make in a new pipeline input in the event that such transformation are not already represented in an existing recipe within the flow
  + histograms
  + easily hide and show columns
  + easily spot dupes, missing values and mismatches
  + easily change data types
  + Reference Files
  + Target Files
  + easy joins

For this particular demo:

|  |  |  |
| --- | --- | --- |
| RAW Stage | REFINED Stage | PRODUCTION Stage |
| * Data Set * Missing Columns * Empty Columns * Null Values * Schema Comparisons * Data Type Mismatches * Unique Values vs. Duplicates * Normalizing | * Calculations * Statistics * Inferred Attributes * Data Joins and Unions * Other enrichments supporting downstream ad hoc reports | * Polished data for automated downstream reporting * Value adding reports |
|  |  |  |