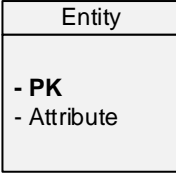
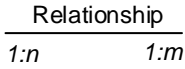
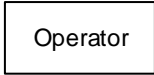
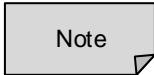
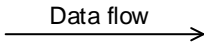
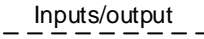


# GR4ML – Data Preparation View

Element	Definition & Symbol:	Questions to ask to identify them*:
<b>Entity</b>	<p>They represent data tables along with their attributes (i.e., fields). They represent the raw data (input to the data preparation flows) as well as the prepared datasets (output of the data preparation flows).</p> 	<ul style="list-style-type: none"> <li>• What kind of data would be relevant for generating the insights and answering the business question at hand?</li> <li>• What data attributes (i.e., features), in what format, and aggregation level are needed for the question goals under consideration?</li> <li>• Explain, to best of your understanding, the attributes, format, and size of the dataset at hand.</li> </ul>
<b>Relationship</b>	<p>They represent conceptual relationship and cardinalities among entities (i.e., data tables).</p> 	<ul style="list-style-type: none"> <li>• Where is the data stored, and what is data schema (i.e., entities and relationships)?</li> </ul>
<b>Operator</b>	<p>They represent an atomic activity that performs (part of) a data preparation task.</p> 	<ul style="list-style-type: none"> <li>• For each attributes, what is the data types, aggregation level, and selection of records (filtering)?</li> <li>• What (sequence of) integration, cleaning, aggregation, filtering and other data preparations are needed for transforming the raw data tables into the prepared data tables?</li> </ul>
<b>Note</b>	<p>Modelers can use the Note elements to attach clarifications and details to each Operation element in the model.</p> 	<ul style="list-style-type: none"> <li>• Are there any sample codes from the data engineering team?</li> <li>• Are there any pseud-code available for data transformation steps?</li> </ul>
<b>Data Flow</b>	<p>Operators are linked by Data Flows to represent the sequence and dependency.</p> 	<ul style="list-style-type: none"> <li>• What order of integration, cleaning, aggregation, and filtering are in place to transform the raw data tables into the prepared data tables?</li> </ul>
<b>Input/output</b>	<p>They connect entities to Operators.</p> 	<ul style="list-style-type: none"> <li>• Where is the raw data coming from?</li> <li>• Where is the final prepared dataset is going to be stored?</li> </ul>

\* These are sample questions and one may extend, modify, or customize them depending on the use case and context. Also they are not sorted in any specific order.