29/08/2025, 18:29 Filter - docetl docs

## Filter Operation

The Filter operation in DocETL is used to selectively process data items based on specific conditions. It behaves similarly to the Map operation, but with a key difference: items that evaluate to false are filtered out of the dataset, allowing you to include or exclude data points from further processing in your pipeline.

#### Motivation

Filtering is crucial when you need to:

- Focus on the most relevant data points
- Remove noise or irrelevant information from your dataset
- · Create subsets of data for specialized analysis
- · Optimize downstream processing by reducing data volume

# Example: Filtering High-Impact News Articles

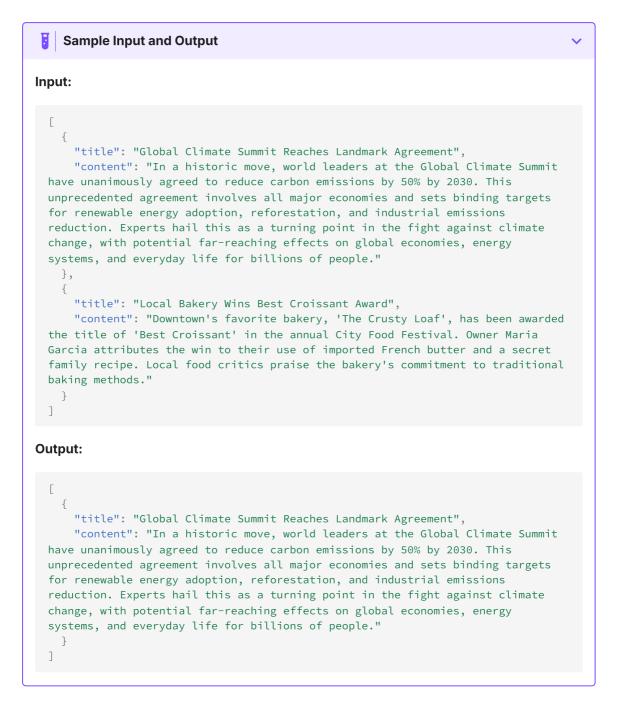
Let's look at a practical example of using the Filter operation to identify high-impact news articles based on certain criteria.

```
- name: filter_high_impact_articles
 type: filter
 prompt: |
   Analyze the following news article:
   Title: "{{ input.title }}"
   Content: "{{ input.content }}"
   Determine if this article is high-impact based on the following criteria:
   1. Covers a significant global or national event
   2. Has potential long-term consequences
   3. Affects a large number of people
   4. Is from a reputable source
    Respond with 'true' if the article meets at least 3 of these criteria,
otherwise respond with 'false'.
 output:
     is_high_impact: boolean
  model: gpt-4-turbo
```

29/08/2025, 18:29 Filter - docetl docs

```
validate:
    - isinstance(output["is_high_impact"], bool)
```

This Filter operation processes news articles and determines whether they are "high-impact" based on specific criteria. Unlike a Map operation, which would process all articles and add an "is\_high\_impact" field to each, this Filter operation will only pass through articles that meet the criteria, effectively removing low-impact articles from the dataset.



This example demonstrates how the Filter operation distinguishes between high-impact news articles and those of more local or limited significance. The climate summit article is retained in the dataset due to its global significance, long-term consequences, and

29/08/2025, 18:29 Filter - docetl docs

wide-ranging effects. The local bakery story, while interesting, doesn't meet the criteria for a high-impact article and is filtered out of the dataset.

## Configuration

#### **Required Parameters**

- name: A unique name for the operation.
- type: Must be set to "filter".
- prompt: The prompt template to use for the filtering condition. Access input variables with input.keyname.
- output: Schema definition for the output from the LLM. It must include only one field, a boolean field.

#### **Optional Parameters**

See map optional parameters for additional configuration options, including batch\_prompt and max\_batch\_size.



#### **Validation**

For more details on validation techniques and implementation, see operators.

### **Best Practices**

- 1. Clear Criteria: Define clear and specific criteria for filtering in your prompt.
- 2. **Boolean Output**: Ensure your prompt guides the LLM to produce a clear boolean output.
- 3. **Data Flow Awareness**: Remember that unlike Map, Filter will reduce the size of your dataset. Ensure this aligns with your pipeline's objectives.