Cluster operation

The Cluster operation in DocETL groups all items into a binary tree using [agglomerative clustering](#) of the embedding of some keys, and annotates each item with the path through this tree down to the item (Note that the path is reversed, starting with the most specific grouping, and ending in the root of the tree, the cluster that encompasses all your input).

Each cluster is summarized using an llm prompt, taking the summaries of its children as inputs (or for the leaf nodes, the actual items).

🚀 Example: Grouping concepts from a knowledge-graph

```yaml
- name: cluster_concepts
  type: cluster
  max_batch_size: 5
  embedding_keys:
    - concept
    - description
  output_key: categories # This is optional, and defaults to "clusters"
  summary_schema:
    concept: str
    description: str
  summary_prompt: |
    The following describes two related concepts. What concept
    encompasses both? Try not to be too broad; it might be that one of
    these two concepts already encompasses the other; in that case,
    you should just use that concept.

    {% for input in inputs %}
    {{input.concept}}:
    {{input.description}}
    {% endfor %}

    Provide the title of the super-concept, and a description.
```

This cluster operation processes a set of concepts, each with a title and a description, and groups them into a tree of categories.

Sample Input and Output

Input:

```
[
  {
    "concept": "Shed",
    "description": "A shed is typically a simple, single-story roofed
structure, often used for storage, for hobbies, or as a workshop, and typically
serving as outbuilding, such as in a back garden or on an allotment. Sheds vary
considerably in their size and complexity of construction, from simple open-
sided ones designed to cover bicycles or garden items to large wood-framed
structures with shingled roofs, windows, and electrical outlets. Sheds used on
farms or in the industry can be large structures. The main types of shed
construction are metal sheathing over a metal frame, plastic sheathing and
frame, all-wood construction (the roof may be asphalt shingled or sheathed in
tin), and vinyl-sided sheds built over a wooden frame. Small sheds may include
a wooden or plastic floor, while more permanent ones may be built on a concrete
pad or foundation. Sheds may be lockable to deter theft or entry by children,
domestic animals, wildlife, etc."
  },
  {
    "concept": "Barn",
    "description": "A barn is an agricultural building usually on farms and
used for various purposes. In North America, a barn refers to structures that
house livestock, including cattle and horses, as well as equipment and fodder,
and often grain.[2] As a result, the term barn is often qualified e.g. tobacco
barn, dairy barn, cow house, sheep barn, potato barn. In the British Isles, the
term barn is restricted mainly to storage structures for unthreshed cereals and
fodder, the terms byre or shippon being applied to cow shelters, whereas horses
are kept in buildings known as stables.[2][3] In mainland Europe, however,
barns were often part of integrated structures known as byre-dwellings (or
housebarns in US literature). In addition, barns may be used for equipment
storage, as a covered workplace, and for activities such as threshing."
  },
  {
    "concept": "Tree house",
    "description": "A tree house, tree fort or treeshed, is a platform or
building constructed around, next to or among the trunk or branches of one or
more mature trees while above ground level. Tree houses can be used for
recreation, work space, habitation, a hangout space and observation. People
occasionally connect ladders or staircases to get up to the platforms."
  },
  {
    "concept": "Castle",
    "description": "A castle is a type of fortified structure built during the
Middle Ages predominantly by the nobility or royalty and by military orders.
Scholars usually consider a castle to be the private fortified residence of a
lord or noble. This is distinct from a mansion, palace, and villa, whose main
purpose was exclusively for pleasance and are not primarily fortresses but may
be fortified.[a] Use of the term has varied over time and, sometimes, has also
been applied to structures such as hill forts and 19th- and 20th-century homes
built to resemble castles. Over the Middle Ages, when genuine castles were
built, they took on a great many forms with many different features, although
some, such as curtain walls, arrowslits, and portcullises, were commonplace."
  },
  {
    "concept": "Fortress",
    "description": "A fortification (also called a fort, fortress, fastness, or
```

```
stronghold) is a military construction designed for the defense of territories
in warfare, and is used to establish rule in a region during peacetime. The
term is derived from Latin fortis ('strong') and facere ('to make'). From very
early history to modern times, defensive walls have often been necessary for
cities to survive in an ever-changing world of invasion and conquest. Some
settlements in the Indus Valley Civilization were the first small cities to be
fortified. In ancient Greece, large stone walls had been built in Mycenaean
Greece, such as the ancient site of Mycenae (known for the huge stone blocks of
its 'cyclopean' walls). A Greek phrourion was a fortified collection of
buildings used as a military garrison, and is the equivalent of the Roman
castellum or fortress. These constructions mainly served the purpose of a watch
tower, to guard certain roads, passes, and borders. Though smaller than a real
fortress, they acted as a border guard rather than a real strongpoint to watch
and maintain the border."
  }
]
```

Output:

```
[
  {
    "concept": "Shed",
    "description": "A shed is typically a simple, single-story roofed
structure, often used for storage, for hobbies, or as a workshop, and typically
serving as outbuilding, such as in a back garden or on an allotment. Sheds vary
considerably in their size and complexity of construction, from simple open-
sided ones designed to cover bicycles or garden items to large wood-framed
structures with shingled roofs, windows, and electrical outlets. Sheds used on
farms or in the industry can be large structures. The main types of shed
construction are metal sheathing over a metal frame, plastic sheathing and
frame, all-wood construction (the roof may be asphalt shingled or sheathed in
tin), and vinyl-sided sheds built over a wooden frame. Small sheds may include
a wooden or plastic floor, while more permanent ones may be built on a concrete
pad or foundation. Sheds may be lockable to deter theft or entry by children,
domestic animals, wildlife, etc.",
    "categories": [
      {
        "distance": 0.9907871670904073,
        "concept": "Outbuildings",
        "description": "Outbuildings are structures that are separate from a
main building, typically located on a property for purposes such as storage,
workshops, or housing animals and equipment. This category includes structures
like sheds and barns, which serve specific functions like storing tools,
equipment, or livestock."
      },
      {
        "distance": 1.148880974178631,
        "concept": "Auxiliary Structures",
        "description": "Auxiliary structures are secondary or additional
buildings that serve various practical purposes related to a main dwelling or
property. This category encompasses structures like tree houses and
outbuildings, which provide functional, recreational, or storage spaces, often
designed to enhance the usability of the property."
      },
      {
        "distance": 1.292957924480073,
        "concept": "Military and Support Structures",
        "description": "Military and support structures refer to various types
of constructions designed for specific functions related to defense and
utility. This concept encompasses fortified structures, such as castles and
```

```
fortresses, built for protection and military purposes, as well as auxiliary
structures that serve practical roles for main buildings, including storage,
recreation, and additional facilities. Together, these structures enhance the
safety, functionality, and usability of a property or territory."
      }
    ]
  },
  {
    "concept": "Barn",
    "description": "A barn is an agricultural building usually on farms and
used for various purposes. In North America, a barn refers to structures that
house livestock, including cattle and horses, as well as equipment and fodder,
and often grain.[2] As a result, the term barn is often qualified e.g. tobacco
barn, dairy barn, cow house, sheep barn, potato barn. In the British Isles, the
term barn is restricted mainly to storage structures for unthreshed cereals and
fodder, the terms byre or shippon being applied to cow shelters, whereas horses
are kept in buildings known as stables.[2][3] In mainland Europe, however,
barns were often part of integrated structures known as byre-dwellings (or
housebarns in US literature). In addition, barns may be used for equipment
storage, as a covered workplace, and for activities such as threshing.",
    "categories": [
      {
        "distance": 0.9907871670904073,
        "concept": "Outbuildings",
        "description": "Outbuildings are structures that are separate from a
main building, typically located on a property for purposes such as storage,
workshops, or housing animals and equipment. This category includes structures
like sheds and barns, which serve specific functions like storing tools,
equipment, or livestock."
      },
      {
        "distance": 1.148880974178631,
        "concept": "Auxiliary Structures",
        "description": "Auxiliary structures are secondary or additional
buildings that serve various practical purposes related to a main dwelling or
property. This category encompasses structures like tree houses and
outbuildings, which provide functional, recreational, or storage spaces, often
designed to enhance the usability of the property."
      },
      {
        "distance": 1.292957924480073,
        "concept": "Military and Support Structures",
        "description": "Military and support structures refer to various types
of constructions designed for specific functions related to defense and
utility. This concept encompasses fortified structures, such as castles and
fortresses, built for protection and military purposes, as well as auxiliary
structures that serve practical roles for main buildings, including storage,
recreation, and additional facilities. Together, these structures enhance the
safety, functionality, and usability of a property or territory."
      }
    ]
  },
  {
    "concept": "Tree house",
    "description": "A tree house, tree fort or treeshed, is a platform or
building constructed around, next to or among the trunk or branches of one or
more mature trees while above ground level. Tree houses can be used for
recreation, work space, habitation, a hangout space and observation. People
occasionally connect ladders or staircases to get up to the platforms.",
    "categories": [
      {
```

```
              "distance": 1.148880974178631,
              "concept": "Auxiliary Structures",
              "description": "Auxiliary structures are secondary or additional
      buildings that serve various practical purposes related to a main dwelling or
      property. This category encompasses structures like tree houses and
      outbuildings, which provide functional, recreational, or storage spaces, often
      designed to enhance the usability of the property."
          },
          {
              "distance": 1.292957924480073,
              "concept": "Military and Support Structures",
              "description": "Military and support structures refer to various types
      of constructions designed for specific functions related to defense and
      utility. This concept encompasses fortified structures, such as castles and
      fortresses, built for protection and military purposes, as well as auxiliary
      structures that serve practical roles for main buildings, including storage,
      recreation, and additional facilities. Together, these structures enhance the
      safety, functionality, and usability of a property or territory."
          }
        ]
      },
      {
        "concept": "Castle",
        "description": "A castle is a type of fortified structure built during the
      Middle Ages predominantly by the nobility or royalty and by military orders.
      Scholars usually consider a castle to be the private fortified residence of a
      lord or noble. This is distinct from a mansion, palace, and villa, whose main
      purpose was exclusively for pleasance and are not primarily fortresses but may
      be fortified.[a] Use of the term has varied over time and, sometimes, has also
      been applied to structures such as hill forts and 19th- and 20th-century homes
      built to resemble castles. Over the Middle Ages, when genuine castles were
      built, they took on a great many forms with many different features, although
      some, such as curtain walls, arrowslits, and portcullises, were commonplace.",
        "categories": [
          {
              "distance": 0.9152435235428339,
              "concept": "Fortified structures",
              "description": "Fortified structures refer to buildings designed to
      protect from attacks and enhance defense. This category encompasses various
      forms of military architecture, including castles and fortresses. Castles serve
      as private residences for nobility or military orders with substantial
      fortification features, while fortresses are broader military constructions
      aimed at defending territories and establishing control. Both types share the
      common purpose of defense against invasion, though they serve different social
      and functional roles."
          },
          {
              "distance": 1.292957924480073,
              "concept": "Military and Support Structures",
              "description": "Military and support structures refer to various types
      of constructions designed for specific functions related to defense and
      utility. This concept encompasses fortified structures, such as castles and
      fortresses, built for protection and military purposes, as well as auxiliary
      structures that serve practical roles for main buildings, including storage,
      recreation, and additional facilities. Together, these structures enhance the
      safety, functionality, and usability of a property or territory."
          }
        ]
      },
      {
        "concept": "Fortress",
```

```
    "description": "A fortification (also called a fort, fortress, fastness, or
stronghold) is a military construction designed for the defense of territories
in warfare, and is used to establish rule in a region during peacetime. The
term is derived from Latin fortis ('strong') and facere ('to make'). From very
early history to modern times, defensive walls have often been necessary for
cities to survive in an ever-changing world of invasion and conquest. Some
settlements in the Indus Valley Civilization were the first small cities to be
fortified. In ancient Greece, large stone walls had been built in Mycenaean
Greece, such as the ancient site of Mycenae (known for the huge stone blocks of
its 'cyclopean' walls). A Greek phrourion was a fortified collection of
buildings used as a military garrison, and is the equivalent of the Roman
castellum or fortress. These constructions mainly served the purpose of a watch
tower, to guard certain roads, passes, and borders. Though smaller than a real
fortress, they acted as a border guard rather than a real strongpoint to watch
and maintain the border.",
    "categories": [
      {
        "distance": 0.9152435235428339,
        "concept": "Fortified structures",
        "description": "Fortified structures refer to buildings designed to
protect from attacks and enhance defense. This category encompasses various
forms of military architecture, including castles and fortresses. Castles serve
as private residences for nobility or military orders with substantial
fortification features, while fortresses are broader military constructions
aimed at defending territories and establishing control. Both types share the
common purpose of defense against invasion, though they serve different social
and functional roles."
      },
      {
        "distance": 1.292957924480073,
        "concept": "Military and Support Structures",
        "description": "Military and support structures refer to various types
of constructions designed for specific functions related to defense and
utility. This concept encompasses fortified structures, such as castles and
fortresses, built for protection and military purposes, as well as auxiliary
structures that serve practical roles for main buildings, including storage,
recreation, and additional facilities. Together, these structures enhance the
safety, functionality, and usability of a property or territory."
      }
    ]
  }
]
```

Required Parameters

- `name` : A unique name for the operation.

- `type` : Must be set to "cluster".

- `embedding_keys` : A list of keys to use for the embedding that is clustered on

- `summary_prompt` : The prompt used to summarize a cluster based on its children. Access input variables by iterating over `inputs` with `{% for input in inputs %}` and accessing properties with `{{input.keyname}}` .

- `summary_schema` : The schema for the summary of each cluster. This is the output schema for the `summary_prompt` based llm call.

Optional Parameters

Parameter	Description	Default
`output_key`	The name of the output key where the cluster path will be inserted in the items.	"clusters"
`model`	The language model to use	Falls back to `default_model`
`embedding_model`	The embedding model to use	"text-embedding-3-small"
`timeout`	Timeout for each LLM call in seconds	120
`max_retries_per_timeout`	Maximum number of retries per timeout	2
`sample`	Number of items to sample for this operation	None
`litellm_completion_kwargs`	Additional parameters to pass to LiteLLM completion calls.	{}