Configuration

params

The **params** variable is a dictionary where you can and should specify some of the ETL's pipeline components such as the *data_source.*

- (Required string) The value of the*data_source* params field should be the sector of activity from where the raw data of the pipeline is coming from, example "yammer", "sharepoint" or "elastic_search". This should not change in time as it is used in paths when saving data in the different lake zones and also in thedatabricks database name where all curated tables for this pipeline will be kept in order to keep things organised.
- (Required dictionary) The *tables*field, a dictionary that can be empty. The specified tables must bestandard.
  - (Optional dictionary) One or more *<table_name>*fields. Simply specifying table names here allows the usage of some of the library's most useful function to save development time by handling common operations in a standard way.
    - (Optional string or list) *unique_key* : A column of unique values or multiple column names for a composite key. Needed for merge operations

```
params = {
  'data_source': 'yammer',
  'tables': {
    'MessagesLikes': {
      'unique_key': ['id', 'message_id'],
      'trusted_incremental_mode': True
    },
    'Messages': {
      'unique_key': 'id',
      'trusted_incremental_mode': False
    },
    'Users': {
      'unique_key': 'id',
      'trusted_incremental_mode': True
    },
    'Groups': {
      'unique_key': 'id',
      'trusted_incremental_mode': False
    }
  }
}
```

Config

A Config instance is needed for most operations.

At the very least, it can take still take no arguments and all mount names will be *raw_default*, *curated_default*and*trusted_default.*

```
from ETL import Config

config = Config()
yammer_config = Config(params)
```