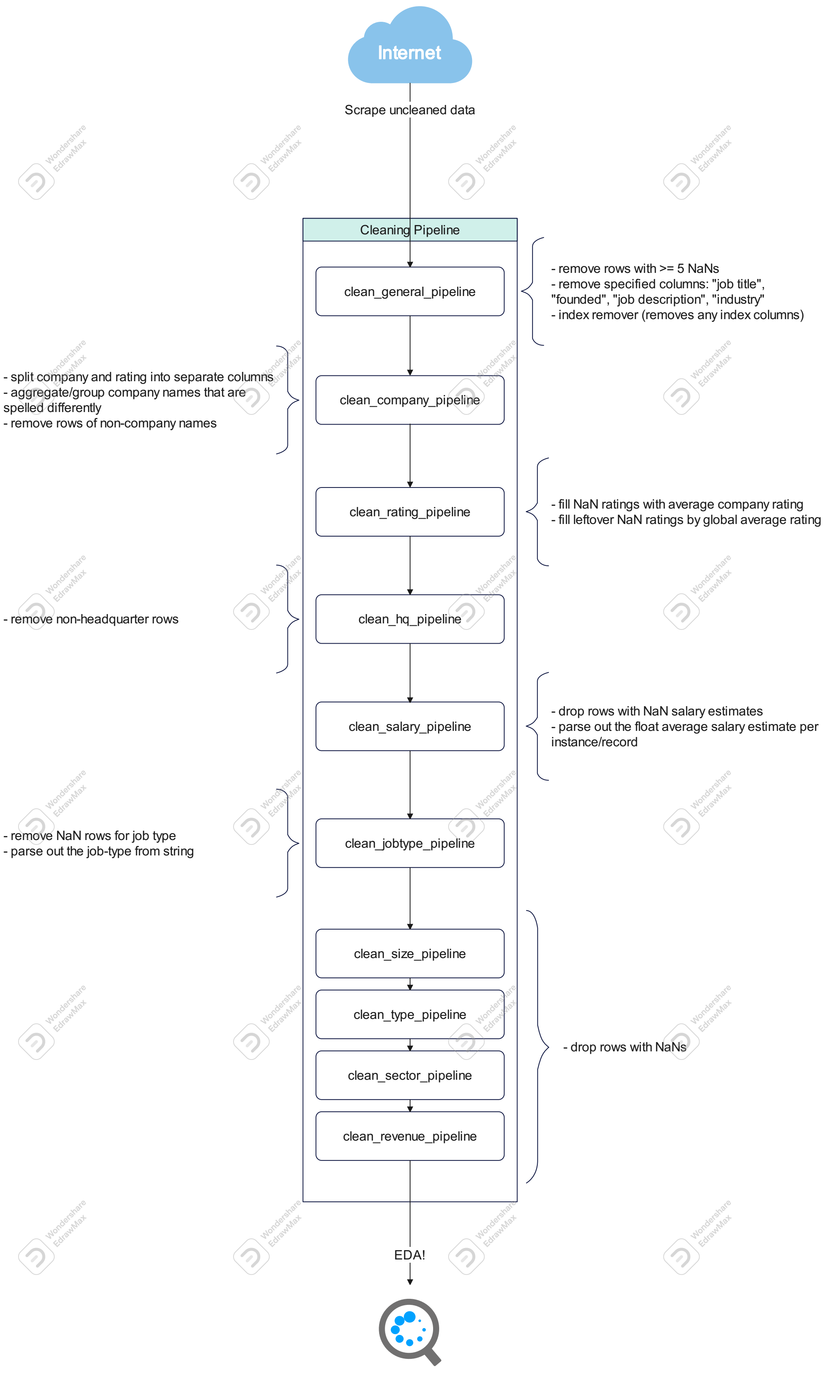
Internet
Scrape uncleaned data
Cleaning Pipeline
clean_general_pipeline
- remove rows with >= 5 NaNs
- remove specified columns: "job title", "founded", "job description", "industry"
- index remover (removes any index columns)
clean_company_pipeline
- split company and rating into separate columns
- aggregate/group company names that are spelled differently
- remove rows of non-company names
clean_rating_pipeline
- fill NaN ratings with average company rating
- fill leftover NaN ratings by global average rating
clean_hq_pipeline
- remove non-headquarter rows
clean_salary_pipeline
- drop rows with NaN salary estimates
- parse out the float average salary estimate per instance/record
clean_jobtype_pipeline
- remove NaN rows for job type
- parse out the job-type from string
clean_size_pipeline
clean_type_pipeline
clean_sector_pipeline
clean_revenue_pipeline
- drop rows with NaNs
EDA!