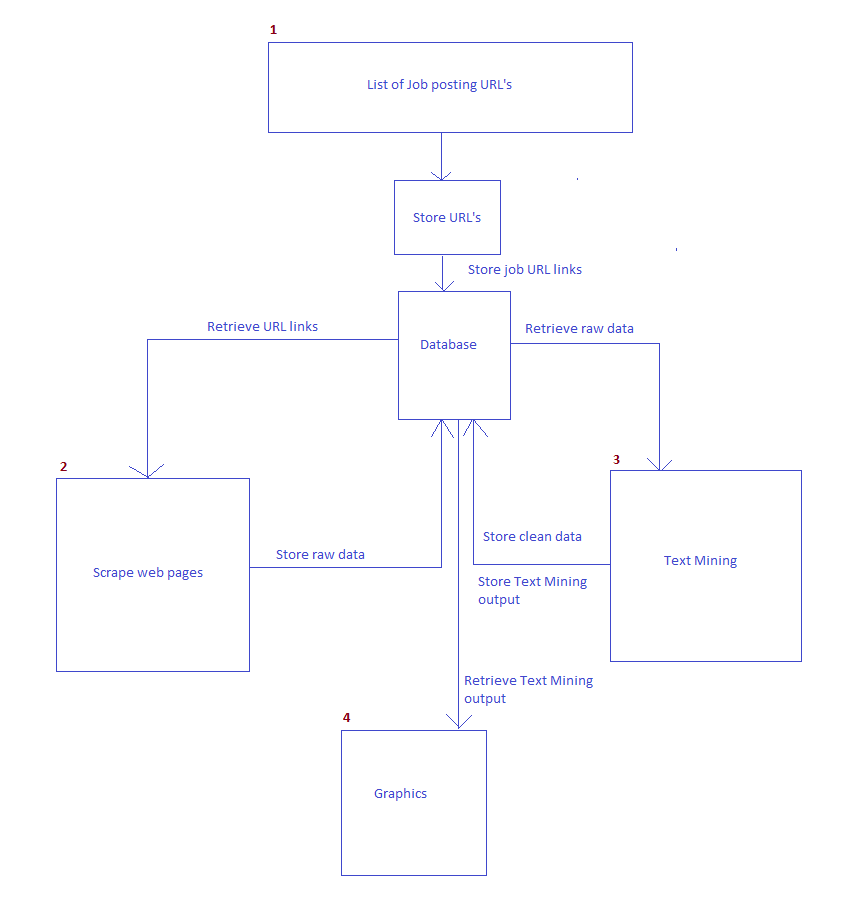
# Data 607: Group 7 Project 3

Due to time constraints, the group has decided not to implement the database. We do not have an existing shared environment that would allow us to access the database from anywhere. To cover the database requirement of the project, we’ve decided to present a design of the concept of how we would incorporate this database in our project.

# Overall Workflow

* Job posting URLs are going to be stored to a table
* Output from web scraping is going to be stored in a table
* Output from text mining is going to be stored in a table
* Clean data
* Term document matrix
* Term Frequency



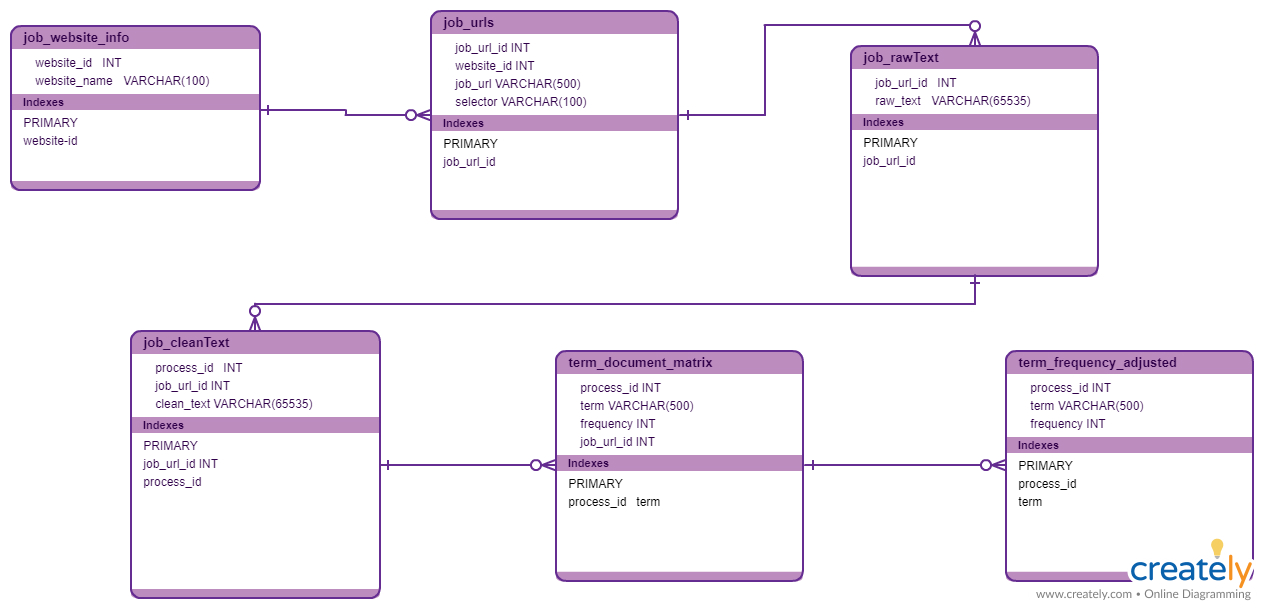
# Database Table Schema

Tables

* job\_website\_info: stores website name
* job\_urls: stores each URL source
* job\_rawText: stores data scraped from the URL location
* job\_cleanText: stores the clean version of the data
* term\_document\_matrix: stores the term document matrix output of the job
* term\_frequency\_adjusted: stores the term frequency output of the job

NOTE:

For data that could change with each run, a *process\_id* is added. This id will be provided by the job. This design would allow us to analyze previous runs if necessary.



# Create Table Statements

CREATE DATABASE DATA607\_GROUP7\_PROJECT3;

USE DATA607\_GROUP7\_PROJECT3;

CREATE TABLE job\_website\_info(

website\_id INT,

website\_name VARCHAR(100),

PRIMARY KEY(website\_id)

);

CREATE TABLE job\_urls (

job\_url\_id INT,

website\_id INT,

job\_url VARCHAR(500),

selector VARCHAR(100),

PRIMARY KEY(job\_url\_id)

);

CREATE TABLE job\_rawText (

job\_url\_id INT,

raw\_text VARCHAR(65535),

PRIMARY KEY(job\_url\_id)

);

CREATE TABLE job\_cleanText (

process\_id INT,

job\_url\_id INT,

clean\_text VARCHAR(65535),

PRIMARY KEY(process\_id, job\_url\_id)

);

CREATE TABLE term\_document\_matrix(

process\_id INT,

term VARCHAR(500),

frequency INT,

job\_url\_id INT,

PRIMARY KEY(process\_id, term)

);

CREATE TABLE term\_frequency\_adjusted(

process\_id INT,

term VARCHAR(500),

frequency INT,

PRIMARY KEY(process\_id, term)

)