**Project 3 (607) – Data Science Thought Leadership**

***Team members (SPARC):***

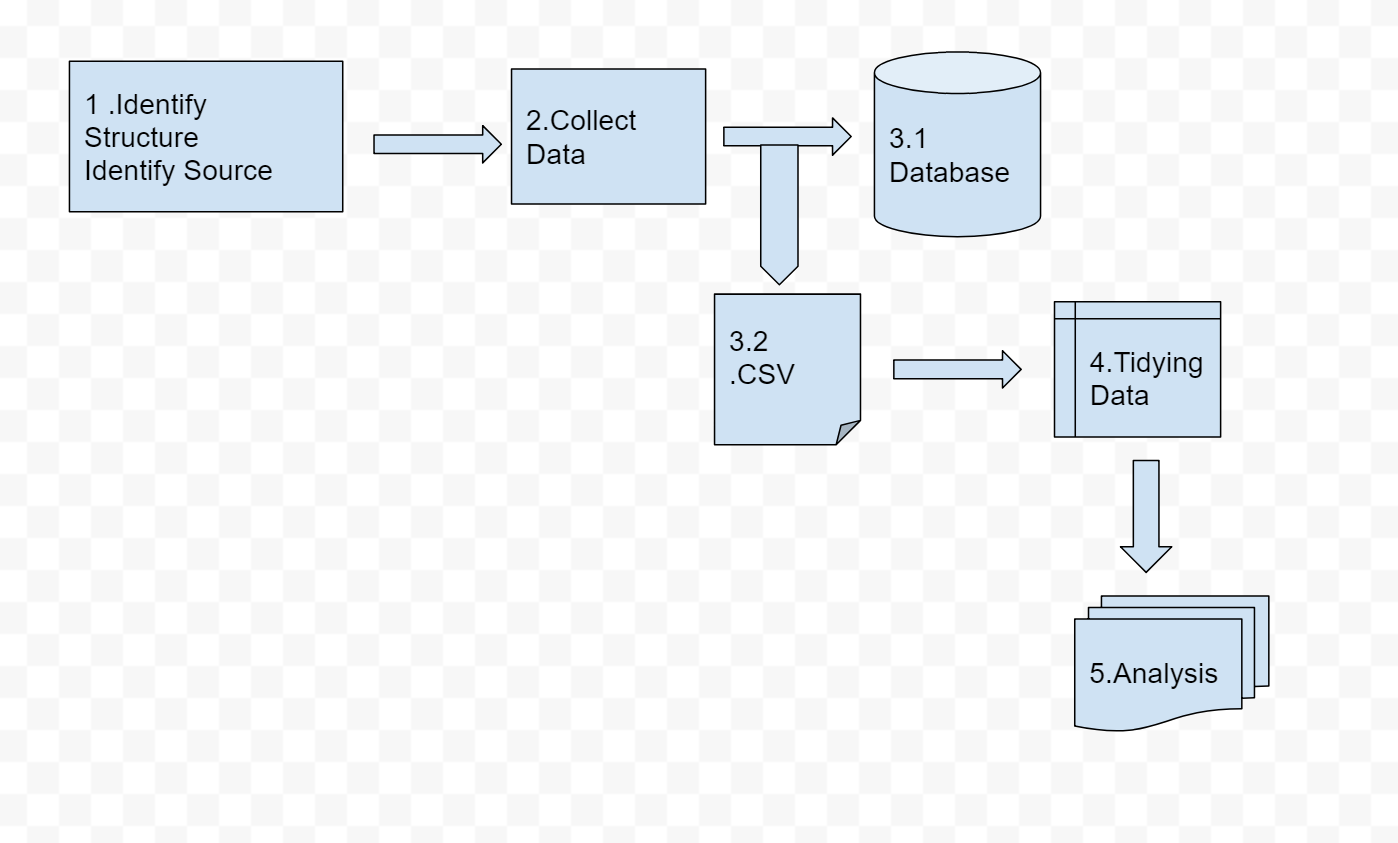
* Rajwant Mishra
* Samantha
* Priya
* Santosh Cheruku

***Motivation:***

The team in the first meet decided to pick the challenging project amongst the 2 project options given to us. Everyone decided to opt for alternate project since the project gave a definite target and to hit this target we had to explore numerous sources of data which weren’t freely available. That would help the team explore various solutions in addressing these issues. Also, team identified there was definite scope to use screen scraping, nosql databases in this solution which would be immensely useful in real time work.

***How did we approach the problem:***

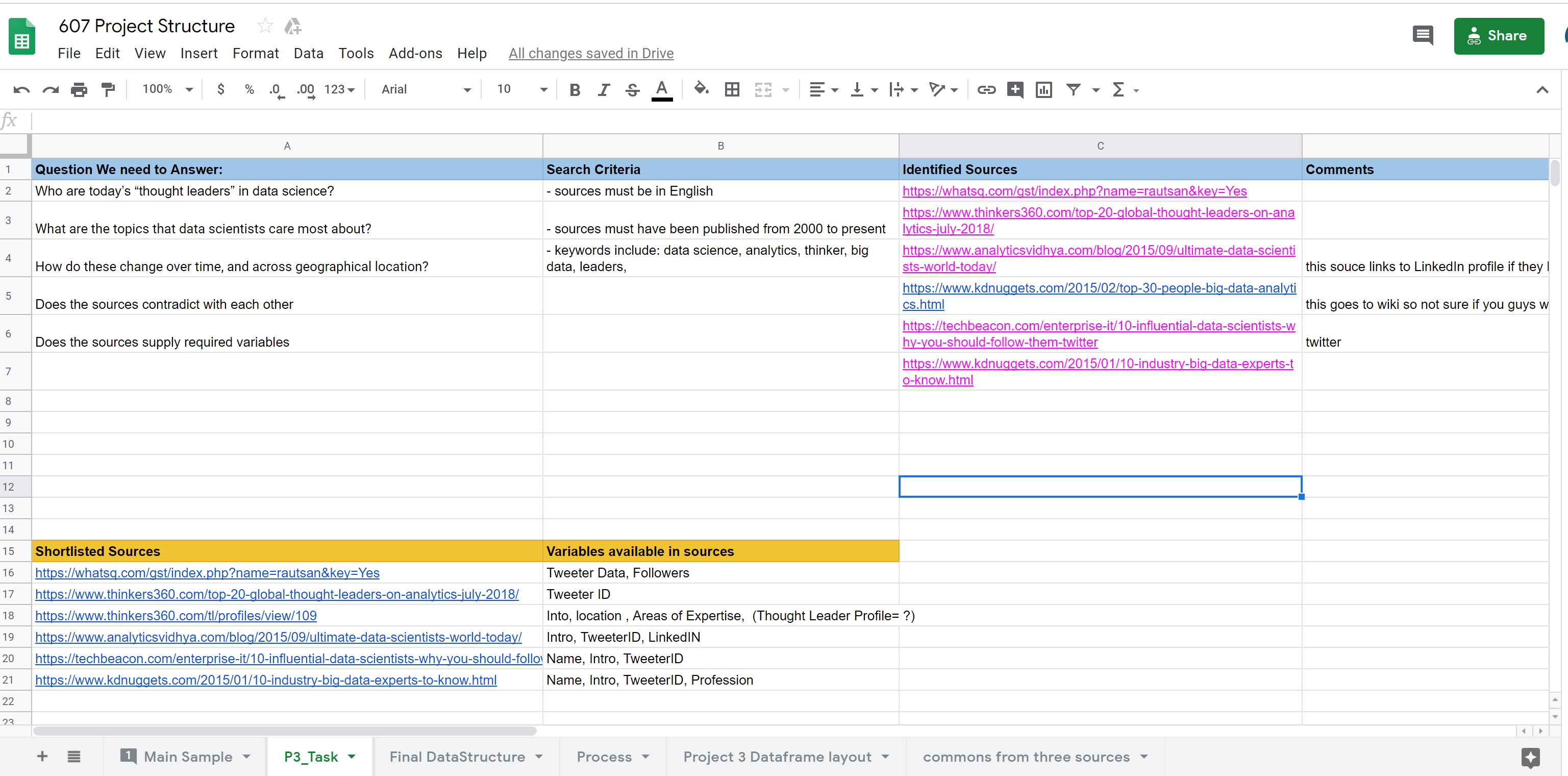
We identified the below steps/tasks to tackle the requirement for this project.

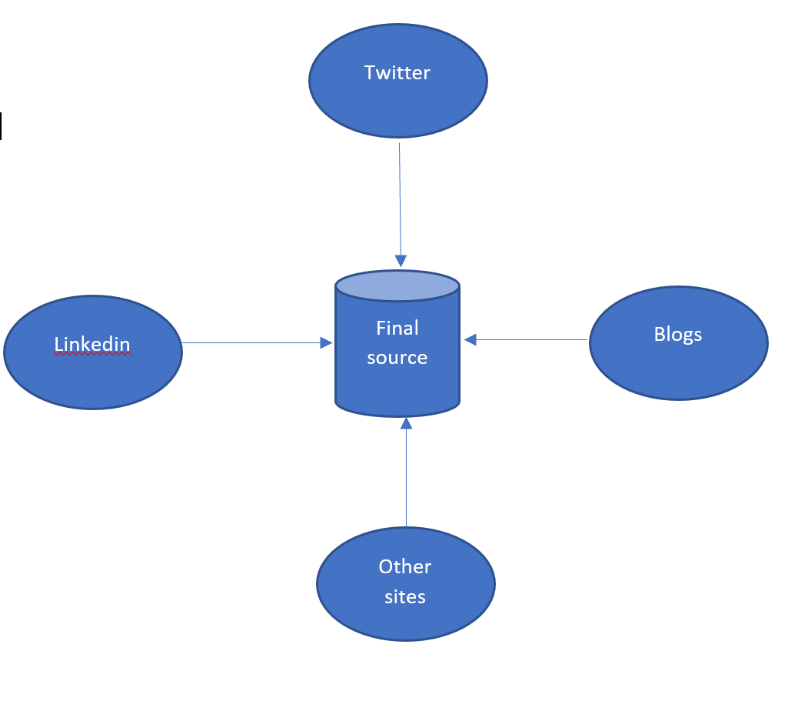


***Data Source Identification:***

**Below Google spreadsheet was shared with the team members**, to share their ideas on what would be best sources to fulfill this requirement and what are the interesting variables we need to look into.

https://docs.google.com/spreadsheets/d/1MQ4My5lz1Qqcsa\_RT8n1cvqH5niKVL2bvfbg7rN3AH4/edit#gid=128945259



Team ran into multiple issues during this search primarily regarding the sources. There was no definite source which could supply all the needed information. So we decided to pull necessary information from various sources such as Twitter, LinkedIn, Blogs and other sites online.

Challenges & solutions:

* There was no single source of data which could supply all the required attributes
  + Each team member took up this task and came up with all the findings/sources and recorded in the Google excel
* How can just one source be trusted?
  + Identified the most commonly mentioned profiles amongst all the sources and discussed which one to select as trusted source
* What if we are given conflicting information from each selected site?
  + Decided to exclude outlier information which was not proposed by majority sites

***Data Collection:***

Team implemented below tools/packages to scrape the data from different sites.

* Google Chrome plugin??
* RVest (R package to scrape web pages)

--Sample code here--

Challenges & Solutions:

* Some of the sites were restricted for data access unless authenticated and some API’s were restricted by limited number of API calls.
  + To overcome, the team created accounts to access the data, and restricted the number of API calls and decided to use other member accounts if we hit a limit.

***Data storage:***

We decided to use a NoSQL DB and decided to utilize MongoDB Atlas (Cloud). We noticed that MongoDB was providing a package for R i.e. mongolite which provided APIs required to connect to cloud and perform data operations.

--Sample code here--

Challenges & Solutions:

* Only one of our team member (Rajwant) was aware of MongoDB Atlas feature, and rest weren’t unaware.
  + Rajwant took a KT session to team to explain the process of connecting to cloud and performing operations
  + Credentials were shared with team members with limited time access.

***Data Tidying & Analysis:***

--Analysis goes here--

--Sample code here--

***Tools used:***

Below tools/packages were actively used for project purposes

* RStudio
* Mongolite
* RVest
* Google APIs

***Sequence of events:***

1. 3/14 – Project initiation discussion
2. 3/16 – First communication in exploring possible options
3. 3/17 – Screen share and KT session
4. 3/19 – Touch base on task status
5. 3/20 – Presentation discussion

***Team work and communication:***

* Along with existing channels like Slack, team also utilized Whatspp, Skype and Google hangouts to explore screen sharing options for discussions.
* The code was shared through Github and Google excel docs.