

WORK PLAN

Gaurav Malhotra

U101115FCS256

B.Tech CSE

2015-19

Mentor:

Manjit

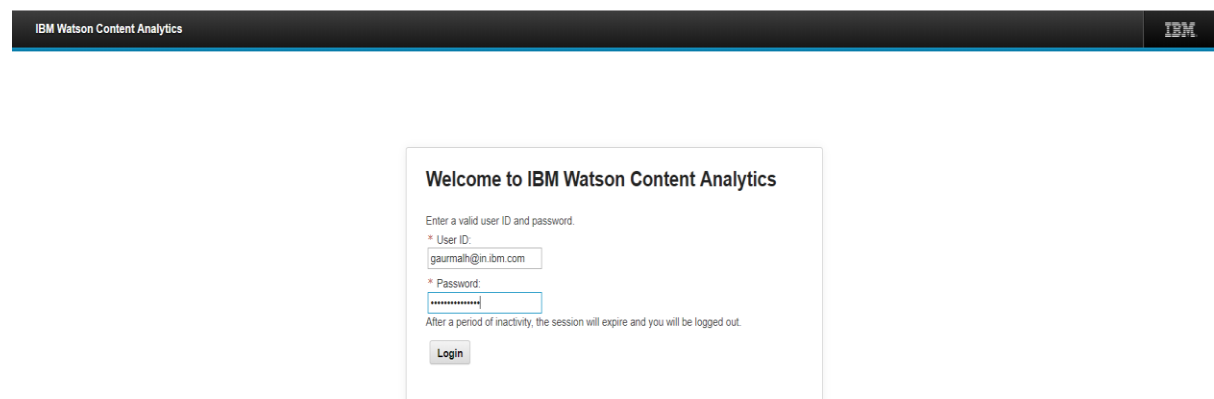
Sodhi

Objective:

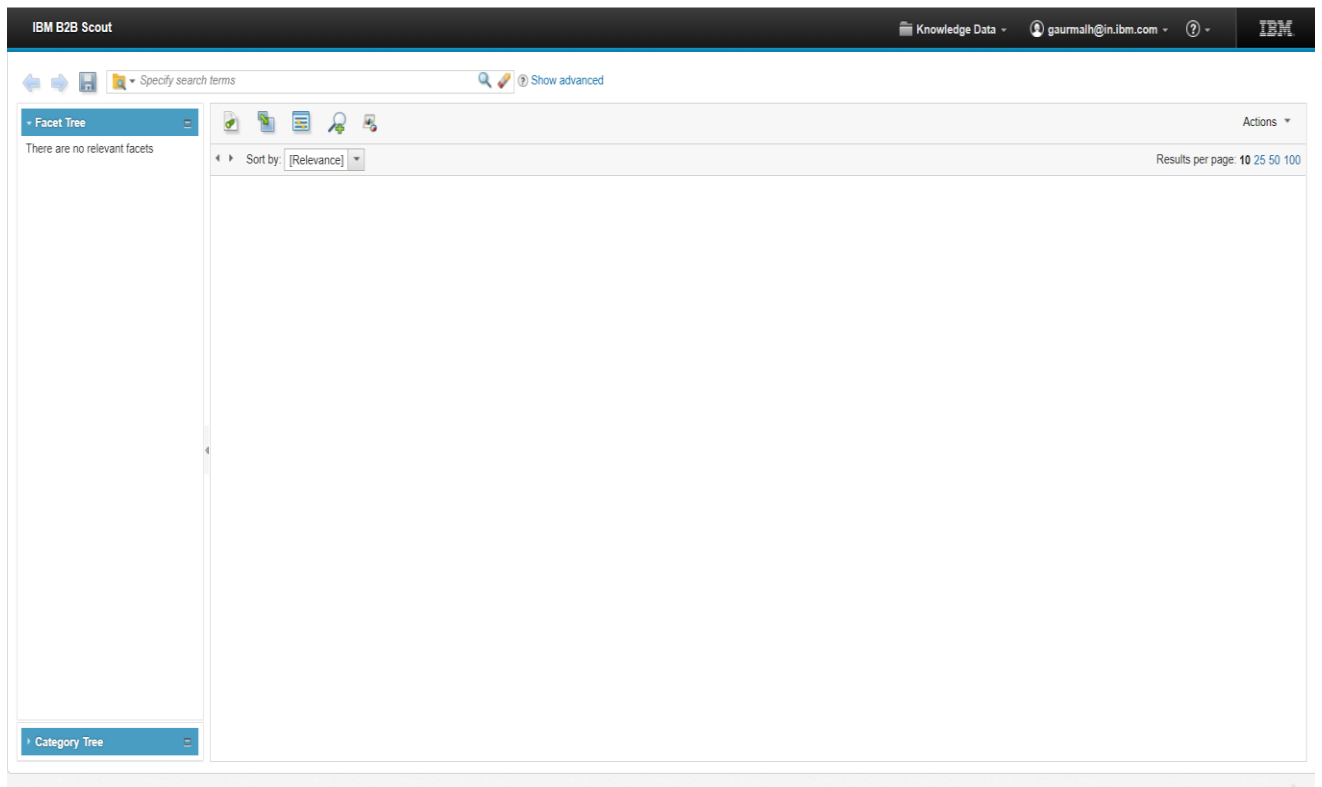
- To build a central repository system where people can fetch data.
- The software will fetch data from multiple sources which will be in multiple formats.
- The software will also understand NLU related queries.
- To perform ranking of all documents shown in the results.

Task Description and its current state (shown via screenshots):

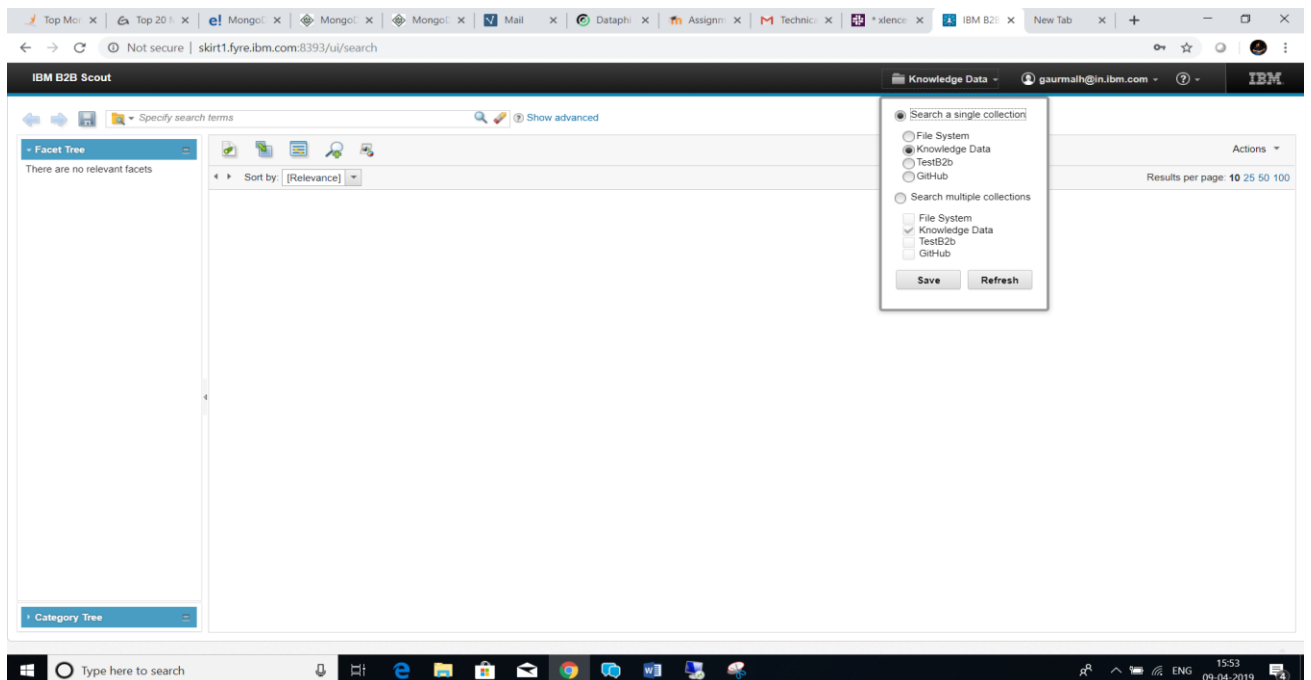
- 1) UI page with LDAP configured.



2) Dashboard for querying



3) Search enabled for Single collection. We can do multiple collection search also as it can be seen in the option "search multiple collection"



4) Displaying single query results

The screenshot shows the IBM B2B Scout interface. The search bar contains the query 'b2b'. The results are displayed in a list format. The first result is titled '供应链 B2B 协作 - 中国 | IBM' and is dated Feb 22, 2019, 11:31:09 PM. The second result is titled 'pom.xml' and is dated Feb 25, 2019, 12:39:35 PM. The third result is titled 'Watson Supply Chain - España | IBM' and is dated Feb 22, 2019, 11:24:30 PM. The fourth result is titled 'Watson Supply Chain - Российская Федерация | IBM' and is dated Feb 22, 2019, 11:24:30 PM. The interface includes a Facet Tree on the left, a search bar at the top, and a results list on the right.

5) Displaying results for the NLU query “What is IBM B2B Integrator”

The screenshot shows the IBM B2B Scout interface with the NLU query 'What is IBM B2B Integrator'. The results are displayed in a list format. The first result is titled 'IBM B2B Integrator - 概述 - 中国' and is dated Feb 22, 2019, 11:39:07 PM. The second result is titled '供应链产品 - 中国 | IBM' and is dated Feb 22, 2019, 11:30:52 PM. The third result is titled 'pom.xml' and is dated Feb 25, 2019, 12:39:35 PM. The fourth result is titled 'Connectivity and Integration Blog' and is dated Feb 21, 2019, 4:40:02 PM. The interface includes a Facet Tree on the left, a search bar at the top, and a results list on the right.

Technical Challenges and solutions adopted:

Challenges:

- 1) Fetching data from different sources was a tough job since most of the websites do not allow us to scrap data.
- 2) The data that is fetched will contain files of different formats. Reading these files to generate results was challenging.
- 3) Since data is in different format, performing NLU is also difficult.
- 4) LDAP connectivity for login was a bit of challenge too.

Solutions:

IBM has developed a tool to solve such problems knows as **Watson Content Analytics**. I have configured the tool accordingly to the challenges faced. It helped us to perform LDAP connectivity and made it secure to use. We can also perform NLU in the query section. Multiple results from the same source can be clubbed together. We can also filter results on the basis of user's need.

It provides us aid to fetch data from multiple resources. We can also query data at the same time from multiple resources. For this, we created clusters for different data source and enabled an option in the user's screen.

Scrapping the data was no longer an issue since **WCA** can easily fetch data and up to multiple layers. We can give the level of depth for which we require the data to be fetched.

Any special software used:

Watson Content Analytics

Any new skills I picked up:

- 1) I got to work with Watson services which was very exciting and got to know its features.

- 2) I learnt how to make rule-based query and create dictionaries for the query operations.
- 3) I also learnt about LDAP and how it helps to secure web-services.

Note: Due to confidentiality issues, I can't share anything related to Administrator part.