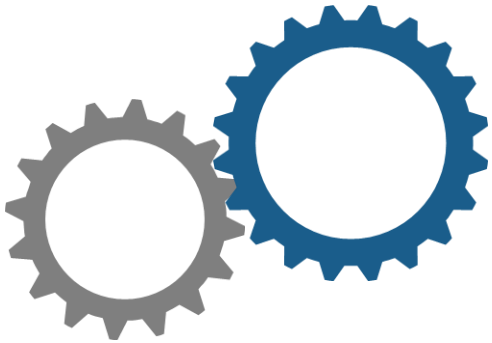
Final Project
Course- Information Architecture (DAV-6100)

Presented By - Group 2

Yeshiva University | KATZ SCHOOL

CONTENTS



Stack Overflow Data Warehouse Solution

Project Overview

- Our project is focused around building a data warehouse solution for the Stack Overflow platform
- The data warehouse will help Stack Overflow achieve their analytical needs like user interests, membership info, active user info & trending technologies/topics etc.
- Stack Overflow can improve their advertising and marketing revenue model by analyzing these users activities, interest and engagement metrics
- This can also help Stack Overflow with providing better job recommendations to end users who are actively looking for change. This improved jobs listing model can help them to increase ad revenue
- The data warehouse will provide a more comprehensive view of Stack Overflow and will provide data which will (in theory) be used to improve site performance

Roles and Responsibilities

Surbhi - AWS Architect:

Project lead and system architect.

Nosson - Data Enginner:

Collect relevant data according to the needs of the system and users.

Yihang - Database Administrator:

Operate, maintain and manage database management systems.

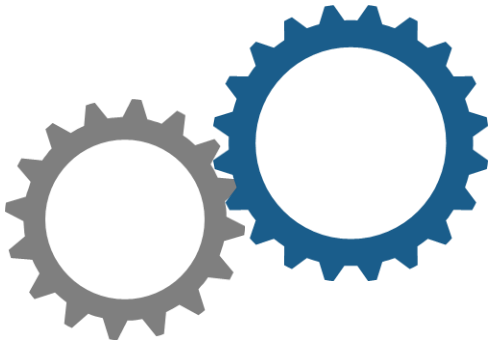
Yifeng - Database developer:

Design and develop database management systems.

Qianwen - Data Analyst:

Analyze and visualize data, and make industry assessments and forecasts based on data.

CONTENTS



About Data

Stack Overflow - Stack Overflow is a question and answer website for professional and enthusiast programmers. It features wide range of topics in the world of computer programming. They also provide membership options along with job portal features for their customers.

Data Source - https://relational.fit.cvut.cz/dataset/Stats

We are acquiring datasets from RELATIONAL DATASET REPOSITORY (An anonymized dump of all user-contributed content on the Stats Stack Overflow network)

Summary of Datasets

Entities in Data Warehouse-

- Posts
- Comments
- Tags
- Users
- Posts History
- Post Links
- Votes
- Badges

Method of access-

For extracting data from the internet open source we are using mysql connector in python environment and loading all the relevant data into the AWS S3 bucket for storage/Staging.

Data Quality-

- There are some missing values in the datasets which will deal according to the requirement.

- Basic Exploratory analysis (EDA) is being performed in python environment.

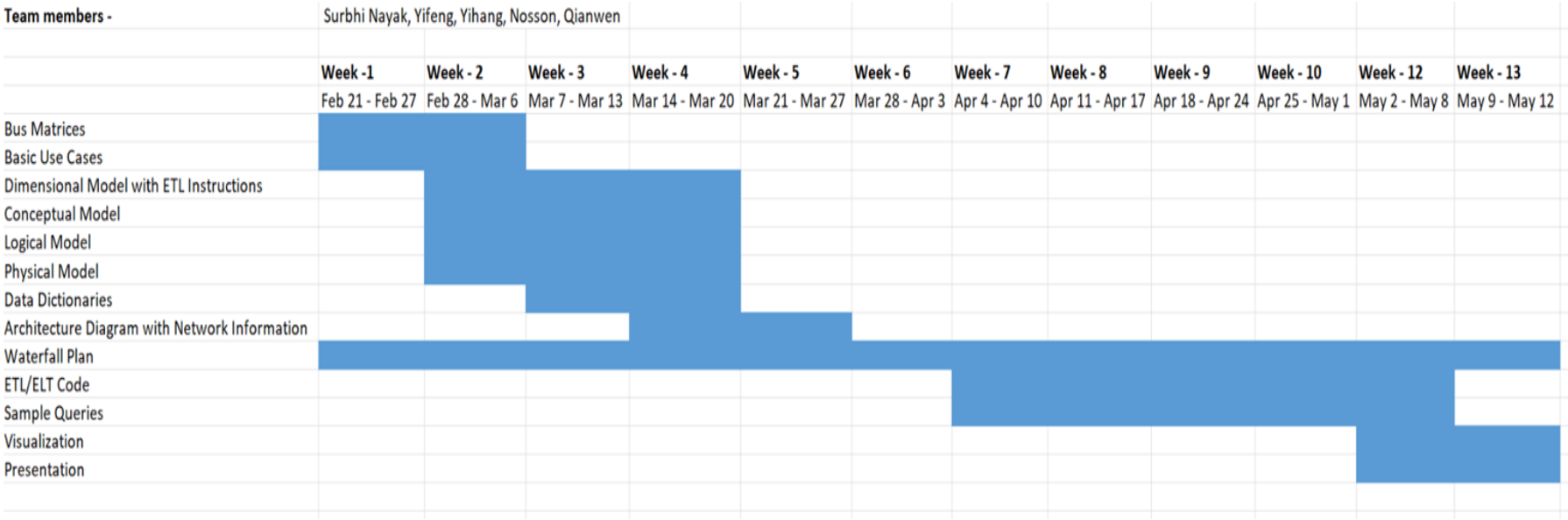
Data Profile

Dataset 1 Summary	
Source of information	https://relational.fit.cvut.cz/dataset/Stats
Number of Table	8
Number of Records	1,027,838 Rows 71 Columns
Data type and structure	Numeric, String, Temporal
Data Acquitsiton Method	MySQL connector

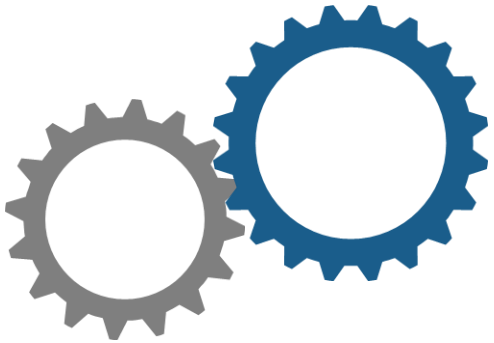
Dataset 2 Summary	
Source of information	https://content.techgig.com/technology/24-highest-paying-programming-languages-for-developers/articleshow/76243434.cms
Number of Table	1
Number of Records	24 Rows
Data type and structure	Numeric, String
Data Acquitsiton Method	Web Scraping using python

Timeline of the Project

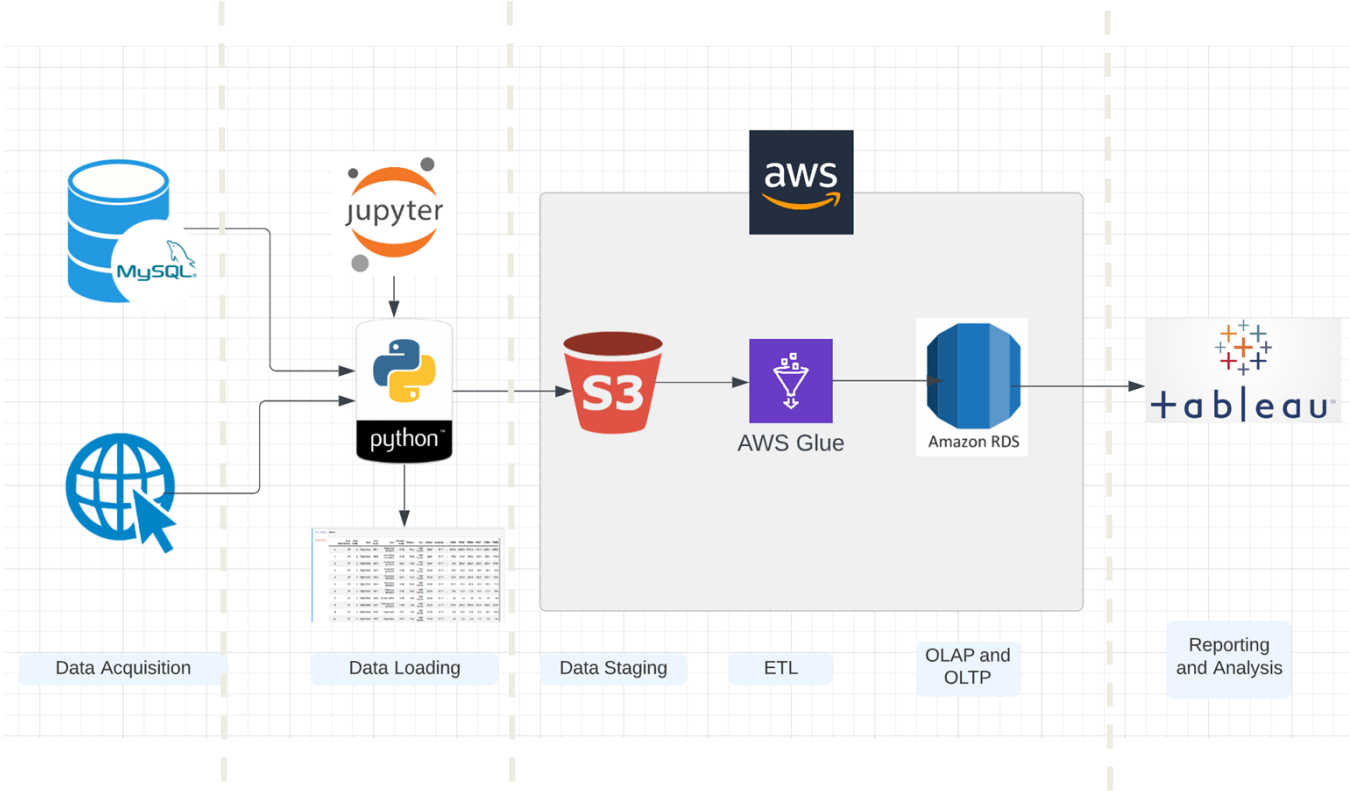
Waterfall Model



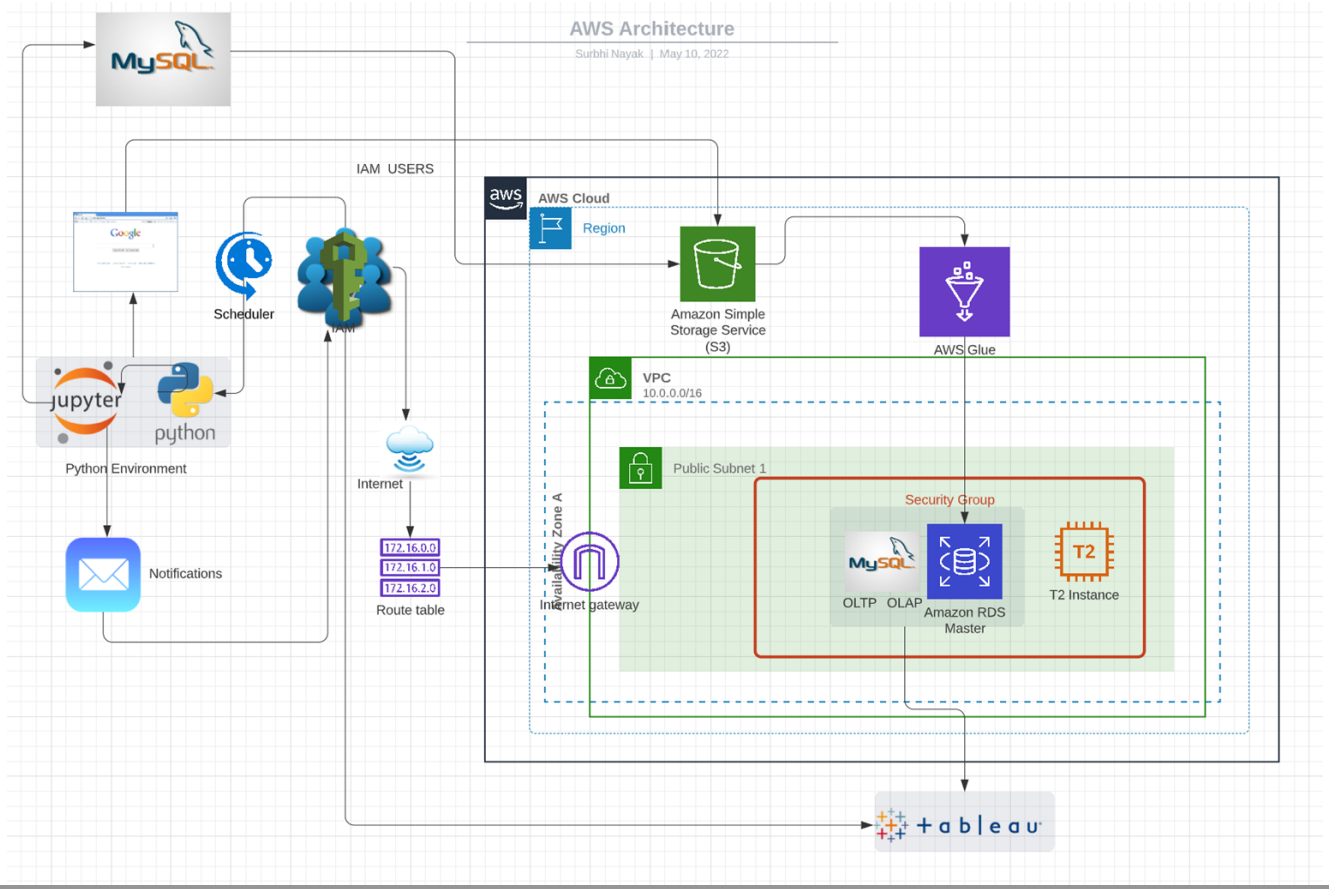
CONTENTS



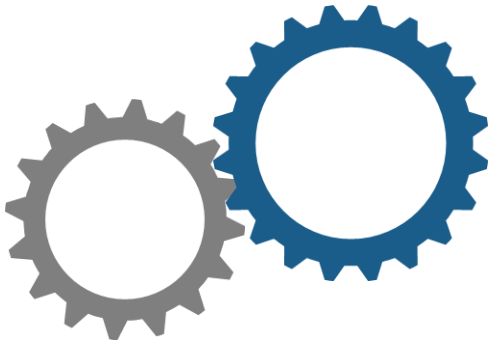
Work Flow


MySQL
jupyter
python
aws
S3
AWS Glue
Amazon RDS
+ableau
Data Acquisition
Data Loading
Data Staging
ETL
OLAP and OLTP
Reporting and Analysis

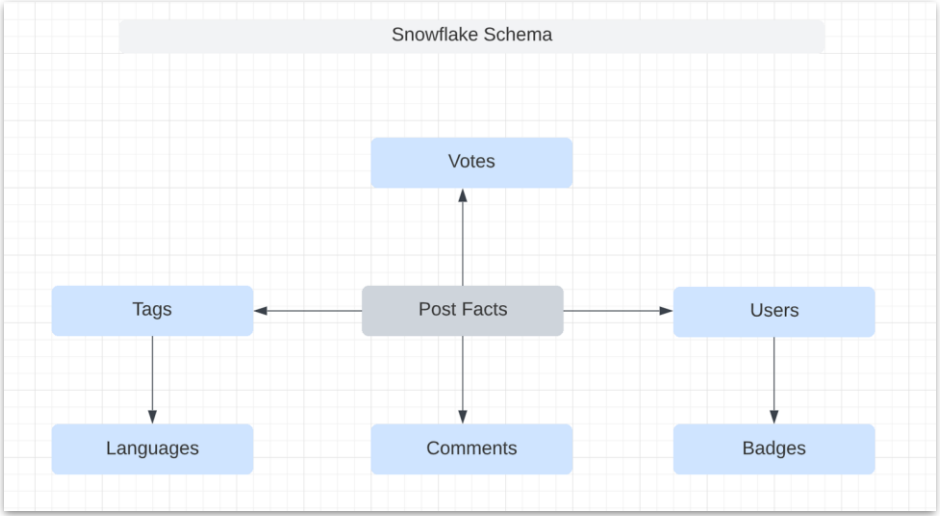
AWS Architecture



CONTENTS



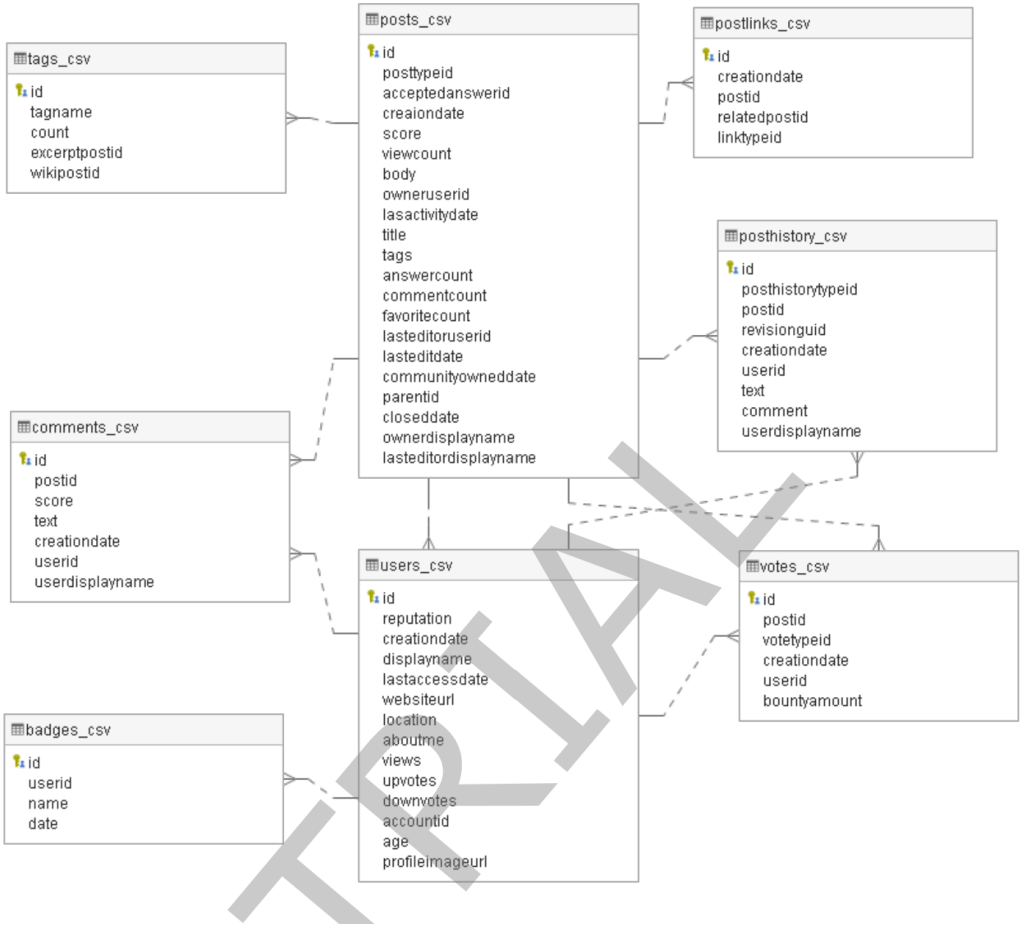
Conceptual Model



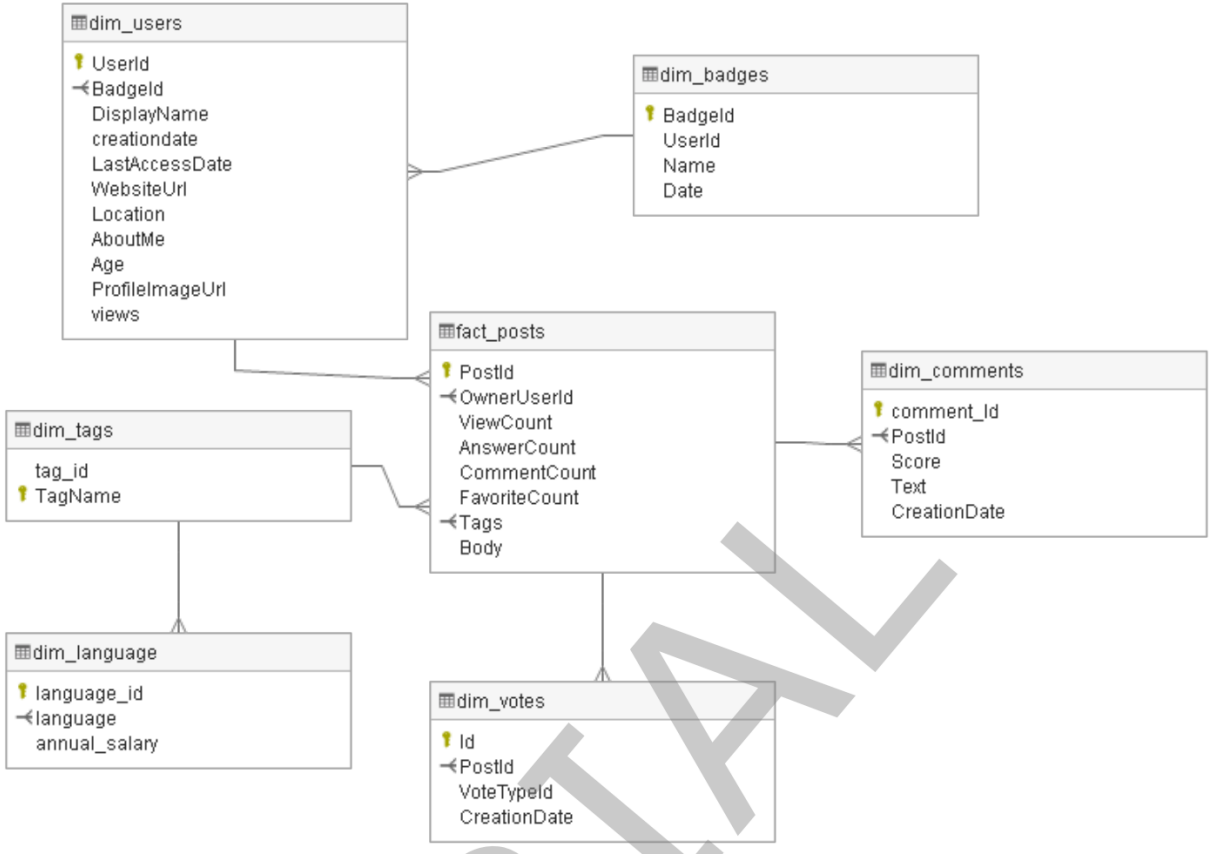
Bus Matrix

	Dimensions				
Facts (Business Processes)	date	text	score	count	link
access	■				■
creation	■				
answer	■	■	■	■	■
comment	■	■		■	■
edit	■	■			■
accept					
vote	■			■	
view	■			■	
close	■				

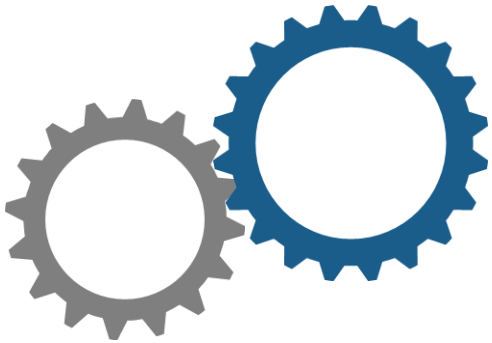
OLTP ERD



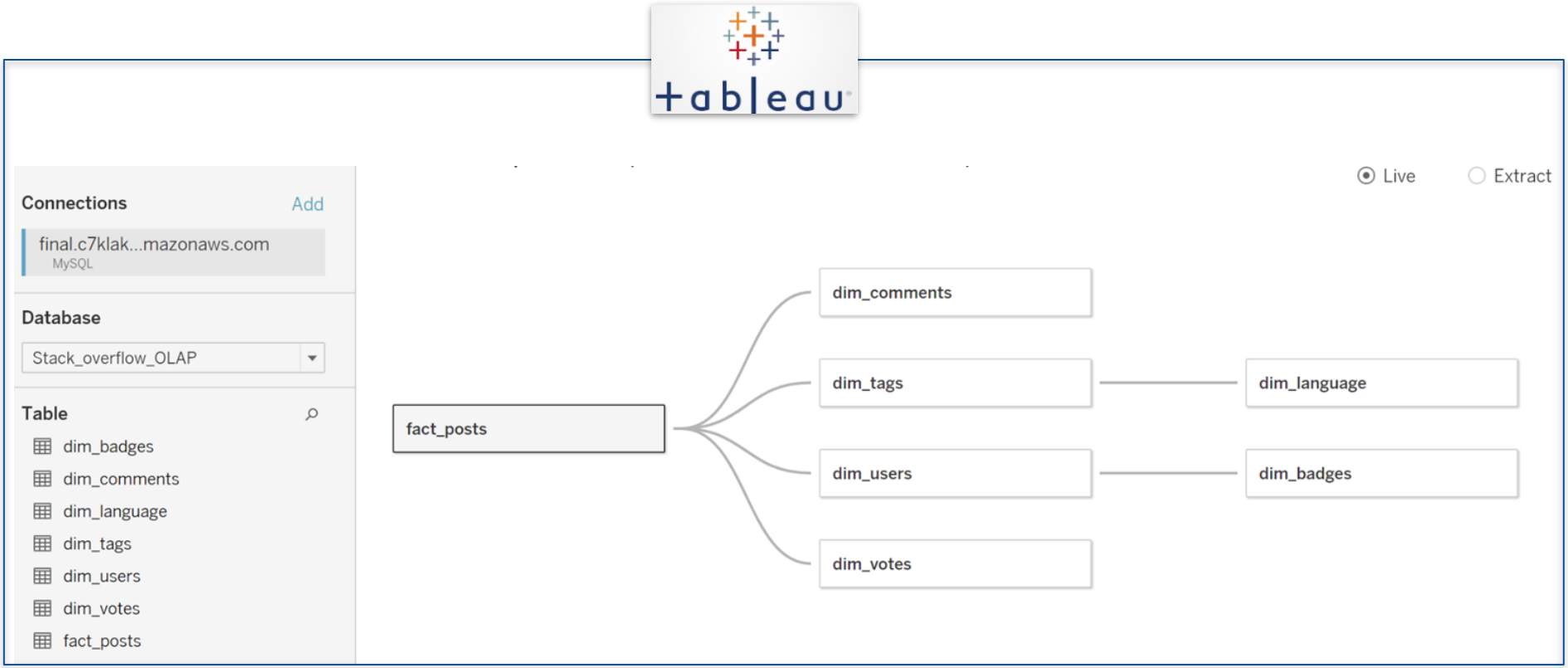
OLAP ERD


dim_users
UserId
BadgeId
DisplayName
creationdate
LastAccessDate
WebsiteUrl
Location
AboutMe
Age
ProfileImageUrl
views
dim_badges
BadgeId
UserId
Name
Date
fact_posts
PostId
OwnerUserId
ViewCount
AnswerCount
CommentCount
FavoriteCount
Tags
Body
dim_comments
comment_Id
PostId
Score
Text
CreationDate
dim_tags
tag_id
TagName
dim_language
language_id
language
annual_salary
dim_votes
Id
PostId
VoteTypeId
CreationDate

CONTENTS

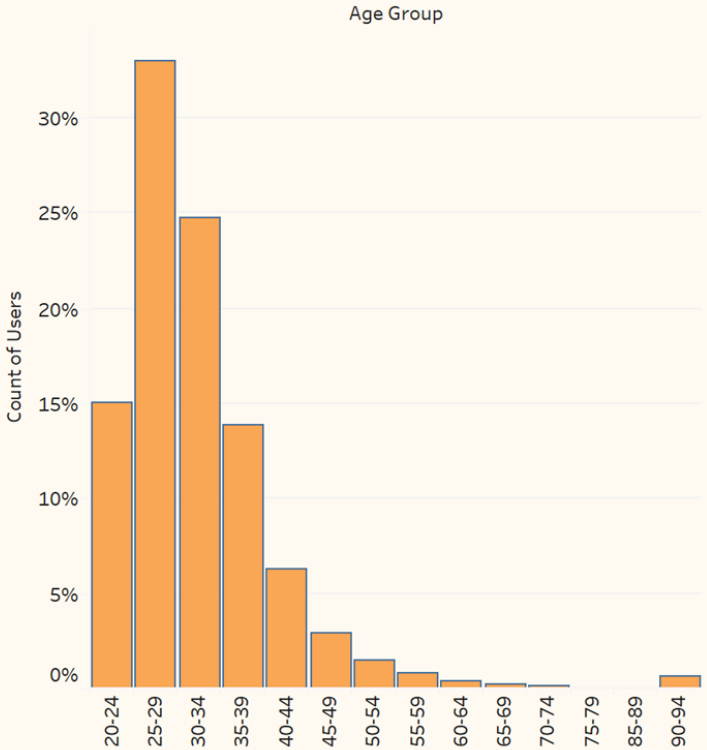


Connection in Tableau

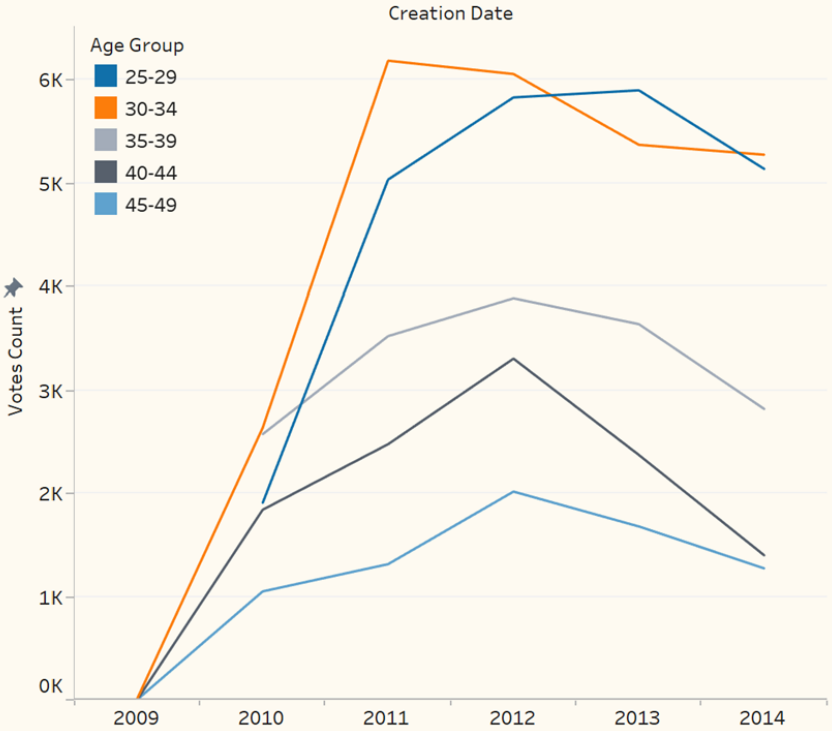


By understanding age groups and increase in user base participation over the years can help understand platform engagement rates and popularity.

Age groups & count of users



Age groups participation



Month wise users count can help to understand user's participation pattern on the stack overflow platform. Also we can analyze badegs popularity among top reputated users.

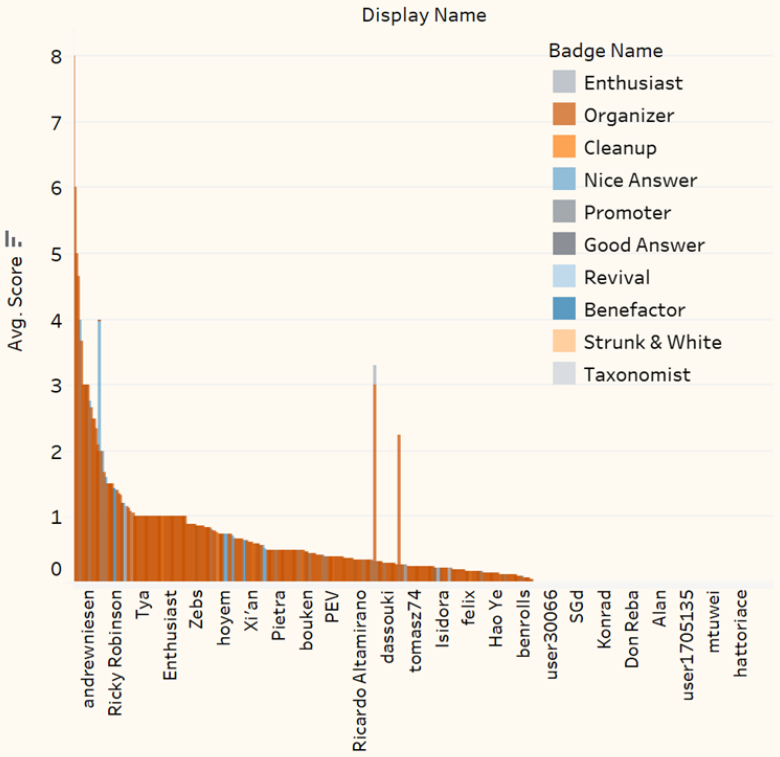
Platform Popularity By Month

Count of User Id

Creationdate 1 289

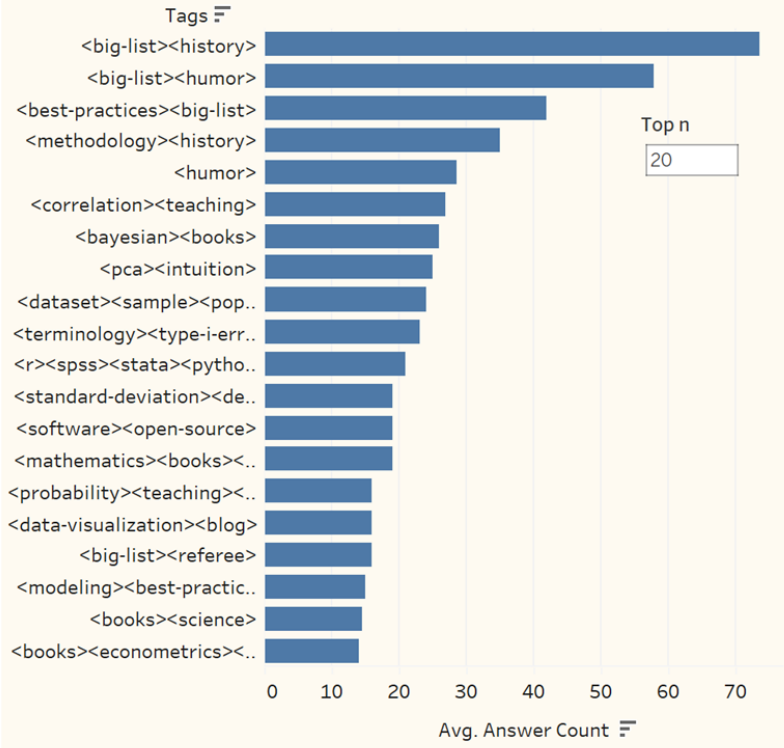
Age Group	July	Mar..	Febr..	Dec..	Nov..	Oct..	Sept..	Aug..	June	May	April	Jan..
15-19	20	10	42	9	6	1	9	10	7	12	9	12
20-24	125	118	110	76	88	68	79	127	117	124	116	89
25-29	289	232	246	194	204	179	191	234	210	254	251	230
30-34	217	157	205	158	157	132	142	197	164	142	171	191
35-39	122	92	100	79	79	78	104	119	87	78	116	90
40-44	70	38	46	34	38	37	38	58	37	36	41	49
45-49	34	21	19	13	16	22	19	21	19	17	18	23
50-54	15	11	9	8	9	9	9	13	7	12	8	12
55-59	5	9	8	6	5	4	7	7	1	3	9	2
60-64	2	4	4		3	1	2	2	3	3	3	5
65-69	3		3	2	1		1	2	1	1	2	3
70-74		1	1	1		3	1	2	2		1	2
75-79	1	1	1									
85-89				1					1			
90-94	5	5	7	1	5	5	3	5	7	4	2	5

Top reputated users

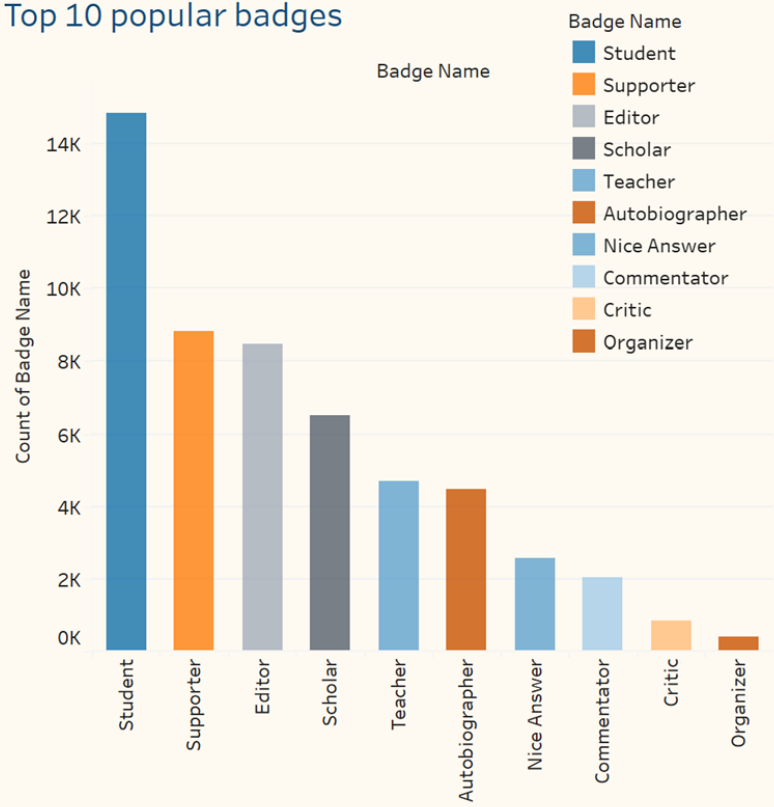


We can analyze various tags asssociated with the posts to help understand latest market trends/topics. Also, badges assigned to users can help further explain the engagement rate on platform.

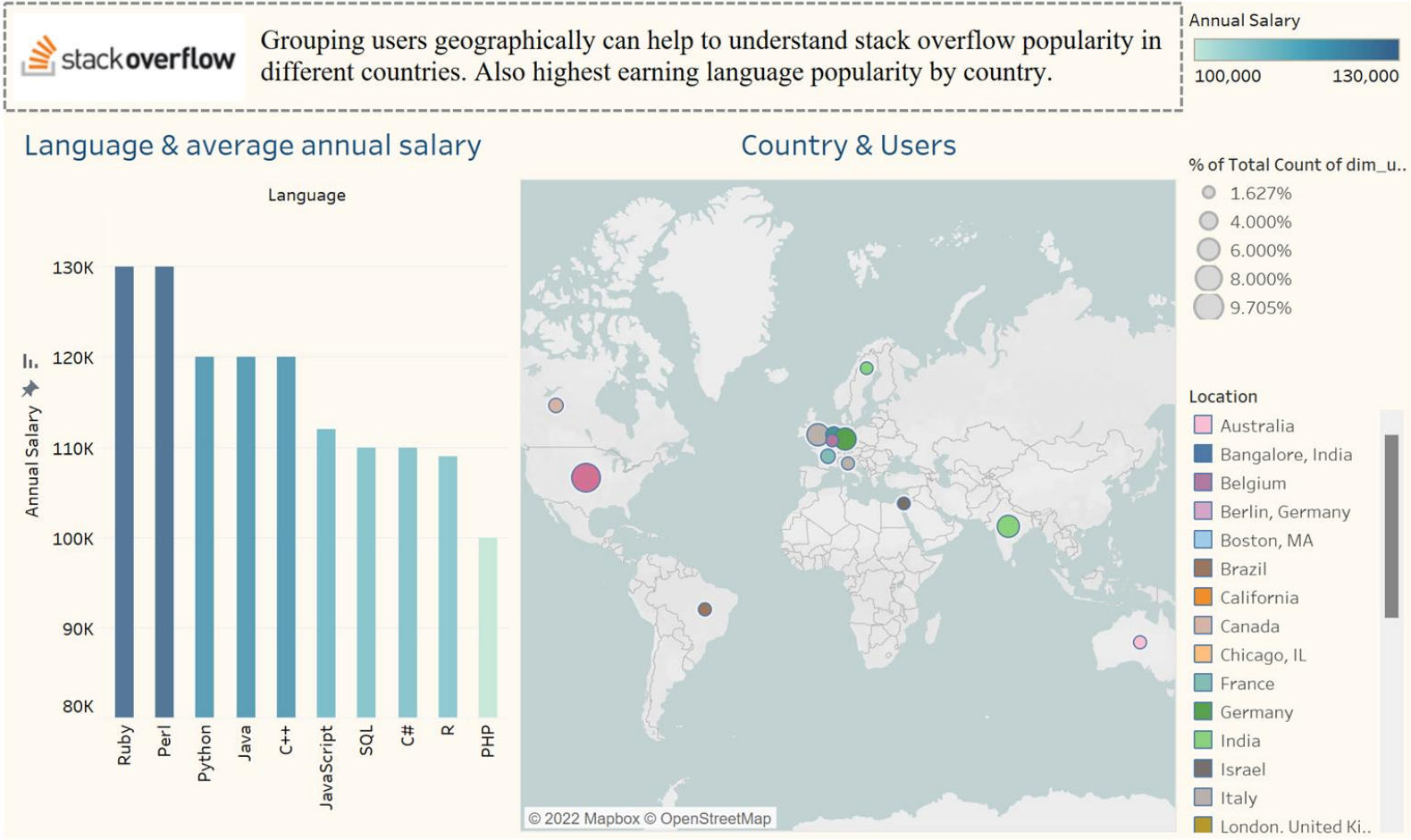
Top 20 Most Answered Tags

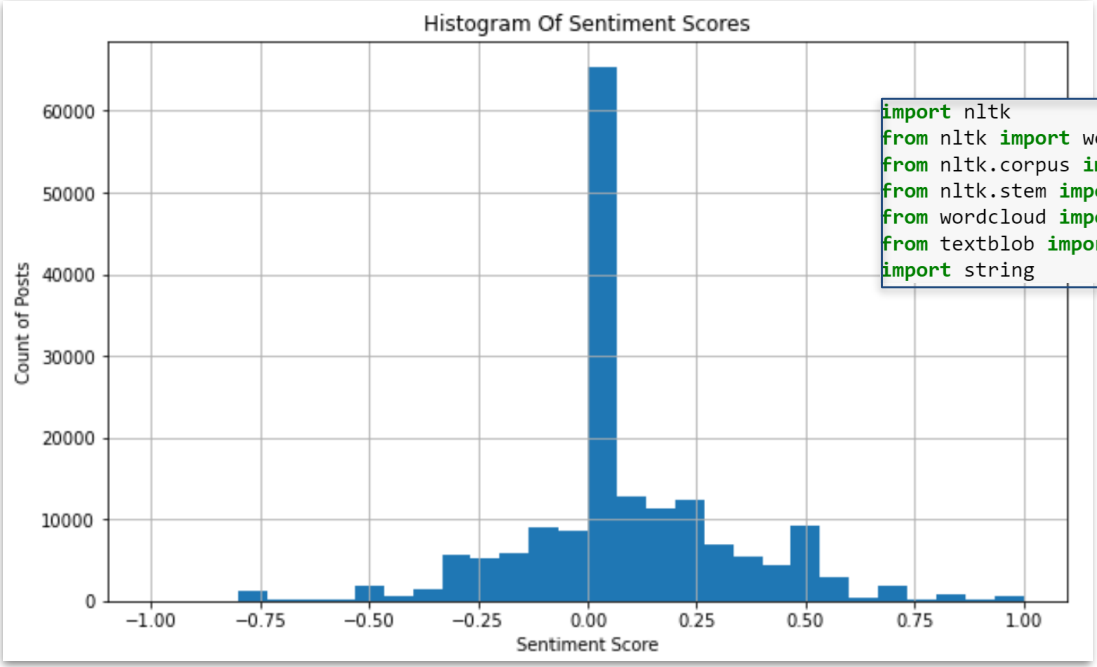


Top 10 popular badges



Grouping users geographically can help to understand stack overflow popularity in different countries. Also highest earning language popularity by country.



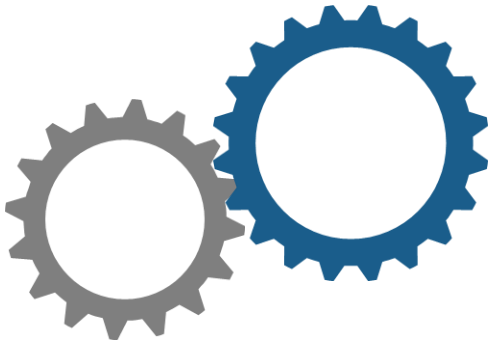



```
import nltk
from nltk import word_tokenize, sent_tokenize
from nltk.corpus import stopwords
from nltk.stem import LancasterStemmer, WordNetLemmatizer, PorterStemmer
from wordcloud import WordCloud, STOPWORDS
from textblob import TextBlob
import string
```

Comment	senti_score_polarity
could poster child fo argumentative subjective least neec	1
yes r nice 'valuable'	1
would convince boss use say excel	1
mature well supported standard within certain scientific	1

Comment	senti_score_polarity
syntax would be: xtmixed studentscore lagstudentscore l	-1
maybe create time series difference two variables	-1
er "if *link* longer works" keyboard misbehaving droppi	-1
use second model get get coefficient estimate effect sch	-1

CONTENTS



Challenges

01 Creating timeline for project

02 Task allocation among the team

03 Finding the data sources and acquisition methods

Thank You!

Any Questions ?