

HW Assignment 2 - CS267-Fall 2023

Hadoop/MapReduce with Data Cleaning Hands-on Exercises

Name : Sanket Kulkarni (017413461)

Assignment Report Content:

Sr. No	Task Name	Page No.
1	Exercise 1 : Web Crawling	2
2	Exercise 2 : Data Cleaning	5
3	Exercise 3: A	15
4	Exercise 3: B	19
5	Exercise 3: C	24
6	Exercise 3: D	32
7	Outcomes	35
8	References	35

Exercise 1: Data Crawling to Collect Data

Web Crawler: Scraping data from all the three links in one go by taking input of the number of pages we want to scrape.

Links:

<http://www.wikicfp.com/cfp/call?conference=Big%20Data>

<http://www.wikicfp.com/cfp/call?conference=machine%20learning>

<http://www.wikicfp.com/cfp/call?conference=Artificial%20Intelligence>

Source Code Snippets: From calling to function wise

```
 79     # To meet the compliance requirement of the web site
80     time.sleep(10)
81
82     # If no more pages, break
83     else:
84         break
85
86     return all_conference_data
87
88 # Get user input for the number of pages to scrape
89 num_pages_input = int(input("Enter the number of pages to scrape (e.g., 3): "))
90
91 try:
92     # universal list to store data of all conference, we will iterate through this to store entire research areas data
93     all_scraped_data = []
94
95     # Data crawling for ML
96     ml_data = scrape_all_conf_data(num_pages_input, 'ML')
97     all_scraped_data.extend(ml_data)
98
99     # Data Crawling for BigData
100    bigdata_data = scrape_all_conf_data(num_pages_input, 'BigData')
101    all_scraped_data.extend(bigdata_data)
102
103    # Data crawl in for AI
104    ai_data = scrape_all_conf_data(num_pages_input, 'AI')
105    all_scraped_data.extend(ai_data)
106
107    # storing all data in below file, this will be stored in python file directory only
108    output_file = "conference_all.tsv"
109
110    # Write the scraped data to a tab-separated file
111    with open(output_file, mode='w', newline='') as tsv_file:
112        fieldnames = ['Conference Acronym', 'Conference Name', 'Conference Location', 'ResearchArea']
113        writer = csv.DictWriter(tsv_file, fieldnames=fieldnames, delimiter='\t')
114
115        # Write the header row
116        writer.writeheader()
117
118        # Write the data rows
119        for item in all_scraped_data:
120            writer.writerow({ 'Conference Acronym': item['Conference Acronym'],
```

Sanket Kulkarni : HW2

The image shows a terminal window with two code snippets. The top snippet is a Python script named `web_crawler.py` which uses BeautifulSoup to scrape conference data from three different websites based on research area. The bottom snippet is another part of the script, likely `city_to_cf_mapper.py`, which processes the scraped data to extract specific information like conference names, locations, and acronyms.

```
from bs4 import BeautifulSoup # for data scraping
import requests # for hitting the pages with HTTP
import time # for storing info about delay
import csv

# Function to scrape conference data for a given page count and research area
def scrape_all_conf_data(num_pages, research_area):
    # Initialize a list which store data of all the conferences for a particular research area
    all_conference_data = []

    # Define the base URL for the selected research area
    if research_area.lower() == 'ml':
        base_url = "http://www.wikicfp.com/cfp/call?conference=Machine%20Learning"
    elif research_area.lower() == 'bigdata':
        base_url = "http://www.wikicfp.com/cfp/call?conference=Big%20Data"
    elif research_area.lower() == 'ai':
        base_url = "http://www.wikicfp.com/cfp/call?conference=Artificial%20Intelligence"
    else:
        raise ValueError("Invalid research area. Please choose 'ML', 'BigData', or 'AI'.")

    # Loop through the specified number of pages
    for page_number in range(1, num_pages + 2):
        # URL for the current page based on page number in the loop
        url = f'{base_url}&page={page_number}'

        # Send an HTTP GET request to the URL
        response = requests.get(url)

        # if we get 200 OK response from the HTTP requests
        if response.status_code == 200:
            # parse the html content
            html = response.text
            # initialisation of soup lib
            soup = BeautifulSoup(html, 'html.parser')

            # The data on the website is in rows and rows are identified based on the color of the row.
            line = soup.find_all('tr', bgcolor=['#f6f6f6', "#e6e6e6"])

            # Initialize a list to store the extracted data for the current page
            conference_data = []

            # Loop through the rows and extract the information (similar to previous code)
            for row in line:
                # Extract the conference acronym name directly from the event name
                conference_acronym = row[0].text

                # Store conf acronym if it exists
                if conference_acronym:
                    conference_data.append({
                        'Conference Acronym': conference_acronym,
                        'Conference Name': row[1].text,
                        'Conference Location': row[2].text,
                        'Research Area': research_area
                    })

            # Appending all the data to one list
            all_conference_data.extend(conference_data)

        # To meet the compliance requirement of the web site
        time.sleep(10)

    # If no more pages, break
    else:
        break

return all_conference_data
```

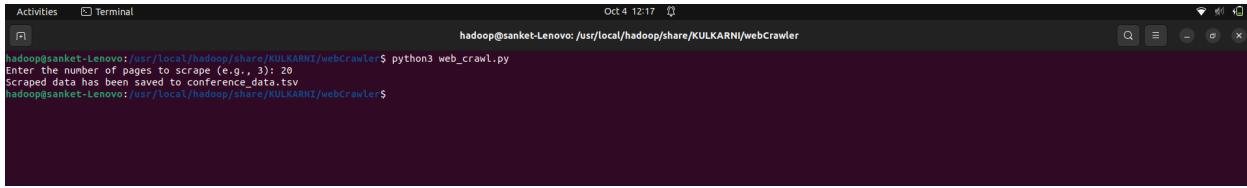
Asking for page inputs:

The terminal window shows the command `python3 web_crawler.py` being run. The user is prompted to enter the number of pages to scrape, with the value '20' being entered.

```
hadoop@sankey-Lenovo: /usr/local/hadoop/share/KULKARNI/webCrawler$ python3 web_crawler.py
Enter the number of pages to scrape (e.g., 3): 20
```

Result after web crawler ran:

Sanket Kulkarni : HW2



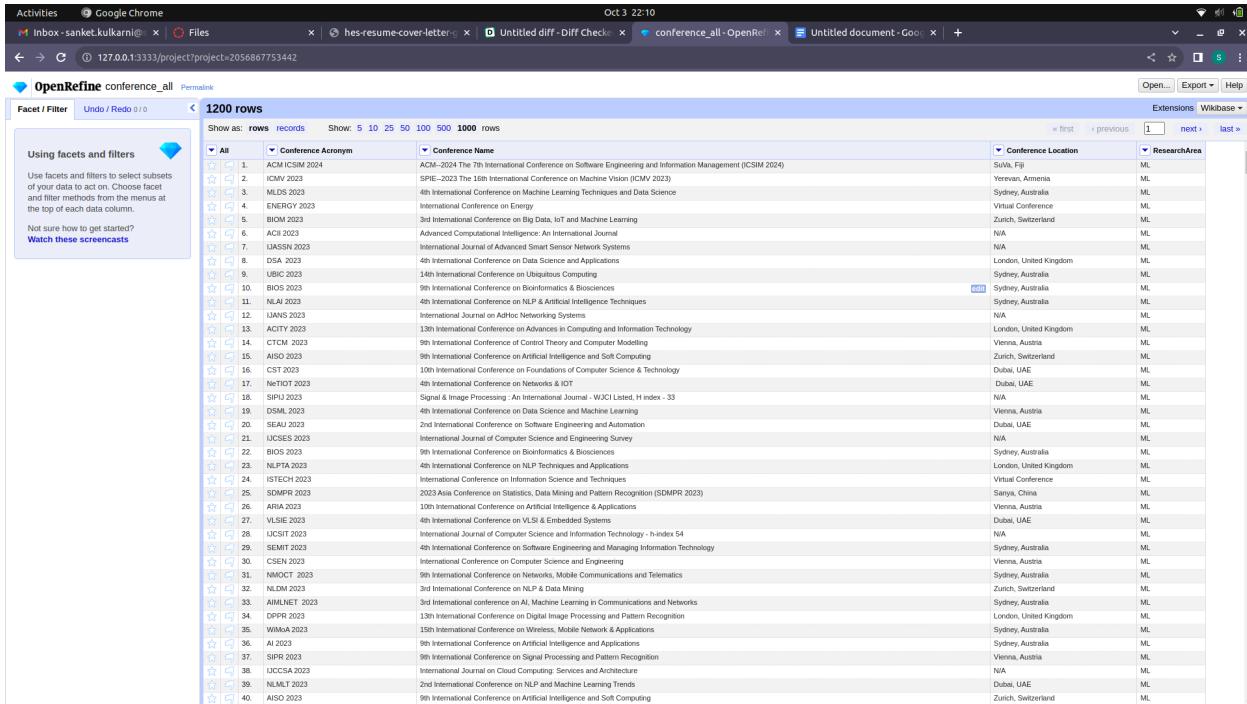
```
Oct 4 12:17
hadoop@saneket-Lenovo:/usr/local/hadoop/share/KULKARNI/webCrawler$ python3 web_crawl.py
Enter the number of pages to scrape (e.g., 3): 20
Scraped data has been saved to conference_data.tsv
hadoop@saneket-Lenovo:/usr/local/hadoop/share/KULKARNI/webCrawler$
```

Web Crawler Working:

- Take the input about how many pages you want to scrape
- I have called web crawler function three times in code to store the data of all research areas
- Used the ‘beautiful soap’ library of python to scrape the data.
- To install beautiful soap I used below command:
- Pip install bs4
- Pip install requests
- Loop through all the research areas one by one and scrape data for it.
- Add a delay of 8-10 seconds to comply with policies of wikicfp
- Now, looping through the table for 20 pages and extracting the rows based on colors
- For each page we have a list, append this to total list
- This total list will then further appended to overall research list
- The output is stored as a tsv file

The output file looks like this:

Generated Conference Data for all 1200 rows for 20 pages per research area (AI/ML/BigData)



1200 rows			
Show as:	rows	records	Show: 5 10 25 50 100 500 1000 rows
All	Conference Acronym	Conference Name	Conference Location
1. ACM-CSIM 2024	ACM-2024 The 7th International Conference on Software Engineering and Information Management (CSIM 2024)	Suva, Fiji	ML
2. ICVM 2023	SPE-2023 The 16th International Conference on Machine Vision (ICMV 2023)	Yerevan, Armenia	ML
3. MLDS 2023	4th International Conference on Machine Learning Techniques and Data Science	Sydney, Australia	ML
4. ENERGY 2023	International Conference on Energy	Virtual Conference	ML
5. BIOM 2023	3rd International Conference on Big Data, IoT and Machine Learning	Zurich, Switzerland	ML
6. ACR 2023	Advanced Computational Intelligence: An International Journal	N/A	ML
7. IJASSN 2023	International Journal of Advanced Smart Sensor Network Systems	N/A	ML
8. DSA 2023	4th International Conference on Data Science and Applications	London, United Kingdom	ML
9. UBIK 2023	14th International Conference on Ubiquitous Computing	Sydney, Australia	ML
10. BIOS 2023	9th International Conference on Bioinformatics & Biosciences	Sydney, Australia	ML
11. NLAD 2023	4th International Conference on NLP & Artificial Intelligence Techniques	Sydney, Australia	ML
12. ICNS 2023	International Journal on Advances in Networking Systems	N/A	ML
13. ACTY 2023	13th International Conference on Advances in Computing and Information Technology	London, United Kingdom	ML
14. CTCM 2023	9th International Conference of Control Theory and Computer Modelling	Vienna, Austria	ML
15. AIGO 2023	9th International Conference on Artificial Intelligence and Soft Computing	Zurich, Switzerland	ML
16. CST 2023	10th International Conference on Foundations of Computer Science & Technology	Dubai, UAE	ML
17. NetTIC 2023	4th International Conference on Networks & IoT	Dubai, UAE	ML
18. SIPU 2023	Signal & Image Processing - An International Journal - WUCI Listed, H index - 33	N/A	ML
19. DSME 2023	4th International Conference on Data Science and Machine Learning	Vienna, Austria	ML
20. SEAU 2023	2nd International Conference on Software Engineering and Automation	Dubai, UAE	ML
21. IUCSES 2023	International Journal of Computer Science and Engineering Survey	N/A	ML
22. BIOS 2023	9th International Conference on Bioinformatics & Biosciences	Sydney, Australia	ML
23. NLPTA 2023	4th International Conference on NLP Techniques and Applications	London, United Kingdom	ML
24. ISTECH 2023	International Conference on Information Science and Techniques	Virtual Conference	ML
25. SDMPR 2023	2023 Asia Conference on Statistics, Data Mining and Pattern Recognition (SDMPR 2023)	Sanya, China	ML
26. ARIA 2023	10th International Conference on Artificial Intelligence & Applications	Vienna, Austria	ML
27. VLSIE 2023	4th International Conference on VLSI & Embedded Systems	Dubai, UAE	ML
28. IUCSIT 2023	International Journal of Computer Science and Information Technology - h-index 54	N/A	ML
29. SEMIT 2023	4th International Conference on Software Engineering and Managing Information Technology	Sydney, Australia	ML
30. CSEN 2023	International Conference on Computer Science and Engineering	Vienna, Austria	ML
31. NWICT 2023	9th International Conference on Networks, Mobile Communications and Telematics	Sydney, Australia	ML
32. NLDM 2023	3rd International Conference on NLP & Data Mining	Zurich, Switzerland	ML
33. AIHNLNET 2023	3rd International conference on AI, Machine Learning in Communications and Networks	Sydney, Australia	ML
34. DPPR 2023	13th International Conference on Digital Image Processing and Pattern Recognition	London, United Kingdom	ML
35. WiMa 2023	15th International Conference on Wireless, Mobile Network & Applications	Sydney, Australia	ML
36. AI 2023	9th International Conference on Artificial Intelligence and Applications	Sydney, Australia	ML
37. SIPR 2023	9th International Conference on Signal Processing and Pattern Recognition	Vienna, Austria	ML
38. IJCCSA 2023	International Journal on Cloud Computing: Services and Architecture	N/A	ML
39. NLMLT 2023	2nd International Conference on NLP and Machine Learning Trends	Dubai, UAE	ML
40. AISO 2023	9th International Conference on Artificial Intelligence and Soft Computing	Zurich, Switzerland	ML

Exercise 2: Data Cleaning using Open Refine

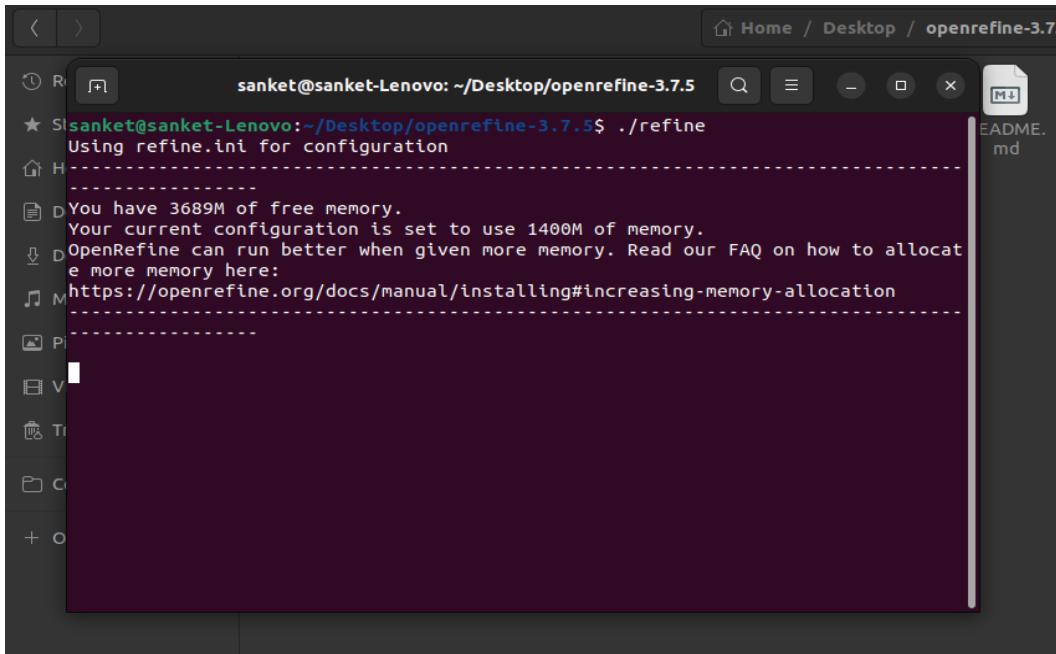
Download Open Refine from google:

The screenshot shows a dark-themed web browser window with multiple tabs open at the top. The active tab is 'openrefine.org/download'. The main content area displays the 'Download OpenRefine' page. It features a prominent blue button labeled 'Download OpenRefine 3.7.5 for Linux' with a downward arrow icon. Below this, a note states 'TAR.GZ file, requires additional Java install'. A link to 'Privacy notice - Release notes' is also present. Further down, a note says 'You also can install OpenRefine on Ubuntu/Debian derivatives with sudo apt install openrefine'. A section titled 'Other platforms and versions' contains a table with three rows, each representing a different version of OpenRefine (3.7.5, 3.6.2, 3.5.2) and listing download links for Windows (including Java), Windows (without Java), Mac OS, and Linux. At the bottom of the page, a link to 'See the full list of releases on GitHub' is provided.

Unzip the folder

Once done, open a terminal from this folder

To run on terminal type `./refine`



This will open a localhost window on to your browser:

Sanket Kulkarni : HW2

Chose the file that we generated from web crawler in exercise 1

It will generate a preview, which would look like this:

Conference Acronym	Conference Name	Conference Location	ResearchArea
1. ACM ICSIM 2024	ACM—2024 The 7th International Conference on Software Engineering and Information Management (ICSIM 2024)	Sarajevo, Bosnia and Herzegovina	ML
2. ICMV 2023	SPIE—2023 The 10th International Conference on Machine Vision (ICMV 2023)	Yerevan, Armenia	ML
3. MLDS 2023	4th International Conference on Machine Learning Techniques and Data Science	Sydney, Australia	ML
4. ENERGY 2023	International Conference on Energy	Virtual Conference	ML
5. BIOM 2023	3rd International Conference on Big Data, IoT and Machine Learning	Zurich, Switzerland	ML
6. ACII 2023	Advanced Computational Intelligence: An International Journal	N/A	ML
7. UJASSN 2023	International Journal of Advanced Smart Sensor Network Systems	N/A	ML
8. DSA 2023	4th International Conference on Data Science and Applications	London, United Kingdom	ML
9. UBiC 2023	14th International Conference on Ubiquitous Computing	Sydney, Australia	ML
10. BIOS 2023	9th International Conference on Bioinformatics & Biosciences	Sydney, Australia	ML
11. NLAI 2023	4th International Conference on NLP & Artificial Intelligence Techniques	Sydney, Australia	ML
12. IJANS 2023	International Journal on Adhoc Networking Systems	N/A	ML
13. ACITY 2023	13th International Conference on Advances in Computing and Information Technology	London, United Kingdom	ML
14. CTCM 2023	9th International Conference of Control Theory and Computer Modelling	Vienna, Austria	ML
15. AISO 2023	9th International Conference on Artificial Intelligence and Soft Computing	Zurich, Switzerland	ML
16. CST 2023	10th International Conference on Foundations of Computer Science & Technology	Dubai, UAE	ML
17. NetIOT 2023	4th International Conference on Networks & IoT	Dubai, UAE	ML
18. SIPU 2023	Signal & Image Processing : An International Journal - WJCI Listed, H index - 33	N/A	ML
19. DSML 2023	4th International Conference on Data Science and Machine Learning	Vienna, Austria	ML
20. SEAU 2023	2nd International Conference on Software Engineering and Automation	Dubai, UAE	ML
21. IJCES 2023	International Journal of Computer Science and Engineering Survey	N/A	ML
22. BIOS 2023	9th International Conference on Bioinformatics & Biosciences	Sydney, Australia	ML
23. NLPTA 2023	4th International Conference on NLP Techniques and Applications	London, United Kingdom	ML
24. ISTECH 2023	International Conference on Information Science and Techniques	Virtual Conference	ML
25. SDMPR 2023	2023 Asia Conference on Statistics, Data Mining and Pattern Recognition (SDMPR 2023)	Sanya, China	ML
26. ARIA 2023	10th International Conference on Artificial Intelligence & Applications	Vienna, Austria	ML
27. VLSIE 2023	4th International Conference on VLSI & Embedded Systems	Dubai, UAE	ML
28. DCIST 2023	International Journal of Computer Science and Information Technology - h-index 54	N/A	ML
29. SEMIT 2023	4th International Conference on Software Engineering and Managing Information Technology	Sydney, Australia	ML
30. CSEN 2023	International Conference on Computer Science and Engineering	Vienna, Austria	ML

Sanket Kulkarni : HW2

Apply a text filter on Conference Location column to remove all N/A locations:
 This generated 151 rows and now select all of them and remove

Conference Name	Conference Location	ResearchArea
Advanced Computational Intelligence: An International Journal	N/A	ML
International Journal of Advanced Smart Sensor Networks	N/A	ML
International Journal on Adhoc Networking Systems	N/A	ML
Signal & Image Processing: An International Journal - WJCI Listed, H index - 33	N/A	ML
International Journal of Computer Science and Engineering Survey	N/A	ML
International Journal of Computer Science and Information Technology - h-index 54	N/A	ML
International Journal on Cloud Computing: Services and Architecture	N/A	ML
Advanced Nanoscience and Technology: An International Journal	N/A	ML
International Journal of Data Mining & Knowledge Management Process	N/A	ML
Machine Learning and Applications: An International Journal	N/A	ML
International Journal of Artificial Intelligence & Applications - H index - 35	N/A	ML
International Journal on Soft Computing	N/A	ML
International Journal of Managing Information Technology - WJCI Indexed, H index 24	N/A	ML
International Journal on Soft Computing, Artificial Intelligence and Applications	N/A	ML
International Journal of Soft Computing, Mathematics and Control	N/A	ML
International Journal of Business Information Systems Strategies	N/A	ML
Special Issue on Cloud Computing and Federated Learning for Smart Cities	N/A	ML
EcoLogy: An International Journal	N/A	ML
Advances in Interconnected Technologies: An International Journal	N/A	ML
Journal of Political Science (JPS)	N/A	ML
International Journal of Modeling, Simulation and Applications	N/A	ML
Advances in Engineering: An International Journal	N/A	ML
International Journal of Law	N/A	ML
Pharmaceutical and Biomedical Sciences: An International Journal	N/A	ML
International Journal of Advanced Dermatology	N/A	ML
Aerospace Engineering: An International Journal	N/A	ML
Earth Sciences: An International Journal	N/A	ML
SI AML CP 2024	N/A	ML
SPECIAL ISSUE on Advancements in Machine Learning for Cybersecurity and Privacy: Algorithms, Models, and Applications	N/A	ML
Super-resolution for remote sensing (Springer, 2024)	N/A	ML
Special Issue on Data Science Methods in Big Data Era in Applied Sciences	N/A	ML
Optimization and Machine Learning in Medical Image Analysis	N/A	ML
Network (MDPI) Special Issue - Blockchain and Machine Learning for IoT Security and Privacy Challenges	N/A	ML
Special Issue on MACHINE LEARNING IN TOURISM - Int. J. of Machine Learning and Cybernetics (Springer)	N/A	ML
SI on Machine Learning-Based Spectrum Occupancy Prediction and Resource Allocation/Management for Wireless Communication Systems	N/A	ML
Explainable AI for Health	N/A	ML
Advanced Medical Sciences: An International Journal	N/A	ML
Emerging Trends in Electrical, Electronics & Instrumentation Engineering: An International Journal	N/A	ML
Mechanical Engineering: An International Journal	N/A	ML
Special Issue of Information Fusion (Elsevier): New Trends of Adversarial Machine Learning for Data Fusion and Intelligent System	N/A	ML
The International Journal of Computational Science, Information Technology and Control Engineering	N/A	ML

Remove leading and trailing spaces in all columns

Conference Name	Conference Location	ResearchArea
ion Management (ICSIM 2024)	ML	ML
	ML	ML
	ML	ML
	Transform...	
	Common transforms	
	Fill down	
	Blank down	
	Split multi-valued cells...	
	Join multi-valued cells...	
	Cluster and edit...	
	Replace...	
2023)	Sanya, China	ML
	Vienna, Austria	ML
	Dubai, UAE	ML
	Sydney, Australia	ML
chnology	Vienna, Austria	ML
	Sydney, Australia	ML

Sanket Kulkarni : HW2

Removing all rows with location as online, you can filter these rows using text filter in column drop down

The screenshot shows the OpenRefine interface with a list of conference records. A facet filter 'Conference Location' is applied with the value 'online'. The results show 1049 rows. The columns include Conference Acronym, Conference Name, Conference Location, and ResearchArea. Many rows have 'Online' listed under Conference Location.

Removing all rows with location as virtual, you can filter these rows using text filter in column drop down

The screenshot shows the OpenRefine interface with a list of conference records. A facet filter 'Conference Location' is applied with the value 'virtual'. The results show 1029 rows. The columns include Conference Acronym, Conference Name, Conference Location, and ResearchArea. Most rows have 'Virtual' listed under Conference Location.

Unselecting the hybrid locations rows from these as guided by the professor

Conference Acronym	Conference Name	Conference Location	ResearchArea
K.2021	The 23rd International Conference on Big Data Analytics and Knowledge Discovery	Linz, Austria (Virtual)	BigData
2021	Intelligent Data Engineering and Automated Learning	Manchester (Virtual)	BigData
3.2023	The Fourth International Workshop on Big Data Reduction	Sorrento, Italy	BigData
TRANT 2023	Special Session on Handling Resource constraints for using ML	Virtual	ML
Star rows	Session on Machine Learning for Graphs	Virtual	ML
Unstar rows	4 Systems in Forensic Engineering	Virtual	ML
Flag rows	International Workshop on Big Data Reduction held with 2022 IEEE International Conference on Big Data	Virtual	BigData
Unflag rows	IEEE International Workshop on Benchmarking, Performance Tuning and Optimization for Big Data Applications	Virtual	BigData
BTSO	International Workshop on Big Data Tools, Methods, and Use Cases for Innovative Scientific Discovery (BTSO) 2022	Virtual	BigData
IWBD	International Workshop on Big Data Reduced held with 2021 IEEE International Conference on Big Data	Virtual	BigData
LegalNet 2022	Legend 2022 - The Fifth Annual Workshop on Applications of Artificial Intelligence in the Legal Industry	Virtual	BigData
IEEE Big Data - BPD021	The Fifth IEEE International Workshop on Benchmarking, Performance Tuning and Optimization for Big Data Applications	Virtual	BigData
BFNDMA 2021	BIG FOOD, NUTRITION AND ENVIRONMENT DATA MANAGEMENT AND ANALYSIS @ IEEE BigData 2021	Virtual	BigData
BFTSD 2023	The 3rd International Workshop on Big Data Tools, Methods, and Use Cases for Innovative Scientific Discovery (BTSO) 2021	Virtual	BigData
CPS-BigData 2021	The 3rd IEEE International Workshop on Application of Big Data Analytics to Cyber-Physical Systems	Virtual	BigData
WITCOM 2021	VIRTUAL 10th Conference in Computer and Telematics	Virtual	BigData
CPF for IEEE Big Data Service 2021	The 7th IEEE International Conference on Big Data Computing Service and Machine Learning Applications	Virtual	BigData
SCSN 2021	The 9th IEEE International Workshop on Semantic Computing for Social Networks and Organization Sciences: from user information to social knowledge	Virtual	BigData
ACG 2023	Advances in Computer Games	Virtual	AI
DaSET 2023	The 2nd International Conference on Data Science and Emerging Technologies 2023	Virtual	AI
CONSTRAINT 2023	Special Session on Handling Resource constraints for using ML	Virtual	AI
SNTA 2021	The Fourth International Workshop on Systems and Network Telemetry and Analytics (SNTA 2021)	Virtual (in conjunction with ACM HPCA)	BigData
ICNS 2022	International Conference on Energy	Virtual Conference	ML
ENCODE 2022	International Conference on Information Science and Techniques	Virtual Conference	ML
ISTECH 2023	International Conference on Information Theory and Machine Learning	Virtual Conference	ML
ISCI 2023	International Conference on Information Theory and Machine Learning	Virtual Conference	ML
CORAI 2023	International Conference on Operations Research and Applications	Virtual Conference	ML
CHENG 2024	International Conference on Advances in Chemistry & Chemical Engineering	Virtual Conference	ML
LIBI 2024	International Conference on Life Sciences	Virtual Conference	ML
HAS 2024	International Conference on Humanities, Arts and Social Studies	Virtual Conference	ML
BIOEN 2024	7th International Conference on Biomedical Engineering and Science	Virtual Conference	ML
NLDNI 2024	3rd International Conference on NLP, Data Mining and Machine Learning	Virtual Conference	ML
MIRIO 2024	3rd International Conference of Multidisciplinary & Interdisciplinary Bioscience	Virtual Conference	ML
CAMS 2023	International Conference on Clinical and Medical Sciences	Virtual Conference	ML
NLAI 2023	International Conference on NLP & AI Information Retrieval	Virtual Conference	ML
ENERGY 2023	International Conference on Energy	Virtual Conference	BigData
CEU 2024	7th International Conference on Civil Engineering and Urban Planning	Virtual Conference	BigData
ACINT 2023	International Conference on Advanced Computational Intelligence	Virtual Conference	BigData
CORAI 2023	International Conference on Operations Research and Applications	Virtual Conference	BigData
CSITAI 2023	International Conference on Computer Science, Information Technology & AI	Virtual Conference	BigData
CSITAI 2023	International Conference on Computer Science, Information Technology & AI	Virtual Conference	BigData

Found below rows for hybrid and cleaning the data manually here, just keeping the city and country and not the hybrid word for the further processing

Conference Acronym	Conference Name	Conference Location	ResearchArea
hybrid	IEEE Big Data AMAG 2023	Hybrid	ML
hybrid	ICDM NeuRec Workshop 2023	Shanghai, China (hybrid)	ML
hybrid	ICDM 2023	Shanghai, China (hybrid)	ML
256.	EuroSymposium 2023	Hybrid (Virtual, Sopot, Poland)	ML
407.	SCSN 2024	Laguna Hills, CA, USA (Hybrid)	BigData
495.	ABCSS 2023	Hybrid Conference Venice, Italy	BigData
518.	IEEE COINS 2023	Hybrid (Virtual, Sopot, Poland)	BigData
539.	SCSN 2023	Hybrid Event (Berlin, Germany Virtual)	BigData
581.	IDEAL 2022	Laguna Hills, CA, USA (Hybrid)	BigData
598.	WITCOM 2022	Manchester UK (hybrid)	BigData
873.	SCSN 2024	Virtual & Hybrid	BigData
873.	SCSN 2024	Laguna Hills, CA, USA (Hybrid)	AI

Sanket Kulkarni : HW2

After cleaning , we have two rows whose location is not mentioned so flagging those

Conference Location	Conference Acronym	Conference Name	Conference Location	ResearchArea
hybrid	IEEE Big Data AMG 2023	IEEE Big Data 2023 Workshop on AI Music Generation	Laguna Hills, USA	BigData
hybrid	SCSN 2024	The 12th IEEE International Workshop on Semantic Computing for Social Networks and Organization Sciences: from user information to social knowledge	Venice, Italy	BigData
hybrid	ABCSS 2023	The 8th International Workshop on Application of Big Data for Computational Social Science	Sopot, Poland	BigData
hybrid	EuroSymposium 2023	EuroSymposium 2023. The Xth EuroSymposium on Digital Transformation, 28th of September, 2023, Hybrid, University of Gdańsk, Poland (LNBPjP proceedings)	Berlin, Germany	BigData
hybrid	IEEE COWS 2023	IEEE COWS 2023 - Berlin, Germany - July 23-25 - Hybrid (In-Person & Virtual) Artificial Intelligence, Internet of Things (IoT), Blockchain, Big Data, Machine Learning	Laguna Hills, USA	BigData
hybrid	SCSN 2023	The 11th IEEE International Workshop on Semantic Computing for Social Networks and Organization Sciences: from user information to social knowledge	Manchester, UK	BigData
hybrid	IDEAL 2022	23rd International Conference on Intelligent Data Engineering and Automated Learning	Virtual & Hybrid 11th Conference in Telematics and Computing, WITCOM	Virtual & Hybrid
hybrid	WITCOM 2022	Virtual & Hybrid 11th Conference in Telematics and Computing, WITCOM	Laguna Hills, USA	BigData
hybrid	SCSN 2024	The 12th IEEE International Workshop on Semantic Computing for Social Networks and Organization Sciences: from user information to social knowledge	Laguna Hills, USA	AI

Removing the flagged two rows

Conference Location	Conference Acronym	Conference Name	Conference Location	ResearchArea
hybrid	Big Data AIMG 2023	IEEE Big Data 2023 Workshop on AI Music Generation	Hybrid	ML
hybrid	WITCOM 2022	Virtual & Hybrid 11th Conference in Telematics and Computing , WITCOM	Virtual & Hybrid	BigData

A context menu is open on the second row, showing options: Transform..., Edit all columns, Facet, Edit rows, Unstar rows, Flag rows, Unflag rows, and Remove matching rows.

Sanket Kulkarni : HW2

Removing few more rows with no city mentioned and only country is given

Conference Acronym	Conference Name	Conference Location	ResearchArea
241. SFGO 2023	11èmes Journées de la Société Africaine de Chimie Informatique	CERN	ML
635. Blockchain 2021	2021 International Conference on Blockchain and Trustworthy Systems	China	BigData
156. CDEM 2023	Consel Financial Engineering 2023 Conference Call For Papers	Cornell School of Financial Engineering	ML
414. CDEM 2023	Consel Financial Engineering 2023 Conference Call For Papers	Cornell School of Financial Engineering	BigData
919. CDEM 2023	Consel Financial Engineering 2023 Conference Call For Papers	Cornell School of Financial Engineering	AI
313. IMSA 2023	IEEE Conference on the Intelligent Methods, Systems, and Applications (IMSA)	Egypt	ML
787. ITK 2024	Interaktiivinen Teknologia Konferenssia	Hämeenlinna, Finland	AI

Making changes to university location as city and country

Conference Acronym	Conference Name	Conference Location	ResearchArea
592. ICFC 2021	2021 International Conference on Future Intelligent Computing	Dengu University, Korea	BigData
366. ICICT 2023	ACM-2023 The 11th International Conference on Information Technology, IoT and Smart City (ICIT 2023)	Kyoto, Japan	BigData
642. RDAPS 2021	Reconciling Data Analytics, Automation, Privacy, and Security: A Big Data Challenge	Hamilton, Canada	BigData
827. ICRCR 2024	IEEE-2024 9th International Conference on Control and Robotics Engineering (ICCRE 2024)	Osaka University, Osaka, Japan	AI
324. ICBDSC 2024	2024 the 7th International Conference on Big Data and Smart Computing (ICBDSC 2024)	The University of Fiji	BigData
908. ICOMA 2023	IEEE-2023 The 11th International Conference on Control, Mechatronics and Automation (ICMA 2023)	University of Agder, Norway	AI
364. IEEE ICM 2024	IEEE-2024 the 10th International Conference on Information Management (ICM 2024)	University of Cambridge, Cambridge, UK	BigData
836. IEEE CCAI 2024	2024 IEEE the 4th International Conference on Computer Communication and Artificial Intelligence (CCAI 2024)	Xidian University, Xian, China	AI

Sanket Kulkarni : HW2

Few locations with Remote space, hence removing those

The screenshot shows the OpenRefine interface with a single row of data. The facet 'Conference Location' has a filter applied for 'Remote'. The row itself shows 'Conference Name' as 'Continual4 Unconference 2023', 'Conference Acronym' as '210...', and 'ResearchArea' as 'ML'.

Splitting the column into 2, separating cities and countries in two columns now as we would feed our mapper a cleaned data of city

The screenshot shows the OpenRefine interface with a context menu open over a column. The 'Edit column' option is selected, and its submenu is visible, with 'Split into several columns...' highlighted. The main table shows various conference names and their locations.

Conference Name	Conference Location	ResearchArea
ference on Software Engineering and Information Management (ICSIM 2024)	Zurich, Switzerland	ML
onference on Machine Vision (ICMV 2023)	Dubai, UAE	ML
hine Learning Techniques and Data Science	Dubai, UAE	ML
Data, IoT and Machine Learning	London, United Kingdom	ML
Science and Applications	Sanya, China	ML
iquitous Computing	Vienna, Austria	ML
nformatics & Biosciences	Dubai, UAE	ML
' & Artificial Intelligence Techniques	Vienna, Austria	ML
vances in Computing and Information Technology	Dubai, UAE	ML
rol Theory and Computer Modelling	Sydney, Australia	ML
cial Intelligence and Soft Computing	London, United Kingdom	ML
undations of Computer Science & Technology	Sanya, China	ML
works & IOT	Vienna, Austria	ML
Science and Machine Learning	Dubai, UAE	ML
ware Engineering and Automation	Vienna, Austria	ML
nformatics & Biosciences	Sydney, Australia	ML
Techniques and Applications	London, United Kingdom	ML
Data Mining and Pattern Recognition (SDMPR 2023)	Sanya, China	ML
ificial Intelligence & Applications	Vienna, Austria	ML
I & Embedded Systems	Dubai, UAE	ML
ware Engineering and Managing Information Technology	Sydney, Australia	ML
er Science and Engineering	Vienna, Austria	ML

9442 records Show: 5 10 25 50 100 500 1000 rows

Conference Acronym

Conference Acronym	Description
ICSM 2024	ACM-2024 The 7th International Conference on Software Engineering and Information Management (ICSM 2024)
SPIE-2023	SPIE-2023 The 16th International Conference on Machine Vision (ICMV 2023)
MLDS 2023	4th International Conference on Machine Learning Techniques and Data Science
BIOM 2023	3rd International Conference on Big Data, IoT and Machine Learning
DSA 2023	4th International Conference on Data Science and Applications
UBIC 2023	14th International Conference on Ubiquitous Computing
BIOS 2023	9th International Conference on Bioinformatics & Biosciences
NLAI 2023	4th International Conference on NLP & Artificial Intelligence Techniques
ACITY 2023	13th International Conference on Advances in Computing and Information Technology
CTCM 2023	9th International Conference of Control Theory and Computer Modeling
ASIO 2023	9th International Conference on Artificial Intelligence and Soft Computing
CST 2023	10th International Conference on Foundations of Computer Science & Technology
NeT-IOT 2023	4th International Conference on Networks & IOT
DISM 2023	4th International Conference on Data Science and Machine Learning
SEAU 2023	2nd International Conference on Software Engineering and Automation
BIOS 2023	9th International Conference on Bioinformatics & Biosciences
NLPA 2023	4th International Conference on NLP Techniques and Applications
SDMPR 2023	2023 Asia Conference on Statistics, Data Mining and Pattern Recognition (SDMPR 2023)
ARIA 2023	10th International Conference on Artificial Intelligence & Applications
VLSIE 2023	4th International Conference on VLSI & Embedded Systems
SEMIT 2023	4th International Conference on Software Engineering and Managing Information Technology
CSEN 2023	International Conference on Computer Science and Engineering
10th International Conference on Artificial Intelligence & Applications	

Split column Conference Location into several columns

How to split column

by separator
 by field lengths

Separator regular expression

Split into columns at most (leave blank for no limit)

After Splitting

Guess cell type
 Remove this column

List of integers separated by commas, e.g., 5, 7, 15

OK Cancel

Result looks like, also renamed city and country as column names

Facet / Filter Undo / Redo 220 / 220 917 rows Show as: rows records Show: 5 10 25 50 100 500 1000 rows

Using facets and filters

Facet / Filter Undo / Redo 221 / 221 917 rows Show as: rows records Show: 5 10 25 50 100 500 1000 rows

Facet / Filter Undo / Redo 221 / 221 917 rows Show as: rows records Show: 5 10 25 50 100 500 1000 rows

Conference Name	City	Country	ResearchArea
ACM-2024 The 7th International Conference on Software Engineering and Information Management (ICSM 2024)	Silvia	Fiji	ML
SPIE-2023 The 16th International Conference on Machine Vision (ICMV 2023)	Yerevan	Armenia	ML
4th International Conference on Machine Learning Techniques and Data Science	Sydney	Australia	ML
3rd International Conference on Big Data, IoT and Machine Learning	Zurich	Switzerland	ML
4th International Conference on Data Science and Applications	London	United Kingdom	ML
14th International Conference on Ubiquitous Computing	Sydney	Australia	ML
9th International Conference on Bioinformatics & Biosciences	Sydney	Australia	ML
4th International Conference on NLP & Artificial Intelligence Techniques	Sydney	Australia	ML
13th International Conference on Advances in Computing and Information Technology	London	United Kingdom	ML
9th International Conference of Control Theory and Computer Modeling	Vienna	Austria	ML
9th International Conference on Artificial Intelligence and Soft Computing	Zurich	Switzerland	ML
10th International Conference on Foundations of Computer Science & Technology	Dubai	UAE	ML
4th International Conference on Networks & IOT	Dubai	UAE	ML
4th International Conference on Data Science and Machine Learning	Vienna	Austria	ML
2nd International Conference on Software Engineering and Automation	Dubai	UAE	ML
9th International Conference on Bioinformatics & Biosciences	Sydney	Australia	ML
4th International Conference on NLP Techniques and Applications	London	United Kingdom	ML
2023 Asia Conference on Statistics, Data Mining and Pattern Recognition (SDMPR 2023)	Sanya	China	ML
10th International Conference on Artificial Intelligence & Applications	Vienna	Austria	ML
4th International Conference on VLSI & Embedded Systems	Dubai	UAE	ML
4th International Conference on Software Engineering and Managing Information Technology	Sydney	Australia	ML
International Conference on Computer Science and Engineering	Vienna	Austria	ML

Splitted acronym in two columns, so that we have year separated from the acronym name

Facet / Filter Undo / Redo 221 / 221 917 rows Show as: rows records Show: 5 10 25 50 100 500 1000 rows

Using facets and filters

Facet / Filter Undo / Redo 221 / 221 917 rows Show as: rows records Show: 5 10 25 50 100 500 1000 rows

Facet / Filter Undo / Redo 221 / 221 917 rows Show as: rows records Show: 5 10 25 50 100 500 1000 rows

Conference Acronym	City	Country	ResearchArea
ACM-2024	Silvia	Fiji	ML
SPIE-2023	Yerevan	Armenia	ML
MLDS 2023	Sydney	Australia	ML
BIOM 2023	Zurich	Switzerland	ML
DSA 2023	London	United Kingdom	ML
UBIC 2023	Sydney	Australia	ML
BIOS 2023	Sydney	Australia	ML
NLAI 2023	Sydney	Australia	ML
ACITY 2023	London	United Kingdom	ML
CTCM 2023	Vienna	Austria	ML
ASIO 2023	Zurich	Switzerland	ML
CST 2023	London	United Kingdom	ML
NeT-IOT 2023	Vienna	Austria	ML
DISM 2023	Zurich	Switzerland	ML
SEAU 2023	Dubai	UAE	ML
BIOS 2023	Dubai	UAE	ML
NLPA 2023	Sydney	Australia	ML
SDMPR 2023	London	United Kingdom	ML
ARIA 2023	Sanya	China	ML
VLSIE 2023	Vienna	Austria	ML
SEMIT 2023	Dubai	UAE	ML
CSEN 2023	Sydney	Australia	ML
International Conference on Computer Science and Engineering	Vienna	Austria	ML

Split column Conference Acronym into several columns

How to split column

by separator
 by field lengths

Separator regular expression

Split into columns at most (leave blank for no limit)

After Splitting

Guess cell type
 Remove this column

List of integers separated by commas, e.g., 5, 7, 15

OK Cancel

Sanket Kulkarni : HW2

Once Splitting Done, I added formatting to the year by using replace and year now has only four digits resembling to its value

917 rows				Extensions Wikibase							
Show as:	rows	records	Show:	5	10	25	50	100	500	1000	rows
All	Conference Acronym	Year	Conference Name								
1.	ACM ICSIM	Facet	ie 7th International Conference on Software Engineering and Information Management (ICSIM 2024)								
2.	ICMV	Text filter	ie 16th International Conference on Machine Vision (ICMV 2023)								
3.	MLDS	Edit cells	4th International Conference on Machine Learning Techniques and Data Science								
4.	BIOM	Edit column	Transforms...								
5.	DSA	Transpose	Common transforms								
6.	DSML		Fill down								
7.	BIOS		Sort...								
8.	NLAI		Blank down								
9.	ACTIV		View								
10.	CTCM		Split multi-valued cells...								
11.	AISO		Reconcile								
12.			Join multi-valued cells...								
13.	CST		Cluster and edit...								
14.	NeTIOT		Replace...								
15.	DSML										
16.	SEAU										
17.	BIOS										
18.	NLPTA										
19.	SOMPR										
20.	ARIA										
21.	VLSIE										
22.	SEMIT										
	CSEN										

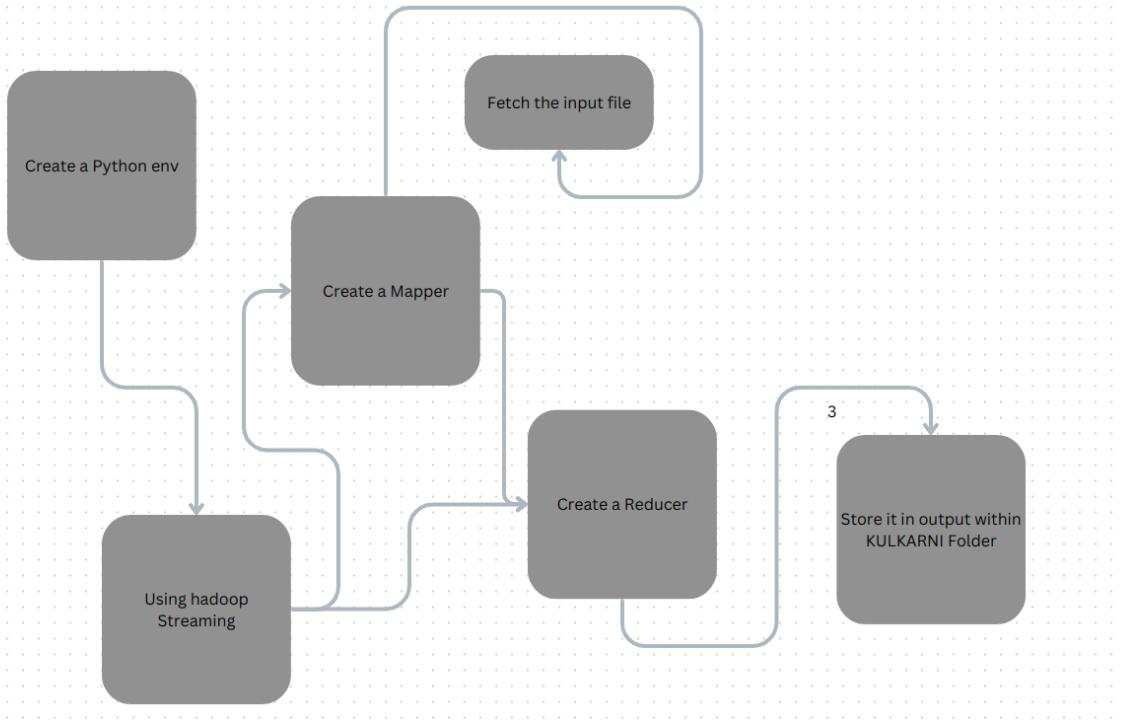
Used the join command to add the conferences with year and downloading this dataset for question 2

7 rows											
Show as:	rows	records	Show:	5	10	25	50	100	500	1000	rows
All	Conference Acronym	Year	Conference Name								
1.	Facet	2024	ACM--2024 The 7th International Conference on Software Engineering and Information Management (ICSIM 2024)								
2.	Text filter	2023	SPIE--2023 The 16th International Conference on Machine Vision (ICMV 2023)								
3.	Edit cells	2023	4th International Conference on Machine Learning Techniques and Data Science								
4.	Edit column	2023	3rd International Conference on Big Data, IoT and Machine Learning								
5.	Transpose	Split into several columns...	Conference on Data Science and Applications								
6.		Join columns...	Conference on Ubiquitous Computing								
7.	Sort...	Add column based on this column...	Conference on Bioinformatics & Biosciences								
8.	View	Add column by fetching URLs...	Conference on NLP & Artificial Intelligence Techniques								
9.		Add columns from reconciled values...	Conference on Advances in Computing and Information Technology								
10.	Reconcile	Rename this column...	Conference of Control Theory and Computer Modelling								
11.	AISO,2023	Remove this column	Conference on Artificial Intelligence and Soft Computing								
12.	CST,2023	Move column to beginning	Conference on Foundations of Computer Science & Technology								
13.	NeTIOT,2023	Move column to end	Conference on Networks & IOT								
14.	DSML,2023	Move column left	Conference on Data Science and Machine Learning								
15.	SEAU,2023	Move column right	Conference on Software Engineering and Automation								
16.	BIOS,2023		Conference on Bioinformatics & Biosciences								
17.	NLPTA,2023		Conference on NLP Techniques and Applications								
18.	SDMPR,2023	2023	2023 Asia Conference on Statistics, Data Mining and Pattern Recognition (SDMPR 2023)								
19.	ARIA,2023	2023	10th International Conference on Artificial Intelligence & Applications								
20.	VLSIE,2023	2023	4th International Conference on VLSI & Embedded Systems								

The downloaded file is tsv file and is stored in the downloads after we export it from OpenRefine. After a lot of data cleaning and manipulations, I was left with around 915 rows, which had proper data in it. As guided by the professor, I focused on removing N/A in locations, spaces leading/trailing, blank / null values, Special characters in name, Made city names same using clustering and few of them manually, Removed university names in locations, acronym splitted into year for processing, etc.

Exercise 3: Hadoop

A) Compute and plot the number of conferences per city. Which are the top 10 locations? (1 Mapper & 1 Reducer)



Initial command to run the mapper and reducer:

```

hadoop@ sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/conference_all $ ls
conference_allmbd.tsv num_conf_percity_mapper.py num_conf_percity_reducer.py
hadoop@ sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/conference_all $ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming*.jar \
-fs num_conf_percity_mapper.py num_conf_percity_reducer.py -mapper "python3 num_conf_percity_mapper.py" -reducer "python3 num_conf_percity_reducer.py" \
-input conference_allmbd.tsv -output num_of_conf_percity
WARNING: All illegal reflective access operations have occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.8.1.jar) to method sun.security.krb5.Config.getInst
andation
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use -illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
23/10/04 00:34:50 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.sessionId
23/10/04 00:34:50 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
23/10/04 00:34:50 INFO jvm.JvmMetrics: Cache initialized JVM Metrics with processName=JobTracker, sessionId= - already initialized
23/10/04 00:34:50 INFO mapreduce.Job: mapred.LocalJobRunner: total number of files to process : 1
23/10/04 00:34:50 INFO mapreduce.JobSubmitter: number of splits:1
23/10/04 00:34:51 INFO mapreduce.Job: Submitting tokens for job: job_local037229615_0001
23/10/04 00:34:51 INFO mapred.LocalDistributedCacheManager: Localized file:/usr/local/hadoop/share/KULKARNI/conference_all/num_conf_percity_mapper.py as file:/tmp/hadoop-hadoop/mapred/local/1696404891483/num_c
f_percity_mapper.py
23/10/04 00:34:51 INFO mapred.LocalDistributedCacheManager: Localized file:/usr/local/hadoop/share/KULKARNI/conference_all/num_conf_percity_reducer.py as file:/tmp/hadoop-hadoop/mapred/local/1696404891484/num_
c_percity_reducer.py
23/10/04 00:34:51 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
23/10/04 00:34:51 INFO mapreduce.Job: Running job: job_local037229615_0001
23/10/04 00:34:51 INFO mapred.LocalJobRunner: OutputCommitter set in config null
23/10/04 00:34:51 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
23/10/04 00:34:51 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
23/10/04 00:34:51 INFO mapred.LocalJobRunner: OutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
23/10/04 00:34:51 INFO mapred.LocalJobRunner: Waiting for map tasks
23/10/04 00:34:51 INFO mapred.LocalJobRunner: Starting task: attempt_local037229615_0001_m_000000_0
23/10/04 00:34:51 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
23/10/04 00:34:51 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
23/10/04 00:34:52 INFO mapred.Task: Using ResourceCalculatorProcessTree: []
23/10/04 00:34:52 INFO mapred.MapTask: Using input file: /usr/local/hadoop/share/KULKARNI/conference_all/conference_allmbd.tsv:0+102318
23/10/04 00:34:52 INFO mapred.MapTask: numReduceTasks: 1
23/10/04 00:34:52 INFO mapred.MapTask: (EQUATOR) 0 kvi_26214396(104857584)
23/10/04 00:34:52 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
23/10/04 00:34:52 INFO mapred.MapTask: soft limit at 83886080
23/10/04 00:34:52 INFO mapred.MapTask: bufstart = 0, bufend = 104857600
23/10/04 00:34:52 INFO mapred.MapTask: Map output collector class: org.apache.hadoop.mapred.MapTask$MapOutputBuffer
23/10/04 00:34:52 INFO mapred.MapTask: Map output collector class: org.apache.hadoop.mapred.MapTask$MapOutputBuffer
23/10/04 00:34:52 INFO streaming.PipeMapRed: PipeMapRed exec [/usr/bin/python3, num.conf.percity.mapper.py]
23/10/04 Configuration.deprecation: map.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
23/10/04 00:34:52 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
23/10/04 00:34:52 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
23/10/04 00:34:52 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
23/10/04 00:34:52 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
23/10/04 00:34:52 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
23/10/04 00:34:52 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skippedrecords
23/10/04 00:34:52 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
23/10/04 00:34:52 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
23/10/04 00:34:52 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
23/10/04 00:34:52 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name

```

Command :

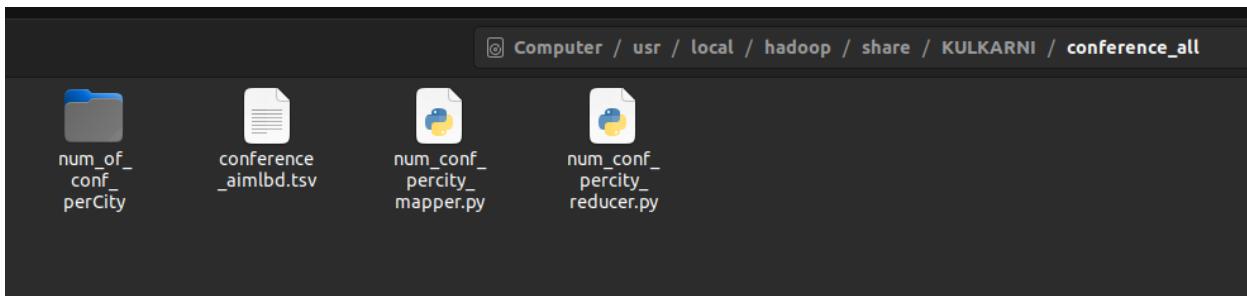
```
hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming*.jar \
-files num_conf_percity_mapper.py,num_conf_percity_reducer.py -mapper "python3
num_conf_percity_mapper.py" -reducer "python3 num_conf_percity_reducer.py" \
-input conference_aimlbd.tsv -output num_of_conf_perCity
```

Successfully ran screenshot:

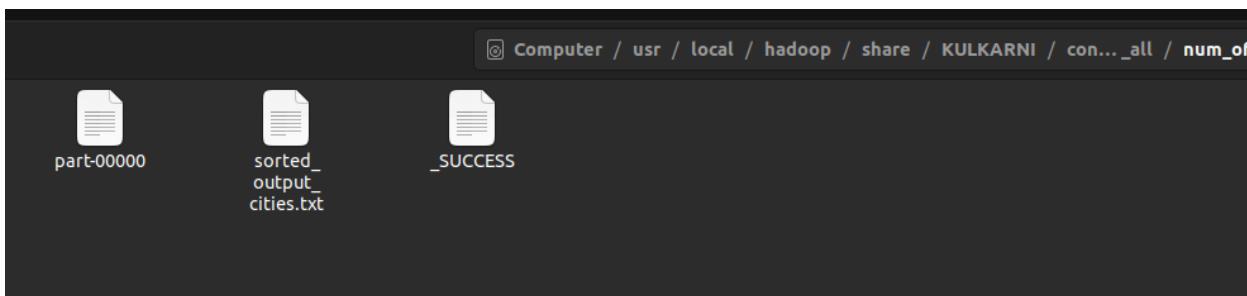
```
hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming*.jar \
-files num_conf_percity_mapper.py,num_conf_percity_reducer.py -mapper "python3
num_conf_percity_mapper.py" -reducer "python3 num_conf_percity_reducer.py" \
-input conference_aimlbd.tsv -output num_of_conf_perCity

23/10/04 00:34:52 INFO mapred.LocalJobRunner: Records R/W-210/1 > reduce
23/10/04 00:34:52 INFO mapred.Task: Task 'attempt_local03729615_0001_r_000000_0' done.
23/10/04 00:34:52 INFO mapred.LocalJobRunner: finishing task: attempt_local03729615_0001_r_000000_0
23/10/04 00:34:52 INFO mapreduce.Job: Job Job_local03729615_0001 running in uber mode : false
23/10/04 00:34:52 INFO mapreduce.Job: map 100% reduce 100%
23/10/04 00:34:52 INFO mapreduce.Job: Job Job_local03729615_0001 completed successfully
23/10/04 00:34:52 INFO mapreduce.Job: counters: 30
File System Counters:
  File system bytes read=499164
  FILE: Number of bytes written=96627
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
Map-Reduce Framework:
  Map output records=918
  Map output bytes=9573
  Map output materialized bytes=11414
  Input split bytes=126
  Combining input records=0
  Data transferred=11414
  Reduce input groups=279
  Reduce shuffle bytes=11414
  Reduce bytes=918
  Reduce output records=279
  Spilled Records=1836
  Failed Records=0
  Failed Shuffles=0
  Merged Map outputs=1
  Total committed heap usage (bytes)=471859200
Shuffle Errors:
  File 0
  CONNECTION=0
  IO_ERROR=0
  NETWORK=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Formats:
  Bytes Read=102318
  File Output Format Counters:
    Bytes Written=3645
23/10/04 00:34:52 INFO streaming.StreamJob: Output directory: num_of_conf_perCity/
hadoop@saneket-Lenovo:/usr/local/hadoop/share/KULKARNI/conference_all$ cd num_of_conf_perCity/
hadoop@saneket-Lenovo:/usr/local/hadoop/share/KULKARNI/conference_all$ hadoop fs -cat part-000* | sort -k2,2nr > sorted_output_cities.txt
part-00000  SUCCESS
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.8.1.jar) to method sun.security.krb5.Config.getInstanc
ANCE()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
hadoop@saneket-Lenovo:/usr/local/hadoop/share/KULKARNI/conference_all$ sudo audit sorted_output_cities.txt
```

Folder Structure where output is generated and program was ran:



Output Location:



Sorted output:

Sanket Kulkarni : HW2

The terminal window shows a sorted list of cities from highest count to lowest. The list includes Singapore (31), Sydney (30), Chengdu (25), Beijing (23), Bangkok (22), Zurich (22), Abu Dhabi (21), Shanghai (20), Laguna Hills (18), Dubai (17), London (17), Vancouver (17), New York (15), Chongqing (12), Lisbon (12), Paris (12), Wuhan (12), Macau (11), Xiamen (10), Istanbul (9), Osaka (9), Sanya (9), Tokyo (9), Madrid (9), Hanoi (8), Copenhagen (8), Nanjing (8), Rome (8), Athens (7), and Barcelona (7). The file path is /usr/local/hadoop/share/KULKARNI/conference_all/_num_of_conf_perCity.

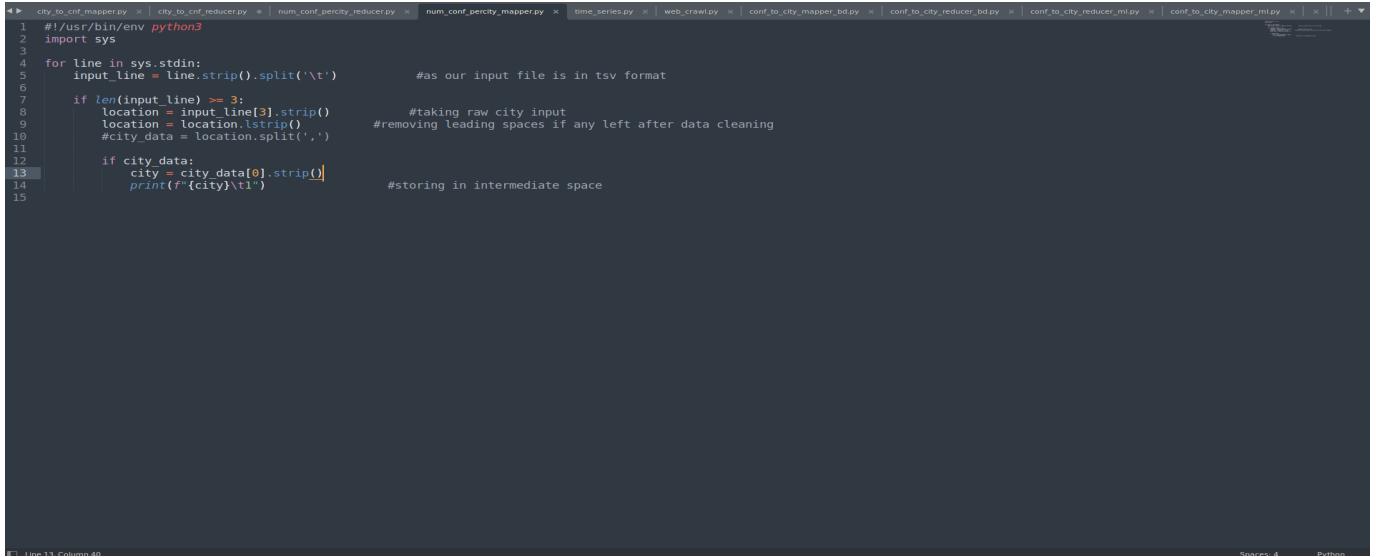
Top 10 locations are:

Sr. NO	City Name	Count
1	Singapore	31
2	Sydney	30
3	Chengdu	25
4	Beijing	23
5	Bangkok	22
6	Zurich	22
7	Abu Dhabi	21
8	Shanghai	20
9	Laguna Hills	18
10	Dubai	17

Code Snippet:

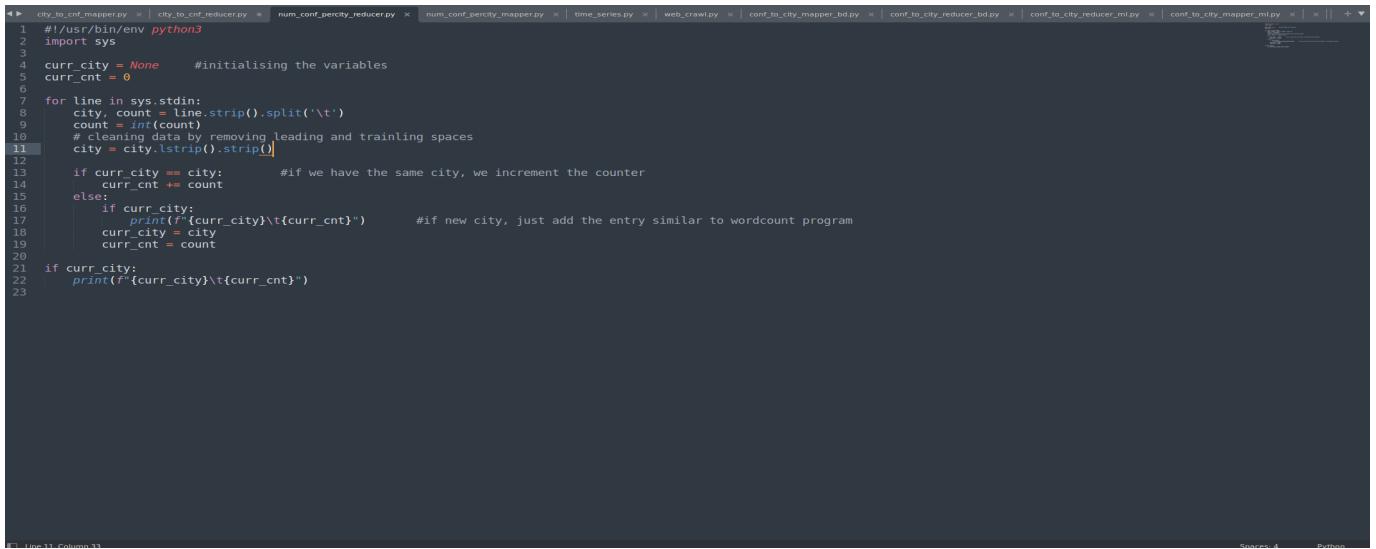
Mapper

Sanket Kulkarni : HW2



```
1 #!/usr/bin/env python3
2 import sys
3
4 for line in sys.stdin:
5     input_line = line.strip().split("\t")           #as our input file is in tsv format
6
7     if len(input_line) >= 3:
8         location = input_line[3].strip()           #taking raw city input
9         location = location.lstrip()
10        #city_data = location.split(',')
11
12        if city_data:
13            city = city_data[0].strip()             #removing leading spaces if any left after data cleaning
14            print(f'{city}\t1')                   #storing in intermediate space
15
```

Reducer:



```
1 #!/usr/bin/env python3
2 import sys
3
4 curr_city = None      #initialising the variables
5 curr_cnt = 0
6
7 for line in sys.stdin:
8     city, count = line.strip().split("\t")
9     count = int(count)
10    # cleaning data by removing leading and trailing spaces
11    city = city.lstrip().strip()
12
13    if curr_city == city:          #if we have the same city, we increment the counter
14        curr_cnt += count
15    else:
16        if curr_city:
17            print(f'{curr_city}\t{curr_cnt}')    #if new city, just add the entry similar to wordcount program
18        curr_city = city
19        curr_cnt = count
20
21 if curr_city:
22     print(f'{curr_city}\t{curr_cnt}')
23
```

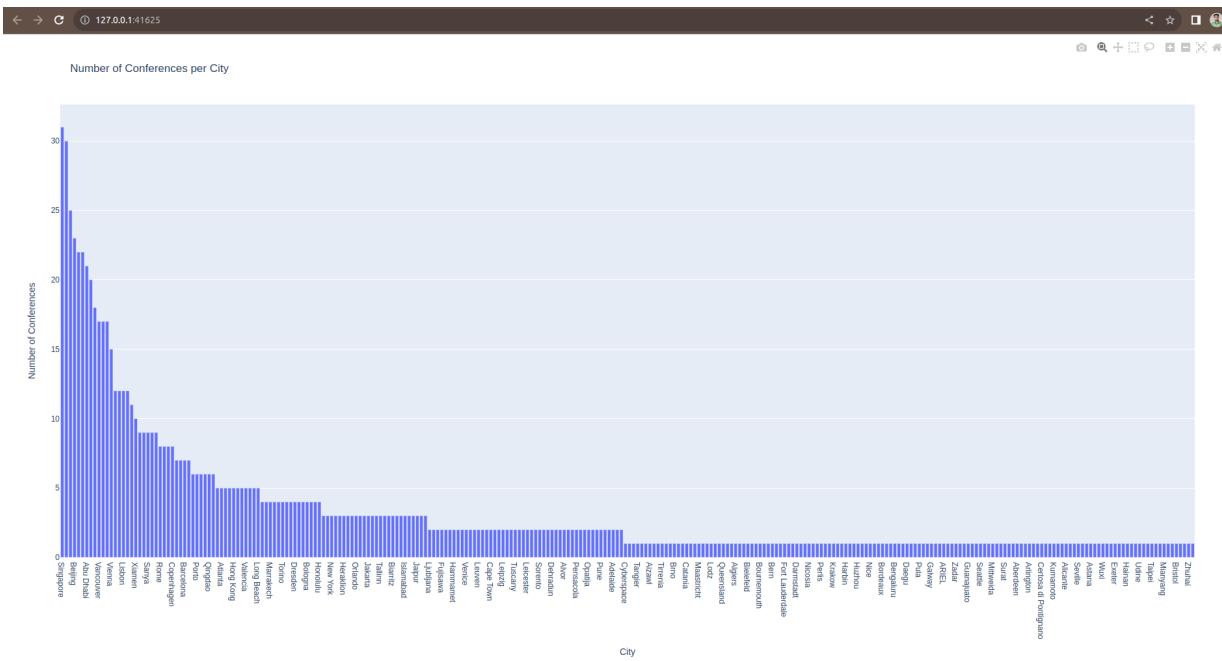
Plotted the graph for the number of conferences per city using plotly library. Have made the graph in descending order so that it visualizes properly.

Code Snippet:

Sanket Kulkarni : HW2

```
city_to_cnf_mapper.py | city_to_cnf_reducer.py | num_conf_perCity_reducer.py | num_conf_perCity_mapper.py | time_series.py | web_crawl.py | conf_to_city_mapper_b6.py | conf_to_city_reducer_m1.py | conf_to_city_mapper_m1.py | city_plot.py | timeseries.py < />
1 import pandas as pd
2 import plotly.express as px
3
4 # Loading the TSV file into a DataFrame
5 df = pd.read_csv('conference_aimlbd.tsv', sep='\t')
6
7 # Counting the city occurrences
8 city_conf_counts = df['City'].value_counts().reset_index()
9 city_conf_counts.columns = ['City', 'Number of Conferences']
10
11 # Creating a bar graph using Plotly
12 fig = px.bar(city_conf_counts, x='City', y='Number of Conferences', title='Number of Conferences per City')
13 fig.update_xaxes(title='City')
14 fig.update_yaxes(title='Number of Conferences')
15
16 # Show the plot
17 fig.show()
18
```

Number of conferences per city



If we hover over the graph we can see all the cities with their corresponding count.

B) Output the list of conferences per city

Steps:

- Take the input line by line
 - Remove leading and trailing spaces if any
 - Consider City and count and conference name inside the mapper that we will put in the intermediate space
 - In the reducer, load the line data and as it was in json format while storing, check for the attributes inside it. Specifically we will look towards location here.
 - If our current city is equal to city, we increment count of that city as well as extend the conference name in that list where city will be our key
 - Finally, put this entire json inside the output file

Mapper Snippet:

```
city_to_cnt_mapper.py x city_to_cnf_reducer.py e | time_series.py x | web_crawl.py x | conf_to_city_mapper_bp.py x | conf_to_city_reducer_bp.py x | conf_to_city_reducer_init.py x | conf_to_city_mapper_mi.py x | city_plot.py x | timeseries.py x | Scraping data from all the three link e
1 #!/usr/bin/env python3
2 import sys
3 import json
4 ''' Storing the output in json format, key will be City and then as per each city, count will be increased in reducer'''
5
6 for line in sys.stdin:
7     items = line.strip().split('\t')      # taking the input of each line
8
9     if len(items) >= 3:
10         city_data = items[3].strip()      # data cleaning by removing leading and trailing spaces if there are any
11         city_data = city_data.lstrip()
12         if city_data:
13             city = city_data.strip()
14             count = 1
15             conference_names = [items[2]] # the conference name is in items[0]
16             city_data = {
17                 "City": city,
18                 "Count": count,
19                 "Conferences": conference_names
20             }
21         print(json.dumps(city_data))
22
```

Reducer Snippet:

```
1 #!/usr/bin/env python3
2 import sys
3 import json
4 """ Reducer for conference data per city"""
5 current_city = None
6 city_data = {}      #json to store the output
7
8 for l in sys.stdin:
9     data = json.loads(l.strip())    #as our mapper was putting data into json chunks, we pick it up using json.load
10    location = data["City"]
11    city = location.lstrip().strip()      #removing leading and trailing spaces to normalise the data
12
13    if current_city == city:           #increment the data count and put if the city is same and conferences are different, append it
14        city_data["Count"] += data["Count"]
15        city_data["Conferences"].extend(data["Conferences"])
16    else:
17        if current_city:
18            print(json.dumps(city_data))
19        current_city = city
20        city_data = {
21            "City": city,
22            "Count": data["Count"],          #this count will help us in storing and cross checking if we have correct number of con per city
23            "Conferences": data["Conferences"]      #final json will be a three level json data, with city having multiple conferences and its number
24        }
25
26 if current_city:
27     print(json.dumps(city_data))
28
```

Program Running command:

```
hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-*jar \
-files city_to_cnf_mapper.py,city_to_cnf_reducer.py -mapper "python3 city_to_cnf_mapper.py" -reducer
"python3 city_to_cnf_reducer.py" \
-input conference_aimlbd.tsv -output city_to_cnf
```

Sanket Kulkarni : HW2

```

Activities Terminal Oct 4 00:59
hadoop@sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/city_to_conference$ hadoop jar /usr/local/hadoop/share/KULKARNI/city_to_conference
cp: cannot stat '/home/sanket/Downloads/conference_all.tsv': No such file or directory
hadoop@sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/city_to_conference$ sudo cp /home/sanket/Downloads/conference_allbd.tsv /usr/local/hadoop/share/KULKARNI/city_to_conference/
hadoop@sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/city_to_conference$ ls
city_to_cnf_mapper.py city_to_cnf_reducer.py conference_allbd.tsv
hadoop@sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/city_to_conference$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming*.jar \
-file city_to_cnf_mapper.py city_to_cnf_reducer.py -mapper "python3 city_to_cnf_mapper.py" \
-reducer "python3 city_to_cnf_reducer.py" \
-input conference_allbd.tsv \
-output city_to_conf
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/common/lib/hadoop-auth-2.8.1.jar) to method sun.security.krb5.Config.getInst
ance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
23/10/04 00:54:19 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session.id
23/10/04 00:54:19 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
23/10/04 00:54:19 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
23/10/04 00:54:19 INFO mapred.FileOutputCommitter: Total number of tasks to process : 1
23/10/04 00:54:19 INFO mapred.JobSubmitter: Number of splits:1
23/10/04 00:54:19 INFO mapred.JobSubmitter: Submitting tokens for job: job_local791898778_0001
23/10/04 00:54:19 INFO mapred.LocalDistributedCacheManager: Localized file:/usr/local/hadoop/share/KULKARNI/city_to_conference/city_to_cnf_mapper.py as file:/tmp/hadoop-hadoop/mapred/local/1696406059747/city_to_cnf_mapper.py
23/10/04 00:54:19 INFO mapred.LocalDistributedCacheManager: Localized file:/usr/local/hadoop/share/KULKARNI/city_to_conference/city_to_cnf_reducer.py as file:/tmp/hadoop-hadoop/mapred/local/1696406059748/city_to_cnf_reducer.py
23/10/04 00:54:20 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
23/10/04 00:54:20 INFO mapred.LocalJobRunner: OutputCommitter set in config null
23/10/04 00:54:20 INFO mapreduce.Job: Running job: job_local791898778_0001
23/10/04 00:54:20 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
23/10/04 00:54:20 INFO output.FileOutputCommitter: FileOutputCommitter Algorithm version is 1
23/10/04 00:54:20 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
23/10/04 00:54:20 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
23/10/04 00:54:20 INFO output.FileOutputCommitter: Starting task: attempt_local791898778_0001_n_000000_0
23/10/04 00:54:20 INFO output.FileOutputCommitter: FileOutputCommitter Algorithm version is 1
23/10/04 00:54:20 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup_temporary folders under output directory:false, ignore cleanup failures: false
23/10/04 00:54:20 INFO mapred.Task: Using DeserializerProcessTree: []
23/10/04 00:54:20 INFO mapred.MapTask: record reader class: file:/usr/local/hadoop/share/KULKARNI/city_to_conference/conference_allbd.tsv:0+102318
23/10/04 00:54:20 INFO mapred.MapTask: numReduceTasks: 1
23/10/04 00:54:20 INFO mapred.MapTask: (EQUATOR) 0 kvl 26214396(194857584)
23/10/04 00:54:20 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 10
23/10/04 00:54:20 INFO mapred.MapTask: soft limit at 83886080
23/10/04 00:54:20 INFO mapred.MapTask: mapreduce.task.io.sort.start = 0
23/10/04 00:54:20 INFO mapred.MapTask: kvCount=2243439, length= 164857600
23/10/04 00:54:20 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
23/10/04 00:54:20 INFO streaming.PipeMapRed: PipeMapRed exec [/usr/bin/python3, city_to_cnf_mapper.py]
23/10/04 00:54:20 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
23/10/04 00:54:20 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
23/10/04 00:54:20 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
23/10/04 00:54:20 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
23/10/04 00:54:20 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
23/10/04 00:54:20 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
23/10/04 00:54:20 INFO Configuration.deprecation: mapred.input.length is deprecated. Instead, use mapreduce.map.input.length
23/10/04 00:54:20 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
23/10/04 00:54:20 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
23/10/04 00:54:20 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
23/10/04 00:54:20 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]

```

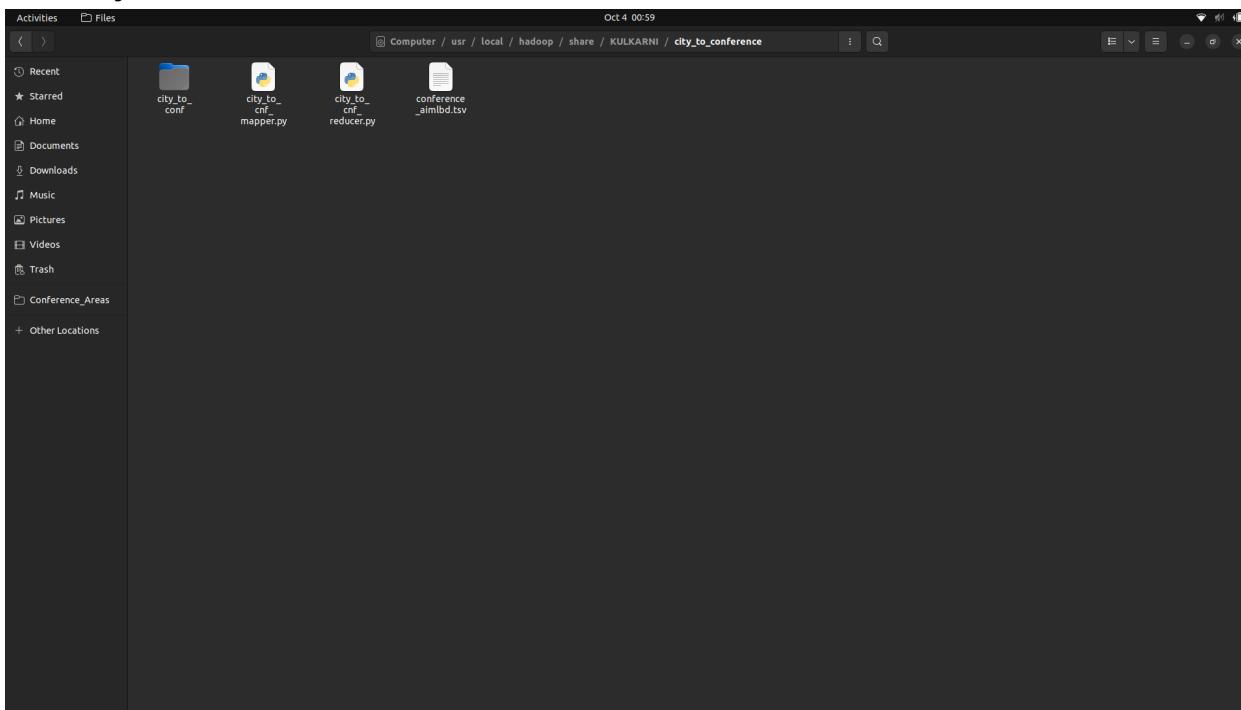
Program Completion event:

```

Activities Terminal Oct 4 00:59
hadoop@sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/city_to_conference$ hadoop jar /usr/local/hadoop/share/KULKARNI/city_to_conference
23/10/04 00:54:21 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
23/10/04 00:54:21 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
23/10/04 00:54:21 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
23/10/04 00:54:21 INFO streaming.PipeMapRed: Records R/W=918/1
23/10/04 00:54:21 INFO streaming.PipeMapRed: MRErrorHandlerThread done
23/10/04 00:54:21 INFO streaming.PipeMapRed: mapreduce.taskattempted
23/10/04 00:54:21 INFO mapred.Task: Task attempt_local791898778_0001_r_000000_0 is done. And is in the process of committing
23/10/04 00:54:21 INFO mapred.LocalJobRunner: 1 / 1 copied
23/10/04 00:54:21 INFO mapred.Task: Task attempt_local791898778_0001_r_000000_0 is allowed to commit now
23/10/04 00:54:21 INFO output.FileOutputCommitter: Saved output of task attempt_local791898778_0001_r_000000_0 to file:/usr/local/hadoop/share/KULKARNI/city_to_conference/city_to_conf/_temporary/0/task_local791898778_0001_r_000000_1
23/10/04 00:54:21 INFO mapred.FileOutputCommitter: Records R/W=918/1 > reduce
23/10/04 00:54:21 INFO mapred.Task: Task attempt_local791898778_0001_r_000000_0 done.
23/10/04 00:54:21 INFO mapred.LocalJobRunner: Finishing task: attempt_local791898778_0001_r_000000_0
23/10/04 00:54:22 INFO mapred.LocalJobRunner: reduce task executor complete.
23/10/04 00:54:22 INFO mapreduce.Job: map 100% reduce 100%
23/10/04 00:54:22 INFO mapreduce.Job: Job job_local791898778_0001 completed successfully
23/10/04 00:54:22 INFO mapreduce.Job: Job: Counters 
  File System Counters
    FILE: Number of bytes read=722438
    FILE: Number of bytes written=1384584
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
  Map-Reduce Framework
    Map input records=918
    Map output records=918
    Map output bytes=120664
    Map output materialized bytes=123004
    Input File Bytes=120664
    Spilling Input records=0
    Combine output records=0
    Reduce input groups=746
    Reduce shuffle bytes=123004
    Reduce input records=918
    Reduce output records=279
    Spilled Records=1036
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=13
    Map output committed heap usage (bytes)=471859200
  Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_TYPE=0
    WRONG_MAGIC=0
    REDUCE=0
  File Input Format Counters
    Bytes Read=102318
  File Output Format Counters
    Bytes Written=88770
23/10/04 00:54:22 INFO streaming.StreamJob: Output directory: city_to_conf
hadoop@sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/city_to_conference$ 

```

Directory Looks like:



Output file is stored in a JSON format for better understanding:

JSON cross verifying Singapore with 31 conferences which matches our earlier output

Json with Json viewer:

```
Activities Google Chrome Oct 4 0:08
↳ inbox (1) · sанкет · Files · hes-resume-cover · Untitled diff-Diff · conference_all · Untitled document · Exploring facets · Library Carpentry · JSON Editor Online · +
<→ Log in https://joneditoronline.org/#left.local.goheju
tree table
1 ↴ [ ]
2 ↴ [ ]
3 ↴ [ ]
4 ↴ [ ]
5 ↴ [ ]
6 ↴ [ ]
7 ↴ [ ]
8 ↴ [ ]
9 ↴ [ ]
10 ↴ [ ]
11 ↴ [ ]
12 ↴ [ ]
13 ↴ [ ]
14 ↴ [ ]
15 ↴ [ ]
16 ↴ [ ]
17 ↴ { "City": "Aberystwyth", "Count": 3, "Conferences": [ "Gold for Papers - EvolUSART 2024 (3-5 April 2024)", "EuroStar 2024 - The Leading European Event on Bloa+Inspired Computation", "EvoStar 2024 - The Leading European Event on Bloa+Inspired Computation" ] }, 18 ↴ [ ]
19 ↴ { "City": "Abu Dhabi", "Count": 21, "Conferences": [ "International Young Scientists Conference in Computational Science 2023", "International Young Scientists Conference in Computational Science 2023", "The 10th International Conference on Social Networks Management and Security", "The 10th International Workshop on Big Data and Social Networks Management and Security", "The 10th International workshop on Big Data and Social Networking Management and Security", "The 10th International workshop on Big Data and Social Networking Management and Security", "The 6th International workshop on Sentiment Analysis and Mining of Social Networks", "The 6th International workshop on Sentiment Analysis and Mining of Social Networks", "The 7th International Workshop on Advances in Natural Language Processing", "The 7th International Workshop on Advances in Natural Language Processing", "The 7th International workshop on Advances in Natural Language Processing", "The 7th International workshop on Data Science Engineering and its Applications", "The 7th International workshop on Data Science Engineering and its Applications", "The 7th International workshop on Data Science Engineering and its Applications", "The 7th International workshop on Online Social Networks Technologies", "The 9th International workshop on Online Social Networks Technologies", "The 9th International Workshop on Cognitive and Neural Systems", "The International Workshop on Cognitive and Neural Systems" ] }, 20 ↴ [ ]
21 ↴ { "City": "Adelaide", "Count": 1, "Conferences": [ "The 8th IEEE International Conference on Agents", "The 8th IEEE International Conference on Agents" ] }, 22 ↴ [ ]
23 ↴ { "City": "Agde", "Count": 1, "Conferences": [ "IEEE -2023 The 11th International Conference on Control, Mechatronics and Automation (ICMA 2023)" ] }
Establishing secure connection...
```

C) For each conference regardless of the year (e.g., Big Data, ML, AI), output the list of cities. (1 Mapper & 1 Reducer for each research field)

Approach :

- As I had collected data of all three research areas in one tsv only, I had mapped these with respective research areas with an extra column, so that we have data about what research area the conference is from
- Created 3 mappers and 3 reducers all of which belong to respective research areas.
- So we will have 3 output files belonging to each of the research areas as well.

Steps:

- Create a mapper which will capture the research area we need to run it for.
- I am making the key based on Acronym of the conference, because that will only be the constant value
- Emitting key value pair for acronym, research area as the key and city as the value
- Create a reducer, split the input line we got from the intermediate space.
- Made a combination key of acronym clubbed with research area
- If current key matches the combined key, we append it to the list else we create a new JSON object and next time we append it n that json object if the key matches
- Finally, put this json in output format

Mapper Snippet:

```
1 #!/usr/bin/env python3
2 import sys
3
4 # Specify the research area to filter (e.g., "Machine Learning")
5 research_area_input = "BigData"
6
7 for line in sys.stdin:
8     # Split the input line by tab
9     columns = line.strip().split('\t')
10
11    if len(columns) == 6:
12        acronym = columns[0]
13        res_area = columns[5]
14        city = columns[3]
15
16        # Check if the research area matches the desired area
17        if res_area == research_area_input:
18            # Emit key-value pair (conference_acronym, res_area) as key
19            # and city as the value
20            print(f'{acronym}\t{res_area}\t{city}')
```

Sanket Kulkarni : HW2

Reducer snippet:

```
conf_to_city_reducer_bd.py
1 #!/usr/bin/env python3
2 import sys
3 import json
4
5 curr_key = None
6 curr_cities = []
7
8 research_area_input = "BigData"
9
10 for line in sys.stdin:
11     # Split the input line by tab
12     key, res_area, city = line.strip().split('\t')
13
14     # Check if the research area matches the desired area
15     if res_area == research_area_input:
16         # Combine conference acronym and research area as the key
17         comb_key = f'{key}\t{res_area}'
18
19         if curr_key == comb_key:
20             curr_cities.append(city)
21         else:
22             if curr_key:
23                 # Build a JSON object for the output
24                 output_data = {
25                     "Conference_Acronym": curr_key.split('\t')[0],
26                     "Research_area": curr_key.split('\t')[1],
27                     "Cities": curr_cities
28                 }
29
30                 # Output the JSON object
31                 print(json.dumps(output_data))
32
33             curr_key = comb_key
34             curr_cities = [city]
35
36     # Output the last JSON object
37 if curr_key:
38     output_data = {
39         "Conference_Acronym": curr_key.split('\t')[0],
40         "Research_area": curr_key.split('\t')[1],
41         "Cities": curr_cities
42     }
43     print(json.dumps(output_data))

```

Similar mappers and reducers will be used for ML, AI. Just we need to change the research_area_input

Command: Similar command for AI and Big data, just change the file names

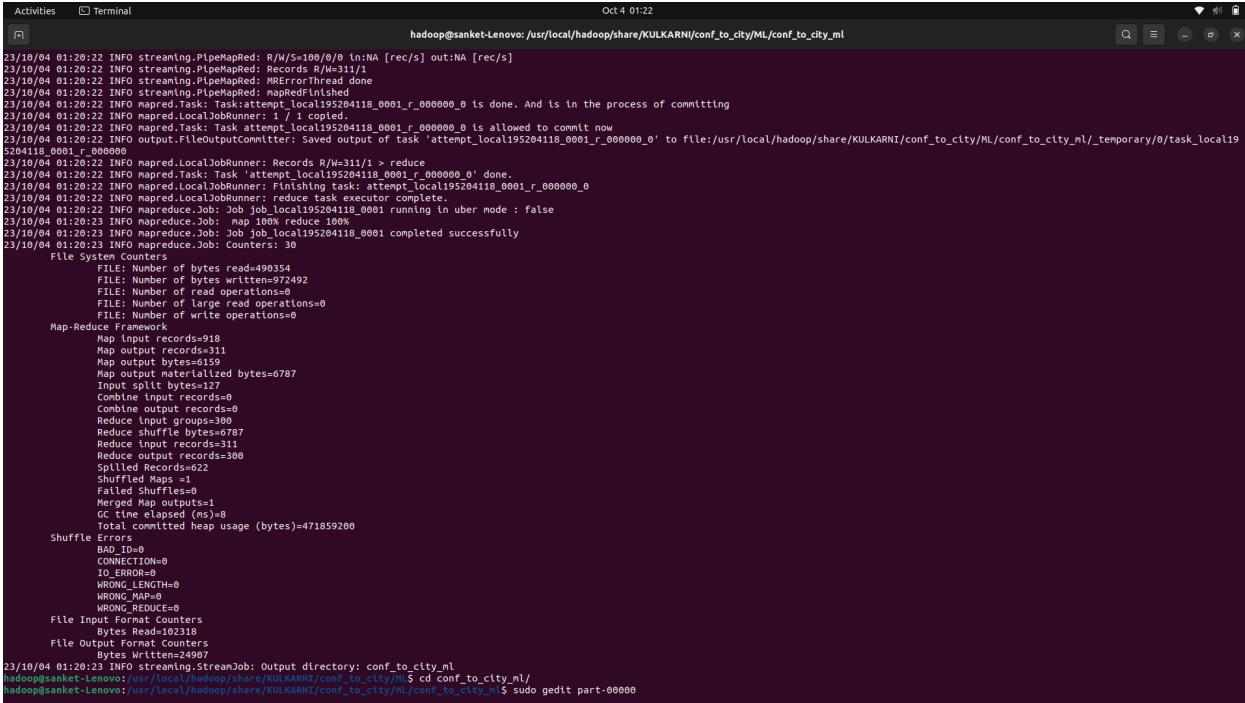
```
hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-*jar \
-files conf_to_city_mapper_ml.py,conf_to_city_reducer_ml.py -mapper "python3 \
conf_to_city_mapper_ml.py" -reducer "python3 conf_to_city_reducer_ml.py" \
-input conference_aimlbd.tsv -output conf_to_city_ml
```

Code execution for ML:

```
Activities Terminal Oct 4 01:23
hadoop@sankey-Lenovo: /usr/local/hadoop/share/KULKARNI/conf_to_city/ML/conf_to_city_ml
conference_aimlbd.tsv conf_to_city_mapper_ml.py conf_to_city_reducer_ml.py
conference_aimlbd.tsv conf_to_city_mapper_ml.py conf_to_city_reducer_ml.py -mapper "python3 conf_to_city_mapper_ml.py" -reducer "python3 conf_to_city_reducer_ml.py"
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.8.1.jar) to method sun.security.krb5.config.getInst
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: All illegal access operations will be denied in a future release
23/10/04 01:20:21 INFO Configuration.deprecation: mapreduce.jobtracker.address is deprecated. Instead, use dfs.namenode.session-id
23/10/04 01:20:21 INFO Configuration.deprecation: mapreduce.jobtracker.address is deprecated. Instead, use dfs.namenode.session-id
23/10/04 01:20:21 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
23/10/04 01:20:21 INFO mapreduce.JobSubmitter: number of splits: 1
23/10/04 01:20:21 INFO mapreduce.JobSubmitter: Submitting tokens for job: local192.168.0.88/0
23/10/04 01:20:21 INFO mapred.LocalDistributedCacheManager: Localized file:/usr/local/hadoop/share/KULKARNI/conf_to_city/ML/conf_to_city_mapper_ml.py as file:/tmp/hadoop-hadoop/mapred/local/1096407621588/conf_to
city_mapper_ml.py
23/10/04 01:20:21 INFO mapred.LocalDistributedCacheManager: Localized file:/usr/local/hadoop/share/KULKARNI/conf_to_city/ML/conf_to_city_reducer_ml.py as file:/tmp/hadoop-hadoop/mapred/local/1096407621589/conf_t
o
city_reducer_ml.py
23/10/04 01:20:21 INFO mapred.MapTask: The url to track the job: http://localhost:8080/
23/10/04 01:20:21 INFO mapred.LocalJobRunner: OutputCommitter set in config null
23/10/04 01:20:21 INFO mapred.MapTask: Running job: job_local192.168.0.88/0001
23/10/04 01:20:21 INFO mapred.FailedOutputCommitter: Failed Output Committer Algorithm version is 1
23/10/04 01:20:21 INFO mapred.FailedOutputCommitter: Failed Output Committer Algorithm version is 1
23/10/04 01:20:22 INFO mapred.LocalJobRunner: Waiting for map tasks
23/10/04 01:20:22 INFO mapred.FailedOutputCommitter: Failed Output Committer Algorithm version is 0
23/10/04 01:20:22 INFO mapred.FailedOutputCommitter: Failed Output Committer Algorithm version is 1
23/10/04 01:20:22 INFO mapred.FailedOutputCommitter: Failed Output Committer skip cleanup _temporary Folders under output directory:false, ignore cleanup failures: false
23/10/04 01:20:22 INFO mapred.MapTask: Processing split: file:/usr/local/hadoop/share/KULKARNI/conf_to_city/ML/conference_aimlbd.tsv:0+02318
23/10/04 01:20:22 INFO mapred.MapTask: mapred.task.to.sort.mib: 100
23/10/04 01:20:22 INFO mapred.MapTask: mapred.task.to.sort.mib: 100
23/10/04 01:20:22 INFO mapred.MapTask: bufsstart = 0 buffwd = 104857600
23/10/04 01:20:22 INFO mapred.MapTask: mapred.task.to.sort.mib: 100
23/10/04 01:20:22 INFO Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
23/10/04 01:20:22 INFO StreamingPipelineMap: pipemapred exec [/usr/bin/python2.7 /opt/codenvy-0.10.0-mapreduce/bin/pipemapred.py]
23/10/04 01:20:22 INFO Configuration.deprecation: map.input is deprecated. Instead, use mapreduce.map.input.dir
23/10/04 01:20:22 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
23/10/04 01:20:22 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
23/10/04 01:20:22 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
23/10/04 01:20:22 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
23/10/04 01:20:22 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
23/10/04 01:20:22 INFO Configuration.deprecation: map.input.records is deprecated. Instead, use mapreduce.map.input.records
23/10/04 01:20:22 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
23/10/04 01:20:22 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
23/10/04 01:20:22 INFO StreamingPipelineMap: /u/s/10/8/inNA [rec/] out/inNA [rec/]
23/10/04 01:20:22 INFO StreamingPipelineMap: /u/s/10/8/inNA [rec/] out/inNA [rec/]
23/10/04 01:20:22 INFO StreamingPipelineMap: Records 0 to 191
```

Sanket Kulkarni : HW2

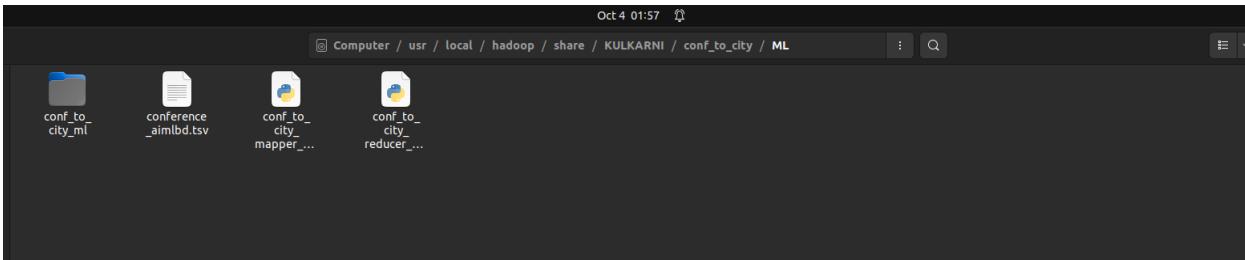
Code Successfully executed ML:



```
Oct 4 01:22
hadoop@sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/conf_to_city/ML/conf_to_city_ml

23/10/04 01:20:22 INFO streaming.PipeMapred: R/W=5/09/0/0 (in:NA [rec/s] out:NA [rec/s])
23/10/04 01:20:22 INFO streaming.PipeMapred: Records R/H=311/1
23/10/04 01:20:22 INFO streaming.PipeMapred: MRErrorThread done
23/10/04 01:20:22 INFO streaming.PipeMapred: mapredFinished
23/10/04 01:20:22 INFO mapred.Task: Task:attempt_local195204118_0001_r_000000_0 is done. And is in the process of committing
23/10/04 01:20:22 INFO mapred.Task: Task:attempt_local195204118_0001_r_000000_0 is allowed to commit now
23/10/04 01:20:22 INFO output.FileOutputCommitter: Saved output of task 'attempt_local195204118_0001_r_000000_0' to file:/usr/local/hadoop/share/KULKARNI/conf_to_city/ML/conf_to_city_ml/_temporary/0/task_local195204118_0001_r_000000
23/10/04 01:20:22 INFO mapred.LocalJobRunner: Records R/W=311/1 > reduce
23/10/04 01:20:22 INFO mapred.Task: Task 'attempt_local195204118_0001_r_000000_0' done.
23/10/04 01:20:22 INFO mapred.LocalJobRunner: Reducing local tasks
23/10/04 01:20:22 INFO mapred.LocalJobRunner: reduce 0 executor complete.
23/10/04 01:20:22 INFO mapreduce.Job: Job job_local195204118_0001 running in uber mode : false
23/10/04 01:20:23 INFO mapreduce.Job: map 100% reduce 100%
23/10/04 01:20:23 INFO mapreduce.Job: Job job_local195204118_0001 completed successfully
23/10/04 01:20:23 INFO mapreduce.Job: Counters: 30
  File System Counters:
    File input bytes read=490354
    File output bytes written=972492
    File: Number of read operations=0
    File: Number of large read operations=0
    File: Number of write operations=0
  Map-Reduce Framework:
    Map input records=918
    Map output records=311
    Map output bytes=6159
    Map output materialized bytes=6787
    Input split bytes=127
    Combine input records=0
    Reduce input records=0
    Reduce input groups=300
    Reduce shuffle bytes=6787
    Reduce input records=311
    Reduce output records=300
    Spilled Records=622
    Shuffled Maps=300
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=8
    Total committed heap usage (bytes)=471859200
  Shuffle Errors:
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAP=0
    WRONG_PARTITION=0
    WRONGReducer=0
  File Input Format Counters:
    Bytes Read=102318
  File Output Format Counters:
    Bytes Written=24907
23/10/04 01:20:23 INFO streaming.StreamJob: Output directory: conf_to_city_ml
hadoop@sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/conf_to_city/ML/conf_to_city_ml$ cd conf_to_city_ml
hadoop@sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/conf_to_city/ML/conf_to_city_ml$ sudo gedit part-00000
```

Directory Structure:



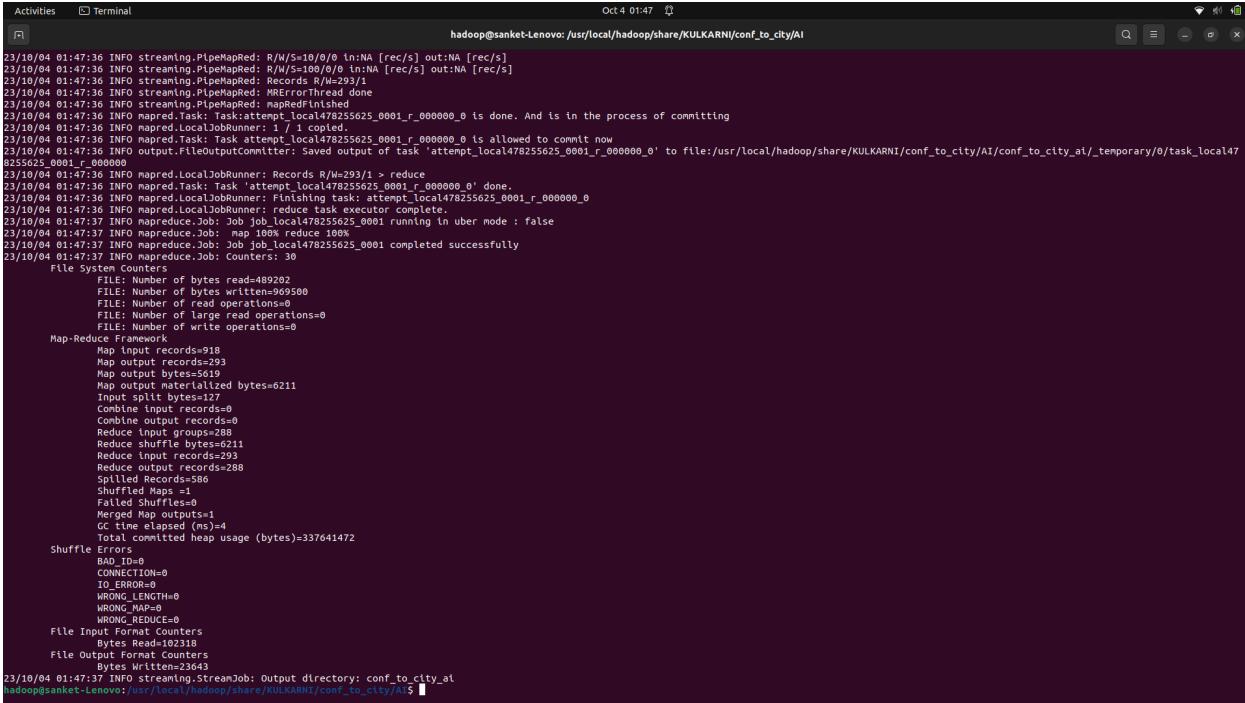
Output for ML as research field:

Activities	Edits	Oct 4 01:22	Save
Open	□	/usr/local/hadoop/share/KULKARNI/conf_to_city/ML/conf_to_city_ml	
1	"Conference_Acronym": "6th AccML", "Research_Area": "ML", "Cities": ["Munich"]}		
2	"Conference_Acronym": "ACML@ECML", "Research_Area": "ML", "Cities": ["Turin"]}		
3	"Conference_Acronym": "ACAI", "Research_Area": "ML", "Cities": ["Jaxing"]}		
4	"Conference_Acronym": "ACDLT", "Research_Area": "ML", "Cities": ["Tuscan"]}		
5	"Conference_Acronym": "ACOSA", "Research_Area": "ML", "Cities": ["Mahe"]}		
6	"Conference_Acronym": "ACTIV", "Research_Area": "ML", "Cities": ["London"]}		
7	"Conference_Acronym": "ACM HPCCT", "Research_Area": "ML", "Cities": ["Jaxing"]}		
8	"Conference_Acronym": "ACM ICML", "Research_Area": "ML", "Cities": ["Shenzhen"]}		
9	"Conference_Acronym": "ACM IJCAI", "Research_Area": "ML", "Cities": ["Singapore"]}		
10	"Conference_Acronym": "ACM ICSIM", "Research_Area": "ML", "Cities": ["Suva"]}		
11	"Conference_Acronym": "ACM IJIPRL", "Research_Area": "ML", "Cities": ["Seoul"]}		
12	"Conference_Acronym": "ACM SPNL", "Research_Area": "ML", "Cities": ["Tianjin"]}		
13	"Conference_Acronym": "ACM-EI/Scopus-AIML", "Research_Area": "ML", "Cities": ["Bangkok"]}		
14	"Conference_Acronym": "ACM-EI/Scopus-MLBDM", "Research_Area": "ML", "Cities": ["Sanya"]}		
15	"Conference_Acronym": "ACM-EI/Scopus-MLDM", "Research_Area": "ML", "Cities": ["Beijing"]}		
16	"Conference_Acronym": "ACM-EI/Scopus-MICMI", "Research_Area": "ML", "Cities": ["Chengdu"]}		
17	"Conference_Acronym": "ACM-EI/Scopus-MLBDM", "Research_Area": "ML", "Cities": ["Sanya"]}		
18	"Conference_Acronym": "ACMLC", "Research_Area": "ML", "Cities": ["Bangkok"]}		
19	"Conference_Acronym": "ACVR", "Research_Area": "ML", "Cities": ["Paris"]}		
20	"Conference_Acronym": "ACWITS", "Research_Area": "ML", "Cities": ["Munich"]}		
21	"Conference_Acronym": "ACOM", "Research_Area": "ML", "Cities": ["Vancouver"]}		
22	"Conference_Acronym": "ADMIT", "Research_Area": "ML", "Cities": ["Chengdu"]}		
23	"Conference_Acronym": "AI", "Research_Area": "ML", "Cities": ["Sydney"]}		
24	"Conference_Acronym": "AI4cyber", "Research_Area": "ML", "Cities": ["Long Beach"]}		
25	"Conference_Acronym": "AI4PMI", "Research_Area": "ML", "Cities": ["Giza"]}		
26	"Conference_Acronym": "AI4PMI", "Research_Area": "ML", "Cities": ["Istanbul"]}		
27	"Conference_Acronym": "AIATA", "Research_Area": "ML", "Cities": ["Kuala Lumpur"]}		
28	"Conference_Acronym": "AIIC", "Research_Area": "ML", "Cities": ["Varanasi"]}		
29	"Conference_Acronym": "AIDBEI", "Research_Area": "ML", "Cities": ["Vancouver"]}		
30	"Conference_Acronym": "AIM@EPIA", "Research_Area": "ML", "Cities": ["Lisbon"]}		
31	"Conference_Acronym": "AIM@MC", "Research_Area": "ML", "Cities": ["Laguna Hills"]}		
32	"Conference_Acronym": "AIMLA", "Research_Area": "ML", "Cities": ["Copenhagen"]}		
34	"Conference_Acronym": "AISI", "Research_Area": "ML", "Cities": ["Porl Salid"]}		
35	"Conference_Acronym": "AISO", "Research_Area": "ML", "Cities": ["Zurich", "Zurich"]}		
36	"Conference_Acronym": "AISTATS", "Research_Area": "ML", "Cities": ["Valencia"]}		
37	"Conference_Acronym": "AITAT", "Research_Area": "ML", "Cities": ["Bangalore"]}		
38	"Conference_Acronym": "AIUP", "Research_Area": "ML", "Cities": ["Kuala Lumpur"]}		
39	"Conference_Acronym": "ANLP", "Research_Area": "ML", "Cities": ["Abu Dhabi"]}		
40	"Conference_Acronym": "ARIA", "Research_Area": "ML", "Cities": ["Vienna"]}		
41	"Conference_Acronym": "ARIAL@ECML", "Research_Area": "ML", "Cities": ["Turin"]}		
42	"Conference_Acronym": "ARRL", "Research_Area": "ML", "Cities": ["Shanghai"]}		
43	"Conference_Acronym": "AST", "Research_Area": "ML", "Cities": ["Koper/Capodistria"]}		
44	"Conference_Acronym": "ATIP", "Research_Area": "ML", "Cities": ["Paris"]}		
45	"Conference_Acronym": "ATVAL", "Research_Area": "ML", "Cities": ["Singapore"]}		
46	"Conference_Acronym": "AU5DM", "Research_Area": "ML", "Cities": ["Auckland"]}		
47	"Conference_Acronym": "Affective_Health", "Research_Area": "ML", "Cities": ["Istanbul"]}		
48	"Conference_Acronym": "AUDSM", "Research_Area": "ML", "Cities": ["Auckland"]}		
49	"Conference_Acronym": "BANDIT", "Research_Area": "ML", "Cities": ["Rome"]}		
50	"Conference_Acronym": "BDN", "Research_Area": "ML", "Cities": ["Kuala Lumpur"]}		
51	"Conference_Acronym": "BDSN", "Research_Area": "ML", "Cities": ["Abu Dhabi"]}		
52	"Conference_Acronym": "BIOM", "Research_Area": "ML", "Cities": ["Zurich"]}		
53	"Conference_Acronym": "BIOS", "Research_Area": "ML", "Cities": ["Sydney", "Sydney"]}		
54	"Conference_Acronym": "BRACIS", "Research_Area": "ML", "Cities": ["Belo Horizonte"]}		

Code Execution AI:

```
Activities Terminal Oct 4 01:48 🔍
hadoop@sankey-Lenovo:/usr/local/hadoop/share/KULKARNI/conf_to_city$ sudo cp /home/sanket/Downloads/conference_ainlbd.tsv /usr/local/hadoop/share/KULKARNI/conf_to_city/AI
hadoop@sankey-Lenovo:/usr/local/hadoop/share/KULKARNI/conf_to_city$ sudo cp /home/sanket/Downloads/conf_to_city_reducer_at.py /usr/local/hadoop/share/KULKARNI/conf_to_city/AI
hadoop@sankey-Lenovo:/usr/local/hadoop/share/KULKARNI/conf_to_city$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming*.jar -files conf_to_city_mapper_at.py,conf_to_city_reducer_at.py -mapper "python3 conf_to_city_mapper_at.py" -reducer "python3 conf_to_city_reducer_at.py" -input conference_ainlbd.tsv -output conf_to_city_st
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.8.1.jar) to method sun.security.krb5.Config.getInst
ance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
[23/10/04 01:47:35 INFO Configuration.deprecation: session_id is deprecated. Instead, use dfs.metrics.session-id
[23/10/04 01:47:35 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
[23/10/04 01:47:35 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
[23/10/04 01:47:35 INFO mapred.JobClient: JobTracker job tracking address: http://localhost:8080
[23/10/04 01:47:35 INFO mapred.JobSubmitter: number of splits:1
[23/10/04 01:47:35 INFO mapred.JobSubmitter: Submitting tokens for job: job_local1478255625_0001
[23/10/04 01:47:35 INFO mapred.LocalDistributedCacheManager: Localized file:/usr/local/hadoop/share/KULKARNI/conf_to_city/AI/conf_to_city_mapper_at.py as file:/tmp/hadoop-hadoop/mapred/local/1696409255779/conf_to
[23/10/04 01:47:35 INFO mapred.LocalDistributedCacheManager: Localized file:/usr/local/hadoop/share/KULKARNI/conf_to_city/AI/conf_to_city_reducer_at.py as file:/tmp/hadoop-hadoop/mapred/local/1696409255780/conf_t
[23/10/04 01:47:36 INFO mapred.Job: The url to track the job: http://localhost:8080/
[23/10/04 01:47:36 INFO mapred.JobRunner: OutputDir set in config null
[23/10/04 01:47:36 INFO mapred.LocalJobController: OutputCommitter org.apache.hadoop.mapred.FileOutputCommitter
[23/10/04 01:47:36 INFO mapred.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
[23/10/04 01:47:36 INFO mapred.LocalJobRunner: Waiting for map tasks
[23/10/04 01:47:36 INFO mapred.LocalJobRunner: Starting task: attempt_local1478255625_0001_m_000000_0
[23/10/04 01:47:36 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
[23/10/04 01:47:36 INFO mapred.Task: ResourceCalculatorProcessTree: []
[23/10/04 01:47:36 INFO mapred.MapTask: Processing split: file:/usr/local/hadoop/share/KULKARNI/conf_to_city/AI/conference_ainlbd.tsv:0+102318
[23/10/04 01:47:36 INFO mapred.MapTask: numReduceTasks: 1
[23/10/04 01:47:36 INFO mapred.MapTask: (EQUATOR) 0 kv1 26214396(104857584)
[23/10/04 01:47:36 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
[23/10/04 01:47:36 INFO mapred.MapTask: mapreduce.task.io.sort.spill.percent: 100
[23/10/04 01:47:36 INFO mapred.MapTask: buffer_size: 104857600
[23/10/04 01:47:36 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
[23/10/04 01:47:36 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapOutputBuffer
[23/10/04 01:47:36 INFO streaming.PipeMapRed: PipeMapRed: exec [/usr/bin/python3, conf_to_city_mapper_at.py]
[23/10/04 01:47:36 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output.dir
[23/10/04 01:47:36 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
[23/10/04 01:47:36 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
[23/10/04 01:47:36 INFO Configuration.deprecation: mapred.partition is deprecated. Instead, use mapreduce.map.input.file
[23/10/04 01:47:36 INFO Configuration.deprecation: mapred.records.on.skip is deprecated. Instead, use mapreduce.map.skiprecords
[23/10/04 01:47:36 INFO Configuration.deprecation: mapred.length is deprecated. Instead, use mapreduce.map.input.length
[23/10/04 01:47:36 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
[23/10/04 01:47:36 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
[23/10/04 01:47:36 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
[23/10/04 01:47:36 INFO Configuration.deprecation: mapred.mapper.input.records is deprecated. Instead, use mapreduce.map.input.records
[23/10/04 01:47:36 INFO streaming.PipeMapRed: R/W=10/P/R.in=[conf_to_city_st],R.out=[conf_to_city_st]
[23/10/04 01:47:36 INFO streaming.PipeMapRed: R/W=10/P/R.in=[conf_to_city_st],R.out=[conf_to_city_st]
```

Code Execution Successfully AI:

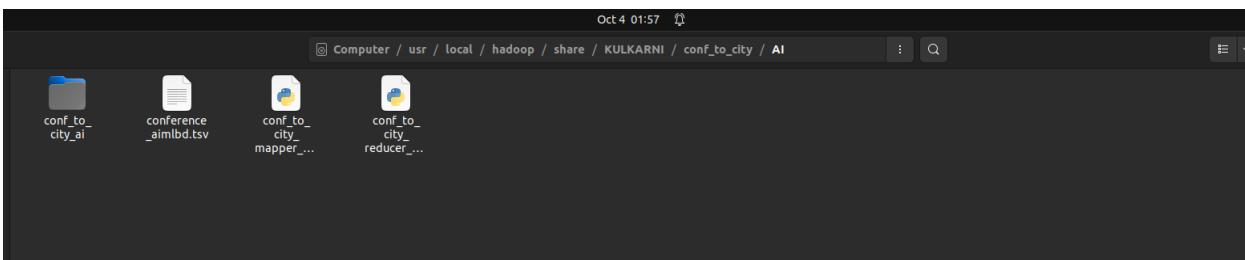


```

Oct 4 01:47
hadoop@sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/conf_to_city/AI

23/10/04 01:47:36 INFO streaming.PipeMapred: R/W/5/10/0/0 in:NA [rec/s] out:NA [rec/s]
23/10/04 01:47:36 INFO streaming.PipeMapred: R/W/s=100/0/8 in:NA [rec/s] out:NA [rec/s]
23/10/04 01:47:36 INFO streaming.PipeMapred: Records R/W=293/1
23/10/04 01:47:36 INFO streaming.PipeMapred: MRErrorThread done
23/10/04 01:47:36 INFO streaming.PipeMapred: mapredFinished
23/10/04 01:47:36 INFO mapred.Task: Task attempt_local478255625_0001_r_000000_0 is done. And is in the process of committing
23/10/04 01:47:36 INFO mapred.Task: Task attempt_local478255625_0001_r_000000_0 is allowed to commit now
23/10/04 01:47:36 INFO mapred.Task: Task attempt_local478255625_0001_r_000000_0 to file:/usr/local/hadoop/share/KULKARNI/conf_to_city/AI/conf_to_city_ai/_temporary/0/task_local478255625_0001_r_000000
23/10/04 01:47:36 INFO output.FileOutputCommitter: Saved output of task 'attempt_local478255625_0001_r_000000_0' to file:/usr/local/hadoop/share/KULKARNI/conf_to_city/AI/conf_to_city_ai/_temporary/0/task_local478255625_0001_r_000000
23/10/04 01:47:36 INFO mapred.LocalJobRunner: Records R/W=293/1 > reduce
23/10/04 01:47:36 INFO mapred.Task: Task attempt_local478255625_0001_r_000000_0 done.
23/10/04 01:47:36 INFO mapred.LocalJobRunner: reduce task executor complete
23/10/04 01:47:37 INFO mapreduce.Job: Job job_local478255625_0001 running in uber mode : false
23/10/04 01:47:37 INFO mapreduce.Job: map 100% reduce 100%
23/10/04 01:47:37 INFO mapreduce.Job: Job job_local478255625_0001 completed successfully
23/10/04 01:47:37 INFO mapreduce.Jobs: Counters: 30
  File System
    FILE: Number of bytes read=489202
    FILE: Number of bytes written=969500
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
  Map-Reduce Framework
    Map input records=918
    Map output records=293
    Map output bytes=5619
    Map output materialized bytes=6211
    Input split bytes=127
    Combine input records=0
    Combine output records=0
    Reduce input groups=288
    Reduce shuffle bytes=6211
    Reduce input records=293
    Reduce output records=288
    Spilled Records=596
    Shuffled Maps =1
    Failed Shuffles=0
    Merged Map outputs=1
    GC time elapsed (ms)=4
    Total committed heap usage (bytes)=337641472
  Shuffle Errors
    BAD_ID=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_MAGIC=0
    WRONG_READ_EOD=0
  File Input Format Counters
    Bytes Read=102318
  File Output Format Counters
    Bytes Written=23643
23/10/04 01:47:37 INFO streaming.StreamJob: Output directory: conf_to_city_ai
hadoop@sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/conf_to_city/AI $ 
```

Folder structure for AI:



Output in Json format for AI:

```

Activities   Edit
Open   part-00000
Oct 4 01:50
Save   x
/us/local/hadoop/share/KULKARNI/conf_to_city/AI/conf_to_city_ai
[{"Conference_Acronym": "AAME-EI", "Research_Area": "AI", "Cities": ["Hong Kong"]}, {"Conference_Acronym": "ACAE", "Research_Area": "AI", "Cities": ["Melbourne"]}, {"Conference_Acronym": "ACAI", "Research_Area": "AI", "Cities": ["Paris"]}, {"Conference_Acronym": "ACDL", "Research_Area": "AI", "Cities": ["Tuscany"]}, {"Conference_Acronym": "ACITY", "Research_Area": "AI", "Cities": ["London"]}, {"Conference_Acronym": "ACKIN", "Research_Area": "AI", "Cities": ["Bangkok"]}, {"Conference_Acronym": "ACM AIEEE", "Research_Area": "AI", "Cities": ["Kyoto"]}, {"Conference_Acronym": "ACM AIEET", "Research_Area": "AI", "Cities": ["Bangkok"]}, {"Conference_Acronym": "ACM CIST", "Research_Area": "AI", "Cities": ["Singapore"]}, {"Conference_Acronym": "ACM CSAT", "Research_Area": "AI", "Cities": ["Beijing"]}, {"Conference_Acronym": "ACM ICRAI", "Research_Area": "AI", "Cities": ["Istanbul"]}, {"Conference_Acronym": "ACM ICDAE", "Research_Area": "AI", "Cities": ["Bangkok"]}, {"Conference_Acronym": "ACM ICDE", "Research_Area": "AI", "Cities": ["Qingdao"]}, {"Conference_Acronym": "ACM ICDE", "Research_Area": "AI", "Cities": ["Beijing"]}, {"Conference_Acronym": "ACM ICML", "Research_Area": "AI", "Cities": ["Shenzhen"]}, {"Conference_Acronym": "ACM ICNA", "Research_Area": "AI", "Cities": ["Singapore"]}, {"Conference_Acronym": "ACM ICRAI", "Research_Area": "AI", "Cities": ["Singapore"]}, {"Conference_Acronym": "ACM ICSCA", "Research_Area": "AI", "Cities": ["Bali"]}, {"Conference_Acronym": "ACM ICSCA", "Research_Area": "AI", "Cities": ["Paris"]}, {"Conference_Acronym": "ACM ICSP", "Research_Area": "AI", "Cities": ["Paris"]}, {"Conference_Acronym": "ACM ICSP", "Research_Area": "AI", "Cities": ["Beijing"]}, {"Conference_Acronym": "ACM ICSP", "Research_Area": "AI", "Cities": ["Shenzhen"]}, {"Conference_Acronym": "ACM ICSP", "Research_Area": "AI", "Cities": ["Singapore"]}, {"Conference_Acronym": "ACM ICSP", "Research_Area": "AI", "Cities": ["Chengdu"]}, {"Conference_Acronym": "ACM ICSP", "Research_Area": "AI", "Cities": ["Bangkok"]}, {"Conference_Acronym": "ACM ICSP", "Research_Area": "AI", "Cities": ["Paris"]}, {"Conference_Acronym": "ACM ICSP", "Research_Area": "AI", "Cities": ["London"]}, {"Conference_Acronym": "ACM ICSP", "Research_Area": "AI", "Cities": ["Tampa"]}, {"Conference_Acronym": "ACM ICSP", "Research_Area": "AI", "Cities": ["Sydney", "Sydney"]}, {"Conference_Acronym": "ACM-EI/Scopus-ACAI", "Research_Area": "AI", "Cities": ["Sanya"]}, {"Conference_Acronym": "ACM-EI/Scopus-AIML", "Research_Area": "AI", "Cities": ["Bangkok"]}, {"Conference_Acronym": "ACM-EI/Scopus-AIHP", "Research_Area": "AI", "Cities": ["Chengdu"]}, {"Conference_Acronym": "ACM-EI/Scopus-AIC", "Research_Area": "AI", "Cities": ["Bangkok"]}, {"Conference_Acronym": "AEICET", "Research_Area": "AI", "Cities": ["Paris"]}, {"Conference_Acronym": "AEIS", "Research_Area": "AI", "Cities": ["London"]}, {"Conference_Acronym": "AGCSSC", "Research_Area": "AI", "Cities": ["Tampa"]}, {"Conference_Acronym": "AHIP", "Research_Area": "AI", "Cities": ["Paris"]}, {"Conference_Acronym": "AM SIGPADS", "Research_Area": "AI", "Cities": ["Atlanta"]}, {"Conference_Acronym": "ANASIG", "Research_Area": "AI", "Cities": ["Sanya"]}, {"Conference_Acronym": "AIAAT", "Research_Area": "AI", "Cities": ["Kuala Lumpur"]}, {"Conference_Acronym": "AIAET", "Research_Area": "AI", "Cities": ["Phuket"]}, {"Conference_Acronym": "AIAET", "Research_Area": "AI", "Cities": ["Bangalore"]}, {"Conference_Acronym": "AIAET", "Research_Area": "AI", "Cities": ["Tokyo"]}, {"Conference_Acronym": "AIBD", "Research_Area": "AI", "Cities": ["Vancouver"]}, {"Conference_Acronym": "AIDBEI", "Research_Area": "AI", "Cities": ["Vancouver"]}, {"Conference_Acronym": "AIFMIA", "Research_Area": "AI", "Cities": ["London"]}, {"Conference_Acronym": "AICMC", "Research_Area": "AI", "Cities": ["Laguna Hills"]}, {"Conference_Acronym": "AIMLA", "Research_Area": "AI", "Cities": ["Copenhagen"]}, {"Conference_Acronym": "AIMLNET", "Research_Area": "AI", "Cities": ["Sydney"]}, {"Conference_Acronym": "AISO", "Research_Area": "AI", "Cities": ["Zurich", "Zurich"]}, {"Conference_Acronym": "AISTAT", "Research_Area": "AI", "Cities": ["Valencia"]}, {"Conference_Acronym": "AISTAT", "Research_Area": "AI", "Cities": ["Maharashtra"]}, {"Conference_Acronym": "AMLP", "Research_Area": "AI", "Cities": ["Abu Dhabi"]}, {"Conference_Acronym": "APCS", "Research_Area": "AI", "Cities": ["Honolulu"]}, {"Conference_Acronym": "APCT", "Research_Area": "AI", "Cities": ["Wuhan"]}, {"Conference_Acronym": "APWG eCrime", "Research_Area": "AI", "Cities": ["Barcelona"]}, {"Conference_Acronym": "ARAEE ET", "Research_Area": "AI", "Cities": ["Shanghai"]}, {"Conference_Acronym": "AFFECTIVE-HEALTH", "Research_Area": "AI", "Cities": ["Vienna"]}, {"Conference_Acronym": "AUSMED", "Research_Area": "AI", "Cities": ["Istanbul"]}, {"Conference_Acronym": "AUSMED", "Research_Area": "AI", "Cities": ["Auckland"]}, {"Conference_Acronym": "BDSN", "Research_Area": "AI", "Cities": ["Abu Dhabi"]}], Plain Text Tab Width: 8 Ln 32, Col 81 INS

```

Code Execution of Big Data:

```

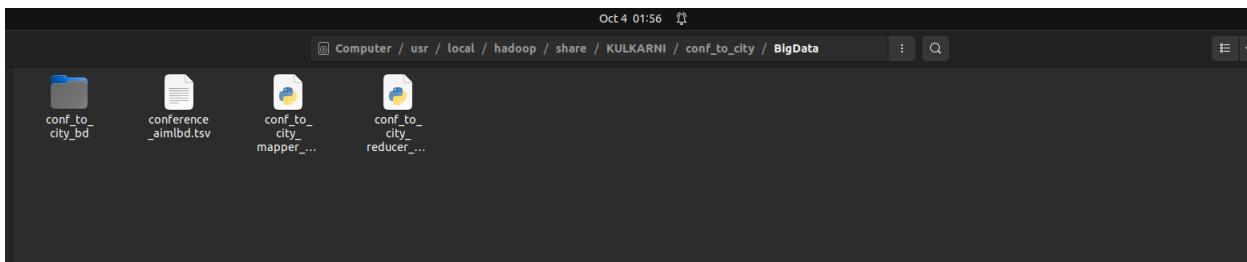
Activities   Terminal
Oct 4 01:55
hadoop@sanket-Lenovo:/us/local/hadoop/share/KULKARNI/conf_to_city/BigData/conf_to_city_bd
hadoop@ sanket-Lenovo: /us/local/hadoop/share/KULKARNI/conf_to_city/BigData$ ls
conference_aimbd.tsv  conf_to_city_mapper_bd.py  conf_to_city_reducer_bd.py
hadoop@ sanket-Lenovo: /us/local/hadoop/share/KULKARNI/conf_to_city/BigData$ hadoop jar /us/local/hadoop/tools/lib/hadoop-streaming*.jar -files conf_to_city_mapper_bd.py,conf_to_city_reducer_bd.py -mapper /usr/bin/python3 /usr/bin/python3 conf_to_city_mapper_bd.py -reducer /usr/bin/python3 conf_to_city_reducer_bd.py -input conference_aimbd.tsv -output conf_to_city_bd
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/us/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.8.1.jar) to method sun.security.krb5.config.getInstnace()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings for further illegal reflective access operations
WARNING: All illegal access operations will be denied starting from Java 9
23/10/04 01:53:35 INFO Configuration.deprecation: session_id is deprecated. Instead, use dfs.metrics.session-id
23/10/04 01:53:35 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=-
23/10/04 01:53:35 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
23/10/04 01:53:36 INFO mapred.FileInputFormat: Total input files to process : 1
23/10/04 01:53:36 INFO mapred.FileInputFormat: Input splits: 1
23/10/04 01:53:36 INFO mapred.JobClient: JobSubmission: Submitting tokens for job: job_local1067668264.0001
23/10/04 01:53:36 INFO mapred.LocalDistributedCacheManager: Localized file:/us/local/hadoop/share/KULKARNI/conf_to_city/BigData/conf_to_city_mapper_bd.py as file:/tmp/hadoop-hadoop/mapred/local/1696409616495/co nf_to_city_mapper_bd.py
23/10/04 01:53:36 INFO mapred.LocalDistributedCacheManager: Localized file:/us/local/hadoop/share/KULKARNI/conf_to_city/BigData/conf_to_city_reducer.bd.py as file:/tmp/hadoop-hadoop/mapred/local/1696409616495/c onf_to_city_reducer_bd.py
23/10/04 01:53:36 INFO mapred.LocalJobRunner: Job: The url to track the job: http://localhost:8080/
23/10/04 01:53:36 INFO mapred.LocalJobRunner: OutputCommitter set in config null
23/10/04 01:53:36 INFO mapred.Job: Running job: job_local1067668264.0001
23/10/04 01:53:36 INFO mapred.LocalJobRunner: OutputCommitter: org.apache.hadoop.mapred.FileOutputCommitter
23/10/04 01:53:36 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
23/10/04 01:53:36 INFO output.FileOutputCommitter: File Output Committer skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
23/10/04 01:53:36 INFO output.FileOutputCommitter: Waiting for tasks
23/10/04 01:53:36 INFO prelocal.JobSubmitter: Submitting tokens for job: job_local1067668264.0001_m_000000_0
23/10/04 01:53:36 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
23/10/04 01:53:36 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
23/10/04 01:53:36 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
23/10/04 01:53:36 INFO mapred.MapTask: Processing split: file:/us/local/hadoop/share/KULKARNI/conf_to_city/BigData/conference_aimbd.tsv:0+102318
23/10/04 01:53:36 INFO mapred.MapTask: Map tasks: 1
23/10/04 01:53:36 INFO mapred.MapTask: Map tasks: 1 (Equation: 1)
23/10/04 01:53:36 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
23/10/04 01:53:37 INFO mapred.MapTask: soft limit at 83886080
23/10/04 01:53:37 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
23/10/04 01:53:37 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
23/10/04 01:53:37 INFO mapred.MapTask: Map task class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
23/10/04 01:53:37 INFO mapred.MapTask: Map task class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
23/10/04 01:53:37 INFO mapred.MapTask: Map task attemptId = 1
23/10/04 01:53:37 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapred.task.id
23/10/04 01:53:37 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapred.cluster.local.dir
23/10/04 01:53:37 INFO Configuration.deprecation: mapred.input.file is deprecated. Instead, use mapred.map.input.file
23/10/04 01:53:37 INFO Configuration.deprecation: mapred.skip.rows is deprecated. Instead, use mapred.job.skiprecords
23/10/04 01:53:37 INFO Configuration.deprecation: mapred.map.output.compress is deprecated. Instead, use mapred.map.output.length
23/10/04 01:53:37 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapred.job.job.id
23/10/04 01:53:37 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapred.job.user.name
23/10/04 01:53:37 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapred.task.partition
23/10/04 01:53:38 INFO streaming.PipeMapRed: R/W=s/10/0/in:NA [rec/s] out:NA [rec/s]
23/10/04 01:53:38 INFO streaming.PipeMapRed: R/W=S/10/0/in:NA [rec/s] out:NA [rec/s]

```

Code execution successful for Big Data:

```
Activities Terminal Oct 4 01:54
hadoop@sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/conf_to_city/BigData/conf_to_city_bd
23/10/04 01:53:38 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
23/10/04 01:53:38 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
23/10/04 01:53:38 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
23/10/04 01:53:38 INFO streaming.PipeMapRed: Reducer thread finished
23/10/04 01:53:38 INFO streaming.PipeMapRed: errorThread done
23/10/04 01:53:38 INFO streaming.PipeMapRed: mapReduced
23/10/04 01:53:38 INFO streaming.PipeMapRed: mapReduced
23/10/04 01:53:38 INFO mapped.Task: Task attempt_local1067668264_0001_r_000000_0 is done. And is in the process of committing
23/10/04 01:53:38 INFO mapped.Task: Task attempt_local1067668264_0001_r_000000_0 is allowed to commit now
23/10/04 01:53:38 INFO output.FileOutputCommitter: Saved output of task attempt_local1067668264_0001_r_000000_0 to file:/usr/local/hadoop/share/KULKARNI/conf_to_city/BigData/conf_to_city_bd/_temporary/0/task_l
ocal1067668264_0001_r_000000_0
23/10/04 01:53:38 INFO mapped.Task: Task attempt_local1067668264_0001_r_000000_0 done.
23/10/04 01:53:38 INFO mapped.LocalJobRunner: Finishing task: attempt_local1067668264_0001_r_000000_0
23/10/04 01:53:38 INFO mapped.Reduce: Reducer executor complete.
23/10/04 01:53:38 INFO mapred.Job: map 100% reduce 100%
23/10/04 01:53:38 INFO mapred.Job: Job job_local1067668264_0001 completed successfully
23/10/04 01:53:38 INFO mapred.Job: Counters: 30
  File System Operations=918
    FILE: Number of bytes read=49316
    FILE: Number of bytes written=97887
    FILE: Number of large file operations=0
    FILE: Number of write operations=0
  Map-Reduce Framework
    Map input records=918
    Map output records=313
    Map output bytes=7531
    Map output materialized bytes=8163
    Input split bytes=132
    Combine output records=0
    Reduce shuffle bytes=8163
    Reduce input records=313
    Reduce input bytes=263
    Spilled Records=26
    Shuffled Maps =1
    Failed Shuffles=0
    Total committed heap usage (bytes)=337641472
  Shuffle Errors
    IOError=0
    CONNECTION=0
    IO_ERROR=0
    WRONG_LENGTH=0
    WRONG_TYPE=0
    WRONG_REDUCE=0
  File Input Format Counters
    Bytes Read=2343
  File Output Format Counters
    Bytes Written=2343
23/10/04 01:53:38 INFO streaming.StreamJob: Output directory: conf_to_city_bd
hadoop@sanket-Lenovo: /usr/local/hadoop/share/KULKARNI/conf_to_city/BigData/conf_to_city_bd/ part-00000
```

Folder structure for bigdata:



Output for Big Data:

```
Activities Terminal Oct 4 01:54
part-00000
29 [{"Conference_Acronym": "WCLCA", "Research_Area": "BigData", "Cities": ["Lyon", "Antwerp"]}, {"Conference_Acronym": "BCD", "Research_Area": "BigData", "Cities": ["Ho Chi Minh"]}, {"Conference_Acronym": "BDCIA", "Research_Area": "BigData", "Cities": ["Beijing"]}, {"Conference_Acronym": "BDCat", "Research_Area": "BigData", "Cities": ["Taormina"]}, {"Conference_Acronym": "BDEE", "Research_Area": "BigData", "Cities": ["Paris"]}, {"Conference_Acronym": "BDEE-EU", "Research_Area": "BigData", "Cities": ["Chengdu"]}, {"Conference_Acronym": "BDIIE", "Research_Area": "BigData", "Cities": ["London", "Singapore"]}, {"Conference_Acronym": "BDIIE", "Research_Area": "BigData", "Cities": ["Wuhan", "Kuwait"]}, {"Conference_Acronym": "BDIOT", "Research_Area": "BigData", "Cities": ["Beijing"]}, {"Conference_Acronym": "BDL", "Research_Area": "BigData", "Cities": ["London", "Rabat"]}, {"Conference_Acronym": "BDML", "Research_Area": "BigData", "Cities": ["Bordeaux"]}, {"Conference_Acronym": "BDML-EL", "Research_Area": "BigData", "Cities": ["Klagenfurt"]}, {"Conference_Acronym": "BDMO", "Research_Area": "BigData", "Cities": ["Rome"]}, {"Conference_Acronym": "BDPC", "Research_Area": "BigData", "Cities": ["Paris"]}, {"Conference_Acronym": "BDSC", "Research_Area": "BigData", "Cities": ["Macau"]}, {"Conference_Acronym": "BDSN", "Research_Area": "BigData", "Cities": ["Abu Dhabi", "Gandia", "Paris"]}, {"Conference_Acronym": "BDT", "Research_Area": "BigData", "Cities": ["Berlin", "Barcelona", "Alicante"]}, {"Conference_Acronym": "BBCS", "Research_Area": "BigData", "Cities": ["Vienna"]}, {"Conference_Acronym": "BDIOM", "Research_Area": "BigData", "Cities": ["Paris"]}, {"Conference_Acronym": "BDIOM", "Research_Area": "BigData", "Cities": ["Cyprus"]}, {"Conference_Acronym": "BDOMA", "Research_Area": "BigData", "Cities": ["Bangkok"]}, {"Conference_Acronym": "BDIOP", "Research_Area": "BigData", "Cities": ["Soronto"]}, {"Conference_Acronym": "BDIOP", "Research_Area": "BigData", "Cities": ["Soronto"]}, {"Conference_Acronym": "BDIOP", "Research_Area": "BigData", "Cities": ["Exeter"]}, {"Conference_Acronym": "BDIOP", "Research_Area": "BigData", "Cities": ["Osaka"]}, {"Conference_Acronym": "BDIOP", "Research_Area": "BigData", "Cities": ["Paris"]}, {"Conference_Acronym": "CAAM", "Research_Area": "BigData", "Cities": ["Austin"]}, {"Conference_Acronym": "CARMA", "Research_Area": "BigData", "Cities": ["Seville"]}, {"Conference_Acronym": "CBDO", "Research_Area": "BigData", "Cities": ["Hannover", "Aix-en-Provence"]}, {"Conference_Acronym": "CCBDIOT", "Research_Area": "BigData", "Cities": ["Berlin"]}, {"Conference_Acronym": "CCBDIOT", "Research_Area": "BigData", "Cities": ["Chengdu"]}, {"Conference_Acronym": "CCGrdLife", "Research_Area": "BigData", "Cities": ["Taormina"]}, {"Conference_Acronym": "CCIO10", "Research_Area": "BigData", "Cities": ["Kuwait"]}, {"Conference_Acronym": "CCIO10", "Research_Area": "BigData", "Cities": ["Copenhagen"]}, {"Conference_Acronym": "CEECCT", "Research_Area": "BigData", "Cities": ["Nanjing"]}, {"Conference_Acronym": "CIEKMA", "Research_Area": "BigData", "Cities": ["Singapore"]}, {"Conference_Acronym": "CIEKMA", "Research_Area": "BigData", "Cities": ["Atlanta"]}, {"Conference_Acronym": "CIEKMA", "Research_Area": "BigData", "Cities": ["Shenzhen"]}, {"Conference_Acronym": "CIEKSD", "Research_Area": "BigData", "Cities": ["Portland"]}, {"Conference_Acronym": "CPS-AEIS", "Research_Area": "BigData", "Cities": ["London"]}, {"Conference_Acronym": "CPS-AEIS", "Research_Area": "BigData", "Cities": ["London"]}, {"Conference_Acronym": "CST", "Research_Area": "BigData", "Cities": ["Dubai"]}, {"Conference_Acronym": "CSoNet", "Research_Area": "BigData", "Cities": ["Montreal"]}, {"Conference_Acronym": "CFCM", "Research_Area": "BigData", "Cities": ["Vienna"]}], [{"Count": 111}]]
```

D) For each city compute and plot a time series of #conferences per year

Approach:

- Create a mapper which will make a key based on the city and year column from the input tsv file.
- In the reducer, keep a count of the number of times city is repeated
- Additionally, after the output is generated we need to plot the graph
- Use plotly to plot the graph, here keep city year on the x axis and count on the y axis
- Once graph is plotted take screenshots with zoom in as well

Code Snippets:

Mapper:

```

File Edit Selection Find View Goto Tools Project Preferences Help
conf_to_city_reducer_bd.py | core-site.xml | mapred-site.xml | hdfs-site.xml | timeseries.py | timeseries_q4.py | city_year_count_reducer.py | city_year_count_mapper.py
1 #!/usr/bin/env python3
2 import sys
3
4 # Process each line from standard input
5 for line in sys.stdin:
6     # splitting the input line to make a tuple below
7     input_line = line.strip().split('\t')
8     if len(input_line) >= 3:
9         location = input_line[3].strip()      #taking raw city input
10        year = input_line[1].strip()        #taking raw year input
11    # intermediate space printing for reducer
12    print(f'{location}\t{year}\t1')
13
14

```

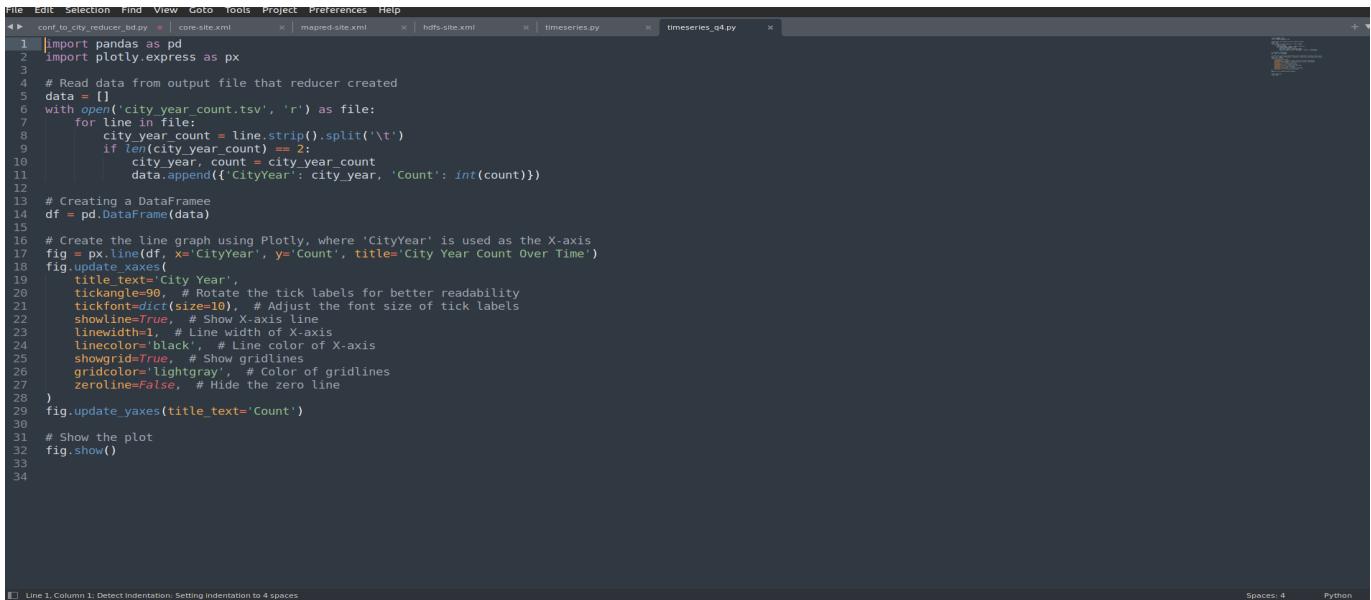
Reducer

```

File Edit Selection Find View Goto Tools Project Preferences Help
conf_to_city_reducer_bd.py | core-site.xml | mapred-site.xml | hdfs-site.xml | timeseries.py | timeseries_q4.py | city_year_count_reducer.py | city_year_count_mapper.py
1 #!/usr/bin/env python3
2 import sys
3
4 current_key = None
5 current_count = 0
6
7 # Process each line from standard input
8 for line in sys.stdin:
9     # Split the line into key and count
10    key, count = line.strip().split('\t')
11
12    # Convert count to an integer
13    count = int(count)
14
15    # If the current key is None, or if it's different from the previous key
16    if current_key is None or key != current_key:
17        # If we have a current key (i.e., not the first key), output its count
18        if current_key is not None:
19            print(f'{current_key}\t{current_count}')
20        # Set the current key and initialize the count
21        current_key = key
22        current_count = count
23    else:
24        # If the key is the same as the current key, increment the count
25        current_count += count
26
27 # Output the last key and its count
28 if current_key is not None:
29    print(f'{current_key}\t{current_count}')
30

```

The input file my plotly is using is part-00000 that hadoop reducer generates, I have just renamed it to city_year_count.tsv

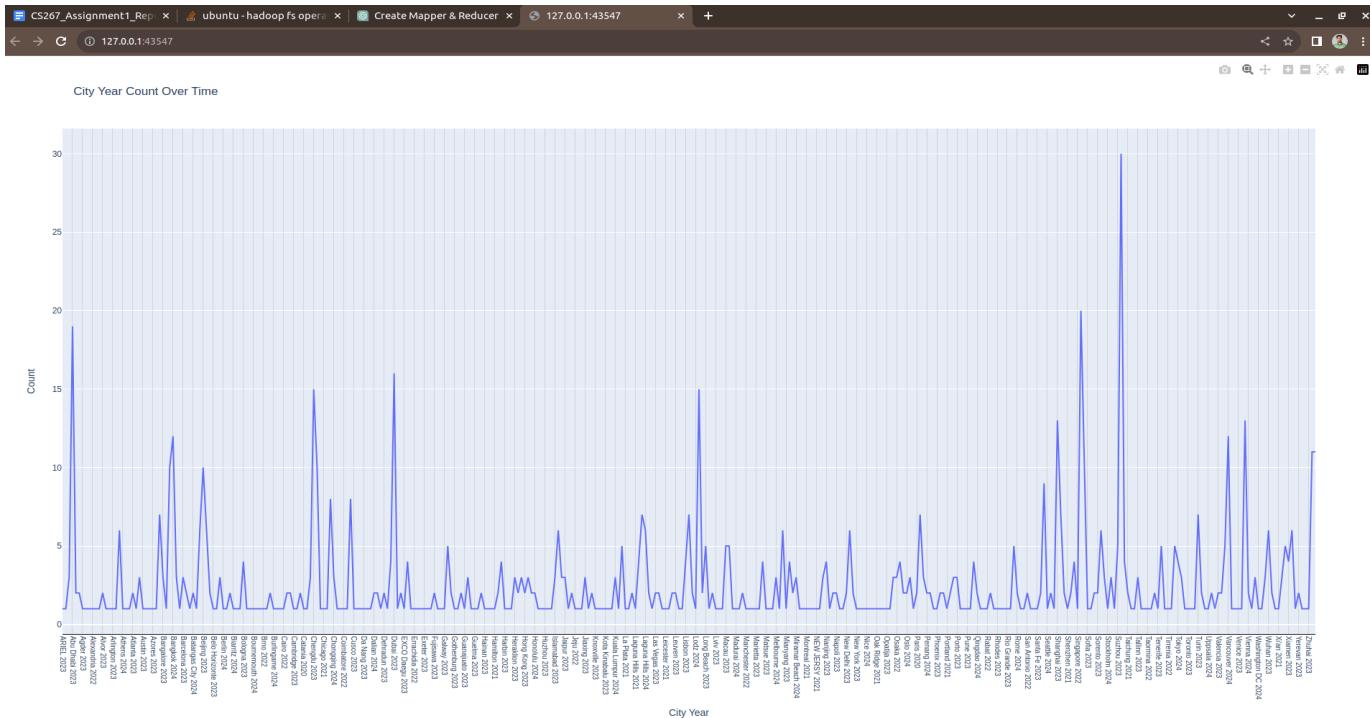


```

File Edit Selection Find View Goto Tools Project Preferences Help
conf_to_city_reducer_bd.py | core-site.xml | hdfs-site.xml | timeseries.py | timeseries_q4.py
1 import pandas as pd
2 import plotly.express as px
3
4 # Read data from output file that reducer created
5 data = []
6 with open('city_year_count.tsv', 'r') as file:
7     for line in file:
8         city_year_count = line.strip().split('\t')
9         if len(city_year_count) == 2:
10             city_year, count = city_year_count
11             data.append({'CityYear': city_year, 'Count': int(count)})
12
13 # Creating a DataFrame
14 df = pd.DataFrame(data)
15
16 # Create the line graph using Plotly, where 'CityYear' is used as the X-axis
17 fig = px.line(df, x='CityYear', y='Count', title='City Year Count Over Time')
18 fig.update_xaxes(
19     title_text='City Year',
20     tickangle=90, # Rotate the tick labels for better readability
21     tickfont_size=10, # Adjust the font size of tick labels
22     showline=True, # Show X-axis line
23     linewidth=1, # Line width of X-axis
24     linecolor='black', # Line color of X-axis
25     showgrid=True, # Show gridlines
26     gridcolor='lightgray', # Color of gridlines
27     zeroline=False, # Hide the zero line
28 )
29 fig.update_yaxes(title_text='Count')
30
31 # Show the plot
32 fig.show()
33
34

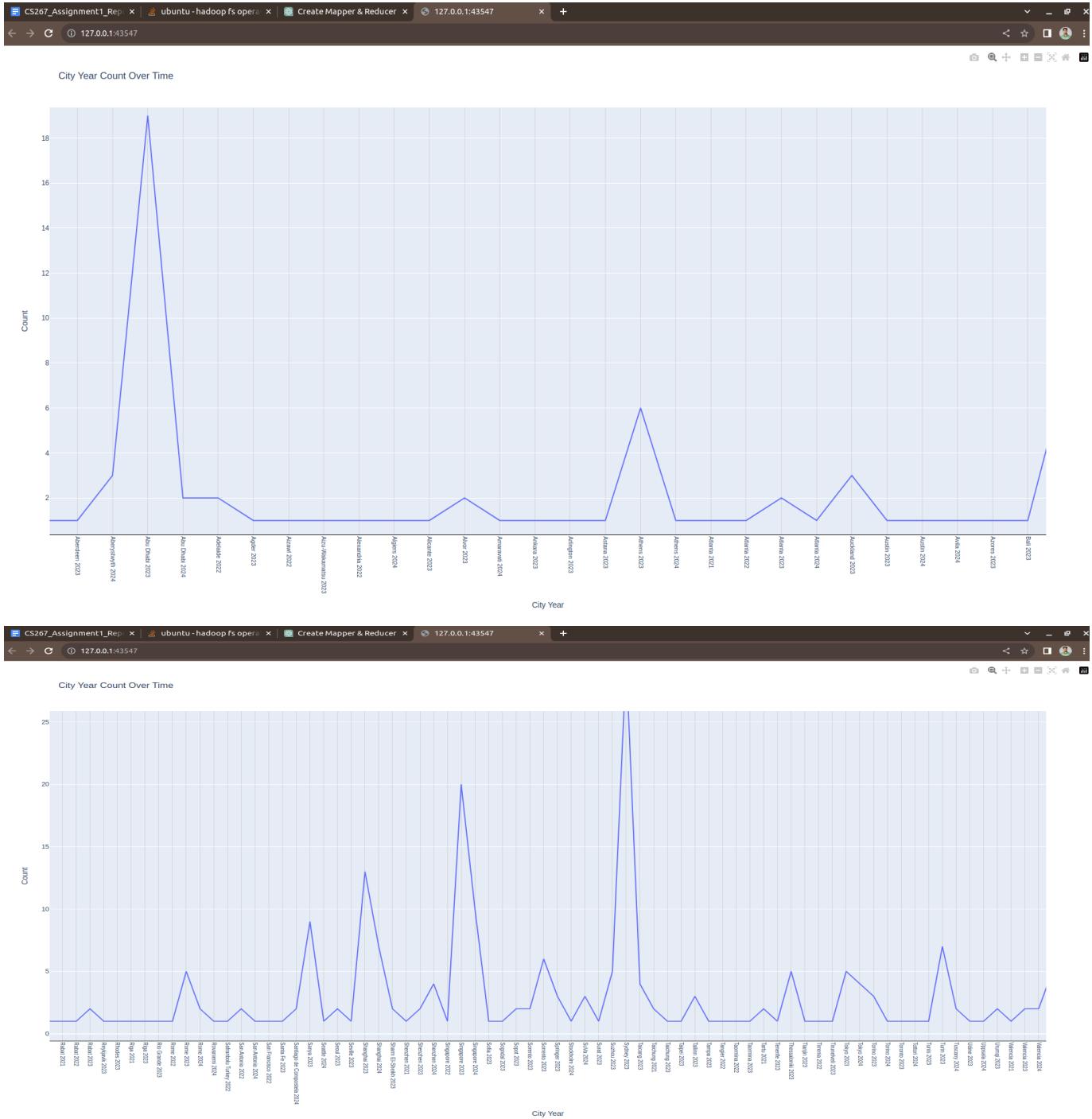
```

Graph of conference count per city per year



Sanket Kulkarni : HW2

Zoom in of the screenshots



Outcomes:

Able to run the web crawlers while complying to the policies. Understood Beautiful Soap of Python for crawling

Understood data cleaning using an open refine tool. Got knowledge of facet creations, text filters and clustering

Understood Hadoop ran for combined keys

Understood Graph plotting using plotly library

References:

<https://openrefine.org/docs/manual/facets>

<https://librarycarpentry.org/lc-open-refine/10-data-transformation.html>

<https://realpython.com/beautiful-soup-web-scraping-with-python/#:~:text=Beautiful%20Soup%20is%20a%20Python,web%20page%20using%20developer%20tools.>

<https://plotly.com/python/>

<https://elixirforum.com/t/create-a-new-map-with-multiple-keys-through-each-iteration-of-map-new/44990>

<https://groups.google.com/g/openrefine/c/9LB9OT0q2rY>

<https://openrefine.org/docs/manual/cellediting#:~:text=One%20way%20of%20doing%20this,the%20cell%20in%20the%20facet.>
