A BIG DATA PROJECT ON

Data-Driven Insights for Enhanced Hospital Care: A Hadoop-Based Analytical Approach

Asmita Mondal [2348018]

PROJECT GUIDE: Dr. Rajesh R

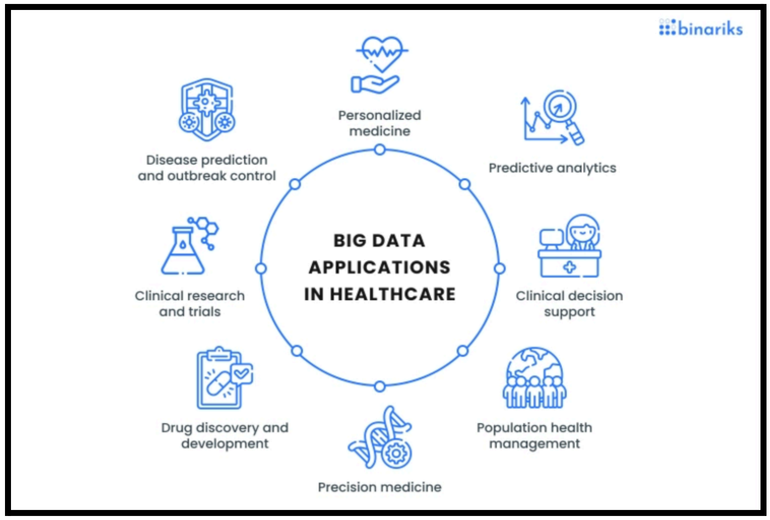
Submitted to the Department of Statistics and Data Science in partial fulfillment of the requirements for the degree of M.Sc. Data Science

TABLE OF CONTENTS

Introduction

Background

Hospitals generate massive amounts of data daily, ranging from patient demographics, admission details, diagnoses, and treatments to outcomes. This data, often unstructured or semi-structured, qualifies as "big data" due to its volume, velocity, and variety. Managing and analyzing such data is a challenge but is essential for driving operational efficiency and improving patient care.



With the rise of chronic diseases, aging populations, and the increasing complexity of healthcare delivery, the need to optimize hospital workflows has become critical. Properly leveraging hospital data can help identify trends, predict resource needs, and ultimately enhance decision-making for patient admission, treatment, and discharge processes.

Managing Hospital Data

Optimizing hospital operations requires analyzing diverse data points, such as patient medical histories, ICU stay durations, chronic conditions, and lifestyle habits. By integrating and processing these datasets, hospitals can gain insights into resource utilization, disease prevalence, and treatment efficacy. For example:

- Predicting ICU bed occupancy to reduce bottlenecks.
- Identifying patient demographics associated with specific outcomes.
- Assessing the influence of lifestyle factors like smoking and alcohol on treatment success rates.

The challenges in managing such data include:

1. **Data Volume**: Hospitals generate terabytes of data, making manual analysis impractical.
2. **Data Integration**: Combining datasets from multiple sources, such as patient records and auxiliary data like health outcomes.
3. **Real-Time Insights**: Deriving actionable insights promptly for time-sensitive decisions.

Problem Statement

With the increasing complexity of healthcare delivery and the need for data-driven decision-making, it becomes essential to uncover actionable insights from patient records. The goal is to analyze hospital data effectively to ensure better resource utilization, improve patient care, and identify key trends influencing hospital operations.

Objectives

1. **Preprocessing and Data Cleansing**: Perform basic preprocessing to handle null values, validate data integrity, and prepare the dataset for meaningful analysis.
2. **Average Duration of Stay Analysis**: Calculate the average duration of stay (DoS) for patients, helping to understand trends in hospitalization durations across various demographics.
3. **Outcome Analysis by Demographics using Partitioner**: Analyze patient outcomes (discharge, expiry, and DAMA) by demographic factors such as age, gender, and location using partitioning to improve data organization.
4. **Chronic Condition and Lifestyle Influence on Outcomes using Partitioner**: Assess the impact of chronic conditions (e.g., diabetes, hypertension) and lifestyle factors (e.g., smoking, alcohol) on patient outcomes using partitioning to segregate results effectively.
5. **ICU Stay Duration Analysis**: Explore the relationship between ICU stay duration and other factors like admission type, type of treatment, and patient outcomes to optimize ICU resource allocation.
6. **Outcome Remarks Integration using Reducer Side Join**: Use a reducer-side join to integrate hospital outcome data with a remarks dataset, providing contextual information for each outcome and facilitating better interpretability of results.

Dataset Description

The dataset, titled "Hospital Admission Data", was sourced from Kaggle and is provided under the Creative Commons License (Attribution-Non-Commercial-Share Alike 4.0 International (CC BY-NC-SA 4.0)).

This dataset was collected from patients admitted over a span of two years, from April 1, 2017, to March 31, 2019, at the Hero DMC Heart Institute, a unit of Dayanand Medical College and Hospital, located in Ludhiana, Punjab, India.

Key facts about the dataset:

- Admissions: The cardiology unit recorded a total of 14,845 admissions during the study period.
- Patients: These admissions correspond to 12,238 unique patients.
- Re-admissions: A subset of 1,921 patients accounted for multiple admissions,highlighting the need for further analysis of readmission trends and their implications.

Data Structure

Column	Sample Values	Description
MRD No.	234735, 234696	Unique patient identifier (Medical Record Number).
AGE	81, 65, 53	Age of the patient in years.
GENDER	M	Gender of the patient (M for Male, F for Female).
RURAL	R, U	Rural (R) or Urban (U) residence of the patient.

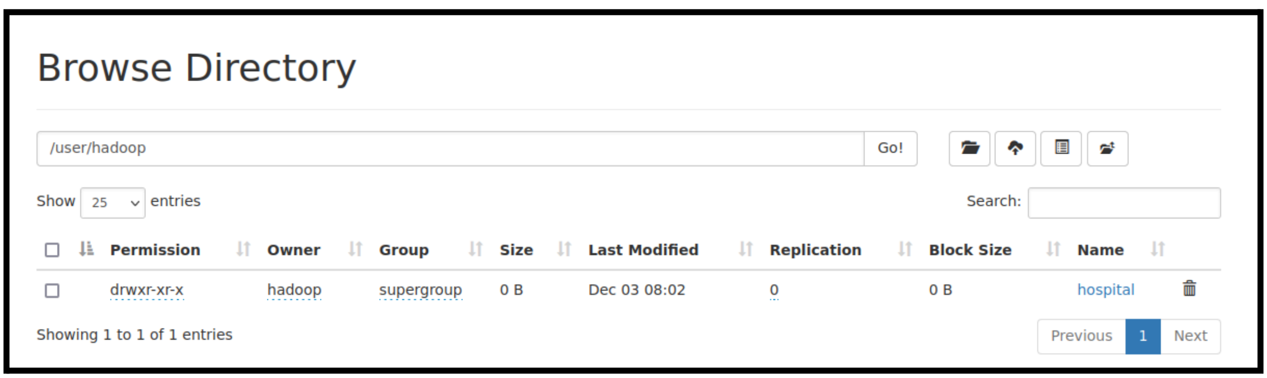
TYPE OF ADMISSION	E (Emergency) O (Outpatient)	Type of hospital admission: Emergency (E) or Outpatient Department (O).
month year	Apr-17	Month and year of admission.
DURATION OF STAY	3, 5	Total number of days the patient stayed in the hospital.
duration of intensive unit stay	2, 3	Number of days the patient spent in the Intensive Care Unit (ICU).
OUTCOME	DISCHARGE, DAMA, EXPIRY	Outcome of the hospital stay (e.g., DISCHARGE, etc.).
SMOKING	0, 1	Smoking history: 1 for smoker, 0 for non-smoker.
ALCOHOL	0, 1	Alcohol consumption: 1 for drinker, 0 for non-drinker.
DM	1, 0	Presence of Diabetes Mellitus: 1 for yes, 0 for no.
HTN	0, 1	Presence of Hypertension: 1 for yes, 0 for no.
CD	0, 1	Likely Coronary Disease (similar to CAD): 1 for yes, 0 for no.

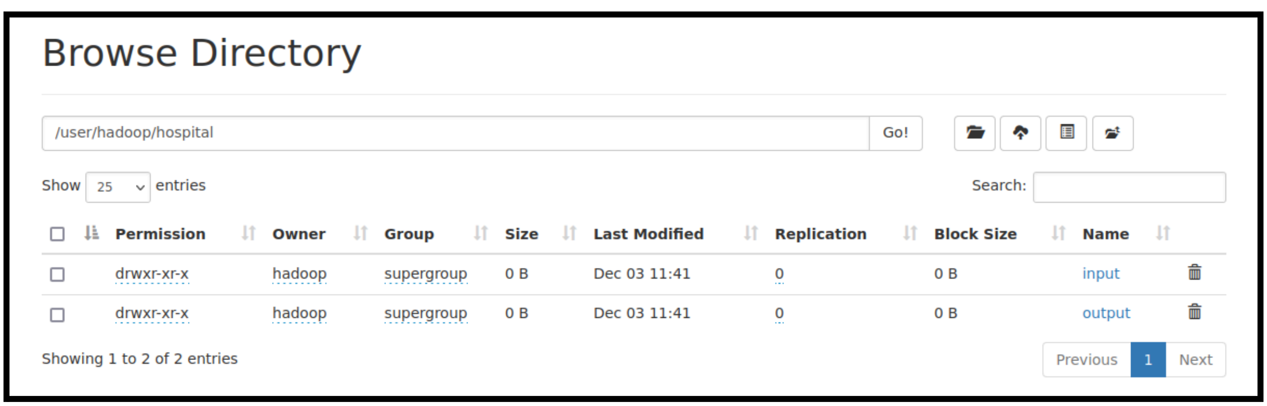
PRIOR CMP	0	Prior cardiac complications: 1 for yes, 0 for no.
CKD	0, 1	Chronic Kidney Disease: 1 for yes, 0 for no.
HB	9.5, 13.7	Hemoglobin level in g/dL.
TLC	16.1, 9	Total Leukocyte Count (WBC count).
PLATELETS	337, 149	Platelets count in thousands per microliter.
GLUCOSE	80, 112, 187	Blood glucose level in mg/dL.
UREA	34, 18, 93	Blood urea level in mg/dL.
CREATININE	0.9, 2.3	Blood creatinine level in mg/dL.
RAISED CARDIAC ENZYMES	1, 0	Presence of raised cardiac enzymes: 1 for yes, 0 for no.
EF	35, 42	Ejection Fraction (percentage of blood pumped out of the heart during each beat).
SEVERE ANAEMIA	0, 1	Severe anemia status: 1 for yes, 0 for no.
ANAEMIA	1, 0	General anemia status: 1 for yes, 0 for no.

Hadoop Setup

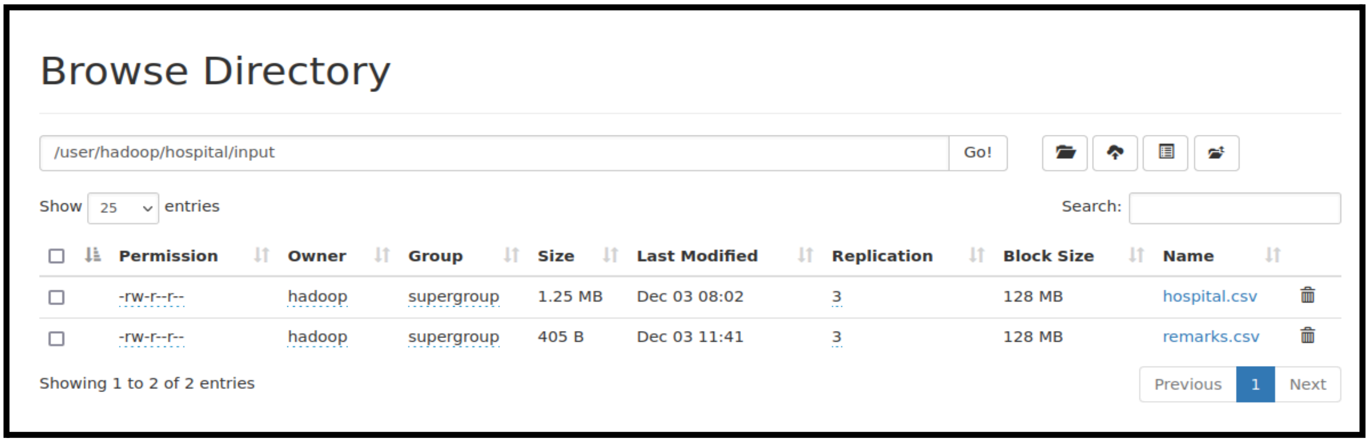
To perform the analysis, Hadoop was set up in a virtual environment, and the following steps were undertaken to configure the system and prepare it for MapReduce operations:

- **Environment Setup**
  - Launched VirtualBox with a pre-configured virtual machine running Ubuntu.
  - Started the Ubuntu operating system and initialized the Hadoop services.
- **Starting Hadoop**
  - Opened the terminal in Ubuntu and executed the following commands to start Hadoop services:
    - start-dfs.sh
    - start-yarn.sh
  - Verified the services were running by accessing the **Hadoop Web UI** through `http://localhost:9870`
- **HDFS Directory Preparation**
  - Created the necessary directories for storing input and output data in the Hadoop Distributed File System (HDFS). The following commands were used:
    - hdfs dfs -mkdir /user/hadoop/hospital
    - hdfs dfs -mkdir /user/hadoop/hospital/input
    - hdfs dfs -mkdir /user/hadoop/hospital/output





- Uploaded the input files (`hospital.csv` and `remarks.csv`) into the `input` directory:
  - hdfs dfs -put /path/to/hospital.csv /user/hadoop/hospital/input
  - hdfs dfs -put /path/to/remarks.csv /user/hadoop/hospital/input



Browse Directory

/user/hadoop/hospital/input Go!

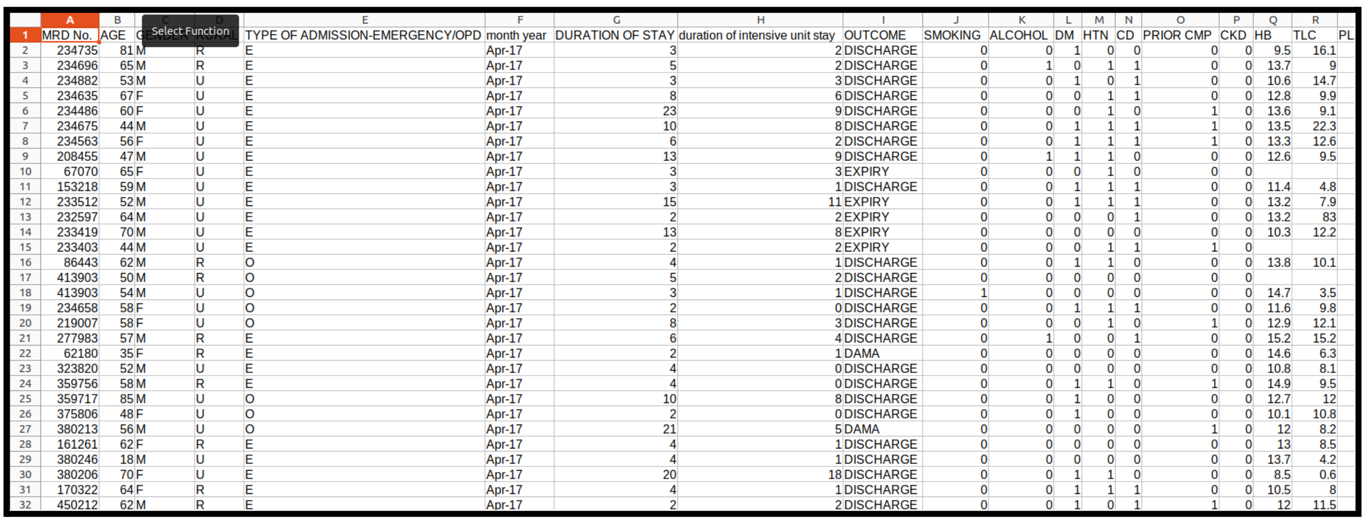
Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	1.25 MB	Dec 03 08:02	3	128 MB	hospital.csv
-rw-r--r--	hadoop	supergroup	405 B	Dec 03 11:41	3	128 MB	remarks.csv

Showing 1 to 2 of 2 entries

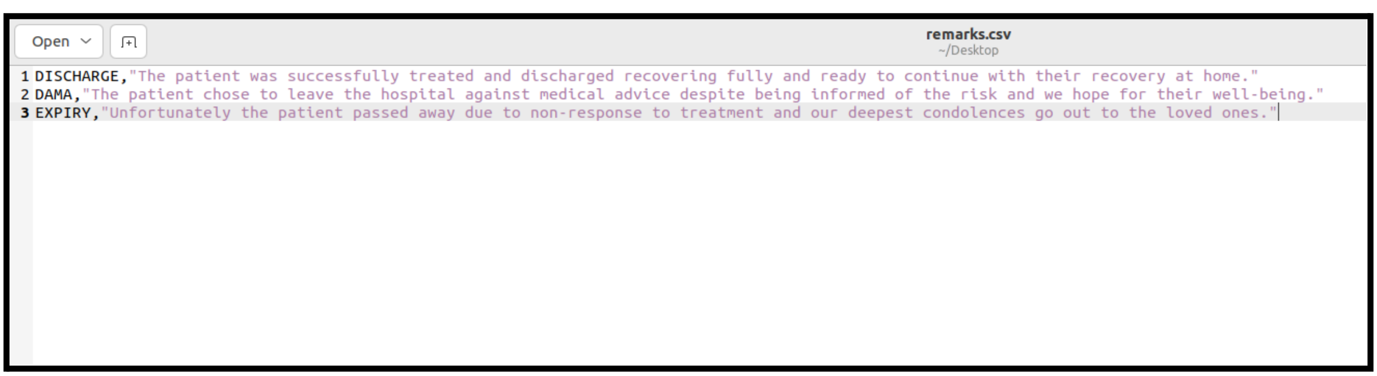
Previous 1 Next

Hospital Data Snippet:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	MRD No.	AGE	GENDER	RURAL	TYPE OF ADMISSION-EMERGENCY/OPD	month year	DURATION OF STAY	duration of intensive unit stay	OUTCOME	SMOKING	ALCOHOL	DM	HTN	CD	PRIOR CMP	CKD	HB	TLC
2	234735	81	M	R	E	Apr-17	3	2	DISCHARGE	0	0	1	0	0	0	0	9.5	16.1
3	234696	65	M	R	E	Apr-17	5	2	DISCHARGE	0	1	0	1	1	0	0	13.7	9
4	234882	53	M	U	E	Apr-17	3	3	DISCHARGE	0	0	1	0	1	0	0	10.6	14.7
5	234635	67	F	U	E	Apr-17	8	6	DISCHARGE	0	0	0	1	1	0	0	12.8	9.9
6	234486	60	F	U	E	Apr-17	23	9	DISCHARGE	0	0	0	1	0	1	0	13.6	9.1
7	234675	44	M	U	E	Apr-17	10	8	DISCHARGE	0	0	1	1	1	1	0	13.5	22.3
8	234563	56	F	U	E	Apr-17	6	2	DISCHARGE	0	0	1	1	1	1	0	13.3	12.6
9	208455	47	M	U	E	Apr-17	13	9	DISCHARGE	0	1	1	1	0	0	0	12.6	9.5
10	67070	65	F	U	E	Apr-17	3	3	EXPIRY	0	0	0	1	0	0	0		
11	153218	59	M	U	E	Apr-17	3	1	DISCHARGE	0	0	1	1	1	0	0	11.4	4.8
12	233512	52	M	U	E	Apr-17	15	11	EXPIRY	0	0	1	1	1	0	0	13.2	7.9
13	232597	64	M	U	E	Apr-17	2	2	EXPIRY	0	0	0	0	1	0	0	13.2	83
14	233419	70	M	U	E	Apr-17	13	8	EXPIRY	0	0	0	0	0	0	0	10.3	12.2
15	233403	44	M	U	E	Apr-17	2	2	EXPIRY	0	0	0	1	1	1	0		
16	86443	62	M	R	O	Apr-17	4	1	DISCHARGE	0	0	1	1	0	0	0	13.8	10.1
17	413903	50	M	R	O	Apr-17	5	2	DISCHARGE	0	0	0	0	0	0	0		
18	413903	54	M	U	O	Apr-17	3	1	DISCHARGE	1	0	0	0	0	0	0	14.7	3.5
19	234658	58	F	U	O	Apr-17	2	0	DISCHARGE	0	0	1	1	1	0	0	11.6	9.8
20	219007	58	F	U	O	Apr-17	8	3	DISCHARGE	0	0	0	1	0	1	0	12.9	12.1
21	277983	57	M	R	E	Apr-17	6	4	DISCHARGE	0	1	0	0	1	0	0	15.2	15.2
22	62180	35	F	R	E	Apr-17	2	1	DAMA	0	0	0	0	0	0	0	14.6	6.3
23	323820	52	M	U	E	Apr-17	4	0	DISCHARGE	0	0	0	0	0	0	0	10.8	8.1
24	359756	58	M	R	E	Apr-17	4	0	DISCHARGE	0	0	1	1	0	1	0	14.9	9.5
25	359717	85	M	U	O	Apr-17	10	8	DISCHARGE	0	0	1	0	0	0	0	12.7	12
26	375806	48	F	U	O	Apr-17	2	0	DISCHARGE	0	0	1	0	0	0	0	10.1	10.8
27	380213	56	M	U	O	Apr-17	21	5	DAMA	0	0	0	0	0	1	0	12	8.2
28	161261	62	F	R	E	Apr-17	4	1	DISCHARGE	0	0	0	0	0	0	0	13	8.5
29	380246	18	M	U	E	Apr-17	4	1	DISCHARGE	0	0	0	0	0	0	0	13.7	4.2
30	380206	70	F	U	E	Apr-17	20	18	DISCHARGE	0	0	1	1	0	0	0	8.5	0.6
31	170322	64	F	R	E	Apr-17	4	1	DISCHARGE	0	0	1	1	1	0	0	10.5	8
32	450212	62	M	R	E	Apr-17	2	2	DISCHARGE	0	0	1	0	1	1	0	12	11.5

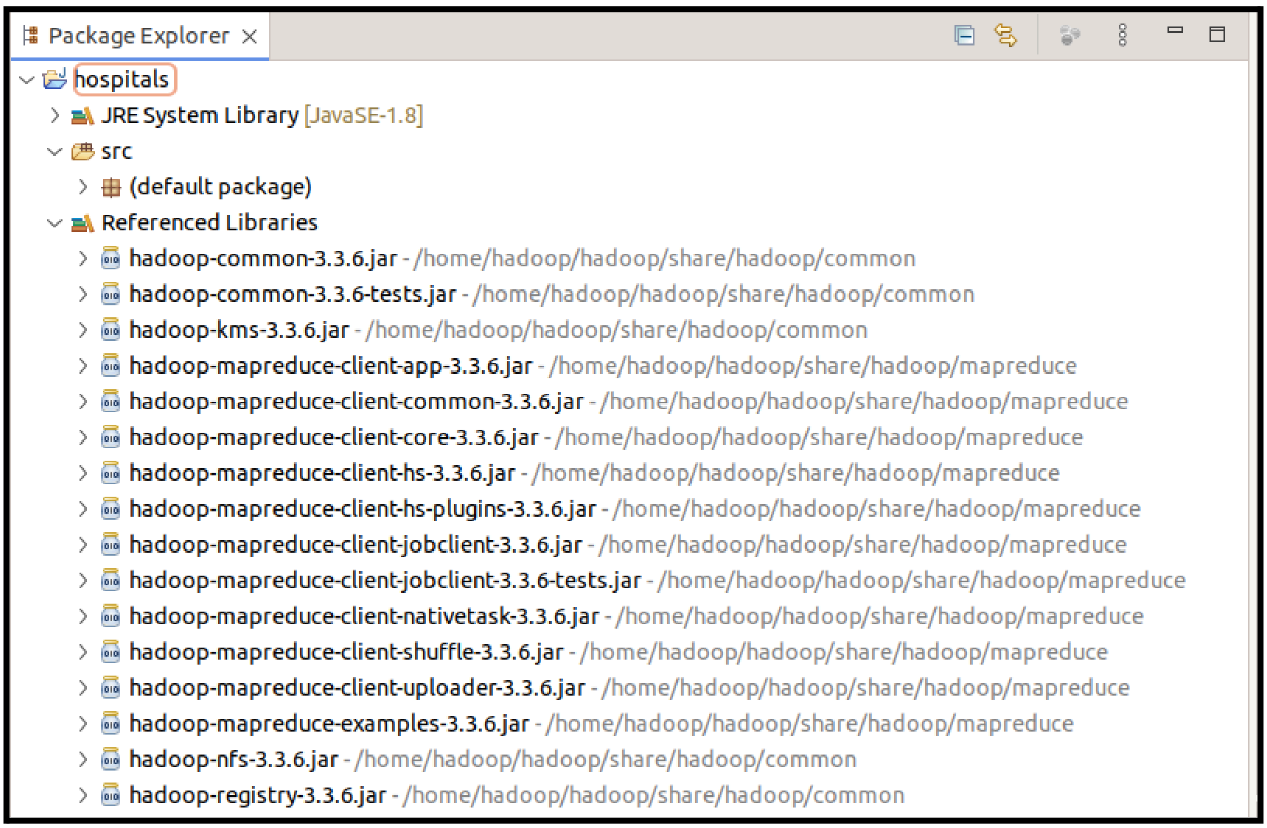
Remarks:



remarks.csv
~/Desktop

```
DISCHARGE,"The patient was successfully treated and discharged recovering fully and ready to continue with their recovery at home."
DAMA,"The patient chose to leave the hospital against medical advice despite being informed of the risk and we hope for their well-being."
EXPIRY,"Unfortunately the patient passed away due to non-response to treatment and our deepest condolences go out to the loved ones."
```

- **Eclipse Configuration for MapReduce Development**
  - Opened Eclipse IDE and created a new Java project named **‘hospitals’**.
  - Worked under the **default** package under the project for organizing all related MapReduce programs.
  - Ensured the Hadoop libraries were correctly referenced to avoid build or runtime errors.
  - Configured the Build Path for the project by adding the necessary external JAR files required for Hadoop and MapReduce operations:
    - hadoop-common-*.jar
    - hadoop-mapreduce-client-core-*.jar
    - commons-cli-*.jar
    - Other related dependencies.



Analysis

A. Preprocessing and Result Display

Method Used

- **MapReduce**: A distributed processing technique used to clean and preprocess the dataset, ensuring that meaningful insights can be extracted from the data.

Code Link

The full code for the preprocessing task is provided here:
https://github.com/AsmitaMondal/hospital-analysis/blob/main/codes/HospitalAnalysis.java

Code Explanation

The program performs preprocessing and categorization tasks using the following logic:

1. **Mapper**:
   - Processes each record in the dataset.
   - Categorizes data based on patient demographics, admission details, and outcomes.
   - Skips invalid or incomplete rows and focuses on valid data points.
2. **Reducer**:
   - Aggregates the counts for each category or key emitted by the mapper.
   - Outputs the final tallies for different demographic, clinical, and outcome-based metrics.
3. **Driver**:
   - Configures the MapReduce job by setting the Mapper, Reducer, and the input/output key-value classes.
   - Specifies the input and output paths for the Hadoop Distributed File System (HDFS).

Inputs

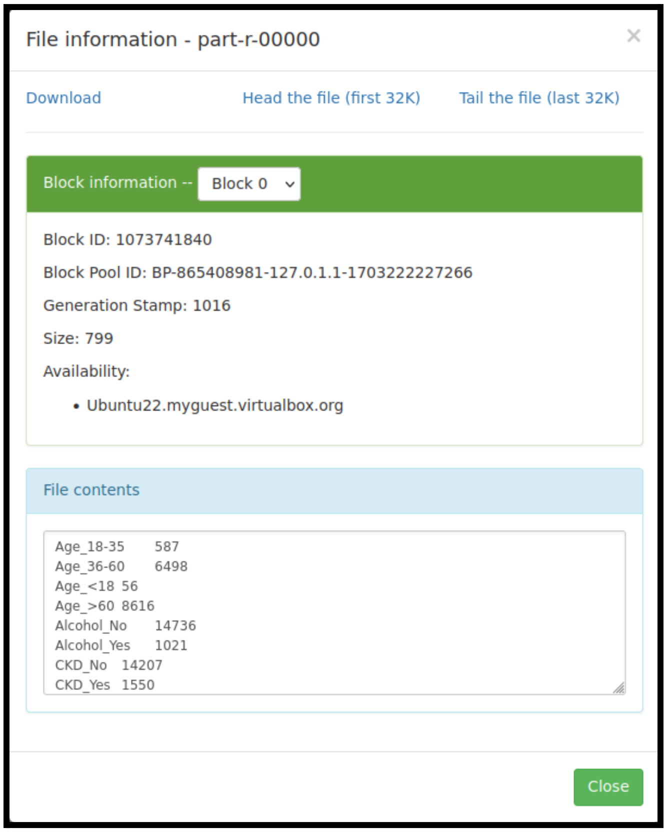
- **Hospital Dataset (`hospital.csv`)**: This dataset contains information about patient demographics, medical conditions, admission details, and outcomes.

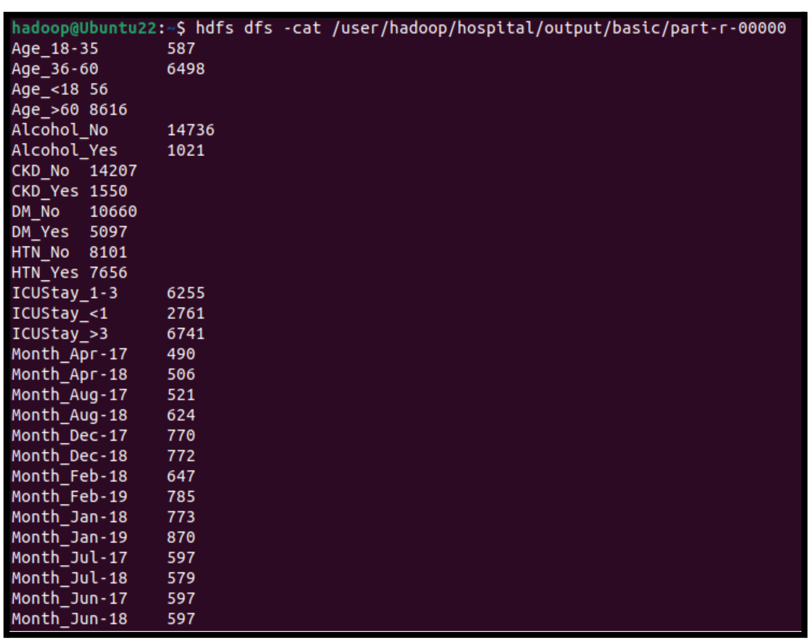
Executing Commands

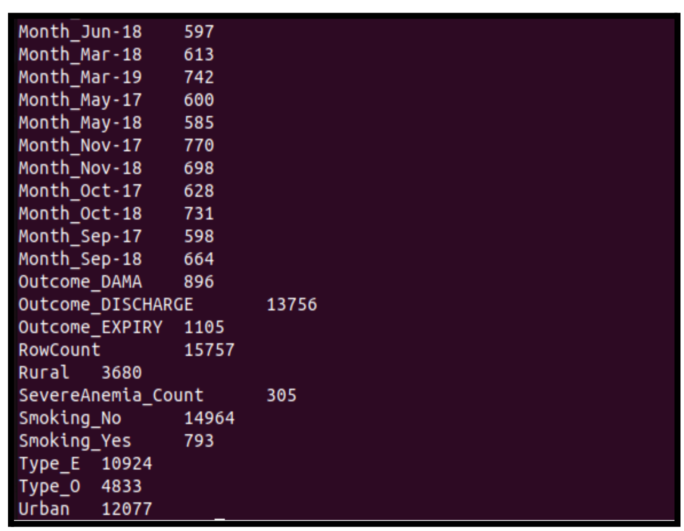
1. **Place Input File in HDFS**:
   ```
   hdfs dfs -put /path/to/hospital.csv /user/hadoop/hospital/input
   ```

2. **Run the MapReduce Job**:
   `hadoop jar HospitalAnalysis.jar HospitalAnalysis /user/hadoop/hospital/input /user/hadoop/hospital/output/basic`
3. **View the Output**:
   `hdfs dfs -cat /user/hadoop/hospital/output/basic/part-r-00000`

Outputs







Interpretation of Output

1. **Age Distribution**:
   - Most patients belong to the 36-60 and >60 age groups, indicating that older adults are the primary demographic.
2. **Alcohol and Smoking**:
   - A significant proportion of patients reported no alcohol (14,736) or smoking habits (14,964).
3. **Chronic Conditions**:
   - Patients without CKD, diabetes (DM), or hypertension (HTN) are in the majority. However, a significant number of patients are affected by these conditions.
4. **ICU Stay Duration**:
   - ICU stays of 1-3 days (6,255) are the most common, followed by longer stays (>3 days).
5. **Outcome Distribution**:
   - The majority of patients were discharged (13,756), followed by cases of DAMA (896) and expiry (1,105).
6. **Geographical Data**:
   - Urban admissions significantly outnumber rural admissions.
7. **Monthly Trends**:
   - Admissions show consistency across months, with slight peaks in December and January.

The preprocessing task ensured that the dataset was cleansed and categorized, paving the way for subsequent analyses. The insights gained provide a broad overview of hospital operations, patient demographics, and clinical outcomes.

B. Average Duration of Stay for Patients

How does the average length of stay in the hospital vary by age group, admission type, and treatment type, and what implications can be drawn for hospital resource management?

Method Used

- **MapReduce**: Utilized to compute the average duration of stay (DoS) for all patients using distributed processing. This involves summing up all stay durations and dividing by the number of patients.

Code Link

The full code for the analysis can be accessed here:
https://github.com/AsmitaMondal/hospital-analysis/blob/main/codes/AverageDurationOfStay.java

Code Explanation

1. **Mapper**:
   - Reads input hospital data and extracts the **Duration of Stay** column (assumed to be at index 6 in the dataset).
   - Skips rows with invalid or missing values using exception handling.
   - Emits a key-value pair where the key is a fixed label (`"DoS"`) and the value is the extracted duration as an `IntWritable`.
2. **Reducer**:
   - Aggregates all durations for the `"DoS"` key received from the mapper.
   - Computes the average duration by dividing the total duration by the number of entries.
   - Emits the key `"DoS"` and the computed average as a `DoubleWritable`.

Inputs

- **Hospital Dataset (`hospital.csv`)**: Contains patient details, including the `Duration of Stay` column.

Executing Commands

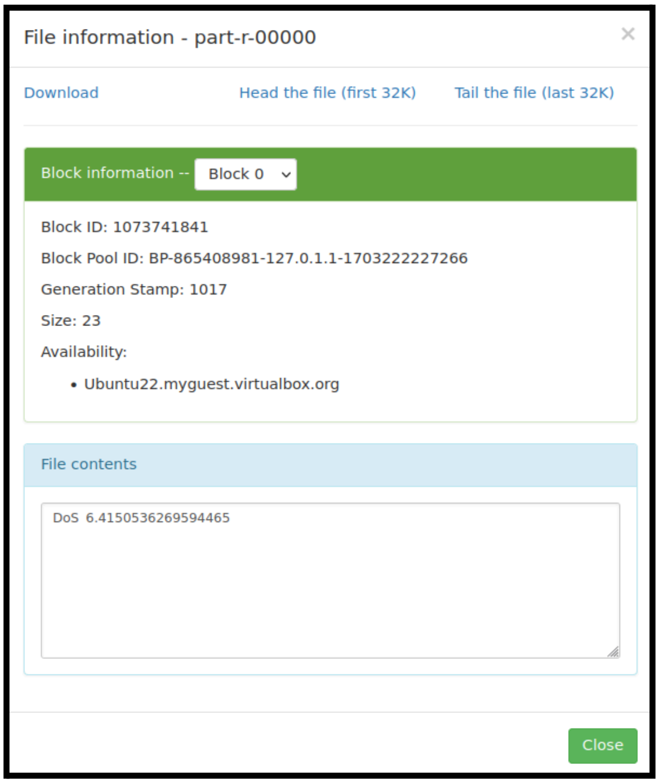
1. **Run the MapReduce Job**:

```
hadoop jar AverageDurationOfStay.jar AverageDurationOfStay
/user/hadoop/hospital/input /user/hadoop/hospital/output/avg_dos
```

2. **View the Output**:

```
hdfs dfs -cat /user/hadoop/hospital/output/avg_dos/part-r-00000
```

Outputs



Interpretation of Output

- **Average Duration of Stay**: The average length of stay for patients in the hospital is approximately **6.42 days**.
- **Insights**:
    - This metric provides a benchmark for hospital administrators to evaluate the efficiency of patient management and treatment protocols.
    - A high average duration could indicate delays in treatment or discharge processes, whereas a very low duration might point to an emphasis on shorter hospital stays.

This analysis forms a foundational step in understanding the overall patient flow and the hospital's capacity to manage patient admissions effectively.

C. Outcome Analysis by Demographics using Partitioner

How do demographic factors (e.g., age, gender) correlate with patient mortality rates, and how can this information help hospitals improve care protocols for high-risk groups?

Method Used

- **MapReduce with Partitioner**: The task involves categorizing hospital outcomes (`DAMA`, `DISCHARGE`, and `EXPIRY`) based on **age group** and **gender**. A **partitioner** is employed to split the outcomes into separate partitions, each handled by its reducer.

Code Link

The full code for the analysis can be accessed here:
https://github.com/AsmitaMondal/hospital-analysis/blob/main/codes/OutcomeAnalysis.java

Code Explanation

1. **Mapper**:
    - Reads the input hospital data and extracts the `AGE`, `GENDER`, and `OUTCOME` fields (assumed to be at indices 1, 2, and 8, respectively).
    - Categorizes `AGE` into groups:
        - `<18`, `18-35`, `36-60`, and `>60`.
    - Generates composite keys in the format: **`Outcome:AgeCategory,Gender`** (e.g., `"DAMA:36-60,F"`).
    - Emits the key and a count of `1` for each record.
2. **Partitioner**:
    - Directs keys to partitions based on the `OUTCOME`:
        - Partition `0`: `DAMA`
        - Partition `1`: `DISCHARGE`
        - Partition `2`: `EXPIRY`
    - Ensures that data for the same `OUTCOME` is processed by the corresponding reducer.
3. **Reducer**:
    - Aggregates the count of occurrences for each demographic category within its assigned partition.
    - Outputs the composite key and the total count for that key.
4. **Driver**:
    - Sets the number of reducers to **3** (one for each outcome).

Inputs

- **Hospital Dataset (`hospital.csv`)**: Contains patient details, including columns for `AGE`, `GENDER`, and `OUTCOME`.

Executing Commands

1. **Run the MapReduce Job**:

```
hadoop jar OutcomeAnalysis.jar OutcomeAnalysis
/user/hadoop/hospital/input
/user/hadoop/hospital/output/outcome_analysis
```

2. **View the Output for Each Partition**:
   - **Partition 0 (DAMA)**:

   ```
   hdfs dfs -cat
   /user/hadoop/hospital/output/outcome_analysis/part-r-00000
   ```

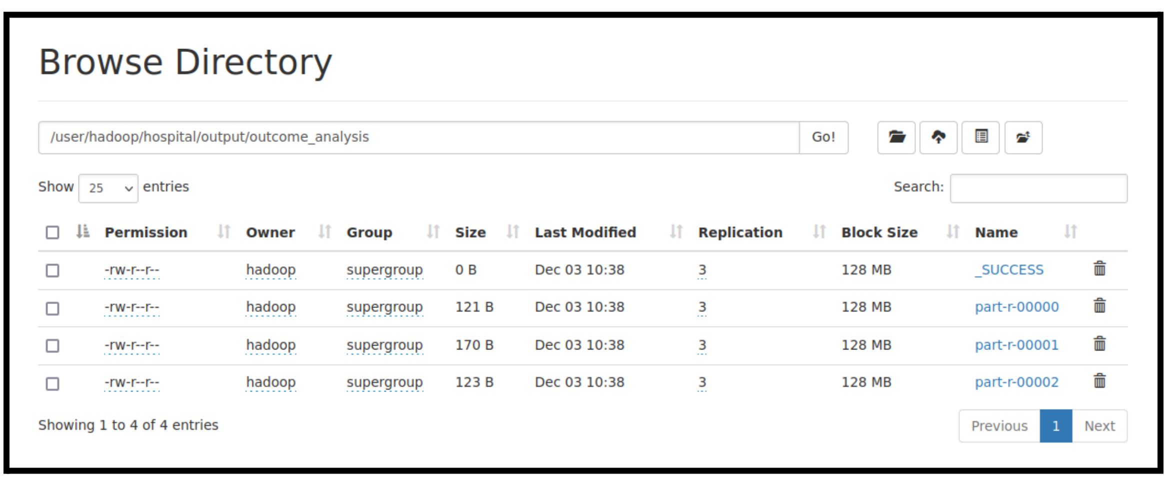
   - **Partition 1 (DISCHARGE)**:

   ```
   hdfs dfs -cat
   /user/hadoop/hospital/output/outcome_analysis/part-r-00001
   ```

   - **Partition 2 (EXPIRY)**:

   ```
   hdfs dfs -cat
   /user/hadoop/hospital/output/outcome_analysis/part-r-00002
   ```

Outputs



	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
☐	-rw-r--r--	hadoop	supergroup	0 B	Dec 03 10:38	3	128 MB	_SUCCESS
☐	-rw-r--r--	hadoop	supergroup	121 B	Dec 03 10:38	3	128 MB	part-r-00000
☐	-rw-r--r--	hadoop	supergroup	170 B	Dec 03 10:38	3	128 MB	part-r-00001
☐	-rw-r--r--	hadoop	supergroup	123 B	Dec 03 10:38	3	128 MB	part-r-00002

```
hadoop@Ubuntu22:~$ hdfs dfs -cat /user/hadoop/hospital/output/outcome_analysis/part-r-00000
DAMA:18-35,F    5
DAMA:18-35,M    32
DAMA:36-60,F    122
DAMA:36-60,M    233
DAMA:<18,F      1
DAMA:<18,M      1
DAMA:>60,F      183
DAMA:>60,M      319
hadoop@Ubuntu22:~$ hdfs dfs -cat /user/hadoop/hospital/output/outcome_analysis/part-r-00001
DISCHARGE:18-35,F       182
DISCHARGE:18-35,M       333
DISCHARGE:36-60,F       2078
DISCHARGE:36-60,M       3729
DISCHARGE:<18,F 28
DISCHARGE:<18,M 23
DISCHARGE:>60,F 2765
DISCHARGE:>60,M 4618
hadoop@Ubuntu22:~$ hdfs dfs -cat /user/hadoop/hospital/output/outcome_analysis/part-r-00002
EXPIRY:18-35,F  13
EXPIRY:18-35,M  22
EXPIRY:36-60,F  125
EXPIRY:36-60,M  211
EXPIRY:<18,M    3
EXPIRY:>60,F    265
EXPIRY:>60,M    466
```

Interpretation of Output

- The output provides insights into demographic trends for each outcome:
  - **DAMA** (Discharge Against Medical Advice): Higher among males aged `>60`.
  - **DISCHARGE**: Most common outcome, especially among males aged `>60`.
  - **EXPIRY**: Most frequent among males aged `>60`, with a significant gender disparity.
  - **Males** consistently have higher counts across all outcomes (`DAMA`, `DISCHARGE`, and `EXPIRY`), especially in the `>60` age group, indicating they might require more targeted healthcare interventions.
  - **Females** show relatively lower counts in `DAMA` and `EXPIRY`, suggesting potential differences in health-seeking behavior or outcomes between genders.
- **Insights**:
  - The data highlights potential age and gender-related trends in hospital outcomes.
  - Administrators can use this analysis to tailor hospital policies for specific demographic groups, particularly for high-risk categories (e.g., males aged `>60`).

D. Chronic Condition and Lifestyle Influence on Outcomes

How do chronic conditions (like Diabetes, Smoking and Hypertension) impact the hospital outcomes (discharge, expiry, etc.) of patients?

Method Used

- **MapReduce**: Applied to analyze the influence of chronic conditions (such as diabetes and hypertension) and lifestyle factors (like smoking and alcohol use) on hospital outcomes (DAMA, DISCHARGE, EXPIRY) using partitioned data for chronic conditions and lifestyle factors.

Code Link

The full code for the analysis can be accessed here:
https://github.com/AsmitaMondal/hospital-analysis/blob/main/codes/ChronicLifestyleAnalysis.java

Code Explanation

1. **Mapper**:
    - Extracts relevant columns for **Chronic Conditions** (DM, HTN) and **Lifestyle Factors** (smoking, alcohol) from the input data.
    - Constructs keys based on combinations of chronic conditions and lifestyle factors, and maps them to the corresponding outcome.
    - Outputs key-value pairs where the key is a string indicating the condition combination (e.g., `Chronic:DM_Yes_HTN_No:DAMA`) and the value is `1`.
2. **Partitioner**:
    - Partitions the data based on whether the key pertains to **Chronic Conditions** or **Lifestyle Factors**, directing them to different reducers for independent processing.
3. **Reducer**:
    - Aggregates the counts for each key (combination of conditions and outcomes).
    - Outputs the key and its corresponding count, representing the number of occurrences for each combination of chronic condition/lifestyle factor and outcome.
4. **Driver**:
    - Sets number of reducer tasks to 3.

Inputs

- **Hospital Dataset (`hospital.csv`)**: Contains patient details, including columns for **Diabetes (DM)**, **Hypertension (HTN)**, **Smoking**, **Alcohol**, and **Outcome**.

Executing Commands

- **Run the MapReduce Job**:

  ```
  hadoop jar ChronicLifestyleAnalysis.jar ChronicLifestyleAnalysis
  /user/hadoop/hospital/input
  /user/hadoop/hospital/output/chronic_lifestyle
  ```
- **View the Output**:

  ```
  a. hdfs dfs -cat
     /user/hadoop/hospital/output/chronic_lifestyle/part-0
  b. hdfs dfs -cat
     /user/hadoop/hospital/output/chronic_lifestyle/part-1
  ```

Outputs

```
hadoop@Ubuntu22:~$ hdfs dfs -cat /user/hadoop/hospital/output/chronic_lifestyle/part-r-0000
0
Chronic:DM_No_HTN_No:DAMA       371
Chronic:DM_No_HTN_No:DISCHARGE  5136
Chronic:DM_No_HTN_No:EXPIRY     539
Chronic:DM_No_HTN_No:OUTCOME    1
Chronic:DM_No_HTN_Yes:DAMA      208
Chronic:DM_No_HTN_Yes:DISCHARGE 4133
Chronic:DM_No_HTN_Yes:EXPIRY    273
Chronic:DM_Yes_HTN_No:DAMA      139
Chronic:DM_Yes_HTN_No:DISCHARGE 1770
Chronic:DM_Yes_HTN_No:EXPIRY    146
Chronic:DM_Yes_HTN_Yes:DAMA     178
Chronic:DM_Yes_HTN_Yes:DISCHARGE        2717
Chronic:DM_Yes_HTN_Yes:EXPIRY   147
```

```
hadoop@Ubuntu22:~$ hdfs dfs -cat /user/hadoop/hospital/output/chronic_lifestyle/part-r-0000
1
Lifestyle:Smoking_No_Alcohol_No:DAMA    793
Lifestyle:Smoking_No_Alcohol_No:DISCHARGE       12402
Lifestyle:Smoking_No_Alcohol_No:EXPIRY  1078
Lifestyle:Smoking_No_Alcohol_No:OUTCOME 1
Lifestyle:Smoking_No_Alcohol_Yes:DAMA   44
Lifestyle:Smoking_No_Alcohol_Yes:DISCHARGE      642
Lifestyle:Smoking_No_Alcohol_Yes:EXPIRY 5
Lifestyle:Smoking_Yes_Alcohol_No:DAMA   28
Lifestyle:Smoking_Yes_Alcohol_No:DISCHARGE      421
Lifestyle:Smoking_Yes_Alcohol_No:EXPIRY 14
Lifestyle:Smoking_Yes_Alcohol_Yes:DAMA  31
Lifestyle:Smoking_Yes_Alcohol_Yes:DISCHARGE     291
Lifestyle:Smoking_Yes_Alcohol_Yes:EXPIRY        8
```

Interpretation of Output

- **Chronic Conditions Analysis**:
    - **Diabetes (DM)** and **Hypertension (HTN)** have a noticeable impact on patient outcomes. Patients with **DM and HTN** tend to have higher occurrences of **DISCHARGE** and **EXPIRY**, especially in the **DM_Yes_HTN_Yes** group, indicating that these conditions may increase the risk of more severe outcomes.
    - In contrast, the **DM_No_HTN_No** group has the highest number of **DISCHARGE** outcomes, suggesting that individuals without chronic conditions may have more favorable outcomes.
- **Lifestyle Factors Analysis**:
    - Patients with **no smoking and no alcohol** (Lifestyle:Smoking_No_Alcohol_No) exhibit the highest **DISCHARGE** outcomes, emphasizing the beneficial effects of maintaining a healthier lifestyle.
    - Conversely, the **Smoking_Yes_Alcohol_Yes** group shows lower **DISCHARGE** counts, which may suggest that smoking and alcohol consumption adversely affect recovery or health outcomes.
- **Suggestions:**
    - **Chronic Condition Patients**: Hospitals should prioritize targeted care and interventions for patients with **diabetes (DM)** and **hypertension (HTN)**, as these conditions are linked to higher **EXPIRY** rates, possibly implementing specialized monitoring and treatment protocols to improve recovery outcomes.
    - **Lifestyle Factor Patients**: For patients with **smoking and alcohol consumption**, hospitals could offer lifestyle modification programs, including counseling and support for quitting smoking and reducing alcohol use, to improve overall health outcomes and reduce hospitalization time.

E. ICU Stay Duration Analysis

How does the type of admission (emergency vs. other) and the treatment type impact the average duration of ICU stays and patient outcomes (discharge, expiry)?

Method Used

- **MapReduce**: This approach was employed to analyze ICU stay duration across different factors such as **Admission Type**, **Treatment Type**, and **Outcome**. The mapper extracts the ICU stay duration and categorizes it based on the respective factors. The reducer then calculates the average stay duration for each category.

Code Link

The full code for the analysis can be accessed here:
https://github.com/AsmitaMondal/hospital-analysis/blob/main/codes/ICUStayDurationAnalysis.java

Code Explanation

1. **Mapper**:
    - Reads input data and extracts relevant columns: **Admission Type**, **Treatment Type**, **Outcome**, and **ICU Stay Duration**.
    - Emits a key-value pair for each category with the corresponding ICU stay duration.
    - The key is formed by combining the category (e.g., `AdmissionType`, `TreatmentType`, or `Outcome`) with the category value (e.g., `E` for emergency, `DAMA` for death after admission).
2. **Reducer**:
    - Aggregates the ICU stay durations for each category (admission type, treatment type, or outcome).
    - Computes the average ICU stay duration by dividing the total duration by the number of entries in that category.
    - Emits the category and its corresponding average ICU stay duration.

Inputs

- **Hospital Dataset (`hospital.csv`)**: Contains patient admission records, including details about the **Admission Type**, **Treatment Type**, **Outcome**, and **ICU Stay Duration**.

Executing Commands

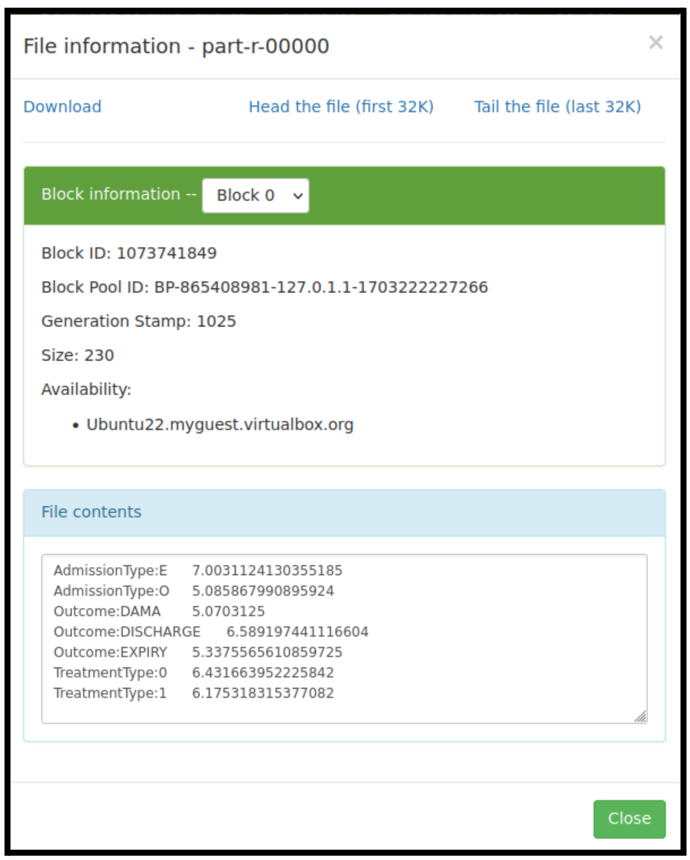
1. **Run the MapReduce Job**:

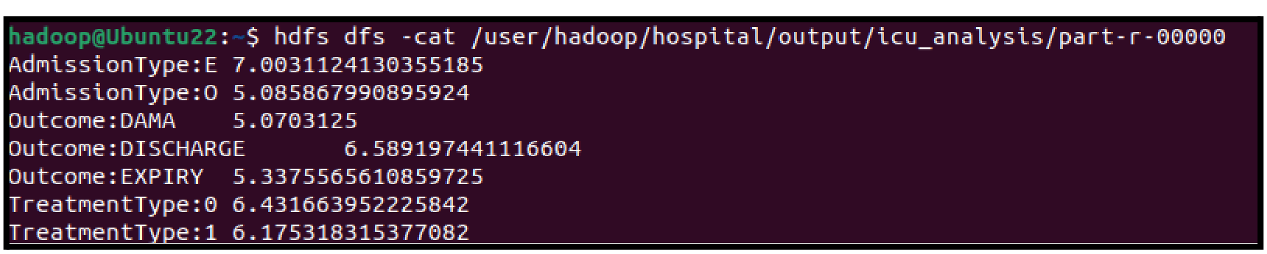
```
hadoop jar ICUStayDurationAnalysis.jar ICUStayDurationAnalysis
/user/hadoop/hospital/input
/user/hadoop/hospital/output/icu_analysis
```

2. **View the Output**:

```
hdfs dfs -cat
/user/hadoop/hospital/output/icu_analysis/part-r-00000
```

Outputs





Interpretation of Output

- **Admission Type**: Emergency (E) admissions have a higher average ICU stay duration (7.00 days) compared to non-emergency (O) admissions (5.09 days). This suggests that emergency patients may require more intensive care.
- **Outcome**: Patients who are discharged (6.59 days) stay longer in ICU compared to those who expire (5.34 days). This indicates that patients with more serious conditions or longer recovery times may stay longer in ICU.
- **Treatment Type**: Treatment type 0 has a slightly higher average ICU stay (6.43 days) than treatment type 1 (6.18 days). This could reflect differences in the severity or complexity of conditions treated with different types of therapies.

Suggestions and Insights

- **For Emergency Admissions**: Hospitals should ensure adequate ICU capacity to manage the longer stay of emergency patients and may need to streamline admission and discharge processes to optimize ICU resource utilization.
- **For Expiry Outcome**: The relatively short ICU stay for patients who expired may highlight the need for earlier intervention or palliative care strategies to address cases where survival prognosis is poor.

F. Outcome Remarks Integration Using Reducer Side Join

How can patient outcomes be linked with detailed remarks to analyze the comments a hospital has to give with respect to outcomes like "discharge," "expiry," or "DAMA" (Discharge Against Medical Advice)?

Method Used

- **Reducer-Side Join**: The approach integrates outcome remarks into the hospital records based on the **Outcome** column. The hospital data and remarks data are processed in separate mappers, and the reducer performs the join by matching the **Outcome** key from both datasets.

Code Link

The full code for this analysis can be accessed here:
https://github.com/AsmitaMondal/hospital-analysis/blob/main/codes/OutcomeRemarkJoin.java

Code Explanation

1. **Hospital Mapper**:
   - Reads the hospital data and emits the **Outcome** column as the key, while tagging the data as "HOSPITAL".
2. **Remarks Mapper**:
   - Reads the remarks data and emits the **Outcome** column as the key, with the value containing the remark text prefixed with "REMARK".
3. **Reducer**:
   - Processes the **Outcome** key, collecting remarks and counting the number of hospital records associated with each outcome.
   - Outputs a formatted result that includes the **Outcome**, the count of hospital records, and the corresponding remark.
4. **Driver**:
   - Configures the MapReduce job, using **MultipleInputs** to handle the two different input files (hospital data and remarks data).
   - Specifies the **OutcomeRemarkReducer** to join the data and writes the output to a specified path.

Inputs

- **Hospital Data (`hospital.csv`)**: Contains patient admission records, including an **Outcome** column.
- **Remarks Data (`remarks.csv`)**: Contains remarks associated with each **Outcome**.

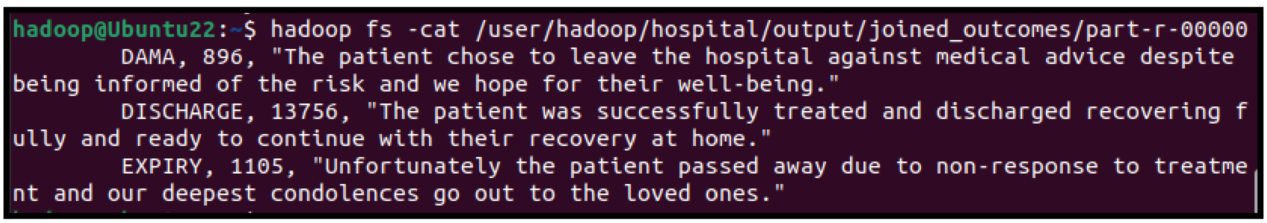
Executing Commands

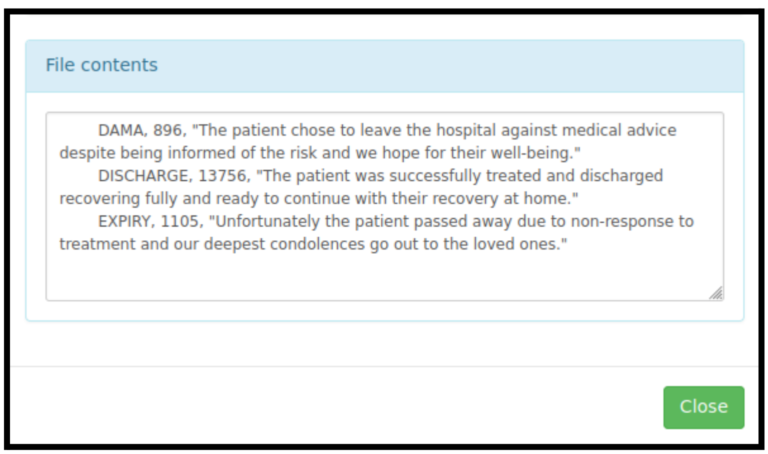
1. **Run the MapReduce Job**:

   ```
   hadoop jar OutcomeRemarkJoin.jar OutcomeRemarkJoin
   /user/hadoop/hospital/input /user/hadoop/remarks/input
   /user/hadoop/output/joined_outcome
   ```
2. **View the Output**:

   ```
   hdfs dfs -cat /user/hadoop/output/joined_outcome/part-r-00000
   ```

Outputs





Interpretation of Output

- **DAMA (Discharge Against Medical Advice)**: There were **896** patients who left the hospital against medical advice, and the remark reflects that they were informed of the risks and the hospital wishes them well.
- **DISCHARGE**: The remark indicates that **13,756** patients successfully completed their treatment and were discharged, fully recovered and ready to continue recovery at home.
- **EXPIRY**: **1,105** patients passed away due to non-response to treatment, and the remark expresses condolences to the families of the deceased.

Conclusion

This analysis successfully explores key aspects of hospital data using MapReduce, providing insights into several critical areas including **ICU Stay Duration**, **Outcome Remarks Integration**, and various patient outcomes. By employing Hadoop's MapReduce framework, we effectively processed large datasets, generating valuable insights that can drive improvements in hospital operations and patient care. The outcomes of this analysis provide actionable recommendations, such as improving discharge protocols and enhancing post-discharge care, while also identifying areas of concern.

Future Scope

1. **Real-Time Monitoring**: Integrating real-time patient data (e.g., vital signs, treatment progress) with the Hadoop ecosystem can provide hospitals with live insights into patient status, enabling proactive interventions.
2. **Integration with External Datasets**: Combining this hospital data with external datasets like insurance claims, weather data, or regional health trends could offer more comprehensive insights into patient outcomes and hospital performance.
3. **Improved Treatment Protocols**: By analyzing larger datasets and identifying patterns in patient outcomes, hospitals can develop more effective treatment guidelines and preventative measures for at-risk patients, ultimately leading to improved care and reduced mortality.

Limitations

1. **Simplified Assumptions**: The analysis relies on certain assumptions, such as using just a few columns for outcome predictions and remarks. Real-world healthcare data is often more complex and would require more granular features to develop more accurate models.
2. **Temporal Variability**: Patient outcomes can vary over time due to evolving medical treatments, policies, and hospital protocols. Incorporating temporal trends into the analysis could provide a more nuanced understanding of healthcare quality and outcomes.
3. **Generalizability**: The findings from this study are specific to the dataset provided and may not generalize to all hospitals. Differences in hospital infrastructure, patient demographics, and healthcare policies could lead to varying results in different settings.

Github Repository

https://github.com/AsmitaMondal/hospital-analysis