UNIVERSITI
MALAYA

WQD7007 Big Data Management

Group 2

Alternative Assessment

Case Study

Name	Low Boon Kiat
Matric Number	17138399
Submission Date	19th January 2024

Table of Contents

1.0 Introduction to the Big Data Resource

1.1 Justification Based on the 7V's Concept of Big Data

The online big data resource of my choice is TripAdvisor. TripAdvisor is an online travel platform that provides information and reviews on travel-related content. It allows users to plan and book various elements of their trips, including accommodations, flights, restaurants, and activities. The given resource is considered big data due to the following justification based on the 7V's concept of big data.

Table 1: Justification based on the 7V's concept of big data.

7V's	Justification
Volume	TripAdvisor collects and processes a massive volume of data. It handles millions of reviews, ratings, and other user-generated content related to hotels, restaurants, attractions, and travel experiences globally.
Velocity	Data on TripAdvisor is generated in real-time as users submit reviews, ratings, and other content. The platform experiences a continuous and high velocity of data flow.
Variety	TripAdvisor data includes a variety of information types, such as structured data like ratings, dates, and unstructured data like text reviews and images. This diversity in data types contributes to the variety aspect of big data.
Veracity	User-generated content may have variations in quality, reliability, and authenticity. TripAdvisor employs algorithms and moderation systems to ensure data veracity by validating and filtering reviews.
Value	The data on TripAdvisor provides valuable insights into customer preferences, opinions, and trends related to travel. This information is highly valuable for businesses and organizations in the tourism industry.
Variability	Data on TripAdvisor can vary based on locations, seasons, and trends in the tourism industry. The variability in data makes it adaptable to changing conditions and preferences.
Visualization	Visualization on TripAdvisor is essential for handling the complexity of the travel-related data it manages. Visualization enhances the user experience, provides valuable insights and facilitates informed decision-making.

1.2 Usefulness to the Tourism Industry

The identified big data resource on TripAdvisor can be useful to tourism industry due to the following justification. First contribution is customer insights. Tourism businesses can analyse TripAdvisor data to gain deep insights into customer preferences, expectations, and satisfaction levels. This information helps in tailoring services to meet customer needs. Second contribution is reputation management. Hotels and restaurants can monitor and manage their online reputation by analysing reviews and ratings on TripAdvisor. This is crucial for maintaining a positive image in the industry. Additionally, tourism industry players can conduct competitive analysis by comparing their performance with competitors on TripAdvisor. This aids in identifying strengths and weaknesses and making informed business decisions. Moreover, businesses can use TripAdvisor data to identify popular attractions, trending destinations, and customer expectations. This information is valuable for designing effective marketing campaigns. Lastly, destination management organizations can utilize TripAdvisor data to understand which attractions are gaining popularity, identify areas that need improvement, and plan infrastructure development based on visitor feedback.

1.3 Dataset Description

The dataset named TripAdvisor Hotel Reviews is obtained from Kaggle. It contains 878,561 reviews from 4333 hotels crawled from TripAdvisor. The dataset consists of 2 files, including ‘offerings’ and ‘reviews’. The ‘offerings’ has 9 attributes whereas the ‘reviews’ has 10 attributes.

Table 2: Attributes of the dataset.

No.	Attribute	Description
		‘offerings’
1	hotel_class	Hotel class (from 1 to 5)
2	region_id	Region ID for hotel location
3	url	URL of hotel webpage on TripAdvisor
4	phone	Contact number of hotel
5	details	Details of hotel
6	address	Address of hotel location
7	type	Type of hotel
8	id	Offering ID
9	name	Name of hotel
		‘reviews’
1	ratings	User ratings on service, cleanliness, overall, value, location, sleep quality, and rooms
2	title	Title of hotel review
3	text	Text of hotel review
4	author	Author who writes the hotel review
5	date_stayed	Date stayed in the hotel
6	offering_id	Offering ID
7	num_helpful_votes	Number of helpful votes from other users
8	date	Date when hotel review is written
9	id	User ID
10	via_mobile	Whether the review is submitted via mobile devices

The dataset contains both structured and unstructured data. In terms of structured data, the 'offerings' file has attributes like hotel_class, region_id, url, phone, details, address, type, id, and name. These attributes follow a clear and predefined structure, and the data in these columns can be organized in a tabular format. This type of data is considered structured. The 'reviews' file includes structured attributes such as ratings, title, author, date_stayed, offering_id, num_helpful_votes, date, id, and via_mobile. These attributes also adhere to a specific format, and the data can be organized in a structured manner.

In terms of unstructured data, the 'reviews' file has the 'text' attribute, which contains the actual text of the hotel review. Text data is inherently unstructured, as it lacks a predefined schema or format. The content of the 'text' attribute can vary greatly, and it requires natural language processing techniques for meaningful analysis. To sum up, while the majority of the attributes in both 'offerings' and 'reviews' files are structured, the 'text' attribute in the 'reviews' file introduces unstructured data.

2.0 Method to Store the Big Data Resource

Choosing the most suitable method to store big data from TripAdvisor involves considering the characteristics of the data and the requirements of the application, as well as advantages and disadvantages of each storage method.

Table 3: Comparison between relational database, MongoDB and cloud-based storage.

Method	Advantage	Disadvantage
Relational Database	• Well-suited for structured data: Relational databases excel at handling structured data with well-defined relationships, which might include information such as user profiles, reviews, and ratings. • ACID Compliance: Relational databases ensure data consistency and integrity through ACID properties (Atomicity, Consistency, Isolation, Durability), making them suitable for applications where transactional integrity is critical. • Mature Technology: Relational databases have been widely used and are a mature technology with a robust ecosystem, including various tools and support.	• Scalability Challenges: Relational databases can face challenges in scaling horizontally to handle massive amounts of data or high write loads, which are common characteristics of big data. • Schema Rigidity: While suitable for structured data, relational databases may face challenges with schema changes when dealing with evolving or semi-structured data.
MongoDB	• Flexible Schema: MongoDB's document-oriented structure allows for flexibility in schema design, accommodating the semi-structured and varied nature of data found on TripAdvisor. • Scalability: MongoDB is designed to scale horizontally, making it suitable for handling large volumes of data and high write loads, common in big data scenarios. • JSON-Like Documents: MongoDB stores data in BSON (Binary JSON) format. This is beneficial for handling diverse data types, such as text reviews, and images.	• Learning Curve: For those familiar with relational databases, adopting a NoSQL solution like MongoDB may have a learning curve. • Maturity: While MongoDB has gained widespread adoption, the NoSQL landscape is generally considered less mature than traditional relational databases.
Cloud-based Storage	• Scalability: Cloud-based storage solutions, such as Google Cloud Storage and Amazon S3 are designed for scalability. They can effortlessly handle the vast volume of data generated by TripAdvisor. • Durability and Reliability: Cloud storage solutions often provide high durability and reliability. Data is redundantly stored across multiple servers and data centres, reducing the risk of data loss.	• Access Speed: Retrieving data from cloud storage may introduce latency compared to accessing data from a local database. This can be a concern for applications requiring real-time data access. • Data Transfer Costs: While storing data in the cloud can be cost-effective, transferring large volumes of data in and

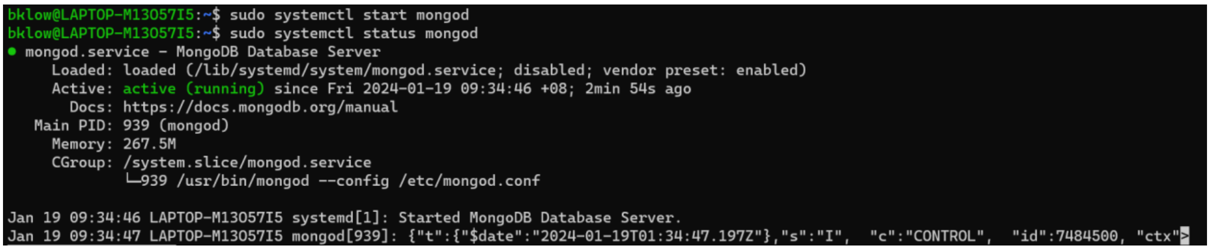
	- Easy Integration: Cloud storage seamlessly integrates with various big data processing and analytics tools available in the cloud ecosystem, facilitating a comprehensive data processing pipeline.	out of the cloud may incur additional costs. Organizations need to consider the implications of data transfer.

Considering the characteristics of TripAdvisor data, which includes structured and unstructured content, MongoDB stands out as a suitable choice. MongoDB is a NoSQL, document-oriented database. Its document-based model allows the storage of data in BSON (Binary JSON) format, a binary representation of JSON-like documents. This is well-suited for the diverse nature of the TripAdvisor data, which includes both structured (e.g., ratings, dates) and unstructured content (e.g., text reviews). Additionally, MongoDB provides flexibility in schema design, allowing for the storage of data without a rigid, predefined structure. This is advantageous for handling unstructured data like the 'text' attribute in the TripAdvisor reviews. Moreover, MongoDB is designed for horizontal scalability, making it suitable for handling large volumes of data and high write loads. TripAdvisor, with its millions of reviews and continuous user interactions, benefits from MongoDB's scalability to accommodate growth without sacrificing performance. To sum up, MongoDB’s flexibility in handling diverse data types and scalability make it well-suited for storing and managing the big data resource from TripAdvisor.

3.0 Demonstration of Data Storage and Data Access

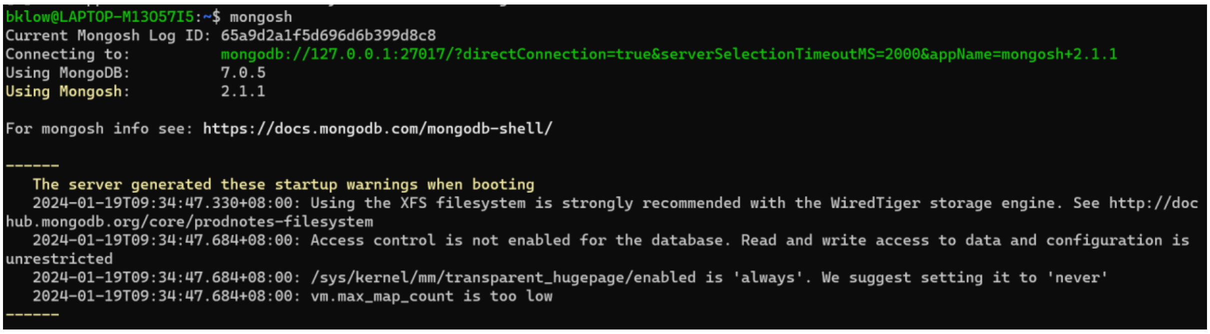
3.1 Data Storage

1. Start MongoDB Server. Ensure that MongoDB server is running.



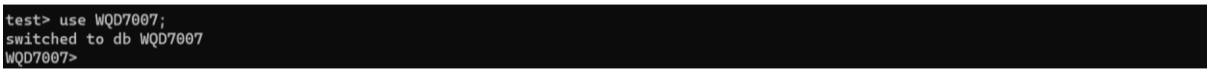
```
bklow@LAPTOP-M13057I5:~$ sudo systemctl start mongod
bklow@LAPTOP-M13057I5:~$ sudo systemctl status mongod
● mongod.service - MongoDB Database Server
     Loaded: loaded (/lib/systemd/system/mongod.service; disabled; vendor preset: enabled)
     Active: active (running) since Fri 2024-01-19 09:34:46 +08; 2min 54s ago
       Docs: https://docs.mongodb.org/manual
   Main PID: 939 (mongod)
     Memory: 267.5M
     CGroup: /system.slice/mongod.service
             └─939 /usr/bin/mongod --config /etc/mongod.conf

Jan 19 09:34:46 LAPTOP-M13057I5 systemd[1]: Started MongoDB Database Server.
Jan 19 09:34:47 LAPTOP-M13057I5 mongod[939]: {"t":{"$date":"2024-01-19T01:34:47.197Z"},"s":"I",  "c":"CONTROL",  "id":7484500, "ctx">
```

2. Use the ‘mongosh’ shell to interact with the MongoDB server.



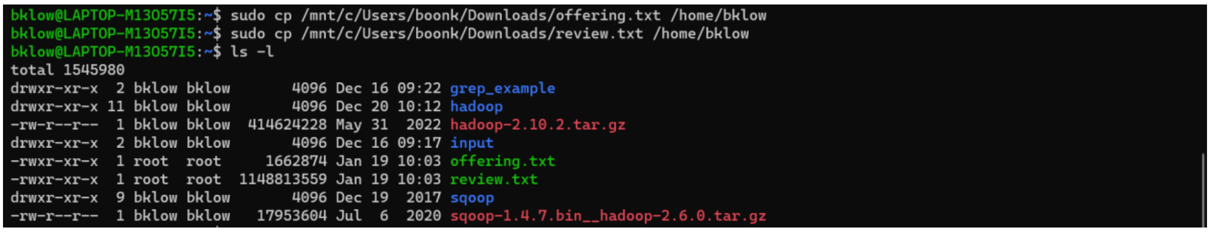
```
bklow@LAPTOP-M13057I5:~$ mongosh
Current Mongosh Log ID: 65a9d2a1f5d696d6b399d8c8
Connecting to:          mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000&appName=mongosh+2.1.1
Using MongoDB:          7.0.5
Using Mongosh:          2.1.1

For mongosh info see: https://docs.mongodb.com/mongodb-shell/

------
   The server generated these startup warnings when booting
   2024-01-19T09:34:47.330+08:00: Using the XFS filesystem is strongly recommended with the WiredTiger storage engine. See http://doc
hub.mongodb.org/core/prodnotes-filesystem
   2024-01-19T09:34:47.684+08:00: Access control is not enabled for the database. Read and write access to data and configuration is
unrestricted
   2024-01-19T09:34:47.684+08:00: /sys/kernel/mm/transparent_hugepage/enabled is 'always'. We suggest setting it to 'never'
   2024-01-19T09:34:47.684+08:00: vm.max_map_count is too low
------
```

3. In the MongoDB shell, create a database (e.g., WQD7007) to store data.



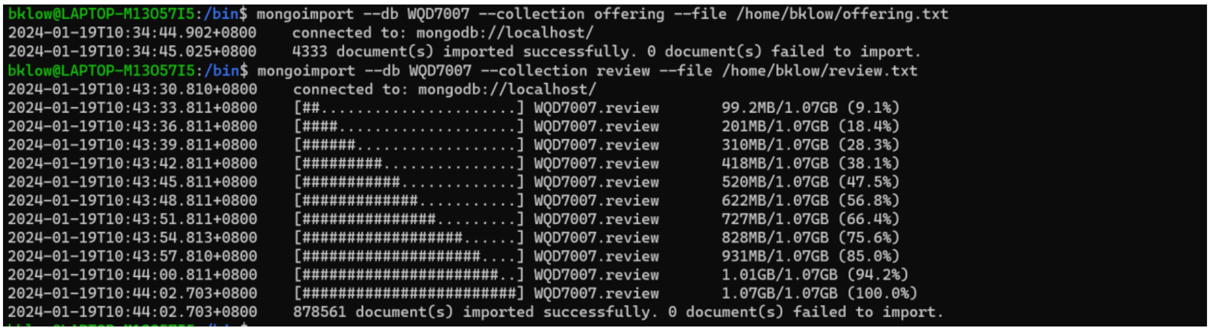
```
test> use WQD7007;
switched to db WQD7007
WQD7007>
```

4. Copy the two files that was downloaded from Kaggle into desired directory.



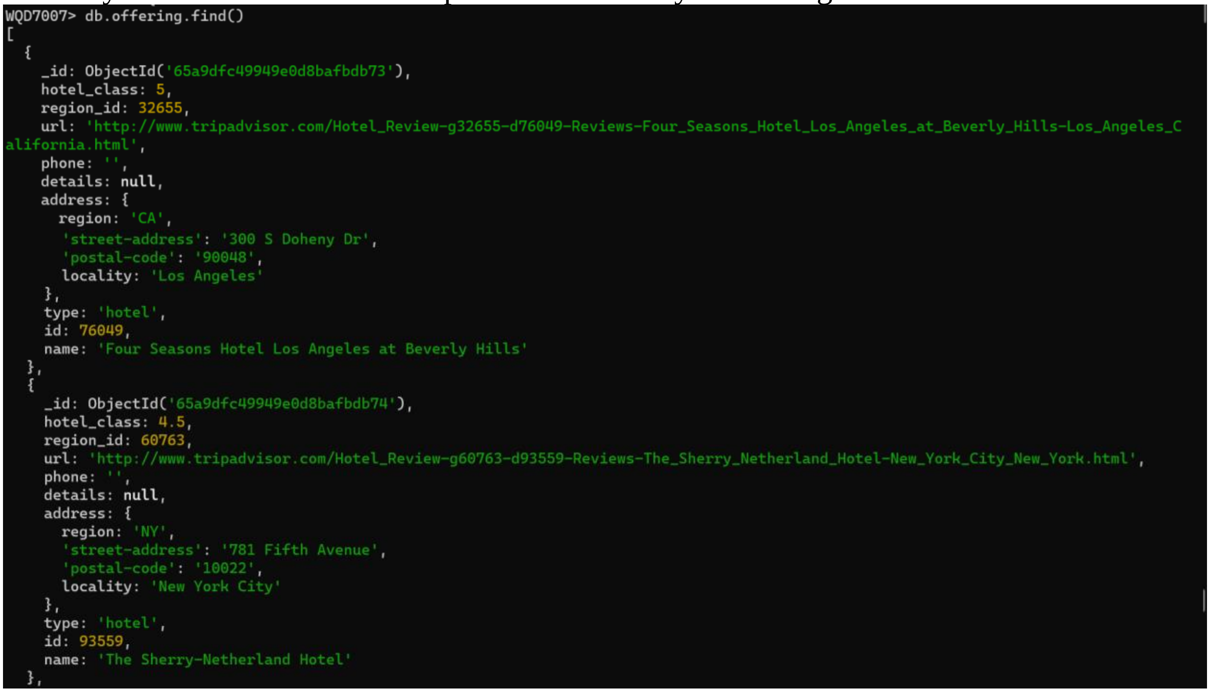
```
bklow@LAPTOP-M13057I5:~$ sudo cp /mnt/c/Users/boonk/Downloads/offering.txt /home/bklow
bklow@LAPTOP-M13057I5:~$ sudo cp /mnt/c/Users/boonk/Downloads/review.txt /home/bklow
bklow@LAPTOP-M13057I5:~$ ls -l
total 1545980
drwxr-xr-x  2 bklow bklow       4096 Dec 16 09:22 grep_example
drwxr-xr-x 11 bklow bklow       4096 Dec 20 10:12 hadoop
-rw-r--r--  1 bklow bklow  414624228 May 31  2022 hadoop-2.10.2.tar.gz
drwxr-xr-x  2 bklow bklow       4096 Dec 16 09:17 input
-rwxr-xr-x  1 root  root     1662874 Jan 19 10:03 offering.txt
-rwxr-xr-x  1 root  root  1148813559 Jan 19 10:03 review.txt
drwxr-xr-x  9 bklow bklow       4096 Dec 19  2017 sqoop
-rw-r--r--  1 bklow bklow   17953604 Jul  6  2020 sqoop-1.4.7.bin__hadoop-2.6.0.tar.gz
```

5. Use ‘mongoimport’ to import the JSON files into their respective collections.

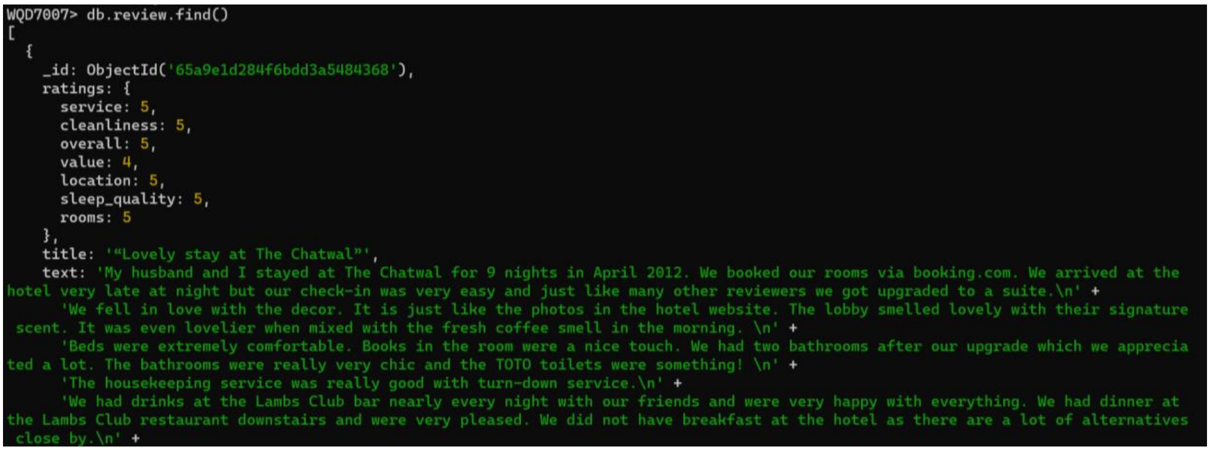


```
bklow@LAPTOP-M13057I5:/bin$ mongoimport --db WQD7007 --collection offering --file /home/bklow/offering.txt
2024-01-19T10:34:44.902+0800    connected to: mongodb://localhost/
2024-01-19T10:34:45.025+0800    4333 document(s) imported successfully. 0 document(s) failed to import.
bklow@LAPTOP-M13057I5:/bin$ mongoimport --db WQD7007 --collection review --file /home/bklow/review.txt
2024-01-19T10:43:30.810+0800    connected to: mongodb://localhost/
2024-01-19T10:43:33.811+0800    [##.....................] WQD7007.review       99.2MB/1.07GB (9.1%)
2024-01-19T10:43:36.811+0800    [####...................] WQD7007.review       201MB/1.07GB (18.4%)
2024-01-19T10:43:39.811+0800    [######.................] WQD7007.review       310MB/1.07GB (28.3%)
2024-01-19T10:43:42.811+0800    [#########..............] WQD7007.review       418MB/1.07GB (38.1%)
2024-01-19T10:43:45.811+0800    [###########............] WQD7007.review       520MB/1.07GB (47.5%)
2024-01-19T10:43:48.811+0800    [#############..........] WQD7007.review       622MB/1.07GB (56.8%)
2024-01-19T10:43:51.811+0800    [###############........] WQD7007.review       727MB/1.07GB (66.4%)
2024-01-19T10:43:54.813+0800    [##################.....] WQD7007.review       828MB/1.07GB (75.6%)
2024-01-19T10:43:57.810+0800    [####################....] WQD7007.review       931MB/1.07GB (85.0%)
2024-01-19T10:44:00.811+0800    [######################..] WQD7007.review       1.01GB/1.07GB (94.2%)
2024-01-19T10:44:02.703+0800    [########################] WQD7007.review       1.07GB/1.07GB (100.0%)
2024-01-19T10:44:02.703+0800    878561 document(s) imported successfully. 0 document(s) failed to import.
```

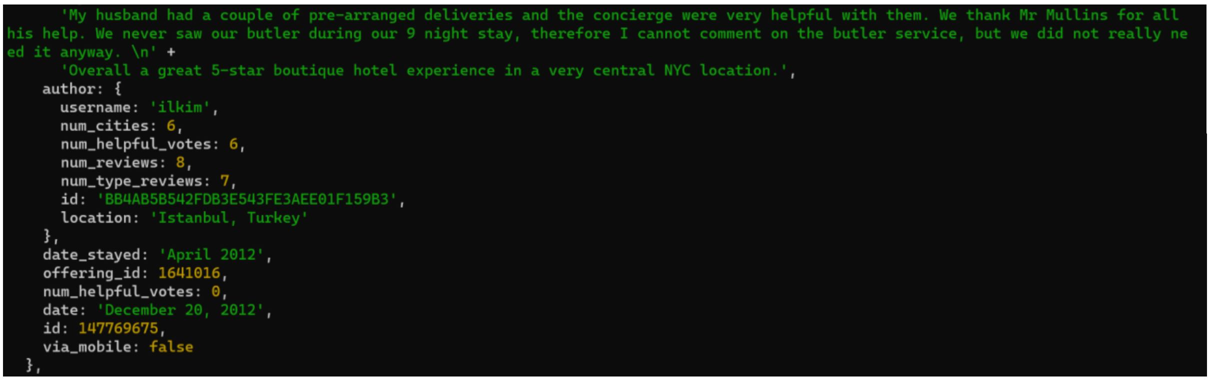
6. Verify that the data has been imported successfully into MongoDB.



```
WQD7007> db.offering.find()
[
  {
    _id: ObjectId('65a9dfc49949e0d8bafbdb73'),
    hotel_class: 5,
    region_id: 32655,
    url: 'http://www.tripadvisor.com/Hotel_Review-g32655-d76049-Reviews-Four_Seasons_Hotel_Los_Angeles_at_Beverly_Hills-Los_Angeles_C
alifornia.html',
    phone: '',
    details: null,
    address: {
      region: 'CA',
      'street-address': '300 S Doheny Dr',
      'postal-code': '90048',
      locality: 'Los Angeles'
    },
    type: 'hotel',
    id: 76049,
    name: 'Four Seasons Hotel Los Angeles at Beverly Hills'
  },
  {
    _id: ObjectId('65a9dfc49949e0d8bafbdb74'),
    hotel_class: 4.5,
    region_id: 60763,
    url: 'http://www.tripadvisor.com/Hotel_Review-g60763-d93559-Reviews-The_Sherry_Netherland_Hotel-New_York_City_New_York.html',
    phone: '',
    details: null,
    address: {
      region: 'NY',
      'street-address': '781 Fifth Avenue',
      'postal-code': '10022',
      locality: 'New York City'
    },
    type: 'hotel',
    id: 93559,
    name: 'The Sherry-Netherland Hotel'
  },
```

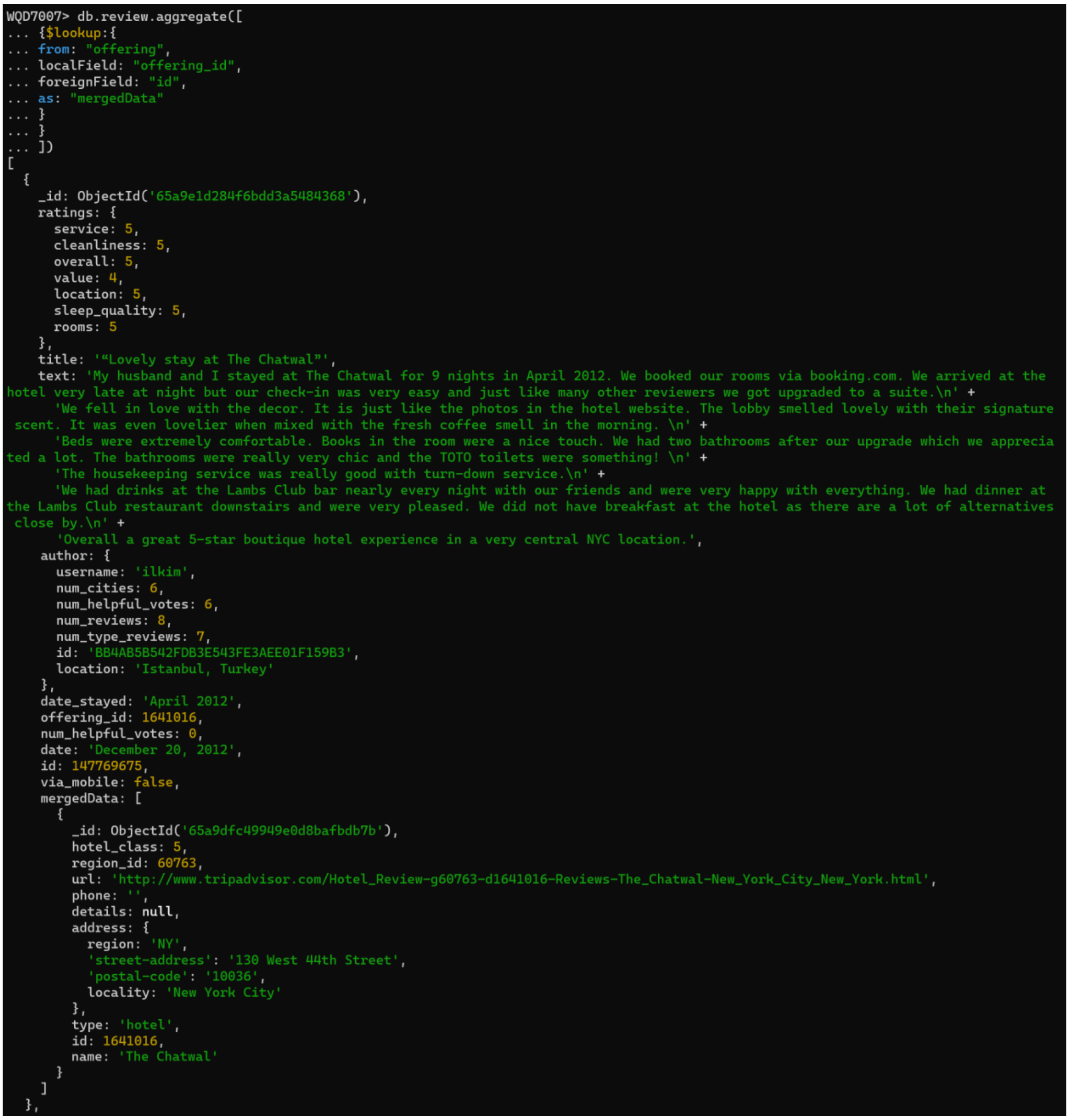


```
WQD7007> db.review.find()
[
  {
    _id: ObjectId('65a9e1d284f6bdd3a5484368'),
    ratings: {
      service: 5,
      cleanliness: 5,
      overall: 5,
      value: 4,
      location: 5,
      sleep_quality: 5,
      rooms: 5
    },
    title: '“Lovely stay at The Chatwal”',
    text: 'My husband and I stayed at The Chatwal for 9 nights in April 2012. We booked our rooms via booking.com. We arrived at the
hotel very late at night but our check-in was very easy and just like many other reviewers we got upgraded to a suite.\n' +
      'We fell in love with the decor. It is just like the photos in the hotel website. The lobby smelled lovely with their signature
 scent. It was even lovelier when mixed with the fresh coffee smell in the morning. \n' +
      'Beds were extremely comfortable. Books in the room were a nice touch. We had two bathrooms after our upgrade which we apprecia
ted a lot. The bathrooms were really very chic and the TOTO toilets were something! \n' +
      'The housekeeping service was really good with turn-down service.\n' +
      'We had drinks at the Lambs Club bar nearly every night with our friends and were very happy with everything. We had dinner at
the Lambs Club restaurant downstairs and were very pleased. We did not have breakfast at the hotel as there are a lot of alternatives
 close by.\n' +
```



```
      'My husband had a couple of pre-arranged deliveries and the concierge were very helpful with them. We thank Mr Mullins for all
his help. We never saw our butler during our 9 night stay, therefore I cannot comment on the butler service, but we did not really ne
ed it anyway. \n' +
      'Overall a great 5-star boutique hotel experience in a very central NYC location.',
    author: {
      username: 'ilkim',
      num_cities: 6,
      num_helpful_votes: 6,
      num_reviews: 8,
      num_type_reviews: 7,
      id: 'BB4AB5B542FDB3E543FE3AEE01F159B3',
      location: 'Istanbul, Turkey'
    },
    date_stayed: 'April 2012',
    offering_id: 1641016,
    num_helpful_votes: 0,
    date: 'December 20, 2012',
    id: 147769675,
    via_mobile: false
  },
```

7. Use the $lookup stage of the aggregation pipeline to perform a left outer join between the ‘review’ collection and the ‘offering’ collection based on the ‘offering_id’ field.

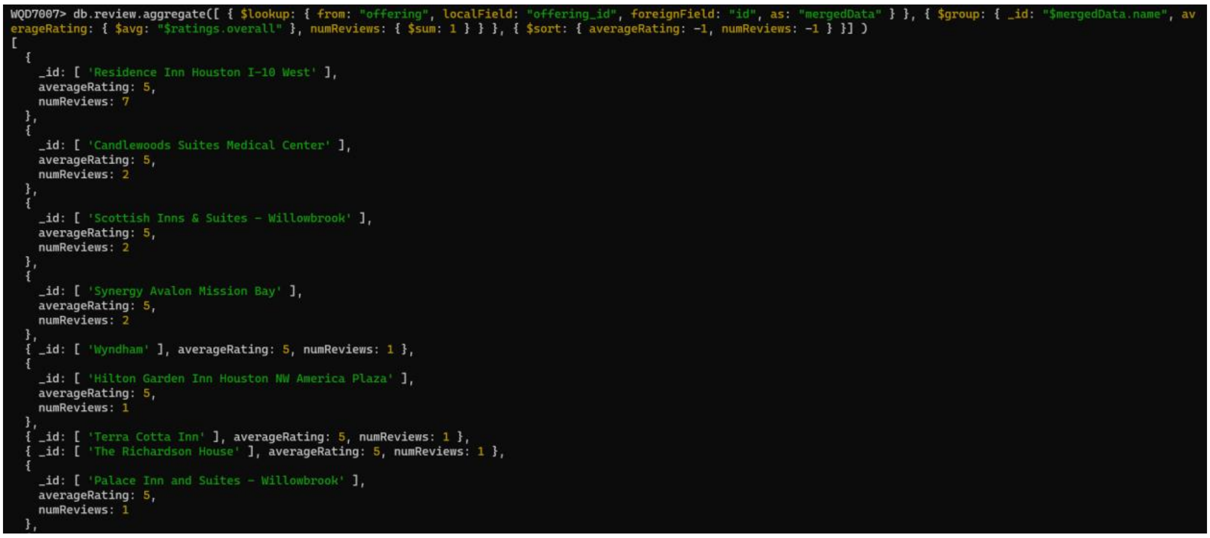


```
WQD7007> db.review.aggregate([
... {$lookup:{
... from: "offering",
... localField: "offering_id",
... foreignField: "id",
... as: "mergedData"
... }
... }
... ])
[
  {
    _id: ObjectId('65a9e1d284f6bdd3a5484368'),
    ratings: {
      service: 5,
      cleanliness: 5,
      overall: 5,
      value: 4,
      location: 5,
      sleep_quality: 5,
      rooms: 5
    },
    title: '“Lovely stay at The Chatwal”',
    text: 'My husband and I stayed at The Chatwal for 9 nights in April 2012. We booked our rooms via booking.com. We arrived at the
hotel very late at night but our check-in was very easy and just like many other reviewers we got upgraded to a suite.\n' +
      'We fell in love with the decor. It is just like the photos in the hotel website. The lobby smelled lovely with their signature
 scent. It was even lovelier when mixed with the fresh coffee smell in the morning. \n' +
      'Beds were extremely comfortable. Books in the room were a nice touch. We had two bathrooms after our upgrade which we apprecia
ted a lot. The bathrooms were really very chic and the TOTO toilets were something! \n' +
      'The housekeeping service was really good with turn-down service.\n' +
      'We had drinks at the Lambs Club bar nearly every night with our friends and were very happy with everything. We had dinner at
the Lambs Club restaurant downstairs and were very pleased. We did not have breakfast at the hotel as there are a lot of alternatives
 close by.\n' +
      'Overall a great 5-star boutique hotel experience in a very central NYC location.',
    author: {
      username: 'ilkim',
      num_cities: 6,
      num_helpful_votes: 6,
      num_reviews: 8,
      num_type_reviews: 7,
      id: 'BB4AB5B542FDB3E543FE3AEE01F159B3',
      location: 'Istanbul, Turkey'
    },
    date_stayed: 'April 2012',
    offering_id: 1641016,
    num_helpful_votes: 0,
    date: 'December 20, 2012',
    id: 147769675,
    via_mobile: false,
    mergedData: [
      {
        _id: ObjectId('65a9dfc49949e0d8bafbdb7b'),
        hotel_class: 5,
        region_id: 60763,
        url: 'http://www.tripadvisor.com/Hotel_Review-g60763-d1641016-Reviews-The_Chatwal-New_York_City_New_York.html',
        phone: '',
        details: null,
        address: {
          region: 'NY',
          'street-address': '130 West 44th Street',
          'postal-code': '10036',
          locality: 'New York City'
        },
        type: 'hotel',
        id: 1641016,
        name: 'The Chatwal'
      }
    ]
  },
```

3.2 Data Access

i. Identifying Top-Rated Hotels

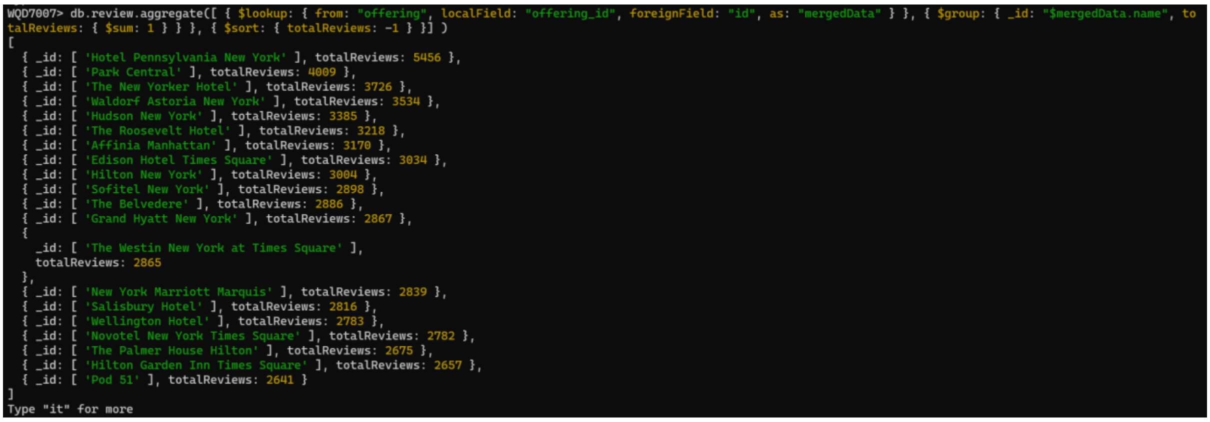
Group the data by the hotel name, calculate the average overall rating and the number of reviews, and finally sort the result by the average overall rating in descending order. This query provides insights into hotel performance by identifying top-rated hotels. Top-rated hotels can leverage their positive ratings in marketing and promotional activities. Highlighting positive customer feedback can attract more customers and enhance the hotel's reputation. Based on query results, Residence Inn Houston I-10 West has the highest average overall rating with highest number of reviews.



```
WQD7007> db.review.aggregate([ { $lookup: { from: "offering", localField: "offering_id", foreignField: "id", as: "mergedData" } }, { $group: { _id: "$mergedData.name", av
erageRating: { $avg: "$ratings.overall" }, numReviews: { $sum: 1 } } }, { $sort: { averageRating: -1, numReviews: -1 } }] )
[
  {
    _id: [ 'Residence Inn Houston I-10 West' ],
    averageRating: 5,
    numReviews: 7
  },
  {
    _id: [ 'Candlewoods Suites Medical Center' ],
    averageRating: 5,
    numReviews: 2
  },
  {
    _id: [ 'Scottish Inns & Suites - Willowbrook' ],
    averageRating: 5,
    numReviews: 2
  },
  {
    _id: [ 'Synergy Avalon Mission Bay' ],
    averageRating: 5,
    numReviews: 2
  },
  { _id: [ 'Wyndham' ], averageRating: 5, numReviews: 1 },
  {
    _id: [ 'Hilton Garden Inn Houston NW America Plaza' ],
    averageRating: 5,
    numReviews: 1
  },
  { _id: [ 'Terra Cotta Inn' ], averageRating: 5, numReviews: 1 },
  { _id: [ 'The Richardson House' ], averageRating: 5, numReviews: 1 },
  {
    _id: [ 'Palace Inn and Suites - Willowbrook' ],
    averageRating: 5,
    numReviews: 1
  },
```

ii. Identifying Popular Hotels

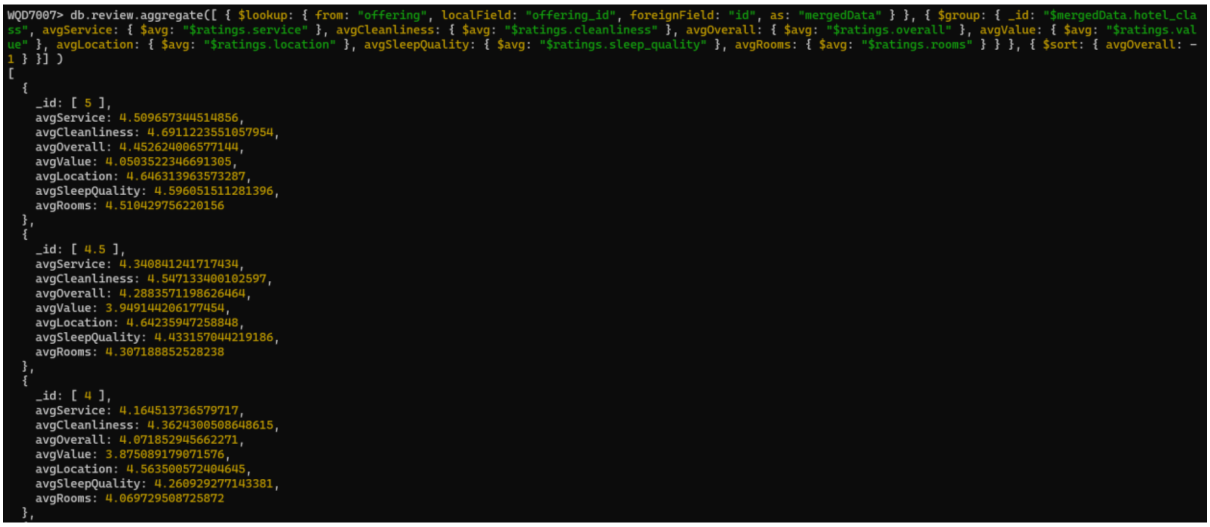
Group the reviews by hotel name and calculate the total number of reviews for each hotel. Sort in descending order based on the total number of reviews. This query can help identify hotels with the highest number of reviews, providing insights into their popularity among users. Understanding which hotels attract more reviews helps businesses allocate resources efficiently. Popular hotels may require additional staff, attention to customer services and specific promotional strategies. Based on query results, Hotel Pennsylvania New York is the most popular hotel with highest number of reviews.



```
WQD7007> db.review.aggregate([ { $lookup: { from: "offering", localField: "offering_id", foreignField: "id", as: "mergedData" } }, { $group: { _id: "$mergedData.name", to
talReviews: { $sum: 1 } } }, { $sort: { totalReviews: -1 } }] )
[
  { _id: [ 'Hotel Pennsylvania New York' ], totalReviews: 5456 },
  { _id: [ 'Park Central' ], totalReviews: 4009 },
  { _id: [ 'The New Yorker Hotel' ], totalReviews: 3726 },
  { _id: [ 'Waldorf Astoria New York' ], totalReviews: 3534 },
  { _id: [ 'Hudson New York' ], totalReviews: 3385 },
  { _id: [ 'The Roosevelt Hotel' ], totalReviews: 3218 },
  { _id: [ 'Affinia Manhattan' ], totalReviews: 3170 },
  { _id: [ 'Edison Hotel Times Square' ], totalReviews: 3034 },
  { _id: [ 'Hilton New York' ], totalReviews: 3004 },
  { _id: [ 'Sofitel New York' ], totalReviews: 2898 },
  { _id: [ 'The Belvedere' ], totalReviews: 2886 },
  { _id: [ 'Grand Hyatt New York' ], totalReviews: 2867 },
  {
    _id: [ 'The Westin New York at Times Square' ],
    totalReviews: 2865
  },
  { _id: [ 'New York Marriott Marquis' ], totalReviews: 2839 },
  { _id: [ 'Salisbury Hotel' ], totalReviews: 2816 },
  { _id: [ 'Wellington Hotel' ], totalReviews: 2783 },
  { _id: [ 'Novotel New York Times Square' ], totalReviews: 2782 },
  { _id: [ 'The Palmer House Hilton' ], totalReviews: 2675 },
  { _id: [ 'Hilton Garden Inn Times Square' ], totalReviews: 2657 },
  { _id: [ 'Pod 51' ], totalReviews: 2641 }
]
Type "it" for more
```

iii. Understanding Customer Preferences Across Hotel Classes

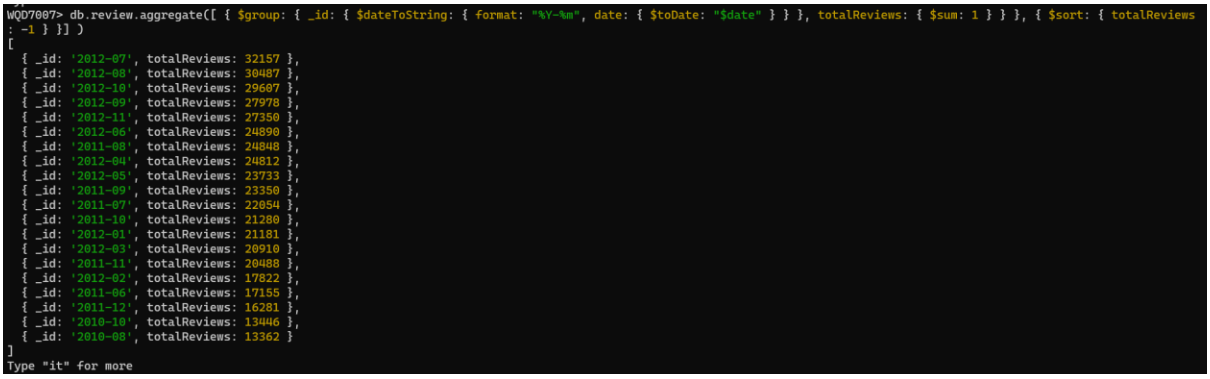
Group the merged data by ‘hotel_class’ and calculate the average ratings for various aspects (service, cleanliness, overall, value, location, sleep quality, and rooms). This query provides insights related to average ratings for various aspects based on the 'hotel_class'. This helps in understanding customer preferences by analysing which aspects are highly valued by customers in different hotel classes. For example, the query results reveal that customers in hotel class 5 prioritize cleanliness, while those in class 4 prioritize location. Hotel managers can use the insights to make strategic decisions related to service improvements or marketing strategies tailored to the preferences of customers in each hotel class.



```
WQD7007> db.review.aggregate([ { $lookup: { from: "offering", localField: "offering_id", foreignField: "id", as: "mergedData" } }, { $group: { _id: "$mergedData.hotel_cla
ss", avgService: { $avg: "$ratings.service" }, avgCleanliness: { $avg: "$ratings.cleanliness" }, avgOverall: { $avg: "$ratings.overall" }, avgValue: { $avg: "$ratings.val
ue" }, avgLocation: { $avg: "$ratings.location" }, avgSleepQuality: { $avg: "$ratings.sleep_quality" }, avgRooms: { $avg: "$ratings.rooms" } } }, { $sort: { avgOverall: -
1 } }] )
[
  {
    _id: [ 5 ],
    avgService: 4.509657344514856,
    avgCleanliness: 4.6911223551057954,
    avgOverall: 4.452624006577144,
    avgValue: 4.0503522346691305,
    avgLocation: 4.646313963573287,
    avgSleepQuality: 4.596051511281396,
    avgRooms: 4.510429756220156
  },
  {
    _id: [ 4.5 ],
    avgService: 4.340841241717434,
    avgCleanliness: 4.547133400102597,
    avgOverall: 4.2883571198626464,
    avgValue: 3.949144206177454,
    avgLocation: 4.64235947258848,
    avgSleepQuality: 4.433157044219186,
    avgRooms: 4.307188852528238
  },
  {
    _id: [ 4 ],
    avgService: 4.164513736579717,
    avgCleanliness: 4.3624300508648615,
    avgOverall: 4.071852945662271,
    avgValue: 3.875089179071576,
    avgLocation: 4.563500572404645,
    avgSleepQuality: 4.260929277143381,
    avgRooms: 4.069729508725872
  },
```

iv. Analysing Temporal Distribution of Reviews

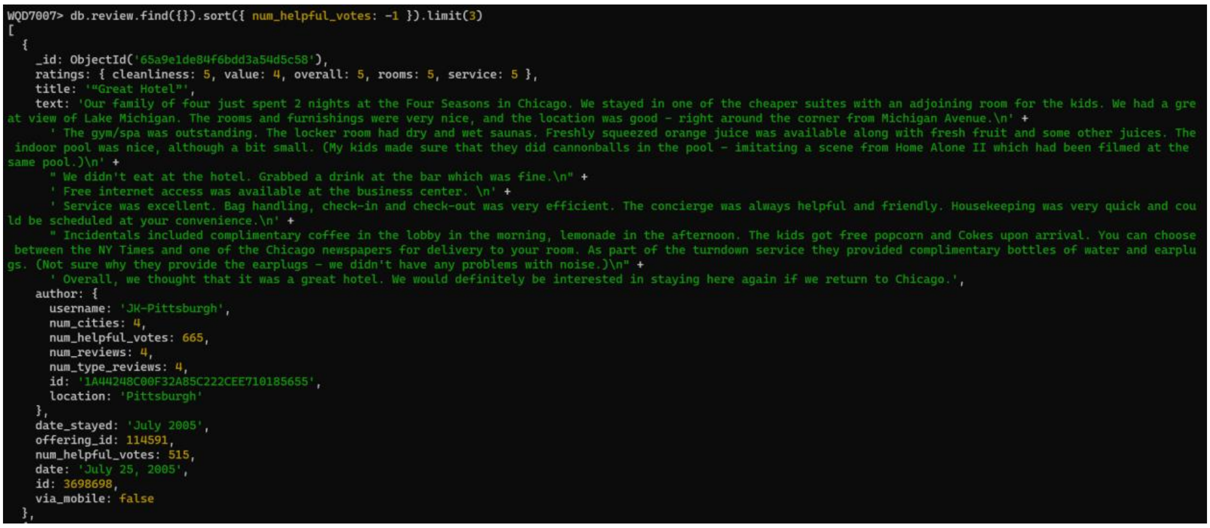
Group the reviews by the year and month of the ‘date’ field and calculate the total number of reviews for each period. Then, sort in descending order based on the total number of reviews. This query helps to provide insights into the distribution of reviews over time, which reveals patterns and seasonality in customer feedback. This information is crucial for businesses to understand when they might expect increased or decreased customer engagement, aiding in resource allocation. Based on query results, hotels receive the highest number of customer engagement in July and August.



```
WQD7007> db.review.aggregate([ { $group: { _id: { $dateToString: { format: "%Y-%m", date: { $toDate: "$date" } } }, totalReviews: { $sum: 1 } } }, { $sort: { totalReviews
: -1 } }] )
[
  { _id: '2012-07', totalReviews: 32157 },
  { _id: '2012-08', totalReviews: 30487 },
  { _id: '2012-10', totalReviews: 29607 },
  { _id: '2012-09', totalReviews: 27978 },
  { _id: '2012-11', totalReviews: 27350 },
  { _id: '2012-06', totalReviews: 24890 },
  { _id: '2011-08', totalReviews: 24848 },
  { _id: '2012-04', totalReviews: 24812 },
  { _id: '2012-05', totalReviews: 23733 },
  { _id: '2011-09', totalReviews: 23350 },
  { _id: '2011-07', totalReviews: 22054 },
  { _id: '2011-10', totalReviews: 21280 },
  { _id: '2012-01', totalReviews: 21181 },
  { _id: '2012-03', totalReviews: 20910 },
  { _id: '2011-11', totalReviews: 20488 },
  { _id: '2012-02', totalReviews: 17822 },
  { _id: '2011-06', totalReviews: 17155 },
  { _id: '2011-12', totalReviews: 16281 },
  { _id: '2010-10', totalReviews: 13446 },
  { _id: '2010-08', totalReviews: 13362 }
]
Type "it" for more
```

v. Understanding User Engagement through Helpful Reviews

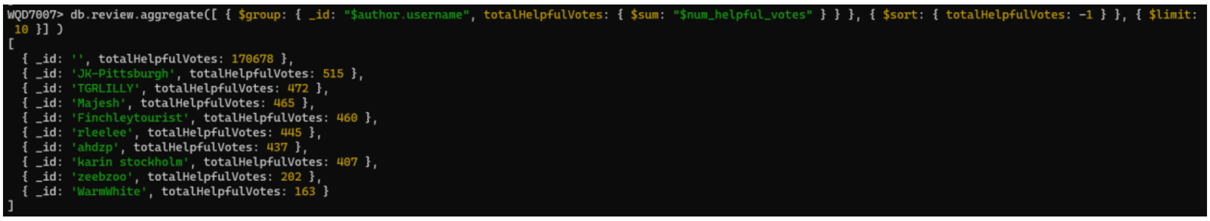
Sort the documents in the ‘review’ collection based on the ‘num_helpful_votes’ field in descending order and limit the result to 3 documents with highest number of helpful votes. This query provides insights into reviews by identifying the most helpful reviews. Understanding which reviews are considered most helpful can help in identifying content that engages users effectively. Platforms can use the insights to improve guidelines for users on creating impactful and informative reviews. Based on query results, review entitled “Great Hotel” has the highest number of helpful votes (515 votes).



```
WQD7007> db.review.find({}).sort({ num_helpful_votes: -1 }).limit(3)
[
  {
    _id: ObjectId('65a9e1de84f6bdd3a54d5c58'),
    ratings: { cleanliness: 5, value: 4, overall: 5, rooms: 5, service: 5 },
    title: '“Great Hotel”',
    text: 'Our family of four just spent 2 nights at the Four Seasons in Chicago. We stayed in one of the cheaper suites with an adjoining room for the kids. We had a gre
at view of Lake Michigan. The rooms and furnishings were very nice, and the location was good - right around the corner from Michigan Avenue.\n' +
      ' The gym/spa was outstanding. The locker room had dry and wet saunas. Freshly squeezed orange juice was available along with fresh fruit and some other juices. The
 indoor pool was nice, although a bit small. (My kids made sure that they did cannonballs in the pool - imitating a scene from Home Alone II which had been filmed at the
same pool.)\n' +
      " We didn't eat at the hotel. Grabbed a drink at the bar which was fine.\n" +
      ' Free internet access was available at the business center. \n' +
      ' Service was excellent. Bag handling, check-in and check-out was very efficient. The concierge was always helpful and friendly. Housekeeping was very quick and cou
ld be scheduled at your convenience.\n' +
      " Incidentals included complimentary coffee in the lobby in the morning, lemonade in the afternoon. The kids got free popcorn and Cokes upon arrival. You can choose
 between the NY Times and one of the Chicago newspapers for delivery to your room. As part of the turndown service they provided complimentary bottles of water and earplu
gs. (Not sure why they provide the earplugs - we didn't have any problems with noise.)\n" +
      ' Overall, we thought that it was a great hotel. We would definitely be interested in staying here again if we return to Chicago.',
    author: {
      username: 'JK-Pittsburgh',
      num_cities: 4,
      num_helpful_votes: 665,
      num_reviews: 4,
      num_type_reviews: 4,
      id: '1A44248C00F32A85C222CEE710185655',
      location: 'Pittsburgh'
    },
    date_stayed: 'July 2005',
    offering_id: 114591,
    num_helpful_votes: 515,
    date: 'July 25, 2005',
    id: 3698698,
    via_mobile: false
  },
```

vi. Recognizing Influential Authors

Group reviews by the username of the author and calculate the total number of helpful votes received by each author. Sort in descending order based on the total number of helpful votes. This query provides insights into the authors who received the most helpful votes, indicating their influence and the quality of their reviews. High helpful vote counts indicates that an author consistently produces high-quality and informative content. Acknowledging and promoting such authors can enhance the overall quality of content on the platform. Based on query result, JK-Pittsburgh is the most popular author on the platform.



```
WQD7007> db.review.aggregate([ { $group: { _id: "$author.username", totalHelpfulVotes: { $sum: "$num_helpful_votes" } } }, { $sort: { totalHelpfulVotes: -1 } }, { $limit:
 10 }] )
[
  { _id: '', totalHelpfulVotes: 170678 },
  { _id: 'JK-Pittsburgh', totalHelpfulVotes: 515 },
  { _id: 'TGRLILLY', totalHelpfulVotes: 472 },
  { _id: 'Majesh', totalHelpfulVotes: 465 },
  { _id: 'Finchleytourist', totalHelpfulVotes: 460 },
  { _id: 'rleelee', totalHelpfulVotes: 445 },
  { _id: 'ahdzp', totalHelpfulVotes: 437 },
  { _id: 'karin stockholm', totalHelpfulVotes: 407 },
  { _id: 'zeebzoo', totalHelpfulVotes: 202 },
  { _id: 'WarmWhite', totalHelpfulVotes: 163 }
]
```

3.3 Meaningful Outcomes for Tourism Industry

The analysis of the TripAdvisor Hotel Reviews dataset has unveiled critical insights for the tourism industry. First, the identification of top-rated hotels allows businesses to leverage positive ratings in marketing efforts, fostering increased customer interest. Identifying popular hotels equips businesses with the knowledge needed for efficient resource allocation and tailored promotional strategies. Moreover, the analysis of average ratings across different hotel classes reveals distinct customer priorities, guiding hotels in tailoring services and marketing strategies accordingly. Temporal trends in customer engagement provide businesses with strategic insights for resource planning and targeted marketing efforts during peak seasons. Understanding the impact of reviews is instrumental for platforms to refine guidelines and elevate the quality of user-generated content. Lastly, recognizing influential authors contributes to enhancing the overall quality of content on platforms by acknowledging and promoting trusted reviewers. In conclusion, these extracted outcomes offer a holistic view of hotel performance, customer behaviour, and content creation within the tourism industry. Businesses can use these insights to refine their marketing strategies, enhance customer experiences, and optimize resource allocation, ultimately contributing to the growth and success of the tourism sector.

4.0 Big Data Pipeline

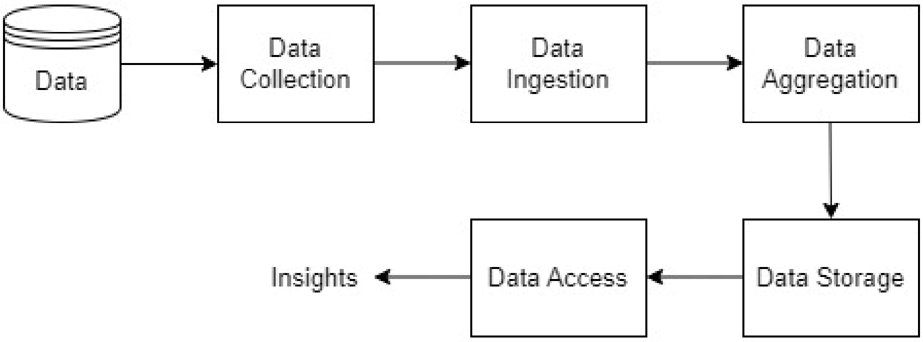


Figure 4.1 Big Data Pipeline for analysing TripAdvisor Hotel Reviews dataset.

The big data pipeline for the TripAdvisor Hotel Reviews dataset involves a systematic process, starting with data collection and culminating in insightful data access for the tourism industry. Initially, the dataset is collected from Kaggle. The subsequent step is data ingestion, where a MongoDB database named WQD7007 is created in the MongoDB shell to store the data. Using the 'mongoimport' command, the JSON files are imported into their respective collections – 'review' and 'offering' within MongoDB, ensuring the data is now available for analysis. In data aggregation, the $lookup stage of the aggregation pipeline in MongoDB is employed to perform a left outer join between the 'review' and 'offering' collections based on the 'offering_id' field, consolidating information across these collections. Subsequently, the data is stored in MongoDB, utilizing its document-based structure. This setup facilitates efficient and flexible data storage, accommodating the diverse attributes of both reviews and hotel offerings. The final stages of the pipeline involve data access, where various queries provide meaningful outcomes for the tourism industry. Collectively, this big data pipeline offers a comprehensive approach to extracting insights, making it a valuable asset for businesses and stakeholders in the tourism sector.