Exam Roll - CSE218056

Class Roll - 001710501029

BCSE - IV & 2nd Semester.

Subject :- Big Data Analytics.

1.

i) The characteristics of big data are as follows:-

1. Volume :- The amount of data handled by big data systems are very huge. Such large amounts cannot be handled by ~~conventional system~~ conventional systems.

2. Velocity :- Since the data involved in big ~~data~~ data is very huge, conventional systems are not able to provide results in reasonable time. Thus providing results on demand at a faster pace is a major aspect of Big Data.

3. Variety :- Big data is generated in multiple varieties. Compared to traditional data like texts, phonenumbers, addresses, the latest trend in data is ~~also~~ also in the form of images, video, audio etc. A high volume of the data is unstructured.

4. Veracity :- Veracity refers to the degree of reliability the data has to offer. Big Data algorithms are required to filter out huge amounts of data and translate ~~the~~ them ~~out as~~ so ~~that~~ that they are relavant and reliable to the particular requirement.

1.

ii) The different types of analytics are :-

1. Descriptive : It involves evaluating past descriptions of entities and their properties. It basically is the summarization of past data in an easily readable form.

2. Diagnostic :- It involves determining whether something has happened based on past data.

3. Predictive :- It involves forecasting events that have not happened yet.

~~4. Cognitive~~

4. Prescriptive :- It involves finding out the best course of action in a given situation.

5. Cognitive :- It involves combining analytics with AI and Machine Learing.

1.

iii) Data ingestion is a process by which data is moved from one or more sources to a destination where it can be stored and further analysed.

Data is collected from different sources in different formats. These non-homogenous data needs to be transformed in a way so that it is possible to analyse it together.

Since Big Data involved data in very high volume, with high variety and velocity, conventional systems fail. Thus Big Data ingestion tools like Apache Kafka, Wavefront, Amazon Kinesis et. are required.

1.

a). Map Reduce is a ~~programming~~ processing technique that can handle large-scale parallel computation in a way that is tolerant to hardware failures.

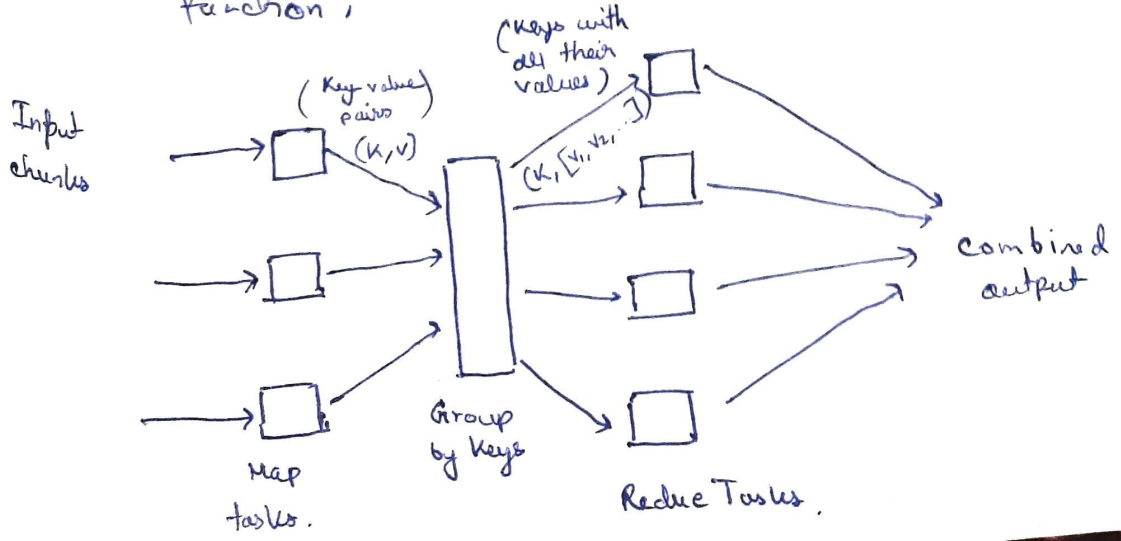
The user needs to write two functions Map & Reduce.

The execution occurs as follows:-

a) Some number of Map tasks are given one or more chunks from a distributed file system. The Map function converts the chunks in a sequence of key-value pairs.

b) The 'key-value' pairs from each Map task are collected by a master-controller and sorted by key. They keys are ~~sorted~~ divided among all the Reduce tasks, so all key-value pairs with the same key wind up at the same Reduce task.

c) The Reduce tasks work on one key at a time, and combine all the values associated with that key in a way specified by the Reduce function.



1.

iv) The broad features of Hadoop distributed file System are:-

→ Can manage very large distributed file system. It can scale upto thousands of nodes and handle large amounts of data.

→ It assumes commodity hardware. Even though a large number of nodes are handled, the nodes themselves are not expected to be very powerfull. Also nodes are expected to fail. Files are replicated to handle hardware failure. The architecture can detect failures and recover them.

→ It is optimized for batch processing. The data locations are exposed so that computations can move to where the data resides. This provides very high aggregate bandwidth.

→ A single namespace is used for the entire cluster.

→ Hadoop has a very robust ecosystem that is well suited to meet the analytical needs of developers and small to large scale organizations.

7. i) Attribute Value Frequency (AVF) Algorithm is a simple & fast algorithm to detect outliers in catagorical data.

It is based on the intuition that outliers are those points which are infrequent in the dataset, and that the ideal outlier point in a categorical dataset is one whose each & every value is extremely irregular or infrequent.

Given a dataset where each point X has m attributes

$X_i = [x_{i1}, x_{i2}, \ldots x_{im}]$ where x_{ij} is the jth attribute of x_i.

$$\therefore \text{AVFScore}(x_i) = \frac{1}{m} \sum_{j=0}^{m} f(x_{ij})$$

Outliers will have infrequent ~~categories~~ attributes, thus will have a lower AVF Score.

Once the AVF Score of all points are calculated we can designate the points with very low scores as outliers.

Using MapReduce architecture we can parallelize AVF and thus can operate on large datasets.

The MR-AVF is comprised of 2-stages.

Stage 1:-

Map:- The first map function associates attribute's each distinct value to the map's output key.

Reduce:- The reduce function computes the ~~attribute~~ freq. count of each attribute value.

~~Solved~~

Stage 2 :-

Map :- AVF score of each data point is calculated.

Reduce :- sorting operation the calculated AVF scores.

PSUDO CODE :-

Input : Database D (n points x m attributes)

Output : K- detected outliers.

```
HashTable H;
map K1 = i, V1 = Di = xi, i = 1.....n begin
    for each L ∈ xi; L = 1.....m do
        collect (xiL, 1);
    end
end.
reduce K2 = xiL, V2 begin
    H(xiL) += V2;
end
```

} stage 1

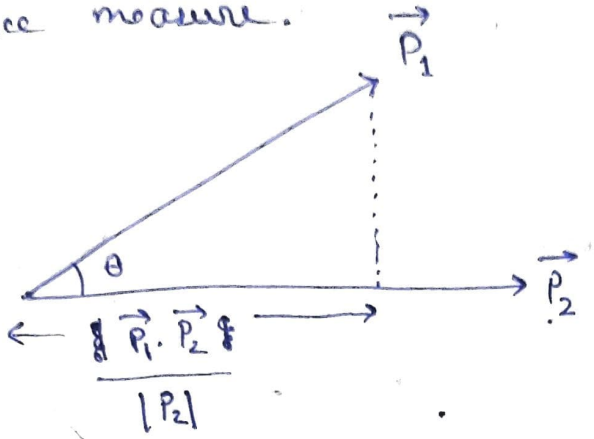
```
map K1 = i, V1 = Di = xi begin
    AVF = Σ_{L=1}^{m} H(xiL);
    collect (K1, AVF);
end
reduce K2 = AVF1, V1 = i;
```

} stage 2.

7. ii) Cosine distance :-

Assume points as vectors from the origin to its location.

Cosine distance simply refers to the angle between such two vectors. It is a non-euclidean distance measure.



$$d(P_1, P_2) = \theta = \arccos\left(\vec{P_1}.\vec{P_2} / |\vec{P_1}|.|\vec{P_2}|\right)$$

We can prove that here d is a distance measure :-

→ $d(x, x) = 0$ because $\arccos(1) = 0$

→ $d(x, y) \geq 0$ because any two intersecting vectors make an angle in the range 0° to 180°.

→ $d(x, y) = d(y, x)$ by symmetry.

→ Triangle inequality: If we rotate an angle x to z and then from z to y, we can't rotate less than from x to y.

∴ $d(x, z) + d(z, y) \geq d(x, y)$

∴ Cosine distance is a distance measure.

2. The main motive behing search autocomplete is to reduce typing time of the users, thus ~~proving~~ providing ~~results~~ relevant results faster.

In most search engines like Google & Yahoo, search engine autocomplete is based on past searches.

The search engine sorts ~~to~~ through trillions of past search queries ~~to~~ that have ~~matching~~ matching phrases • to provide suggestions.

One way of implementing this auto complete feature is through Fuzzy Search.

Fuzzy search is implemented ~~using~~ by ~~first~~ breaking the search query into ~~k-shingles~~ k-shingles.

After Shingling the search queries are converted into bit vectors and then using MinHashing their hashes are stored.

When a ~~user~~ user searches a new query, its MinHash is ~~compared~~ compared with other ~~pre~~ previous searches in the database.

An efficient method of comparing the ~~xxx~~ ~~hashes~~ hashes. is ~~&~~ using ~~Locatid~~ Locality Sensitive Hashing.
The MinHash signatures of the search ~~xxxx~~ queries are broken down into a number of bands. This bands are then again hashed ~~into~~ ~~multiple buckets~~. to a hash table with k buckets. Queries that hash to the same bucket are likely to be similar.

Once we have the list of all ~~similat~~ similar ~~past~~ ~~queries~~ queries another algorithm is used to ~~filter~~ ~~filters~~ filter those out in ~~order~~ order to show ~~the~~ only the most relevant queries to the user.
The queries may be sorted by relevance and the most relevant queries are shown to the user.

In some search engines like Google, if the user is logged in, the ~~engine~~ first few suggestion may the users own ~~xxxx~~ past search queries.

~~Queries are also sorted~~ Relevance of a query may be affected by various ~~xx~~ factors. Some of these factors include trending searches, location, language used to search etc.

In the screen shots we can see that the search suggestions of Yahoo & Google are quite different, although (although there are a few common suggestion in case of "ce").

This difference may arrise due to difference in filtering algorithms of both the search engines.

Also due to difference in user base, the search query history database of both the search engines is expected to be different.

If the user is logged in on the search engine (has an account made on the platform) user data like post searches & other preferences can also play a role in these suggestions.

3. Advertisers bid for query strings on search engines. The search engines shows adds on the search results that maximizes their ~~profits~~ revenue.

The expected revenue by displaying an add is calculated as :- Bid * CTR

where Bid is the amount that the advertiser bid on the search query, and CTR is Click Through Rate i.e. the probability that once an ad is shown a user clicks on the ad.

Based on this Bid * CTR value a search engine can choose which ads to show. Various algorithms like Greedy appraech or Balance algorithm can be used in order to select the ads.

~~In the first screenshot~~

If we look into the screenshots of the first query string "t-shirts for men" :-

→ Google displays adds from "The House" & "Bewakoof.com".

→ Yahoo displays adds from "Myntra" & "Shoppers Stop". Along with that Yahoo also displays websites on the right.

In the second query "pulse oximeter" :-

→ Google displays various ads for Oximeter from various companies like "Groqii technologies", "Dr Trust", "Flipkart" etc.

→ In case of yahoo all the ads on the top are from "Healthmg.com". And there are other related web sites that are shown on the right.

The difference in ads ~~may be due~~ in both search engines may be due to different bids from the compains. Like in the 2nd query, there is only one company in the Yahoo page results.

The difference may also arise from different CTR on the different search engines.

The search engines may also use different strategies for optimizing their revenue, thus the order of ads appearing can also be different.

4. Shingles used by Google :-

→ Central Government
→ discussions
→ Pfizer
→ American Pharmaceutical Company
→ provide its COVID-19 vaccine possibly
→ July, informed Dr. VK Paul, chair
→ National Expert Group
→ vaccine Administration (NEGVAC)

Shingles used by Yahoo :-

→ Pfizer
→ pharmaceutical
→ its COVID-19 vaccine.
→ vaccine
→ Nation Expert Group vaccine Administration (NEGVAC)

In Google, most of the articles on the first sequence of websites ~~almost~~ ~~about~~ have ~~some~~ almost complete match with the paragraph. As we gradually go down, the ~~matches~~ matches become ~~scarce~~ ~~scarce~~ ~~more and more~~ ~~rat~~ less often..

On the other side in Yahoo, even the first website doesn't have a complete match with the paragraph. (This may be because the ~~web~~ ~~webpage~~ web pages containing the news article are not indexed yet by Yahoo).

But here also, as we ~~down~~ go down, relevance of the sites decreases.

The webpages on the internet are indexed by the search engines. In order to measure relevance of content the content of the web sites are converted to K-shingles. These shingles are then Hashed and the hashes are stored.

When a search query is made the search string is broken into shingles and converted into a bit array. This bit array is hashed & compared with the hashes of indexed webpages.

Different search engines use different comparison algorithms. One such algorithm can be LSH (Locality Sensitive Hashing).

Various search engines use various algorithms in order to rank the webpages that have content similar to the search query. Some of the factors that contribute to the ranking may include, degree of match with the search string, data, location of the search, user preferences etc.

5. (A) → Authority (H) → Hubs

a) "Kalyani University" in Google

1. www.klyuniv.ac.in → (A)
2. collegedunia.com → (H)
3. en.wikipedia.com → (H)
4. west-bengal.result91.com → (H)
5. shiksha.com → (H)
6. educationdunia.com → (H)
7. examtc.com → (H)
8. career360.com → (H)
9. college admissions.in → (H)
10. dl.acm.org → (H)
11. www.indiatvnews.com → (A)
12. timesofindia.indiatimes.com → (A)
13. researchgate.net → (H)
14. pincode.net.in → (H)
15. iiitkalyani.ac.in → (A)

"Kalyani University" in Yahoo.

1. klyuniv.ac.in → (A)
2. collegedunia.com → (H)
3. career360.com → (H)
4. ~~klyuniv.ac.in/conf~~
4. www.buruniv.ac.in → (A)
5. www.knu.ac.in → (A)
6. in.seekweb.com → (H)
7. in.zapmeta search.com → (H)
8. www.americancollegespain.com → (H)
9. www.agoda.com → (H)
10. www.icbse.com → (H)
11. education i connect.com → (H)
12. www.sarvgyan.com → (H)
13. education dunia.com → (H)
14. www.freshersline.com → (H)
15. institutes.aglasem.com → (H)

Exam Roll- CSE218056, Md, Sahil, 00171050102 9 (A)

6). "Gurusaday Museum" in Google

1. en.wikipedia.org — (H)
2. www.gurusadaymuseum.org → (A)
3. scroll.in — (H)
4. www.tripadvisor.in — (H)
5. facebook.com — (H)
6. www.telegraphindia.com — (A)
7. www.museumsofindia.com → (H)
8. www.getbengal.com → (H)
9. www.ixigo.com → (H)
10. www.sahapedia.org → (H)
11. www.inspirock.com → (H)
12. www.anineus.in → (A)
13. www.kolkataonwheels.com → (H)
14. bbb.in → (H)
15. indiantribalheritage.org → (H)

"Gurusaday Museum" in Yahoo

1. www.museumsofindia.org → (H)
2. www.gurusadaymuseum.org. → (A)
3. indianetzone.com → (H)
4. bbb.in → (H)
5. www.tripadvisor.in → (H)
6. www.republicworld.com → (H)
7. www.livehistoryindia.com → (H)
8. transindiatravels.com → (H)
9. www.tripadvisor.com → (H)
10. www.betto.org. → (H)
11. www.museumsofindia.org → (H)
12. yaha.com → (H)
13. www.kolkataonwheels.org. → (H)
14. en.wikipedia.org → (H)
15. www.indianetzone.com → (H)

Good quality Hubs have higher hub scores & good quality of content..

Hubs are pages that link to authorities.

Good hubs are those that have appropriate content and links to authorities. Ex- wikipedia.

Sites that have content but very few links to authorities are bad hubs.

Sites that have links to authorities but improper or unverified content are bad hubs, Ex- social media sites.

The hubbiness 'h' and authoritativeness 'a' of a web page can be computed using HITS algorithm.

HITS algo:

Each page i has 2 scores

authority score : a_i

Hub score : h_i

HITS algorithm :-

→ Initialize : $a_j^{(0)} = 1/\sqrt{N}$, $h_j^{(0)} = 1/\sqrt{N}$ for $j = 1 \ldots N$;

→ Keep iterating till convergence:

$\forall i$: Authority : $a_i^{(t+1)} = \sum_{j \to i} h_j^{(t)}$

$\forall i$: Hub : $h_i^{(t+1)} = \sum_{i \to j} a_j^{(t)}$

$\forall i$: Normalize:

$$\sum_i \left(a_i^{(t+1)}\right)^2 = 1, \sum_j \left(h_j^{(t+1)}\right)^2 = 1$$

Here h_i → denotes the hubbiness of the webpage

6. Both Amazon and flipkart list a lot of books

6. In Amazon's search results almost all the books in the first page has the phrase "They all fall down" in the book title.

But in flipcart all the books in the first page has the phrase "all fall down" in the books.

Books titled "They all fall down" appear only on the 6th row of the search results.

Like most search engines, Amazon & Flipkart also perform searches based on key words.

Based on the above results one may conclude that the difference in search results of the two may be because flipkart is considering the word "they" as unimportant while searching for keywords.

Both Amazon & flipkart after filtering out books with the keywords given in the search query, sort the books based on relevance (by default). relevance can consist of various factors, like degree of match with the search query, ratings, number of people who bought it, availability on the location etc.

When we select a book on Amazon

There were mostly 3 types of recommendation :-

1. "Frequently bought together" (Amazon) :-

These recomendations contains items that are very closely related to each other. Like other books in the series. These recomendation can be generated using frequent items set mining algorithms. To find a frequent item set, we compare baskets which had selected books and the frequent items in those baskets can be used as a recomendation. This can be done by algorithms like A-priori Algorithm or PCY algorithm.

2. "Customers who bought this item also bought" (Amazon) :-

These recomendations are generated by the help of user-user collaborative filtering. Here we filter the user who also bought the selected books and based on their purchase history, display recomendations products that are frequently bought.

3. "Products related to the item" (Amazon)

or

"Similar products" (Flipkart) :-

These recomendations are generated based on key words loosely related to the current book/product.

It can be done using searching based on key words or even content based to labeling... labeling.

7.

iii) Future of big data :-

a) Business strategies :- With the expected increase in population, the customers for businesses are also increasing. Thus increasing data for businesses to handle. More and more businesses will understat the importance of big data analytics in proving a more tuned and streamlined products & services to the customers.

Big data tools will become more & more common & easily accessible even to need small startups.

b) Calamity Avoidance :- The COVID-19 pandemic is not the first pandemic that the world has faced and nor will it be the last. Big data and other related distributed application can play a very crucial role in avoiding preventing and recovering from there natural calamities.

We may see more and more predictive analysis taking place and if there are more pandemics in the future prescriptive analysis can help.

e

c) User Privacy & Security :- With more & more Bigdata applications collecting user data & preferences & preferences using big data analytics &, data breaches are becoming more & more common. Various organizations like the EU & county goverments & are stepping in and making laws on how user data is being handled.