USN | | | | | | | | | | |

Internal Assessment Test 3 scheme and solution – Jan 2022

Sub:	Big Data Analytics					Sub Code:	18CS72	Branch:	CSE
Date:	25/1/2022	Duration:	90 mins	Max Marks:	50	Sem / Sec:	VII/A,B,C	OBE	

		MARKS	CO	RBT
1	Explain Hubs and Authorities. Explain HITS algorithm. Hubs and Authorities -5 marks HITS -5 marks	[10]	CO5	L2
2	Explain web content mining tasks. web content mining tasks -10 marks	[10]	CO5	L2
3	Explain five phases in a process pipeline in Text mining. Diagram- 1 marks Pipeline phases-9 marks	[10]	CO5	L2
4	Solve the matrix multiplication using MapReduce. Write algorithm for matrix multiplication with one step. $A = \begin{matrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \end{matrix}$ $B = \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix}$ Solve problem- 6 marks Algorithm- 4 marks	[10]	CO4	L4
5	Explain with diagram Components of Hive architecture. components- 9 marks diagram-1 marks	[10]	CO4	L2
6	Write HIVE queries for CREATE, SHOW, DROP and PIG queries for LOAD , STORE. Each query – 2 marks	[10]	CO4	L3
7	Explain with diagram Hive Integration and Workflow Steps. Diagram -1 marks Workflow steps- 9 marks	[10]	CO4	L3

Solution:

1. Explain Hubs and Authorities. Explain HITS algorithm.

A hub is an index page that out-links to a number of content pages. A content page is topic authority. An authority is a page that has recognition due to its useful, reliable and significant information.

Figure 9.10(a) shows hubs (shaded circles) with the number of out-links associated with each hub. Figure 9.10(b) shows authorities (dotted circles) with the number of in-links and out-links associated with each link.

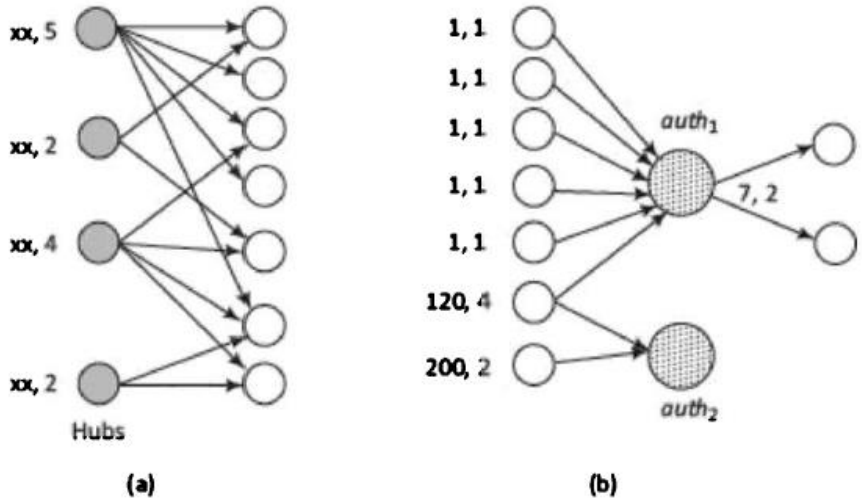


Figure 9.10 (a) Hubs (shaded circles) and (b) Authorities (dotted circles)

In-degrees (number of in-edges from other vertices) can be one of the measures for the authority. However, in-degrees do not distinguish between an in-link from a greater authority or lesser authority.

Authority, auth1 in Figure 9.10(b) has in-links from 6 vertices (in-degrees = 6) and auth2 has in-links to just 2 (in-degree = 2). However, auth 1 has link with six vertices with in-degrees = 1, 1, 1, 1, 1 and 120 (total = 125). Authority, auth 2 has links with two vertices with in-degrees= 120 and 200 (total= 220). Aut h2 has association with greater authorities. Therefore, in-degrees may not be a good measure as compared to authority.

Kleinberg (1998) developed the Hypertext-Induced Topic Selection (HITS) algorithm.6 The algorithm computes the hubs and authorities on a specific topic t. The HITS analyses a sub-graph of web, which is relevant to t.

Basis of computation is

(i) hubs are the ones, which out-link to number of authorities, and

(ii) authorities are the ones, which in-link to number of hubs. A bipartite graph exists for the hubs and authorities.

Consider a specifically queried topic t. Following are the steps:

1. Let a set of pages discover a root set R using standard search engine. Root pages may limit to top 200 for t.
2. Find a sub-graph of pages S, using a query that provides relevant pages fort and pointed by pages at R. Sub-graph S pages form Set for computations as it includes the children of parent R and limit to a random set of maximum 50 pages returned by a "reverse link" query.
3. Eliminate purely navigational links and links between two pages on the same host.
4. Consider only u (llull" 4-8) pages from a given hyperlink as pointer to any individual page. (Section 9.4.2)

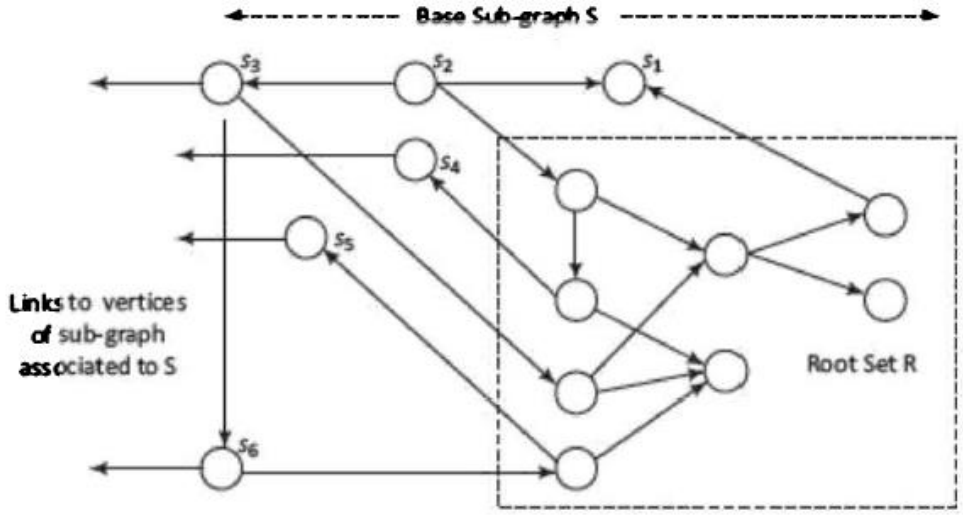


Figure 9.11 Sub-graph for HITS consisting of root set R of pages and base sub-graph S including all the pages pointed to by any page of R.

****

2. Explain web content mining tasks.

Mining Tasks for Web Content Analytics

1. Classification - A supervised technique which:

(i) Identifies the class or category a new web documents belongs to from the set of predefined classes or categories

(ii) Categories in the form of a term vector that are produced during a "training" phase

(iii)Employs algorithms using term vector to categorize the new data according to the observations at the training set.

2. Clustering - An unsupervised technique:

- Groups the web documents (clustered) with similar features using some similarity measure
- Uses no pre-defined perception of what the groups should be
- Measures most common similarity using the dot product between two web document vectors.
- Identifying the association between web documents - Association rules help to identify correlation between web pages that occur mostly together.

The other significant mining tasks are:

1. *Topic identification, tracking and drift analysis* - A way of organizing the large amount of information retrieved from the web is categorizing the web pages into distinct topics. The categorization can be based on a similarity metric, which includes textual information and co-citation relations. Clustering or classification techniques can automatically and effectively identify relevant topics and add them in a topic-wise collection library.

Adding a new document to a collection library includes:

(i) Assigning each document to an existing topic (category)

(ii) Re-checking of collection for the emergence of new topics

(iii) Tracking the number of views to a collection

(iv) Identifying the drift in a topic(s)

2. *Concept hierarchy creation* - Concept hierarchy is an important tool for capturing the general relationship among web documents. Creation of concept hierarchies is important to understand a category and sub-categories to which a document belongs. The clustering algorithms leverage more than two clusters, which merge into a cluster. That is merging the sub Document relevance descriclusters into a cluster.

- Important factors for creation of concept hierarchy include:

  i. Identifying the organization of categories, such as flat, tree or network

  ii. Planning the maximum number of categories per document

  iii. Building category dimensions, such as domain, location, time, application and privileges.

3. *Relevance of content* - Relevance or the applicability of web content can be measured with respect to any of the following basis:

   (i) Document relevance describes the usefulness of a given document in a specified situation.

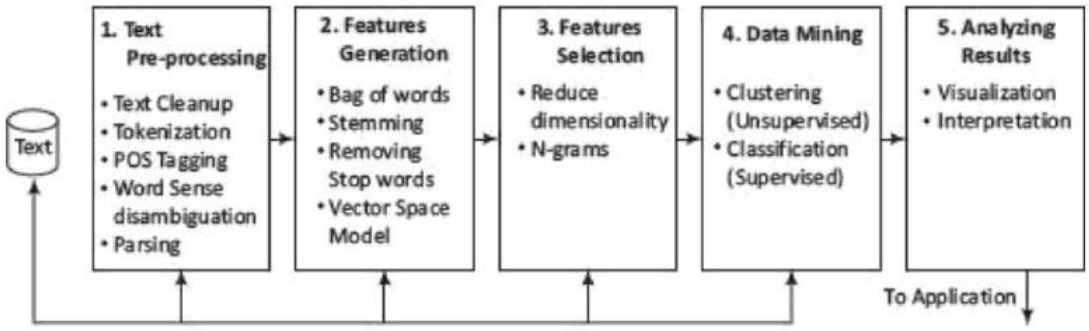
   (ii) Query-based relevance is the most useful method to assess the relevance of web pages. Query-based relevance is used in information retrieval tools such as search engines. The method calculates the similarity between query (search) keywords and document. Similarity, results can be refined through additional information such as popularity metric as seen in Google or the term positions in Altavista.

   (iii) User-based relevance is useful in personal aspects. User profiles are maintained, and similarity between the user profile and document is calculated. The relevance is often used in push notification services.

   (iv) Role/task-based relevance is quite similar to user-based relevance. Instead of a user, here the profile is based on a particular role or task. Multiple users can provide input to profile.

****

3. Explain five phases in a process pipeline in Text mining.



- The five phases for processing text are as follows:

Phase 1: Text pre-processing enables Syntactic/Semantic text-analysis and does the followings:

- Text *cleanup* is a process of removing unnecessary or unwanted information. Text cleanup converts the raw data by filling up the missing values, identifies and removes outliers, and resolves the inconsistencies. For example, removing comments, removing or escaping "%20" from URL for the web pages or cleanup the typing error, such as teh (the), don't (do not) [%20 specifies space in a URL].

- *Tokenization* is a process of splitting the cleanup text into tokens (words) using white spaces and punctuation marks as delimiters.
- *Part of Speech (POS) tagging* is a method that attempts labeling of each token (word) with an appropriate POS. Tagging helps in recognizing names of people, places, organizations and titles. English language set includes the noun, verb, adverb, adjective, prepositions and conjunctions. Part of Speech encoded in the annotation system of the Penn Treebank Project has 36 POS tags.4
- *Word sense disambiguation* is a method, which identifies the sense of a word used in a sentence; that gives meaning in case the word has multiple meanings. The methods, which resolve the ambiguity of words can be context or proximity based. Some examples of such words are bear, bank, cell and bass.
- *Parsing* is a method, which generates a parse-tree for each sentence. Parsing attempts and infers the precise grammatical relationships between different words in a given sentence.

Phase 2: Features Generation is a process which first defines features (variables, predictors). Some of the ways of feature generations are:

1. *Bag of* words-Order of words is not that important for certain applications. Text document is represented by the words it contains (and their occurrences). Document classification methods commonly use the bag-of-words model. Document classification methods then use the occurrence (frequency) of each word as a feature for training a classifier. Algorithms do not directly apply on the bag of words, but use the frequencies.
2. Stemming-identifies a word by its root.
    - Normalizes or unifies variations of the same concept, such as *speak* for three variations, i.e., speaking, speaks, speakers denoted by [speaking, speaks, speaker -+ speak]
    - Removes plurals, normalizes verb tenses and remove affixes.
3. Stemming reduces the word to its most basic element. For example, impurification -+ pure.
4. *Removing stop words* from the feature space-they are the common words, unlikely to help text mining. The search program tries to ignore stop words. For example, ignores *a, at, for,* it, *in* and *are.*
5. *Vector Space Model* (VSM)-is an algebraic model for representing text documents as vector of identifiers, word frequencies or terms in the document index. VSM uses the method of term frequency-inverse document frequency (TF-IDF) and evaluates how important is a word in a document.
6. When used in document classification, VSM also refers to the bag-of-words model. This bag of words is required to be converted into a term-vector in VSM. The term vector provides the numeric values corresponding to each term appearing in a document. The term vector is very helpful in feature generation and selection.

Phase 3: Features Selection is the process that selects a subset of features by rejecting irrelevant and/or redundant features (variables, predictors or dimension) according to defined criteria. Feature selection process does the following:

1. *Dimensionality* reduction-Feature selection is one of the methods of division and therefore, dimension reduction. The basic objective is to eliminate irrelevant and redundant data. Redundant features are those, which provide no extra information. Irrelevant features provide no useful or relevant information in any context.
2. Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) are dimension reduction methods. Discrimination ability of a feature measures relevancy of features. Correlation

helps in finding the redundancy of the feature. Two features are redundant to each other if their values correlate with each other.

3. *N-gram evaluation-finding* the number of consecutive words of interest and extract them. For example, 2-gram is a two words sequence, ["tasty food", "Good one"]. 3-gram is a three words sequence, ["Crime Investigation Department"].

4. *Noise detection and evaluation of outliers* methods do the identification of unusual or suspicious items, events or observations from the data set. This step helps in cleaning the data.

The feature selection algorithm reduces dimensionality that not only improves the performance of learning algorithm but also reduces the storage requirement for a dataset. The process enhances data understanding and its visualization.

Phase 4: Data mining techniques enable insights about the structured database that resulted from the previous phases. Examples of techniques are:

1. Unsupervised learning (for example, clustering)
    1. The class labels (categories) of training data are unknown
    2. Establish the existence of groups or clusters in the data

Good clustering methods use high intra-cluster similarity and low inter-cluster similarity. Examples of uses - blogs, patterns and trends.

2. *Supervised learning (for example, classification)*
    1. The training data is labeled indicating the class
    2. New data is classified based on the training set

Classification is correct when the known label of test sample is identical with the resulting class computed from the classification model.

Examples of uses are *news filtering application,* where it is required to automatically assign incoming documents to pre-defined categories; *email spam filtering,* where it is identified whether incoming email messages are spam or not.

Example of text classification methods are *Naive Bayes Classifier* and *SVMs.*

3. *Identifying evolutionary patterns* in temporal text streams-the method is useful in a wide range of applications, such as summarizing of events in news articles and extracting the research trends in the scientific literature.

Tf/Idf

Phase 5: Analysing results

(i) Evaluate the outcome of the complete process.

(ii) Interpretation of Result- If acceptable then results obtained can be used as an input for next set of sequences. Else, the result can be discarded, and try to understand what and why the process failed.

(iii) Visualization - Prepare visuals from data, and build a prototype.

(iv) Use the results for further improvement in activities at the enterprise, industry or institution.

****

4.Solve the matrix multiplication using MapReduce. Write algorithm for matrix multiplication with one step.

$$A = \begin{matrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \end{matrix} \quad B = \begin{matrix} 1 & 2 \\ 1 & 3 \end{matrix}$$

I/P file

matrix	i/k	j	value
A	0	0	1
A	0	1	2
A	1	0	2
A	1	1	1
A	2	0	3
A	2	1	4
B	0	0	1
B	0	1	~~2~~ 1
B	1	0	~~1~~ 2
B	1	1	3

Read ← direction for second matrix

Map function

i/k	value (matrix, j, value)
0	(A, 0, 1)
0	(A, 1, 2)
1	(A, 0, 2)
1	(A, 1, 1)
2	(A, 0, 3)
2	(A, 1, 4)
0	(B, 0, 1)
0	(B, 1, 2)
1	(B, 0, 1)
1	(B, 1, 3)

Interchange second last step to third last step **

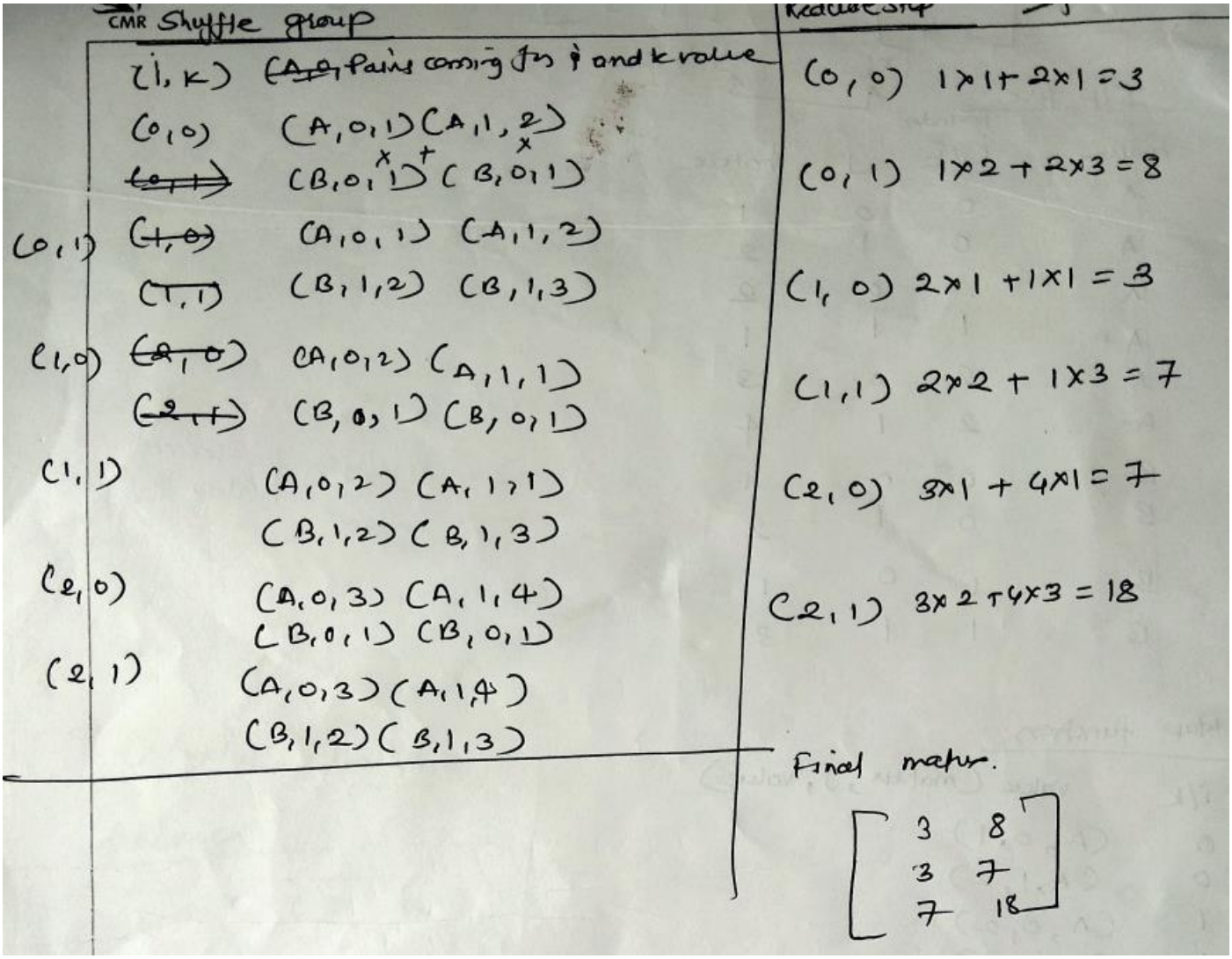
remove error

applies to any matrix even 3×3.

0	B, 0, 1
0	B, 1, 1
1	B, 0, 2
1	B, 1, 3

First matrix will be read Rowwise

Second matrix will be read columnwise



The product A.B = Natural join of tuples in the relations RA and RB followed by grouping and aggregation. Natural join of A (I,J, va) and B (J, K, vb), having only attribute J in common = Tuples (i, j, k, va, vb) from each tuple (i, j, va) in A and tuples (0,k, vb) in B.

l. MapReduce tasks for Steps 5 and 6: Five-component tuple represents the pair of matrix elements (aij, bjk). Requirement is product of these elements. That means four component tuple (i,j, k, va x vb), from equation (4.4) for elements Cik = Sum (aij.bjk) j= 1 to j

(a) Mapper Function: (i) Mapper emits the key-value pairs 0, (A, i, aij)) for each matrix element aij, and (ii) Mapper emits the key-value pair 0, (B, k, bjk)) for each matrix element aij.

(b) Reduce Function: Consider the tuples of A = (A, i, aij) for each key j, consider tuples of B = (B, k, bjk) for each key j. Produce a key-value pair with key equal to (i, k) and value= aij x bjk

1. A and B are just the names, may be represented by 0101 and 1010.

2. Next MapReduce Steps 7: Perform <I, K> i SUM (va x vb). That means do grouping and aggregation, with I and K as the grouping attributes and the sum of vax vb as the aggregation.

(c) The Mapper emits the key-value pairs (i, k, Vc) for each matrix element of C inputs with key i and k, and Vc from earlier task of the reducer Va x Vb.

(d)Reducer groups (i, k, vc) in C using [C, i, k, sum (Vc)J from aggregated values of Vc from sum (vJ Aggregation uses the same memory locations as used by elements Vc- C is just the name, may be represented by 1111.

MapReduce tasks for Steps 5 to 7 in a single step.

(e) Map Function: For each element aij of A, the Mapper emits all the key value pairs [(i, k), (A, j, aij)] for k = 1, 2, ... , up to the number of columns of B. Similarly, emits all the key-value pairs [(i, k), (B, j, bjk)] for i = 1, 2, ... , up to the number of rows of A. for each element bjk of B.

(f) Reduce Function: Consider the tuples of A = (A, i, aij) for each key j. Consider tuples of B = (B, k, bjk) for each key j. Emits the key-value pairs with key equal to (i, k) and value= sum of (aij x bjk) for all values j.

Memory required in one step MapReduce is large as compared to two steps in cascade. This is due to the need to store intermediate values of Vc and then sum them in the same Reducer step.

5. Explain with diagram Components of Hive architecture.

Hive was created by Facebook. Hive is a data warehousing tool and is also a data store on the top of Hadoop. An enterprise uses a data warehouse as large data repositories that are designed to enable the tracking, managing, and analyzing the data.

- **Hive Server (Thrift) - An optional service that allows a remote client to submit requests to Hive and retrieve results. Requests can use a variety of programming languages. Thrift Server exposes a very simple client API to execute HiveQL statements.**
- **Hive CLI (Command Line Interface)** - Popular interface to interact with Hive. Hive runs in local mode that uses local storage when running the CLI on a Hadoop cluster instead of HDFS.
- **Web Interface** - Hive can be accessed using a web browser as well. This requires a HWI Server running on some designated code. The URL http:// hadoop:<port no.> / hwi command can be used to access Hive through the web.
- **Metastore** - It is the system catalog. All other components of Hive interact with the Metastore. It stores the schema or metadata of tables, databases, columns in a table, their data types and HDFS mapping.
- **Hive Driver** - It manages the life cycle of a HiveQL statement during compilation, optimization and execution.

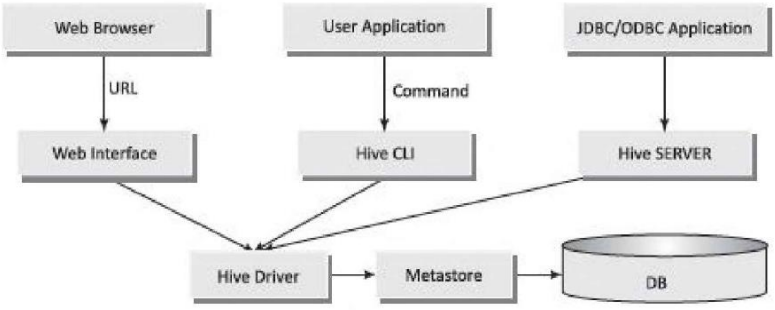


Figure 4.10 Hive architecture

6. Write HIVE queries for CREATE, SHOW, DROP and PIG queries for LOAD , STORE

LOAD Command The first step to a dataflow is to specify the input. Load statement in Pig Latin loads the data from PigStorage.

To load data from HBase: book load 'MyBook' using HBaseStorage();

For reading CSV file, PigStorage takes an argument which indicates which character to use as a separator. For example, book = LOAD 'PigDemo/Data/Input/myBook.csv' USING PigStorage (,);

For reading text data line by line:

'PigDemo/Data/Input/myBook.txt' USING book = PigStorage() LOAD AS (lines: chararray);

To specify the data-schema for loading: book = LOAD 'MyBook' AS (name, author, edition, publisher);

Store Command

Pig provides the store statement for writing the processed data after the processing is complete. It is the **mirror image** of the load statement in certain ways.

By default, Pig stores data on HDFS in a tab-delimited file using **PigStorage:**

STORE processed into '/PigDemo/Data/Output/Processed';

To store in HBaseStorage with a *using* clause: STORE processed into 'processed' **using HBaseStorage();**

To store data as comma-separated text data, PigStorage takes an argument to indicate which character to use as a separator: STORE processed into 'processed' using PigStorage(',');

CREATE [TEMPORARY] [EXTERNAL] TABLE [IF NOT EXISTS] [<database name>.] <table name>

[(<column name> <data type> [COMMENT <column comment>], ...)] [COMMENT <table comment>]

[ROW FORMAT <row format>] [STORED AS <file format>]

HiveQL database commands for data definition for the DBs and Tables are CREATE DATABASE, SHOW DATABASE (list of all DBs), CREATE SCHEMA, CREATE TABLE.

CREATE DATABASE|SCHEMA [IF NOT EXISTS] <database name>;

IF NOT EXISTS is an optional clause. The clause notifies the user that a database with the same name already exists. SCHEMA can be also created in place of DATABASE using this command

A command is written to get the list of all existing databases.

SHOW DATABASES;

A command is written to delete an existing database.

DROP (DATABASE|SCHEMA) [IF EXISTS] <database name> [RESTRICTICASCADE];

Q6 Explain with diagram Hive Integration and Workflow Steps.

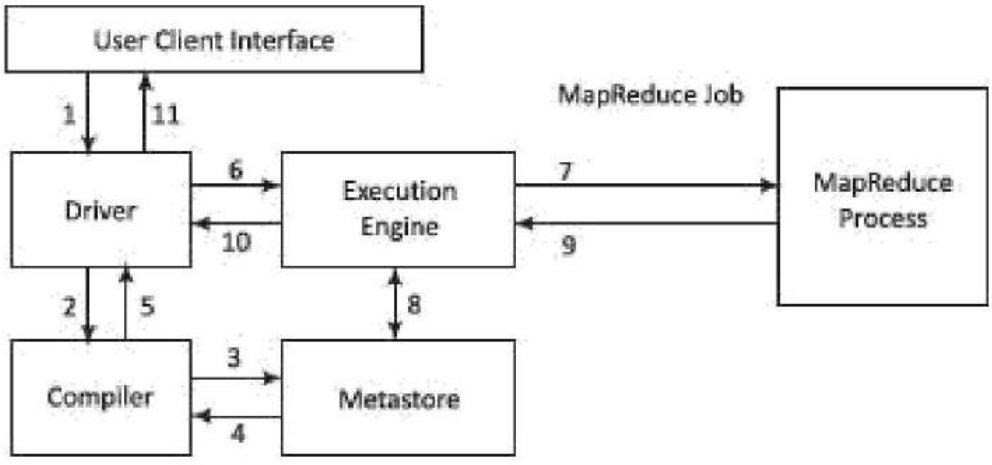


Figure 4.11 Dataflow sequences and workflow steps

Step No.	OPERATION
1	Execute Query: Hive interface (CLI or Web Interface) sends a query to Database Driver to execute the query.

2	Get Plan: Driver sends the query to query compiler that parses the query to check the syntax and query plan or the requirement of the query.
3	Get Metadata: Compiler sends metadata request to Metastore (of any database, such as MySQL).
4	Send Metadata: Metastore sends metadata as a response to compiler.
5	Send Plan: Compiler checks the requirement and resends the plan to driver. The parsing and compiling of the query is complete at this place.
6	Execute Plan: Driver sends the execute plan to execution engine.
7	Execute Job: Internally, the process of execution job is a MapReduce job. The execution engine sends the job to JobTracker, which is in Name node and it assigns this job to TaskTracker, which is in Data node. Then, the query executes the job.
8	Metadata Operations: Meanwhile the execution engine can execute the metadata operations with Metastore.
9	Fetch Result: Execution engine receives the results from Data nodes.
10	Send Results: Execution engine sends the result to Driver.
11	Send Results: Driver sends the results to Hive Interfaces.