GUJARAT TECHNOLOGICAL UNIVERSITY

Syllabus for Master of Computer Applications, 2nd Semester
Subject Name: Big Data Tools
Subject Code: 629407

With effective from academic year 2020-21

1. **Learning Objectives:**
   - **a.** To understand basics of Big Data
   - **b.** To understand MongoDB, Hadoop, Map reduce, Pig and Hive

2. **Prerequisites:** Working knowledge of Programming Language and Database Concepts

3. **Contents:**

Unit	Course Content	Weightage percentage
Unit I	**Unit 1: Introduction to Big Data** Types of Digital Data: Classification of Data (Structured, Semi-structured and Unstructured), Characteristics of Data, Evolution of Big Data, Definition of Big Data, Challenges of Big Data, Characteristics of Big Data (Volume, Velocity, Variety), Other characteristics of Big Data which are not Definitional Traits of Big Data, Why Big Data? Are we Information Consumer or Producer? Traditional BI vs Big Data, Typical Data Warehouse Environment, Typical Hadoop Environment, what is Changing in Realms of Big Data? Terminologies used in Big Data Environments	**15%**
Unit II	**Unit 2: Introduction to NoSQL and Hadoop** **NoSQL:** Introduction: Where is it used? What is it? Types of NoSQL databases, Why NoSQL?, Advantages of NoSQL, Use of NoSQL in Industry, SQL vs NoSQL, NewSQL **Hadoop:** Introduction, Distributed Computing Challenges, History of Hadoop, Overview of Hadoop and Hadoop Ecosystems, Features and key advantages of Hadoop, Versions of Hadoop, Hadoop distributions, RDBMS versus Hadoop, Hadoop vs SQL, Integrated Hadoop Systems offered by leading market vendors, Cloud based Hadoop solutions, HDFS, Processing data with Hadoop, Managing Resources and applications with Hadoop YARN, Interacting with Hadoop Ecosystem	**25%**
Unit III	**Unit 3: Introduction to MongoDB and Map Reduce** **MongoDB**: Introduction: What is MongoDB? Why Mongo DB? (using JSON, Creating or generating a unique key, Support for Dynamic Queries, Storing Binary Data, Replication, Shading, Updating information in –place), Terms used in RDBMS and Mongo DB, Data types in Mongo DB, MongoDB Query Language **Map Reduce:** Data Flow, Map, Shuffle, Sort, Reduce, Hadoop Streaming, mrjob, Installation, word count in mrjob, Executing mrjob	**25%**
Unit IV	**Unit 4: Introduction to HIVE and Pig** **HIVE:** Introduction: What is HIVE? HIVE Architecture, HIVE data Types, HIVE File Formats, HIVE Query Language, RCFile implementation, SerDe, User-Defined Functions (UDF)	**25%**

	Pig: Introduction: What is Pig? The anatomy of Pig, Pig on Hadoop, Pig philosophy, Use Case for Pig- ETL Processing, Pig Latin overview, Data types in Pig, Running Pig, Execution modes of Pig, HDFS commands, Relational operators, Eval function, Complex Data Types, Piggy Bank, User-defined Functions, Parameter substitution, Diagnostic Operator, Word Count Example using Pig, when to use and not use Pig? ,Pig at Yahoo, Pig vs HIVE.	
Unit V	**Unit 5: Overview of SPARK** Introduction to Data Analysis with Spark, Downloading Spark and Getting Started, Programming with RDDs	**10%**

4. Text Book(s):

1) Seema Acharya, Subhashini Chellappan, " Big Data and Analytics", Wiley India Pvt. Ltd.,2015
2) Matei Zaharia, Patrick Wendell, Andy Konwinski, Holden Karau, "Learning Spark", O'Reilly Media,2015
3) Zachary Radtka and Donald Miner, "Hadoop with Python'', O'Reilly Media,2016
(Free eBook is available on the following link) (As on 12-10-2018)
https://www.oreilly.com/programming/free/hadoop-with-python.csp

5. Reference Books:

1) Shashank Tiwari, "Professional NoSQL", Wiley India Pvt. Ltd.,2011
2) Kyle Banker,Peter Bakkum,Shaun Verch,Douglas Garrett,Tim Hawkins,"MongoDB in Action", DreamTech Press, 2nd Edition ,2016
3) Chris Eaton,Paul Zikopoulos,Tom Deutsch,George Lapis,Dirk Deroos,"Understanding Big Data : Analytics for Enterprise Class Hadoop and Streaming Data", Mcgraw Hill Education (India)Pvt.Ltd.,2012
4) Tom White,"Hadoop: The Definitive Guide",O'Reilly Media,4th Edition,2015
5) Vignesh Prajapati,"Big Data Analytics With R and Hadoop", Packt Pub Ltd ,2013
6) Services, "Big Data - Black Book", Dreamtech Press,2016

Web Resources:

1) http://www.bigdatauniversity.com
2) http://www.mongodb.com
3) http://hadoop.apache.org/

6. Unit wise coverage from Textbook(s):

Unit 1	**Book#**	**Topics**
I	1	Chapter. 1, 2, 3.12
II	1	Chapter 4,5
III	1,3	Chapter 6 (Book 1), Chapter 2 (Book 3)
IV	1	Chapter 9,10
V	2	Chapter 1,2 and 3 (For Chapter 2 and 3, only Python, No Java, No

		Scala)

7. Accomplishment

Upon completion of this course, students will be able to do the following:

- Students will learn difference between conventional SQL query language and NoSQL basic concepts
- Students will be able to design and build MongoDB based Big Data Applications and learn MongoDB query language
- Students will be able to write Map-Reduce based Applications

Practical List

Part I: Mongo DB

- Learn to Use MongoDB Atlas (The Cloud Version of MongoDB)
- Install and configure MongoDB

MongoDB Shell Commands / Queries: View all databases, Create new database, Drop existing database, View current database, Switch over to a given database, db. Help(), Display statistics of a given database, Display current version of MongoDB Server, Display list of collections in current database, Create Collection, Drop Collection, CRUD operations (Create, Read, Update, Delete), Insert, Update else insert, save, update, remove, Find, Dealing with Using NULL Values, Count, Limit, Sort, Skip, Arrays and Array Operations, Aggregate

1) Create a Student Master database with a collection called "Student" containing documents with some or all of the following fields: StudentRollNo, StudentName, Grade, Hobbies, and DOJ. Perform the following operations on the database:

a) Insert 10 Records in the database.
b) Find the document wherein the "StudName" has value "Ajay Rathod".
c) Find all documents in proper format. (Without _Id field)
d) Retrieve only Student Name and Grade.
e) Retrieve Student Name and Grade of student who is having _id column is 1.
f) Add new field "Address" in Student Collection.
g) Find those documents where the Grade is set to 'VII'.
h) Find those documents where the Grade is not set to 'VII'.
i) Find those documents where the Hobbies is set to either 'Chess'or is set to 'Dancing".
j) Find those documents where the Hobbies is set neither to 'Chess' nor is set to 'Dancing".
k) Find those documents where the student name begins with 'M'.
l) Find those documents where the student name has an "e" in any position.
m) Find those documents where the student name ends in "a".
n) Find total number of documents.
o) Find total the number of documents where Grade is 'VII'.
p) Sort the documents in ascending order of student name.
q) Display the last two records.

2) Create a MovieMaker Database with a collection called "Movies "containing documents with some or all of the following fields: titles, directors, years, actors. Perform the following operations on the database (either in the console or using any programming language):

a) Retrieve all documents
b) Retrieve all documents with Director set to "Quentin Tarantino"
c) Retrieve all documents where actors include "Brad Pitt".
d) Retrieve all movies released before the year 2000 or after 2010.
e) Add a synopsis to "The Hobbit: An Unexpected Journey": "A reluctant hobbit, Bilbo Baggins, sets out to the Lonely Mountain with a spirited group of dwarves to reclaim their mountain home - and the gold within it - from the dragon Smaug."
f) Add a synopsis to "The Hobbit: The Desolation of Smaug": "The dwarves, along with Bilbo Baggins and Gandalf the Grey, continue their quest to reclaim Erebor, their homeland, from Smaug. Bilbo Baggins is in possession of a mysterious and magical ring."
g) Add an actor named "Samuel L. Jackson" to the movie "Pulp Fiction"
h) Find all movies that have a synopsis that contains the word "Bilbo"
i) Find all movies that have a synopsis that contains the word "Gandalf"
j) Find all movies that have a synopsis that contains the word "Bilbo" and not the word "Gandalf"
k) Find all movies that have a synopsis that contains the word "dwarves" or "hobbit"
l) Find all movies that have a synopsis that contains the word "gold" and "dragon".
m) Delete the movie "Pee Wee Herman's Big Adventure"

3) Create a database named "BookStore" in MongoDB with a collection called "Books" containing documents with some or all of the following fields: bookId, bookTitle, authors (containing fields: authorName), publicationYear, publisher, Orders (containing fields: OrderedId, orderDate, customerName, price, quantityOrdered, discount).

4) Note that a book may have one or more authors and orders. Also, the same Ordered can be present in one or more books. Perform the following operations on the database (either in the console or using any programming language):

a) Insert records for 10 books from 5 authors, and at least 20 orders in total.
b) Update the title of a particular book.
c) Display all the books having less than 3 authors and sort by book name.
d) Display the number of books from each publisher.
e) Use Map Reduce function to display the total quantity of books ordered for each date.
f) Use Map Reduce function to display the discount offered to a particular customer.

5) Create a database named "Store" in MongoDB with a collection called "Sales" containing documents with some or all of the following fields: customerId, customerName, gender, dataOfBirth, contactNumber, address (containing fields: houseNo, street, area, city, pincode), orders (containing fields: orderId, orderDate, items (containing fields: itemId, itemName, itemPrice, quantityOrdered, discount)). Note that some customers may not provide their date of birth and/or contact number. Also, not all products would be sold at a discount. Perform the following operations on the database (either in the console or using any programming language):

a) Insert records for 3 customers and 5 items in at least 20 orders.
b) Update the contact number of a particular customer.
c) Display customerId, customerName, gender, contactNumber, of customers residing in "Ahmedabad".
d) Display city-wise count of customers
e) Use MapReduce function to display the number of times each item was sold.

6) Create a database "BookStore" with a collection called "Books" containing documents with some or all of the following fields: Category, BookName, Author, quantity, price, pages. Perform the following operations on the database:

a) Insert Records for 5 books.
b) Write Map & Reduce functions to split the books into the following two categories: Bigbooks, Smallbooks. (Books which have more than 300 pages should be in the Big books category. Books which have less than 300 pages should be in the Small books category.)
c) Count the number of books in each category
d) Store the output as follow as documents in a new collection called "Book Result".

Book Category	**Count of the Books**
Big books	2
Small books	3

Part II: Hadoop HDFS

- Installation and configuration for: Apache Hadoop Stand-Alone Mode and Pseudo
- Distributed Mode
- Installation and configuration for: Apache Hadoop Real Cluster consisting of a single
- Master and Two Slave nodes.
- Test the above set-up with sample examples bundled along with the downloaded package.
- To develop and execute sample programs like word-count, maximum temperature, etc. Using Python with Map-Reduce in Hadoop
- HDFS Commands: -ls, -ls -R, -mkdir, -put, -get, -copyFromLocal, -copyToLocal, -cat, -cp, -rm-r

1) Create a file "Sample" in a local file system and export it to the HDFS File System.
2) Write the HDFS command for copying a "Sample" file from HDFS to local File System.
3) Write HDFS commands for creating "Test" directory in HDFS and then removing that directory.
4) Write HDFS command to display complete list of directories and files of HDFS.
5) Write HDFS command for displaying the contents of "Sample" text file in HDFS on screen.
6) Write HDFS command for copying an existing "Sample" file in a "Test" HDFS directory to some another HDFS directory.

Part III: MapReduce

1) Prepare an "input" folder containing multiple text files. Create a program using MapReduce that would accept the path to the "input" folder and generate an "output" folder having a text file containing the total number of occurrences of each single word present in text document. For example, if the text containing in input files is as follows:
2) "We thank you for your visit to Ahmedabad. We hope that you would visit us again."
3) The Output should be as follow:

Word	Word Length	No of occurrences
We	2	2
To	2	1
Us	2	1
You	3	2
For	3	1
Your	4	1
That	4	1
Hope	4	1
Thank	5	1
Visit	5	2
Would	5	1
Again	5	1
Ahmedabad	9	1

1) Write a program for Matrix Vector Multiplication using MapReduce.
2) Write a program to perform Union, Intersection and Difference operation using
3) MapReduce on following files.

Input files:

1) Content of file 1 (apple, orange, mango, apple, banana)
2) Content of file 2 (apple, apple, plum, kiwi, kiwi, mango, mango)
3) Content of file 3 (orange, orange, plum, grapes, kiwi, mango, apple)

Part IV: Pig

- Install and configure Apache Pig
- Test the Pig Installation for local and map-reduce mode execution
- Test the Pig Installation for Interactive (Grunt Shell) and Batch Mode (. pig file) Execution
- Develop UDF (User Defined Function) in Python for Pig

Working with Pig Operators/Functions (LOAD, DUMP, FOREACH, GROUP, DISTINCT, LIMIT, ORDER BY, JOIN, UNION, SPLIT, SAMPLE, AVG, MAX, COUNT, TUPLE, MAP,PIGGY BANK, PARAMETER SUBSTITUTION, DESCRIBE, Simple Problems like Word Count using PIG

1. Write a pig script to load and store "Student data". (Student file contain Roll no, Name, Marks and GPA).
    a. Filter all the students who are having GPA>5.

b. Display the name of all Students in Uppercase.
c. Group tuples of students based on their GPA.
d. Remove duplicates tuple of Student list.
e. Display first three tuples from "student" relation.
f. Display the names of students in ascending order.
g. Join two relations namely Student and department (Rno, DeptNo, DeptName) based on the values contain in the roll no column.
h. Merge content of two relation Student and department.
i. Partition a relation based on the GPA's acquired by students.
j. To calculate the average marks for each student.
k. Calculate maximum marks of each student.
l. Count the number of tuples in a bag.

2. Load the file menu.csv (Category, Name, Price) and write one Pig script

a. Which meals cost more than 30.00?
b. Which meals contain the word "Panner"?
c. Which are the 10 most expensive meals?
d. For every day, what's the average price for a meal?
e. For every day, what's the most expensive meal?

- Write a program to count Word on Pig.
- Write a pig script to spilt customers for reward program based on their life time values.

If Life time values is >1000 and < =2000 then Silver Program
If Life time values is >20000 then Gold Program
Input:

Customers	Lifetime value
Jack	25000
Smith	8000
David	12000
John	15000
Scott	12000
Lucy	28000
Ajay	12000
Vinay	30000
Joseph	21000
Joshi	25000

1) Create a data file for below schemas:
Order: CustomerId, ItemId,ItemName,OrderDate,DelivaryDate
Customer: CustomerId,CustomerName, Address,City,State,Country

a. Load Order and Customer Data.
b. Write a pig latin Script to determine number of items bought by each customer.

- Do the Following:

1. Create a file which contains bag dataset as shown below.

User Id	From	To

user1001	user1001@sample.com	{(user003@sample.com),(user004@sample.com), (user006@sample.com)}
user1002	user1002@sample.com	{(user005@sample.com),(user006@sample.com)}
user1003	user1003@sample.com	{(user001@sample.com),(user005@sample.com)}

2. Write a pig latin statement to display the names of all users who have sent emails and also a list of the people that have sent the email to.
3. Store the result in a file.

7) Create a UDF to convert name into uppercase.

Part V: Hive

- Install and configure Apache Hive
- SerDe and User Defined Function Creation in Hive using Java

Create database, display list of existing databases, describe database, describe extended database, alter database properties, to make a given database as current database, drop database, create managed table, create external table, loading data into a table, working with collection data types, querying a table using select, querying collection data types, create static partition and load data into it from original table, static partition creation using alter, create dynamic partition, load data into dynamic partition, create bucket, create view, query view, drop view, sub-query, joins, Aggregation, Group By and Having, RC File Implementation

1. Create a data file for below schemas
Order: CustId,ItemId,ItemName,OrderDate,Delivary Date
Customer:CustId,CustName,Address,City,State,Country

1) Create a table for Orders and Customer Data.
2) Write Hive Query Language to find number of items bought by each customer.

2. Create a partition table for Customer Schema to reward customer based on their life time value.

Customer Id	Customers	Lifetime value
1001	Jack	25000
1002	Smith	8000
1003	David	12000
1004	John	15000
1005	Scott	12000
1006	Lucy	28000
1007	Ajay	12000
1008	Vinay	30000
1009	Joseph	21000
1010	Joshi	25000

1) Create partition table if life time value is 12000.
2) Create partition table for all life time values.