Work Integrated Learning Programmes

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Computer Science & Information Systems

Assignment- Big Data Systems MM: 15

1. PIG

Describe a problem statement and choose relevant data set to perform analysis using PIG Commands.
The Problem statement should

a) Match to a real life system.
b) The data set for problem should contain large number of records (Approx. 10000).
c) The data set should have sufficient number of attributes.

Identify 5 different kinds of relevant analysis and Write PIG commands for the same. The analysis should be of reasonable complexity (which involves writing multiple PIG commands). Clearly mention the input and output for each kind of analysis. [5]

2. MongoDB

Pick a Json data-setthat can be stored in MongoDB and perform analysis using MongoDB queries. The dataset should

a) Match to a real life system.
b) Should contain more than 1000 json records
c) The data set should have adequate number of attributes

Identify 25 analysis scenarios (5 each student) and write MongoDB queries for the same. The analysis should be meaningful and should have adequate complexity. Kindly make use of various kinds of operators and statistical functions. [5]

3. Connecting MongoDBAtlas with Programming Language

a) Setup a MongoDB cluster using Mongo Atlas service. Further connect MongoDB cluster with programming language (Python/Java) and demonstrate CRUD operations on MongoDB by embedding MongoDB queries in programming language. [2]

b) Configure the MongoDB cluster for different read/write consistencies and observe the performance of read/write operations with respect to different consistency configurations. [3]

Note

- This is group assignment. Each group should choose unique problem statements and datasets for PIG commands and MongoDB queries
- Kindly make use of Mongo shell for executing retrieval queries.
- Do not choose any problem statement that has been discussed in text book(s), reference book(s), or lectures.
- The solutions for the problem statement should not be readily available on the Internet/Books. The same would be likely to be rejected. However, you can refer to internet sources for data-sets.

Deliverables

1. Source of PIG and Json data-sets.
2. Description of data-sets, analysis, PIG commands and MongoDB queries
3. PIG Commands with their output snapshots
4. MongoDB Queries with their output snapshot
5. Programming language code to connect to MongoDB Atlas.
6. CRUD operations using Programming language on MongoDB cluster with output snapshots.
7. Consistency Configurations of Read/Write on MongoDB cluster and snapshots of read/write performance under each consistency.

There may be viva/presentation post submission.

How to Submit

1. Create one document of all deliverables.
2. Convert it into Pdf format.
3. Upload the pdf on e-learn portal. Name your file as **"Group-[number]"**

Evaluation Criteria

1. Scope of problem statements and analysis.
2. Fully functional PIG analysis, MongoDB analysis queries, programming language connectivity with MongoDB CRUD operations using programming language on MongoDB dataset.
3. Fully functional Read/Write Consistency configurations on MongoDB.
4. Presentation in the submission document.

Timelines for Submission

Kindly submit the assignment latest by Nov 12, 2023.

Plagiarism cases will be penalized strictly.