BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
CS/SS G527 Cloud Computing
I Semester 2016-17
Assignment-1
Weightage: 10% (70M) **Due Date of Submission:** 29-SEP-2015

==

Important to Note:

1. Group of maximum 3 students.
2. 2 programming problems and one case study.
3. For any clarifications please contact me (khari@pilani.bits-pilani.ac.in).

Plagiarism will be thoroughly penalized.

==

P1. Performance evaluation of Hadoop, MPI and Spark platforms on a search problem described below. [30M]

- The problem is about searching a huge number of text files using a pre-built word-level inverted index. This index is built by a background job using Hadoop, MPI or Spark platforms. This index is used for searching. CLI or GUI should be provided to search for a phrase, and it should list all the files where the phrase is found.
- Build a cluster of N nodes (N>=3).
- Dataset (with M files) to be used is available here. Description of the dataset is here. Each node in the cluster processes M/N files (approx) to build the word-level inverted index. Time and memory consumed by the processes for building the index should be noted for each of the platforms. These statistics should be reported in the form of a graph (on X-axis there should be 3 markings corresponding to Hadoop, MPI and Spark).
- Query can be submitted (via CLI or GUI) to any of the nodes. Time and memory consumed to execute the query should be noted and presented in the form of a graph. The results should be presented for at least 1000 queries (i.e. on X-axis there should be 1...999 markings).

Deliverables:

- Brief Design Documents (2 pages, typewritten) in design.doc/.pdf
- Code in folder named Code
- Test cases in testcases.txt
- Performance Plots in plots.doc/.pdf
- Data files should not be submitted

P2. Understanding Consensus Algorithms [24M]

In this problem, implement 2PC and Paxos Commit protocols to ensure data consistency among the replicas. The same data is replicated across all nodes in the cluster. Cluster should have at least 3 nodes. The problem is described below.

- Use MPI as the platform for programming.
- There is one text file "store.txt" which contains one line per user. The line has fields
  <userid>$<balance_amount>$<last_transaction_timestamp>\n
  OR
  Alternatively SQLite database can be used to store the table.
- store.txt/SQLite is replicated on all nodes before any transaction starts.

- User can submit his transaction [view balance/debit/deposit] on any node. Provide CLI/GUI for this.
- The implementation should ensure consistency of store.txt in all nodes in the following cases:
    - user can submit transactions simultaneously(approx.) on two nodes.
    - one of the nodes can be restarted
    - multiple users can operate transactions at the same time.

Deliverables:

- Brief Design Documents (2 pages, typewritten) in design.doc/.pdf
- Code in folder named Code
- Test cases in testcases.txt
- store.txt/SQLite db

P3. Describe the consensus/consistency algorithms (along with safety and liveness properties) used in the following cloud services/products. [16M]

- Apache Cassandra
- Apache Zookeeper
- Amazon DynamoDB
- Amazon S3
- Google Spanner
- Google Chubby
- Apache HBase
- Facebook Hydrabase

Deliverables:

- Assignment should be hand-written in A4 sheets and submitted during class hour on 29th Sep.

P4. [Required to be done only by 4 member group]

Extend P2 by implementing Byzantine Agreement protocol and 3PC. Rest of the details are same as that of P2. [20M] [Total 90M will be scaled down to 70M].

Deliverables:

- Brief Design Documents (2 pages, typewritten) in design.doc/.pdf
- Code in folder named Code
- Test cases in testcases.txt
- store.txt/SQLite db

How to upload?

- ▪ Create group.txt file and put idno, name of members into this file.
- ▪ Make a directory for each problem like P1, P2 etc and copy deliverables into these directories.
- ▪ Tar all of them including group.txt into idno1_idno2_idno3_assignment1.tar
- Upload at the course webpage.

===End of assignment===