Internship Objectives

Evaluate of Hadoop technologies for huge data problems in CSIT.

Problem Set A:

Managing of huge data from web captures

1. Storage
   - Data: Web content, meta information, user value-add information
   - Traffic: 30 GB of Web content daily
   - Volume: 30 million records, 5% growth annually
   - Format: HTML, XML, PDF, MS OFFICE, Meta Data
   - Operations: Write many, read many
   - Compare to commercial Content Management Systems (CMS): EMC Documentum, IBM Lotus Notes
2. Application
   - Entity Extraction
   - Link Analysis
   - De-duplication
   - Search
   - Horizon Tracking - Timeline of events

Problem Set B:

Managing of huge archival data

1. Storage
   - Data: Web content, meta information, user value-add information
   - Volume: Trillion of records
   - Format: HTML, XML, PDF, MS OFFICE, Meta Data
   - Operations: Write once, read many
   - Migration (Mass importing from legacy systems)
2. Application
   - Entity Extraction
   - Link Analysis
   - Search
   - Horizon Tracking - Timeline of events

Todo:

1. Building foundation (March)
    a. Database, Content management system (CMS), Apache, Filesystem, Open source, … etc
    b. Deliverables:
        i. Tech Foundation Mindmap
2. Understand Hadoop - (March - April)
    a. MapReduce, HDFS, Architecture
    b. Use cases of Hadoop deployments, customers, and service providers
    c. Deliverables:
        i. Hadoop Mindmap
        ii. Report Scoping
        iii. Findings Report1
3. Explore Hadoop related technologies (April-May)
    a. Hbase, Cassandra,Zookeeper...
    b. Deliverables:
        i. Hbase Mindmap
        ii. Cassandra Mindmap
        iii. Report Scoping
        iv. Findings Report2
4. Evaluate Hadoop for CSIT problem sets (June-July)
5. Final Reports (~1 Month)

References:

1. http://en.wikipedia.org/wiki/Apache_Hadoop
2. http://www.cloudera.com/
3. E:\Cloudera Hadoop Training
4. http://hadoop.apache.org/
5. http://thecloudtutorial.com/hadoop-tutorial.html
6. http://www.cloudera.com/resource/intorduction-hbase-todd-lipcon/ - Intro to HBase
7. http://www.youtube.com/watch?v=egwgKhH94z8 - Intro to Cassandra