Understanding Big Data Fundamentals To Prepare Future Data Scientists
Hema Bahirwani(hgb1348@rit.edu)
Advisor: Professor Carlos Rivero(crr@cs.rit.edu)

**Goal**
Teaching industry standards of working with Big Data so that students get experience solving real world data science challenges.

**Structured Guideline**
Designing a structured guideline to strengthen the foundation of traditional database systems and to introduce Big Data concepts.

**Week 1:** Big Data overview
Big Data and its significance, discuss the course structure and introduce RDBMS

**Week 2:** Data Modeling
Three schema architecture, ER concepts: entity, relationships, cardinalities, Conceptual Model, Logical Model, Physical Models

**Week 3:** Introduction to SQL
DCL, DDL, DML, constraints, CRUD, operations, assertions, actions as triggers

**Week 4:** More on SQL
Joins, normalization 1NF, 2NF, 3NF, aggregation, indexing

**Week 5:** NoSQL
Unstructured data, nosql DBs, data modeling, collections, documents, setting up mongo

**Week 6:** More on MongoDB
CRUD operations, indexing, RDBMS to MongoDB

**Common Theme: Internet Movie DB**
Devising a common, interesting and relevant theme, and designing contents encouraging the use of Big Data concepts

**Functional Dependency**

**Data Load**
Loading about 50 Million data records from a TSV file
for every data instance:
```
INSERT INTO <TABLE> (col1, col2, col3)
VALUES(value1, value2, value3);
```
LOAD DATA LOCAL INFILE <FILE_PATH>
INTO TABLE <TABLE_NAME>
FIELDS TERMINATED BY '\t';

**Time Taken:** ~ 8 Hours to load Title Information
**Time Taken:** ~ 5 minutes to load Title Information

**SQL -> MongoDB**
Pycharm Heap Space
Min: 2GB, Max: 4GB
Result: Out of Memory

**Find all the distinct Actors who are known for the titles of horror movies**
Increase the Server Timeout Span to at least 2500 s
JOIN is efficient than using IN SELECT statements for millions of records

**Sorting a huge collection in MongoDB**
Encouraging indexing or limiting the number of records

**Inserting data into a table from existing tables**
Result: The number of locks exceeds the lock table size