CIS442F Big Data

Course for MS Business Analytics Students

Dr. David Tilson
David Tilson mini-bio

Joined University of Rochester in 2007
PhD in Information Systems from Case Western Reserve University (2008)

Prior industry experience

- Consulting with McKinsey & Co. in high-tech, finance, insurance, energy...
- Product/project manager / system architect in telecom and broadcast R&D

Teaching

- Developed courses for MS/MBA
  - Big Data Analytics
  - Business Process Analysis & Design
  - Consulting-based experiential courses
- Also
  - Statistics and problem-solving for EMBA
  - Executive teaching on project management, teamwork, and problem-solving

Research

- Digital infrastructures and platforms
- Mobile communications and computing
- Technical standards
- Healthcare operations
- Big data???
Big Data class scheduled at end of MSBA program

Key implication: Students already familiar with R, Python, Machine Learning, and SQL
Course taught over 9 weeks (2 x 90min sessions per week)

<table>
<thead>
<tr>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
<th>Week 4</th>
<th>Week 5</th>
</tr>
</thead>
</table>
| Intro to big data  
• Distributed computing  
• Hadoop & MapReduce  
• HDFS and S3  
• Intro to Linux  
• Connecting to cluster | Apache Pig  
• Pig Latin  
• Zeppelin notebooks  
• ETL  
• Weblog analysis  
• Storage formats | Apache Hive  
• HiveQL (HQL)  
• Metadata store  
• Data ingestion  
• Twitter sentiment analysis using Hive | | |

Week 6  
<table>
<thead>
<tr>
<th>Week 7</th>
<th>Week 8</th>
<th>Week 9</th>
</tr>
</thead>
</table>
| Apache Spark 2.x  
• Transformations, actions & lazy evaluation  
• Spark SQL  
• Working with DataFrames (not RDDs)  
• PySpark (and some scala)  
• General data processing | Machine Learning in Spark MLlib  
• Feature extraction and selection  
• Regression and classification models  
• Hyperparameter tuning  
• ML pipelines  
• Cluster models and geo data | Exam  
• SparkR  
• Sparklyr  
• Big data on the cloud |
Syllabus and other resources available from this link


Includes links to other resources – including videos on just enough Unix command line

- Books about Hadoop and Big Data (none are required)
- Connecting to Hadoop and Configuring Software (Windows users)
- Connecting to Hadoop and Configuring Software (Mac version)
- Getting started with Linux
- Apache Pig (additional video with tips for running Pig scripts from 1)
- Apache Hive
- Apache Spark

Course elements

- 18 x 90min class sessions
  - Lectures on principles, tech, examples
  - Code walk-throughs
  - Some hands-on in class
- 8 coding homeworks (~one per week)
- 14 in-class quizzes (based on lecture and homeworks)
- Final exam (could do project)
There are many things to learn about to teach such a course.

Big data contexts in business, CS of distributed computing, Hadoop, Hdfs, MapReduce, Java, Linux CLI, bash scripting, AWS, EMR, AWS CLI, Pig Latin, SQL, Hive, Python, Pandas, Scikit-Learn, Scala, pySpark, Hue, Angular (javascript), Zeppelin, markdown...
Cloudera have supported the class. I drew upon many other resources.

Also input from De Liu (Minnesota) and Miriam Allalouf (Tel Aviv University)
We built a cluster using Amazon Web Services

**Need to use S3 specific software**
(also need IAM user keys)

**Easy form scripts**
but clumsy from command line

**S3 Buckets**

**hdfs dfs commands**

**SFTP**

**Hadoop Cluster**

**Rest of world**

**Ingestion**

**Laptop**
- SFTP client (e.g. WinSCP or Cyberduck)
- S3 client (e.g. Cloudberry Explorer for S3 or Cyberduck)

**Use s3:// URLs in scripts and Zeppelin notebooks**
(also need IAM user keys)
Using a cluster is not for the faint-hearted

**Estimated cost**
- $6k cluster time (Spark was most expensive part)
- $2k admin time
- $10k consulting time (one-off)

Asking AWS for $5k credit (~50 students) and they are considering more turn-key solution (no promises yest)

>50% cost reduction by active management (off at night, reset). Could be less than $50/student at 50 students for 10 weeks

**AWS provide good support (solution architects)**
- But they were learning too
- Multi-user different and complex across AWS, Linux, HDFS, Pig, Hive, S3, Hue, Spark, and Zeppelin
- Will use us as case study in multi-tenancy cluster operation

We used Zeppelin notebooks. They are very similar to Jupyter notebooks.

Download and import this notebook from Blackboard.
Code walkthrough much easier with notebooks than shell scripts and some visualization possible

---

**Run Linear Regression**

```
% pyspark
from pyspark.ml import Pipeline
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.regression import LinearRegression, LinearRegressionModel

# Set Features
features = VectorAssembler(inputCols=['X'], outputCol='features')

linreg = LinearRegression().setLabelCol("y")
pipeline = Pipeline(stages=[features, linreg])
pipeline_model = pipeline.fit(data)
```

 Took 22 sec. Last updated by simonstudent65 at May 08 2018, 11:21 AM.

---

**% pyspark**

# We can see that `pipeline_model` is indeed a pipeline model with several stages
print('pipeline_model is a {}'.format(type(pipeline_model)))

pipeline_model is a `<class 'pyspark.ml.pipeline.PipelineModel'>`

 Took 0 sec. Last updated by simonstudent65 at May 08 2018, 11:24 AM.

---

**% pyspark**

# We can access the stages of the pipeline
# Here we access the linear regression stage
linRegModel = pipeline_model.stages[-1]
type(linRegModel)

 Took 0 sec. Last updated by simonstudent65 at May 08 2018, 11:24 AM.

---

**Summarize model training**

```
% pyspark
print("RMSE: ", format(linRegModel.summary.rootMeanSquaredError))
print("R2: ", format(linRegModel.summary.r2))
print("Model: Y = \{X + \}.
linRegModel.summary.residuals.show()
```

---

**SQL**

```
SELECT * FROM clustered_view
WHERE sex IS NOT NULL
```

---

![Graph showing grouped and stacked data with male and female categories.](image_url)