**UTC - Intro to Hadoop**

Course:

* Advanced Queries & Business Reports
* Dr. Beni Asllani
* Presented by Joe France
  + @joefrance
  + [joefrance@numerigy.com](mailto:joefrance@numerigy.com)

**Lesson Plan**

The lesson will cover two technologies:

* Hadoop 5:45 PM to 7:15 PM - with break(s)
* GitHub 7:15 PM to 7:45 PM

Intro to Hadoop

* High Level Description of Hadoop Ecosystem
  + Created by Doug Cutting
  + Named for his son's toy elephant
  + Scalable, batch processing with reliable, redundant storage.
  + Built on two fundamental components
    - HDFS (Hadoop Distributed File System)
      * Based on concepts from Google’s GFS paper:
        + [The Google File System](http://static.googleusercontent.com/media/research.google.com/en/us/archive/gfs-sosp2003.pdf)
    - MapReduce framework
      * Based on 2004 paper from Google on MapReduce:
        + [MapReduce: Simplified Data Processing on Large Clusters](http://static.googleusercontent.com/media/research.google.com/en/us/archive/mapreduce-osdi04.pdf)
      * Java-based (starting with v1)
      * Streaming, allows non-Java coding (starting with MRv2/YARN)
* Examples of previous work and planned work
  + Baseball streaming
    - Listener stats
  + Handyman service
    - Error log analysis
    - Job and Craftsman skills match
    - Potential for marketing/zip code analysis
  + Ceiling fan manufacturer (future plans)
    - Failure analysis
    - Sentiment analysis of AS/400 text-based call logs from customer call-in
* Hadoop distributions
  + Cloudera
  + Hortonworks
  + HDInsight
  + Amazon Elastic MapReduce

Discussion

* HDFS
  + Default of 3 copies of data
  + Contains large block size to handle large numbers of files
  + Deep storage
  + Schema-on-read
* MapReduce
  + Java/JVM based paradigm
  + MRv2/YARN - Streaming, can use ruby, other languages
  + This structure brings the code to the data

MapReduce Example

* Review WeatherData Java code
* Run WeatherData example
  + Point out Map/Reduce percentage as job runs
* Discuss WeatherData example
* Q & A about WeatherData exampe

Overview of Hive and Pig

* Hive
  + High-level query atop MapReduce
  + Similar to SQL
  + HiveQL compatible with many SQL flavors
  + Compiles queries to MapReduce
  + Can see MapReduce progress on longer queries
  + Run Hive examples
* Pig
  + High-level "scripting" atop MapReduce
  + Run Pig examples
  + Can see MapReduce progress on longer queries

Local Big Data User Groups

* CHadoop
  + <http://meetup.com/CHadoop>
  + <https://github.com/CHadoop>
* Chattanooga Data Science R Users and Machine Learning
  + <http://www.meetup.com/Chattanooga-Data-Science-R-Users-and-Machine-Learning>

Installing Hadoop for Yourself

* Installing Hadoop on MS Windows, Mac OS X, and Ubuntu
* Download instructions and installation files from:
  + [CHadoop](https://github.com/CHadoop/InstallationGuide)

Intro to GitHub

* Brief history
  + Git is a source control system
  + Used by Linux creator Linus Torvalds and the kernel team
  + GitHub is a popular website frontend
  + Can create and share code in repositories (repo)
* Free Kindle tutorial
  + Ry's Git Tutorial
* Quick reference
  + git init
  + git status
  + git add
  + git revert
  + git rm
  + git commit -m "Comment about changes"
  + git log
  + git config ...
    - git config --global user.name ""
    - git config --global user.email
* Creating an account
* Installation
  + Mac
    - <https://mac.github.com/>
      * Choose to "Install Command Line Tools" during installation
  + Windows
    - <https://windows.github.com/>
* Adding a repo
  + The README.md file
* Alternatives to GitHub
  + <http://www.boxuk.com/blog/a-tale-of-three-git-systems/>
  + Gitorious
    - <https://gitorious.org/>
  + GitLab
    - <https://about.gitlab.com/>