Midwest Big Data Summer School 2016

Dr. Beth Plale

2:10 - 3:10pm Campanile Room Management, Access, and Use of Big and Complex Data. Part I Data Pipelines in e-Science

Key questions: What is a data pipeline? Why are they particularly important today? What role do they play in science?

Video: https://mix.office.com/watch/1tju6hnnthxvw

What is a data pipeline? Data rarely instantly show up ready to use in whatever exploratory purpose a science researcher may have in mind. Data from creation to use undergoes numerous steps, some of which are end products in themselves. This session discusses data lifecycle, data pipeline, e-Science, cyberinfrastructure, Big Oh notation, and data analysis.

Discussion Readings:

Jim Gray on eScience: A Transformed Scientific Method, Edited by Tony Hey, Stewart Tansley, and Kirstin Tolle, in *The Fourth Paradigm: Data Intensive Scientific Discovery*, Tony Hey, Stewart Tansley, and Kritsin Tolle eds., Microsoft Research, 2009, pp. xix – xxxiii. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf

Jim Gray's Fourth Paradigm and the Construction of the Scientific Record, Clifford Lynch, in *The Fourth Paradigm: Data Intensive Scientific Discovery*, Tony Hey, Stewart Tansley, and Kritsin Tolle eds., Microsoft Research, 2009, pp. 177-183. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part4_lynch.pdf

Understanding execution time complexity: the Selection Sort versus the Heap Sort

http://en.wikipedia.org/wiki/Selection_sort

http://nova.umuc.edu/~jarc/idsv/lesson3.html

3:40 - 5:30pm Campanile Room Management, Access, and Use of Big and Complex Data. Part II Pipelines in Business

Key questions: how does business view the data pipeline? What is the role of cloud computing in the business view of data pipelines?

Video: https://mix.office.com/watch/1p4ed0ozhea8x

This session introduces the business perspective of data pipelines. It draws inspiration from a 2011 talk by Wernert Vogels *"Data Without Limits".* Vogels is CTO of Amazon, and in this nice 2011 talk discusses data pipelines in context of business computing. He argues that cloud computing is core to a business model "without limits". The pipeline he proposes is: collect | store | organize | analyze | share.

Vogels talks about mapreduce extensively during his discussion of analysis. If you're not familiar with MapReduce, a decent primer on MapReduce (Hadoop really; MapReduce is built into the open source Hadoop tool) can be found here: http://readwrite.com/2013/05/23/hadoop-what-it-is-and-how-it-works

Discussion Readings:

Three best practices for building successful data pipelines, Michael Li, Sep 2015. http://radar.oreilly.com/2015/09/three-best-practices-for-building-successful-data-pipelines.html

> "At The Data Incubator, our team has trained more than 100 talented Ph.D. data science fellows who are now data scientists at a wide range of companies, including Capital One, the New York Times, AIG, and Palantir. We commonly hear from Data Incubator alumni and hiring managers that one of their biggest challenges is also implementing their own ETL pipelines. Drawn from their experiences and my own, I've identified three key areas that are often overlooked in data pipelines, and those are making your analysis: Reproducible, Consistent, and Productionizable"

Examples of Data Pipelines You Can Build Today, Robert Dempsey Blog, Nov 2015. http://robertwdempsey.com/data-pipeline-examples/

> Over the past few months I started hearing the term "data pipeline" more and more at the local data meetups. Curious as to just what that meant, I looked it up. In this post I'm going to tell you what I found, and more importantly provide real-world examples of data pipelines you can use for your data projects.