Lecture 06 - Introduction to Workflows
Before getting into today’s lecture, let us look at what we have done so far.

Learning Objectives in Syllabus

- Describe clearly the basic concepts of distributed systems and know how to use the Cloud to scale the setup for distributed systems.
- Implement data collection modules to collect data.
- Integrate data storage platforms through calling APIs.
Before getting into today’s lecture, let us look at what have we done so far.

Learning Objectives in Syllabus

- Build distributed workflow driven approach to process big data using the Cloud.
- Familiarize in using scalable techniques on existing distributed systems in the NoSQL world.
- Understand and present scholarly work in the area of big data and cloud computing.
- Develop stand-alone big data application using a NoSQL-based distributed systems
Grand idea behind these discussions

- Take a problem and revolve our discussion around the problem and realize how we could use distributed system to tackle the problem.

- Problem 1: how do you collect large amount of data from web (IMPORTANT Problem). Can you collect it in one machine? NOOOOOOOO...... WHY?
Grand idea behind these discussions

how did we solve the problem?

- collect the data from web? Jsoup
- ability to use multiple machines? AWS Cloud, VM Provisioner, Jsch Library.
- store and process intermediate datasets efficiently to perform efficient computations? MySQL databases and MySQL queries.
- store initial and final datasets, which is huge? outsource to external data storage platforms, Dropbox.
What is the next problem that we are going to discuss?

Problem 2: how do you develop a framework which is usable by novice users and at the same time scalable to perform large scale computations fastly? WORKFLOWS !!

- Easily package computations into prebuilt tasks that can be used by other users.
- Allow computations built in different programming languages and different execution environments to be interoperable.
- Easily represent large scale research problem into different tasks and create a graphical representation of the problem.
- Automatic scaling feature to execute the computation block in a time and cost efficient manner.
Business Workflow

Goal is to reduce human resources (and other costs) and increase revenue.

Control flow oriented with the workflow orchestration performed to setup the logical flow in order to solve the given business problem.
Backgrounds of Workflows

Scientific Workflow

Lin et al. [a1] Architecture

- Goal is to reduce both human and computation costs and hence accelerate the innovation and discovery process.
- Data flow oriented with the workflow orchestration performed to setup a series of computation methods in order to solve a given scientific problem.
Backgrounds of Workflows


- One of the most cited paper: 136 citations
Big Data

- Big Data is the amount of data just beyond technology’s capability to store, manage and process efficiently.

“Ah, but a man’s reach should exceed his grasp, Or what’s a heaven for?”
- Robert Browning
Where is Big Data?
Where is Big Data?

Energy Efficient Homes

Healthcare Personalization

Particles Discovery

Pacemaker Monitoring
Importance of Big Data

“I expect big data to be important for a long time to come.”

Michael Stonebraker
The need for distributed system?

- Build virtually any distributed system that:
  1. support any workload regardless of volume, velocity, and variety of data.
  2. with 50+ services and hundreds of features added every year, AWS Cloud provides everything you need to collect, store, process, analyze, and visualize big data on the cloud.
  3. high scalability (unbound number of machines to process compute and data intensive workload).
Big Data Workflows

- Big data workflows are the next generation of scientific workflows to address the challenges of big data analytics, including volume, velocity, and variety.
- We are now in the big data era.
- Big data is important and challenging.
- One way to address such challenge is using big data workflow, which is "The computerized modeling and automation of a process consisting of a set of computational tasks and their data interdependencies to process and analyze data of ever increasing in scale, complexity, and rate of acquisition."
Big Data Workflows

Scientific Domains

Oceanography  Astronomy  Biology  Neurology

Montage Big Data Workflow

Dr. Aravind Mohan  (Allegheny College)
Montage Workflow

- Driven by the mosaic images from the sky
Montage Workflow

- Running the workflow produces fits and jpeg mosaics for each band, as well as a combined color one:

![Montage Image]
Interested to see Montage code in Python?

- checkout git page:
  
  https://github.com/pegasus-isimontage-workflow-v2
DRAW Workflow

Driven by the OpenXC dataset
Interested to know more about DRAW workflow:

- checkout our paper:
Research articles accepted this Spring 2018

- Our paper titled "Big Data Workflows: A Reference Architecture and The Dataview System" in IEEE Services Transactions on Big Data (STBD) journal. This paper is published by collaboration Andrey Kashlev at Eastern Mich and Dr. Lu at Wayne State.

- Our paper titled "Scheduling Big Data Workflows in the Cloud under Deadline Constraints" accepted to be published at IEEE Big Data Application and Service. This paper is published by collaboration with Mahdi Ebrahimi at Lawrence Tech and Dr. Lu at Wayne State.
We will discuss specific challenges in workflows next time.
Questions