Big Data Open Source Software and Projects Syllabus

Data Science Curriculum
January 6 2016

Geoffrey Fox
gcf@indiana.edu
http://www.infomall.org
School of Informatics and Computing
Digital Science Center
Indiana University Bloomington
CLASS INTRODUCTION

Stress Programming Expertise
Python for scripting and Java
Class Overview I

- I590 (I524) Sections 12388 (online), 33365: Big Data & Open Source Software Projects (Offered as part of Data Science Curriculum) – Geoffrey Fox
  - Section 33366 is discussion section associated with residential class 33366

- This course studies software used in many commercial activities to study Big Data. The backdrop for course is the ~350 software subsystems illustrated at [http://hpc-abds.org/kaleidoscope/](http://hpc-abds.org/kaleidoscope/). We will describe the software architecture represented by this collection which we term HPC-ABDS (High Performance Computing enhanced - Apache Big Data Stack). We will cover:
  - The cloud computing architecture underlying ABDS and contrast of this with HPC.
  - The software architecture with its different layers at [http://hpc-abds.org/kaleidoscope/](http://hpc-abds.org/kaleidoscope/) covering broad functionality and rationale for each layer.

- This discussion will be interleaved with a technology discussion describing OpenStack based virtual machines and give examples of HPC-ABDS. We will also cover DevOps tools centered around Ansible to manage Big Data software deployment

- Students will chose one or more other open source member of Kaleidoscope each and deploy as illustrated in class

- The final part of the course will be building a significant project using multiple HPC-ABDS subsystems combined with user code and data.

- Projects will be suggested or students can chose their own but they must use HPC-ABDS software
Class Overview II
(Technology)

• Technology Prerequisites
• Elementary knowledge in a scripting language needed (if not available this can be acquired as part of this course)
• Basic knowledge of Python desirable (if not available this can be acquired as part of this course)
• Ability to (learn to) use the Linux/Unix command shell (we will have lesson on this)
• Basic understanding on how to install packages and programs on Linux (we will have a lesson on this)
• Use of Java to build new and modify big data modules
Class Overview III (Technology)

- Then as part of the technology component of the course you will get experience in

- **DevOps**: "software deployment automation"
- Linux command shell
- Elementary usage of ssh
- Use of Github to store software packages and documentation
- The reproducible installation of sophisticated platforms on virtual clusters.
  - This is facilitated either by scripts developed in Python, Openstack Heat, or a DevOps framework such as Ansible (preferred), Chef, or Puppet.
  - Which framework is chosen will depend on the experience level of the student.
- You will learn utility of the key parts of Big Data Stack
Course Construction I

• Class Materials
  – Combination with 6 Units of Theory and 11 Units of Technology
  – 1 Unit (26 lessons and 8.5 hours) on details of HPC-ABDS optional
  – Via OpenEdx course Website

• Grading
  – Assignments (40%)
  – Project (60%)

• Online section Discussions (Adobe Connect)
  – Will be scheduled (30mins – 1 hour)
  – Residential section has a discussion session once a week

• Class policy
  – No penalty for late submission
  – Software project Only (No Paper accepted)
Course Construction II

- Instructor
  - Dr. Geoffrey C. Fox
- Lead AI
  - Hyungro Lee (lee212@Indiana.edu)
- ALWAYS send email to BDOSSPcoursehelp@googlegroups.com
  - Do not send it to individuals

<table>
<thead>
<tr>
<th>Topics</th>
<th>Web site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignment Release, Grades</td>
<td>IU Canvas <a href="https://canvas.iu.edu/">https://canvas.iu.edu/</a></td>
</tr>
<tr>
<td>Lecture Videos, Resources</td>
<td>OpenEdx Course Website <a href="http://datascience.scholargrid.org/">http://datascience.scholargrid.org/</a></td>
</tr>
<tr>
<td>General Questions and Discussions</td>
<td>Slack Website <a href="http://bdossspring2016.slack.com/">http://bdossspring2016.slack.com/</a></td>
</tr>
<tr>
<td>Questions and Help (Course related, technical issues)</td>
<td>Course help email <a href="mailto:BDOSSPcoursehelp@googlegroups.com">BDOSSPcoursehelp@googlegroups.com</a></td>
</tr>
<tr>
<td>Code Submission</td>
<td>Github <a href="https://github.com/futuresystems-courses">https://github.com/futuresystems-courses</a></td>
</tr>
<tr>
<td>Week</td>
<td>Topic</td>
</tr>
<tr>
<td>--------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>Week 1</td>
<td>INTRODUCTION</td>
</tr>
<tr>
<td>Week 2</td>
<td>CORE TECHNOLOGIES</td>
</tr>
<tr>
<td>Week 3</td>
<td>HPC-ABDS</td>
</tr>
<tr>
<td>Week 4</td>
<td>CLOUD COMPUTING INFRASTRUCTURE</td>
</tr>
<tr>
<td>Week 5</td>
<td>BIG DATA APPLICATIONS</td>
</tr>
<tr>
<td>Week 6</td>
<td>DevOps TOOLS</td>
</tr>
<tr>
<td>Week 7</td>
<td>VIRTUAL CLUSTERS I</td>
</tr>
<tr>
<td>Week 8</td>
<td>NIST SOFTWARE</td>
</tr>
<tr>
<td>Week 9</td>
<td>VIRTUAL CLUSTERS II</td>
</tr>
</tbody>
</table>

Detailed content subject to change!
<table>
<thead>
<tr>
<th>Week</th>
<th>Topic</th>
<th>Unit</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 10</td>
<td>SPRING BREAK</td>
<td>Due: Final Project Proposal</td>
<td></td>
</tr>
<tr>
<td>Week 11</td>
<td>DATABASE</td>
<td>Lessons: Database (HBase, Cassandra, MongoDB)</td>
<td>Technology</td>
</tr>
<tr>
<td>Week 12</td>
<td>PRESENTATION WEEK</td>
<td>DEMO (3-5 min each team in 2 sessions; e.g. Tue, Thu)</td>
<td></td>
</tr>
<tr>
<td>Week 13</td>
<td>CONTAINER CLUSTERS</td>
<td>Lessons: Container clusters (docker swarm, rocket fleet, lxd) &amp; Schedulers (mesos, kubernetes)</td>
<td>Technology</td>
</tr>
<tr>
<td>Week 14</td>
<td>BIG DATA SOFTWARE I</td>
<td>Lessons: Docker, Vagrant, Storm, Pig, Hive</td>
<td>Technology</td>
</tr>
<tr>
<td>Week 15</td>
<td>BIG DATA SOFTWARE II</td>
<td>Lessons: File system, network</td>
<td>Technology</td>
</tr>
<tr>
<td>Week 16</td>
<td>FINAL PROJECT DUE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Detailed content subject to change!
CLASS CONTENTS

Overview and Section 1
Overall Structure

• 4 Sections:
  – 0: Course Overview
  – 1: Background of field divided into 6 units
  – 2: Detailed discussion of technology for using clouds and HPC-ABDS. This will use back end cloud (FutureSystems) and/or local machines
  – 3: Detailed discussion of members of HPC-ABDS (Optional)
## Kaleidoscope of (Apache) Big Data Stack (ABDS) and HPC Technologies

### Cross-Cutting Functions
- **1) Message and Data Protocols:** Kafka, RabbitMQ, Amazon SQS, Apache Qpid, Apache Kafka
- **2) Distributed Coordination:** ZK, Apache ZooKeeper, Google Spanner, Amazon DynamoDB, IBM DS
db
- **3) Security & Privacy:** Kerberos, SSL/TLS, OAuth, OpenID Connect
- **4) Monitoring:** Nagios, Prometheus, Ganglia, Zabbix, InfluxDB

### 21 layers Over 350 Software Packages

### May 15 2015
Section 1 Units 1 and 2

- **Unit 1: Introduction**
  - Lesson 1A: Digital Data and Cloud backdrop
  - Lesson 1B, C: Real World Big Data
  - Lesson 1D: Basic Trends and Jobs

- **Unit 2: Data Access Patterns and Introduction to using HPC-ABDS**
  - Lesson 2A: Introduction to HPC-ABDS and Typical Data Interaction Scenarios
  - Lesson 2B: Data Access Patterns: Science Examples
  - Lesson 2C: Remaining General Access Patterns
  - Lessons 2D-G: Using the HPC-ABDS Stack (summary of different layers)
Section 1 Units 3 and 4

• **Unit 3: Big Data Applications Structure: NIST Big Data Use Cases and their Features**
  – Lesson 3A: Introduction to NIST Big Data Initiative and Use Cases
  – Lesson 3B: Sources of Parallelism and Big Data Features
  – Lesson 3C: Streaming Big Data and More Features
  – Lesson 3D: Machine Learning and yet more Features

• **Unit 4: Aspects of Big Data Applications**
  – Lesson 4A: Other sources of Use Cases and Data System Examples
  – Lesson 4B: NoSQL, Global Machine Learning, Implementing Big Data Applications
  – Lesson 4C: Clouds and HPC Systems, Data Analysis and Simulations
Section 1 Units 5 and 6

• Unit 5: Big Data Applications and Generalizing their Structure
  – Lesson 5A: NIST Use Cases Reviewed and Image-based Applications I
  – Lesson 5B: Image-based Applications II
  – Lesson 5C: Internet of Things based Applications
  – Lesson 5D: Previous Application Classifications and Ogre Facets I
  – Lesson 5E: Ogre Facets II

• Unit 6: Aspects of Big Data Applications
  – Lesson 6A: Implementing the Software Stack
CLASS CONTENTS

Technology Section 2
Section 2: Practical Use of Clouds and HPC-ABDS III: Core Technologies

- **Unit Tech0: Collaboration Tools**
  - Overview and Introduction
  - Google: Google+, Hangout, Remote Desktop
  - Github
  - VirtualBox with the course image

- **Unit Tech1: System Access to FutureSystems**
  - ssh-keygen
  - Shell Access
  - Account Creation
  - Remote Login
  - Resources (Where to find information)
Section 2: Practical Use of Clouds and HPC-ABDS III: Core Technologies

• Unit Tech2: The Basics of OpenStack on FutureSystems
  – Introduction and Overview
  – OpenStack for Beginners – Compute Engine (Nova)
  – OpenStack for Beginners – Storage (Swift)
  – Network (Neutron)
  – Web Dashboard (Horizon)
  – Housekeeping (clean up shared resources)
Section 2: Practical Use of Clouds and HPC-ABDS IV: Comparison Of Cloud Computing Infrastructure

- **Unit –Tech3: Introduction to Cloud Computing**
- **Unit 3-1: IaaS - Infrastructure as a service frameworks**
  - OpenStack
  - Amazon Web Services (AWS)
  - Microsoft Azure
- **Optional Unit 3-2: Many Compute Engines Review**
  - Compute Engine (Nova)
  - Microsoft Azure

- **Unit –Tech4: Hybrid Clouds and Other Efforts**
- Cloudmesh
- Comet
Section 2: Practical Use of Clouds and HPC-ABDS V: DevOps Tools

• Unit-Tech5: IT Operations - Automation and Orchestration
• Ansible
• SaltStack
• Puppet
• Chef
• OpenStack Heat
• Ubuntu Juju
Section 2: Practical Use of Clouds and HPC-ABDS VI
Virtual Clusters I

- Unit-Tech6: Virtual Clusters I - First appearance of Hadoop
- Introduction and Overview
- Dynamic Deployment of Arbitrary X Software on Virtual Cluster
- Deploying Hadoop Cluster
- Hadoop Example: Word Count
- Deploying MongoDB Sharded Cluster
Traditional Compute Cluster with virtual clusters superimposed. These are either isolated or share nodes. SR-IOV support ensures that maintain high performance communication with Infiniband interconnect.

Lustre, GPFS, Object Store
This cluster has large disk on each node; Infiniband interconnect
Data copied from Object store to HDFS at start of session
Section 2: Practical Use of Clouds and HPC-ABDS VII
Virtual Clusters II

• Unit-Tech7: Virtual Clusters II
• Composite Cluster with Sub-Clusters
• Apache Hadoop YARN
• Apache ZooKeeper
• OpenMPI Virtual Cluster

• HPC Queuing System (optional)
Section 2: Practical Use of Clouds and HPC-ABDS VIII
HPC-ABDS Technologies

- Several Units Covering
- Database (HBase, Cassandra, MongoDB)
- Container clusters (docker swarm, rocket fleet, lxd) & Schedulers (mesos, kubernetes)
- Docker Basics
- VM Software – Vagrant
- Hadoop MRv2
- Apache Spark
- Apache Storm
- Apache Pig
- Apache Hive
CLASS CONTENTS

Section 3
Section 3

- This goes through 21 layers in HPC-ABDS software stack with roughly one slide each on the ~350 software systems in HPC-ABDS.
- This is divided into 26 lessons covering initially 4 crosscutting layers and then 17 stack layers in order starting with Layer 5 (IaaS) and ending with Layer 17 (orchestration).
21 Layers of HPC-ABDS

1) Message Protocols:
2) Distributed Coordination:
3) Security & Privacy:
4) Monitoring:
5) IaaS Management from HPC to hypervisors:
6) DevOps:
7) Interoperability:  
   Here are 21 functionalities. (including 11, 14, 15 subparts)
8) File systems:
9) Cluster Resource Management:
10) Data Transport:  
    4 Cross cutting at top
11) A) File management  
    17 in order of layered diagram starting at bottom  
    B) NoSQL  
    C) SQL
12) In-memory databases&caches / Object-relational mapping / Extraction Tools
13) Inter process communication Collectives, point-to-point, publish-subscribe, MPI:
14) A) Basic Programming model and runtime, SPMD, MapReduce:  
    B) Streaming:
15) A) High level Programming:  
    B) Frameworks
16) Application and Analytics:
17) Workflow-Orchestration: