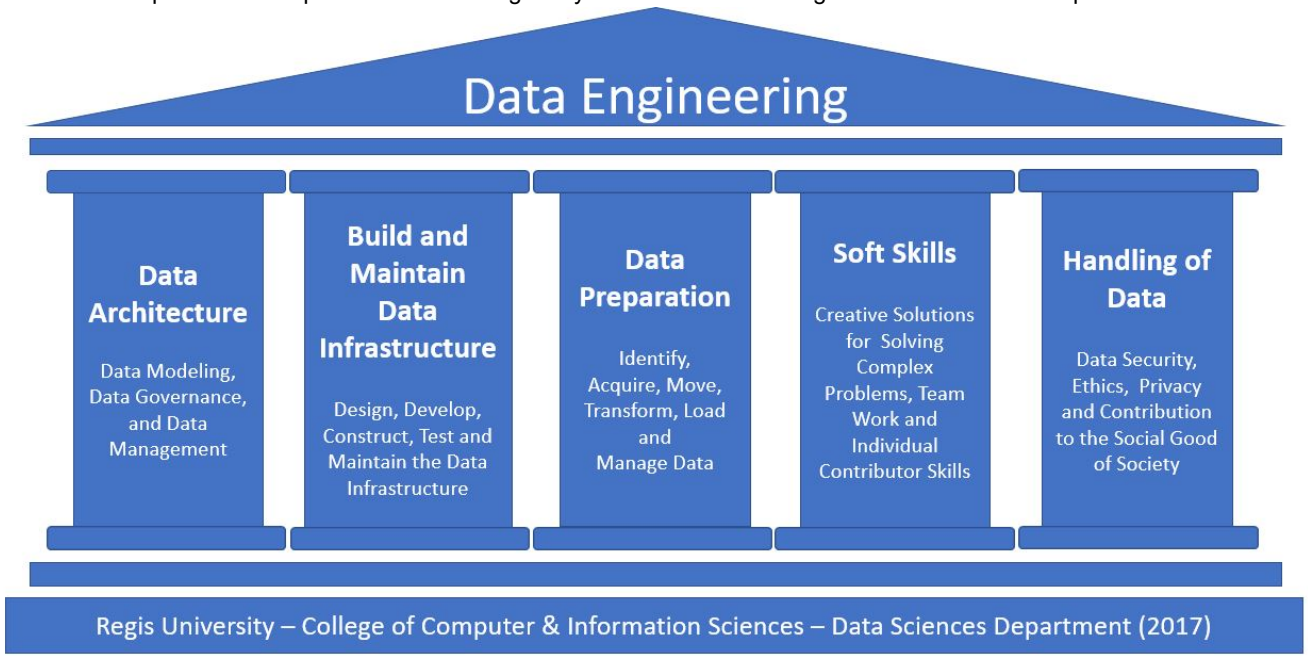
Overview and Outcomes

Week 1: Data Engineering and Hadoop - Single Node Setup in Pseudo-Distributed Mode

Overview

This week you will learn about Data Engineering and install Hadoop using a single node setup in the pseudo-distributed mode of operation.

Data Engineering is increasingly in demand with employers as the number of Data Scientists increase. Data Engineers are responsible for the creation and maintenance of the data infrastructure that facilitates the analysis of data by Data Scientists. While there is a certain amount of overlap when it comes to skills and responsibilities between Data Scientists and Data Engineers, these two positions are being increasingly separated into distinct roles. Data Scientists are much more focused on the interaction with the data infrastructure for analytical purposes, whereas Data Engineers are responsible for the development, construction, and maintenance and testing of the data infrastructure, such as databases and large-scale processing systems. Data Engineers are also responsible for the creation of data set processes used in data architecture, mining, acquisition, loading and verification. A reasonable analogy is that a Data Scientist is a painter that explains the data using analytical tools. A Data Engineer is considered the plumber that builds the data infrastructure which can then facilitate data analytics.



Cloud computing is having a huge impact on the the IT industry and Data Engineering. This week you will review an article about Big Data and Cloud Computing opportunities and challenges.

In regards to the Week 1 lab assignment (installing Hadoop), in this configuration of the Hadoop services, those related to HDFS (Hadoop Distributed File System) and YARN (Yet Another Resource Negotiator), MapReduce or Spark, run in individual Java Virtual Machine (JVM) processes on a single node, i.e., a single computer or a single virtual machine.

Hadoop supports three modes of operation. First, there is Standalone Operation where Hadoop is configured to run in single Java process on a single computer. This mode is useful for development of data processing software that will subsequently be deployed to a cluster. Second, there is the pseudo-distributed mode of operation. This is still a non-distributed mode of operation. However, each Hadoop daemon runs in a separate Java process. This mode of operation mimics the operation of a cluster on a single machine. Third, there is the Fully-Distributed Operation, which is a distributed mode of operation for clusters.

For more information see Hadoop: Setting up a Single Node Cluster. [http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html]

Learner Outcomes

Upon the completion of this week, you should be able to:

Evaluate Data Engineering as discipline of study.

Differentiate Data Engineering from Data Science.

Compare Data Engineer job responsibilities and the skills needed.

Appraise the Job growth in Data Engineering.

Evaluate Cloud computing opportunities and challenges.

Install and configure a Linux server in a virtual machine.

Install and configure the Java Software Development Kit on a Linux server.

Install and configure Hadoop in pseudo-distributed mode.

Perform tests to verify correct installation of all software.

Activity Details

Task: View this topic