Big Data Analytics with Python

Overview

The ever increasing growth of volume, complexity and speed in data drives the need for scalable data analytic algorithms and systems which are different from traditional frameworks. In order to survive and be productive in this environment, every data scientist needs to be able to handle large scale data whether it's when faced with basic analytics tasks or building machine learning models. In this course, we study algorithms and frameworks which will enable you to handle the challenges presented by large datasets. First, we will define Big Data and introduce you to the Hadoop ecosystem, which is a core suite of technologies for interacting with large scale data. The rest of the course will focus on giving you hands-on skills to use Apache Spark to interact with large scale data.

This course aims to teach you all the core concepts of Big Data and to equip you with the knowledge of the most popular tools and techniques for interacting with large scale datasets. We have chosen Apache Spark as a large scale data processing framework to use in this course because it is arguably the most widely used engine for working with Big Bata at the moment. We will work through real-life datasets, not toy datasets. At the end of this course, you will be confident to handle what can be considered as a “large dataset” and you will have the knowledge of most common technologies and platforms (e.g., Cloudera, Databricks, Hadoop, HDFS) to store and manage Big Data. Furthermore, you will gain hands-on skills on

how to process, query, analyze and build machine learning models when you have a large dataset using Apache Spark's MLlib.

Approach

This course takes a very practical approach to equip you with the most essential tools in the shortest possible time. Overall, the lessons start with a conceptual introduction to key topics, and then quickly jump into practical examples using real life datasets. The course ensures that you end up with knowledge and skills of how to store, manage, analyze and report on large datasets. Furthermore, the learners will learn how to build machine learning (ML) models when faced with Big Data. The key features of the course are as follows:

- Ensures you understand what we mean by big data and other commonly used keywords before getting into practical aspects
- Use real life large datasets to enable learners practice skills and concepts
- Teach learners state of the art when it comes to

Expected Learning Outcomes

At the end of this course, you will be able to:

- Demonstrate understanding of what Big Data means and explain why we need different sets of frameworks and algorithms to deal with it. Furthermore, you should be able to identify key characteristics of Big Data.

- Explain what functional programming means and how it is different from the procedural programming paradigm. Furthermore, students should be able to appreciate why functional programming is crucial for distributed data processing.
- Describe how Python supports functional programming
- Write simple functional programming style programs in Python
- Explain the role of Hadoop in Big Data processing systems. Furthermore, you should be able to define Hadoop, review its history, including its relationship with Big Data processing and its core components (e.g., HDFS)
- Understand Apache Spark architecture and its key components. You should further understand the different ways to deploy Spark programs.
- Write and deploy PySpark programs to perform core data wrangling tasks such as ingestion, transformations, cleaning and more
- Use the Spark MLlib to build and evaluate models for large scale datasets
- Use Pyspark on Azure using Databricks

Author Biography

Dr. Dunstan Matekenya is a consummate Data Scientist with over 10 years' experience in both traditional statistics and modern data science methods. Currently, he works as a *Data Scientist* in the Big Data team at the World Bank Group in Washington DC. Prior to joining the WBG, Dunstan completed his PhD at the University of Tokyo in 2016. His PhD research focused on use of machine learning methods to study human mobility from large scale mobile

phone datasets. Prior to this, he worked as a Statistician at the National Statistical Office in Malawi from 2007 up until 2017. While there he actively contributed to flagship projects such as the 2008 Malawi Population and Housing Census and also led the GIS unit. His passion includes contributing to modernization of official statistics in developing countries with use of alternative data sources such as mobile phone data as well improving capacity in data science.

Outline

In this course, we will cover 5 modules as follows: Big Data basics?; the hadoop ecosystem; introduction to apache spark; data wrangling with apache spark; machine learning with Apache Spark.

Big Data Basics

In this first module, the focus is to answer questions such as what we mean by big data, how it's being used in industry (applications), the different technologies associated with storage, management, analytics and how to do machine learning with big data.

- What's Big Data?
- Sources of Big Data
- What can we do with Big Data? (Applications)
- Tasks associated with big data: storage, analytics, ML
- Overview of Big Data platforms, technologies and tools
- How to work with Big Data-other big data processing packages in Python (e.g., Dask, pandas, vaex and datatable)

Functional Programming and Distributed Data Processing

This is a brief module to introduce students to the core knowledge in functional programming which is crucial for parallel data processing. We will cover the following topics

- What is functional programming (FP) including its advantages
- How FP is supported in Python
- Writing FP programs in Python
- Distributed data processing-parallel processing, scaling

Data Gathering from APIs and the Web

This is also a brief module which will cover best practices for harvesting data from APIs as well as web scraping. We will cover the practical aspects of web scraping and accessing APIs as follows:

- Data collection from APIs
- Web scraping

The Hadoop Ecosystem

In this module, we will focus on the fundamentals of Hadoop which is the foundational suite of open-source software tools for reliable, scalable, distributed computing technologies to work with large scale datasets. We will cover the architecture, components, ecosystem, practices, and commonly used applications including Distributed File System (HDFS), MapReduce, HIVE and HBase.

- Introduction to Hadoop and HDFS
- Hadoop ecosystem

- Distributed storage and processing
- Distributed programming models
- Mapreduce
- Hive, HBase

Introduction to Apache Spark

Here, we will start covering what's arguably the most popular Big Data processing engine: Apache Spark. When used with Python, we also call it PySpark. We will cover the architecture of Apache Spark and basics of distributed computing. We will dive into the data structures used by Spark: Resilient Distributed Datasets (RDDs), DataFrames, SparkSQL and more.

- The Spark advantage: Spark vs. Mapreduce
- Spark architecture and components
- Spark data structures (RDDs, Datasets, Dataframes, SparkSQL)
- How to run spark programs

Data Wrangling with Spark's Structured API

This will be a hands-on module focusing on how to perform typical data science tasks such as data ingestion, data cleaning, exploratory data analysis (EDA) with Apache Spark through the DataFrames API. In this module, we will focus on how to perform common data wrangling tasks with spark such as data ingestion, ETL, sampling, descriptive statistics generation, transformations and more.

- RDDs
- DataFrames and SparkSQL
- Working with Spark DataFrames

- Spark Application processes including, Jobs, Tasks, Stages, and cluster deploy modes.

Machine Learning with Apache Spark

In this module, we will learn about how to build machine learning models when faced with a big dataset. For this, we will focus on MLlib- which is Spark's machine learning (ML) library.

- Introduction to MLlib
- Featurization: feature extraction, transformation, dimensionality reduction, and selection
- Model building and evaluation

Bonus Module

The following modules will be covered subject to time availability

Assessments

In order to check if students are keeping up with the materials, the course will utilize the following for assessment:

- Weekly quizzes and coding exercises
- Project

Required Datasets

TBD

Prerequisite Python Knowledge

We expect that the learners will have at the minimum beginner level knowledge in Python and essential Python packages for data science.

Python basics

We expect that the learners will have at the minimum beginner level knowledge in Python and essential Python packages for data science.

- Data structures,
- Control flow
- Functions
- Scripts

If you need a refresher in Python, here is a good free course.

Python data science stack

The learners are expected to have familiarity with the essential packages used to perform data science tasks in Python. For instance, pandas, numpy, matplotlib, scikit-learn.

Hardware and Software Requirements

Cloud compute requirements

For some of the tasks in this course, it will be important to access cloud based platforms. Please use your student email address to create cloud accounts and attempt to get student free credits.

- **Microsoft Azure**-please follow this link to create a free account and access student credits
- **Databricks**- Follow this link to create Databricks account

Hardware requirements

For an optimal student experience, we recommend the following hardware configuration:

- OS: Windows 7 SP1 64-bit, Windows 8.1 64-bit or Windows 10 64-bit, Ubuntu Linux, or the latest version of OS X

- Processor: Intel Core i5 or equivalent
- Memory: 8GB RAM
- Storage: 50 GB available space (100 GB preferred)

Software requirements

You'll also need the following software installed in advance:

- **Browser:** Google Chrome/Mozilla Firefox Latest Version
- **IDE/Text editor:** Visual Studio Code/Sublime Text as IDE
- **Python:** Python 3.9+. Best to install with Anaconda
- **Python libraries:** If you install Python without Anaconda, ensure you install these packages: (Jupyter, Numpy, Pandas, Matplotlib, BeautifulSoup4, and so)
- **Hortonworks HDP:** TBD
- **Apache Spark/Pyspark:** Follow these instructions or any other to install pyspark.
- **Other big data packages:** Dask, vaex

Course Materials

All source code is publicly available on GitHub and fully referenced within the training material.

- Github repo: TBD
- Google Drive with course materials: TBD

Keywords

- Big data
- Hadoop
- Apache Spark
- Pyspark
- HDFS