Cheatsheets / **Introduction to Big Data with PySpark**

codecademy

Introduction to Big Data

HDFS Overview

One system for big data storage is called Hadoop Distributed File Storage (HDFS). In this system, a cluster of computing resources stores the data. This cluster consists of a *manager node*, which sends commands to the *worker nodes* that house the data.

MapReduce Overview

MapReduce is a framework that can be used to process large datasets stored in a Hadoop Distributed File System (HDFS) cluster. MapReduce consists of two main functions, *map* and *reduce*, which can perform complex operations over a distributed system.

MapReduce Process Overview

MapReduce works by sending commands from the *manager node* down to the numerous *worker nodes*, which process subsets of data in parallel. This speeds up processing when compared to traditional data processing frameworks.

codecademy

Big Data Definition

Big data is a term that refers to data that are too large and complex to be processed by normal computing capabilities. Describing data as “big” is relative to our modern computing power.

Big Data 3 Vs

Big Data can be characterized by what are known as the 3 Vs:

1. **Volume**: the size of big data is larger than the amount of available computing power.
2. **Velocity**: big data grows rapidly as it becomes faster, cheaper, and easier to collect automatically and continuously.
3. **Variety**: big data comes in a variety of formats, such as structured (data tables with rows and columns), semi-structured (think JSON files with nested data), and unstructured (audio, image, and video data).

Big Data and RAM

Big data analysis is limited by the amount of Random Access Memory (RAM) that the available computing resources have. Many big data systems will use a computing cluster to increase the amount of total RAM.

Print Share

codecademy