ILLINOIS DATA SCIENCE INITIATIVE

TECHNICAL REPORTS

PySpark on Python 3: Configuration and Package Mangement Guide

Author:
Benjamin Congdon

February 24, 2017

PySpark on Python 3: Configuration and Package Mangement Guide

Benjamin Congdon[1] **and The Big Dog**[2]

[1] *National Center For Supercomputing Applications (NCSA)*
[2] *Laboratory for Computation, Data, and Machine Learning*

Compiled February 24, 2017

PySpark requires requires Python 2.6 or later, and does not officially support Python 3. However, with only minor configuration changes, it is possible to successfully run a Spark cluster using Python 3 for both PySpark drivers and workers.

https://github.com/lcdm-uiuc

1. INTRODUCTION

2. ASSUMPTIONS

This technical report will make the assumption that you have already set up a Spark Cluster with a CentOS worker base image.

3. EXAMPLES OF ARTICLE COMPONENTS

The sections below show examples of different article components.

FUNDING INFORMATION

ACKNOWLEDGMENTS

SUPPLEMENTAL DOCUMENTS

REFERENCES