Syracuse University
School of Information Studies

Applications of Opensource in Big Data Management with Examples

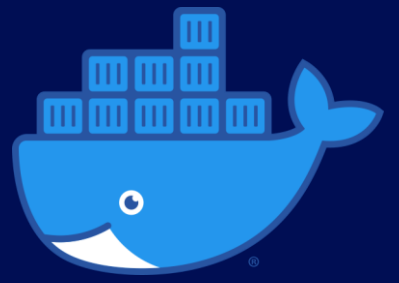
Michael Fudge
Professor of Practice
Syracuse University
CASL Workshop, December 2022

Code Walkthroughs and Demonstrations

1. Setting up your computer
2. Basic docker commands
3. Jupyter and Spark Introduction
4. Object Storage and HDFS
5. Spark Data Transformations
6. Spark SQL
7. MongoDb Basics
8. Accessing MongoDb from Spark
9. Neo4j Basics
10. Accessing Neo4j from Spark

Setting Up Your Computer

Basic Docker Commands



Jupyter and Spark Introduction

Jupyter Notes

- Access at http://localhost:8888
- Password is SU2orange!

Object Storage and HDFS

Object Storage / HDFS Notes

- The Minio S3 compatible storage http://localhost:9001
  - User name is **minio**
  - Password is **SU2orange!**
- HDFS is accessible at http://localhost:50070/

Spark Data Transformations

Two Part Video On Transformations

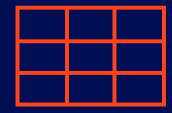
PART I

- File Formats
- Adding Schema
- 3 Ways to Reference a column.
- Column Transformations
- Row Transformations
- Sorting

PART II

- Aggregates
- Joins
- Unions
- Nested Columns
- Explain / DAG

Spark SQL



MongoDb Basics

MongoDb Notes

- The MongoDb Server is http://localhost:27017
- The Mongo Admin UI is http://localhost:8881
  - User name is **admin**
  - Password is **mongopw**

Accessing MongoDb from Spark

Neo4j Basics

Neo4j Notes

- The Neo4j Server is http://localhost:7687
- The Neo4j Admin UI is http://localhost:7474
  - No Username or Password

Accessing Neo4j from Spark

Syracuse University
School of Information Studies

Applications of Opensource in Big Data Management with Examples

Michael Fudge
mafudge@syr.edu
Professor of Practice
Syracuse University
CASL Workshop, December 2022