Massively Scalable EDM with Spark

Tristan Nixon
Institute for Intelligent Systems
University of Memphis
365 Innovation Drive
Memphis, TN, USA, 38152
t.nixon@memphis.edu

1. INTRODUCTION

The creation and availability of ever-larger datasets is motivating the development of new distributed technologies to store and process data across clusters of servers. Apache Spark has emerged as the new standard platform for developing highly scalable cluster computing applications. It offers a wide range of connectors to numerous databases and enterprise data management systems, an ever growing library of machine-learning algorithms and the ability to process streaming data in near-realtime. Developers can write their applications in Java, Scala, Python and R. Applications can be run locally (for easy development and testing), and deployed to dedicated clusters or on clusters leased from cloud-computing providers.

2. TUTORIAL

This day-long tutorial will provide a hands-on introduction to developing massively scalable machine learning and data mining applications with Spark. Participants will be expected to follow along with all examples on their own laptops throughout the tutorial, and to collaborate in small groups. All code used in the tutorial will either be taken from publicly available examples, or be available for download from the IEDMS github repository[1], and made available under a very liberal open source license. All examples will be designed to process a modestly sized sample of the KDD cup dataset available from the PSLC DataShop[2].

In advance of the day, participants will be given instructions on how to install and configure Spark and Scala on their laptops, so that they might arrive at the tutorial ready to begin. Throughout the tutorial, participants will be given exercises and problems to solve in small groups. This will give them experience with the material as it is presented and hands-on practice with structuring a distributed application in Spark.

2.1 Outline

The following material will be covered in the course of the tutorial:

- An overview and history of cluster computing and the development of map-reduce
- An example of a very simple map-reduce algorithm (distributed word-count) in Spark
- An introduction to the Spark runtime model, including:
    - Basic import and export operations
    - Resilient distributed datasets (RDDs)
    - RDD transformations and actions
    - How Spark optimizes the execution of distributed computation
- An overview to the different deployment options for Spark, including:
    - Launching and using the interactive spark command-line shell program
    - Running spark programs locally on a single machine
    - Launching a Spark cluster on Amazon Web Services
    - Submitting applications to remote clusters
- An introduction to Spark streaming
- An introduction to SparkSQL and working with DataFrames
    - How to load and manipulate an EDM dataset (KDD cup data)
    - Data representations needed to fit various EDM algorithms
- An introduction to Spark's Machine learning library MLib, including:
    - Transformers and Estimators
    - Chaining transformers into machine-learning pipelines
    - Examples of common EDM algorithms in Spark:
        - IRT algorithms using logistic regression (AFM, PFM, IFM)
        - BKT parameter fitting: (brute-force, HMMs)

Any remaining time will be devoted to discussing potential applications that participants may have in mind for their own data or projects.

[1] https://github.com/IEDMS/spark-tutorial

[2] https://pslcdatashop.web.cmu.edu/KDDCup/