**605.788 Big Data Processing Using Hadoop**

**Presentation Summary– Apache Spark**

**Guan Yue Wang**

**JHU ID: gwang39**

**December 10th, 2018**

**Presentation Topic: Apache Spark**

Apache Spark is a unified analytics engine for large-scale data processing. It supports interaction with various languages such as Java, Python, and Scala as well as powers various libraries for SQL, streaming, machine learning, and complex analytics. Furthermore, Spark can be ran on standalone cluster, Hadoop YARN, EC2, etc. Consequently, Spark becomes one of the most popular tools from the Hadoop ecosystem in the field of machine learning and data science.

In my presentation, I went through the history of Apache spark, from starting as a research project in UC Berkeley to becoming one of the most active open source big data project. After that, I provided an overview of the Spark, including definitions, advantages, running environment. Furthermore, I also talked about the major component of the Spark framework such as Spark SQL, Spark Streaming, MLlib, and Graph X. In addition to conceptual contents, I did a short video demonstration the installation of Spark as well as some spark shell command line examples.

Overall, the demonstration section went really well, I was able to get Spark working follow the similar instructions in the Hive and Pig lecture with some documentations in the Spark website. The demonstration in the spark-shell gives me a better idea of the Scala language as well as the performance of the jobs in Scala. Regarding the presentation content, I’m only able to cover some brief history, definitions, and framework components. If I have more time, I would like to dig deeper into more use cases, real life examples of Spark. Furthurmore, I want to research more on the major project inside the Spark framework.

**Appendix – Spark installation and command line demonstration contents**

Installation prereq: JDK is installed

1 - Download Apache Spark from Apache website

* Website: <https://spark.apache.org/downloads.html>
* Spark release:2.4.0
* Spark package type: Pre-built for Apache Spark 2.6

2 - Go to installation location and extract Spark project

* Go to expected install location
* $ tar -xvf ~/Downloads/spark-2.4.0-bin-hadoop2.6.tgz

3 - Add Spark to Path (SPARK\_HOME and PATH)

* Set SPARK\_HOME environment variable in your .bash\_rc
  + export SPARK\_HOME=~/spark-2.4.0-bin-hadoop2.6
* Add $SPARK\_HOME/bin to your PATH
  + export PATH=$PATH:$SPARK\_HOME/bin

4 - Launch Shell

* Spark-shell

5 - Spark Command Line Demonstration

// “sc” is a “Spark context” – this transforms the file into an RDD

val textFile = sc.textFile("AliceInWonderland.txt")

// Return number of items (lines) in this RDD; count() is an action

textFile.count()

// Filtering.  Filter is a tranform.  By itself this does no real work

val linesWithMy = textFile.filter(line => line.contains("Alice"))

// Chaining – how many lines contain “Spark”?  count() is an action.

textFile.filter(line => line.contains("Alice")).count()

// Word count – traditional map-reduce.  collect() is an action

val wordCounts = textFile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey((a, b) => a + b)

wordCounts.collect.foreach(println)