Apache Spark and Apache SystemML for Survival Analysis on Stackexchange

Mateo Álvarez Calvo

May 13, 2017

Introduction & main goals

Many studies have been done over the Stackexchange community [such as], one of the biggest Q&A sites in the world. The present is yet another study over the data of the famous site, but in this case, the study has two particularities, the use of Apache Spark with the library of Apache SystemML for the processing in a parallel environment, and the use of Survival Analysis to analyze the impact of the variables in the time an answer is accepted for each question, the "survival of each question" in the community.

1 Main technologies

As one of the biggest Q&A communities, Stackexchange has large amount of data of each interaction. Stackexchange is separated in several communities, regarding diferent topics. These communities can be small, as [] or really big, as Stackoverflow, the developers community. This particularity makes necessary the use technologies prepared to process large amounts of data, in the later case.

For this purpose, Apache Spark has been chosen, the latest distributed open-source processing technology, to parallelize the operations on the data.

Spark ML is the machine learning library of spark, which contains lots of algorithms. It also includes some Survival Analysis algorithms, but just for parametric modelling. This gives an excuse to use the recently adopted by the Apache Foundation SystemML, a machine learning library developed by IBM, which has non, semi and parametric algorithms for survival analysis.

1.1 Apache Spark

Apache Spark is a distributed processing technology developed in Scala by Databricks that represents the next step of Apache Hadoop, including the best parts of it, such as the Hadoop File System, but under a complete new paradigm that allows operations diferent from the famous map-reduce, using RAM as storage for results rather than

writting to disk and lazy and optimized execution of tasks, to mention some of the main features.

1.2 Apache SystemML

Recently included in the Apache Foundation Incubating program, Apache SystemML is a machine learning library that works over distributed frameworks, Spark or Hadoop, written in a self made language, Distributed Machine Learning (dml) and with a python, scala or R like syntaxis API.

This library provides many distributed implementations of important algorithms as well as a syntaxis to create new algorithms in a distributed mode.

2 The Stackexchange data

Stackexchange facilitates raw partially processed? data from the communities database every once in a while or every three months? for data scientists and people in general to download and analyze. The data is available in a torrent file and each package has about 35 - 40 GB of compressed information.

This compressed file has data from different communities for a certain period of time. In our case, the analysis is done over Scyfy community, wich is a median size community for science-fiction Q&A.

3 Main Objectives

For this study several objectives have been proposed:

- Use Spark to make the data cleaning to create a script for further research on the Stackexchange site.

- Verify SystemML integration with Spark for further research and scalability.

- Use SystemML survival analysis algorithms to analize Stackexchange's data and obtain conclussions on the main variables affecting the fast resolution of questions.

Technologies
The downloaded data from Stackexchange for the analysis [weights?] 40 GB, wich is enough ammount [size] to consider distributed processing.

4 Apache Spark

The latest distributed processing technology for data processing is Apache Spark, born in [2015] in [Ucla?], wich is the precursor of the well known processing framework Apache Hadoop, with all it's ecosystem.

Comparing to Hadoop, Spark is able to speedup calculations, with a higher variety of operations, not limited to Map-Reduce, preserving the fault tolerance and [].