Published on *CERN openlab* (http://test-static-05.web.cern.ch)



# Evaluation of Apache Spark as Analytics as framework for CERN's Big Data Analytics [1]

**Date published:**
Tuesday, 1 September, 2015
**Document type:**
Summer student report
**Author(s):**
S. Ganju
I present an evaluation of Apache Spark for streamlining predictive models which use information from CMS data -services. The models are used to predict which datasets will become popular over time. This will help to replicate the datasets that are most heavily accessed, which will improve the efficiency of physics analysis in CMS. The evaluation will cover implementation on Apache Spark framework to evaluate quality of individual models, make ensembles and choose best predictive model(s) for new set of data. The task in this project is to predict popular datasets. Finding the popular datasets is helpful in a two-fold way. Firstly, it helps in providing expeditious access to datasets that might be required and secondly, it helps in finding which might become the ?hot topics? in high energy physics. It is also necessary to define what a popular dataset is. Based on the data collected, some parameters such as nusers, naccesses and tot_cpu can be said to define popularity because the curve between all the parameters and popularity is mostly dependent on them. To find the numerical value of the threshold limit beyond which a dataset is termed popular, a graph is plotted. This graph is plotted on the log scale so that all values can be plotted within the region represented by the graph. After calculating the threshold values, transformation into a classification problem is done. Now, a rolling forecast is performed. This helps us to predict binary popularity values for each week. Each week?s data is added to the existing data and a new model is created. This follows the notion, more data leads to better data analysis. Prediction can be done in various ways following implementation of several machine learning algorithms, mainly, Naive Bayes, Stochastic Gradient Descent and Random Forest. Their models are then combined into an ensemble to check which algorithm offers the best true positive, true negative, false positive or false negative value. This project also includes plotting the results obtained against the time scale to get a notion of how accuracy scores change with each week. These include sensitivity, specificity, precision, and recall and fallout rate against time scale.

**Report on ZENODO:**
Document on ZENODO [2]

- Visit Us
- RSS Feeds




---



**Links**
[1] http://test-static-05.web.cern.ch/publications/technical_documents/evaluation-apache-spark-analytics-framework-cerns-big-data
[2] http://zenodo.org/record/31861