J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

Leveraging Big Data for Spark Deep Learning to Predict Rating

Monika Mishra[1] Mingoo Kang[2] Jongwook Woo[1*]

ABSTRACT

The paper is to build recommendation systems leveraging Deep Learning and Big Data platform, Spark to predict item ratings of the Amazon e-commerce site. Recommendation system in e-commerce has become extremely popular in recent years and it is very important for both customers and sellers in daily life. It means providing the users with products and services they are interested in. Therecommendation systems need users' previous shopping activities and digital footprints to make best recommendation purpose for next item shopping. We developed the recommendation models in Amazon AWS Cloud services to predict the users' ratings for the items with the massive data set of Amazon customer reviews. We also present Big Data architecture to afford the large scale data set for storing and computation. And, we adopted deep learning for machine learning community as it is known that it has higher accuracy for the massive data set. In the end, a comparative conclusion in terms of the accuracy as well as the performance is illustrated with the Deep Learning architecture with Spark ML and the traditional Big Data architecture, Spark ML alone.

☞ keyword : Big Data, Deep Learning, Spark, Analytics Zoo, Amazon EMR, Machine Learning, Recommendation

1. Introduction

Amazon has been popullar in the e-commerce markets and expands to be a leader at the IT industry providing many services in the cloud computing. Many e-commerce companies like Amazon have made the online shopping much more accessible, easy and covenient. Although e-commerce market continues to grow rapidly, it lacks in giving buyer the experience of the brick-and-mortar stores, seeing and trying out the products before committing to purchase. [1-5]. This is where having best-in-class consumer-generated ratings is more important than ever.[6].

Recommendation is a popular method of predictive analysis that provides close recommendations based on user information such as history of purchases, clicks, and ratings. Google and Amazon use these methods to display a list of recommended items for their users, based on the information from their past actions. The items are filtered in the e-commerce database by predicting how a user might rate them. This helps in connecting the users with the right items. This technique is useful in two ways: If there is a massive database of items, the user may or may not find item relevant to his choices. Also, by recommending the relevant item, one can increase consumption and get more users. [7]

Recently, not only the quantity of data has exponentially increased, the type and nature of this data have also dramatically evolved. All this has made the need for an efficient and effective Big Data processing platform a necessity. Big Data is defined as non-expensive frameworks, mostly on distributed parallel computing systems, which can store a large-scale data and process it in parallel. A large-scale massive data set means a data of giga-bytes or more, which cannot be processed or stored using traditional computing systems [8]. Hadoop and Spark are the popular Big Data platforms and lately NoSQL DB and search engine such as Elasticsearch are regarded as Big Data frameworks.

Spark has been adopted to Big Data platform as an efficient in-memory distributed computing engine. It is designed for fast computation. Furthermore, its Machine Leaning library, Spark ML, allows developers to use Spark for data processing at scale while building machine learning

[1] Dept. of Computer Information Systems, California State University Los Angeles, California, United States
[2] Dept. of Computer Information Systems, Hanshin University, Korea
* Corresponding author: jwoo5@exchange.calstatela.edu
[Received 01 June 2020, Reviewed 30 September 2020, Accepted 27 Oct 2020]
☆ This work was supported by the AWS Cloud Credits for Research Program.

http://dx.doi.org/10.7472/jksii.2020.21.6.33

models. Itis becoming the de-facto platform for building machine learning algorithms and applications.

Deep Learning has received highlights past several years, mostly after Google shares its TensorFlow library and NVidia's GPU become non-expensive as multi-core parallel computing processor in a single chip. It is a branch of machine learning that uses algorithms to model high-level abstractions in data. These methods are based on artificial neural network topologies and can scale with larger data sets.

As Deep Learning grows popular, it has had many different architectures to integrate multi-core GPU systemsto distributed systems, which are TensorFlowOnSpark by Yahoo, DeepLearning Pipeline for Apache Spark by Databricks, BigDL/Analytics Zoo by Intel, DL4J by Skymind, Distributed DeepLearning with Keras & Spark by Elephas. This paper adopts Analytics Zoo and Amazon EMR to execute the models using Spark ML and Analytics Zoo.

We adopt the BigDL architecture by Intel, which is integrated into Analytics Zoo using AWS EMR Big Data cloud service. It is a platform to integrate the scalable Spark computing engine and deep learning algorithms, which can leverage the strength of distributed computing systems and multi-core parallel computing systems.

The paper presentsRelated Work at the section 2. The section 3 illustrates the dataset and background. The section 4 illustrate the Big Data Deep Learning architecture. The sections 5 relates to experiment system and results. The last section is the conclusion part.

2. Related Work

Monika et al [9] presenteddescriptive and predictive analysis using Big Data on Cloud Computing by building recommendation models using traditional system.

Bhavesh [10] classified product review of Amazon to positive and negative in the traditional sequential systems. He performed sentimental analysis for one of the baby products. And, he mainly concentrated on one product category: baby.

Max [11] only performed descriptive analysis using Sparklyr platform.

Predictive analysis models are presented in our paper to predict users' ratings using several models of Big Data Spark ML. Besides, predictive analysis is done by leveraging Big Data Unified Analytics Platform and adopting the integrated systems of Deep Learning and Spark to compare the performance and accuracy.

3. Backgrounds

The big data analysis and prediction is mostly based on Hadoop and Spark. Big Data is defined as non-expensive frameworks, mostly on distributed parallel computing systems, which can store a large-scale data and process it in parallel [8]. Hadoop & Spark clusters have mostly selected as the solutions for Big Data analysis and prediction in the world.

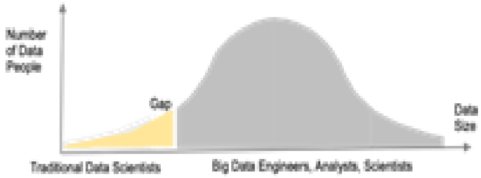
In the paper, the data set is uploaded to Hadoop Distributed File systems and Amazon AWS S3 and transformed to be analyzed using Spark and Analytics Zoo. The data set is acquired from the amazonaws site [12]. The data has details about the products reviewed on Amazon site between 2005 and 2015 in the United States. The Amazon product review datasetcontains 15 attributes and has about 6.93 million records. The total file size is 3.63 GB, which is too big to be handled by the traditional systems in terms of storage and computation time for the prediction..

Our models for predictive analysis is to implement the rating prediction in both Spark ML and Analytics Zoo. Each metric of the models is evaluated in terms of accuracy, MAE (*Mean Absolute Error*) and performance, *Time*.

4. Big Data Deep Learning Architecture

Spark is in-memory distributed computing engine with linear scalibilty and it has been popular as integrated to Big Data plaforms such as Hadoop and NoSQL DB.

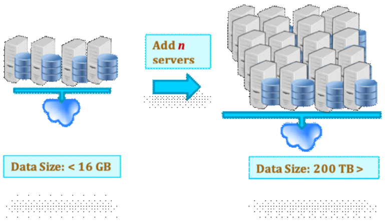
Apache Spark is a distributed parallel and cluster computing systems and supports MLlib machine learning APIs. Spark computing engine in the paper uses Hadoop Big Data systems for its HDFS file systems and YARN resource management, which provides a unified data analytics (UDA) platform. UDA platform is an integrated system for data storage, analysis, and prediction, especially for massive datasets.



(Figure 1) Gap between the Deep Learning & Data Scientists and the professionals in Big Data

People have talked about deep learningand machine learning lately. However, it is mostly for the traditional data science using deep learning and machine learning in Python and Rfor the small dataset, which can process up to Mega-Bytes of data. Figure 1 shows the gap (underrepresented number) between the deep learningand people in Big Data. Industry needs more data people as the data grows exponentially, which includes Big Data engineers, Analysts, and Scientists. When there is data higher than Giga-bytes, the deep learning method cannot efficiently implement models with the massive data set because of its storage limitation – actually, it is not possible. However, the Big Data platforms can store and compute the massive dataset, which is higher than Giga-bytes, for data engineering, analysis, and even for deeplearning.

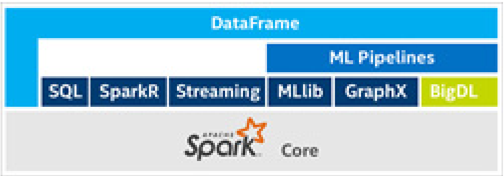
Figure 2 shows that the Spark Hadoop cluster can grow by adding more servers while collecting more data. For example, the dataset increases from the 16 Gigabytes of data to 200 Terabyte of data. Then, the systems require additional storage and computing engines by adding more servers, and it still works well with the TBs of data by well-supported resource management. The cluster is called Unified Data Analytics (UDA) platform,which provides data engineering, data analysis and prediction with massive data set.



(Figure 2) Spark Hadoop Big Data UDA platform and the scalability

Purush et al. showed that the Big Data architecture is linearly scalable so that if the architecture can store and compute several Gigabytes of data, it should work with hundreds of Gigabytes of dataset and more [15].

As Deep Learning grows exponentially on single chip with multi-core computing power, it has had many different architectures to integrate Spark and Deep Learning: DeepLearning Pipeline for Apache Spark by Databricks, TensorFlowOnSpark by Yahoo, BigDL/Analytics Zoo by Intel, DL4J by Skymind, Distributed DeepLearning with Keras & Spark by Elephas.



(Figure 2) BigDL Architecture [13]

Figure 2 shows the BigDL architecture by Intel, which is integrated intoAnalytics Zoo. In the paper, Analytics Zoo is adopted as a experimental platform using AWS EMR cloud service. Analytics Zoo supports Deep Learning models, Keras and Tensor Flow as well as BigDL and can run on Hadoop/Spark cluster.

5. Experimental System and Results

5.1 Details of Data Set

We acquired the data set for the paper from the Amazon AWS site ("S3.amazonaws.com", n.d). The data set presents the details about the products reviewed on Amazon sites between 2005 and 2015 in the United States. The analyzed Amazon product review dataset contains 15 attributes and has about 6.93 million records. The total file size is 3.63

GB, which is vast so that the traditional systems cannot afford or takes a long time to compute the prediction. The format of the file is Tab Separated Values (TSV).

The list of columns and Table 1 describes the details of the columns. The data set is of both quantitative and qualitative nature. The data set size is 3.63GB, having 6.93 million records, which describes the property of quantitative. The opinions present in *review_headline* and *review_body* columns by the Amazon customers represent the qualitative nature of the dataset. It presents their experiences regarding the products on the Amazon website.

〈Table 1〉 Column name and with column details of the dataset

Column Name	Column Details
marketplace	Country code - US
customer_id	The ID of the customer
review_id	The unique ID of the review.
product_id	The unique Product ID the review pertains to
product_parent	Random identifier used to aggregate reviews for same product
product_title	Title of the product
product_category	Broad product category that can be used to group reviews
star_rating	The 1 - 5 star rating of the review
helpful_votes	Number of helpful votes.
total_votes	Number of total votes the review received
vine	Review was written as part of the Vine program
verified_purchase	The review is on a verified purchase
review_headline	The title of the review
review_body	The review text
review_date	The date the review was written

5.2 Cloud Computing Systems

We used big data EMR services and storage supported by Amazon AWS. We developed our models on Hadoop and Spark system components in AWS EMR. Amazon Elastic MapReduce (EMR) is a tool for big data processing and analysis. Amazon EMR offers the expandable low-configuration service as an easier alternative to running in-house cluster computing.The EMR H/W instances of the Hadoop/Spark cluster details are given below:

Number of Nodes: 3
EMR Instances: r3.2xlarge
CPU: 8 vCPU
CPU speed: 3.1 GHz
Memory size: 183 GB (= 61 GB x 3)
Storage: 960 GB (= 2 x 160GB x 3)

5.3 Prediction Programming

For predictive analysis, we implement the models in Python 2.7 programming language in Spark ML (Machine Learning) and Analytics Zoo,which provides the machine/deep leaning libraries. The features that affect the prediction to build models are [user, item] and the label is [rating] in the range 1 to 5.

In Spark ML, we have used Alternating Least Squares (ALS) algorithm to build the recommender. Spark ML pipeline is composed of parameters and *pipeline fit* method is used to train the model. For the accuracy, we evaluate the model by computing MAE in the rating prediction for each product by the user.

In Analytics Zoo integrated into the Spark cluster, Neural Collaborative Filtering (NCF) with explict feedback is used as a Recommender API in Analytics, which is a neural network recommendation system. NCF leverages a multi-layer perceptrons to learn the *user* and *item* interaction function. Then, NCF can express and generalize matrix factorization and the *users* can build a NCF with or without matrix factorization [14]. Additionally, optimizer of BigDL is adopted to train the model.

5.4 Experimental Results

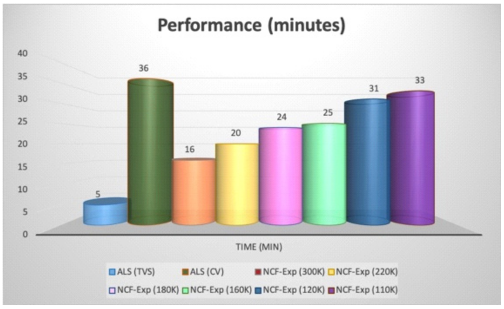
As a process of data engineering, the dataset is split into 80:20 ratios for training and testing. Initially, we train the model with *TrainValidation* method. Then, we used

CrossValidation method with 8 folds. We applied the combination of Parameters to train the ALS models as follows:

Rank: [1, 5]
Maximum Iteration: [5, 10]
Regularization Parameters: [0.3, 0.1, 0.01]
Alpha: [2.0, 3.0]

The pipeline is used as an estimator to train the model. The models are then evaluated using *RegressionEvaluator* for the Mean Absolute Error (MAE).

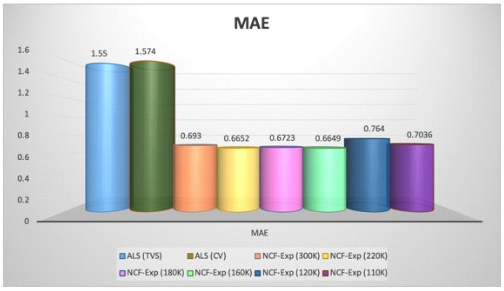
In Analytics Zoo on the same Hadoop Spark cluster, the data set is also randomly split into 80:20 ratios for training and testing. NCF neural network algorithms is adopted with *Optimizer* for training a model. Then, several batch sizes are given in the range of 110K to 300K for more about 6.5 million records with 10 epochs.



(Figure 3) Performance

Figure 3shows the experimental result of the computing times when evaluating the models with the training and test data set. ALS modes in PySpark take 5 and 36 minutes for TVS (*TrainValidationSplit)* and CV (*CrossValidation)* respectively. For Analytics Zoo, it takes 16 to 33 minutes as decreasing the batch size from 300,000 to 110,000.

Figure 4 shows the MAEs as a measurement to evaluate the accuracy of the models. ALS modes in PySpark has 1.55 and 1.574 for TVS and CV respectively. For Analytics Zoo, the MAEs varies from 0.693 to 0.7036 while decreasing the batch size from 300,000 to 110,000. When the batch size is 220,000, MAE becomes the minimum value 0.6652. We observe that the computing times in Analytics Zoo are similar to build an ALS model with Cross Validation in Spark ML. Furthermore, integrating deep learning with Spark has 55% less MAE than Spark alone



(Figure 4) Mean Absolute Error (MAE)

6. Conclusion

In the paper, Amazon S3 and EMR Big Data systems are adopted to store and analyze about 3.63 GB amazon data set, which is linearly scalable when the data set grows further.

Analytics Zoo provides a unified analytics and AI platform that seamlessly unites Spark, TensorFlow, Keras, and BigDL programs into an integrated pipeline. The entire pipeline can then transparently scale out to a large Hadoop/Spark cluster for distributed training or inference.

The paper presents the experimental result of the Spark ML and Deep Learning using Analytics Zoo. Recommendation model has been implemented to predict ratings of the items and a comparison has been made between the Spark ML and Analytics Zoo on the basis of accuracy and the performance of these models.

We developed prediction models byintegrating Spark cluster and multi-core deep learning, which leverages the existing Big Data Spark cluster, and which can obtaina distributed and single-chip parallel computing systems. Then, we show that Deep Learning Spark models present 55% more accurate predictions than Spark ML while the computing time is similar, especially with Cross Validation modeling.

References

[1] A. S. Jain, "Top 10 Benefits of Online Shopping (and 10 Disadvantages),"*ToughNickel*, 2018. [Online]. Available: https://toughnickel.com/frugal-living/Online-shopping-sites-benefits.

[2] C. Morah, "Shopping Online: Convenience, Bargains And A Few Scams,"*Investopedia*, 2018. [Online]. Available: https://www.investopedia.com/articles/pf/08/buy-sell-online.asp.

[3] D. L. Montaldo, "When It Is Best to Shop Online and When It Is Not?," *The Balance*, 2019. [Online]. Available: https://www.thebalance.com/the-pros-and-cons-of-online-shopping-939775.

[4] Valencia, S. (2018, June 19). An Introductory Recommender Systems Tutorial - AI Society. Retrieved from https://medium.com/ai-society/a-concise-recommender-systems-tutorial-fa40d5a9c0fa

[5] Ward, S. (2019, June 25). How Do Brick and Mortar Stores Compare with Online Retail Sites? Retrieved May 10, 2020, from https://www.thebalancesmb.com/compare-brick-and-mortar-stores-vs-online-retail-sites-4571050

[6] Cahill, Jannie. "Why Ratings and Reviews Matter." Profitero, 17 Oct. 2019, www.profitero.com/2017/06/why-ratings-and-reviews-matter/.

[7] Roy, S., Sharma, M., & Singh, S. K., ”Movie Recommendation System Using Semi- Supervised Learning,“ In 2019 Global Conference for Advancement in Technology (GCAT)1-5). IEEE, 2019, October 2019.

[8] Woo, Jongwook, & Xu, Yuhang (July 18-21, 2011), Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing, The 2011 international Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2011), Las Vegas.

[9] Monika Mishra, Jaydeep Chopde, Maitri Shah, Pankti Parikh, Rakshith Chandan Babu, Jongwook Woo, "Big Data Predictive Analysis of Amazon Product Review", KSII The 14th Asia Pacific International Conference on Information Science and Technology (APIC-IST) 2019, pp141-147, 2019.

[10] B. Patel, "Predicting Amazon product reviews' ratings,"*Towards Data Science*, 26-Apr-2017. [Online]. Available: https://towardsdatascience.com/predicting-sentiment-of-amazon-product-reviews-6370f466fa73.

[11] M. Woolf, "Playing with 80 Million Amazon Product Review Ratings Using Apache Spark,"*minimaxir*, 02-Jan-2017. [Online]. Available: https://minimaxir.com/2017/01/amazon-spark/.

[12] S3.amazonaws.com (n.d), Amazon Reviews Multilingual Dataset from https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_multilingual_US_v1_00.tsv.gz

[13] BigDL: Distributed Deep Learning on Apache Spark, https://software.intel.com/en-us/articles/bigdl-distributed-deep-learning-on-apache-spark

[14] Intel Analytics Zoo, https://software.intel.com/en-us/ai/analytics-zoo

[15] Purushu P, Melcher N, Bhagwat B, Woo J, ”Predictive Analysis of Finan-cial Fraud Detection using Azure and Spark ML,“ Asia Pacific Journal of In-formation Systems (APJIS), VOL.28, NO.4, pp.308~319, December 2018.

[16] Monika Mishra, Mingoo Kang, Jongwook Woo, "Rating Prediction using Deep Learning and Spark", The 11th International Conference on Internet (ICONI 2019), pp307-310, Dec 15-18 2019.

[17] D. Dauletbak and J. Woo, "Big Data Analysis and Prediction of Traffic in Los Angeles", KSII Transactions on Internet and Information Systems, Vol. 14, No. 2, pp. 841-854, 2020. https://doi.org/10.3837/tiis.2020.02.021

[18] Qiuhua Wang, Xiaoqin Ouyang, and Jiacheng Zhan. ”A Classification Algorithm Based on Data Clustering and Data Reduction for Intrusion Detection System over Big Data,“ KSII Transactions on Internet and Information Systems, Vol.13, No.7, pp.3714-3732, 2019. https://doi.org/10.3837/tiis.2019.07.021

[19] Fidalcastro.A and Baburaj.E. “Sequential Pattern Mining for Intrusion Detection System with Feature Selection on Big Data,“ KSII Transactions on Internet

and Information Systems, Vol.11, No.10, pp.5023-5038, 2017.
https://doi.org10.3837/tiis.2017.10.018

◐ 저 자 소 개 ◑

Monika Mishra

Ms. Monika Mishra is a recent graduate from California State University, Los Angeles. She has extensive work experience in the Oracle E-Business Suite Applications. Her interest lies in the Big Data Science and Analytics field. Currently, she is working as a Business Intelligence Developer in the healthcare industry.

Mingoo Kang

Dr. MinGoo Kang is a professor (Dean of Academic affairs) in the Dept.of IT Contents at Hanshin University, Osan South Korea from 2000. He has received the B.S., M.S., and Ph.D. degrees from Yonsei University, Seoul, Korea all in Electronic Engineering in 1986, 1989 and 1994, respectively. He was a research engineer at Samsung Electronics from 1985 to 1997. His research interests include wireless communication algorithm, mobile devices, and Smart UX for Mobile TV & DTV. He also served a chair of the Korean Society of Internet Information from 2018 to 2019.

Jongwook Woo

Dr. Jongwook Woo received his Ph.D. from USC and went to Yonsei University. He is a Professor at CIS Department of California State University Los Angeles. He serves as a Technical Advisor of Tim Solution, Council Member of IBM Spark Technology Center, and president at KSEA-SC. He has consulted companies in Hollywood: CitySearch, ARM, E!, Warner Bros, SBC Interactive. He published more than 40 papers, and his research interests include Big Data Analysis and Prediction. He has been awarded Teradata TUN faculty Scholarship and received grants from Amazon, Google Cloud Platform, IBM, Oracle, MicroSoft, DataBricks, Cloudera, SAS, QlikView, Tableau. He is a founder of Hemosoo Inc and The Big Link.