Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 18 Issue 3 Version 1.0 Year 2018
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Spark Big Data Analysis of World Development Indicators

By Kunal Pritwani, Knox Wasley & Jongwook Woo

California State University

Abstract- We would like to analyze different development indicators such as life expectancy of a country, patent applications by residents, and trademark of applications which serves as great analyzation for the Business Analysts, Financial Analysts, and Data Scientists. Data set is collected from the World Bank websites, for this analysis, which has world development indicators with respect to the country. The data is analyzed based on the yearly timeline, geographical locations on the map and also top 10 countries for a particular world development indicator. Moreover, it has found that the countries where the life expectancy is high the people are more creative and the patent applications are created on a huge scale. Also, the trademark applications are more where the life expectancy is higher. This analysis provides insights on the world development indicators. In the paper, data analysis is done on a huge dataset by using Spark on Hadoop Big Data cluster and its visualization charts are presented.

Keywords: *big data, spark, databricks, life expectancy, trademark applications, patent applications, tableau hadoop, world development indicators data analysis.*

GJCST-C Classification: *H.3.m*

Strictly as per the compliance and regulations of:

Spark Big Data Analysis of World Development Indicators

Kunal Pritwani [α], Knox Wasley [σ] & Jongwook Woo [ρ]

Abstract- We would like to analyze different development indicators such as life expectancy of a country, patent applications by residents, and trademark of applications which serves as great analyzation for the Business Analysts, Financial Analysts, and Data Scientists. Data set is collected from the World Bank websites, for this analysis, which has world development indicators with respect to the country. The data is analyzed based on the yearly timeline, geographical locations on the map and also top 10 countries for a particular world development indicator. Moreover, it has found that the countries where the life expectancy is high the people are more creative and the patent applications are created on a huge scale. Also, the trademark applications are more where the life expectancy is higher. This analysis provides insights on the world development indicators. In the paper, data analysis is done on a huge dataset by using Spark on Hadoop Big Data cluster and its visualization charts are presented.

Keywords: big data, spark, databricks, life expectancy, trademark applications, patent applications, tableau hadoop, world development indicators data analysis.

I. Introduction

Our goal was to analyze the world development indicators which are important factors in big data analytics. This kind of data is analyzed by big name analysts for big money as this kind of analysis provides insights on different aspects of development indicators. The outcome of this analysis will help Business Analysts, Financial Analysts, and Data Scientists to compare between different development indicators and select the world development indicator that would have a better impact on the economy of the country.

Big Data is defined as non-expensive frameworks that can store a large scale data and process it in parallel [4, 5]. A large scale data means really a big data, this data cannot be processed using traditional computing techniques. Data is getting generated everyday through social media, websites, mobile applications etc. To analyze and store data we use Hadoop, which is an open source framework which provides distributed storage on the commodity hardware. Hadoop has two major components which are MapReduce and HDFS (Hadoop Distributed File System).

Authors α σ ρ: Department of Computer Information Systems, College of Business and Economics, California State University, Los Angeles.
e-mails: kpritwa@calstatela.edu, kwasley@calstatela.edu, Andjwoo5@calstatela.edu

Spark and tableau are adopted for data analyzation and visualization. Technically, to process a huge chunk of data a fast processor is required to process the big data. Hadoop is a technique which creates a cluster on the commodity hardware and stores and process the data on multiple nodes while having multiple copies of the same node on the cluster. Old Hadoop clusters use map reduce technique to map and store the data but Apache Spark is a newer version which is faster and runs on top of Hadoop architecture and it runs in memory.

Apache Spark runs 100 times faster than Hadoop but it doesn't have its own HDFS. So it uses HDFS as its file system and runs on top of Hadoop by using memory. Spark uses RDD (Resilient Distributed Datasets) which replaces the Map Reduce functionality to write the data to physical storage every time.

II. Related Works

Miranda and Michael does the data analysis using statistical techniques to find the correlation between different columns. But, we have used spark to manipulate and visualize the data to get useful insights [Chen, Ching 2000]. Life expectancy is analyzed by selecting the multiple columns using statistical techniques to find the correlation and by using scatter plots for visualization [Chen, Ching 2000]. We simply used geographical visualization to show top earning states. Paul uses basic approach for analysis of top countries for patent applications by using bar graph and we used geographical visualization with time series analysis using historical data [6]. Besides, Spark computation is less time consuming to process the results.

We have used Big Data Spark platform to store and analyze the data and BI tool such as tableau for visualizations. By analyzing the 247 countries data of 54 years, we have different results as we analyzed a very huge dataset. We have the detailed analysis for 247 countries and they have analysis for around top 10 countries. We have found that top countries where people are being more creative and innovative due to high life expectancy which helps the economic growth of the country. Spark helps to process the queries and gives the results fast and also spark has a very less lines of code as compared to other technologies.

III. Methods

First, we collected the data from an online community dedicated to data scientists where the dataset comprises of historical data of 5,656,458 rows which contain over a thousand annual indicators of economic development from hundreds of countries around the world. Further, by using the Spark technique to find different terminologies like we would like to analyze different development indicators like Life expectancy of a country, Patent applications by residents and Trademark of applications. Detailed Analysis of world development indicators has been performed using data visualization tools.

a) Specification of Data Set

The data is collected from an online community kaggle. We have historical data of about 5,656,458 rows which contain over a thousand annual indicators of economic development from hundreds of countries around the world. To be precise, there are 1344 indicators and 247 countries in this dataset spanning of 54 years. The data size is 574 MB and file is in CSV (Comma Separated Values) format.

b) Platforms

Data Analysis tools used are Apache Spark cluster on Databricks cloud platform, and visualization tool Tableau 9.2 is used for detailed data analysis for daily and yearly records.

i. *Cloud Computing: Databricks*

In order to collect and analyze data, cloud computing service from Databricks is adopted. Databricks provides a very fast data platform which runs on top of apache spark that helps to easily create big data advanced analytics solution. It can *be connected directly to the existing storage apache clusters and Databricks services in the cloud. It provides highly integrated workspace to* create dashboards by using notebooks. Also it provides a functionality to use third party Business Intelligence tools and create custom spark applications with spark production jobs. Basically Databricks provides a very easy to use cloud platform which reduces the use of high end and expensive hardware. There is no need to install the Apache Spark environment which is a huge time saver for Data scientists, Data analysts and Data engineers. It also provides option to choose different spark versions like Spark 1.6.1(Hadoop 1).

There are various programing languages available on Databricks in form of different notebooks like Scala, python, SQL and R. The instances are highly scalable which can be modified as per the user. Also the visualization can be done instantly instead of exporting it to other visualization tools. Notebook access can be given to multiple users to edit or read [2].

Figure 1. Shows the different languages spark API supports like R, SQL, Python, Scala and Java. Also shows that spark supports data frames, streaming, machine learning and GraphX usability.

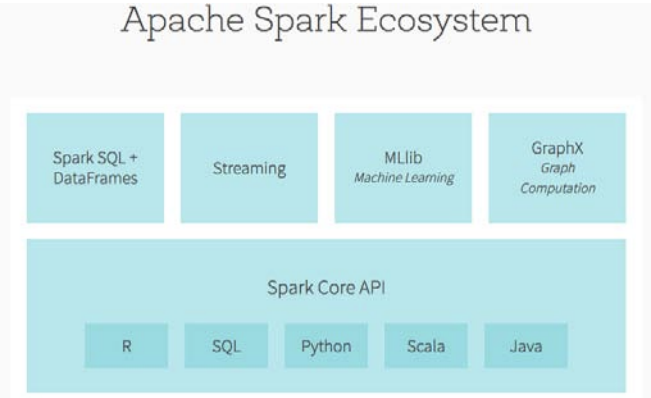

Figure 1: Apache Spark Ecosystem

In the Spark service of Databricks, the code is written in the ipython notebook. The spark cluster is accessed by the notebook, on which the query processing is done. In to the cluster, we have to upload the data file, in this case its 'indicators_csv'. We can change the data_type also during the creation of the table. We have to create RDDs using SQL context and run the queries which are discussed in the analysis part.

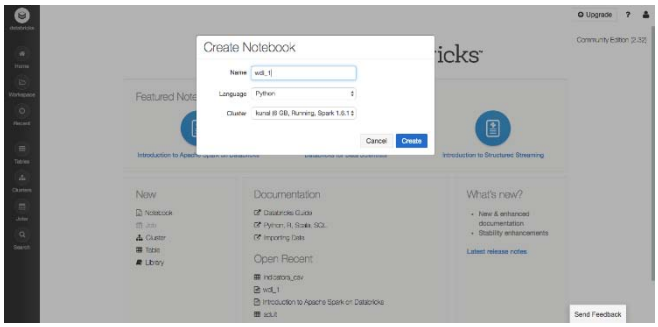

Figure 2: Creating a notebook in Databricks

Figure 2. Shows a screenshot of a notebook creation where it gives you the option of selecting a language, In this case it's python also called as 'PySpark'. Also gives you the option of selecting a cluster in this case its Spark 1.6.1. This cluster is Apache Spark Version (Spark 1.6.1) of Databricks. Memory is 6GB with 0.88 Cores CPU cores and 1 CPU master node.

Figure 3: Table format in Databricks

Figure 3. Shows a table which was imported 'indicators_csv' and here we can also modify the datatype of the column. The PySpark code is built, where Databricks automatically takes a default datatype.

ii. *Visualization: Tableau*

Tableau is adopted for visualizing the result data set that is computed in Spark, which is a business intelligence tool. It is easy to use and produce interactive visualizations to get the insights using data analytics techniques. Tableau provides to the traditional small data set a user friendly and powerful environment for Data Scientist, Data Analyst and Data Engineers. It can produce visualization from relational database, cloud databases and excel files. However, it cannot compute huge data set. The data analysis code is built and run in Spark in order to generate and find out insights, which is the result data set and small amount of data. Thus, Tableau can take the result data set of the data analysis in Spark and it can produce the graphs in the next sections. It provides a number of types of graphs like bar graphs, pie charts, line graphs, geographical maps, Gantt chart etc. Tableau is used to get the hidden insights from the data which can help to improve the world by implementing changes to the world development indicators [9].

c) *Terminology*

i. *Life expectancy at birth, total (years)*

Derived from male and female future during childbirth from sources, for example, (1) United Nations Population Division. Total populace Prospects, (2) Census reports and other factual productions from national measurable workplaces, (3) Eurostat: Demographic Statistics, (4) United Nations Statistical Division. Populace and Vital Statistics Report (different years), (5) U.S. Enumeration Bureau: International Database, and (6) Secretariat of the Pacific Community: Statistics and Demography Program [3].

ii. *Patent applications, residents*

World Intellectual Property Organization (WIPO), WIPO Patent Report: Statistics on Worldwide Patent Activity. The International Bureau of WIPO accepts no accountability concerning the change of these information [3].

iii. *Trademark applications, total*

Trademark applications documented are applications to enlist a trademark with a national or local Intellectual Property (IP) office. A trademark is an unmistakable sign which distinguishes certain merchandise or administrations as those created or gave by a particular individual or venture. A trademark gives assurance to the proprietor of the check by guaranteeing the select appropriate to utilize it to distinguish products or benefits, or to approve another to utilize it as an end-result of installment. The time of security fluctuates, however a trademark can be reestablished uncertainly past as far as possible on installment of extra charges [3,11].

IV. Detail Data Analysis Results

a) *Life expectancy at birth, total (years)*

This formula selects columns CountryName and Value with a filter on Indicator Name as "Life expectancy at birth, total (years)" Results are stored in 'results' RDD and then displayed using Spark Display command [8].

➔ Results= sqlContext.sql('Select CountryName, Value from indicators_csv where IndicatorName = "Life expectancy at birth, total (years)" order by Value desc')

➔ Display (results)

Figure 4. Shows the geographical view of life expectancy on the map. Life expectancy has a high value and the lighter regions have the less value. United States is dark green so it means that the Life expectancy is good in United States and In Africa the area is light green which means the life expectancy is less.

Figure 5 shows the top countries which have high value for life expectancy. In this case San Marino has the highest average value in world for life expectancy as 81.49.

Figure 6. Shows the Lowest Life Expectancy at birth, Top 10 countries. In this case it shows that Seirra Leone has the lowest value for life expectancy as 38.68.

Figure 7. Shows the life expectancy of United States from 1960 to 2013. The trend shows that the life expectancy is increasing which increases the United States Growth.

b) *Patent applications, residents*

This formula selects columns CountryName and Patent applications, residents with the filter on column IndicatorName as "Patent applications, residents". Results are stored in 'results' RDD and then displayed using Spark Display command [8]. Refer the code at Github[12].

Figure 8. Shows Patent applications by residents, Top 10 countries. In this graph we can see that Japan and USA has the highest Average value all around the world.

Figure 9. Shows Patent applications by residents, Lowest Top 10 countries. In this graph we can see that Aruba has the lowest Average value all around the world.

Figure 10. Shows the geographical view of Patent applications by residents on the map. Patent applications which have a high value are darker regions and the lighter regions have the less value. Japan, United States are dark blue so It means that the Patent applications are more in United States, China and Japan and in Africa the area is light blue which means the life expectancy is less.

Figure 11. Shows the Patent applications by residents of United States from 1960 to 2013. The trend shows that the Patent applications is increasing which increases the United States Growth.

c) Trademark applications, total

This formula selects columns CountryName and Trademark applications, residents with a filter on IndicatorName as "Trademark applications, total". Results are stored in 'results' RDD and then displayed using Spark Display command [8].Refer the code at Github[12].

Figure 12. Shows Trademark applications for Top 10 countries. In this graph we can see that China, Japan and USA has the highest Average value all around the world.

Figure 13. Shows the geographical view of life expectancy on the map. Trademark applications have which have a high value are darker regions and the lighter regions have the less value. China, Japan, United States are dark yellow so It means that the Trademark applications are more in China and In Africa the area is light which means less number of Trademark applications.

Figure 14. Shows the Trademark applications of United States from 1960 to 2013. The trend shows that the Trademark applications are increasing which increases the United States Growth.

Figure 4: Life Expectancy at birth on the Map

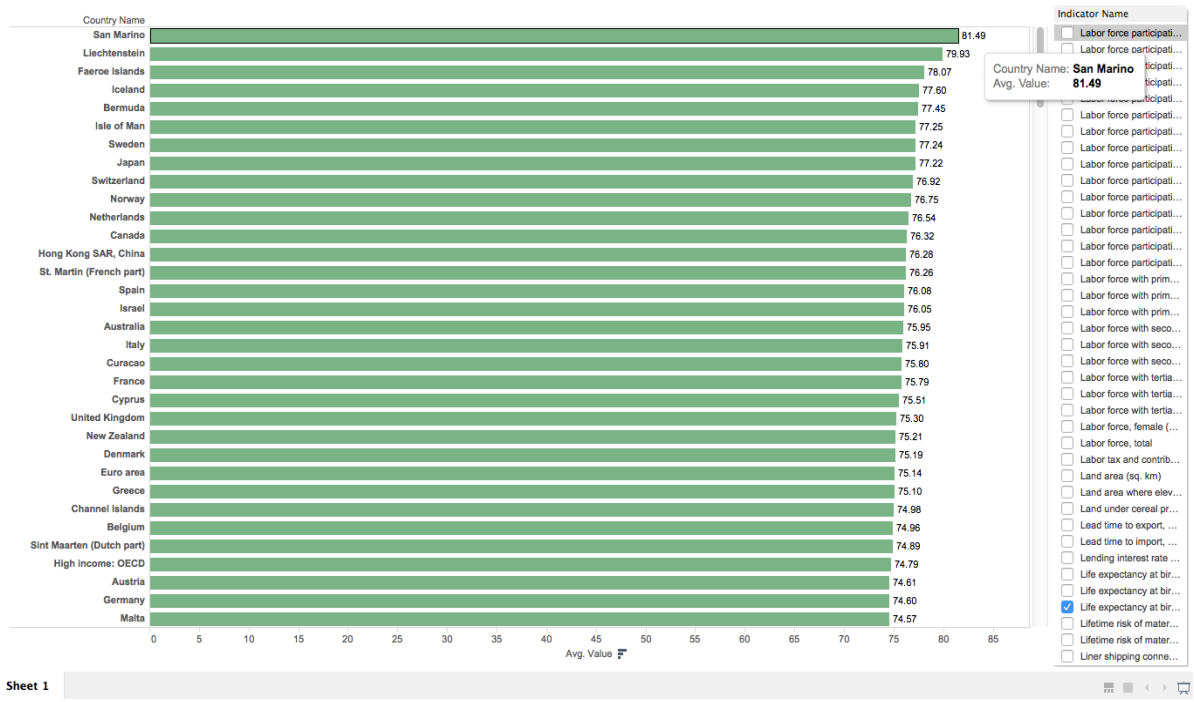

Figure 5: Life Expectancy at birth, Top 10 countries

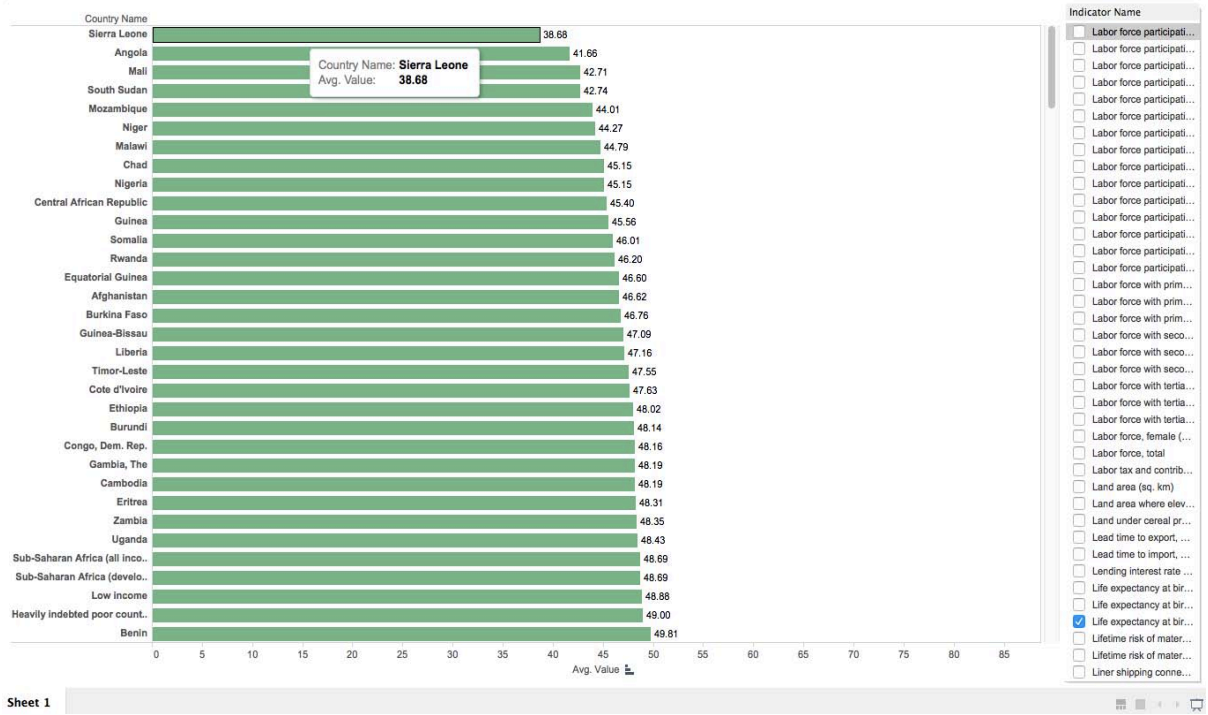

Figure 6: Lowest Life Expectancy at birth, Top 10 countries

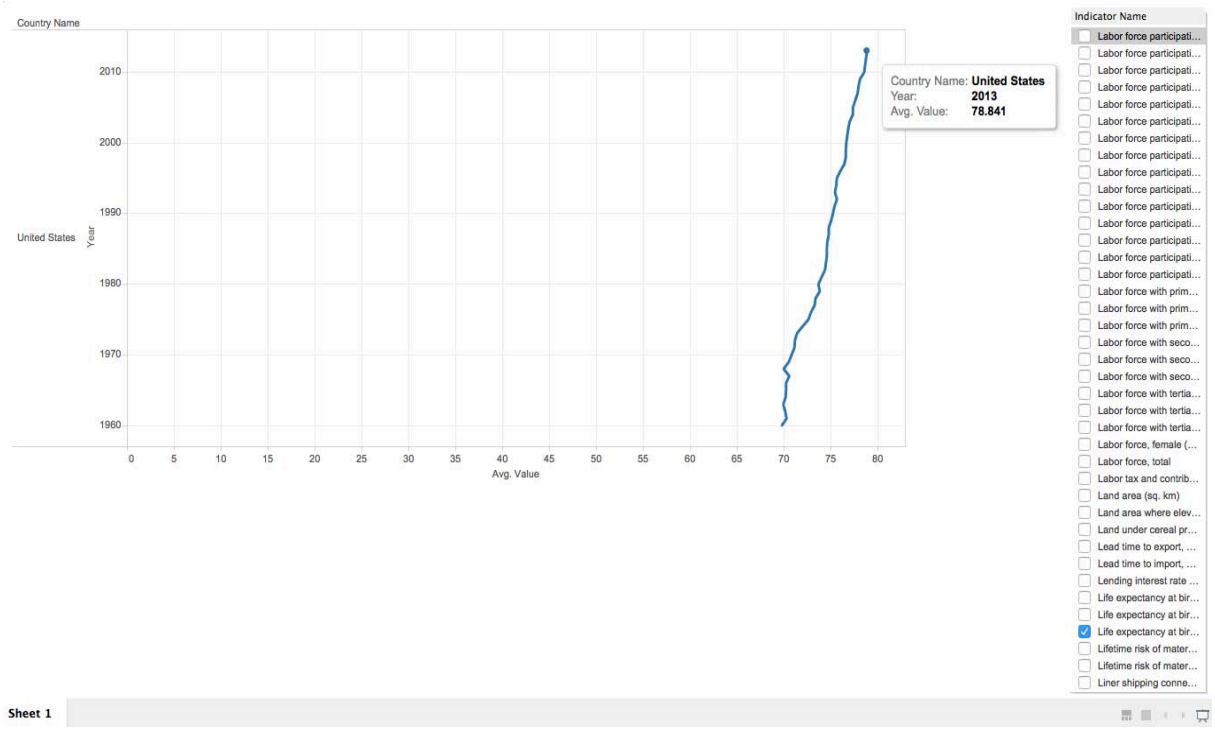

Figure 7: Life Expectancy at birth, from 1960 to 2013(United States)

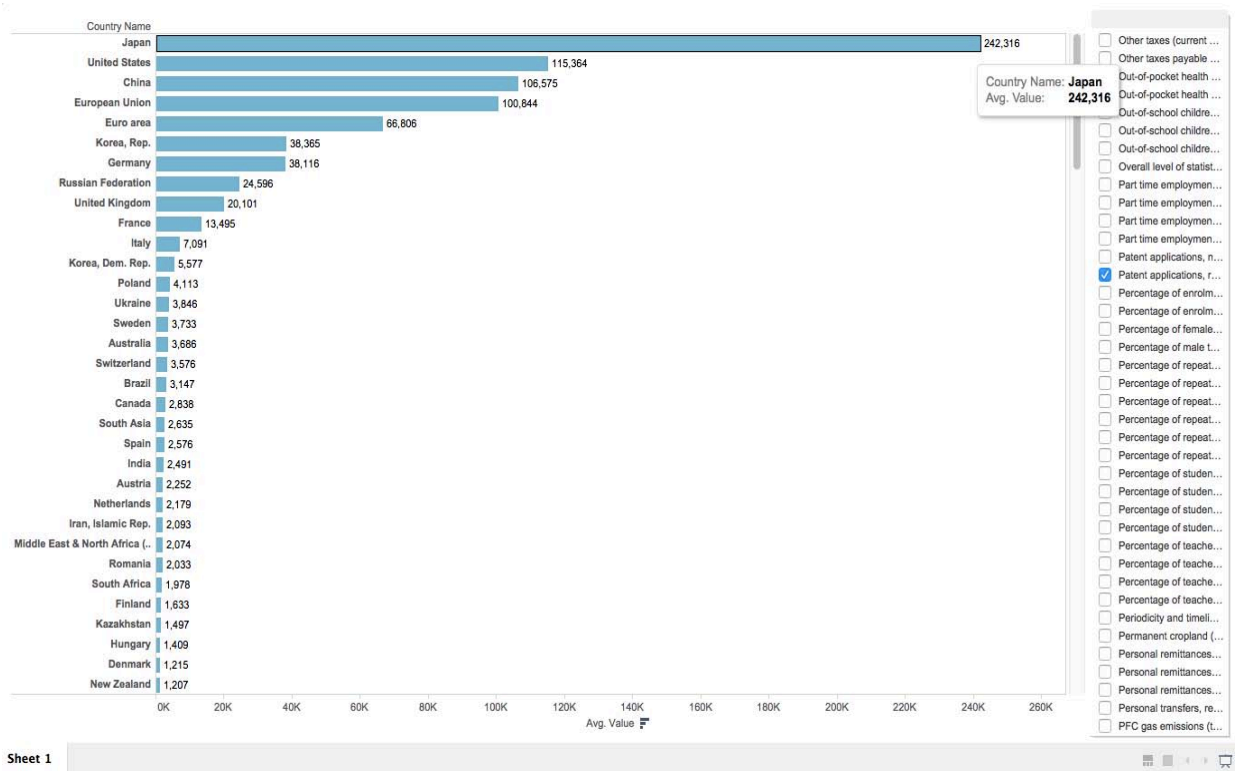

Figure 8: Patent applications by residents, Top 10 countries

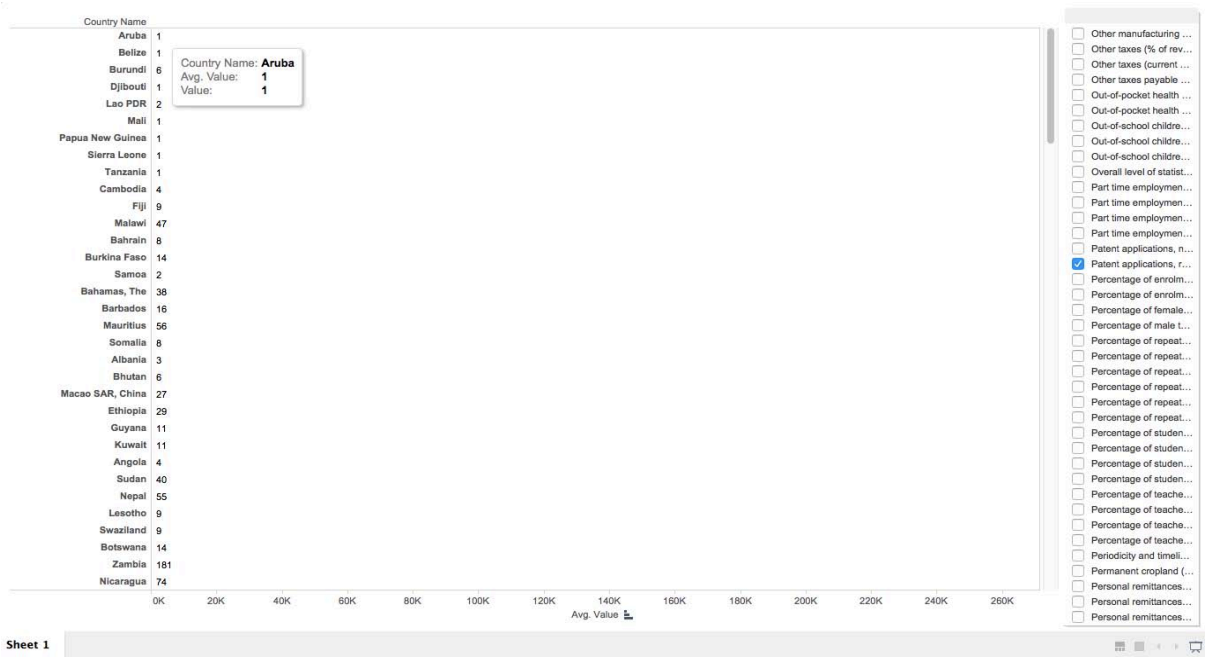

Figure 9: Patent applications by residents, Lowest Top 10 countries

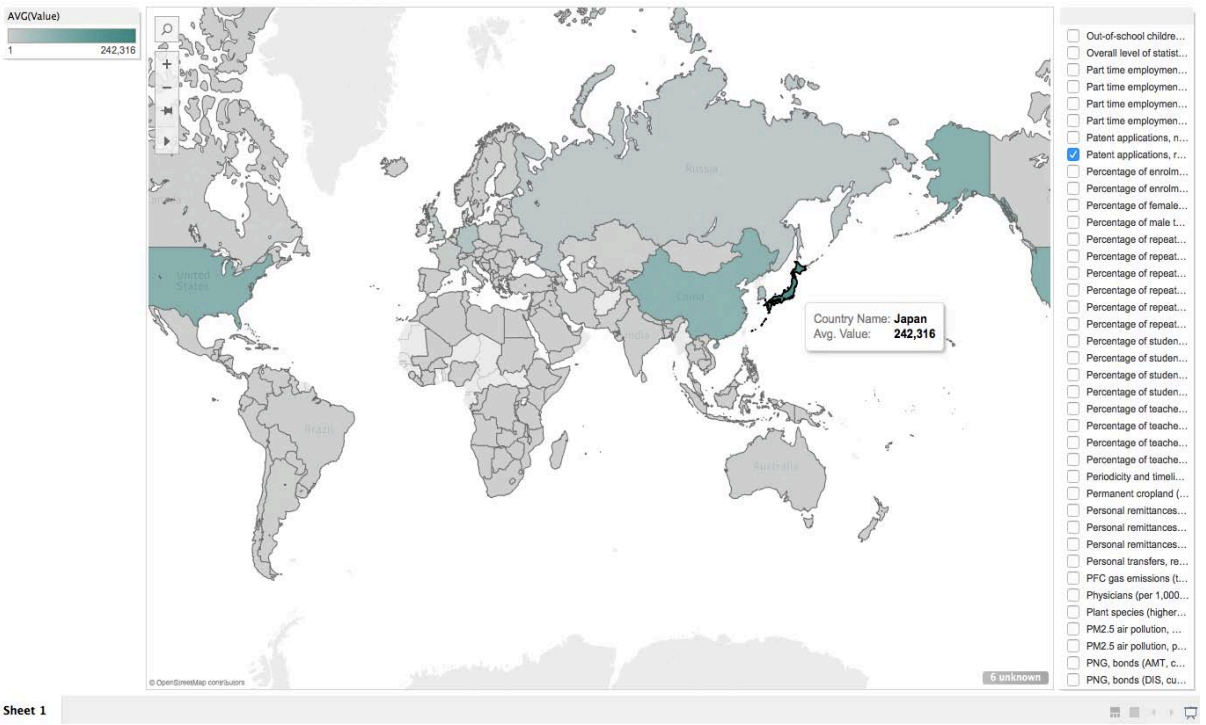

Figure 10: Patent applications by residents on map

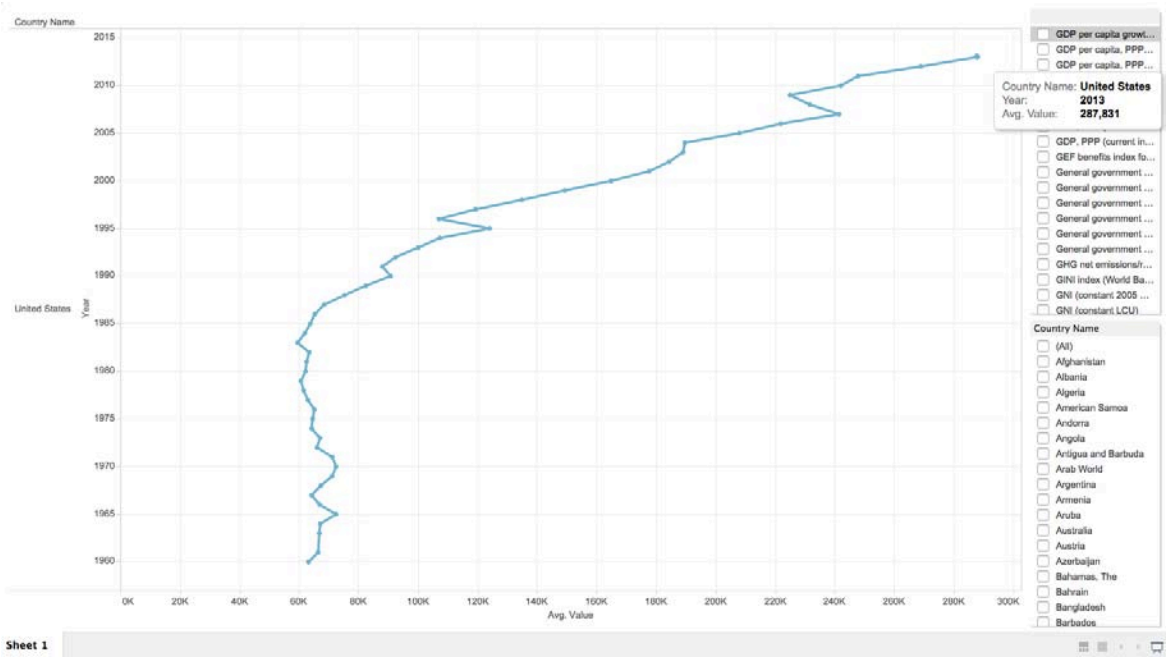

Figure 11: Patent applications by residents, from 1960 to 2013(United States)

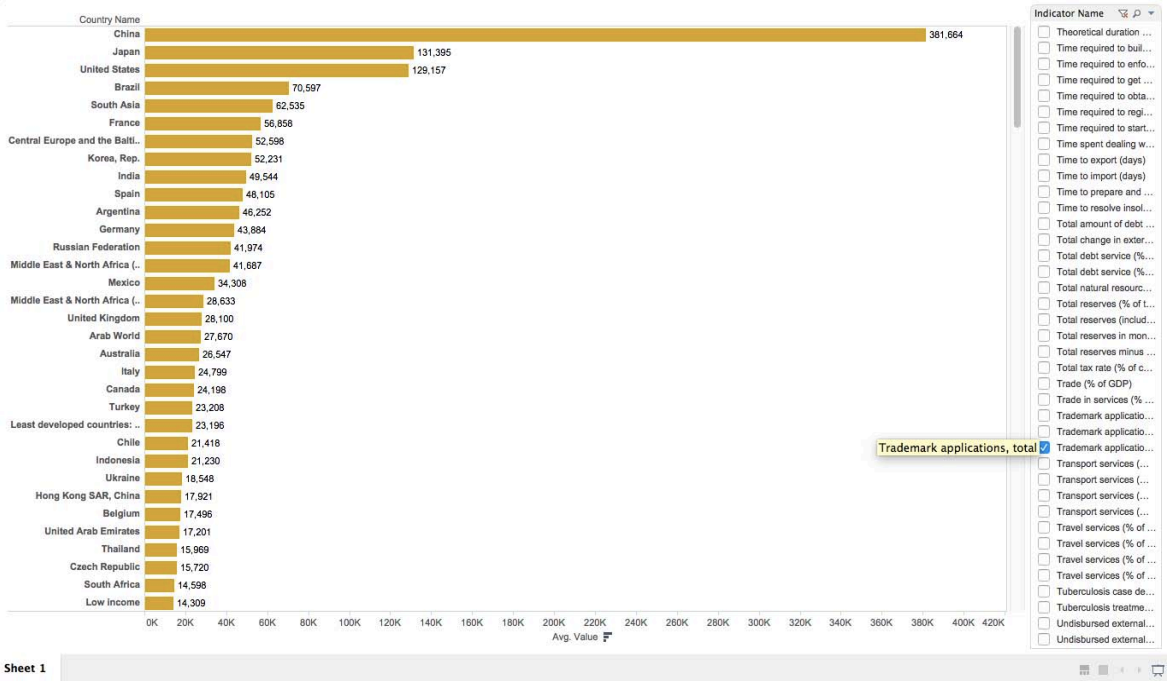

Figure 12: Trademark applications, Top 10 countries

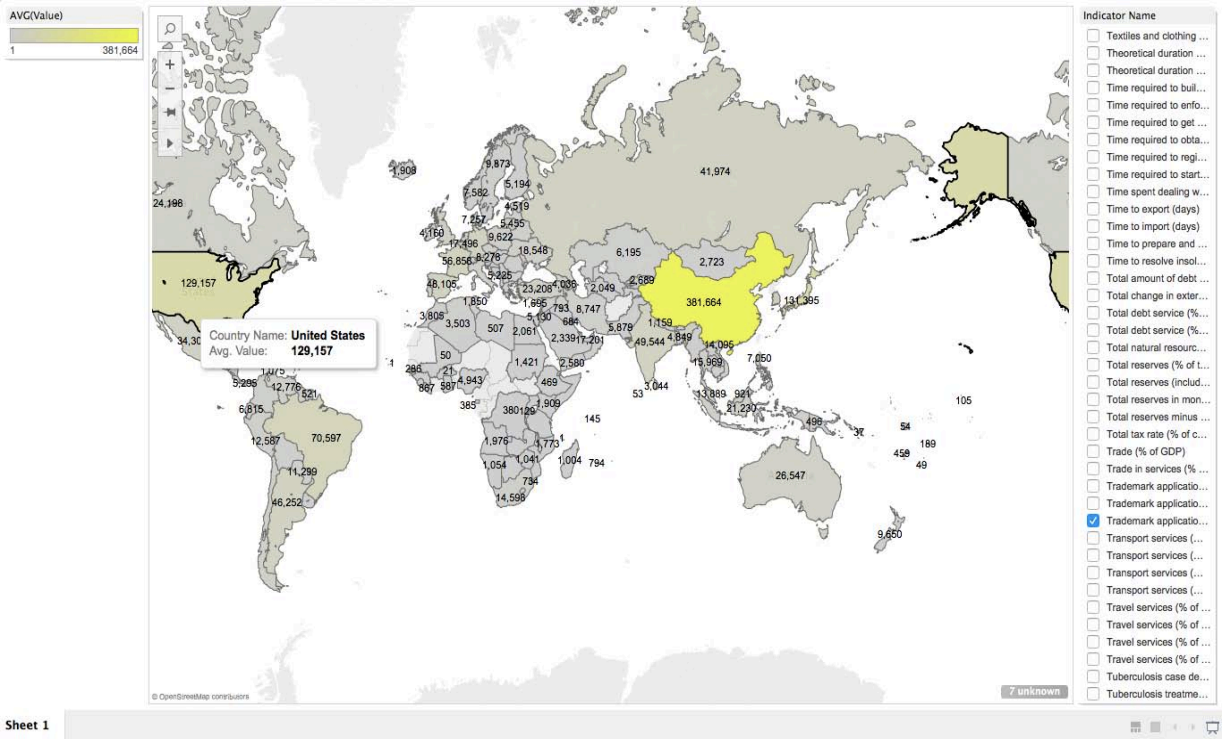

Figure 13: Trademark applications on map

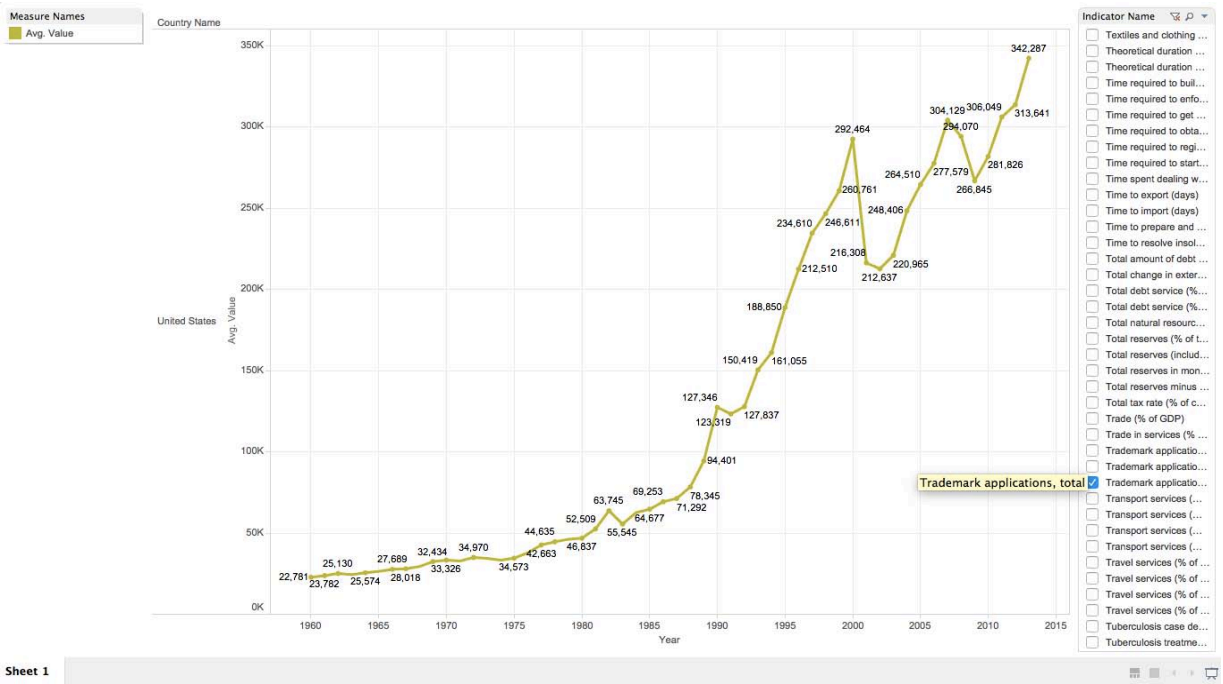

Figure 14: Trademark applications, from 1960 to 2013(United States)

V. Conclusion

To conclude, world development indicators reflect the growth of the country using big data analytics. This kind of insight will be charged huge sum by data analyst for what we just presented with the analysis insights. It has found that the countries where the life expectancy is high, the people are more creative and the patent applications are created on a huge scale. Also the trademark applications are more where the life expectancy is higher. This analysis provides insights on the world development indicators. These analysis is helpful for decision making at a higher level where the growth factors of the country are planned.

It is found that life expectancy and creativity of the people are correlated. The more people live, the more creative they become. They will create more creative applications and trademarks for the betterment of the country and world. Technology plays a huge role in finding these insights as it helps the analysts for decision making. Big data technology is the future for decision making systems as the data is getting bigger and bigger every day and we need to analyze, process, store and fault tolerance this big data. Databricks helps users to easily create Apache Spark Hadoop clusters and run the queries on huge chunks of data. Tableau provides a wide variety of data visualization options to gather some use full insights like "finding a gold in Data Lake" so that better decisions can be made for the development of the country and world.

Acknowledgement

This research was supported by Amazon AWS research grant.

References Références Referencias

1. Chen, M., and Ching, M. (2000, December 18). "A STATISTICAL ANALYSIS OF LIFE EXPECTANCY ACROSS COUNTRIES USING MULTIPLE REGRESSION". Retrieved April 15, 2017, from http://www.seas.upenn.edu/~ese302/Projects/Project_2 pdf
2. Databricks. (2016). "What is Apache Spark?" Retrieved November 02, 2016, from https://databricks.com/spark/about
3. Worldbank. (2013). "Life expectancy at birth, total (years)". Retrieved November 10, 2016, from http://data.worldbank.org/indicator/SP.DYN.LE00.IN
4. Woo, J., and Xu Y. 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 (July 18-21, 2011).
5. Woo, J. (2013) "Market Basket Analysis Algorithms with MapReduce", DMKD-00150, Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, Oct 28 2013, Volume 3, Issue 6, pp445-452, ISSN 1942-4795.
6. Muggeridge, P. (2014, May 17). "Which countries file the most patent applications?". Retrieved April 16, 2017, from https://www.weforum.org/agenda/2015/09/which-countries-file-the-most-patent-applications/
7. Worldbank. (2013). "Patent applications, residents". Retrieved December November 14, 2016, from http://data.worldbank.org/indicator/IP.PAT.RESD
8. Spark Apache. (2016). "Spark SQL Data Frames and Datasets Guide". Retrieved November 29, 2016, from http://spark.apache.org/docs/latest/sql-programming-guide.html
9. Tableau Software. (2016). "Business intelligence for your people". Retrieved November 02, 2016, from https://www.tableau.com/resource/business-intelligence
10. Worldbank. (2013) "Trademark applications, total". Retrieved November 19, 2016, from http://data.worldbank.org/indicator/IP.TMK.TOTL
11. Trademark Economics. (2014). "Trademark applications - total in World". Retrieved November 27, 2016, from http://www.tradingeconomics.com/world/trademark-applications-total-wb-data.html
12. Github, "https://github.com/pritwanikunal/Big-Data-Analysis-of-World-Development-Indicators-using-Apache-Spark"