Big Data & Apache Spark Models

Data Intensive Architectures – 100% Project

Polina Prinii
Postgraduate Diploma in Science in Data Analytics
School of Computing, National College of Ireland, Mayor Street Lower, IFSC, Dublin1, D01Y300, Ireland
Email: x21137757@student.ncirl.ie

Abstract—**This document serves the purpose of a hypothesis interim paper outlining the project idea and an overview of the proposed processing model to address the big data datasets at hand. The hypothesis for the project holds no grading weight.**

I. Question

Over the past number of years overpopulation has been a controversial concern with the global population steadily rising since the 1950's. The current global population today stands at a whopping 7.7 billion, resulting in an increase of 5.2 billion since 1951. [1]

The project proposes the following hypothesis:

"***Is overpopulation still a concern with the steady decline of birth rates?***"

The project looks to research the proposed by utilizing Apache Spark, an open-source distributed processing system designed to work with big data workloads. Many various programming languages can be implemented to achieve the proposed research. [2]

The following datasets gathered from OurWorldinData.org. will accompany the project, they are:

- Fertility rates vs contraceptive prevalence which outlines the percentage of females practicing or whose partners are practicing a form of contraceptive. [3]
- Total Fertility Rates, which outlines the average number of live births per woman if they were to experience the exact age-specific fertility rates and live till the end of their reproductive life. [3]

II. Value

It is assumed that in the next 70 years the global population is set to hit a whooping 10.8 billion [4], that is considering the fertility rates decline steadily as they have been doing so since the 1950's. There are many factors which contribute to the steady growth of the global population, factors such as: low mortality rates with the high-paced growth in medicine as well as underutilization of contraception.

The project believes it can present value by clarifying the above hypothesis. At present one can assume that the answer to the hypothesis is not self-evident as multiple factors are considered when discussing the steady growth of the global population.

Therefore, the project will look to additionally include the following dataset from OurWorldinData:

- Annual number of deaths by world region which outlines yearly recorded deaths per country commencing from 1950 to 2019. [5]

The project hopes to undertake an analysis into the outlined three datasets to allow for the understanding if population is still a standing concern in line with the steady growth of the world's population.

III. Justification

Apache Spark has built its reputation for big data workloads being open-sourced and compatible with multiple programing language [6]. It is to no surprise at the vast number of time series application of through Apache Spark. From predictive analysis using a mixture of Spark and R [7] to utilizing machine learning algorithms such as k-weighted nearest neighbors along-side Spark [8].

Following the vast number of scholars available for consumption to the readers, the project failed to identify a research topic surrounding the prediction of fertility rates, ultimately failing to identify if overpopulation is still a concern.

The project believes there is novelty in researching the proposed hypothesis and assumes conclusion can be reached with machine learning algorithms such as Time Series forecasting as ultimately the data at hand is of timely nature.

With the compatibility function as offered by Apache Spark the project will look to store the three proposed datasets within its environment and further apply analysis using the R programming language.

Further if time allows the project will hope to apply an additional machine learning algorithm of Linear Regression to ultimately identify if a relationship exists between the various data points once the proposed datasets are merged.

References

[1] worldometer, "World Population by Year", Accessed on Apr 2nd 2022, [Online], Available at: https://www.worldometers.info/world-population/world-population-by-year/

[2] AWS, "Introduction to Apache Spark", Accessed on Apr 2nd 2022, [Online], Available at: https://aws.amazon.com/big-data/what-is-spark/

[3] Our World in Data, "Fertility Rate", Accessed on Apr 2nd 2022, [Online], Available at: https://ourworldindata.org/fertility-rate

[4] Population Media Centre, "Overpopulation: Cause and Effect", Accessed on Apr 2nd 2022, [Online], Available at: https://info.populationmedia.org/blog/overpopulation-cause-and-effect

[5] Our World in Data, "Annual number of deaths by world region", Accessed on Apr 2nd 2022, [Online], Available at: https://ourworldindata.org/grapher/annual-number-of-deaths-by-world-region

[6] Salloum, S., Dautov, R., Chen, X. *et al.* Big data analytics on Apache Spark. *Int J Data Sci Anal* **1,** 145–164 (2016). https://doi.org/10.1007/s41060-016-0027-9

[7] Krome, C., Sander, V. Time series analysis with apache spark and its applications to energy informatics. *Energy Inform* **1,** 40 (2018). https://doi.org/10.1186/s42162-018-0043-1

[8] Talavera-Llames, R., Pérez-Chacón, R., Troncoso, A. and Martínez-Álvarez, F., 2018. Big data time series forecasting based on nearest neighbours distributed computing with Spark. Knowledge-Based Systems, 161, pp.12-25. https://doi.org/10.1016/j.knosys.2018.07.026

[9]