Storage Solutions and Data Analytics: RDBM,

Hadoop and APIs in Neural Networks Contexts

Jose Maria Rico Leal  
*Master of Science in Data Analytics*  
sba23021@student.cct.ie

*Abstract*

*This study investigates the relationship between Big Data management tools such as Relational Database Management System (RDBMS), Hadoop, and APIs, and how they can be separately interlinked with advanced data analytics, specifically neural networks. The purpose of the study is to compare a RDBMS with Hadoop when processing a 1.31 GB dataset, and then apply a Neural Network. To expand the scope, this study will also include the usage of APIs (Keras library) for implementing Neural Networks. This study was conducted using my personal laptop to load a 1.31 GB dataset into a RDBMS and Spark. I utilized Jupyter Notebooks to interact with these two technologies, exploring computing times, roadblocks faced, and other insights. Following this, we applied the same Neural Network to predict if certain jobs are more popular based on gender. Another aspect of the study involves utilizing an API; for this, we are employing Keras and a Convolutional Neural Network (CNN). Our aim is to evaluate the performance of the CNN model in classifying movie reviews as positive or negative based on their sentiment. The research findings indicate that using RDBMS or Hadoop for data processing is not as quick and straightforward as using an API like Keras, where you simply import the data without the need to worry about how to push it into databases, this becomes clear when modelling data using NN via Jupyter Notebooks.*

Keywords: Relational Database Management System, Hadoop, API, Keras, Neural Network (NN), Convolutional Neural Network (CNN)

# Introduction

Relational Database Management Systems have been well-established since the late 1970s; at that time, the concept of Big Data was not the same as it is today. As technology rapidly advanced, the industry needed to process large amounts of data. To address this need, an open-source framework for writing and running distributed applications, called Hadoop, entered the scene (Lam, 2010). These two technologies, RDMS and Hadoop, are great; however, the implementation of both requires a high level of technical software skill. This is where APIs offer a solution to this problem, which the industry refers to as Machine Learning as a Service (MLaaS), e.g., Azure ML or AWS ML, just to mention a few (Atakan Cetinsoy et al., 2016).

The intention of this paper is to explore all three technologies—RDBMS, Hadoop, and APIs—to determine which one is the best fit for data extraction and processing in the context of Neural Networks implementation. This consideration is crucial, given that many individuals interested in Machine Learning are not software developers, and the need for a 'plug-in' to deploy their ML models is evident.

# Topic overview

The chosen topic is Big Data and Neural Networks, with NN being considered a type of Machine Learning (ML) process known as Deep Learning (Mishra and Gupta, 2017). The field of Big Data is constantly growing and encompasses a need for efficient data management and processing tools. Two well-known tools for handling and analyzing large datasets are Relational Database Management Systems (RDBMS) and Hadoop. However, the rampant advancement of Machine Learning and Neural Networks, the integration of these data management tools with advanced analytics technologies is the focus of this paper.

## Objectives

* Examine the current state of RDBMS, Hadoop, and APIs when used in modeling NN.
* Store a 1.31 GB dataset in both an RDBMS (SQL) and Hadoop, and then retrieve the data into a Jupyter Notebook to model a neural network.
* Utilize an API (Keras) to model a neural network and compare its performance in conjunction with RDBMS and Hadoop.
* Discuss the rationale behind the selection of the NN model for both scenarios

## Research question

How do Relational Database Management Systems (RDBMS) and Hadoop compare in terms of efficiency and effectiveness in processing large datasets for the application of neural networks, and how can APIs, particularly the Keras library, streamline the implementation of neural network models in data analytics (B Arnold, 2017).

# State of the art

## RDBMS

The current state of RDBMS has evolved substantially with enhancements in storage, speed, and scalability by using cloud-based solutions (Li et al., 2016). The future holds a shift for RDBMS transitioning to a NoSQL database (Candel, Sevilla Ruiz and García-Molina, 2022). To understand why NoSQL is taking over RDBMS, it is crucial to talk about: Schemas, where NoSQL uses dynamic instead of static schemas; the type of data to be stored, with NoSQL databases offering advantages for hierarchical data storage due to their flexible data models and scalability, while RDBMS are not that flexible; scalability, with NoSQL depending on horizontal scalability and RDBMS on vertical scalability; and other points where NoSQL surpasses RDBMS, including data warehouse, complexity, cloud, and big data handling, and output performance (Palanisamy and SuvithaVani, 2020).

## Hadhoop

## ML APIs

## Deep Learning Neural Networks

## Research methodologies and key papers

## Topic Overview

* Objectives
* Research questions
* State of Art

# Literature review

## Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the Microsoft Word, Letter file.

## Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

# Implementation

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.

Identify applicable funding agency here. If none, delete this text box.

* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Critical evaluation

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

# Future work

# Conclusions

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

Arnold, B. and T., 2017. kerasR: R Interface to the Keras Deep Learning Library. *The Journal of Open Source Software*, 2(14), 296. <https://doi.org/10.21105/joss.00296>.

Candel, C.J.F., Sevilla Ruiz, D. and García-Molina, J.J., 2022. A unified metamodel for NoSQL and relational databases. Information Systems, 104, 101898. https://doi.org/10.1016/j.is.2021.101898.

Cetinsoy, A., Martin, F.J., Ortega, J.A. and Petersen, P., 2016. The Past, Present, and Future of Machine Learning APIs. In*: PAPIs 2015 - Proceedings of the 2015 Conference on Predictive APIs and Apps*, vol. 50, JMLR: Workshop and Conference Proceedings, pp.43-49.‌

Lam, C. (2010). *Hadoop in Action*. Simon and Schuster.

Li, F., Das, S., Syamala, M. and Narasayya, V.R., 2016. Accelerating Relational Databases by Leveraging Remote Memory and RDMA. In: *Proceedings of the 2016 International Conference on Management of Data - SIGMOD ’16.* https://doi.org/10.1145/2882903.2882949.

Mishra, C. and Gupta, D.L., 2017. Deep Machine Learning and Neural Networks: An Overview. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 6(2), pp.66-73. <https://doi.org/10.11591/ijai.v6.i2.pp66-73>.

Palanisamy, S. and SuvithaVani, P., 2020. A survey on RDBMS and NoSQL Databases: MySQL vs MongoDB. In: *2020 International Conference on Computer Communication and Informatics (ICCCI -2020)*, 22-24 January 2020, Coimbatore, India. IEEE.

‌