-ORIGINAL ARTICLE-

Data Science & Engineering into Food Science: A novel Big Data Platform for Low Molecular Weight Gelators' Behavioral Analysis

Ciencia de Datos e Ingeniería en Ciencia de Alimentos: una Plataforma novedosa de Big Data para Análisis de Comportamiento de Gelantes de Bajo Peso Molecular

Verónica Cuello[1], María G. Corradini[2], Michael Rogers[2], and Gonzalo Zarza[1]

[1]*Insituto de Tecnología (INTEC), Universidad Argentina de la Empresa (UADE), Buenos Aires, Argentina*
{ccuello, gzarza}@uade.edu.ar
[2]*Department of Food Science & Arrell Food Institute, University of Guelph, Ontario, Canada*
{mcorradi, mroger09}@uoguelph.ca

Abstract

The objective of this article is to introduce a comprehensive end-to-end solution aimed at enabling the application of state-of-the-art Data Science and Analytic methodologies to a food science related problem. The problem refers to the automation of load, homogenization, complex processing and real-time accessibility to low molecular-weight gelators (LMWGs) data to gain insights into their assembly behavior, i.e. whether a gel can be mixed with an appropriate solvent or not. Most of the work within the field of Colloidal and Food Science in relation to LMWGs have centered on identifying adequate solvents that can generate stable gels and evaluating how the LMWG characteristics can affect gelation. As a result, extensive databases have been methodically and manually registered, storing results from different laboratory experiments. The complexity of those databases, and the errors caused by manual data entry, can interfere with the analysis and visualization of relations and patterns, limiting the utility of the experimental work. Due to the above mentioned, we have proposed a scalable and flexible Big Data solution to enable the unification, homogenization and availability of the data through the application of tools and methodologies. This approach contributes to optimize data acquisition during LMWG research and reduce redundant data processing and analysis, while also enabling researchers to explore a wider range of testing conditions and push forward the frontier in Food Science research.

Keywords: big data, data science, food science, colloidal science, self assembly, low molecular weight gelators.

Resumen

Este trabajo tiene como finalidad presentar una solución integral destinada a permitir la aplicación de metodologías de analítica y ciencia de datos de última generación a un problema relacionado con la ciencia de los alimentos. Dicho problema se refiere a la automatización de la carga, la homogeneización, el procesamiento complejo y el acceso en tiempo real a los datos de los gelantes de bajo peso molecular (LMWG por sus siglas en inglés) para obtener información sobre su comportamiento de ensanblaje, es decir, si un gel se puede mezclar con un solvente apropiado o no. La mayor parte del trabajo en el campo de la ciencia coloidal y de alimentos en relación con los LMWG se ha centrado en identificar los solventes adecuados que pueden generar geles estables, y evaluar cómo las características del LMWG pueden afectar la gelificación. Como resultado, se han registrado de forma metódica y manual extensas bases de datos, que almacenan los resultados de diferentes experimentos de laboratorio. La complejidad de esas bases de datos, y los errores causados por la entrada manual de datos, pueden interferir con el análisis y visualización de relaciones y patrones, limitando la utilidad del trabajo experimental. Por los motivos antes mencionados, hemos propuesto una solución de Big Data escalable y flexible para permitir la unificación, homogeneización y disponibilidad de los datos mediante la aplicación de herramientas y metodologías de datos. Este enfoque contribuye a optimizar la adquisición de datos durante la investigación de LMWG y reduce el procesamiento y análisis de datos redundantes, al tiempo que permite a los investigadores explorar una gama más amplia de condiciones de prueba y avanzar la frontera en la investigación de la ciencia de los alimentos.

Palabras claves: big data, ciencia de datos, ciencia de los alimentos, coloides, auto-ensamblado, gelantes de bajo peso molecular.

1 Introduction

This study seeks to make available a Big Data architecture that can contribute to streamline research in Food Science, particularly in the area of low molecular weight gelators (LMWGs). LMWGs are small molecules that in contact with a solvent can form a molecular gel. Molecular gels are soft materials that can be potentially be introduced into foods to generate novel products, such as healthier butter-like spreads or low caloric foods.

The resulting data platform could drive real-time data acquisition and processing and automation of manual activities to analyze LMWGs behavior, enabling opportunities for improving their performances and even providing feedback during experimental activities. To this end, Data Science into Food Science is a multidisciplinary research project that fosters the collaboration of food and data scientists, and engineers at the Technology Institute in Buenos Aires [1]

Rogers et al. have systematically produced and collected data related to a variety of LMWGs. [2]. These databases were manually generated, following specific traditional protocols that records inputs and outputs of an experiment. In this particular case, the major inputs are the components used, i.e. LMWGs and solvents, and the outputs the systems that their combination produces, i.e. solution, gel, or precipitate. A gel can be defined as a semi-solid colloidal dispersion of a liquid entrapped on a solid network (like jelly). A solution is a homogeneous mixture composed of at least one solute, in this case the LGWM, dissolved in another substance, the solvent. In this context, the term precipitate refers to a system presenting two distinct visible phases, a solid at the bottom and a solution on top.

In addition to the above mentioned information, each database also includes complementary data regarding potentially relevant physicochemical properties solvents. These properties are significant to this proposal since they are, to different extents, related to the outcome of each LMWG-solvent combination, namely solution, gel, or precipitate. The databases are concrete and full of interest parameters, but they are split into several structures; that is why, so far, they have been partially analyzed, and only limited information and correlations have been obtained.

To improve the utility of these actual databases and experimental data, we propose to apply a set of Big Data solutions, specifically designed to face this kind of data problems where there is a requirement to efficiently deal with at least one of the three main data attributes: volume or amount of data, variety in the source or format of the data, and speed of generation or consumption of data [3]. Although the most relevant attributes of this project are the variety and format of the data, the proposed solution will also enable processing huge volumes of data as well. Consequently, the Big Data Platform will enable scientists to exploit the potential of currently (and future) available data.

This paper is organized as follows. First, the motivation of our proposal and the state of the art of Food Science data is detailed. Then, a brief analysis on how a Big Data solution can contribute to exploit the potential of the data, is presented in Section 2. The proposed design and its components are described in Section 3. The platform and the preliminary results along with the next steps are introduced in Section 4. Finally, conclusions are drawn in Section 5.

1.1 Low molecular weight gelators

Molecular gels are formed by low molecular weight organic compounds, i.e. LMWG. The resulting gels are the result of the assembly of highly organized architectures through different types of chemical non-covalent bonds. The formation of a molecular gel depends on the solubility of the gelator in a solvent, and hence on the solvent characteristics. The ability of a LMWG to form a gel relies on a delicate balance of interactions between its own molecules and the solvent. The LMWG should be soluble in the solvent to avoid precipitation, but not excessively soluble so that a solution, instead of a gel, is formed. The selection of the solvent is, thus, crucial to establish enough gelator-gelator and gelator-solvent to promote the formation of a network capable of entrapping the solvent. The resulting gels have practical application in fields like food, medicine, and other industries. For example, in medicine and foods, molecular gels can be used as delivery systems to transport drugs and bioactive compounds, e.g., vitamins, respectively. Additionally, in foods their utilization as fat mimetic, systems that can produce similar sensations as fat carrying less calories, are being studied Among environmental uses, LMWGs have been proposed to control oil spills since they can gel in contact with oil reducing their spread and facilitating their collection.

1.2 The unsolved question

The ability of a LMWG to gel a solvent has been empirically explored, mostly using a trial and error method. The diversity of forces acting on both gelators and solvent determine the complexity of the system and the inability to accurately predict *a priori* the outcome of a gelator-solvent combination. Thus, time-consuming experiments with tens of gelators often results in no *one gel*. However, studies of solvent properties and gelator structures have advanced a better understanding of why certain gelators form gels in specific solvent types and others do not [4].

1.3 State of the art of Food Science data

The data sources considered in this work were assembled from the literature or directly collected by the

Rogers' lab at the University of Guelph. The compiled databases were comprised of published (and self-collected) information from leading research groups in LMWGs [2]. Such datasets have been used to populate Microsoft Excel spreadsheets. These files have a mostly unified structure among them, similar to the one exemplified in Fig. 1, including entries for each combination of solvent and gelator: solvent identification, gelator identification, enumeration and description of solvent properties, result (gel, precipitate, or solution). Although the structure of the tables was similar across all cases, they had a set of differences and anomalies due to the manual data entry origin.

2 What Data Science & Engineering can do for Food Science

The data provided in compiled spreadsheets on LMWGs showed individually (for each solvent and gelator combination) the outcome obtained, i.e. a solution, a gel, or a precipitate. This way of handling and processing data makes difficult to analyze trends and common patterns. In contrast, if all the inputs and outputs of the gelation process could be homogenized and analyzed together, researchers would be able to discover such trends more effectively.

The motivation of this study is to enable gathering and exploiting all these heterogeneous data sources in a unique, homogeneous, reliable and scalable platform to exploit their potential in order to unravel the mechanisms of LMWGs. Not only will this sallow unifying current and newly obtained data sources but also it will efficiently provide additional -and most sophisticated- computing solutions for this complex datasets. This enablement may lead, for example, to the discovery of relationships, conditions, or patterns of gelators behavior.

To address the questions derived from the Data Science into Food Science project, a Kappa Architecture has been proposed. This Big Data Architecture allows a seamlessly orchestration of the consumption, processing, and availability of current and future data sources for its exploitation, in a simple and scalable way. Over the last few years, Big Data solutions were commonly based on Lambda and Kappa Architectures [3]. The Kappa Architecture appeared in 2014 with Jay Kreps, who identified some weaknesses of the Lambda Architecture and and evolution to solve them, which he called Kappa Architecture [5].

3 The Food Science Big Data Platform

In the previous sections, we have introduced some bottlenecks and inefficiencies in the Food Science research methodology that could be tackled by introducing (and applying) some of the Data Science & Engineering best practices and techniques. In order to address this challenge, we have designed and implemented a Big Data Platform to overcome the aforementioned issues, as shown in Fig. 2.

This work has been carried out within the frame of the *Data Science into Food Science* research project of the Universidad Argentina de la Empresa (UADE) by a multidisciplinary team composed by Data Scientists and Engineers, Food Scientists, and Chemistry Scientists [6]. The four main parts of the Big Data Platform, as shown in Fig. 3, are:

1. **Data Ingestion**: the solution begins through the consumption of the data available in the *de-facto* standard in the field of Food Science, Microsoft Excel spreadsheets. In addition, the modular design of the platform has been conceived to enable the seamless addition of advanced ingestion sub-modules, tailored design to extract data from laboratory and industry equipment.

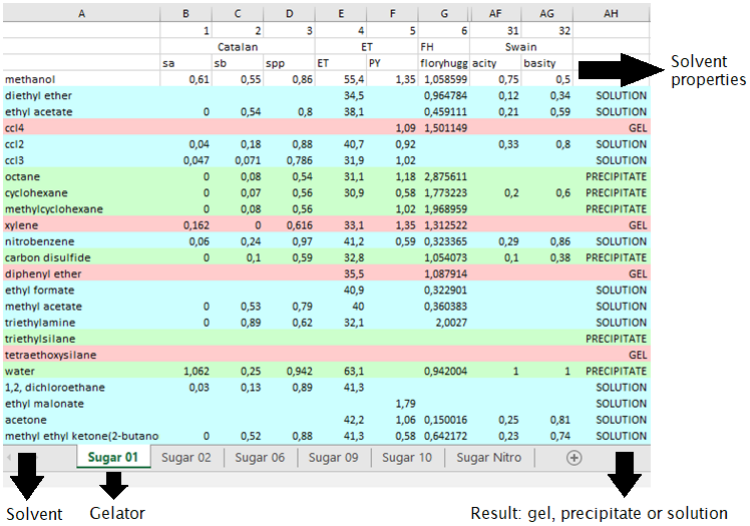
2. **Data Staging**: the data obtained is stored in a specifically designed staging area within the platform in a way that allows to efficiently manipulating, consolidating and transforming many records in various formats, which is then used for further processing. This module is specially relevant given the hard scalability and flexibility requirements that gave origin to the research project. The development of Food Science research fields explicitly stated the necessity of enabling the addition of new and highly heterogeneous data sources, ranging from manual research data entry up to results from advanced automatized laboratory equipment in real-time.

3. **Data Integration and Processing**: since the data sources could be arbitrarily large and diverse in terms of their data characteristics[1], they need to be pre-processed in order to prepare them for an efficient exploitation. This processing involves reading the data in the storage, processing (enhancing) them, and generating new outputs in different storage layers, including solutions designed to perform complex and time consuming batch processing and also real-time interactive visualization tools.

4. **Data Serving layer**: a subset of the data previously processed by means of a set of customized ETL (Extract, Transform, Load) routines is served through a high-performance repository in a pre-agreed structured format[2] to enable its access and querying using the current state-of-the-art Food Science analytical tools. Identifying

[1]The format and structure of Food Science data is not only constrained by the different research groups that originated them, but also from the proprietary data protocols from the wide set of laboratory and industrial equipment being used.

[2]Data is also available in non-structured formats for additional purposes and complex Data Science & Analytics processes.



	1	2	3	4	5	6	31	32	
	Catalan			ET		FH	Swain		
	sa	sb	spp	ET	PY	floryhugg	acity	basity	
methanol	0,61	0,55	0,86	55,4	1,35	1,058599	0,75	0,5	
diethyl ether				34,5		0,964784	0,12	0,34	SOLUTION
ethyl acetate	0	0,54	0,8	38,1		0,459111	0,21	0,59	SOLUTION
ccl4					1,09	1,501149			GEL
ccl2	0,04	0,18	0,88	40,7	0,92		0,33	0,8	SOLUTION
ccl3	0,047	0,071	0,786	31,9	1,02				SOLUTION
octane	0	0,08	0,54	31,1	1,18	2,875611			PRECIPITATE
cyclohexane	0	0,07	0,56	30,9	0,58	1,773223	0,2	0,6	PRECIPITATE
methylcyclohexane	0	0,08	0,56		1,02	1,968959			PRECIPITATE
xylene	0,162	0	0,616	33,1	1,35	1,312522			GEL
nitrobenzene	0,06	0,24	0,97	41,2	0,59	0,323365	0,29	0,86	SOLUTION
carbon disulfide	0	0,1	0,59	32,8		1,054073	0,1	0,38	PRECIPITATE
diphenyl ether				35,5		1,087914			GEL
ethyl formate				40,9		0,322901			SOLUTION
methyl acetate	0	0,53	0,79	40		0,360383			SOLUTION
triethylamine	0	0,89	0,62	32,1		2,0027			SOLUTION
triethylsilane									PRECIPITATE
tetraethoxysilane									GEL
water	1,062	0,25	0,942	63,1		0,942004	1	1	PRECIPITATE
1,2, dichloroethane	0,03	0,13	0,89	41,3					SOLUTION
ethyl malonate					1,79				SOLUTION
acetone				42,2	1,06	0,150016	0,25	0,81	SOLUTION
methyl ethyl ketone(2-butano	0	0,52	0,88	41,3	0,58	0,642172	0,23	0,74	SOLUTION

Figure 1: Sample from a real data source (food science spreadsheet).

the relevant data, and how it should be served, was one of the main tasks at the beginning of the project.

The requirements and challenges originated from the above-mentioned modules, have played a mayor role in the evaluation and design of the resulting architecture. The selection of the Data Platform Big Data tools to be introduced in Section 4 was driven by two prevailing factors. On one hand, the need of efficiently processing state-of-the-art Food Science data sources, characterized by their non-uniform structure and their usual data sparsity. Despite the fact that such problems could be addressed with certain degree of efficiency by SQL-based solutions, adopting a hard relational model is far from being an optimal solution and it also imposes specific scalability constraints given the fixed relational nature of SQL-based data models. On the other hand, the necessity of increasingly and continuously enhancing the Food Science datasets by including a wider range of newer non-relational and highly-heterogeneous data sources denotes the need of a flexible and modular solution that could effectively address both, relational and non-relational (or even streams-based) processing.

3.1 Data Integration and Processing

The Data Integration and Processing phase performs the computation that transforms the incoming data in order to fulfill the pre-agreed structure, with a special emphasis on the integration, uniformity, structure and cleanliness of the resulting data set. This process is divided into 3 main phases:

- **Gelators, solvents and their properties alignment.** Because the name of the gelators and solvents, and their properties, may vary for various reasons (for example, the existence of different denominations of the same compound, or typing errors) in this step the solution introduces homogenization to this master data. To accomplish this objective, mapping structures were built containing frequent names for the most commonly used solvents and gelators in the context of Data Science into Food Science project. This mapping of data denominations is not restrictive. It seeks to homogenize and improve the name of those solvents that have some error or deviation, but does not exclude from the solution those records that were not identified in the mapping. In these cases, the name of the new solvent, property or gelator is added to the solution as reported in the data source, without alterations. This allows users to progressively incorporate experiences with new solvents or gelators without needing to adapt the mapping. But, if they want to add the new denomination discovered to the mapping, it is possible; and doing that will immediately take effect enabling new incoming compounds to be automatically adapted into the solution (whether they are solvents, solvent properties, or gelators), ensuring homogeneity.
- **Improvement of solvent properties database.** This improvement only happens in the case of increasing precision in numerical values; any other difference is considered an exception. If a solvent property in the incoming data is reported to be more accurate than the available value for the same property in the model, the property is automatically updated, storing the new value.
- **Exception Handling.** As a result of the data processing, exceptions were identified (unexpected data or combination of data). For those exceptions, the data is specially classified within the model, with the aim of being evaluated in the fu-

ture and act upon requirement. For this purpose, an entry in a log structure is generated for each exception detected, indicating what type of exception is and how data is affected, which allows a detailed examination of the record after processing in order to improve data quality. There are two types of exceptions in this processing: informative exceptions and blocking exceptions. The informative exceptions are intended to report an inconvenience that may require attention to proactively improve the performance of the solution in the future, but does not stop the data processing (in this case, the data that produced the exception is not excluded from the results). Examples of informative exceptions are unidentified gelators, unidentified solvents, and solvent properties with doubtful value. The blocking exceptions reactively determine that there is an anomaly in certain data that may deteriorate the integrity, and therefore it is important to exclude such anomalous data from the processing. Examples of blocking exceptions are contradictory solvent properties, contradictory results, and lack of results.

3.2 Data Serving Layer

The Data Serving layer is based on a dimensional model, as shown in Fig. 4, containing the facts, the dimensions, and the exceptions for the solution. A dimensional model is a type of data design that is oriented to face queries, based on comprehensibility and performance.

A fact table is the main table in a dimensional model, where the performance measurement of the business process is stored, without duplication. The term fact is used to represent a measure, taken at the intersection of all dimensions. This list of dimensions is what defines the granularity of the fact table. In this work, the fact table is the Results table shown in Fig. 4. The most common facts are numerical and can have aggregations (for example: searching for the sum, the minimum value, the maximum, or the average of a given value). However, it may be possible for a measured event to be textual (although it is rare). When the measure is textual, it turns out to be the description of something, and is extracted from a discrete list of values. In this design, the measure is textual, because it is reduced to the result of experiences between gelators and solvents, and can only take a discrete value within the following enumeration: gel, precipitate, or solution.

The dimensions go along with the table of facts, containing the textual descriptors of the different attributes that enable understanding the measures and the business process integrally. Above all, the dimensions are the entry points in the fact table, offering possibilities of "cutting" the data [7]. An example of a dimension in this model is the solvents data, among others.

As part of the data that helps describing the business process, exceptions were integrated as dimensions in the data repository. Informative exceptions, optionally, may be part of a fact. On the other hand, the blocking exceptions are not part of a fact because its presence drives to the refuse of the data. However, for control purposes, the blocking exceptions are stored, detailing the date of occurrence and the data that caused the exception. The tables that stores the exception data are Informative Exceptions and Blocking Exceptions.

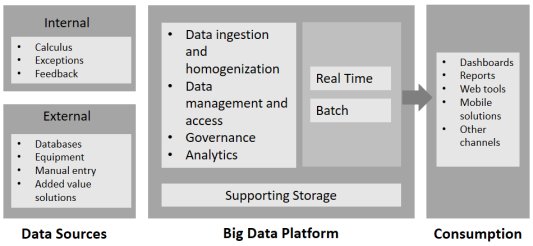


Figure 2: Big Data Platform Capabilities.

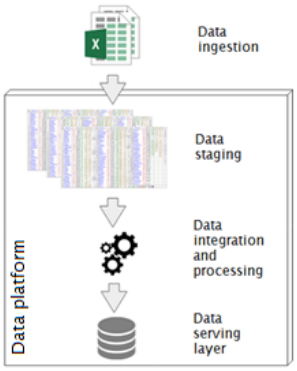


Figure 3: The four phases of the proposed solution.

4 Data Platform Architecture

Since the development of the Food Science Big Data Platform is framed within the Data Science into Food Science research project, there are several interdependencies between the different teams involved in the project. To overcome this limitation and also to reduce the overall development time, we have followed an agile development methodology organized around the release of a series of well-defined Minimum Viable Products (MVP) [8]. In order to do this, we have followed the incremental three-step development approach proposed by López Murphy and Zarza [3]: Proof-of-Concept (POC) to Prototype to MVP.

To carry out the first MVP iteration of the proposed architecture, a PAAS (Platform As A Service) solution was chosen, allowing a quick start-up, without the implications of the investment and management of a hardware and software infrastructure at the beginning of the project. Additionally, this decision enabled flexibility to scale and continue extending the solution

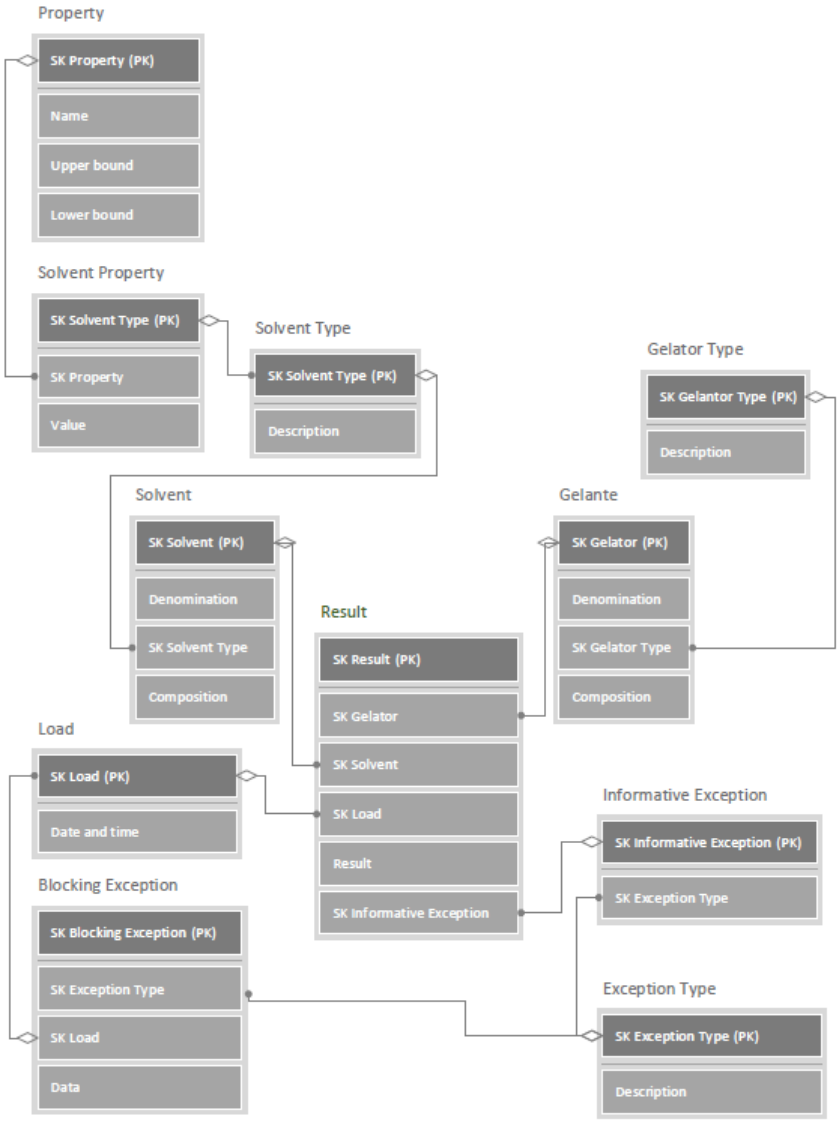


Figure 4: Dimensional Data Model.

to the future. As stated by Cearley [9]: "Organizations that do not have a high-level strategy in the cloud, driven by their business strategy, will significantly increase the risk of failure and lost investment".

Regarding PAAS providers, Amazon Web Services (AWS) leads the industry in offering the widest range of functionality and maturity in PAAS solutions, with strong proposals for large organizations, and large infrastructure. Additionally, AWS has a solution called AWS Educate, chosen to implement the MVP based on the Data Platform architecture introduced in this work. The AWS Educate program [10] was launched in 2015 aimed on providing students and educators the necessary resources to accelerate learning efforts related to cloud solutions.

AWS was the PAAS solution provider chosen for implementing this platform, after performing a detailed comparative exercise with its main competitors: Google Cloud Platform and Microsoft Azure.

Nowadays, state-of-the-art Big Data solutions are commonly developed on distributed storage systems, where HDFS (Hadoop Distributed File System)[3] is the most widely used solution, and was the chosen one for this work, as shown in the dataflow depicted in Fig. 5 [11]. HDFS is a distributed, scalable and portable file system, implemented in Java, which was initially designed to be used as part of the Hadoop framework. Apache Hadoop is an open source framework that enables distributed processing of large data sets in computer groups, using simple programming models and affordable commodity hardware [12]. It was designed to scale from individual servers to thousands of machines, where each one offers local computing and storage.

In order to perform the data preparation and processing within the proposed Data Platform, we have used

[3]More information available at https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html.

Apache Hive, a data storage tool built on Hadoop to provide grouping, query, and analysis of large data sets [13]. Hive supports the Hadoop file system (HDFS) and the Amazon S3 file system [14], and offers an SQL-based query language to read and manipulate queries. Apache Hive was primarily used to query data from the dimensional model. Amazon S3, an object storage service designed to store and retrieve in the cloud any volume of data, from any location, is focused on offering very high durability and offers comprehensive compliance and safety capabilities that meet the most severe requirements. It provides flexibility to manage data in relation to cost optimization, access control and compliance.

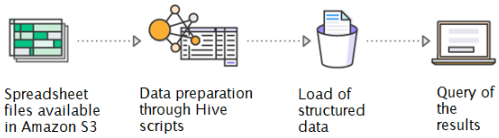


Figure 5: Data flow within the proposed solution.

4.1 Preliminary results and next steps

At the time of writing this paper, the first MVP iteration of the platform is up and running on AWS instances. It is being used to process data from the Food Science into Data Science project. Like never before, records consumed from the original data sources (spreadsheets) are classified to enhance and ease the result of the chemical process, for instance whether the combination resulted in a gel, a precipitate, or a solution, as exemplified in the output shown in Fig. 6.

In a context where all the analyses that were done before were null or partial, completely handmade, the Food Science team now has a tool that, even from the MVP, collaborates in its work.

From now, iterations will be focused on improving the scope of the solution and getting retrospective insights from the outcoming Data and the Food Scientist, primarily to improve the overall accuracy of Data Science models.

5 Final Discussion

In this proposal, the scientific method has been used as a framework to provide insights into a withstanding problem in colloidal/food science and understanding the behavior of LMWGs. The scientific method follows these steps [15]: place the questions in the context of existing knowledge (#1); formulate a possible answer (#2); deduces consequences and predictions (#3); test the hypotheses in the specific field for which it has been raised (#4), demonstrating that the solution is useful and shows progress against the existing landscape; when the solution gets consistency, it is confirmed (#5); and, finally, the theory obtained may again be the subject of new questions (#6).

Following the steps of the scientific method, the first activities of this work were directly focused on understanding the chemical concepts, the actual needs, and the opportunities for improvement (#1). Then, based on that understanding, a data architecture was designed (#2) that aimed to solve each of the needs detected, without losing focus on scalability to allow the continuity of the project (#3).

The proposed solution showed enablement of the incorporation of existing data in the Data Science into Food Science project to the designed data platform, allowing the processing and improvement of that data, and structuring the information in a way that is distinguished by its classification, homogeneity, and consistency (#4). In this way, the work generates value and maximizes the potential of the data, making them available for use, either for study or for incorporating them into other IT solutions (#5).

During the development of this proposal, some outstanding opportunities for evolution and growth of the proposed solution were noted (#6). Regarding solvents and gelators denomination and properties, there is an opportunity for implementing visual tools that allow final users to frequently enter valuable data to feed (and feedback) the platform. Besides, data than nowadays come from spreadsheets could be directly consumed from their source (a laboratory instrument, or a calculus performed by another tool). Regarding laboratory instruments, each equipment provides only one value; so it will be needed to compile all the relevant data for LMWG. The platform presented by this paper will help to this goal providing a scalable way to be directly updated from different data sources

Finally, since the original question to answer was what gelator-solvent combinations would result in a gel, the more important challenge is to design a Data Science solution that helps predicting what are the most likely cases to try in the laboratory.

Competing interests

The authors have declared that no competing interests exist.

Authors' contribution

VC and GZ designed and developed the solution, and revised the manuscript. MC and MR conceived the idea, conducted the experiments and analyzed the results. All authors read and approved the final manuscript.

Acknowledgements

We are grateful for the funding from INTEC at Universidad Argentina de la Empresa (UADE) for the project Data Science into Food Science (P17T04). We would also like to acknowledge support from all project team members, specially Bibiana Rossi, Marta Gozzi, Cinthia Santo Domingo, and Tomás Tecce.

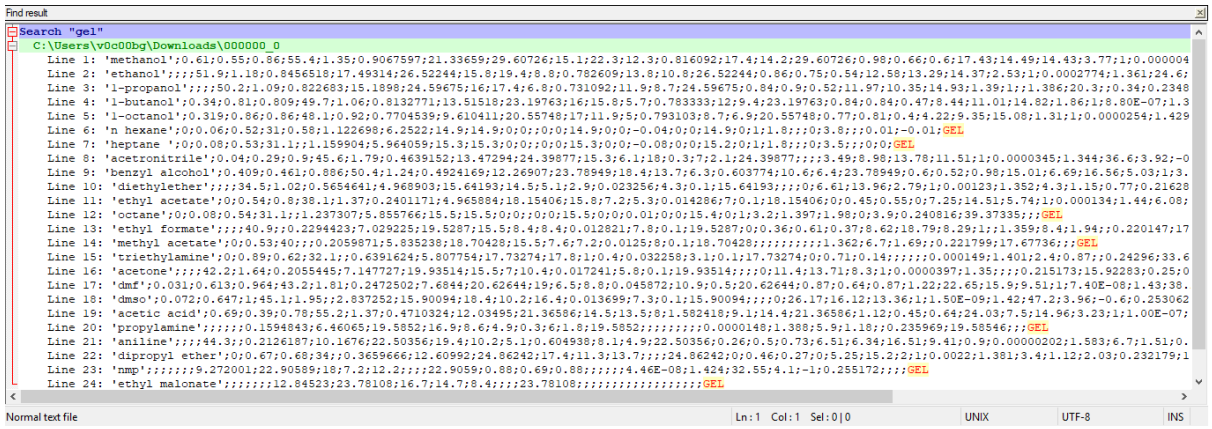


Figure 6: Example of a real raw data extract obtained from the online ETL process.

References

[1] Universidad Argentina de la Empresa (UADE), "Instituto de Tecnología (INTEC) - Proyectos," 2020. [online; accessed on March 2020].

[2] M. A. Rogers, Q. Feng, V. Ladizhansky, D. B. Good, A. K. Smith, M. Corridini, D. A. S. Grahame, B. C. Bryksa, P. D. Jadhav, S. Sammynaiken, L.-T. Lim, B. Guild, Y. Y. Shim, P.-G. Burnett, and M. J. T. Reaney, "Self-assembled fibrillar networks comprised of a naturally-occurring cyclic peptide–LOB3," *RSC Advances*, vol. 6, no. 47, pp. 40765–40776, 2016.

[3] J. L. Murphy and G. Zarza, *La ingeniería del Big Data, Cómo trabajar con datos*. Barcelona: Editorial UOC, Oct. 2017.

[4] Y. Lan, M. G. Corradini, R. G. Weiss, S. R. Raghavan, and M. A. Rogers, "To gel or not to gel: correlating molecular gelation with solvent parameters," *Chemical Society Reviews*, vol. 44, no. 17, pp. 6035–6058, 2015.

[5] J. Kreps, "Questioning the Lambda Architecture," July 2014. [online; accessed on March 2020].

[6] C. Santo Domingo, M. Gozzi, G. Zarza, T. Tecce, M. Rogers, and M. Corradini, "Data Science Computational Strategies in Food Science and Engineering," tech. rep., Instituto de Tecnología (INTEC), UADE, 2018.

[7] R. Kimball and M. Ross, *The Data Warehouse Toolkit*. Wiley John & Sons, 2013.

[8] E. Ries, "Minimum Viable Product: a guide," Aug. 2009. [online; accessed on March 2020].

[9] D. Cearley, "Advance Cloud Computing Capabilities," 2018. [online; accessed on March 2020].

[10] Amazon, "AWS Educate," 2015. [online; accessed on March 2020].

[11] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, IEEE, may 2010.

[12] Apache Software Foundation, "Apache Hadoop," 2006. [online; accessed on March 2020].

[13] Apache Software Foundation, "Apache Hive," 2011. [online; accessed on March 2020].

[14] Amazon, "Amazon S3," 2020. [online; accessed on March 2020].

[15] G. Dodig-Crnkovic, "Scientific methods in computer science," in *Proceedings of the Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, Skövde, Suecia*, pp. 126–130, 2002. [online; accessed on March 2020].

Citation: V. Cuello, M. Corradini, M. Rogers and G. Zarza. *Data Science & Engineering into Food Science: A novel Big Data Platform for Low Molecular Weight Gelators' Behavioral Analysis*. Journal of Computer Science & Technology, vol. 20, no. 2, pp. 72–79, 2020.
DOI: 10.24215/16666038.20.e08.
Received: April 1, 2020 **Accepted:** September 9, 2020.