# Towards a Big Data and Data Science Vision at Adventist Media Ministries

Germán H. Alférez, Ph.D., *Professor, School of Engineering and Technology, Montemorelos University*

**Abstract**—Netflix and other media corporations use Big Data and Data Science as two strategic assets. Media companies strongly depend on data to understand their customers and to create empathy with them. In fact, the lack of a strong Big data and Data Science Vision can lead to: 1) an inefficient or insufficient Big Data computing infrastructure that cannot handle the volume, velocity, and variety of audience-related data to be analyzed; and 2) isolated data analysis components that barely indicate the activities, roles, and tools that are required for data management. The contribution of this document is to present why a vision for Big Data and Data Science is necessary in Adventist media ministries. Future work will cover the definition and materialization of the vision.

**Index Terms**—Big Data, Data Science, Machine Learning, Seventh-day Adventist Church, Media Ministry, Architecture, Methodology, Netflix

## 1 Introduction

Big Data has the ability to change the nature of a business. In fact, there are many firms whose sole existence is based upon their capability to generate insights that only Big Data can deliver. This is specially true in media companies, which understand that Big Data is not just about technology—it is also about how these technologies can propel an organization forward. Big Data initiatives are strategic in nature and should be business-driven. The adoption of Big Data can be transformative but is more often innovative. Big Data is not a technology related to business transformation; instead, it enables innovation within an enterprise on the condition that the enterprise acts upon its insights. Moreover, Big Data is not simply "business as usual," and that the decision to adopt Big Data must take into account many business and technology considerations [1].

In order to take advantage of Big Data, it is necessary to count on solid mechanisms for data analysis. This foundation is given by Data Science, which is the science about data [2]. Since Big Data and Data Science are not trivial in nature, there are many aspects that need to be considered and planned from the very beginning. Otherwise, Big Data and Data Science could be used just as buzzwords without knowing their real implications in business operations.

Recently, the Seventh-day Adventist (SDA) Church has started to recognize the importance of Big Data and Data Science as two strategic assets. For instance, the North-American Division (NAD) has a department for Big Data. Recently, that department asked the School of Engineering and Technology, Montemorelos University, to analyze data related to church members, churches, and schools at Washington Conference. Moreover, the relevance of Big Data and Data Science to support the Adventist mission has been evident in several presentations. For instance, at the NAD's Presidents of Large Conferences Retreat (2017) and at the Global Adventist Internet Network (GAiN) forum (2016). A complete list of international presentations in these areas is available at www.harveyalferez.com/relevant_presentations. Also, at the Adventist education system, the School of Engineering at Technology has played a key role in guiding Loma Linda University and the Peruvian Union University with research projects and courses related to Big Data and Data Science.

In addition to church-related data projects, the School of Engineering and Technology, Montemorelos University, has worked on projects focused on solving relevant problems in health, smart cities, geoscience, sentiment analysis, autonomous systems, gesture analysis, and self-driving vehicles. These projects make a strong use of Data Science, Artificial Intelligence, and Software Engineering. A comprehensive list of the resulting top-tier scholar publications is available at www.harveyalferez.com/publications.

Adventist media ministries can use Big Data and Data Science to understand their audiences and be more effective in reaching them. As any engineering project, the application of Big Data and Data Science requires a strong vision. Such a vision will help Adventist media ministries to see the "big picture" to acquire, process, analyze and visualize data. This vision is a necessary step before starting any software development or hardware configuration to manage data. The lack of this vision will lead to the following problems: 1) an inefficient or insufficient Big Data computing infrastructure that will fall short to handle the volume, velocity, and variety of audience-related data to be analyzed; and 2) isolated components for data analysis will barely indicate the activities, roles, and tools that are required to materialize a strong data management vision. In order to avoid these situations, the contribution of this document is to present a preliminary work that shows why a strong vision for Big Data and Data Science is strategical for Adventist media ministries. This document does *not* presents the vision itself but prepares the path towards the definition and material-

- *Germán H. Alférez: School of Engineering and Technology, Montemorelos University, Montemorelos, Mexico*
  *E-mail: harveyalferez@um.edu.mx, Website: www.harveyalferez.com.*

ization of this vision.

This document is organized as follows. The second section presents the essentials of Big Data and Data Science. The comprehension of these concepts is key to build a common language among stakeholders at Adventist media ministries. The third section presents how Netflix is using Big Data and Data Science. This successful case study can help Adventist media ministries realize the impact of data analysis in media companies. Finally, this document presents arguments about the need for future steps to define and materialize a vision for Big Data and Data Science at Adventist media ministries.

## 2 Big Data and Data Science 101

Big data and Data Science are concepts commonly used nowadays in the news or in meetings. But what do they mean? In order to answer these questions, this section describes the generalities of Big Data and Data Science. Also, it describes why a fully incremental architecture is not enough for Big Data.

### 2.1 What is Big Data?

Big data is a term that can be used to describe datasets so large and complex that they become difficult to work with using standard techniques [3]. The digital universe is huge, doubling in size every two years. By 2020 it will reach 44 zettabytes, or 44 trillion gigabytes [4]. This fact has motivated companies and scientists around the world to find new ways to understand Big Data in the digital universe. Big Data is definitely the next big thing, so much so that people are saying Big Data is the new oil [5].

There are tons of Big Data that the church can dig into, both internally and externally. In the case of external data, the church could make use of open data, which is data that anyone can access, use or share [6]. For example, Data.gov offers a lot of ready-to-use open datasets offered by the U.S. Government. By means of Data.gov, church management, pastors, and members could find datasets related to health, education, and much more.

Big data opens new opportunities for the SDA Church. For instance, a previous research work published on Adventist Review shows how Big Data was used to understand the people's perceptions about the church's fundamental beliefs [7]. The dataset used in the experiments of that research work was composed of digitized texts containing about 4 percent of all books ever printed between 1800 and 2008.

#### 2.1.1 *The 4 Vs of Big Data*

Big datasets tend to be more unstructured, distributed, and complex than ever before. At the level of business, data generated by business operations, can be analyzed to generate insights that can help the business make better decisions. This makes the business grow bigger, and generate even more data, and the cycle continues. This is represented by the blue cycle on the top-right of Figure 1.

On the other level, Big Data is different from traditional data in every way: space, time, and function. The quantity of Big Data is 1,000 times more than that of traditional data. The speed of data generation and transmission is 1,000 times faster. The forms and functions of Big Data are 10 times more diverse: from numbers to text, pictures, audio, videos, Web logs, machine data, and more. There are many more sources of data, from individuals to organizations to governments, using a range of devices from mobile phones to computers to industrial machines. Not all of Big Data will be of equal quality and value. This is represented by the cycle on the bottom left of Figure 1.

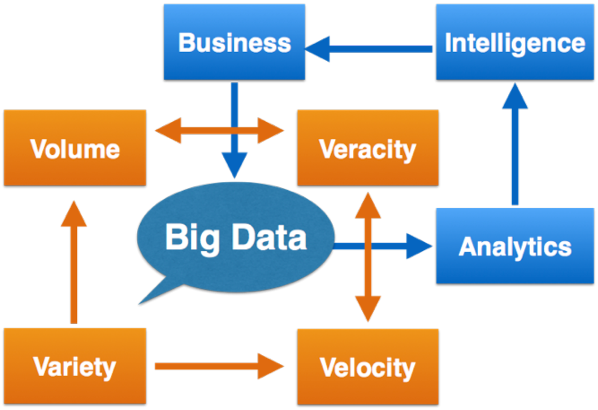

Fig. 1. Big Data context [8]

As depicted in Figure 1, the most relevant characteristics of Big Data can fall into four dimensions [2], [9], [8]:

1) the *volume* of information that systems must ingest, process, and disseminate. On one hand, traditional data can be measured in Gigabytes and Terabytes. On the other hand, Big Data is measured in Petabytes and Exabytes (1 Exabyte = 1 million TB). The primary reason for the growth of data is the dramatic reduction in the cost of storing data (30-40% every year).
2) the *velocity* at which information grows or disappears. If traditional data is like a lake, Big Data is like a fast-flowing river. Big Data is being generated by billions of devices, and communicated at the speed of the light, through the Internet. Ingesting all this data is like drinking from a fire hose. One does not have any control over how fast the data will come. A huge unpredictable data-stream is the new metaphor for thinking about Big Data. There are two reasons for the increased velocity of data: 1) increase in Internet speed (from 10 MB/sec to 1 GB/sec – 100 times faster); and 2) increase variety of sources, such as mobile devices.
3) the *variety* of data sources and formats, which covers structured (e.g. databases), semi-structured (e.g. server logs), and unstructured (e.g. tweets from Twitter). According to experts, unstructured data accounts for 80% to 90% of enterprise data [10].
4) the *veracity* of uncertain data. Veracity relates to the truthfulness, believability, and quality of data. Big Data tends to be messy. Moreover, in some cases the source of information may not be authoritative (e.g. Wikipedia is useful, but not all pages are equally reliable).

If data were only growing too large, or only moving too fast, or only becoming too diverse, it would have been relatively easy. However, when the four Vs arrive together in an interactive manner, it creates a high complexity. While the volume and velocity of data drive the major technological concerns and the costs of managing Big Data, these two Vs are themselves being driven by the variety of forms and functions and sources of data. The varying veracity and value of data complicate the situation further [8].

### 2.2 What is Data Science?

Data Science can be defined as the study of the generalizable extraction of knowledge from data. Data Science seeks actionable and consistent pattern for predictive uses [11]. Data Science calls for multi-disciplinary approaches that incorporate theories and methods from many fields including mathematics, statistics, pattern recognition, knowledge engineering, machine learning, high performance computing, etc. Furthermore, Data Science is the science about data [2]. Broadly, Data Science is in the intersection of the areas described in Figure 2.

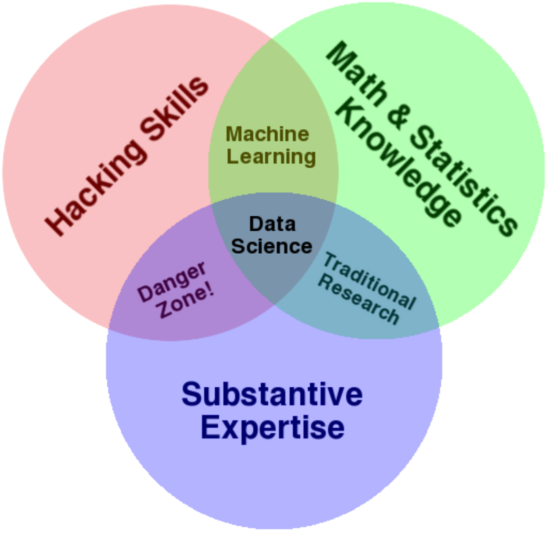

Fig. 2. The Data Science Venn Diagram [12]

In an article published in Harvard Business Review [13], Davenport and Patil state that the data scientist possesses the training and curiosity to make discoveries in data. He/she is a hybrid of data hacker, analyst, communicator, and trusted adviser. More than anything, what data scientists do is make discoveries while swimming in data.

It is important to mention that Data Science is not restricted to only Big Data. However, the fact that data is scaling up and the invention of new tools for Big Data analysis open a new era for Data Science [2].

A key element in Figure 2 is machine learning. Machine learning is a subfield of artificial intelligence that gives computers the ability to learn without being explicitly programmed. The term "machine learning" refers to the methods or algorithms that can be used as an alternative to traditional statistical methods. With traditional statistics, we define the model specification prior to working with the data. With machine learning, the model specification is defined by applying algorithms to the data (i.e., data-adaptive). With machine learning, few assumptions are made about the underlying distributions of the data.

An article in Adventist Review [14] describes how Data Science has been used to understand the needs of people in New York City. This metropolis has central significance in our church's ongoing Mission to the Cities project. Specifically, machine learning was used to analyze the sentiments of people from tweets related to several topics.

### 2.3 A Fully Incremental Architecture is Not Enough for Big Data

A typical system for data analysis is a Web analytics application. This application tracks the number of page views for any URL a customer wishes to track. The customer's Web page pings the application's Web server with its URL every time a page view is received. Additionally, the application should be able to tell at any point what the top 100 URLs are by number of page views.

At the highest level, a traditional architecture for this solution looks like Figure 3. What characterizes these architectures is the use of read/write databases and maintaining the state in those databases incrementally as new data is seen. However, there are several problems with fully incremental architectures, such as operational complexity (e.g. online compaction), extreme complexity of achieving eventual consistency, and lack of human-fault tolerance [15].

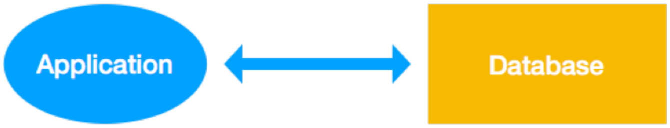

Fig. 3. Fully incremental architecture [15]

Database systems, such as relational databases, have been pushed to the limit. In an increasing number of cases these systems are breaking under the pressures of "Big Data." In fact, data management techniques associated with them, have failed to scale to Big Data.

Due to the problems of fully incremental architectures, new approaches have been taken by companies to try to deal with Big Data. These companies understand the complexity and scalability of Big Data. Not only must a Big Data system perform well and be resource-efficient, it must be easy to reason about as well. The properties of a Big Data system are described as follows:

- *Robustness and fault tolerance:* In the 90s, the idea for handling large volumes of data was to use one "big box" computer for processing and storage. However, this solution was expensive, it only supported a low volume of data (not big enough), and the hardware had to be "premium". In order to solve this, the idea is to use consumer-grade hardware (not "gold plated") for Big Data and to manage complexity in software. Nevertheless, building systems that "do the right thing" is difficult in the face of the challenges of distributed systems. Systems need to behave correctly despite machines going down randomly, the complex semantics of consistency in distributed databases, duplicated data, concurrency, and more. These challenges make it difficult even to reason about what a system is doing. Part of making a Big Data system robust is avoiding these complexities to easily reason about the system.

- *Low latency reads and updates:* The vast majority of applications require reads to be satisfied with very low latency, typically between a few milliseconds to a few hundred milliseconds. On the other hand, the update latency requirements vary a great deal between applications. Some applications require updates to propagate immediately, but in other applications a latency of a few hours is fine. Regardless, it is important to be able to achieve low latency updates when the company needs them in the Big Data system. More importantly, it is important to be able to achieve low latency reads and updates without compromising the robustness of the system.
- *Scalability:* Scalability is the ability to maintain performance in the face of increasing data or load by adding resources to the system. It is important to count on a horizontally scalable architecture across all layers of the system stack. Scaling is accomplished by adding more machines.
- *Extensibility:* Extensible systems allow functionality to be added with a minimal development cost. Oftentimes a new feature or a change to an existing feature requires a migration of old data into a new format. Part of making a system extensible is making it easy to do large-scale migrations.
- *Ad hoc queries:* Being able to do ad hoc queries on data is extremely important. Nearly every large dataset has unanticipated value within it. Being able to mine a dataset arbitrarily gives opportunities for business optimization and new applications. Ultimately, it is not possible to discover interesting things to do with data unless we can ask arbitrary questions of it.
- *Minimal maintenance:* Maintenance is the work required to keep a system running smoothly. This includes anticipating when to add machines to scale, keeping processes up and running, and debugging anything that goes wrong in production. The creation of autonomous architectures can help to prevent and fix problems without human intervention (a comprehensive list of previous work on autonomous systems is available at www.harveyalferez.com/publications).

In terms of database management systems, a new generation of scalable data management technologies has emerged in the last ten years. Relational Database Management Systems (RDBMS), which provide strong data-consistency guarantees based on vertical scaling of computer and storage hardware, are being replaced by NoSQL (variously interpreted as "No SQL", or "Not Only SQL") data stores running on horizontally-scaled commodity hardware. These NoSQL databases achieve high scalability and performance using simpler data models, clusters of low-cost hardware, and mechanisms for relaxed data consistency that enhance performance and availability [16]. Table 1 shows the differences between RDBMS and NoSQL.

According to the Consistency, Availability, Partition Tolerance (CAP) Theorem (also named Brewer's Theorem), it is impossible for a distributed computer system to simultaneously provide all three: 1) [data] consistency [all clients see the same data at the same time]; 2) [data] availability [guaranteed server response: success or failure]; and 3) partition tolerance [nodes/messages may fail/get lost/unreachable]. ACID focuses on consistency and availability. BASE focuses on partition tolerance and availability and throws consistency out the window. It sacrifices consistency to gain faster responses [17], [18].

It is also important to mention that traditional RDBMSs are not just insufficient in terms of Big Data but also in terms of data analysis. In fact, they "are not suited for knowledge discovery because they are optimized for fast access and summarization of data, given what the user wants to ask, or a query, not discovery of patterns in massive swaths of data when users lack a well-formulated query. Unlike database querying, which asks 'What data satisfies this pattern (query)?' discovery asks 'What patterns satisfy this data?' Specifically, our concern is finding interesting and robust patterns that satisfy the data, where 'interesting' is usually something unexpected and actionable and 'robust' is a pattern expected to occur in the future" [11]. This discovery process is known as Data Science.

## 3 How Netflix is Using Big Data and Data Science

It is interesting to see how traditional television viewership is on the decline [19], and fewer people are actually going to the movies [20]. Meanwhile, streaming video services like Netflix, Amazon's Instant Video, and Hulu keep adding subscribers and original programming. In Netflix, for instance, Big Data is used to optimize the quality and stability of video streams, and to assess customer entertainment preferences. In this way, Netflix can do a better job for targeting its users with offers for shows they might like to see [21]. A quick glance at the jobs page of Netflix can give an idea of how seriously data and analytics is taken in that company [22].

In order to have a technical view of the role of Big Data and Data Science at Netflix, its Big Data architecture is presented in Figure 4. This architecture is described in [23] as follows:

> "The infrastructure is executed entirely in Amazon's cloud domain. Netflix has divided computation to online, nearline, and offline parts based on different real-time requirements. Services in the online computation have requirements for maximum latency, when responding to client applications. The nearline computation is similar to the online computation with the exception that computed results can be stored instead of immediate serving of end users. The offline computation has most relaxed requirements for timing.
>
> End user interacts with Netflix by executing operations (e.g. Play, Rate etc.) in the user interface of the service. Recommendations are provided to the user based on other users' behavior.

TABLE 1
RDBMS vs. NoSQL

| *Feature* | *RDBMS* | *NoSQL* |
|---|---|---|
| Applications | Mostly centralized applications (e.g. Enterprise Resource Planning) | Mostly designed for the decentralized applications (e.g. Web and mobile) |
| Availability | Moderate to high | Continuous availability to receive and serve data |
| Velocity | Moderate velocity of data | High velocity of data (e.g. social media). Low latency of access |
| Data Volume | Moderate size; achieved after for a certain period | Huge volume of data, stored mostly for a long time or forever. Linearly scalable database |
| Data Sources | Data arrives from one or few, mostly predictable sources | Data arrives from multiple locations. Data is unpredictable |
| Data Type | Data is mostly structured | Structured or unstructured data |
| Rigor | Support ACID properties for transaction processing.<br>ACID: Pessimistic behavior: force consistency at end of the transaction.<br>*Atomicity:* all or nothing (of the *n* actions): commit or rollback<br>*Consistency:* transactions never cause inconsistent data<br>*Isolation:* transactions are not aware of concurrent transactions<br>*Durability:* acknowledged transactions persist in all events | Support BASE properties for approximate reporting.<br>BASE: Optimistic behavior: accept temporary database inconsistencies.<br>*Basically Available:* the data store is available all the time whenever it is accessed, even if parts of it are unavailable<br>*Soft-state:* it does not need to be consistent always and can tolerate inconsistency for a certain time period<br>*Eventually consistent:* after a certain time period, the data store comes to a consistent state |

> Recommendation algorithms require data, models, and signals as input. The data is previously processed information, which has been stored into a database. The signals are fresh information from online services. The models are comprised of parameters, which are usually trained initially offline, but are enhanced based on incremental machine learning. Events from end users are distributed via the Chukwa framework for the offline processing, or via user event queue for the nearline processing. Chukwa consists of agents, which transmit events in HTTP POSTs to collectors, which write data to HDFS file system. Manhattan is a distributed messaging system developed by Netflix. Hermes is a publish/subscribe framework, which is used for delivering of data to multiple subscribers in near real-time.
>
> Netflix has two streaming data sources: online data service and Netflix user events. The event-based data from Netflix users and signals from online data service are modeled as streaming data. The format of extracted streaming data is unknown. Chukwa agent can be considered as a stream extraction process, and Chukwa collector as a Stream temp data store. Hadoop HDFS can be modeled as a Raw data store. Execution of offline Pig jobs is modeled as Deep analytics. Hermes is a Publish & Subscribe store for storing and retrieving of offline analysis results.
>
> User data queue is modeled as a Stream temp data store. Data transformation by Manhattan is modeled as Stream processing for nearline computation. Intermediate analysis results are stored (processed streaming data) into Stream data stores (Cassandra, MySQL, EVCache). Online and nearline computation is modeled as a Stream analysis process. Algorithm service acts as a Serving data store for Netflix UI clients (End user application). Additionally, models are trained with machine learning algorithms."

According to Figure 4, Netflix makes a strong use of machine learning in different parts of the architecture. As mentioned in Section 2.2, machine learning is a key component of Data Science. This kind of approach has helped them to be successful in recommending content [24], [25], [26], [27], [28], optimize the stream quality of experience [29], A/B testing [30], and even to design covers [31]. One of the most successful cases of the application of Big Data and Data Science at Netflix was in the production of House of Cards. By analyzing viewer data – 30 million plays, 4 million ratings, 3 million searches – the company was able to determine that fans of the original House of Cards were also watching movies that starred Kevin Spacey and were directed by David Fincher, who is one of the show's executive producers. Netflix analyzes everything from when the audience watch a show to when they pause it or turn it off [32].

## 4 Need for a Vision in Big Data and Data Science at Adventist Media Ministries

The technological platforms of Adventist media ministries are growing in content and audience. Therefore, it is a must for them to count on solid platforms that support the volume, variety, and velocity of audience-related data. As described in Section 3, media companies such as Netflix have demonstrated that Big Data and Data Science are strategic assets. A Big Data and Data Science vision will help Adventist media ministries to create the software architectures and computer infrastructures to understand their

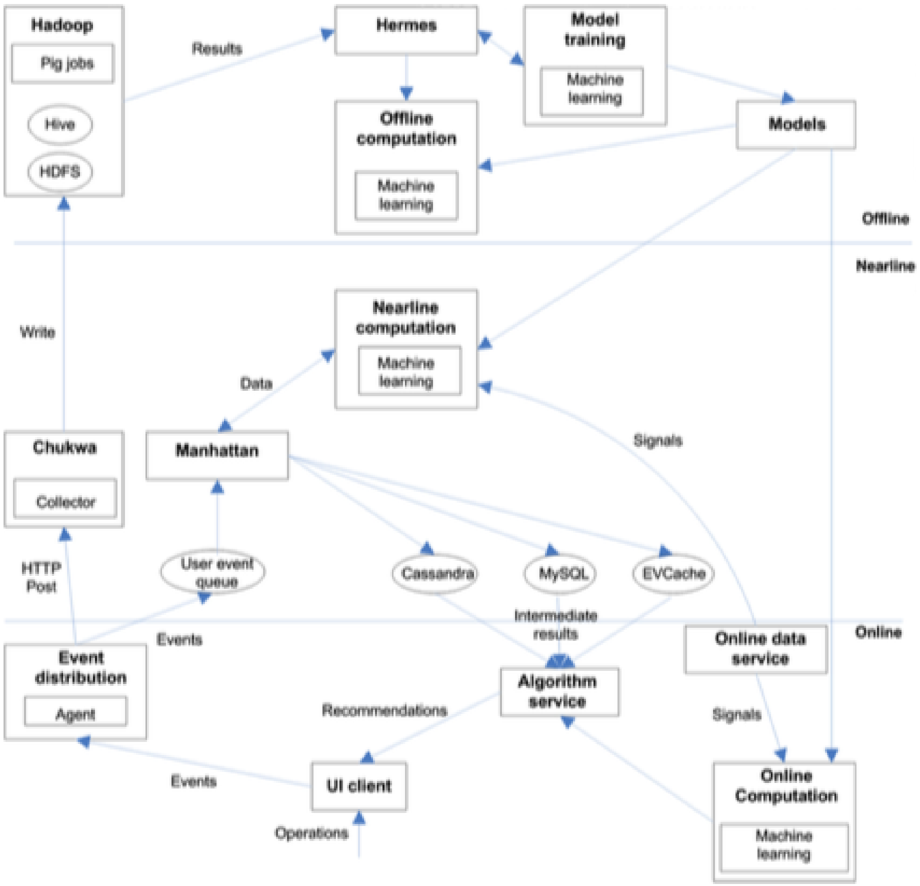

Fig. 4. Data analytics infrastructure at Netflix [23]

customers and improve their experience by providing the content they want. In fact, no organization can long survive without retaining and finding customers. In the context or Adventist media ministries, their audience is composed by their viewers or listeners. The process of retaining and finding new users begins by learning as much as possible about the audience.

Big Data initiatives are strategic in nature and should be business-driven. The adoption of Big Data can be transformative but is more often innovative. Innovation requires a shift in mindset because it will fundamentally alter the structure of a business either in its products, services or organization. This is the power of Big Data adoption; it can enable this sort of change [1].

In any engineering project, the most important part is the beginning, "the vision". In the visioning step the problem is understood and a solution is visualized. Therefore, a well-stablished data-management vision at Adventist media ministries is key for having success in the next steps: to set up the software architecture and the computer infrastructure, and to maintain and extend them.

A vision for Big Data and Data Science at Adventist media ministries needs to describe the components (architecture) and steps (methodology) to be carried out for implementing Big Data and Data Science:

- *Architectural View:* Define the Big Data software architecture for acquiring, processing, analyzing, and visualizing data related to/for the benefit of the media ministry. It is important to understand which components, both built in-house or commercial, are necessary to support data analysis in the near and far future. The architecture can be conceptualized in terms of software quality attributes (e.g. scalability, availability, performance, etc.).
- *Methodological View:* Propose the methodologies that can be followed to apply Data Science to 1) enhance user experience and 2) understand the audience. A methodological view describes how people roles, tools, and activities can be integrated to achieve particular solutions based on data analysis at different points of the process.

### 4.1 What is Next?

After defining a strong Big Data and Data Science vision for the Adventist media ministry, the next steps are as follows. It is important to mention that these tasks cannot be carried out just in one sequence (i.e., waterfall model). They have to be carried out iteratively and incrementally:

1) Purchase or build software to acquire data from several sources:
    - Social media
    - When users pause, rewind, or fast forward
    - What day they watch/listen content

- The date they watch/listen
- What time they watch/listen content
- Where they watch/listen (zip code)
- What device they use to watch (Do they like to use their tablet for TV programs and their Roku for movies? Do people access a particular TV program more on their iPads, etc.?)
- When people pause and leave content (and if they ever come back)
- The ratings given
- Searches
- Browsing and scrolling behavior
- The volume, colors, and scenery to find out what users like

Also, it is necessary to filter, clean, and aggregate the collected data. These steps are required before the analysis even occurs.

2) Create the Big Data software architecture for data collection (e.g. data streaming), analysis, and visualization.
3) Create (or rent) the Big Data computer infrastructure on which the Big Data architecture will run on.
4) Create machine learning software for: video/audio recommendation, automatic analysis of user behavior, analysis of market segments through clustering, gesture analysis, sentiment analysis, geolocation analysis, classification of people's needs, etc. Previous and current work demonstrate the potential of machine learning in these areas:

    a) In terms of the analysis of user behavior, there is strong evidence that demonstrate that machine learning can be used to automatize the process [33], [34], [35], [36], [37], [38], [39].
    b) Currently, at the School of Engineering and Technology of Montemorelos University, we are developing a tool that automatically analyzes and classifies the gestures of people who are watching videos by means of deep neural networks. The results will be available in May 2018.
    c) Cluster analysis involves finding groups in data. Cluster analysis takes many user-related variables and uses them to represent differences between users. When two users have similar values for variables like time of visit and location, they are seen as being similar. Cluster analysis looks at the differences among users–their distances from one another–to identify groups of users. For instance, in terms of discovering people's needs, cluster analysis was satisfactorily used in our previous work to understand segments of individuals who were removed from membership in the Inter-Oceanic Mexican Union Conference from 2005 to 2013 [40]. In addition, cluster analysis was used to discover mission-oriented patterns with open data in New York City [41] and to discover hidden patterns in US health-related open data [42]. Also, our preliminary results demonstrate the applicability of cluster analysis to discover hidden correlations among the listeners of Adventist World Radio.
    d) Sentiment analysis was satisfactorily used to understand the needs of people in New York City by analyzing their tweets [14].
    e) In our previous work, geolocation analysis was used for studying the distribution of church members, churches, and schools in the territory of Washington Conference. Moreover, it was used to analyze geochemical data in a joint research project with Loma Linda University [43].
    f) In terms of detecting and classifying people's needs, image recognition and artificial neural networks were used for the early detection of melanoma [44]. Also, open data and machine learning were used for the early classification of causes of maternal mortality in Mexico [45]. Currently, we are developing a tool for discovering the needs of people at the 10/40 Window with open data and machine learning. The results will be available in May 2018.

5) Maintain and extend the Big Data computer infrastructure and the Big Data software architecture. Machine learning algorithms need to be tried, deployed, and evaluated constantly. Otherwise, the generated models will not clearly reflect the current or future needs of the audience.
6) Periodically analyze and visualize the data collected by the machine learning software. This is a manual step.

## 5 Conclusion

This document presented the foundations towards a Big Data and Data Science vision in Adventist media ministries. The definition and materialization of this vision are parts of future work. As described in this document, implementing Big Data and Data Science in a company is not an easy task. However, this vision is a must for Adventist media ministries to avoid the creation and maintenance of inefficient and ineffective solutions for data analysis (e.g. the ones based on fully incremental architectures). This vision has to be defined from the architectural and methodological points of view. Beyond being just a technological vision, this is a strategical one. This vision needs to be permeated throughout the decisions and operations of the ministry, even on a daily basis. Most importantly, this vision will help Adventist media ministries to better understand their audiences and present the Gospel to them with empathy.

## References

[1] T. Erl, *Big Data Fundamentals: Concepts, Drivers & Techniques*. Prentice Hall, 2016.

[2] Z. Wu and O. B. Chin, "From big data to data science: A multidisciplinary perspective," *Big Data Research*, vol. 1, no. Supplement C, p. 1, 2014, special Issue on Scalable Computing for Big Data.

[3] C. Snijders, U. Matzat, and U.-D. Reips, “"big data": Big gaps of knowledge in the field of internet science,” *International Journal of Internet Science*, vol. 7, pp. 1–5, Jan. 2012.

[4] E. Corporation, “The digital universe of opportunities: Rich data and the increasing value of the internet of things,” 2014. [Online]. Available: www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

[5] P. Rotella, “Is data the new oil?” [Online]. Available: https://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/#7df09ede7db3

[6] O. D. Institute, “What is open data?” [Online]. Available: http://theodi.org/what-is-open-data

[7] G. H. Alférez, “Big data for reaching a big world,” *Adventist Review*, vol. 192, no. 11, pp. 47–51, 2015.

[8] A. Maheshwari, *Big Data Made Accessible*. Amazon Digital Services LLC, 2016.

[9] IBM, “The four v’s of data.” [Online]. Available: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[10] M. Wilczek, “From Big Data to good data: closing the gap between data governance and business insights.” [Online]. Available: https://www.bloomberg.com/professional/blog/big-data-good-data-closing-gap-data-governance-business-insights

[11] V. Dhar, “Data science and prediction,” *Commun. ACM*, vol. 56, no. 12, pp. 64–73, Dec. 2013.

[12] D. Conway, “The data science venn diagram,” 2010. [Online]. Available: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

[13] T. H. Davenport and D. Patil, “Data scientist: the sexiest job of the 21st century,” *Harvard Business Review*, vol. 90, no. 10, pp. 71–76, 2012.

[14] G. H. Alférez, “Tweeting in New York City - data science can teach us to sympathize,” *Adventist Review*, vol. 193, no. 2, pp. 47–49, 2016.

[15] N. Marz and J. Warren, *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications, 2015.

[16] I. Gorton, “Addressing the software engineering challenges of big data.” [Online]. Available: https://insights.sei.cmu.edu/sei_blog/2013/10/addressing-the-software-engineering-challenges-of-big-data.html

[17] P. Vanroose and K. V. Thillo, “Acid or base? - the case of nosql.” [Online]. Available: http://www.abis.be/resources/presentations/gsebedb220140612nosql.pdf

[18] K. Grolinger, W. A. Higashino, A. Tiwari, and M. A. Capretz, “Data management in cloud environments: Nosql and newsql data stores,” *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 2, no. 1, p. 22, Dec. 2013.

[19] J. Lupis, “The state of traditional TV: Updated with Q2 2017 data,” 2017. [Online]. Available: https://www.marketingcharts.com/featured-24817

[20] A. Levy, “Why 2015 wasn’t as good for theaters as the box office says,” 2015. [Online]. Available: https://www.fool.com/investing/general/2016/01/06/why-2015-wasnt-as-good-for-theaters-as-the-box-off.aspx

[21] J. Schectman, “Netflix uses big data to improve streaming video,” 2012. [Online]. Available: https://blogs.wsj.com/cio/2012/10/26/netflix-uses-big-data-to-improve-streaming-video/

[22] Netflix, “Netflix jobs.” [Online]. Available: https://jobs.netflix.com

[23] P. Pääkkönen and D. Pakkala, “Reference architecture and classification of technologies, products and services for big data systems,” *Big Data Research*, vol. 2, no. 4, pp. 166 – 186, 2015.

[24] C. A. Gomez-Uribe and N. Hunt, “The Netflix recommender system: Algorithms, business value, and innovation,” *ACM Trans. Manage. Inf. Syst.*, vol. 6, no. 4, pp. 13:1–13:19, Dec. 2015.

[25] Y. Raimond and J. Basilico, “Recommending for the world,” 2016. [Online]. Available: https://medium.com/netflix-techblog/recommending-for-the-world-8da8cbcf051b

[26] X. Amatriain and J. Basilico, “Netflix recommendations: Beyond the 5 stars,” 2012. [Online]. Available: https://medium.com/netflix-techblog/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429

[27] ——, “Netflix recommendations: Beyond the 5 stars (part 2),” 2012. [Online]. Available: https://medium.com/netflix-techblog/netflix-recommendations-beyond-the-5-stars-part-2-d9b96aa399f5

[28] G. K. Prasanna Padmanabhan, Kedar Sadekar, “What’s trending on Netflix?” 2015. [Online]. Available: https://medium.com/netflix-techblog/whats-trending-on-netflix-f00b4b037f61

[29] A. Berglund, “How data science helps power worldwide delivery of Netflix content,” 2017. [Online]. Available: http://bit.ly/2sapG36

[30] N. Govind, “A/b testing and beyond: Improving the Netflix streaming experience with experimentation and data science,” 2017. [Online]. Available: http://bit.ly/2sGmrRu

[31] P. Simon, “Big data lessons from Netflix,” 2014. [Online]. Available: https://www.wired.com/insights/2014/03/big-data-lessons-netflix/

[32] G. Petraetis, “How Netflix built a house of cards with big data,” 2017. [Online]. Available: https://www.cio.com/article/3207670/big-data/how-netflix-built-a-house-of-cards-with-big-data.html

[33] G. I. Webb, M. J. Pazzani, and D. Billsus, “Machine learning for user modeling,” *User Modeling and User-Adapted Interaction*, vol. 11, no. 1, pp. 19–29, Mar 2001.

[34] S. Bidel, L. Lemoine, F. Piat, T. Artières, and P. Gallinari, “Statistical machine learning for tracking hypermedia user behavior,” in *2nd Workshop on Machine Learning, Information Retrieval and User Modeling, 9th Int. Conf. in UM, 2003*, 2003.

[35] Y. C. Yang, “Web user behavioral profiling for user identification,” *Decision Support Systems*, vol. 49, no. 3, pp. 261 – 271, 2010.

[36] M. Narvekar and S. S. Banu, “Predicting user’s web navigation behavior using hybrid approach,” *Procedia Computer Science*, vol. 45, no. Supplement C, pp. 3 – 12, 2015, international Conference on Advanced Computing Technologies and Applications (ICACTA).

[37] O. Raphaeli, A. Goldstein, and L. Fink, “Analyzing online consumer behavior in mobile and pc devices: A novel web usage mining approach,” *Electronic Commerce Research and Applications*, vol. 26, no. Supplement C, pp. 1 – 12, 2017.

[38] K. Dembczyński, W. Kotlowski, and M. Sydow, “Effective prediction of web user behaviour with user-level models,” *Fundam. Inf.*, vol. 89, no. 2-3, pp. 189–206, Apr. 2008.

[39] A. Markitanis, D. Corapi, A. Russo, and E. C. Lupu, *Learning User Behaviours in Real Mobile Domains*. Imperial College Press, 2014, ch. 6, pp. 43–51.

[40] G. H. Alférez, E. Zebadúa, and E. Cruz, “Using data science to understand segments of individuals who have been removed from membership in the Inter-Oceanic Mexican Union Conference from 2005 to 2013,” School of Engineering and Technology, Montemorelos University, Tech. Rep., 2016. [Online]. Available: http://www.harveyalferez.com/publications/TechnicalReport_June_23_2016_GSL_UM.pdf

[41] G. H. Alférez, “Discovering mission-oriented patterns with open data in New York City,” 2017. [Online]. Available: http://bit.ly/2BYrbpR

[42] ——, “Discovering hidden patterns in US health-related open data with machine learning,” 2016. [Online]. Available: http://bit.ly/2DtEPzf

[43] G. H. Alférez, J. Rodríguez, L. Pompe, and B. Clausen, “Interpreting the geochemistry of southern california granitic rocks using machine learning,” in *Proceedings of the 2015 International Conference on Artificial Intelligence (ICAI 2015)*, 2015.

[44] C. Marín, G. H. Alférez, J. Córdova, and V. González, *Detection of melanoma through image recognition and artificial neural networks*. Cham: Springer International Publishing, 2015, pp. 832–835.

[45] R. Domínguez, “Aplicación de ciencia de datos para la creación de software predictivo de morbimortalidad materna en México,” Master’s thesis, School of Engineering and Technology, Montemorelos University, 2017. [Online]. Available: http://dspace.biblioteca.um.edu.mx/xmlui/handle/20.500.11972/858?locale-attribute=en

**Dr. Germán H. Alférez** is a professor at the School of Engineering and Technology, Montemorelos University, Mexico. He holds a Ph.D. in Computer Science (Summa Cum Laude) from Technical University of Valencia (Spain), a MSc in Information and Communication Technology from Assumption University (Thailand), and a BSc in Computer Science Engineering from EAFIT University (Colombia). His research interests include Services Computing, Model-Driven Engineering, Models at Runtime, Autonomic Computing, Data Science, Big Data, and Dynamic Software Product Lines. He has contributed to publications in top journals, book chapters, and international conferences (Scopus h-index: 6). He has worked in universities, IT companies, and research groups of four continents (America, Asia, Australia, and Europe). He leads the Global Software Lab at the School of Engineering and Technology, Universidad de Montemorelos. Also, he is a member of the IEEE Monterrey Section and the IEEE Computer Society.

His research contributions have been recognized by the National Council of Science and Technology (CONACYT), Government of Mexico by awarding him a distinction in the National System of Researchers (SNI) during 2018-20. Also, he is a research fellow of Peru's National Council for Science, Technology and Technological Innovation (CONCYTEC) during 2016-18.