Ph.D. Proposal

**Title: How to use unstructured data in a digital twin system for an oil field.**

**Subtitle: How to ingest and process unstructured data to use in data-driven and hybrid models.**

By Fábio Corrêa Cordeiro - Petrobras Research and Development Center

# Introduction

The automobile was the symbol of the industrial age. Ford’s assembly line shaped the twentieth-century industry, cities, and society. In the same way, the self-driving car is the symbol of the digital transformation. A self-driving car is a physical object, however, besides its mechanical and electrical components, it is powered by many software and artificial intelligence algorithms. Furthermore, communication protocols allow the exchange of information with other cars, the road, traffic signals, satellites, the internet as well as other existing systems.

However, what makes it possible for a car to drive itself is data, a lot of data. A self-driving car has much more sensors than an ordinary car. The sensors receive information about engine rotation, velocity, fuel level, and brake temperature. There are several cameras and 3D scannings that interpret images of other vehicles, pedestrians and traffic signals. They access information about localization from Global Positioning System (GPS) satellites, update maps and routes from the internet, as well as information from social networks.

The same phenomenon that is happening with ordinary cars is happening or will happen with almost all equipment and machines. The concept that all systems consist of two systems, one physical and the other virtual, is called digital twins (Grieve, 2017). Digital twins are data-intensive systems, so it is very important to pay special attention to unstructured data. Some studies estimate that they are about 80% of all generated data (Chelmis et al., 2013).

For a long time, companies have made a big effort to integrate their structured data. However, unstructured data is still being underused. For that reason, this Ph.D. proposal is about how to use unstructured data in a digital twin for an oil and gas field. More specifically, how to ingest and process this data to use in the models that will support this digital twin.

First of all, it is necessary to understand what digital technology is and the difference between the physical and virtual layers in its architecture. Another important aspect is the definition of complex and cognitive systems and digital twins.

This proposal is for a specific kind of digital twin, a digital oil field. The components of an oil field are the geological reservoir, the wells, submarine equipment, pipes and risers where the oil flows from the reservoir to the surface and the topside equipment in the oil platform. There are many sources of data and sensors in this complex system. All this data has the potential to increase the precision and reliability of the digital twin. For this reason, the purpose of this proposal is to understand how to process unstructured and semi-structured data for the ingestion of a digital oil field.

# Digital Technology

In the last decades, the transformation of technology and its impact on society was so intense that the term fourth industrial revolution has become common. The World Economic Forum Annual meeting in 2016 was held under the theme “Mastering the Fourth Industrial Revolution” (World Economic Forum, 2016). There are many digital technologies supporting this transformation. In this kind of technology, it is possible to divide the physical and the virtual, the device and the service provided, the physical network and the data that flows in it.

The digital transformation is in the center of the fourth industrial revolution. The Gartner Group estimate about 20 billion devices connected to the internet in 2020 (Hung, 2017), and every day 2.5 quintillion bytes of data worldwide are created (Marr, 2018). The challenge for the companies is how to properly use all these devices and data.

It is common to enumerate digital technologies instead of defining them. Some examples are devices connected to the internet (Internet of Things), platforms for distributed storage and process a large volume of data (Big Data), on-demand availability of computer system resources (Cloud Computer), and many algorithms that mimic cognitive functions (Artificial Intelligence) (Qi and Tao, 2018).

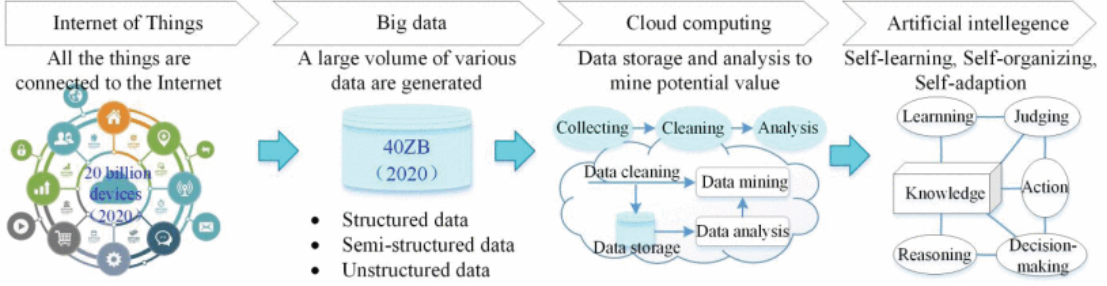


Fig 1 - New IT and their application (Qi and Tao, 2018)

One way to define digital technologies is by the characteristic of their layered architecture (Yoo, Henfridsson and Lyytinen, 2010). In digital technologies, it is possible to separate the physical and virtual components. In a smartphone, for example, there is the physical device and the apps downloaded from an app store. In the same way, there is a network and the content. The network has protocols shared by several companies whereas the content is stored on the device or on the cloud. The separation of the device and the services, the apps in the case of a smartphone, allows the re-programmability of the equipment. Likewise, the homogenization of the data allows the split of the network and the content.

The layered architecture consists of four layers: device, network, services, and content (Yoo, Henfridsson and Lyytinen, 2010). The device layer is divided into physical - the hardware part, and logical - the operational system. In the same manner, the network layer has physical components, like cables, radio spectrum, transmitters, and a logical part with the protocols and standards of communication. The applications and functionalities are in the service layer. And finally, the content layer deals with data and metadata.

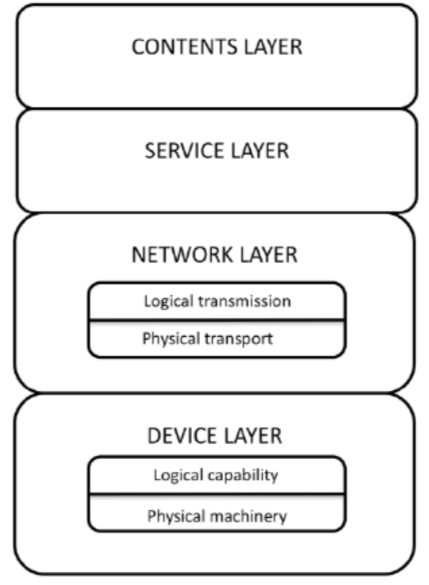


Figure 2 - The layered architecture of digital technology (Yoo, Henfridsson and Lyytinen, 2010)

Modular architecture is another important aspect of digital technology. Modular designs emphasize standardized interfaces, and how the product can be decomposed and recombined. So, in a modular layered architecture, it is desirable: products agnostic components; fluid boundary and meaning between components and products; shared standards and protocols; and multiple design hierarchies.

The rise of the volume of data and the Internet of Things devices increase the possibilities of products with a modular layered architecture. The layered architecture, with their physical and virtual components, is ideal for complex and cognitive systems.

# Complex and cognitive systems

Architecture is the representation of a product, its components, connections and the relationship between them. In the modular layer architecture of digital technologies, these components could be physical or virtual. The set of physical and virtual components is considered a system.

*“A system is two or more components that combine together to produce from one or more inputs one or more results that could not be obtained from the components individually”* (Grieves and Vickers, 2017)

It is possible to classify the systems into three classes: simple, complicated, and complex. Simple systems are predictable and it is not difficult to identify their components. It is simple to discern each part of the system. The inputs and outputs are clear and the actions are visible. Also, in the case of complicated systems, the inputs are visible and output predictable. However, in complicated systems, the number of components is bigger. Finally, a complex system is composed of a large network of components and several communication paths interacting with each other. In a complex system, it is not easy to describe and discern its components. There is sophisticated information processing and the outputs are not readily predictable.

In the case of the term *Cognitive System,* it is used for systems that mimic human intelligence. The term became famous after IBM launches Watson, a platform for artificial intelligence services, tools, and applications (Ferrucci et al., 2010). In 2011, Watson won Jeopardy!, a popular game show of question and answer in natural language. The question and answer tasks are very difficult problems, one of the most challenging in computer science and artificial intelligence domain (Daily and Peterson, 2016). In general, question and answer tasks are related to synthesizing information retrieval, natural language processing, knowledge representation and reasoning, machine learning, and computer-human interfaces.

After Watson had won the popular trivia game, many systems had added a cognitive layer in their architecture. One special case is the digital twins, a complex system composed of physical and virtual components, powered by artificial intelligence applications, with a semantic backbone.

# Digital Twins

To achieve business outcomes that matter, Daily and Peterson (2016) said it is crucial to bring together machines, data, insights, and people. It is necessary to connect these four elements with an integrated view of the operation. Thus, the digital twin aims to be a system that does this connection.

A digital twin is a system that connects physical products and their virtual representation. One definition of a digital twin is:

*“a set of virtual information constructs that fully describes a potential or actual physical manufactured product from the micro atomic level to the macro geometrical level. At its optimum, any information that could be obtained from inspecting a physical manufactured product can be obtained from its Digital Twin”* (Grieves and Vickers, 2017).

The granularity of the representation depends on the business goals. A digital twin can represent a component, a piece of equipment, an asset or a fleet of assets. For example, in the case of the aeronautic industry, the digital twin can represent a blade in a jet engine, a jet engine, an entire airplane, or a fleet of planes (GE, 2018).

One challenge of a space project is the virtual representation of a real asset that should be operated from a long distance. In 1970, when Apollo 13 was in trouble, NASA already had a mirrored system that helped the earth team to find a way to rescue the mission (Marr, 2017). However, the actual concept of digital twins come from the Product Lifecycle Management (PLM) center in the University of Michigan. In 2002, they proposed that the “*digital information would be a ‘twin’ of the information that was embedded within the physical system itself and be linked with that physical system through the entire lifecycle of the system*” (Grieves and Vickers, 2017).

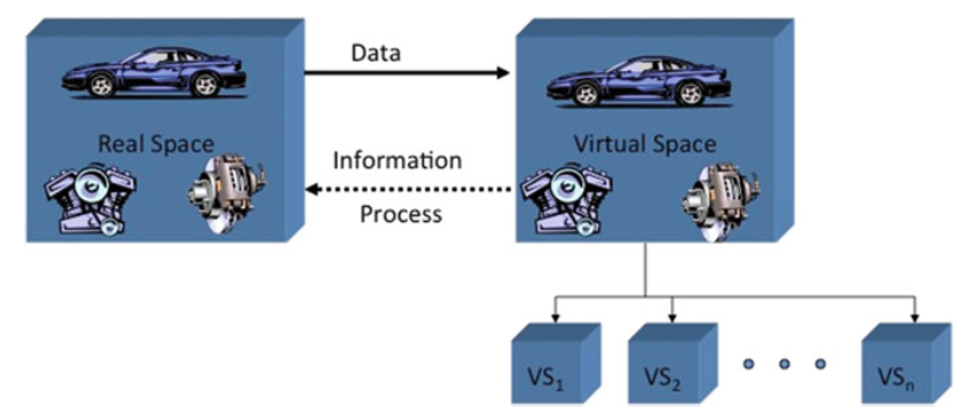


Figure 3: Conceptual idea for Product Life Cycle (Grieves and Vickers, 2017)

The architecture of a digital twin system consists of three major elements: assets models that describe systems and subsystems; analytics that describe, predict and prescribe the behavior of the assets; and a knowledge base of data sources and derived insights (GE, 2018). In the General Electric platform PREDIX, they proposed an architecture with components that ingest data from several sources, an industrial platform centered in the asset model and the industrial applications. This architecture is very consistent with the layered modular architecture for digital technologies proposed by Yoo, Henfridsson, and Lyytinen (2010). It is possible to identify the components of the four layers: device, network, service, and contents.

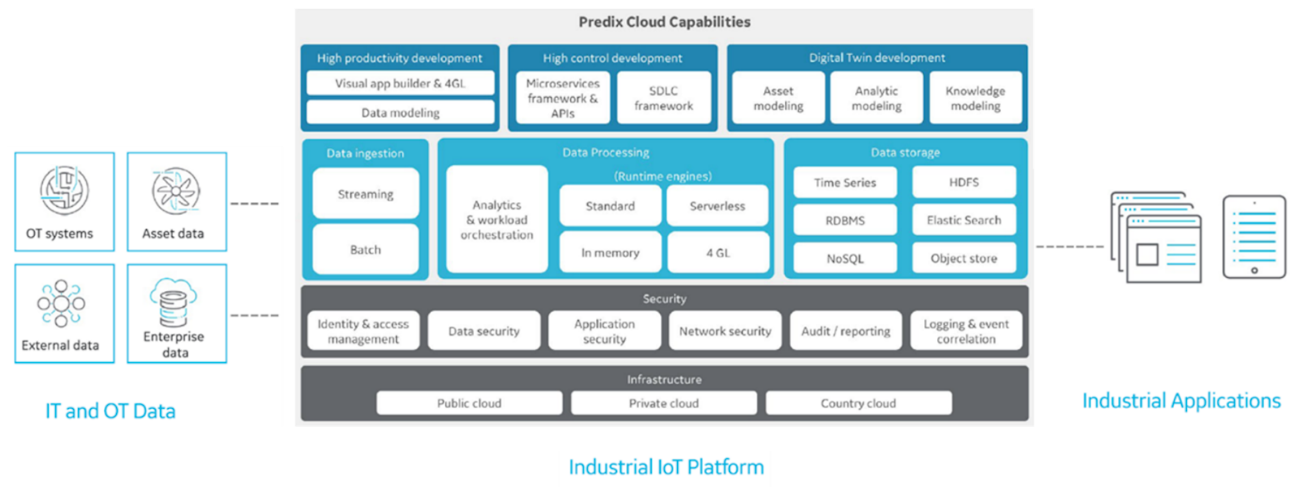


Figure 4: The architecture of General Electric platform PREDIX (GE, 2018)

The digital twin is a virtual representation of the product, so the acquisition of data from the physical components is essential. The sensors and the Internet of Things devices are the main source of input data. Another important source of information is the enterprise management systems. However, the digital twin should ingest data, from any source, which makes the models more predictable. The main data sources are industrial control systems, sensors, structured databases, data warehouse, enterprise’s unstructured data, and external sources.

The components can be divided into infrastructure, security, data running, and development environment. The infrastructure is in charge of the management of all system and data, including the cloud providers. In a digital twin system, the equipment is much more connected to the internet, for this reason, some cybersecurity components must be available. In the data running components are the ingestion of streaming and batch data, the processing of all this data and their storage.  The data store component could be a simple relational database, but, in general, it is a technology related to big data or cloud. The virtual twin models and the industrial application will use the components of the development environment.

In the development environment are the three main components of the digital twin: asset, analytic and knowledge modeling. The asset model is a logical construct that represents the physical system. This logical construct could be a physics-based model, in other words, it uses mathematics and physics equations in their numerical solutions. Another option is the data-driven model, solutions based on collected data and machine learning algorithms. Eventually, the models can be a hybrid model of physics-based and data-driven.

The analytic model of the digital twin aims to describe a past situation or to predict a future state. In the case of descriptive analytics, the objective is to summarize past results. Otherwise, the predictive analytics guess the future states and helps to make a decision in complex situations. Finally, when the control of what is modeled is good enough, it is possible to prescribe an action after predicting the future state. The prescribed action can be a suggestion for a human operator or it can be automatized.

The last component of the digital twin is a knowledge model, the semantic backbone of the system. It is composed of a knowledge base of data sources and derived insights. A domain ontology will help with reasoning and information retrieval tasks. The knowledge model is the component that allows the digital twin to become a cognitive system.

The description above is a generic representation of a digital twin, but every complex piece of equipment has its own peculiarities. The object of this proposal is a digital twin of an oil field, with thousands of pieces, dozens of wells, and a geological reservoir. Therefore, this digital twin has many specificities.

# Digital Oil Field

An oil field is a complex system that aims to extract oil, and also gas in many cases, from a subterranean reservoir during a long period of time. The components of this system are geologic rock formations; fluids, like oil and gas, but also water and carbon dioxide; kilometers of drilled wells, valves, chokes, wellheads, pipelines; treatment facilities; and storage and offloading tanks. In this proposal, the focus is deepwater offshore oil fields. Therefore, the components of this system are in one or more oil platforms, more than a hundred kilometers from the coast. The wellheads and submarine equipment are in the seafloor about two thousand meters below sea level. The rock formation can be drilled by dozens of wells across a wide salt layer, and the oil and gas reservoir can be more than six thousand meters from the oil platform in the surface (ANP, 2019).

An oil field is a data-intensive system. For example, Chevron reported for only one field 9,000 wells that recorded one million data points per day. About 30% of valuable information of this field is in structured databases, the rest are in spreadsheets and other unstructured files on desktops. Their workflow has nine datasets and eleven different tools (Chelmis et al., 2013).

The Digital Oil Field is a digital twin for an oil field system. This digital twin can embrace all the oil field system or only part of that (Mayani, Svendsen and Oedegaard, 2018). For example, depending on the business priority, the digital twin can be only one well, all the subsea equipment or all the oil field. Besides, the implementation of the digital twin can be gradual, starting with a few subsystems and progressively adding more. Some examples of projects that aim to implement a digital oil field are the *i-Field* program of Chevron, *Smart Fields* of Shell, the *Field of the Future* program of BP, and *Integrated Operation of the High-North* (IOHN) (Chelmis et al., 2013).

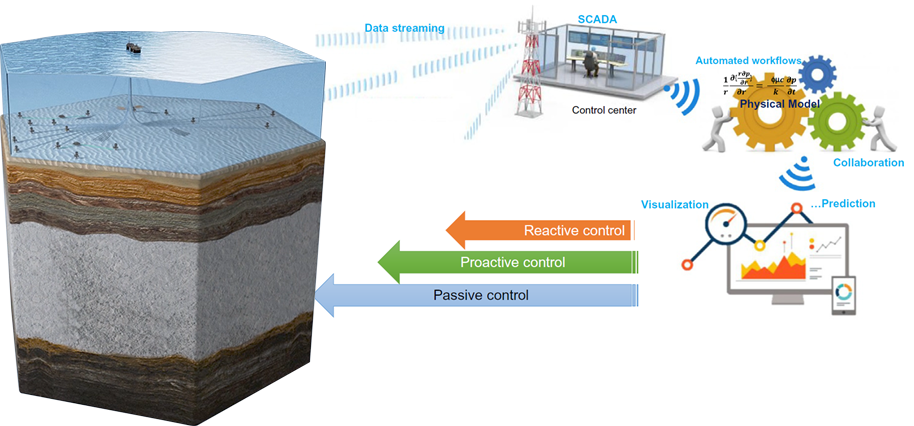


Figure 5: Digital Oil Field (adapted from Carvajal, Maucec and Cullick, 2017)

Some components of a digital twin for an oil and gas field are the sensors to collect data, the control center, the automated workflows, the analytic components, and the automated control. The sensors receive data from the platform facilities, subsea equipment, and subsurface. This amount of collected data is combined with enterprise data, engineer drawings, laboratory analysis, and other documents to be used by the digital twin models. The data is sent to the control center, in the platform or onshore. The control center is in charge of the automated workflows and automated controls. The analytic components could help the operation of the oil field or aggregate information for its management.

The uncertainty in the oil business is constant. It is necessary to build a huge industrial complex, to spend billions of dollars before the oil flow from the reservoir is seen. Decisions are made based on sparse data. Historically, the models are built using only well known structured data. However, there is a considerable amount of documents and information waiting to be used by these models.

# Data Sources in digital oil fields

Recently, big data became a buzzword and is used also as a synonym of digital transformation. Nevertheless, big data is the capacity of storing, processing and using a huge volume and variety of data. This data should be processed in a fast enough velocity to bring value for the organization, and in many cases, it means real-time. In a data-intensive company, gathering the right information is an essential part of the work of many professionals. The data can be structured, unstructured or semi-structured. For a long time, the attention was in the structured databases, but the unstructured and semi-structured documents are about 80% of the business data (Chelmis, 2013). The oil companies are full of unstructured data ready to be incorporated into the workflows.

Some authors define big data by the 4Vs: Volume, Variety, Velocity, and Value (Qi and Tao, 2018). The volume of data is the most obvious face of a big data system. However, a huge data warehouse not necessarily can be called big data. A second characteristic is the variety of type, format, content, and size. For a big data system, it is important to get information from heterogeneous documents. For many applications, the data should be processed in real-time, so velocity is essential. And value is the main objective as well as great opportunity. A big data system should seek to balance the following requirements: an acceptable installation overhead; end users must be able to access the data; scalability; and combination of data sources, with static and streaming data (Giese et al., 2015).

It is possible to understand the centrality of data manipulation by the example of the service centers for power plants from Siemens Energy. Their main task is monitoring and diagnosing assets like gas and steam turbines, generators, and compressors, common equipment in an oil platform. The engineer’s routine is: a) receive an issue about an asset, b) gather relevant data about the issue, c) analyze the data, and d) report ways to address the issue. The second step, gather the data, is the bottleneck and consumes 80% of the time (Kharlamov et al., 2017). So, finding data is essential to give a better and quicker answer for an issue.

There are three types of data, structured, semi-structured and unstructured. The structured type refers to data with a table format, in a relational database or spreadsheets. Unstructured data are the files without a structure at all, like raw text, images, videos, audio, and logs. At last, the semi-structured documents follow self-describing structures that use tags or other marks to enforce hierarchies and semantic elements. The most common semi-structured formats are the Extensible Markup Language (XML) and JavaScript Object Notation (JSON) formats (Qi and Tao, 2018).

Some examples of unstructured documents useful for a digital oil field are the operational and laboratory reports, radio messages and camera images. Even the relational databases have some textual sections. These sections store historical events of breakages, failures and, maintenances, per example (Furtado, 2017). In the same way, the reports produced by the laboratory teams can provide rich information about the oil field and bring insights derived from other fields with a similar condition. The radio messages exchanged by the operators can be stored using some speech-to-text application. These messages have the most up-to-date information about the operational situation and, using sentiment analysis, it is feasible to discern the level of stress of the operation team.

Image is a type of unstructured document very common in industrial plants. Deep neural networks already overtook human skill in image recognition (The Economist, 2016), and using these algorithms to process videos can help the control centers. In the topside facilities, there are cameras to scan the activities looking for unsafe events and abnormal equipment behaviours. Some newer oil platforms already have infrared cameras, which enable a new layer of observation. The infrared images allow tracking the temperature of each piece of equipment. This information extracted from the videos can be registered in a structured database for predictive or prescriptive analysis. Another very useful type of video is the recording made by from underwater vehicles. They permit access to the subsea equipment such as risers, wellheads, manifolds, and pipes. The underwater videos can help to monitor the seabed as well as the marine environment.

An oil and gas field produces a lot of semi-structured data. In general, these data are transmitted and stored in an XML format. The industry tries to standardize some of those data formats, some examples are the Energistics standards. They elaborated three XML formats very useful: WITSML, PRODML, and RESQML. The WITSML data come from the drilling world, it covers static information, snapshots, and sensor information. PRODML organize data from production and surface. Finally, the RESQML standard is for moving data in the earth modeling chain (Energistics, 2019).

The great variety of unstructured and semi-structured data available in an oil field is a great opportunity to improve the asset models. A digital twin for an oil field must have good models and should not waste 80% of its data.

**Project Proposal**

The objective of this Ph.D. proposal is to study the best way to ingest unstructured and semi-structured data in the scope of a digital twin for an oil and gas field. Specifically, to understand how to process these documents, join them with structured data and run models with better results. In order to achieve this result, it is necessary to find a study case and apply all the processes from data ingestion to running the model.

When it is necessary to mix structured and unstructured data, one possible way to do it is to structure the unstructured data. An example is the Two Sigma competition on Kaggle website (Kaggle, 2019). Kaggle is a platform of machine learning competition, and they launched a challenge for training a predictive model for the stock market. The winning model must use stock prices, a structured time series, and news information, an unstructured data. The objective is to predict which company will increase or decrease their stock value more than the market average. In the case of this competition, the texts already had been transformed. For instance, they tagged the news using some predefined label structure. They also applied sentiment analysis techniques to define if the text was positive, negative or neutral. Finally, descriptive statistics of the text and metadata was added in the final dataset. In the end, the datasheet available for the competitors is not the raw text of the newspaper, but a structured data enriched by semantics and context.

The same procedure can be made when processing unstructured and semi-structured documents in an oil field, and it is the main aim of this project. The objective is to apply semantics and context in texts, audio, images, videos, XML and JSON documents for a digital oil field. This files must to be processed and enriched using some technique like auto-tagging, entity and relation recognition, and events classification, that can help to structure these files. At the end of the project, this enriched new data should enable a better model for the oil field digital twin system.

**Conclusion**

In the context of the fourth industrial revolution, companies that pay attention to digital technologies achieved better results. These technologies can be presented in a modular layered architecture with physical and virtual components. Its modular characteristic allows an easier connection, exchange, and upgrade of the components. It is a good framework for complex and cognitive systems.

Digital twins are cognitive systems that represent all information of an asset through virtual representation. The physical and virtual components are linked by sensors and models, so all information and behavior of the real asset are accessed by the digital twin. A digital twin is a data-intensive system and, nowadays, 80% of the business data generated is unstructured or semi-structured. Therefore, this project aims to understand what is the best method for using unstructured and semi-structured data in a digital twin for an oil and gas field.

**Candidate**

Fábio Corrêa Cordeiro is an Industrial Engineer graduate of the Federal University of Rio de Janeiro (UFRJ), postgraduate in Strategic Management of Technological Innovation of the State University of Campinas (UNICAMP), and a Business Intelligence Master of Pontificia Universidade Catolica (PUC-Rio) (finishing in December, 2019). During the last twelve years, he has worked for the largest Brazilian research center where he had the opportunity to follow several R&D projects and to interact with the leading  Brazilian universities. Presently he works on R&D projects for Digital Technologies, mainly in areas of semantic search and natural language processing.

LinkedIn: <https://www.linkedin.com/in/fabio-correa-cordeiro/>

Git Hub: [github.com/fabiocorreacordeiro](https://github.com/fabiocorreacordeiro)

E-mail: [fabio.cordeiro@petrobras.com.br](mailto:fabio.cordeiro@petrobras.com.br)

**References**

ANP (2019). Lula - 2018, Plano de Desenvolvimento Aprovado. [online] Anp.gov.br. Available at: http://www.anp.gov.br/images/EXPLORACAO\_E\_PRODUCAO\_DE\_OLEO\_E\_GAS/Gestao\_Contratos/Fase\_Producao/Planos\_Desenvolvimento/sumario\_lula.pdf [Accessed 27 May 2019].

Carvajal, G., Maucec, M. and Cullick, S. (2017). Intelligent Digital Oil and Gas Fields. San Diego: Gulf Professional Publishing [Imprint].

Chelmis, C., Zhao, J., Sorathia, V., Suchindra, A. and Prasanna, V. (2013). Toward an Automatic Metadata Management Framework for Smart Oil Fields. SPE Economics & Management, 5(01), pp.33-43.

Daily, J. and Peterson, J. (2016). Predictive Maintenance: How Big Data Analysis Can Improve Maintenance. Supply Chain Integration Challenges in Commercial Aerospace, pp.267-278.

Energistics. (2019). Energistics standards are developing and integrating. [online] Energistics.org. Available at: https://www.energistics.org/wp-content/uploads/2019/05/DEJ-Feb-Apr19-pages-16-and17.pdf [Accessed 27 May 2019].

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J., Nyberg, E., Prager, J., Schlaefer, N. and Welty, C. (2010). Building Watson: An Overview of the DeepQA Project. AI Magazine, 31(3), p.59.

Giese, M., Soylu, A., Vega-Gorgojo, G., Waaler, A., Haase, P., Jimenez-Ruiz, E., Lanti, D., Rezk, M., Xiao, G., Ozcep, O. and Rosati, R. (2015). Optique: Zooming in on Big Data. Computer, 48(3), pp.60-67.

Grieves M., Vickers J. (2017) Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. In: Kahlen FJ., Flumerfelt S., Alves A. (eds) Transdisciplinary Perspectives on Complex Systems. Springer, Cham

Hung, M. (2017). Leading the IOT, Gartner insight on how to lead in a connected world. [online] Gartner.com. Available at: https://www.gartner.com/imagesrv/books/iot/iotEbook\_digital.pdf [Accessed 27 May 2019].

Kaggle. (2019). Two Sigma: Using News to Predict Stock Movements. [online] Kaggle.com. Available at: https://www.kaggle.com/c/two-sigma-financial-news/data [Accessed 27 May 2019].

Kharlamov, E., Mailis, T., Mehdi, G., Neuenstadt, C., Ozcep, O., Roshchin, M., Solomakhina, N., Soylu, A., Christoforos Svingos, C., Brandt, S., Giese, M., Ioannidis, Y., Lamparter, S., Moller, R., Kotidis, Y. and Waaler, A. (2017). Semantic Access to Streaming and Static Data at Siemens. SSRN Electronic Journal.

Marr, B. (2017). What Is Digital Twin Technology - And Why Is It So Important? [online] Forbes.com. Available at: https://www.forbes.com/sites/bernardmarr/2017/03/06/what-is-digital-twin-technology-and-why-is-it-so-important/#5956aaca2e2a [Accessed 27 May 2019].

Marr, B. (2018). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. [online] Forbes.com. Available at: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#6396e6f960ba [Accessed 27 May 2019].

Mayani, M., Svendsen, M. and Oedegaard, S. (2018). Drilling Digital Twin Success Stories the Last 10 Years. SPE Norway One Day Seminar.

Qi, Q. and Tao, F. (2018). Digital Twin and Big Data Towards Smart Manufacturing and Industry 4.0: 360 Degree Comparison. IEEE Access, 6, pp.3585-3593.

(Qi and Tao, 2018)

SIRIUS (2019). Digital Twins - Sirius. [online] Sirius. Available at: https://sirius-labs.no/digital-twins/ [Accessed 27 May 2019].

The Economist. (2016). From not working to neural networking. [online] Economist.com. Available at: https://www.economist.com/special-report/2016/06/23/from-not-working-to-neural-networking [Accessed 27 May 2019].

World Economic Forum (2016). World Economic Forum Annual Meeting 2016. Mastering the Fourth Industrial Revolution. [online] Www3.weforum.org. Available at: http://www3.weforum.org/docs/WEF\_AM16\_Report.pdf [Accessed 27 May 2019].

Yoo, Y., Henfridsson, O. and Lyytinen, K. (2010). Research Commentary—The New Organizing Logic of Digital Innovation: An Agenda for Information Systems Research. Information Systems Research, 21(4), pp.724-735.