

**Herramientas de Software para Big Data**

**M7A ID**

**Obligatorio 2022:**

**Sistema de análisis de información de compañías bajo Big Data**

**245650 – Fernanda Secinaro**

**241896 – Juan Pablo Rodríguez Sotto**

Índice

[Introducción 2](#_Toc44772675)

[Planteamiento del Problema 2](#_Toc44772676)

[Objetivo General 2](#_Toc44772677)

[Objetivos Específicos 3](#_Toc44772678)

[Alcance 3](#_Toc44772679)

[Propuesta de Valor 3](#_Toc44772680)

[Metodología 4](#_Toc44772681)

[Ingeniería de atributos 5](#_Toc44772682)

[Análisis cualitativo 5](#_Toc44772683)

[Análisis de nulidad y componentes estadísticos 6](#_Toc44772684)

[Análisis de correlación 6](#_Toc44772685)

[Arquitectura de la Solución 7](#_Toc44772686)

[Conclusiones 8](#_Toc44772687)

[Restricciones 9](#_Toc44772688)

[Bibliografía 10](#_Toc44772689)

# Introducción

El presente documento consiste en una especificación del caso de negocio abordado para la realización del trabajo obligatorio de la materia Herramientas de Software para Big Data, más una descripción detallada de las operaciones realizadas a lo largo del análisis.

Luego de analizar varios dataset, se decidió optar por (<https://www.kaggle.com/datasets/peopledatalabssf/free-7-million-company-dataset?select=companies_sorted.csv>)

Consta de un archivo, llamado companies\_sorted.csv que cuenta con 7.173.438

|  |  |
| --- | --- |
| CAMPO | DESCRIPCIÓN |
| id | Id que identifica a la compañía |
| Name | Nombre de la compañía |
| domain | Sitio web de la compañía |
| Year\_founded | Año de fundación de la compañía |
| industry | Industria a la cual pertenece la compañía |
| Size range | Rango de tamaño |
| locality | Localidad (ciudad, estado, país) |
| country | País donde la compañía reside |
| Linkedin\_url | URL del linkedin asociado a la compañía |
| Current employee estimate | Cantidad actual de empleados estimados |
| Total employee estimate | Estimación de empleados totales |

# Planteamiento del Problema

En base a los datos disponibles se intentará analizar un volumen de aproximadamente siete millones de compañías, las cuales presentan una gran variedad de países, industrias, tamaños, etc. Analizar este volumen de información, sin un sistema de análisis como este, sería casi imposible por la cantidad de tiempo que requeriría. Pero nuestro sistema permite incrementar exponencialmente la velocidad, llegando a resultados en un tiempo razonable.

En base a esto intentaremos determinar la industria principal de cada país y de todos los países en general, determinar las principales localidades donde residen las compañías y determinar cuáles son las indutrias de mayor tamaño. Además, analizar la tendencia de qué tipo de industria se crea por año.

Los resultados de este análisis pueden ser muy útiles para un emprendedor que quiera crear una nueva start-up o una compañía ya existente que quiera hacer un rebranding. Por ejemplo, a partir de este análisis, puede determinar cuáles son las mejores localidades o países para una industria en particular.

# Objetivo General

Desarrollar un sistema de análisis bajo big data en near to real time para determinar las principales tendencias de las compañías, y a partir de esto, poder brindarle información útil a un emprendedor o compañías ya existentes.

# Objetivos Específicos

Se plantean los siguientes objetivos específicos:

* Realizar ingeniería de atributos, analizando y determinando atributos que pueden tener mayor poder explicativo.
* Desarrollar una arquitectura de big data que permita el análisis de las compañías.
* Analizar las tendencias de las compañías.
* Proponer acciones y políticas en base a los resultado obtenidos.

# Alcance

El presente trabajo representa un ensayo teórico, basado en datos reales.

Su objetivo es validar los conceptos, metodologías y herramientas vistas a lo largo del curso, y por lo tanto, se limitará a:

* Determinar la industria con más compañías en Uruguay, Estados Unidos y a nivel mundial.
* Determinar cuál es el país con más compañías.
* Determinar la localidad donde residen más compañías (a nivel mundial, en Uruguay y en Estados Unidos).
* Determinar la industria con mayor cantidad de empleados (a nivel mundial, en Uruguay y en Estados Unidos).
* Analizar la tendencia de qué tipo de industria se crea por año.

No se incluye en este enfoque:

* Evaluación de las mejores herramientas a utilizar para este tipo de análisis.
* Verificación o comprobación de los resultados obtenidos.
* Mejoramiento o enriquecimiento de los datos en caso de ser necesario.
* Análisis de inversión y propuesta de plan de acción para la implementación de las acciones sugeridas.

# Propuesta de Valor

En base a un volumen de aproximadamente siete millones, y con datos con una gran variedad, es que se va a poder determinar estadísticas varias que podrían ser cruciales para el nuevo emprendedor.

Nuestro enfoque propone un abordaje metódico y estadístico que posibilite acciones a la hora de crear o hacer un rebranding de una comañía. Sus resultados, combinados con el juicio experto de profesionales en la materia, pueden ayudar a crear nuevas y mejoras start-ups en diferentes industrias de todo el mundo. Además de que también provee información para mejorar las compañías ya existentes.

# Metodología

El presente trabajo se realizará utilizando la metodología de desarrollo ágil denominada Scrum. La misma se estructura en base a períodos de trabajo establecidos, denominados sprints, los cuales cuentan con un conjunto de tareas y actividades a realizar dentro del mismo. En este caso el scrum master será Pedro Bonillo y se realizarán cuatro sprints:

1. Determinación de objetivos y alcance del trabajo.
2. Armado de arquitectura de big data.
3. Armado de algoritmo de análisis.
4. Análisis de resultados y armado de conclusiones.

Los entregables de este trabajo son:

* Informe de trabajo realizado.
* Documentación de ingeniería de atributos realizados de forma “externa”
* Presentación power point para defensa.
* Jupyter Notebook con implementación del modelo.

# Plan de Trabajo

Al momento de realizar la tarea se realizan las siguientes actividades:

* Documentación
* Ingeniería de atributos. Este proceso se realiza de forma independiente al streaming con un desarrollo en Python, usando Jupyter Notebooks.
* Realizar las consultas necesarias para cumplir con nuestro alcance

# Ingeniería de atributos

A continuación se realizan los siguientes análisis sobre los atributos del dataset:

* Análisis cualitativo de las variables a utilizar, en base a la presunción de impacto en el modelo.
* Análisis de campos nulos y componentes estadísticos.
* Análisis de correlación entre variables para evitar errores de ponderación en el modelo.

## Análisis cualitativo

|  |  |
| --- | --- |
| **Atributo** | **Descripción** |
| id | Identificador - No aplica para el análisis |
| Name | Se requiere para conocer el nombre de las empresas al realizar consultas |
| domain | No aplica para el análisis |
| Year\_founded | Se utilizará para determinar tendencia de qué tipo de industria se crea por año |
| industry | Se utilizará para determinar las industrias de mayor tamaño |
| Size range | No aplica - Asumimos no tiene poder explicativo |
| locality | Se utilizará para determinar las localidades de mayor tamaño |
| country | Se utilizará para determinar las industrias de mayor tamaño en un país, etc. |
| Linkedin\_url | No aplica para el análisis |
| Current employee estimate | Se utilizará para determinar las industrias de mayor tamaño |
| Total employee estimate | No aplica - Se utilizará current employee estimate |

## Análisis de nulidad y componentes estadísticos

Se presentan los principales hallazgos del análisis realizado:

* No se borraron filas por tener un campo nulo, pero si se tuvieron en cuenta a la hora de realizar consultas, para no hacer el análisis utilizando campos nulos. Esto fue porque había muchas filas que, por ejemplo, tenián el país nulo y la industria no nula, o viceversa, entonces a dichas filas las usamos para realizar consultas sobre el campo no nulo, que no tienen en cuenta al que sí es nulo. Por ejemplo, en el caso de ver cuál es la industria principal a nivel global, nos quedamos con las filas que el valor de la industria no es nulo, y como en este caso, no importa que el país sea nulo, dejamos aquellas filas que tenían país nulo (e industria no nula).

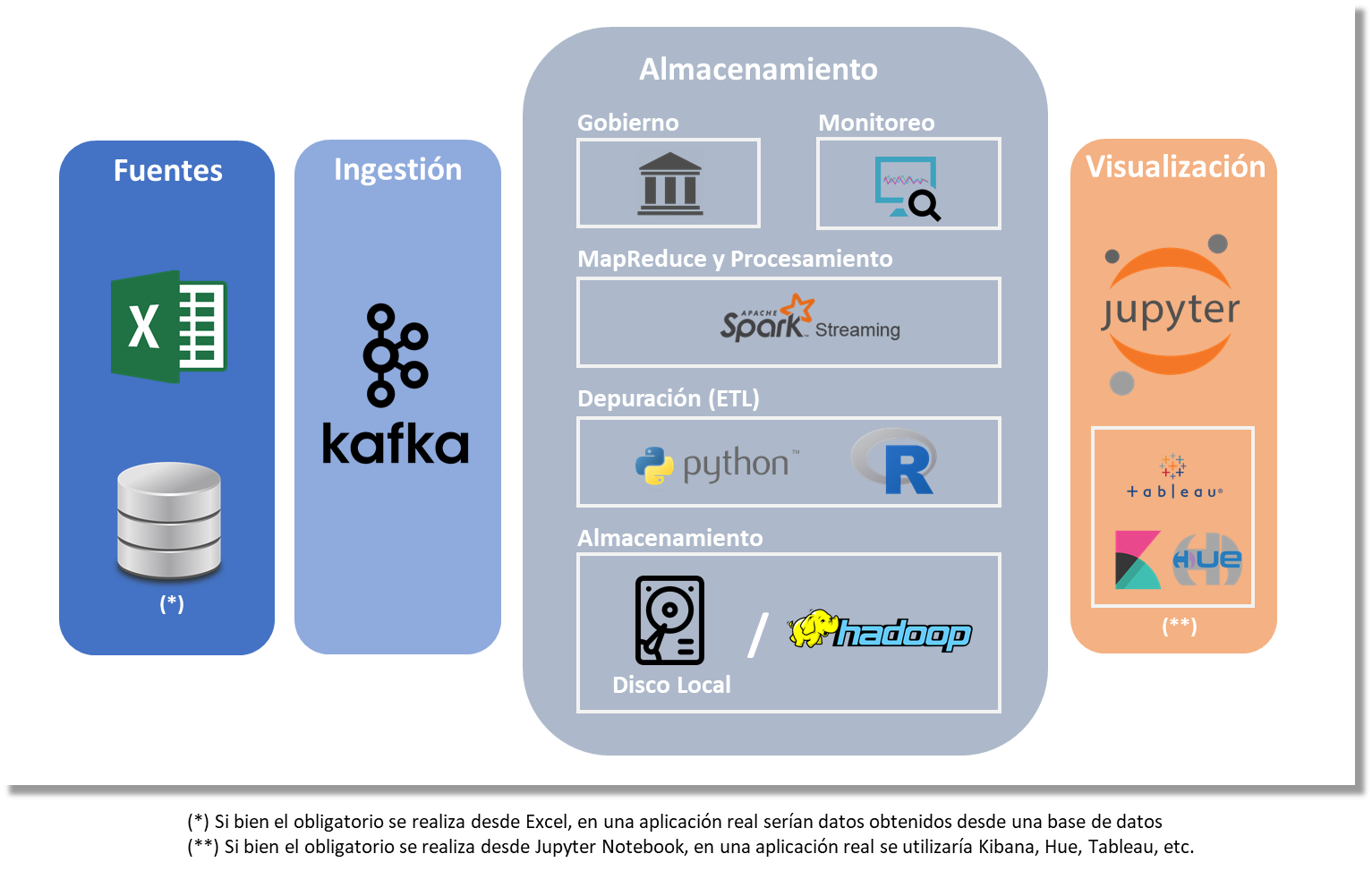
## Análisis de correlación

Tal como se comentó previamente se presenta un análisis de correlación para las variables seleccionadas.



Como se puede apreciar en la matriz, existe una correlación significativa entre las “current employee estimate” y “total employee estimate”, por lo que se procede a quedarnos únicamente con la primera.

# Arquitectura de la Solución

Para lograr los objetivos antes mencionados se propone la siguiente arquitectura:

Pasos del proceso:

1. Utilizando Jupyter Notebooks se realiza la ingeniería de atributos, seleccionando los atributos significativos.

Esta herramienta nos permite automatizar el análisis, aplicando funciones y programas pre desarrollados que facilitan sustancialmente la tarea. Adicionalmente es una herramientas abierta y robusta, que no requiere inversión en licencia.

1. Se toman los datos desde archivo formato plano y se procesan en Kafka para simular un streaming de datos. Se realiza la lectura del archivo y se publican los datos a través del productor de Kafka.

Se utiliza Kafka debido a su capacidad de procesar un streaming de datos en alta performance y dada su potencialidad de integración con Python para la aplicación de modelos de Machine Learning.

1. Se utiliza Spark Streaming y MLlib para procesar el archivo, generando un modelo no estructurado de streaming y ejecutando un modelo de clusterización KMeans.

Se utilizan las herramientas de Spark dada su robustez y amplia documentación acerca de la aplicación de algoritmos de clustering (Kmeans). Existen diversos artículos al respecto de su validez para soluciones similares al caso planteado.

1. Se realiza un ploteo de los datos para visualizar los centroides obtenidos.

Se utilizan las herramientas de ploteo de Jupyter Notebook debido a su practicidad y uso extendido a lo largo de la industria.

# Conclusiones

Las conclusiones obtenidas son:

**Realizar ingeniería de atributos, analizando y determinando atributos que pueden tener mayor poder explicativo para aplicarlos al modelo.**

Luego de un análisis cualitativo y cuantitativo de los atributos se decidió seleccionar los siguientes atributos:

* Nombre
* Año de fundación
* Industria
* Cantidad actual de empleados
* Localidad
* País

El país nos permite realizar consultas para países en específico, en nuestro caso, Estados Unidos y Uruguay. El año de fundación nos permite realizar un análisis de la tendencia de qué tipo de industria se crea en ciertos períodos de años. La industria la utilizamos para conocer cuál es la mayor industria con más empresas a nivel global, en Estados Unidos y Uruguay. Además, también se utilizó para conocer las industrias con mayor cantidad de empleados, junto con la cantidad actual de empleados. La localidad se utilizó para conocer cuál es la localidad con más compañías.

Finalmente, se determinó que las variables “current employee estimate” y “total employee estimate” cuentan con una correlación significativa, por lo que nos quedamos únicamente con la primera.

**Desarrollar una estructura de streaming de datos no estructurado que permita procesar la información en modalidad near to real time.**

La arquitectura propuesta y las herramientas seleccionadas demostraron ser útiles para el procesamiento del streaming de datos requerido. Pese a lo anterior, existe un conjunto de herramientas adicionales (pertenecientes y no pertenecientes al ecosistema de Apache Hadoop) que pueden potenciar este tipo de trabajos y brindar mejores niveles de performance, precisión y capacidad de visualización.

Es importante destacar que algunas de las tareas de ETL y análisis de datos se realizaron por fuera de la estructura del streaming debido a la restricción de tiempo antes mencionada. Es posible mejorar sustancialmente el modelo incorporando dichos pasos al streaming y ajustando los parámetros del algoritmo en base a la totalidad de los datos.

**Resultados Obtenidos**

**Determinar los hotspots de ocurrencia de accidentes, utilizando algoritmos de Machine Learning adecuados (Spark MLlib).**

Los clusters obtenidos son los siguientes:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster** | **# accidents** | **avg casualties** | **avg speed limit** | **avg light cond.** | **avg w cond.** |
| **1** | 11 | 1.18 | 10.00 | 2.00 | 1.63 |
| **2** | 7,324 | 1.18 | 19.98 | 1.72 | 1.52 |
| **3** | 667,632 | 1.28 | 30.00 | 1.84 | 1.58 |
| **4** | 86,144 | 1.44 | 40.00 | 1.94 | 1.62 |
| **5** | 287,464 | 1.53 | 61.69 | 2.23 | 1.70 |

Del análisis de los mismos se puede concluir:

1. Los clusters número 3 y 5 agrupan la mayoría de los accidentes de tránsito, representando cerca del 90% de los casos.
2. Alienado con la hipótesis previa, se detecta una tendencia de aumento de severidad de los accidentes (mayor cantidad de víctimas fatales) a medida que aumenta el promedio de velocidad de los vehículos involucrados.
3. Las condiciones de iluminación y clima tienen el mismo impacto en cada una de las zonas identificadas.

**Conclusiones en base a los resultados obtenidos.**

En base a los resultados obtenidos previamente llegamos a las siguientes conclusiones:

* A nivel mundial y en Uruguay la industria con más compañías es “Information and Technology Services”, esto indica que es una realidad que la tecnología es la industria con mayor demanda. Por otro lado, en Estados Unidos la industria con más compañías es Construcción, lo que refleja que es un país con mucho crecimiento a nivel de infraestructura, ya que por ejemplo, comparado con Uruguay, construcción no entra ni siquiera en el top 10 de compañías con más empresas.
* El país con más empresas por lejos es Estados Unidos, esto también puede reflejar que es uno de los países con mayor “facilidad” para crear una empresa, cuenta con mucha proyección y es un buen lugar para emprender.
* La localidad con más empresas a nivel global es London, Greater London, United Kingdom. Esto fue algo que es bastante llamativo, ya que creímos que iba a ser San Francisco, pero resultó estar en el 9no lugar. Por otro lado, en Uruguay la localidad con más empresas es Montevideo, esto era esperable y refleja que hay mucha centralización en la capital del país.
* A nivel mundial la industria con mayor cantidad de empleados es “Information and Technology Services”, al igual que antes, es un indicador de que la tecnología es la industria con mayor demanda, y hay un nivel de desempleo 0.
* A nivel global podemos concluir que fue cambiando la tendencia de qué tipo de empresa se creaba mayormente a lo largo de los años. Antes de 1980, la más popular era la construcción, y luego pasó a ser Information and Technology Services”, y podemos ver que cada vez el incremento de empresas de este tipo es mayor. En Uruguay está pasando exactamente lo mismo.

La conclusión general del trabajo es que aplicando las herramientas y metodologías antes mencionadas es posible determinar cuáles son los mejores lugares (tanto país como localidad) para crear una startup, o mudar una empresa ya existente. Además, también con el año de fundación e industria pudimos analizar cómo fue evolucionando qué tipo de industria es la más creada, algo importante para conocer cuáles son las industrias que más se adecuan a los días de hoy, y cuáles fueron disminuyendo con el correr de los años.

También es posible determinar cuál es la industria con más demanda a nivel de empleados, por lo que esta información le sirve también a una persona que está buscando estudiar una profesión con mucha demanda laboral.

# Restricciones

El presente análisis representa un trabajo de laboratorio, con el objetivo de validar los conceptos, metodologías y herramientas vistos en el curso. Sobre este trabajo aplican las siguientes restricciones:

* Tiempo: Existe escasa disponibilidad de tiempo tanto por contar con una fecha de entrega preestablecida, como por no contar con dedicación total a la tarea por parte de los realizadores.
* Teletrabajo: Dada la situación particular generada por la pandemia el curso completo y las sesiones de scrum se realizaron de forma virtual. Se requirió una alta dedicación de tiempo para la instalación y configuración de las herramientas requeridas para el análisis.
* Equipos disponibles: Se trabajó de forma local, con equipos de escasos recursos, lo que limitó la capacidad de procesamiento y aumentó los tiempos considerablemente.
* Experiencia: El equipo de trabajo asignado no cuenta con experiencia relevante en este tipo de análisis, siendo su primera aproximación al concepto práctico de Big Data.

# Bibliografía

Materiales de clase ORT:

* <https://aulas.ort.edu.uy/course/view.php?id=3894&section=0>
* <https://aulas.ort.edu.uy/course/view.php?id=3939>
* <https://aulas.ort.edu.uy/course/view.php?id=4303>

Libros

* [Shumway, Stoffer, Time Series Analysis and its Applications (4th ed.)](https://aulas.ort.edu.uy/mod/url/view.php?id=232570)
* Tom M. Mitchell. McGraw-Hill, 1997

Artículos

* <https://www.linkedin.com/pulse/propuesta-de-una-arquitectura-gesti%C3%B3n-grandes-datos-la-bonillo-ramos/>