MADISON INTELLIGENCE

PROGRAMA DE IMPLEMENTACIÓN TECNOLÓGICA

Mayo 2020

Etapa	Sub-etapa	Descripción	Equipo de trabajo "básico"			Equipo de trabajo "intermedio"			Equipo de trabajo "avanzado"		
			Grado de automatización	Competencias	Perfil	Grado de automatización	Competencias	Perfil	Grado de automatización	Competencias	Perfil
I. Estrategia y diseño	Análisis y exploración de las diversas fuentes de datos	Se llevará a cabo un análisis exhaustivo de las diversas fuentes de datos de interés, tanto públicas como privadas, llenando una ficha o "scorecard" de acuerdo a sus requerimientos técnicos en función de su naturaleza, estructura y calidad	**Dirección - expertos**								
	Diseño y presentación de la estrategia global de automatización en el corto, mediano y largo plazo	Una vez exploradas y analizadas las fuentes de datos, se tendrán los elementos para el diseño completo del desarrollo									
II. Extracción	Extracción de fuentes de internet	Según el tipo de fuente, se definirán las tecnologías más adecuadas para la automatización en la extracción de los datos	Automatización en la extracción ("Scraping") de páginas web con estructuras similares (RSS, html), principalmente de los procesos de alta prioridad y con alta periodicidad (diaria, semanal)	Conocimiento en el uso de plugins de extracción nivel básico (ej. ImportHTML, Scraper-Chrome, Parsers, OutWit Hub)	**Ingeniero de Software Jr./Ingeniero de Datos Jr.**	Automatización en la extracción ("Scraping") de diferentes tipos de fuentes principalmente de medios masivos. Se pueden analizar y extraer datos de algunas fuentes con mayor grado de complejidad, las cuales requieren una intervención más "especifica"	Conocimiento en el uso de herramientas de extracción nivel intermedio (ej. Import.io, Webscraper.io, Octoparse, ParseHub)	**Ingeniero de Datos Jr.**	Grado de automatización óptima basado en progrmación ad hoc (códigos en distintos niveles de prioridad) para la ejecución e integración de los diversos algoritmos de extracción de cada una de las fuentes de datos.Creación de rutinas y sub-rutinas para la extracción con el uso de diversas librerías.	Dominio y experiencia en progrmación con Python en el uso de herramientas y librerías de extracción nivel avanzado (ej. Scrapy, BeautifulSoup, Selenium, Puppeteer)	**Ingeniero de Datos Sr**
III. Pre-procesamiento	Almacenamiento de la información	Se analizarán las mejores alternativas para el alacenamiento de la información en función de la cantidad y características de los datos.	Almacenamiento en infraestructura local. Se define y configura un repositorio con campos básicos como: fecha, título, palabras claves, actor de interés (personaje, institución), palabras claves, etc.	Almacenamiento en infraestructura actual con tecnologías y métodos convenientes anticipando un futuro escalamiento y aspectos de seguridad		Semi-automatización en el desarrollo de infraestructura de almacenamiento de datos, detallando e implementando los requerimientos y configuraciones necesarios para el futuro escalamiento en Nube o servidores	Conocimientos en la configuración y manejo de una o varias tecnologías con servicios en la NUBE (ej. AWS, Google Cloud, Watson, Azure). Alternativamente, tener conocimientos en el manejo y configuración de servidores.		Automatización integral en el proceso de almacenamiento basada en tecnología robusta y escalable. Se tienen definidas completamente los protocolos de comunicación, configuraciones de conexión y pleno conocimiento del esquema de costos en función de los recursos requeridos	Especialización en el manejo y configuración integral de bases de datos en servidores y Nube para el escalamiento. Conocimientos de ciber-seguridad.	**Ingeniero de Datos Jr.**
	Modelo y estructura de bases de datos	Dependiendo del proceso de negocio y la naturaleza de los datos (estructurados, semi-estructurados, no-estructurados) se definirá el o los modelos de bases de datos óptimos que garantice la robustez y escalabilidad. Desarrollo de protocolos de integración y consultas	Infraestructura básica y empleo de tecnologías de uso común. Bajo grado de automatización para los procesos de negocio específicos de la empresa	Manejo de bases de datos más comunes (ej. Excel, MySQL)		Integración de bases de datos, relacionando y no relacionales. Se define un modelo de bases de datos con el potencial de ser escalado y robusto en el largo plazo	Conocimiento en el manejo y control de acceso de bases de datos estructurados y no estructurados (ej. MongoDb, Cassandra, Postgresql)		Automatización integral de la amplia variedad y volumén de datos, big data.	Conocimientos de tecnologías avanzadas para el manejo de Big Data (ej. Data lake, data WareHouse, ETL)	
	Técnicas de análisis de la calidad de la información y detección de inconsistencias y errores	Evaluación de la calidad de la información, exploración completa de los datos en búsqueda de errores, inconsistencias, unidades de medición, posibles relaciones internas, etc.	NA	Manejo de técnicas estadísticas descriptivas básicas		Algoritmos computacionales y estadísticas semi-automatizados para la evaluación de la calidad de la información	Manejo de técnicas estadísticas descriptivas e inferenciales básicas (pruebas de hipótesis, intervalos de confianza, pruebas específicas		Técnicas computacionales y estadísticas avanzadas para la evaluación de la calidad de la información	Conocimiento profundo de técnicas de estadística inferencial y multivariantes	
IV. Análisis	Métodos analíticos "Web scraping de noticias"	Se elegirán las técnicas de análisis más adecuadas según las características definidas en los "scorecards" y los objetivos específicos de negocio	Estadísticas básicas de minería de textos, por ejemplo: conteo de incidencias, asociaciones básicas entre palabras o frases, normalización básica (mayúsculas-minúsculas, conversión de números a letras, espacios y caracteres espaciales). Se definenen los criterios para la generación de catálogos de los procesos de negocio más relevantes	Técnicas básicas de "Minería de datos", particularmente "Minería de textos" (incidencias/conteos/normalización/clasificación de textos, etc.)	**Data scientist Sr.**	Se desarrollan algoritmos automatizados de normalización y clasificación con técnicas de nivel intermedio y avanzadas. Se hace énfasis en aprendizaje "no-supervisado" y para tareas ya validadas (conceptos y procesos bien definidos) se pueden aplicar algunas técnicas de aprendizaje "supervisado". Se disponen de catálogos completos de los procesos de negocios más relevantes	Aplicación de normalización y segmentación de textos. Conocimientos en técnicas de clasificación de uso común y de mediana complejidad (regresión logística, árboles de decisión, k-means)	**Ingeniero de Software Jr./Ingeniero de Datos Jr./Data scientist Jr., (1) Data Scientist Sr.**	Se desarrollan algoritmos complejos de Procesamiento de Lenguaje Natural y de Inteligencia Artificial. Se disponen en este momento de catálogos con criterios de búsqueda completos como guías para la ejecución de los algoritmos	Aplicación de técnicas avanzadas de análisis de textos, como: Procesamiento de Lenguaje Natural (PLN), técnicas avanzadas de normalización (n-gramas, segmentación), clasificación mediante Máquinas de Soporte Vectorial (SVM), "Random Forest" o STM, finalmente la Representación de Documentos mediante enfoque "Léxico" o "Semántico", Análisis de Sentimientos	**Ingeniero de Software Jr./Ingeniero de Datos Jr./Data scientist Jr., (1) Data Scientist Sr.**
	Sistema de Alertas	Deberán desarrollarse algoritmos que permitan identificar comportamientos atípicos de los resultados . Asimismo, aplicación de técnicas para la medición de la calidad y precisión de los resultados en función de otros datos de referencia o históricos. En este sentido es importante definir "alertas" basados en fundamentos estadísticos o reglas de decisión arbitrarios según el expertise del analista	NA	Conocimientos de estadística inferencial básica		Generación de alertas para indicadores básicos (incidencias, asociaciones entre palabras-frases de interés) de fuentes validadas (estructura e información robusta y consistente)	Conocimientos de estadística inferencial para la comparativa y evaluación de resultados basada en criterios de calidad, robustez y precisión		Algoritmos para la identificación automática de cambios estructurales o eventos atípicos con relación a criterios como el histórico o referencias específicas basado en umbrales de tolerancias o reglas de decisión	Conocimientos avanzados de estadística inferencial para la comparativa y evaluación de resultados basada en criterios de calidad, robustez y precisión	
V. Visualización	Métodos para la visualización	Algunos resultados tendrán atributos geoespaciales, otros serán indicadores simples y compuestsos de seguimiento o monitoreo, otros serán índices representativos de un fenómenos más complejo y estarán en una escala definida (ej. 1-100), o bien "rankings", "ratings", "scores", "probabilidades", etc. Por lo tanto, resulta muy importante seleccionar correctamente las técnicas para representar adecuadamente el fenómeno de estudio	Carga manual o semi-automática de archivos en softwarres especializados (comerciales u "open source") según el tipo de archivo "output", ya sea geoespacial o "data-frame". Disponibilidad de herramientas limitadas.	Generación de elementos visuales personalizados basados en plataformas científicas (ej. Python, R, Matlab)	*** Ingeniero de Software Jr./Ingeniero de Datos Jr./Data scientist Jr.**	Semi-automatización de la integración entre los algoritmos de los lenguajes de programación y las diversas herramientas de visualización. Se trabaja particularmente en dar una estructura y formato convenientes a los "outputs" de los análisis para una lectura "automática" y "actualizable"	Conocimientos en el manejo de softwares especializados como: GIS (geoespacial), Tableau (general), Shiny (ej. R). Deberá saber extraer y generar conexiones automatizadas con los algoritmos generadores	**Data scientist Jr.**	Integración completa de los "outputs" de los análisis con sus respectivas visualizaciones finales, sin la necesidad de intervenir en la transformación de formatos/estructuras, ni en la carga o transferencia de los datos	Alta especialización para el desarrollo de sistemas integrados que interactúen con los algoritmos generadores de tablas o archivos. Aglunos conocimientos deseables son: R-Shiny nivel avanzado, desarrollo de sistemas integrales locales (C, Java, etc.) y en web (html, CSS, etc.), manejo e integración de datos geoespaciales (Geoserver, cesium, etc.).	**Data scientist Sr./Ingeniero de Software Jr.**