**Categorías de las Herramientas de la Ciencia de los Datos**

Las herramientas de código abierto están disponibles para diversas tareas de la ciencia de los datos.

En este video, veremos las diferentes tareas de la ciencia de los datos.

En los siguientes videos veremos las herramientas de código abierto más usadas para esas tareas. Las herramientas más importantes se cubren a lo largo de este curso.

La Gestión de Datos es el proceso de persistencia y recuperación de datos.

La Integración y Transformación de Datos, comúnmente conocida como Extracción, Transformación y Carga o por sus siglas en inglés “ETL”, es el proceso de recuperación de datos de sistemas de gestión de datos remotos.

La transformación y carga de los datos en un Sistema local de gestión de datos también es parte de la Integración y Transformación de Datos.

La Visualización de Datos es parte de un proceso inicial de exploración de datos, así como parte de un producto final.

La Construcción de Modelo es el proceso de creación de un modelo de aprendizaje automático o aprendizaje profundo usando un algoritmo apropiado con muchos datos.

El despliegue del modelo permite poner a disposición a las aplicaciones de terceros un modelo de aprendizaje automático o de aprendizaje profundo.

El seguimiento y evaluación de los modelos, asegura un control continuo de la calidad del rendimiento de los modelos desplegados.

Estas verificaciones son para la exactitud, equidad y la robustez del adversario.

La gestión de activos de código usa el control de versiones y otras características de colaboración para facilitar el trabajo en equipo.

La gestión de activos de datos implica los mismos componentes de versiones y de colaboración en los datos.

La gestión de activos de datos a su vez apoya la replicación, copia de seguridad y la gestión de los derechos de acceso.

Los entornos de desarrollo, comúnmente conocidos como Entornos de Desarrollo Integrado o “IDEs”, son herramientas que ayudan al científico de datos a implementar, ejecutar, probar y desplegar su trabajo.

Los entornos de ejecución son herramientas en las que se realiza el preprocesamiento de datos, la capacitación de modelos y el despliegue.

Finalmente, hay disponible una herramienta visual completamente integrada que cubre todos los componentes de la herramienta previa, bien sea parcial o por completo.

**Herramientas de Código Abierto para la Ciencia de los Datos - Parte 1**

En la parte uno de esta serie de dos partes, cubriremos la gestión de datos, integración de datos de código abierto, transformación y las herramientas de visualización.

Las herramientas de gestión de datos de código abierto más usadas son las bases de datos relacionales como MySQL y PostgreSQL; NoSQL, MongoDB Apache CouchDB y Apache Cassandra

y las herramientas basadas en archivos como Hadoop File System o los sistemas de archivos en la nube como Ceph.

Por último, Elasticsearch es principalmente usado para almacenar datos de texto y crear un índice de búsqueda para una rápida recuperación de documentos.

La tarea de integración de datos en el mundo clásico de almacenamiento de datos es llamado ETL que según sus siglas quiere decir “extraer, transformar y cargar”.

Hoy en día los científicos de datos generalmente proponen el término “ELT” – Extraer, Cargar y Transformar “ELT”

resalta el hecho de que los datos son volcados en algún lugar y el ingeniero de datos o el mismo científico de datos es el responsable de los datos.

Hay otro termino para este proceso que ha surgido actualmente: “refinamiento y limpieza de datos”.

Estas son las herramientas de integración y transformación de datos de código abierto más usadas:

Apache AirFlow, originalmente creada por AirBNB; KubeFlow la cual permite ejecutar las tuberías de datos de la ciencia de los datos en Kubernetes; Apache Kafka la cual se originó de LinkedIn,

Apache Nifi la cual ofrece un editor visual amigable, Apache SparkSQL (el cual te permite usar ANSI SQL y escalar para computar clústeres de 1000s de nodos) y NodeRED el cual también proporciona un editor visual.

NodeRED consume tan pocos recursos que incluso funciona en pequeños dispositivos como Raspberry Pi.

Bueno, ahora presentaremos las herramientas de visualización de datos de código abierto más usadas.

Debemos distinguir entre las librerías de programación en las que debes usar el código y las herramientas que contienen una interfaz de usuario.

Las librerías más populares son cubiertas en los próximos videos.

Pixie Dust es una librería, pero viene con una interfaz de usuario la cual ayuda enormemente a cualquiera quien necesite graficar en Python.

Hue tiene un enfoque similar el cual puede crear visualizaciones a partir de consultas en SQL.

Kibana, es una aplicación web de exploración y visualización de datos está limitada a Elasticsearch

(el proveedor de datos).

Finalmente, Apache Superset es una aplicación web de exploración y visualización de datos.

El despliegue del modelo es extremadamente importante. Una vez que has creado un modelo de aprendizaje automático capaz de predecir algunos aspectos claves del futuro, deberías hacer que ese modelo sea consumible por otros desarrolladores luego convertirlo a una API.

* Apache PredictionIO actualmente solo soporta modelos Apache Spark ML para el despliegue pero que sea soportada por cualquier tipo de librería está en camino.
* Seldon es un producto interesante ya que soporta todos los marcos de trabajo concluyendo TensorFlow, Apache SparkML, R, y scikit-learn.
* Seldon puede funcionar en Kubernetes y Redhat OpenShift.
* Otra manera de desplegar los modelos de SparkML es utilizando MLeap.
* Finalmente, TensorFlow puede servir con cualquiera de sus modelos utilizando el servicio de TensorFlow. Puedes desplegarlo en un dispositivo integrado como Raspberry Pi o un teléfono inteligente usando TensorFlow Lite e incluso desplegarlo en un navegador web usando TensorFlow dot JS. El monitoreo del modelo es otro paso crucial.

Una vez que has desplegado un modelo de aprendizaje automático, es necesario hacer un seguimiento de su desempeño de predicción a medida que los nuevos datos llegan con el fin de mantener los modelos desactualizados.

A continuación, encontrarás algunos ejemplos de herramientas de seguimiento de modelos:

ModelDB es una base de metadatos de modelos de máquinas donde la información sobre los modelos puede ser almacenada y consultada.

Nativamente soporta Apache Spark ML Pipelines y scikit-learn.

Una herramienta genérica multi propósitos llamada Prometheus es también usada ampliamente para el seguimiento del modelo de aprendizaje automático, aunque no está hecha específicamente para este propósito. El desempeño del modelo no es medido exclusivamente con respecto a la precisión.

Los modelos de tendencias contra grupos protegidos como el género o la raza también son importantes.

El kit de herramientas de código abierto de IBM AI Fairness 360 hace precisamente esto.

Detecta y mitiga las tendencias en los modelos de aprendizaje automático.

Los modelos de aprendizaje automático, especialmente los modelos de aprendizaje profundo basados en las redes neuronales que pueden ser objeto de ataques adversos donde un atacante trata de engañar al modelo con datos manipulados o manipulando el propio modelo.

El IBM Adversarial Robustness 360 Toolbox puede ser usado para detectar la vulnerabilidad de los ataques adversarios y ayuda a hacer mucho más robusto el modelo.

Los modelos de aprendizaje automático frecuentemente son considerados como una caja negra que aplica algo de “magia” misteriosa.

El kit de herramientas IBM AI Explainability 360 hace el proceso de aprendizaje automático más entendible al encontrar ejemplos similares dentro del conjunto de datos que puede ser presentado a un usuario para su comparación de forma manual.

El kit de herramientas IBM AI Explainability 360 también puede ilustrar la capacitación de un modelo de aprendizaje automático más simple al explicar cómo las diferentes variables de entrada afectan la decisión final del modelo.

Las opciones de las herramientas de gestión de activos de código han sido ampliamente simplificadas:

Para la gestión de activos de código – también conocida como gestión de versiones o control de versiones

– Git es hoy en día el estándar.

Han surgido múltiples servicios para soportar a Git, siendo claramente el más destacado GitHub, el cual proporciona alojamiento para la gestión de versiones de desarrollo de software.

El segundo puesto es definitivamente para GitLab el cual ha tenido la ventaja de ser una plataforma completamente de código abierto que puedes alojar y administrar tú mismo.

Otra opción interesante es Bitbucket.

La gestión de activos de datos también conocida como gobernanza de datos o linaje de datos es otra parte crucial de la ciencia de los datos a nivel empresarial. Los datos tienen que ser versionados y anotados con metadatos.

Apache Atlas es una herramienta que soporta esta tarea.

Otro proyecto interesante es ODPi Egeria el cual es gestionado a través de Linux Foundation y es un entorno abierto. Este ofrece un conjunto de APIs abiertas, tipos y protocolos de intercambio que los repositorios de metadatos usan para compartir e intercambiar los datos.

Y por último, Kylo que es una plataforma de software de código abierto para la gestión de Data Lake que provee un amplio soporte para la gran variedad de tareas de gestión de activos de datos.

**Herramientas de Código Abierto para la Ciencia de los Datos - Parte 2**

Bienvenido a la segunda parte de esta serie.

En esta sección, vamos a cubrir el entorno de desarrollo, la integración de datos de código abierto, transformación y herramientas de visualización.

Uno de los entornos de desarrollo actualmente más populares que los científicos de datos están utilizando es “Jupyter”.

Jupyter surgió primeramente como una herramienta para la programación interactiva en Python, ahora soporta más de cien lenguajes de programación por medio de “kernels”.

Kernels no deberían ser confundidos con los kernels del sistema operativo.

Los Jupyter kernels están encapsulando los distintos interpretes interactivos para los diferentes lenguajes de programación.

Una propiedad clave de Jupyter Notebooks es la habilidad de unificar la documentación, código, salida del código, los interpretes de comandos y las visualizaciones en un solo documento.

JupyterLab es la siguiente generación de Jupyter Notebooks y a largo plazo reemplazará

a Jupyter Notebooks.

Los cambios a nivel de arquitectura que se están introduciendo en JupyterLab hacen que Jupyter sea más moderno y modular.

Desde la perspectiva del usuario, la principal diferencia introducida por JupyterLab es la habilidad de abrir distintos tipos de archivos, incluyendo Jupyter Notebooks, datos y terminales.

Entonces puedes arreglar estos archivos en el canvas.

Aunque Apache Zeppelin ha sido completamente reimplantado, este está inspirado por Jupyter Notebooks y proporciona una experiencia similar.

Un elemento clave es la capacidad de graficar integrado. En Jupyter Notebooks, es necesario el uso de librerías externas en Apache Zeppelin y el graficar no requiere de un código.

También puedes extender estas capacidades mediante el uso de librerías adicionales. RStudio es uno de los entornos de desarrollo más antiguos para la estadística y la ciencia de los datos, siendo introducida en el año 2011. Es exclusivamente ejecutada en R y todas las librerías asociadas a R.

Sin embargo, el desarrollo en Python es posible y R está estrechamente integrado en esta

herramienta para proporcionar una experiencia de usuario óptima.

RStudio unifica la programación, ejecución, depuración, acceso remoto a los datos, exploración de datos y visualización en una sola herramienta.

Spyder intenta imitar el comportamiento de RStudio para llevar su funcionalidad al mundo de Python. Sin embargo, Spyder no tiene el mismo nivel de funcionalidad que RStudio, los científicos de datos lo consideran como una alternativa. Pero en el mundo de Python, Jupyter es usado con más frecuencia.

Este diagrama muestra cómo Spyder integra el código, la documentación, las visualizaciones y otros componentes en un simple canvas.

A veces tus datos no entran en la capacidad de almacenamiento o memoria principal de una sola computadora. Es allí donde entran los entornos de ejecución de clúster.

El bien conocido marco de trabajo de computación en clúster Apache Spark es uno de los proyectos Apache más activos y es usado a lo largo de todas las industrias, incluyendo muchas compañías en la lista de Fortune 500.

La propiedad clave de Apache Spark es la escalabilidad lineal. Lo que quiere decir que, si duplicas el número de servidores en un clúster, también duplicarás su desempeño.

Después de que Apache Spark comenzó a ganar cuota de mercado y Apache Flink fue creado.

La diferencia clave entre Apache Spark y Apache Flink es que Apache Spark es un motor

de procesamiento de datos por lotes, capaz de procesar grandes cantidades de archivos de datos por archivo.

Apache Flink, por otra parte, es una imagen de procesamiento de flujo de datos con su enfoque principal en el procesamiento de flujos de datos en tiempo real.

Aunque el motor soporta ambos paradigmas de procesamientos de datos, Apache Spark es generalmente la opción en la mayoría de los casos de uso.

Uno de los últimos desarrollado en los entornos de ejecución de la ciencia de los datos es llamado “Ray”, el cual se enfoca claramente en la formación de modelos de aprendizaje profundo a gran escala.

Echemos un vistazo a las herramientas de código abierto para los científicos de datos que están completamente integradas y son visuales.

Con estas herramientas, no necesitas conocimientos de programación.

Las tareas más importantes son soportadas por estas herramientas, estas tareas incluyen la integración de datos, la transformación, la visualización de datos y la construcción de modelos.

KNIME desarrollado en la Universidad de Constanza en el año 2004. Como puedes ver KNIME tiene una interfaz de usuario visual con la habilidad de arrastrar y soltar.

También tiene capacidades de visualización ya incorporadas. Knime puede ser extendida para programar con R y Python y tiene conectores con Apache Spark.

Otro ejemplo de este grupo de herramientas es Orange. Es menos flexible que KNIME pero mucho más fácil de usar.

En esta sección, has aprendido acerca de las tareas más comunes en la ciencia de los datos y cuales herramientas de código abierto son relevantes para estas tareas.

En el próximo video, discutiremos algunas herramientas comerciales establecidas que encontrarás en tu experiencia con la ciencia de datos. Vamos al siguiente video para obtener más detalles.

**Herramientas Comerciales para la Ciencia de los Datos**

Previamente cubrimos las herramientas de código abierto para la ciencia de los datos.

Ahora, veamos las opciones comerciales que encontrarás en muchos proyectos empresariales.

Vamos a repasar nuestra visión general de las distintas categorías de herramientas.

En la gestión de datos, la mayoría de los datos relevantes de una empresa se almacenan en una base de datos como lo son Oracle Database, Microsoft SQL Server, o IBM Db2.

Aunque las bases de datos de código abierto están ganando popularidad, esos tres productos de gestión de datos se siguen considerando como los estándares de la industria.

Lo que quiere decir que no desaparecerán en un futuro cercano. No se trata solamente de funcionalidad. Los datos son el corazón de cualquier organización y la disponibilidad de soportes comerciales juegan un papel muy importante.

El soporte comercial es ofrecido directamente por los proveedores de software, los socios influyentes y las redes de soporte.

Cuando nos enfocamos en las herramientas comerciales de integración de datos, hablamos de herramientas de “extracción, transformación y carga” o “ETL”.

De acuerdo a Gartner Magic Quadrant (Cuadrante Mágico de Gartner), Informatica Powercenter e IBM InfoSphere DataStage son los líderes, seguidos por los productos de SAP, Oracle, SAS, Talend y Microsoft.

Estas herramientas apoyan el diseño y despliegue de las tuberías de procesamiento de datos de ETL a través de una interfaz gráfica.

También proporcionan conectores a la mayoría de los sistemas de información de destino comercial al igual que para los de código abierto.

Finalmente, Watson Studio Desktop incluye un componente llamado Data Refinery, el cual permite La definición y ejecución de procesos de integración de datos en un estilo de hoja de cálculo.

En el ámbito comercial, las visualizaciones de datos utilizan herramientas de inteligencia empresarial o “BI”. Su enfoque principal es crear informes visualmente atractivos y fáciles de comprender y tableros de control en vivo. Los ejemplos comerciales más importantes son: Tableau, Microsoft Power BI e IBM Cognos Analytics.

Otro tipo de visualización se destina a los científicos de los datos en lugar de a los usuarios habituales.

Un problema de muestra podría ser “¿Cómo pueden relacionarse las distintas columnas de una tabla?”. Este tipo de funcionalidad está incluida en Watson Studio Desktop.

Si quieres construir un modelo de aprendizaje automático usando una herramienta comercial, deberías considerar el uso de un producto de minería de datos.

Los más importantes de estos tipos de productos son: SPSS Modeler y SAS Enterprise Miner.

Además, una versión de SPSS Modeler está igualmente disponible en Watson Studio Desktop, basada en la versión de la nube la de herramienta.

Discutiremos más sobre las herramientas basadas en la nube en el próximo video.

En el software comercial, el despliegue de modelos está estrechamente integrado en el proceso de construcción de modelos. Este diagrama muestra un ejemplo de SPSS Collaboration and Deployment Services en el cual son usados para desplegar cualquier tipo de activo creado por la suite de herramientas de Software SPSS. Otros proveedores usan el mismo tipo de proceso.

Los softwares comerciales igualmente pueden exportar modelos en formato abierto. Ejemplo, SPSS Modeler permite la exportación de modelos como Predictive Model Markup Language,

o PMML, el cual puede ser leído por numerosos paquetes de software comerciales y abiertos.

Model monitoring es una nueva disciplina y actualmente no existen herramientas comerciales relevantes disponibles. Por lo tanto, el código abierto es la primera opción. Esto mismo pasa con la gestión de activos de código. El código abierto con Git y GitHub es el estándar efectivo.

La gestión de activos de código, con frecuencia llamado gobernanza de datos o linaje de datos, es una parte fundamental de la ciencia de datos a nivel empresarial.

Los datos deben ser versionados y anotados utilizando metadatos. Los proveedores, incluyendo Informatica Enterprise Data Governance e IBM, proporcionan herramientas para estas tareas específicas.

IBM InfoSphere Information Governance Catalog cubre las funciones como el diccionario de datos, lo cual facilita el descubrimiento de los activos de datos.

* Cada activo de datos se asigna a un administrador de datos - - el propietario de los datos.
* El propietario de los datos es responsable de ese activo de datos y puede ser contactado.
* El linaje de datos también es cubierto; esto permite al usuario hacer un seguimiento de los pasos de transformación seguidos en la creación de los activos de datos.
* El linaje de datos también incluye una referencia a los daos de la fuente real.
* Se pueden agregar reglas y políticas para reflejar los complejos requisitos normativos y comerciales de privacidad y retención de datos.

Watson Studio es un entorno de desarrollo completamente integrado para los científicos de datos. Es normalmente consumido a través de la nube, y cubriremos más sobre este en una siguiente lección. También hay una versión de escritorio disponible.

Watson Studio Desktop combina Jupyter Notebooks con herramientas gráficas para maximizar el rendimiento de los científicos de datos.

Watson Studio, conjuntamente con Watson Open Scale, es una herramienta completamente integrada que cubre el ciclo de vida completa de la ciencia de los datos y todas las tareas que hemos discutido previamente.

Hablaremos más de ambos en la siguiente lección.

Tan solo ten en cuenta que pueden ser desplegados en un centro de datos local que pueden ser Kubernetes o RedHat OpenShift.

Otro ejemplo de una herramienta comercial totalmente integrada es H2O Driverless AI, la cual cubre el ciclo de vida completo de la ciencia de los datos.

En esta lección, has aprendido cómo las tareas más comunes de la ciencia de los datos son soportadas por las herramientas comerciales.

En el siguiente video, descubriremos herramientas de la ciencia de los datos que están disponibles exclusivamente en la nube.

**Herramientas basadas en la nube para la ciencia de los datos**

Ya que hemos cubierto previamente las herramientas de código abierto para la ciencia de los datos, veamos las opciones comerciales que encontrarás en los diferentes proyectos empresariales.

Echa otro vistazo a la visión general de las diferentes categorías de herramientas. Debido a que los productos de nubes son una especie más reciente, estos siguen la tendencia de tener múltiples tareas integradas en las herramientas. Esto es especialmente válido para las tareas marcadas con el color verde en el diagrama.

Comencemos con la categoría de herramientas visuales completamente integradas. Ya que estas herramientas introducen un componente donde la ejecución a gran escala de flujos de trabajo de la ciencia de los datos ocurre en un clúster computacional, hemos cambiado el título y agregado la palabra “Plataforma”.

Estos clústeres están compuestos por múltiples servidores, de forma transparente para el usuario, en segundo plano.

Watson Studio, junto con Watson OpenScale, cubre el ciclo de vida completa de desarrollo para todas las tareas de la ciencia de los datos, aprendizaje automático e inteligencia artificial.

Otro ejemplo particular es Microsoft Azure Machine Learning. Esta es también una opción alojada en la nube que apoya el ciclo de vida completo de desarrollo de todas las tareas de ciencia de los datos, aprendizaje automático e inteligencia artificial. Y por último, otro ejemplo es H2O Driverless AI, el cual ya hemos presentado en el último video.

Aunque es un producto que se descarga e instala, el despliegue con un solo clic está disponible para los proveedores de servicios de nubes comunes. Debido a que las operaciones y mantenimiento no son hechas por el proveedor de la nube, como es el caso de Watson Studio, Open Scale, y Azure Machine Learning, este modelo de entrega no debería ser confundido con la plataforma o el Software como un Servicio -- PaaS o SaaS.

En la gestión de datos, con algunas excepciones, existen versiones de SaaS de las herramientas de código abierto y comerciales existentes. Recuerda, SaaS quiere decir “Software como un servicio”. Lo que quiere decir que el proveedor de la nube opera la herramienta por ti en la nube. Como por ejemplo, el proveedor de la nube opera el producto haciendo una copia de seguridad de tus datos, configuración e instalación de actualizaciones.

Como se ha mencionado anteriormente, hay herramientas de propiedad, que sólo están disponibles como un producto de la nube. A veces sólo está disponible por un único proveedor de la nube.

Un ejemplo particular de este servicio es Amazon Web Services DynamoDB, que es una base de datos NoSQL que permite el almacenamiento y recuperación de los datos en un formato de valor clave (key-value) o de almacenamiento de documentos. La estructura de datos de documento más relevante es JSON (pronunciado “yey-son”).

Otro ejemplo de este servicio es Cloudant, el cual es una base de datos como un servicio. Pero, en su interior está basado en el código abierto Apache CouchDB. Tiene una ventaja: aunque las tareas operativas complejas como la actualización, copia de seguridad, restauración y el escalado, las realiza el proveedor de la nube, en su interior esta opción es compatible con CouchDB. Por lo tanto, la aplicación puede ser migrada a otro servidor de CouchDB sin cambiar la aplicación.

De igual forma, IBM ofrece Db2 como un servicio. Este es un ejemplo de una base de datos comercial que se ofrece como un software como un servicio en la nube, quitándole al usuario las tareas operativas.

Cuando se trata de herramientas de integración de datos comerciales, no sólo hablamos de herramientas de “extracción, transformación y carga” o “ETL”, sino también de herramientas “extracción, carga y transformación” o mejor dicho “ELT”. Esto significa que los pasos de transformación no son hechos por un equipo de integración de datos, sino que son impulsados hacia el dominio del científico de datos o ingeniero de datos.

Dos herramientas de integración de datos comerciales muy usadas son informática Cloud Data Integration y Data Refinery de IBM. Data Refinery permite transformar grandes cantidades de datos brutos en información consumible y de calidad en una interfaz de usuario similar a una hoja de cálculo. Data Refinery es parte de IBM Watson Studio.

El mercado de herramientas de visualización de datos en la nube es inmenso y todos los principales proveedores de nubes tienen una.

Un ejemplo de una empresa más pequeña basada en la herramienta de visualización en la nube es DataMeer. IBM ofrece su famosa suite de Cognos Business Intelligence como una solución en la nube también.

IBM Data Refinery también ofrece la funcionalidad de exploración y visualización de en Watson Studio.

Una vez más, estos son sólo algunos ejemplos de un ecosistema comercial en rápido cambio y crecimiento entre un enorme número de proveedores ya establecidos al igual que los emergentes.

En Watson Studio, se pueden usar una gran cantidad de visualizaciones diferentes para comprender mejor los datos. Por ejemplo, este gráfico de barras en 3D permite visualizar un valor objetivo en la dimensión vertical, que depende de otros dos valores en las dimensiones horizontales. El color te permite visualizar una tercera dimensión. La agrupación de limites jerárquicos te permiten visualizar las correlaciones y afiliaciones entre entidades. Si es suficiente, un gráfico de barras clásico puede hacer el trabajo, mientras que un gráfico de dispersión en 2D con un mapa de calor muestra dos campos de datos dependientes, uno en el eje “Y” y otro como intensidad de color.

Un diagrama de árbol muestra la distribución de los subconjuntos dentro de un conjunto, el famoso gráfico de torta hace lo mismo, pero no de forma jerárquica y finalmente, una nube de palabras hace aparecer términos significativos en un corpus de documentos.

La construcción de modelos se puede hacer usando un servicio como Watson Machine Learning. Watson Machine Learning puede entrenar y construir modelos usando diferentes librerías de código abierto.

Google tiene un servicio similar en su nube llamado AI Platform Training. Casi todos los proveedores de nubes tienen una solución para esta tarea. El despliegue de modelos en los softwares comerciales usualmente está estrechamente integrado en el proceso de construcción de modelos.

Aquí está un ejemplo de SPSS Collaboration and Deployment Services, el cual puede ser usado para desplegar cualquier tipo de activo creado por la suite de herramientas de Software SPSS.

Lo mismo ocurre con otros proveedores. Además, los softwares comerciales pueden exportar modelos en un formato abierto. Por ejemplo, SPSS Modeler permite la exportación de modelos como Predictive Model Markup Language, o “PMML,” el cual puede ser leído por numerosos paquetes de software comerciales.

Watson Machine Learning también puede ser usado para desplegarse un modelo y ponerlo a disposición de los consumidores utilizando una interfaz REST.

Amazon SageMaker Model Monitor es un ejemplo de una herramienta de nube que hace seguimiento continuamente de los modelos de aprendizaje automático y aprendizaje profundo. De nuevo, todos los grandes proveedores de nubes tienen herramientas similares. También es el caso para Watson OpenScale. OpenScale y Watson Studio… …unifican el entorno.

Todo lo marcado en color verde puede ser hecho usando Watson Studio y Watson OpenScale. Discutiremos Open Scale en un próximo video. Has aprendido cómo las tareas más comunes en la ciencia de los datos son soportadas por las herramientas comerciales de la nube.

La integración nos proporciona la capacidad de utilizar las mismas herramientas para múltiples tareas. En los próximos videos, discutiremos los paquetes, APIs, conjuntos de datos y los modelos para la ciencia de los datos.