HERRAMIENTAS PARA DATA SCIENCE

TABLA DE CONTENIDO

1. CATEGORÍAS DE HERRAMIENTAS.....3
2. HERRAMIENTAS DE CODIGO LIBRE.....5
3. HERRAMIENTAS EN LA NUBE.....10

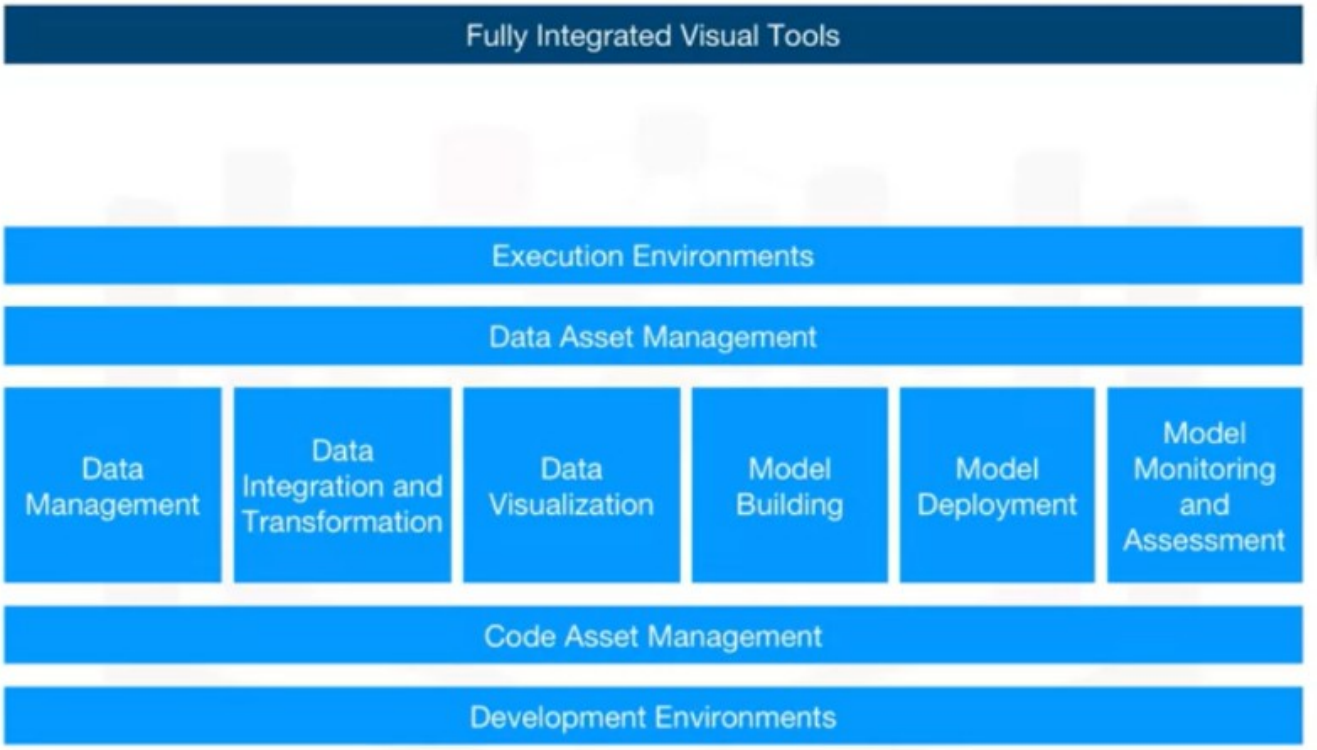
1. CATEGORÍAS DE HERRAMIENTAS

Se tiene:

- La administración de datos
  - Proceso de persistencia y recuperación de datos.
- Integración y transformación de datos (ETL)
  - Extraer, transformar y cargar.
  - Es el proceso de recuperación de datos de sistemas remotos, transformarlos y cargarlos en un sistema de gestión local.
- Visualización de datos
  - Es parte de un proceso inicial de exploración de datos, así como de una entrga final.
- Cosntrucción de modelos
  - Proceso de creación de un modelo de aprendizaje automático o profundo utilizando un algoritmo apropiado y determinada cantidad de datos.
- Implementación de modelos
  - Hace que el modelo esté disponible para aplicaciones de terceros.
- Supervisión y evaluación de modelos
  - Garantizan controles continuos de calidad del rendimiento en los modelos implementados.
  - Son para exactitud, equidad y solidez.
- Gestión de activos de código
  - Utiliza el control de versiones y otras funciones colaborativas para facilitar el trabajo en equipo.
- Administración de activos de datos
  - Aporta los mismos componentes de control de versiones y colaboración a los datos.
  - También admite replicación, backup y administración de derechos de acceso.
- IDEs
  - Ayudan al desarrollador a implementar, ejecutar y testear su trabajo.
- Entornos de ejecución

- Son herramientas en las que se lleva a cabo el preprocesamiento de datos, la formación de modelos y la implementación.

- Herramientas visuales integradas que cubren todos los componentes de herramientas anteriores ya sea parcial o completamente.



2. HERRAMIENTAS DE CODIGO LIBRE

Las **herramientas de gestión de datos** de código abierto más utilizadas son:

- Bases de datos relacionales
  - MySQL, PostgreSQL,..
- Bases de datos no relacionales
  - MongoDB, Apache CouchDB, Apache Cassandra
- Herramientas basadas en archivos
  - Hadoop File System, Cloud File Systems (ej. Ceph)
- Elasticserach
  - Se utiliza principalmente para almacenar datos de texto y crear un índice de búsqueda para la recuperación rápida de documentos.

La tarea de **ETL** hace hincapié en el hecho de que los datos se vuelcan en algún lugar y el ingeniero de datos es responsable de los mismos. Otro término para este proceso ha surgido ahora: «refinería y limpieza de datos».

Las herramientas son:

- Apache AirFlow
- KubeFlow
  - Permite ejecutar canalizaciones de ciencia de datos sobre Kubernetes.
- ApacheKafka
- Apache Nifi
- Apache SparkSQL
  - Permite usar ANSI SQL y escalar hasta calcular clústeres de 1000s de nodos.
- NodeRed
  - Consume tan poco en recursos que incluso funciona en dispositivos pequeños como una Raspberry Pi.

Ahora presentaremos las **herramientas de visualización de datos**:

- Hue
  - Puede crear visualizaciones a partir de consultas SQL.
- Kibana.

    - Una aplicación web de exploración y visualización de datos, está limitada a Elasticsearch (el proveedor de datos).
- Apache Superset

La **implementación de modelos** es extremadamente importante. Una vez que haya creado un modelo de aprendizaje automático capaz de predecir algunos aspectos clave del futuro, debe hacer que ese modelo sea consumible por otros desarrolladores y convertirlo en una API.
Las herramientas son:

- Apache PredictionIO
    - Actualmente solo admite modelos de Apache Spark ML para la implementación, pero el soporte para todo tipo de bibliotecas está en la hoja de ruta.
- Seldon
    - Soporta casi todos los marcos, incluyendo TensorFlow, Apache SparkML, R y scikit-learn.
    - Puede correr encima de Kubernetes y Redhat OpenShift.
- Mleap
    - Implementa modelos SparkML.
- TensorFlow
    - Puede servir a cualquiera de sus modelos utilizando el servicio TensorFlow.
    - Puede implementarse en un dispositivo incrustado como un Raspberry Pi o un teléfono inteligente usando TensorFlow Lite, e incluso en un navegador web usando TensorFlow.js.

La **supervisión de modelos** es otro paso crucial. Una vez que haya implementado un modelo de aprendizaje automático, debe realizar un seguimiento de su rendimiento de predicción a medida que lleguen nuevos datos para mantener modelos obsoletos.
Las herramientas son:

- ModelDB
    - Es una base de metadatos del modelo de máquina donde se almacena información sobre los modelos y se puede consultar.

    - Es compatible de forma nativa con Apache Spark ML Pipelines y scikit-learn.
- Prometheus
    - Se utiliza ampliamente para el monitoreo de modelos de aprendizaje automático, aunque no está hecha específicamente para este propósito.

El rendimiento del modelo no se mide exclusivamente a través de la precisión. También es importante el sesgo modelo contra grupos protegidos como el género o la raza. El kit de herramientas de código abierto IBM AI Fairness 360 hace exactamente esto. Detecta y mitiga contra el sesgo en los modelos de aprendizaje automático. Los modelos de aprendizaje automático, especialmente los modelos de aprendizaje profundo basados en redes neuronales, pueden ser objeto de ataques adversarios, donde un atacante intenta engañar al modelo con datos manipulados o manipulando el modelo en sí. IBM Adversarial Robustness 360 Toolbox se puede utilizar para detectar la vulnerabilidad a los ataques adversarios y ayudar a que el modelo sea más robusto. Los modos de aprendizaje automático a menudo se consideran como una caja negra que aplica alguna misteriosa «magia». IBM AI Explainability 360 Toolkit hace que el proceso de aprendizaje automático sea más comprensible al encontrar ejemplos similares dentro de un conjunto de datos que se pueden presentar a un usuario para la comparación manual. El IBM AI Explainability 360 Toolkit también puede ilustrar la formación para un modelo de aprendizaje automático más simple explicando cómo las diferentes variables de entrada afectan a la decisión final del modelo.

Las opciones para **herramientas de gestión** de activos de código se han simplificado enormemente: para la gestión de activos de código, también conocida como gestión de versiones o control de versiones , **Git** es ahora el estándar. Varios servicios han surgido para dar soporte a Git, siendo el más destacado GitHub, que proporciona alojamiento para la gestión de versiones de desarrollo de software. El finalista es sin duda GitLab, que tiene la ventaja de ser una plataforma de código abierto que puede hospedar y gestionar usted mismo. Otra opción 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 datos de nivel empresarial. Los datos deben ser versionados y anotados con metadatos.

Las herramientas son:

- Apache Atlas

- ODPi Egeria
  - Ofrece un conjunto de API abiertas, tipos y protocolos de intercambio que los repositorios de metadatos utilizan para compartir e intercambiar datos.
- Kylo
  - Es una plataforma de software de gestión del lago de datos de código abierto que proporciona un amplio soporte para una amplia gama de tareas de gestión de activos de datos.

Uno de los entornos de desarrollo actuales más populares que los científicos de datos están utilizando es «Jupyter». Jupyter surgió por primera vez como una herramienta para la programación interactiva de Python; ahora soporta más de un centenar de lenguajes de programación diferentes a través de «kernels». Los kernels no deben confundirse con los kernels del sistema operativo. Los kernels de Jupyter están encapsulando los diferentes intérpretes interactivos para los diferentes lenguajes de programación. Una propiedad clave de Jupyter Notebooks es la capacidad de unificar documentación, código, salida del código, comandos de shell y visualizaciones en un solo documento. JupyterLab es la próxima generación de Jupyter Notebooks y, a largo plazo, en realidad reemplazará a Jupyter Notebooks. Los cambios arquitectónicos introducidos en JupyterLab hacen que Jupyter sea más moderno y modular. Desde la perspectiva del usuario, la principal diferencia introducida por JupyterLab es la capacidad de abrir diferentes tipos de archivos, incluyendo Jupyter Notebooks, datos y terminales.

A continuación, puede organizar estos archivos en el lienzo. Aunque Apache Zeppelin ha sido completamente reimplementado, está inspirado en Jupyter Notebooks y proporciona una experiencia similar. Un diferenciador clave es la capacidad de trazado integrada. En Jupyter Notebooks, es necesario usar bibliotecas externas en Apache Zeppelin, y el trazado no requiere codificación. También puede ampliar estas capacidades mediante el uso de bibliotecas adicionales.

RStudio es uno de los entornos de desarrollo más antiguos para la estadística y la ciencia de datos, ya que se introdujo en 2011. Ejecuta exclusivamente R y todas las bibliotecas R asociadas. Sin embargo, el desarrollo de Python es posible y, por lo tanto, R está estrechamente integrado en esta herramienta para proporcionar una experiencia de usuario

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

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

A veces, los datos no encajan en el almacenamiento de un solo ordenador o en la capacidad de memoria principal. Ahí es donde entran los entornos de ejecución de clústeres. El conocido marco de computación en clústeres Apache Spark se encuentra entre los proyectos más activos de Apache y se utiliza en todas las industrias, incluidas muchas empresas de la lista Fortune 500. **La propiedad clave de Apache Spark es la escalabilidad lineal.** Esto significa que, si duplica el número de servidores de un clúster, también duplicará aproximadamente su rendimiento.

Después de que Apache Spark comenzó a ganar cuota de mercado, 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 datos archivo por archivo. Apache Flink, por otro lado, es una imagen de procesamiento de secuencias, con su enfoque principal en el procesamiento de flujos de datos en tiempo real. Aunque el motor admite ambos paradigmas de procesamiento de datos, Apache Spark suele ser la elección en la mayoría de los casos de uso.

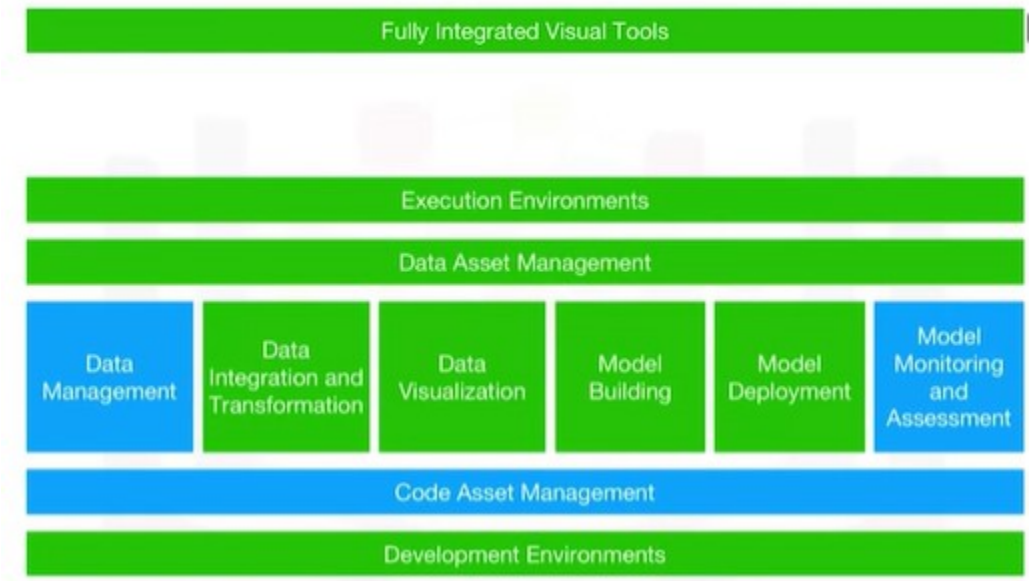
Uno de los últimos desarrollos en los entornos de ejecución de la ciencia de datos se llama «Ray», que tiene un claro enfoque en la formación de modelos de aprendizaje profundo a gran escala.

Herramientas que pueden utilizarse sin tener conocimientos de programación son:

- KNIME
- Orange

3. HERRAMIENTAS EN LA NUBE

Eche otro vistazo a la descripción general de las diferentes categorías de herramientas.



Dado que los productos en la nube son una especie más nueva, siguen la tendencia de tener múltiples tareas integradas en herramientas. Esto es especialmente cierto para las tareas marcadas en verde en el diagrama.

Comencemos con la categoría de **herramientas visuales totalmente integradas**. Dado que estas herramientas introducen un componente donde la ejecución a gran escala de flujos de trabajo de ciencia de datos ocurre en clústeres informáticos, hemos cambiado el título aquí y añadido la palabra «Plataforma». Estos clústeres se componen de múltiples máquinas servidor, de forma transparente para el usuario, en segundo plano.

Las herramientas son:

- Watson Studio
  - Que junto con Watson OpenScale cubren el ciclo de vida completo del desarrollo para todas las tareas de ciencia de datos, aprendizaje automático y IA.
- Microsoft Azure Machine Learning
  - Esta es también una oferta totalmente alojada en la nube que admite el ciclo de vida completo del desarrollo de todas las tareas de ciencia de datos, aprendizaje automático y IA.

- H2O Driverless AI

Dado que las operaciones y el mantenimiento no son realizadas por el proveedor de la nube, como es el caso de Watson Studio, Open Scale y Azure Machine Learning, este modelo de entrega no debe confundirse con Platform o Software as a Service (PaaS o SaaS).

En la administración de datos, con algunas excepciones, existen versiones SaaS de herramientas comerciales y de código abierto existentes.

Recuerde, **SaaS** significa «software como servicio», esto significa que el proveedor de la nube opera la herramienta para usted en la nube. Como ejemplo, el proveedor de la nube opera el producto realizando copias de seguridad de sus datos y configuración e instalando actualizaciones.

Como se mencionó, hay herramientas patentadas, que solo está disponible como un producto en la nube. A veces solo está disponible desde un único proveedor de nube. Un ejemplo de este servicio es Amazon Web Services DynamoDB, una base de datos NoSQL que permite el almacenamiento y la recuperación de datos en un formato clave o de almacén de documentos.

La estructura de datos del documento más prominente es JSON.

Cloudant es una oferta de base de datos como servicio. Pero, debajo del capó se basa en el código abierto Apache CouchDB. Tiene una ventaja: aunque las tareas operativas complejas como la actualización, copia de seguridad, restauración y escalado son realizadas por el proveedor de la nube, bajo el capó esta oferta es compatible con CouchDB. Por lo tanto, la aplicación se puede migrar a otro servidor CouchDB sin cambiar la aplicación.

IBM también ofrece Db2 como servicio. Este es un ejemplo de una base de datos comercial disponible como una oferta de software como servicio en la nube, quitando las tareas operativas al usuario.

Cuando se trata de herramientas de integración de datos comerciales, hablamos no solo de «extraer , transformar y cargar» o «ETL», sino también de herramientas de «extraer, cargar y transformar» o «ELT». Esto significa que los pasos de transformación no son realizados por un equipo de integración de datos, sino que se empujan hacia el dominio del científico de datos o el ingeniero de datos.

Dos herramientas de integración de datos comerciales ampliamente utilizadas son:

- Informatica Cloud Data Integration
- Data Refinery de IBM.
    - Permite la transformación de grandes cantidades de datos sin procesar en información consumible y de calidad en una interfaz de usuario similar a una hoja de cálculo.
    - Es parte de Watson Studio.

El mercado de las herramientas de visualización de datos en la nube es enorme, y cada proveedor importante de la nube tiene una. Un ejemplo de la herramienta de visualización de datos basada en la nube de una empresa más pequeña es DataMeer. IBM también ofrece su famosa suite de inteligencia empresarial de Cognos como solución en la nube. IBM Data Refinery también ofrece funcionalidad de exploración y visualización de datos en Watson Studio.

La construcción de modelos se puede hacer utilizando un servicio como Watson Machine Learning. Watson Machine Learning puede entrenar y construir modelos utilizando varias bibliotecas de código abierto. Google tiene un servicio similar en su nube llamado AI Platform Training. Casi todos los proveedores de nube tienen una solución para esta tarea.

La implementación de modelos en software comercial suele estar estrechamente integrada en el proceso de construcción de modelos. A continuación se muestra un ejemplo de los Servicios de colaboración e implementación de SPSS, que se pueden utilizar para implementar cualquier tipo de activo creado por la suite de herramientas de software SPSS. Lo mismo ocurre con otros proveedores. Además, el software comercial puede exportar modelos en un formato abierto. Como ejemplo, SPSS Modeler admite la exportación de modelos como Predictive Model Markup Language, o «PMML», que pueden ser leídos por muchos otros paquetes de software comerciales y abiertos. Watson Machine Learning también se puede utilizar para implementar un modelo y ponerlo a disposición de los consumidores mediante una interfaz REST. Amazon SageMaker Model Monitor es un ejemplo de una herramienta en la nube que supervisa continuamente los modelos de aprendizaje automático y aprendizaje profundo implementados. Una vez más, todos los principales proveedores de nube tienen herramientas similares. Este es también el caso de Watson OpenScale. OpenScale y Watson Studio ... unifican el paisaje.