Categorías de herramientas de ciencia de datos

Antes de que puedan ser útiles, los datos sin procesar deben pasar por varias categorías de tareas de ciencia de datos, como gestión de datos, integración y transformación de datos, visualización de datos, construcción de modelos, implementación de modelos y monitoreo y evaluación de modelos. La gestión de datos es el proceso de recopilar, conservar y recuperar datos de forma segura, eficiente, y rentable. La integración y transformación de datos es el proceso de extracción, transformación y carga de datos, y esto se llama ETL.

Los almacenes de datos se utilizan principalmente para recopilar y almacenar grandes cantidades de datos para su análisis.

La transformación de datos es el proceso de transformar los valores, la estructura, y el formato de los datos. Después de extraer los datos, el siguiente paso es transformar los datos. Los datos transformados se vuelven a cargar en el almacén de datos. La visualización de datos es la representación gráfica de datos e información. Puede utilizar la visualización para representar datos en forma de gráficos, diagramas, mapas, animaciones, etc.

Gráfico de barras que compara el tamaño de cada componente, un mapa de árbol que muestra datos jerárquicos, un gráfico de líneas que traza una serie de puntos de datos a lo largo del tiempo, y un gráfico de mapa que muestra datos por ubicación.

La construcción del modelo es el siguiente paso. Aquí es donde se entrenan los datos y se analizan patrones con algoritmos de aprendizaje automático. El sistema aprende a proporcionar predicciones o decisiones por sí mismo. Luego puedes usar este modelo para hacer predicciones sobre datos nuevos no vistos.

El siguiente paso es la implementación del modelo, el proceso de integrar un modelo desarrollado en un entorno de producción. En la implementación de modelos, un modelo de aprendizaje automático se pone a disposición de aplicaciones de terceros a través de API.

El monitoreo y la evaluación de modelos ejecutan controles de calidad continuos para garantizar la precisión, imparcialidad y solidez del modelo. La supervisión de modelos utiliza herramientas como **Fiddler** para rastrear el rendimiento de los modelos implementados en un entorno de producción.

La evaluación de modelos utiliza métricas de evaluación como la puntuación F1, la tasa de positivos verdaderos, o la suma del error al cuadrado para comprender el rendimiento de un modelo.

La gestión de activos de datos, también llamada gestión de activos digitales o DAM es la organización y gestión de datos importantes recopilados de diferentes fuentes. La gestión de activos digitales se realiza en una plataforma DAM que permite el control de versiones y la colaboración. Las plataformas DAM también admiten la replicación, la copia de seguridad y la gestión de derechos de acceso para los datos almacenados.

Los entornos de desarrollo, también llamados entornos de desarrollo integrados o IDE, proporcionan un espacio de trabajo y herramientas para desarrollar, implementar ejecutar, probar, e implementar código fuente.

Herramienta de código abierto para la ciencia de datos

Las herramientas de Administración de datos de código abierto más utilizadas son las Base de datos relacional como **MySQL** y **PostgreSQL**. También existen bases de datos NoSQL como **MongoDB**, **Apache CouchDB**, y **Apache Cassandra**. Además, hay herramientas basadas en archivos como **Hadoop** File System o UO en la nube como **Ceph**.

La tarea de integración y transformación de datos en el mundo del almacenamiento de datos clásico es extraer, transformar y cargar (ETL). Los científicos de datos a menudo proponen extraer, cargar, transformar, o ELT a medida que los datos se descargan en algún lugar y el ingeniero de datos o el científico de datos maneja la transformación de los datos; surgió otro término para este proceso: refinación y limpieza de datos. Las herramientas de integración y transformación de datos de código abierto más utilizadas son las siguientes **Apache Airflow, Kubeflow**; que permite la ejecución de pipelines de ciencia de datos sobre Kubernetes, **Apache Kafka**, **Apache NiFi**; que ofrece un editor visual muy agradable, **Apache Spark SQL** le permite utilizar **ANSI SQL** y escalar para computar clústeres de miles de nodos, y **Node-RED** también ofrece un editor visual; consume tan pocos recursos que incluso funciona en dispositivos pequeños.

Las herramientas de visualización de datos de código abierto más utilizadas. **Pixie Dust** es una biblioteca que tiene una interfaz de usuario que facilita la creación de gráficos en Python. Un enfoque similar utiliza **Hue**, que puede crear visualizaciones a partir de consultas SQL, mientras que **Kibana**, una aplicación web de exploración y visualización de datos, se limita a Elasticsearch, o al proveedor de datos. **Apache Superset** es una aplicación web de exploración y visualización de datos. La implementación del modelo es un paso crucial.

Una vez que hayas creado un modelo de aprendizaje automático capaz de predecir algunos aspectos críticos del futuro, debes hacerlo utilizable por otros desarrolladores y convertirlo en una API.

Selden es un producto interesante ya que admite casi todos los marcos, incluidos TensorFlow, Apache Spark ML R y Scikit-learn. Curiosamente, puede ejecutarse sobre Kubernetes y Red Hat OpenShift. Otra forma de desplegar modelos de Spark ML es **MLEAP**, y finalmente, **TensorFlow** puede servir a cualquier modelo de TensorFlow utilizando el servicio TensorFlow. Puede ser un dispositivo integrado como una Raspberry Pi o un teléfono inteligente que utilice TensorFlow Lite e implementado en un navegador web mediante TensorFlow.js.

El monitoreo del modelo es un paso importante también. Una vez que haya implementado un modelo de aprendizaje automático, querrá realizar un seguimiento de su rendimiento de predicción a medida que llegan nuevos datos para mantener los modelos obsoletos. Algunos ejemplos son los siguientes. **ModelDB** es una base de metadatos de modelos de máquina donde se almacena y consulta información sobre los modelos. Admite de forma nativa los pipelines de Apache Spark ML y Scikit-learn y CA. También se utiliza ampliamente una herramienta multipropósito genérica llamada **Prometheus**. Aunque no está diseñado específicamente para monitorear modelos de aprendizaje automático, se utiliza para este propósito.

La elección de herramientas de gestión de activos de código se ha vuelto bastante sencilla. **Git** es ahora el estándar de facto para la gestión de activos de código, también conocida como gestión de versiones o Control de versión. Alrededor de Git surgieron varios servicios. El más destacado es **GitHub**, pero el segundo lugar lo ocupa **GitLab**, con la ventaja de que la plataforma es completamente de código abierto y puede alojarse y administrarse por su cuenta. Otra opción es **Bitbucket**.

La gestión de activos de datos, también conocida como gobernanza de datos o linaje de datos, es una parte crucial de la ciencia de datos de nivel empresarial.

Los datos deben estar versionados y anotados con metadatos. **Apache Atlas** es una herramienta que respalda esta tarea. Otro proyecto interesante es ODPI Egeria, gestionado a través de la Fundación Linux. Es un ecosistema abierto que ofrece un conjunto de API abiertas, tipos y protocolos de intercambio que los repositorios de metadatos utilizan para compartir e intercambiar datos. Y, por último, **Kylo** es una plataforma de software de gestión de datos de código abierto con amplio soporte para tareas de gestión de activos de datos.

Las herramientas de gestión de datos son MySQL, PostgreSQL, MongoDB, Apache CouchDB, Apache Cassandra, Hadoop File System, Ceph y Elasticsearch.

Las herramientas de Integración de datos y transformación de datos son Apache Airflow, Kubeflow, Apache Kafka Apache NiFi, Apache Spark SQL, y Node-RED.

Las herramientas de visualización de datos son Pixie Dust, Hue, Kibana, y Apache Superset.

Las herramientas de implementación de modelos son Apache Prediction I.O., Selden, Kubernetes, Red Hat OpenShift, MLEAP, TensorFlow Service, TensorFlow Lite y TensorFlow.js.

Las herramientas de monitoreo de modelos son ModelDB, Prometheus, IBM AI Fairness 360, IBM Adversarial Robustness 360 Toolbox, y IBM AI Explainability 360.

Las herramientas de gestión de activos de código son Git, GitHub, GitLab, y Bitbucket. Y finalmente, las herramientas de gestión de activos de datos son Apache Atlas, ODPI Egeria, y Kylo.

Desarrollo de modelos

IBM ofrece varias herramientas y plataformas adaptadas para el desarrollo de modelos en diversos ámbitos. He aquí algunos ejemplos:

- IBM Watson Studio: Diseñado como un entorno integrado, Watson Studio simplifica el desarrollo, la formación y el despliegue de modelos. Ofrece compatibilidad con varios lenguajes y marcos de trabajo, como Python, R y TensorFlow, además de funciones de colaboración, herramientas de preparación de datos y opciones de despliegue versátiles.
- IBM AutoAI: IBM AutoAI, una función notable integrada en Watson Studio, agiliza el proceso de construcción de modelos de aprendizaje automático. Mediante la exploración dinámica de varios algoritmos e hiperparámetros, pretende identificar el modelo óptimo para un conjunto de datos determinado.
- IBM Watson OpenScale: Como plataforma para supervisar y gestionar modelos de IA en producción, Watson OpenScale desempeña un papel fundamental a la hora de garantizar la equidad, la explicabilidad y la mitigación de sesgos de los modelos. Proporciona información sobre el rendimiento del modelo y su evolución a lo largo del tiempo, lo que facilita la toma de decisiones informadas.
- IBM Watson Aprendizaje automático: Watson Aprendizaje automático, disponible como servicio en la plataforma IBM Cloud, permite a los usuarios escalar su formación y despliegue de modelos de aprendizaje automático. Da soporte sin problemas a marcos populares como TensorFlow, PyTorch y Scikit-learn, y ofrece API para una integración perfecta con otras aplicaciones.

El entorno de desarrollo más famoso que están utilizando los científicos de datos es Jupyter, que surgió como una herramienta para la programación interactiva en Python. Jupyter ahora admite más de 100 lenguajes de programación diferentes a través de kernels. Esto encapsula el entorno de ejecución para diferentes lenguajes de programación; una propiedad clave de los cuadernos Jupyter es unificar la documentación, el código, la salida del código, los comandos de shell, y las visualizaciones en un solo documento. JupyterLab es la siguiente versión de Notebook de Jupyter y, a largo plazo, reemplazará a Notebook de Jupyter.

La principal diferencia entre JupyterLab y Notebook de Jupyter es la capacidad de abrir diferentes tipos de archivos, incluyendo Notebook de Jupyter, datos, y terminales.

Apache Zeppelin se inspiró en Notebook de Jupyter y ofrece una experiencia similar, aunque un diferenciador clave es la capacidad de trazado integrada.

RStudio es uno de los entornos de desarrollo más antiguos para estadística y ciencia de datos. Ejecuta exclusivamente R y todas sus bibliotecas R asociadas. En el entorno R, es posible el desarrollo en Python. R está estrechamente integrado en la herramienta Jupyter y proporciona una experiencia de usuario óptima. RStudio unifica programación, ejecución, depuración, acceso remoto a datos, exploración de datos, y visualización en una sola herramienta.

Spyder intenta imitar el comportamiento de RStudio para llevar su funcionalidad al mundo Python.

A veces, los datos no caben en una sola computadora, almacenamiento o capacidad de memoria principal. Por lo tanto, existen entornos de ejecución en clúster. El ampliamente famoso Apache Spark está entre los proyectos de Apache más activos que se utilizan en todas las industrias. La propiedad clave de Apache Spark es la escalabilidad lineal, esto significa que, si duplicas el número de servidores en un clúster, aproximadamente duplicas su rendimiento.

Apache Flink se desarrolló después de que Apache Spark continuara ganando participación de mercado. 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 uno por uno o archivo por archivo, mientras que Apache Flink es una imagen de procesamiento de flujo cuyo foco principal es el procesamiento de flujos de datos en tiempo real. Y aunque ambos motores admiten ambos paradigmas de procesamiento de datos al mismo tiempo, Apache Spark es la opción para la mayoría de los casos de uso.

Después de Apache Spark y Apache Flink, Ray es uno de los últimos desarrollos en los entornos de ejecución de ciencia de datos y tiene un enfoque claro en el entrenamiento de modelos de aprendizaje profundo a gran escala.

Herramientas comerciales para la ciencia de datos

En la gestión de datos, la mayoría de los datos relevantes de la empresa se almacenan en una base de datos Oracle, en Microsoft SQL Server, o en IBM DB2. Aunque las bases de datos de código abierto están cobrando importancia, estos tres productos de gestión de datos se consideran estándares de la industria y seguirán existiendo por un tiempo. Además, no se trata sólo de funcionalidad. Dado que los datos son el corazón de cada organización la disponibilidad de soporte comercial juega un papel importante. Los soportes comerciales son proporcionados directamente por proveedores de software, socios influyentes y redes de soporte.

Las herramientas de integración de datos comerciales que incluyen herramientas de extracción, transformación y carga (ETL). Según el Cuadrante Mágico de Gartner, Informática Power Center e IBM InfoSphere Data Stage son los líderes. A continuación, están los productos de SAP, Oracle, SAS Talend, y Microsoft. Estas herramientas apoyan el diseño y la implementación de canales de procesamiento de datos ETL a través de una interfaz gráfica. Traen consigo conectores para la mayoría de los sistemas de información comerciales y de código abierto. *Watson Studio Desktop incluye un componente llamado Data Refinery, que permite la definición y ejecución de procesos de integración de datos en estilo hoja de cálculo*

En el entorno comercial, las visualizaciones de datos utilizan herramientas de inteligencia empresarial o BI El enfoque de estas herramientas es crear informes visuales y paneles en vivo. Los representantes comerciales más destacados son Tableau, Microsoft Power BI e IBM Cognos Analytics.

Si quieres construir un modelo de aprendizaje automático con una herramienta comercial, deberías usar un producto de minería de datos. Los productos más destacados en ese ámbito son SPSS Modeler y SAS Enterprise Miner. Además, SPSS Modeler también está disponible en Watson Studio Desktop, basado en la versión en la nube de la herramienta. Ahora, la implementación de modelos en software comercial está estrechamente integrada en el proceso de construcción de modelos. A continuación, se muestra un ejemplo de los servicios de colaboración e implementación de SPSS, utilizados para implementar activos creados por SPSS. Y lo mismo ocurre con otros proveedores. Además, el software comercial puede exportar modelos en un formato abierto. Por ejemplo, SPSS Modeler admite la exportación de modelos como lenguaje de marcado de modelos

predictivos, o PMML, que pueden leer una gran cantidad de otros paquetes de software comerciales y abiertos.

El monitoreo de modelos es una disciplina muy nueva, y actualmente, no existen herramientas comerciales relevantes disponibles. Por lo tanto, el código abierto es la primera opción. Lo mismo ocurre con la gestión de activos de código. El código abierto con Git y GitHub es el estándar de facto. La gestión de activos de datos, a menudo denominada gobernanza de datos o linaje de datos es una parte crucial de la ciencia de datos de nivel empresarial. Datos versionados y anotados con metadatos. Proveedores, como Informática Enterprise Data Governance e IBM, ofrecen Gobierno de datos

El catálogo de gobernanza de la información cubre funciones como un diccionario de datos, que facilita el descubrimiento de activos de datos. Cada activo de datos se asigna a un administrador de datos o al propietario de los datos. El propietario de los datos es responsable de ese activo de datos y se puede contactar con él. Luego, se cubre el linaje de datos, lo que permite rastrear los pasos de transformación en la creación de los activos de datos. El linaje de datos también incluye una referencia a los datos fuente reales. Además, se pueden agregar reglas y políticas para reflejar requisitos comerciales y regulatorios complejos para la privacidad y retención de datos.

Watson Studio es un entorno de desarrollo totalmente integrado para científicos de datos. La mayoría de la gente lo consume a través de la nube. Y 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, junto con Watson OpenScale, es una herramienta totalmente integrada que cubre el ciclo de vida de la ciencia de datos que involucra todas las tareas discutidas anteriormente. Se pueden implementar en un centro de datos local sobre Kubernetes o Red Hat OpenShift. Otro ejemplo de una herramienta comercial totalmente integrada es H2O Driverless AI, que cubre el ciclo de vida completo de la ciencia de datos.

Herramientas basadas en la nube para la ciencia de datos

Herramientas visuales. Dado que estas herramientas introducen un componente donde la ejecución a gran escala de flujos de trabajo de ciencia de datos ocurre en clústeres de cómputo, hemos cambiado el título y agregado la palabra plataforma. Estos clústeres están compuestos por múltiples máquinas servidor de forma transparente para el usuario en segundo plano. Watson Studio y Watson OpenScale cubren el ciclo de vida de desarrollo completo para todas las tareas de ciencia de datos, aprendizaje automático, e inteligencia artificial o IA.

Otro ejemplo es Microsoft Azure Aprendizaje automático. También es una oferta completa alojada en la nube, que respalda el ciclo de vida de desarrollo completo de todas las tareas de ciencia de datos, aprendizaje automático e inteligencia artificial. Y finalmente, otro ejemplo es la inteligencia artificial sin conductor H2O, si bien es un producto que se descarga e instala, existe una implementación con un solo clic para los proveedores de servicios en la nube estándar. Dado que el proveedor de la nube no realiza operaciones ni mantenimiento, como ocurre con Watson Studio, OpenScale, y Azure Machine Learning, este modelo de entrega debe ser distinto del de plataforma o software como servicio, o PaaS, o SaaS.

En la gestión de datos, con algunas excepciones, existen versiones de software como servicio, o SaaS, de herramientas comerciales y de código abierto existentes. El proveedor de la nube opera la herramienta para usted en la nube. Algunas herramientas propietarias solo están disponibles a través de un único proveedor de nube. Un ejemplo de este tipo de servicio es Amazon Web Services DynamoDB, que es una base de datos NoSQL y Servicios web, permite almacenar y recuperar datos en un valor clave o en un formato de almacén de documentos. La estructura de datos de documentos más destacada es JSON.

Ahora bien, otra variante de dicho servicio es Cloudant que es una base de datos como oferta de servicio, pero en el fondo, se basa en el Apache CouchDB de código abierto. La ventaja es que las tareas operativas complejas como actualización, copia de seguridad, restauración, y escalamiento las realiza el proveedor de la nube, sin embargo, la oferta de nube y servicios es compatible con CouchDB. Por lo tanto, la aplicación migra a otro servidor CouchDB sin realizar ningún cambio en la aplicación.

IBM también ofrece DB2 como servicio. Es un ejemplo de una base de datos comercial disponible como oferta SaaS en la nube, quitándole al usuario las tareas operativas.

Las herramientas comerciales de integración de datos que incluyen herramientas de Extraer, Transformar, y Cargar, o ETL. Los pasos de transformación no los realiza un equipo de integración de datos, sino que se trasladan al dominio del científico o ingeniero de datos. Dos herramientas de integración de datos comerciales ampliamente utilizadas son Informática Cloud Data Integration y Data Refinery de IBM. Data Refinery es parte de IBM Watson Studio, permite transformar grandes cantidades de datos sin procesar en información de calidad consumible en una interfaz de usuario similar a una hoja de cálculo.

El mercado de herramientas de visualización de datos en la nube es enorme, y cada proveedor importante de la nube tiene una. Un ejemplo de una empresa pequeña que ofrece una herramienta de visualización de datos basada en la nube

es DataMirror. IBM Data Refinery ofrece funcionalidad de exploración y visualización de datos en Watson Studio.

Una nube de palabras resalta términos significativos en un corpus de documentos. Ahora, la construcción de modelos se puede hacer utilizando un servicio IU. Un ejemplo de un servicio es Watson Aprendizaje automático. Watson Aprendizaje automático puede entrenar y construir modelos utilizando varias bibliotecas de código abierto. Google tiene un servicio similar en su nube llamado IA Platform Training, y cada proveedor de nube tiene una solución para esta tarea. La implementación de modelos en software comercial generalmente está estrechamente integrada en el proceso de construcción del modelo.

Resumen:

- **Herramientas de Gestión de Datos** - Facilitan el almacenamiento, organización y recuperación de datos. Incluyen Bases de Datos Relacionales, Bases de Datos NoSQL y plataformas de Big Data.
- **Herramientas de Integración y Transformación de Datos** - Optimiza los flujos de trabajo de procesamiento de datos y automatiza los pipelines de datos. La tarea de integración y transformación de datos en el mundo clásico de los data warehouses es Extraer, Transformar y Cargar (ETL).
- **Herramientas de Visualización de Datos** - Proporciona representación gráfica de los datos y ayuda a comunicar hallazgos.
- **Herramientas de Despliegue, Monitoreo y Evaluación de Modelos** - Soporta la construcción, despliegue, monitoreo y evaluación de modelos de datos y aprendizaje automático.
- **Herramientas de Gestión de Activos de Datos** - Organiza y gestiona datos, aplica controles de acceso y asegura copias de seguridad de los activos.
- **Herramientas de Desarrollo y Ejecución de Código** - Proporcionan entornos para desarrollar, probar y desplegar código, ofreciendo recursos computacionales para ejecutarlo.
- **Herramientas de Gestión de Activos de Código** - Permiten el almacenamiento y gestión de código, rastrean cambios y apoyan el desarrollo colaborativo.