Capítulo 2
El Big Data y la Ciencia de Datos

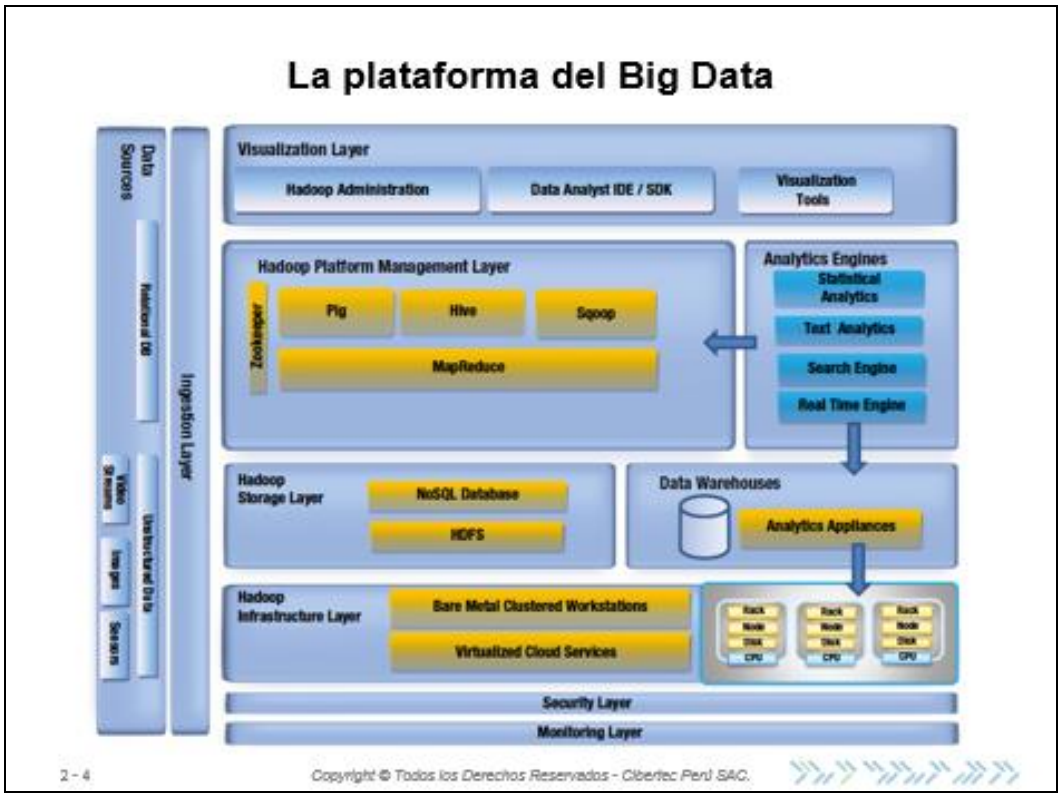
Al finalizar el capítulo, el alumno podrá:

- Comprender el proceso del Big Data y el rol del científico de datos.

Temas

1. La plataforma del Big Data
2. El científico de Datos
3. El proceso de la Ciencia de Datos

1. La Plataforma de Big Data

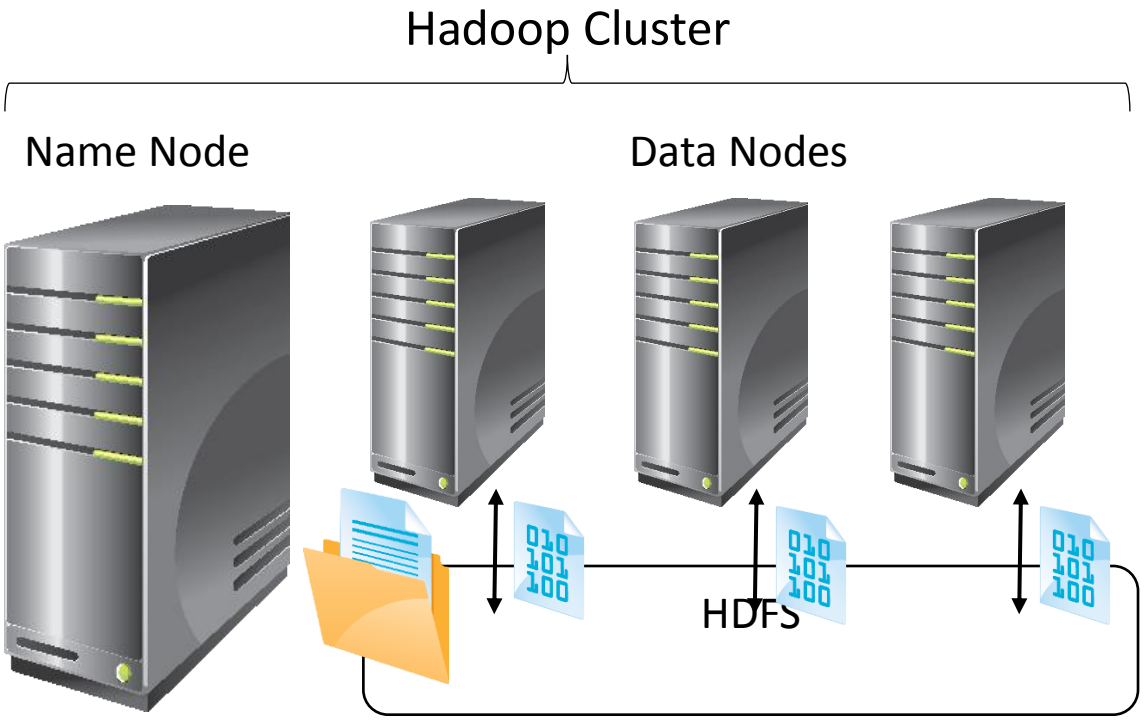


¿Qué es Hadoop?

Apache Hadoop, es un proyecto open source que tiene ya casi diez años de existencia que se caracteriza por dos temas principales, el primero es qué maneja un sistema de archivos distribuidos que por sus siglas lo vamos a conocer como HDFS o Hadoop Distributed File System y la segunda característica, es el procesamiento distribuido.

Las tres principales distribuciones que existen actualmente en el mercado:

- Cloudera
- Hortonworks
- MAPR



El sistema de archivos distribuido Hadoop (HDFS) permite a las aplicaciones ejecutarse en varios servidores. HDFS tiene una alta tolerancia a fallos, se ejecuta en hardware económico y proporciona acceso a datos con un gran rendimiento.

Los datos en un clúster Hadoop se dividen en partes más pequeñas llamadas *bloques* y, a continuación, se distribuyen en todo el clúster. Los bloques y copias de bloques, se almacenan en otros servidores en el clúster hadoop. Es decir, un archivo individual se almacena como bloques más pequeños que se replican entre varios servidores en el clúster.

Cada HDFS clúster tiene un número de *DataNodes* teniendo un DataNode para cada nodo en el clúster. Los DataNodes gestionan el almacenamiento que se adjunta a los nodos en los que se ejecutan. Cuando se divide un archivo en bloques, éstos se almacenan en un conjunto de DataNodes que se distribuyen en todo el clúster. Los DataNodes son responsables de servir las solicitudes de lectura y escritura de los clientes en el sistema de archivos, y también gestionan la creación, supresión y replicación de bloques.

Un clúster HDFS da soporte a *NameNodes*, un NameNode activo y un NameNode en espera, que es la configuración común para la alta disponibilidad. El NameNode regula el acceso a los archivos por parte de los clientes, y rastrea todos los archivos de datos en HDFS. El NameNode determina la correlación de bloques con DataNodes y gestiona operaciones como abrir, cerrar y renombrar archivos y directorios. Toda la información del NameNode se almacena en memoria, lo que permite tiempos de respuesta rápidos al añadir almacenamiento o leer solicitudes. El NameNode es el repositorio para todos los metadatos de HDFS.

Un despliegue típico de HDFS tiene un sistema dedicado que ejecuta solo el NameNode, porque el debido a almacena los metadatos en memoria. Si el sistema que ejecuta el NameNode falla, se pierden todos los metadatos del clúster, por lo que dicho equipo suele ser más robusto que el resto del clúster.
El sistema de procesamiento de datos de Hadoop es un sistema que permite el procesamiento distribuido de big data a través de clústeres de servidores usando modelos de programación sencillos. Está diseñado para escalar de servidores individuales a miles de máquinas, cada una ofreciendo computación y almacenamiento local.

En lugar de confiar en el hardware para ofrecer alta disponibilidad, el propio sistema de procesamiento de datos de Hadoop está diseñado para detectar y manejar fallos en la capa de aplicación, por lo que ofrece un servicio altamente disponible sobre un clúster de servidores, cada uno de los cuales puede ser propenso a fallos.
Por lo tanto, se trata de un sistema de procesamiento de datos diseñado para ser robusto, en el cual el procesamiento de Big Data seguirá funcionando incluso cuando los servidores individuales o clústeres fallen. Y también está diseñado para ser eficiente ya que no requiere que sus aplicaciones transporten enormes volúmenes de datos a través de su red.

Cuando los bits de información van entrando en el sistema de procesamiento de datos distribuido de Hadoop el camino que recorren es el siguiente:

- División de los datos entrantes en segmentos.
- **Distribución de los segmentos en nodos diferentes** que soportan el procesamiento paralelo.
- **Replicación de cada segmento en múltiples nodos de datos** de forma que dos copias queden alojadas en nodos del mismo grupo y una adicional se envíe a un nodo situado en un rack distinto.
- Agrupación de los nodos en clústeres HDFS.

Además, **una de las ventajas del sistema de procesamiento de datos distribuido de Hadoop es que, gracias a esa replicación que tiene lugar, la información queda protegida frente a bastantes tipos de fallos.** Así, si un nodo tuviera problemas y no permitiese el acceso a los datos contenidos en sus segmentos, el procesamiento no se detendría, ya que podría continuar recurriéndose a los nodos de cualquier otro grupo.

Sin embargo, no hay que olvidar que todavía podrían plantearse algunos inconvenientes. **La tolerancia a fallos del sistema de procesamiento de datos tiene una excepción**, y es que se **requiere un único NameNode y éste se ubica en un único servidor**, por lo que, caso de producirse un fallo que le afectase, todo el sistema de archivos quedaría inaccesible.

Podría decirse que el sistema estaría cerrado, al menos, hasta que pudiera reiniciarse el servidor gracias a los datos que un NameNode secundario va guardando cada vez que lleva a cabo una copia de seguridad periódica del principal. Algo que, en ningún caso serviría para mantener las operaciones en ejecución.
Por último, queda citar **un componente importante en el sistema de procesamiento de datos distribuido de Hadoop: MapReduce**. Precisamente MapR Technologies anunció recientemente un sistema de archivos compatible con Hadoop y que, entre sus principales características, **cuenta con un NameNode distribuido que elimina el único punto de fallo presente en HDFS.**

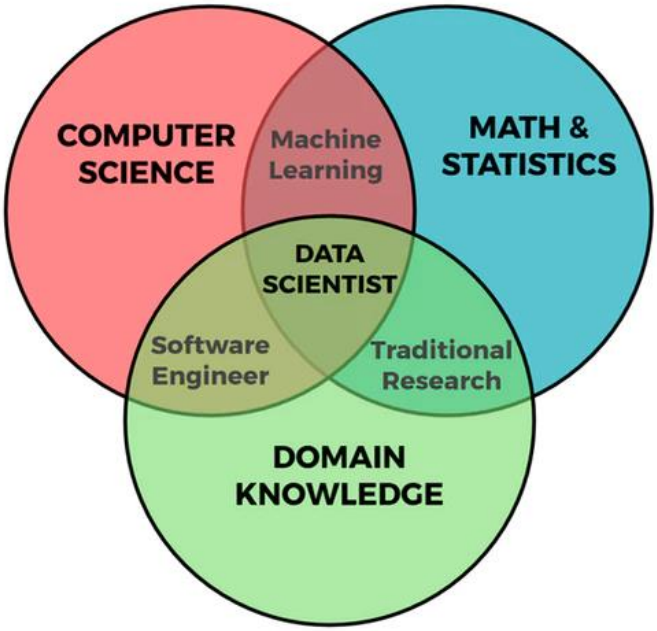
Si bien, la función por la que se conoce a MapReduce es por su papel en la gestión del procesamiento de datos distribuido. Su funcionamiento permite que los trabajos se envíen a un JobTracker capaz de asignar una tarea a un nodo TaskTracker gracias a su conocimiento de la ubicación de cada segmento de datos.

De esta forma, **el sistema de procesamiento de datos de** Hadoop **permite multiplicar la eficiencia de los procesos** gracias a su carácter distribuido que hace posible un trabajo más ágil, en menor tiempo, con un riesgo mínimo y que puede fácilmente escalarse.

2. El científico de datos

El científico de datos es una nueva profesión que hoy es considerada clave en el mundo de las tecnologías y es una de las mejor pagadas. Se trata de una persona formada en las ciencias matemáticas y las estadísticas que domina la programación y sus diferentes lenguajes, ciencias de la computación y analítica.

El profesional de la ciencia de datos también debe tener la capacidad y los conocimientos necesarios para comunicar sus hallazgos a medida que los tiene, no sólo al área de tecnología sino además al sector de los negocios. Debe dominar la tecnología y las bases de datos para modificar y mejorar la orientación de los negocios de la empresa para la que trabaja.



¿Por qué una carrera en ciencia de datos?

- Data Science es un campo en el que puede generar enormes impactos aprovechando algunas ideas fundamentales.
- ¡Poder extraer ideas clave de los datos para ayudar a construir un futuro mejor es una de las carreras más gratificantes!
- A menudo, las personas solo asocian la ciencia de datos con las grandes compañías de tecnología, ¡pero casi todas las industrias pueden obtener una ventaja sustancial con la ciencia de datos!
- ¡También hay muchas posiciones que se relacionan con las ideas en este campo!
- Según Forbes, los trabajos que requieren habilidades de aprendizaje automático están pagando un promedio de $ 114,000.
- Los trabajos de científicos de datos anunciados pagan un promedio de $ 105,000 y los trabajos de ingeniería de datos anunciados pagan un promedio de $ 117,000.
- La demanda anual de los nuevos roles de rápido crecimiento de científicos de datos, desarrolladores de datos e ingenieros de datos alcanzará cerca de 700,000 vacantes para 2020.
- Para el año 2020, el número de empleos para todos los profesionales de datos de EE. UU. Aumentará en 364,000 aperturas a 2, 720,000 según IBM.

- Razones principales para la carrera de ciencias de datos
  - Cumpliendo con el trabajo
  - Temas técnicos muy interesantes
  - Alta demanda
  - Gran paga y beneficios

Títulos del científico de datos

- El término "Científico de datos" no es el único título de trabajo que utilizan las empresas.
- Muchas compañías usan diferentes títulos para diferentes enfoques para roles en el campo de la ciencia de datos.
- Las empresas también usan el mismo término de manera diferente, un "Científico de datos" en una empresa pequeña puede llamarse "Analista de datos" en otra compañía.
- Repasemos algunos títulos y sus descripciones generales
    - Product Analyst / Product Data Scientist
        - Analiza datos de usuario para crear informes para gerentes de producto
        - Por lo general, utiliza la codificación como herramienta principal, pero las empresas pueden preferir una herramienta de software predefinida
    - Business Analyst / Business Intelligence
        - Crear ideas y análisis a partir de datos comerciales
        - A menudo se enfoca en una herramienta específica como Tableau o Excel
    - Machine Learning Engineer
        - Papel muy técnico, se requieren conocimientos teóricos y buenas habilidades de codificación
        - Crea modelos de algoritmos de aprendizaje de máquina personalizados para el equipo
    - Data Engineer
        - Rol técnico centrado en codificación y herramientas, no tanto teoría.
        - Crea conductos que conectan el almacén de datos con el análisis de datos o sistemas de aprendizaje automático.
    - Data Scientist
        - Requiere una combinación de habilidades de codificación y teoría.
        - A menudo, la empresa publicará un caso de uso más detallado en la publicación de trabajo real.

Educación del científico de datos

- Una pregunta común es: ¿necesito una maestría o un doctorado para ser un científico de datos?
- Muchos científicos de datos, tienen estudios de posgrado, y otros científicos de datos de grandes empresas privadas no tienen títulos de posgrado.

Herramientas del científico de datos

- Analicemos algunas herramientas técnicas que las personas usan en la industria de las ciencias de datos
    - Lenguajes de programación
        - Java
        - SQL
        - Python
        - R
    - Frameworks

        - SciKit-Learn (Python)
        - Caret (R)
        - MLlib (Spark)
        - Lots of other libraries (e.g. pandas, numpy, various R libraries, etc…)
    - Herramientas de software de nivel superior
        - Weka
        - H2O
        - AWS ML Services
        - Turi
        - Tableau

Conocimientos de la teoría

- El conocimiento general de los conceptos básicos sobre ciencia de datos y matemática es aún más importante que las herramientas de software utilizadas.
- Vamos a cubrir algunos de los conceptos clave que debe saber al aplicar
- Comencemos discutiendo algunos de los temas matemáticos más básicos que debe comprender.
- ¡Tenga en cuenta que pocas personas son expertas en todo esto y que constantemente aprenderá más en el trabajo!
    - Cálculo
    - Álgebra lineal
    - Teoría de probabilidad
    - Estadística
        - Inferencia estadística
        - Distribuciones y estadísticas descriptivas
    - Temas matemáticos específicos
        - Gráfico y análisis de red
        - Estadísticas bayesianas
        - Prueba A | B

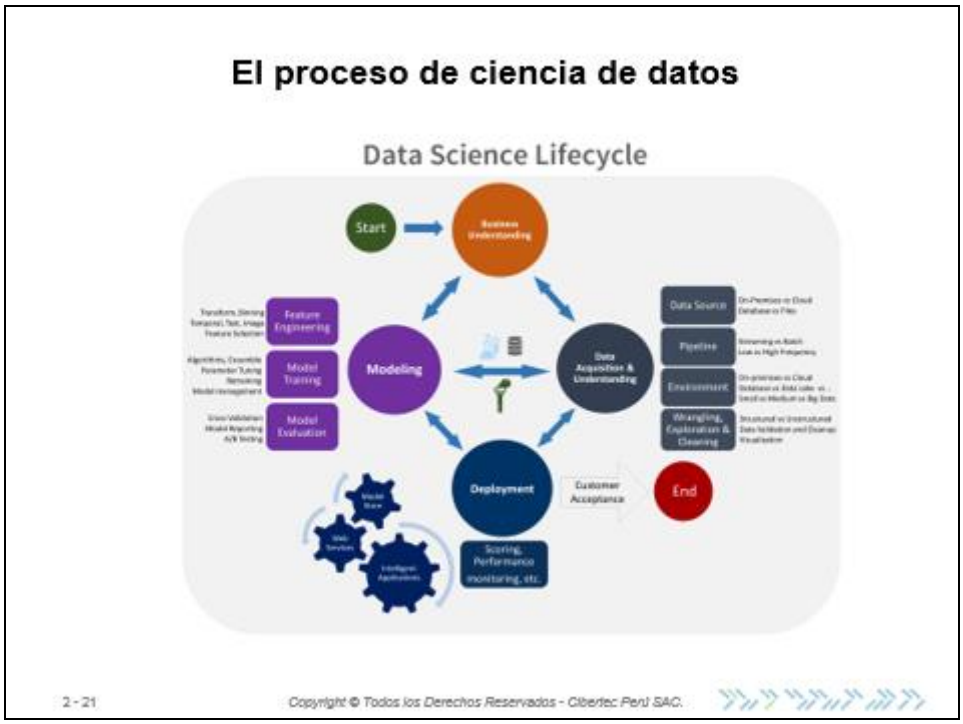
Conocimientos de Machine Learning

- La mayoría de las funciones de Data Science requerirán que entiendas varios aspectos del Machine Learning, tales como:
    - Supervised Learning
    - Linear Regression, SVM, Random Forest, Logistic Regression, KNN, etc...
    - Unsupervised Learning
    - K-Means Clustering, PCA, etc…
    - NLP, Model Validation, K-Folds, etc...
    - Bias-Variance Trade-Off
    - Gradient Descent
    - L1 / L2 Regularization
    - Bagging / Boosting

Conocimientos de Ingeniería de Software

- Entre las principales habilidades de programación de un científico de datos, tenemos las siguientes:
    - Algorithms and Data Structures
    - Databases (SQL)
    - Distributed Computed (Spark)
    - Data Visualization Products or Services

3. El proceso de la Ciencia de Datos



El proceso de ciencia de datos en equipo (TDSP - The Microsoft Team Data Science Process) es una metodología de ciencia de datos ágil e iterativa para proporcionar soluciones de análisis predictivo y aplicaciones inteligentes de manera eficiente. TDSP ayuda a mejorar la colaboración en equipo y el aprendizaje. Contiene una extracción de los mejores procedimientos y estructuras de Microsoft y otros fabricantes del sector que facilitan la correcta implementación de iniciativas de ciencia de datos. El objetivo es ayudar a las empresas a que se den cuenta de las ventajas de su programa de análisis.

En este capítulo se proporciona una introducción a TDSP y sus componentes principales. Aquí se ofrece una descripción genérica del proceso que se puede implementar con diversas herramientas.

Principales componentes del TDSP

El TDSP consta de los siguientes componentes clave:

- Una definición de **ciclo de vida de ciencia de datos**
- Una **estructura de proyecto estandarizada**
- **Infraestructura y recursos** para proyectos de ciencia de datos
- **Herramientas y utilidades** para la ejecución de proyectos

Ciclo de vida de ciencia de datos

El proceso de ciencia de datos en equipo (TDSP) proporciona un ciclo de vida para estructurar el desarrollo de los proyectos de ciencia de datos. El ciclo de vida describe el proceso, de principio a fin, que suelen seguir los proyectos al ejecutarlos.

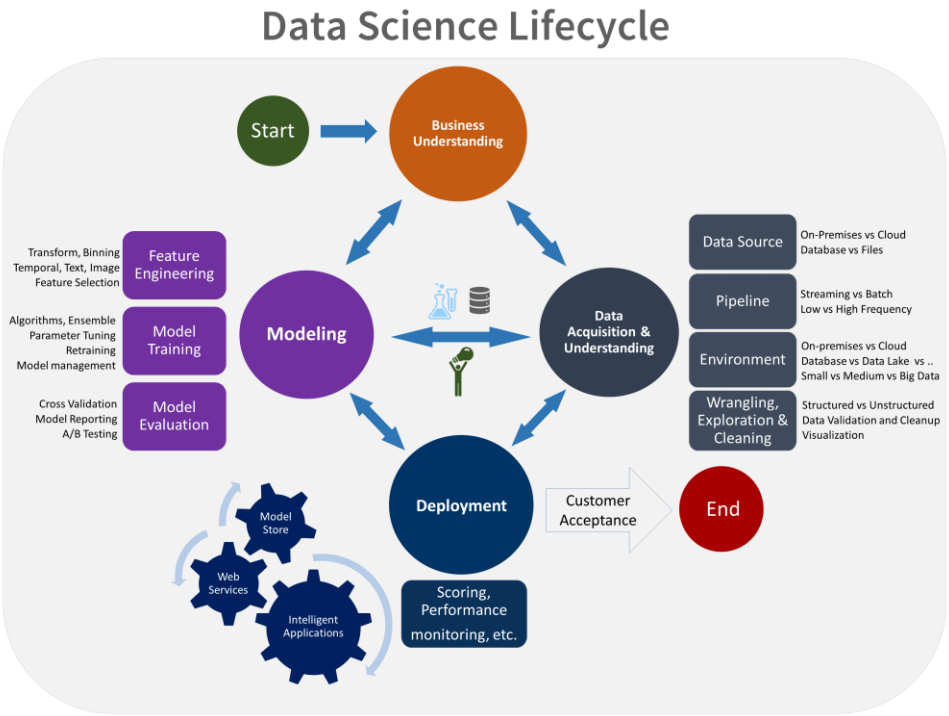
Aunque esté usando otro ciclo de vida de ciencia de datos, como CRISP-DM, KDD o el proceso personalizado de su organización, puede usar también el TDSP basado en tareas en el contexto de esos ciclos de vida de desarrollo. En un nivel alto, estas distintas metodologías tienen mucho en común.

Este ciclo de vida se ha diseñado para proyectos de ciencia de datos que se enviarán como parte de aplicaciones inteligentes. Estas aplicaciones implementan modelos de aprendizaje o inteligencia artificial de máquina para realizar un análisis predictivo. Los proyectos de ciencia de datos exploratorios o proyectos de análisis ad hoc también se pueden beneficiar del uso de este proceso. Pero, en estos casos, puede que algunos de los pasos descritos no sean necesarios.

El ciclo de vida describe las fases principales por las que pasan normalmente los proyectos, a menudo de forma iterativa:

- Conocimiento del negocio
- Adquisición y comprensión de los datos
- Modelado
- Implementación
- Aceptación del cliente

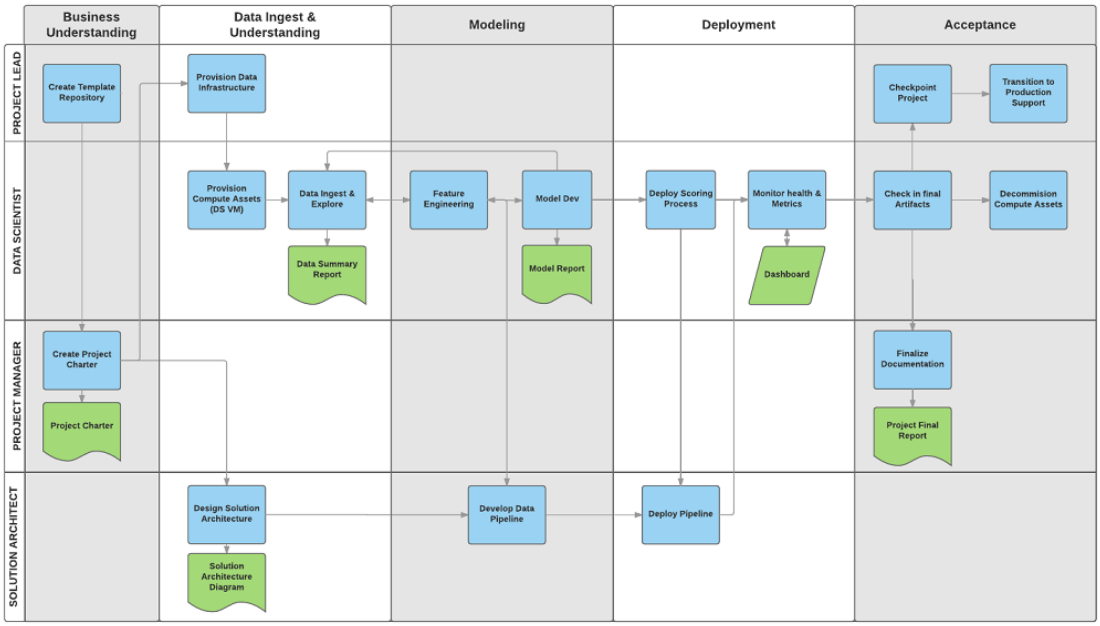
Esta es una representación visual del **ciclo de vida del proceso de ciencia de datos en equipo**.



En el tema [Team Data Science Process lifecycle] (Ciclo de vida del proceso de ciencia de datos en equipo) se describen los objetivos, las tareas y los artefactos de documentación de cada fase del ciclo de vida de TDSP. Estas tareas y artefactos están asociados con roles de proyecto:

- Arquitecto de soluciones
- Jefe de proyecto
- Científico de datos
- Responsable de proyecto

En el siguiente diagrama se proporciona una vista de cuadrícula de las tareas (en azul) y los artefactos (en verde) asociados con cada fase del ciclo de vida (eje horizontal) de estos roles (eje vertical).

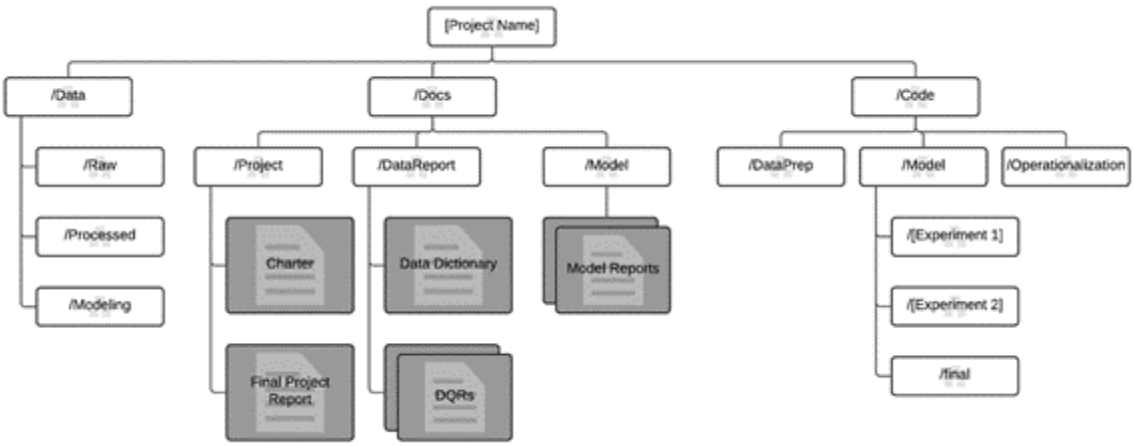


Estructura de proyecto estandarizada

Cuando todos los proyectos comparten una estructura de directorio y usan plantillas para los documentos de proyecto, resulta fácil para los miembros del equipo encontrar información sobre sus proyectos. Todo el código y los documentos se almacenan en un sistema de control de versiones (VCS), como Git, TFS o Subversion para permitir la colaboración en equipo. El seguimiento de las tareas y las características en un sistema de seguimiento de proyectos ágil, como Jira, Rally o Visual Studio Team Services permite seguir más de cerca el código para conocer sus características individuales. Este seguimiento también permite a los equipos obtener mejores estimaciones de los costos. TDSP recomienda crear un repositorio independiente para cada proyecto en el VCS de cara al control de versiones, la seguridad de la información y la colaboración. La estructura estandarizada para todos los proyectos ayuda a crear conocimiento institucional en toda la organización.

Se proporcionan plantillas para la estructura de carpetas y los documentos necesarios en ubicaciones estándar. Esta estructura de carpetas organiza los archivos que contienen código para la exploración de datos y la extracción de características, y los que registran las iteraciones de los modelos. Estas plantillas permiten a los miembros del equipo comprender el trabajo que otros realizan, y agregar nuevos miembros a los equipos de forma fácil. Las plantillas de documento se pueden ver y actualizar fácilmente en formato de marcado. Use plantillas para proporcionar listas de comprobación con preguntas clave en cada proyecto y de esta forma garantizar que el problema esté bien definido y que los resultados entregados satisfagan la calidad

esperada.



La estructura de directorio se puede clonar desde Github.

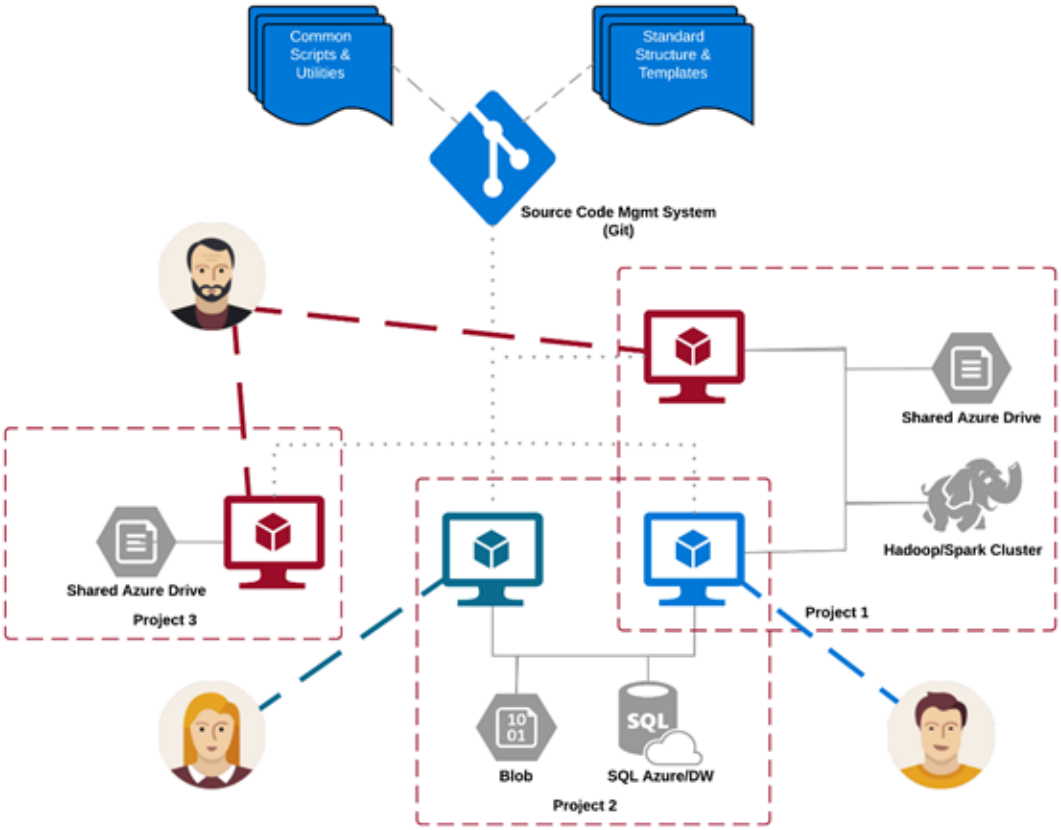
Infraestructura y recursos para los proyectos de ciencia de datos

TDSP proporciona recomendaciones para administrar análisis compartido e infraestructura de almacenamiento, por ejemplo:

- sistemas de archivos en la nube para almacenar conjuntos de datos
- bases de datos
- clústeres de macrodatos (Hadoop o Spark)
- servicios de aprendizaje automático.

La infraestructura de análisis y almacenamiento puede estar en la nube o en el entorno local. Aquí es donde se almacenan los conjuntos de datos sin procesar y procesados. Esta infraestructura permite un análisis reproducible. También evita la duplicación, lo que puede llevar a incoherencias y costos de infraestructura innecesarios. Se proporcionan herramientas para aprovisionar los recursos compartidos, realizar un seguimiento de ellos y permitir que cada miembro del equipo se conecte a dichos recursos de forma segura. También es una buena práctica pedir a los miembros del proyecto que creen un entorno de proceso coherente. Luego, diferentes miembros del equipo pueden replicar y validar los experimentos.

Este es un ejemplo de un equipo que trabaja en varios proyectos y que comparte diversos componentes de la infraestructura de análisis.



Herramientas y utilidades para la ejecución de proyectos

En la mayoría de las organizaciones la introducción de procesos presenta ciertos desafíos. Las herramientas proporcionadas para implementar el proceso y el ciclo de vida de ciencia de datos ayudan a reducir las barreras a su adopción y la normalizan. TDSP proporciona un conjunto inicial de herramientas y scripts para impulsar la adopción de TDSP dentro de un equipo. También ayuda a automatizar algunas de las tareas comunes del ciclo de vida de ciencia de datos, como la exploración de datos y el modelado de línea de base. Existe una estructura bien definida que se proporciona a los individuos para que contribuyan con herramientas y utilidades compartidas al repositorio de código compartido de su equipo. Estos recursos se pueden aprovechar luego en otros proyectos dentro del equipo o en la organización. TDSP también tiene previsto habilitar las contribuciones de herramientas y utilidades a toda la comunidad. Las utilidades de TDSP se pueden clonar desde Github.