

1. Para que las fuentes de datos se carguen en el almacén de datos, los datos deben entenderse bien, estructurado y normalizado con las definiciones de tipo de datos adecuadas. Aunque este tipo de centralización permite la seguridad, la copia de seguridad y la comutación por error de datos muy críticos, también significa que los datos normalmente deben pasar por un preprocesamiento significativo y puntos de control antes de que puedan ingresar a este tipo de entorno controlado, que no se presta a la exploración de datos ni a la iteración analítica.
2. Como resultado de este nivel de control en el EDW, pueden surgir sistemas locales adicionales en forma de almacenes departamentales y mercados de datos locales que los usuarios comerciales crean para adaptarse a sus necesidades de análisis flexible. Es posible que estos mercados de datos locales no tengan las mismas restricciones de seguridad y estructura que el EDW principal y permitan a los usuarios realizar algún nivel de análisis más profundo. Sin embargo, estos sistemas únicos residen de forma aislada, a menudo no están sincronizados o integrados con otros almacenes de datos y es posible que no se realice una copia de seguridad.
3. Una vez en el almacén de datos, las aplicaciones adicionales de la empresa leen los datos para BI y propósitos de informes. Estos son procesos operativos de alta prioridad que obtienen fuentes de datos críticos de los almacenes de datos y los repositorios.
4. Al final de este flujo de trabajo, los analistas obtienen datos para sus análisis posteriores. Debido a que los usuarios generalmente no pueden ejecutar análisis personalizados o intensivos en bases de datos de producción, los analistas crean extractos de datos de EDW para analizar datos fuera de línea en R u otras herramientas analíticas locales. Muchas veces, estas herramientas se limitan a análisis en memoria en equipos de escritorio que analizan muestras de datos, en lugar de a toda la población de un conjunto de datos. Debido a que estos análisis se basan en extractos de datos, residen en una ubicación separada, y los resultados del análisis, y cualquier conocimiento sobre la calidad de los datos o anomalías, rara vez se retroalimentan en el repositorio de datos principal.

Debido a que las nuevas fuentes de datos se acumulan lentamente en el EDW debido al riguroso proceso de validación y estructuración de datos, los datos tardan en trasladarse al EDW y el esquema de datos cambia lentamente.

Los almacenes de datos departamentales pueden haber sido diseñados originalmente para un propósito específico y un conjunto de necesidades comerciales, pero con el tiempo evolucionaron para albergar más y más datos, algunos de los cuales pueden ser forzados a los esquemas existentes para permitir BI y la creación de cubos OLAP para análisis e informes. Aunque el EDWach logra el objetivo de informar y, a veces, la creación de paneles de control, los EDW generalmente limitan la capacidad de los analistas para iterar sobre los datos en un entorno de no producción separado donde pueden realizar análisis en profundidad o realizar análisis sobre datos no estructurados.

Las arquitecturas de datos típicas que se acaban de describir están diseñadas para almacenar y procesar datos de misión crítica, admitir aplicaciones empresariales y habilitar actividades de informes corporativos. Aunque los informes y los cuadros de mando siguen siendo importantes para las organizaciones, la mayoría de las arquitecturas de datos tradicionales inhiben la exploración de datos y un análisis más sofisticado. Además, las arquitecturas de datos tradicionales tienen varias implicaciones adicionales para los científicos de datos.

- Los datos de alto valor son difíciles de alcanzar y aprovechar, y las actividades de análisis predictivo y minería de datos son las últimas en la fila de datos. Debido a que los EDW están diseñados para la gestión centralizada de datos y la generación de informes, los datos que desean analizar se priorizan generalmente después de los procesos operativos.
- Los datos se mueven en lotes desde la EDW a las herramientas analíticas locales. Este flujo de trabajo significa que los científicos de datos están limitados a realizar análisis en memoria (como con R, SAS, SPSS o Excel), lo que restringirá el tamaño de los conjuntos de datos que pueden usar. Como tal, el análisis puede estar sujeto a restricciones de muestreo, que pueden sesgar la precisión del modelo.
- Los proyectos de ciencia de datos permanecerán aislados y ad hoc, en lugar de administrados de forma centralizada. La implicación de este aislamiento es que la organización nunca puede aprovechar el poder de la analítica avanzada de una manera escalable, y los proyectos de ciencia de datos existirán como iniciativas no estándar, que con frecuencia no están alineadas con los objetivos o la estrategia empresarial corporativa.

Todos estos síntomas de la arquitectura de datos tradicional dan como resultado un "tiempo de comprensión" lento y un impacto comercial menor que el que se podría lograr si los datos fueran más fácilmente accesibles y apoyados por un entorno que promoviera la analítica avanzada. Como se mencionó anteriormente, una solución a este problema es introducir cajas de arena analíticas para permitir que los científicos de datos realicen análisis avanzados de manera controlada y autorizada. Mientras tanto, las soluciones actuales de almacenamiento de datos continúan ofreciendo informes y servicios de BI para respaldar la gestión y las operaciones de misión crítica.

### 1.2.3 Impulsores de Big Data

Para comprender mejor los impulsores del mercado relacionados con Big Data, es útil comprender primero algunos antecedentes de los almacenes de datos y los tipos de repositorios y herramientas para administrar estos almacenes de datos.

Como se muestra en la Figura 1-10, en la década de 1990 el volumen de información se media a menudo en terabytes. La mayoría de las organizaciones analizaron datos estructurados en filas y columnas y utilizaron bases de datos relacionales y almacenes de datos para gestionar grandes almacenes de información empresarial. La década siguiente vio una proliferación de diferentes tipos de fuentes de datos, principalmente herramientas de productividad y publicación, como repositorios de administración de contenido y sistemas de almacenamiento adjuntos en red, que administran este tipo de información, y los datos comenzaron a aumentar de tamaño y comenzaron a medirse a escalas de petabytes.. En la década de 2010, la información que las organizaciones intentan gestionar se ha ampliado para incluir muchos otros tipos de datos. En esta era, todo el mundo y todo está dejando una huella digital. La Figura 1-10 muestra una perspectiva resumida sobre las fuentes de Big Data generadas por nuevas aplicaciones y la escala y tasa de crecimiento de los datos. Estas aplicaciones, que generan volúmenes de datos que se pueden medir en una escala de exabytes, brindan oportunidades para nuevos análisis e impulsan un nuevo valor para las organizaciones. Los datos ahora provienen de varias fuentes, como estas:

- Información médica, como secuenciación genómica e imágenes de diagnóstico.
- Fotos y secuencias de video subidas a WorldWideWeb
- Videovigilancia, como las miles de cámaras de video repartidas por una ciudad.
- Dispositivos móviles, que proporcionan datos de ubicación geoespacial de los usuarios, así como metadatos sobre mensajes de texto, llamadas telefónicas y uso de aplicaciones en teléfonos inteligentes.
- Dispositivos inteligentes, que proporcionan una recopilación de información basada en sensores de redes eléctricas inteligentes, edificios inteligentes y muchas otras infraestructuras públicas e industriales.

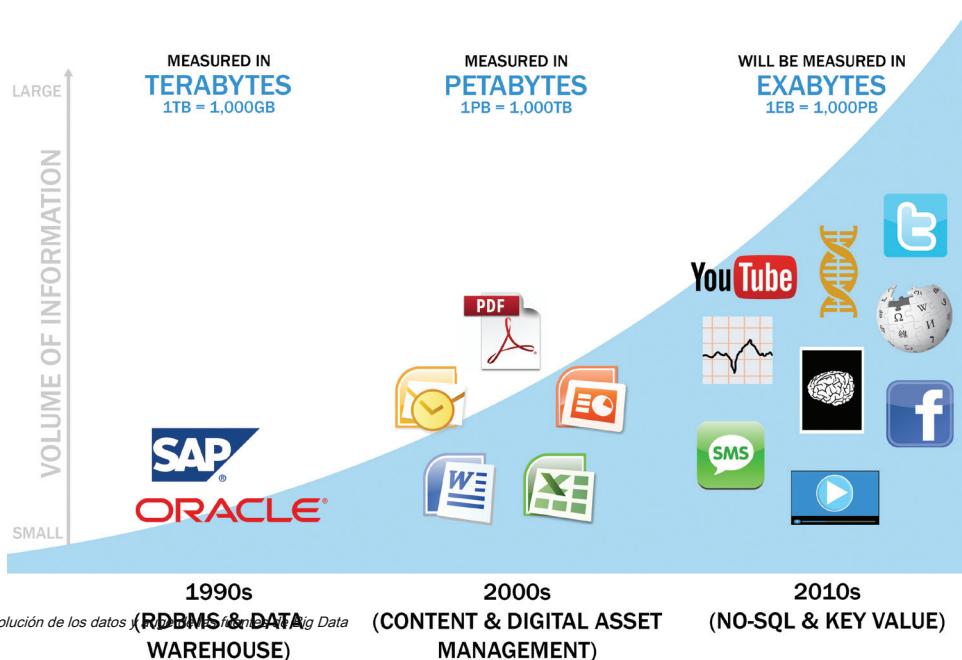


FIGURE 1-10 Evolución de los datos (Big Data vs. Big Data)

La tendencia de Big Data está generando una enorme cantidad de información de muchas fuentes nuevas. Esta avalancha de datos requiere análisis avanzados y nuevos actores del mercado para aprovechar estas oportunidades y la nueva dinámica del mercado, que se analizarán en la siguiente sección.

## 1.2.4 Ecosistema emergente de Big Data y un nuevo enfoque de análisis

Las organizaciones y los recolectores de datos se están dando cuenta de que los datos que pueden recopilar de las personas contienen un valor intrínseco y, como resultado, está surgiendo una nueva economía. A medida que esta nueva economía digital continúa

evolucionan, el mercado ve la introducción de proveedores de datos y limpiadores de datos que utilizan el crowdsourcing (como Mechanical Turk y GalaxyZoo) para probar los resultados de las técnicas de aprendizaje automático. Otros proveedores ofrecen valor agregado al volver a empaquetar las herramientas de código abierto de una manera más simple y llevar las herramientas al mercado. Proveedores como Cloudera, Hortonworks y Pivotal han proporcionado este valor agregado para el marco de código abierto Hadoop.

A medida que el nuevo ecosistema toma forma, hay cuatro grupos principales de jugadores dentro de esta red interconectada. Estos se muestran en la Figura 1-11.

- **Dispositivos de datos** [que se muestra en la sección (1) de la Figura 1-11] y "SensorNet" recopilan datos de varias ubicaciones y generan continuamente nuevos datos sobre estos datos. Por cada gigabyte de nuevos datos creados, se crea un petabyte adicional de datos sobre esos datos. [2]
  - Por ejemplo, considere a alguien que juega un videojuego en línea a través de una PC, consola de juegos o teléfono inteligente. En este caso, el proveedor de videojuegos captura datos sobre la habilidad y los niveles alcanzados por el jugador. Los sistemas inteligentes monitorean y registran cómo y cuándo el usuario juega el juego. Como consecuencia, el proveedor del juego puede ajustar con precisión la dificultad del juego, sugerir otros juegos relacionados que probablemente interesarían al usuario y ofrecer equipo adicional y mejoras para el personaje según la edad, el género y los intereses del usuario. Esta información puede almacenarse localmente o cargarse en la nube del proveedor del juego para analizar los hábitos de juego y las oportunidades de venta adicional y venta cruzada, e identificar perfiles arquetípicos de tipos específicos de usuarios.
  - Los teléfonos inteligentes proporcionan otra rica fuente de datos. Además de la mensajería y el uso básico del teléfono, almacenan y transmiten datos sobre el uso de Internet, el uso de SMS y la ubicación en tiempo real. Estos metadatos se pueden utilizar para analizar patrones de tráfico mediante el escaneo de la densidad de teléfonos inteligentes en ubicaciones para rastrear la velocidad de los automóviles o la congestión relativa del tráfico en carreteras con mucho tráfico. De esta manera, los dispositivos GPS en los automóviles pueden brindar a los conductores actualizaciones en tiempo real y ofrecer rutas alternativas para evitar retrasos en el tráfico.
  - Las tarjetas de fidelización de compras minoristas registran no solo la cantidad que gasta una persona, sino también las ubicaciones de las tiendas que visita, los tipos de productos comprados, las tiendas donde se compran los bienes con mayor frecuencia y las combinaciones de productos comprados juntos. La recopilación de estos datos proporciona información sobre los hábitos de compra y viaje y la probabilidad de una orientación publicitaria exitosa para ciertos tipos de promociones minoristas.
- **Recopiladores de datos** [los óvalos azules, identificados como (2) en la Figura 1-11] incluyen entidades de muestra que recopilan datos del dispositivo y los usuarios.
  - Los datos son el resultado de un proveedor de televisión por cable que rastrea los programas que ve una persona, qué canales de televisión alguien pagará y no pagará para verlos a pedido, y los precios que alguien está dispuesto a pagar por contenido de televisión premium.
  - Tiendas minoristas que rastrean el camino que toma un cliente a través de su tienda mientras empujan un carrito de compras con un chip RFID para que puedan medir qué productos obtienen más tráfico utilizando datos geoespaciales recopilados de los chips RFID
- **Agregadores de datos** (los óvalos de color gris oscuro en la Figura 1-11, marcados como (3)) dan sentido a los datos recopilados de las diversas entidades de "SensorNet" o "Internet de las cosas". Estas organizaciones recopilan datos de los dispositivos y patrones de uso recopilados por agencias gubernamentales, tiendas minoristas,

y sitios web. A su vez, pueden optar por transformar y empaquetar los datos como productos para vender a los corredores de listas, que pueden querer generar listas de marketing de personas que pueden ser buenos objetivos para campañas publicitarias específicas.

- **Usuarios y compradores de datos** se indican con (4) en la Figura 1-11. Estos grupos se benefician directamente de los datos recopilados y agregados por otros dentro de la cadena de valor de los datos.

- Los bancos minoristas, actuando como compradores de datos, pueden querer saber qué clientes tienen la mayor probabilidad de solicitar una segunda hipoteca o una línea de crédito con garantía hipotecaria. Para proporcionar información para este análisis, los bancos minoristas pueden comprar datos de un agregador de datos. Este tipo de datos puede incluir información demográfica sobre personas que viven en lugares específicos; personas que parecen tener un nivel específico de deuda, pero que aún tienen puntajes crediticios sólidos (u otras características como pagar las facturas a tiempo y tener cuentas de ahorro) que se pueden usar para inferir la solvencia crediticia; y aquellos que buscan en la web información sobre el pago de deudas o proyectos de remodelación de viviendas. La obtención de datos de estas diversas fuentes y agregadores permitirá una campaña de marketing más dirigida,

- Al utilizar tecnologías como Hadoop para realizar el procesamiento del lenguaje natural en datos textuales no estructurados de sitios web de redes sociales, los usuarios pueden medir la reacción a eventos como las campañas presidenciales. La gente puede, por ejemplo, querer determinar los sentimientos del público hacia un candidato analizando blogs relacionados y comentarios en línea. De manera similar, los usuarios de datos pueden querer rastrear y prepararse para desastres naturales identificando qué áreas

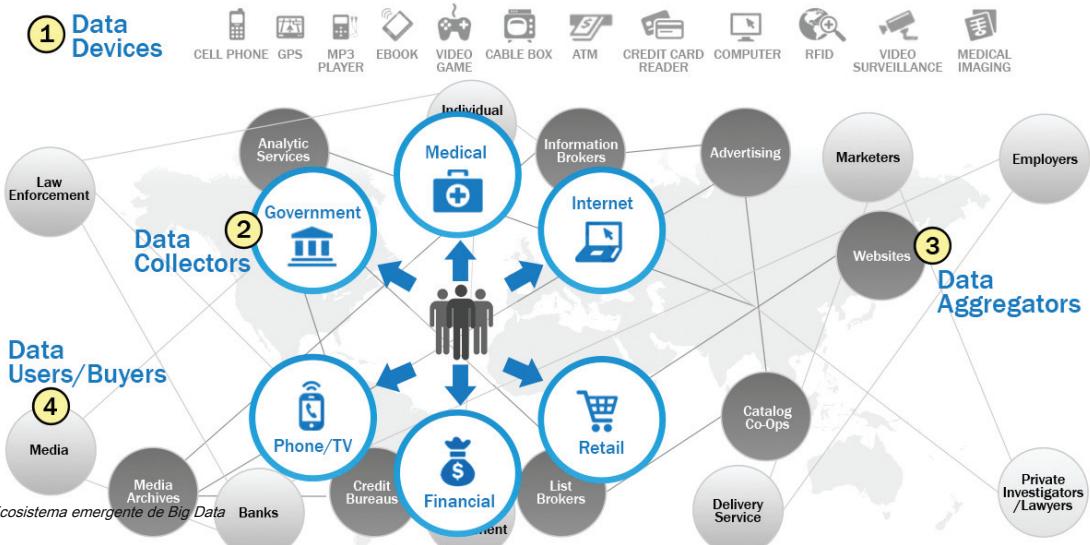


FIGURE 1-11 Ecosistema emergente de Big Data

Como ilustra este ecosistema emergente de Big Data, los tipos de datos y la dinámica del mercado relacionada varían mucho. Estos conjuntos de datos pueden incluir datos de sensores, texto, conjuntos de datos estructurados y redes sociales. Teniendo esto en cuenta, vale la pena recordar que estos conjuntos de datos no funcionarán bien dentro de los EDW tradicionales, que se diseñaron para optimizar los informes y los paneles de control y se administrarán de forma centralizada. En cambio, los problemas y proyectos de BigData requieren diferentes enfoques para tener éxito.

Los analistas deben asociarse con TI y administradores de bases de datos para obtener los datos que necesitan dentro de una caja de arena analítica. Una caja de arena analítica típica contiene datos sin procesar, datos agregados y datos con varios tipos de estructura. El sandbox permite una exploración sólida de datos y requiere un usuario experto para aprovechar y aprovechar los datos en el entorno del sandbox.

## 1.3 Funciones clave para el ecosistema de NewBig Data

Como se explica en el contexto del ecosistema BigData en la Sección 1.2.4, han surgido nuevos jugadores para seleccionar, almacenar, producir, limpiar y realizar transacciones de datos. Además, la necesidad de aplicar técnicas analíticas más avanzadas a problemas empresariales cada vez más complejos ha impulsado la aparición de nuevos roles, nuevas plataformas tecnológicas y nuevos métodos analíticos. Esta sección explora los nuevos roles que abordan estas necesidades, y los capítulos siguientes exploran algunos de los métodos analíticos y plataformas tecnológicas.

1-12. Estos roles fueron

### Three Key Roles of The New Data Ecosystem

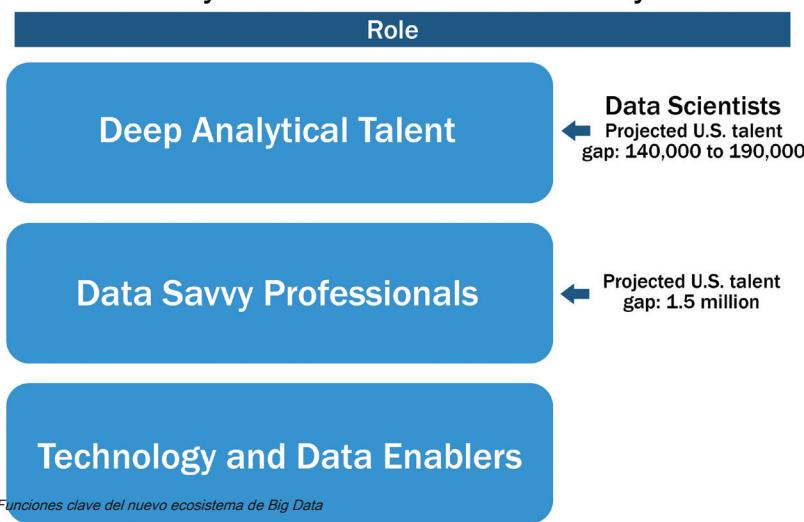


FIGURE 1-12 Funciones clave del nuevo ecosistema de Big Data

Note: Figures above reflect a projected talent gap in US in 2018, as shown in McKinsey May 2011 article "Big Data: The Next Frontier for Innovation, Competition, and Productivity"

El primer grupo, DeepAnalytical Talent, es técnicamente experto, con fuertes habilidades analíticas. Los miembros poseen una combinación de habilidades para manejar datos sin procesar, no estructurados y para aplicar técnicas analíticas complejas en

escalas masivas. Este grupo cuenta con formación avanzada en disciplinas cuantitativas, como matemáticas, estadística y aprendizaje automático.

Para hacer su trabajo, los miembros necesitan acceso a un espacio de trabajo o una caja de arena analítica sólida donde puedan realizar experimentos de datos analíticos a gran escala. Ejemplos de profesiones actuales que encajan en este grupo incluyen estadísticos, economistas, matemáticos y el nuevo rol del científico de datos.

El estudio de McKinsey prevé que para el año 2018, Estados Unidos tendrá una brecha de talento de 140.000 190.000 personas con profundo talento analítico. Esto no representa la cantidad de personas necesarias con un profundo talento analítico; más bien, este rango representa la diferencia entre lo que estará disponible en la fuerza laboral en comparación con lo que se necesitará. Además, estas estimaciones solo reflejan la escasez de talento prevista en los Estados Unidos; el número sería mucho mayor a nivel mundial.

El segundo grupo, los profesionales expertos en datos, tiene menos profundidad técnica, pero tiene un conocimiento básico de estadística o aprendizaje automático y puede definir preguntas clave que pueden responderse mediante análisis avanzados. Estas personas tienden a tener un conocimiento básico del trabajo con datos, o aprecian parte del trabajo que realizan los científicos de datos y otros con un profundo talento analítico. Ejemplos de profesionales conocedores de datos incluyen analistas financieros, analistas de investigación de mercado, científicos de vida, gerentes de operaciones y gerentes funcionales y de negocios.

El estudio de McKinsey pronostica que la brecha de talento estadounidense proyectada para este grupo será de 1,5 millones de personas para el año 2018. En un nivel alto, esto significa que para cada perfil de Data Scientist necesario, la brecha será diez veces mayor para los profesionales expertos en datos. Pasar a convertirse en un profesional conocedor de los datos es un paso fundamental para ampliar la perspectiva de los gerentes, directores y líderes, ya que esto proporciona una idea de los tipos de preguntas que se pueden resolver con los datos.

La tercera categoría de personas mencionadas en el estudio es la tecnología y los habilitadores de datos. Este grupo representa a personas que brindan experiencia técnica para respaldar proyectos analíticos, como el aprovisionamiento y la administración de entornos sandbox analíticos, y la administración de arquitecturas de datos a gran escala que permiten un análisis generalizado dentro de las empresas y otras organizaciones. Esta función requiere habilidades relacionadas con la ingeniería informática, la programación y la administración de bases de datos.

Estos tres grupos deben trabajar juntos de cerca para resolver desafíos complejos de BigData. La mayoría de las organizaciones están familiarizadas con las personas de los dos últimos grupos mencionados, pero el primer grupo, Deep Analytical Talent, tiende a ser el rol más nuevo para la mayoría y para los menos comprendidos. Para simplificar, esta discusión se centra en el papel emergente del científico de datos. Describe los tipos de actividades que realiza ese rol y proporciona una visión más detallada de las habilidades necesarias para cumplir ese rol.

Hay tres conjuntos recurrentes de actividades que realizan los científicos de datos:

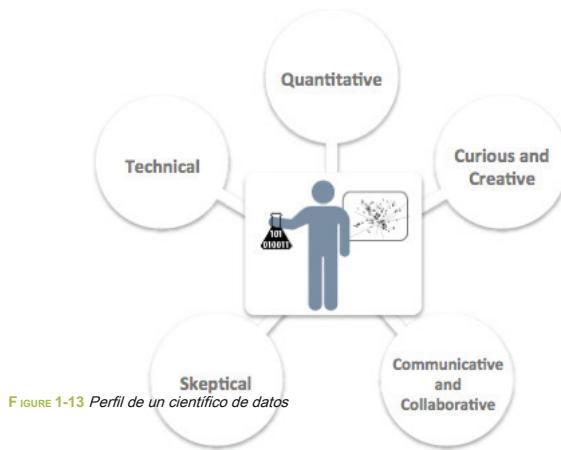
- **Replantear los desafíos comerciales como desafíos analíticos.** Específicamente, esta es una habilidad para diagnosticar problemas comerciales, considerar el núcleo de un problema dado y determinar qué tipos de métodos analíticos candidatos se pueden aplicar para resolverlo. Este concepto se explora con más detalle en el Capítulo 2, "Ciclo de vida del análisis de datos".
- **Diseñar, implementar y desplegar modelos estadísticos y técnicas de minería de datos en BigData.** Este conjunto de actividades es principalmente lo que la gente piensa cuando considera el papel del científico de datos:

es decir, aplicar métodos analíticos complejos o avanzados a una variedad de problemas comerciales utilizando datos. El capítulo 3 al capítulo 11 de este libro presenta al lector muchas de las técnicas y herramientas analíticas más populares en esta área.

- **Desarrolle conocimientos que conduzcan a recomendaciones prácticas.** Es fundamental tener en cuenta que la aplicación de métodos avanzados a los problemas actuales no genera necesariamente un nuevo valor empresarial. En cambio, es importante aprender a extraer conocimientos de los datos y comunicar el tema de manera efectiva. El Capítulo 12, "El final del juego, o ponerlo todo junto", tiene una breve descripción general de las técnicas para hacer esto.

Generalmente se piensa que los científicos de datos tienen cinco conjuntos principales de habilidades y características de comportamiento, como se muestra en la Figura 1-13:

- **Habilidad cuantitativa:** como matemáticas o estadística
  - **Aptitud técnica:** a saber, ingeniería de software, aprendizaje automático y habilidades de programación
  - **Mentalidad escéptica y pensamiento crítico:** Es importante que los científicos de datos puedan examinar su trabajo de manera crítica en lugar de hacerlo de forma unilateral.
  - **Curioso y creativo:** A los científicos de datos les apasionan los datos y encuentran formas creativas de resolver problemas y representar información.
  - **Comunicativo y colaborativo:** Los científicos de datos deben poder articular el valor empresarial en CA
- incluidos los patrocinadores del proyecto y  
estaca



Los científicos de datos generalmente se sienten cómodos usando esta combinación de habilidades para adquirir, administrar, analizar y visualizar datos y contar historias convincentes sobre ellos. La siguiente sección incluye ejemplos de lo que los equipos de ciencia de datos han creado para impulsar nuevo valor o innovación con Big Data.

## 1.4 Ejemplos de análisis de Big Data

Después de describir el ecosistema emergente de BigData y los nuevos roles necesarios para respaldar su crecimiento, esta sección proporciona tres ejemplos de Big Data Analytics en diferentes áreas: comercio minorista, infraestructura de TI y redes sociales.

Como se mencionó anteriormente, Big Data presenta muchas oportunidades para mejorar las ventas y los análisis de marketing. Un ejemplo de esto es el minorista estadounidense Target. El libro de Charles Duhigg *El poder del hábito* [4] analiza cómo Target utilizó Big Data y métodos analíticos avanzados para generar nuevos ingresos. Después de analizar el comportamiento de compra del consumidor, los estadísticos de Target determinaron que el minorista ganó una gran cantidad de dinero con tres situaciones principales de eventos de la vida.

- Matrimonio, cuando la gente tiende a comprar muchos productos nuevos
- Divorcio, cuando las personas compran productos nuevos y cambian sus hábitos de gasto.
- Embarazo, cuando las personas tienen muchas cosas nuevas que comprar y tienen la urgencia de comprarlas

Target determinó que el más lucrativo de estos eventos de la vida es la tercera situación: el embarazo. Con los datos recopilados de los compradores, Target pudo identificar este hecho y predecir cuáles de sus compradores estaban embarazadas. En un caso, Target supo que una compradora estaba embarazada incluso antes de que su familia lo supiera [5]. Este tipo de conocimiento permitió a Target ofrecer cupones e incentivos específicos a sus compradores embarazadas. De hecho, Target no solo pudo determinar si una compradora estaba embarazada, sino en qué mes de embarazo podría estarla. Esto permitió a Target administrar su inventario, sabiendo que habría demanda de productos específicos y probablemente variaría de un mes a otro durante los próximos ciclos de nueve a diez meses.

Hadoop [6] representa otro ejemplo de innovación de BigData en la infraestructura de TI. Apache Hadoop es un marco de código abierto que permite a las empresas procesar grandes cantidades de información de forma altamente paralelizada. Hadoop representa una implementación específica del paradigma MapReduce y fue diseñado por Doug Cutting y Mike Cafarella en 2005 para usar datos con diferentes estructuras. Es un marco técnico ideal para muchos proyectos de Big Data, que se basan en conjuntos de datos grandes o difíciles de manejar con estructuras de datos no convencionales. Uno de los principales beneficios de Hadoop es que emplea un sistema de archivos distribuidos, lo que significa que puede usar un clúster distribuido de servidores y hardware básico para procesar grandes cantidades de datos. Algunos de los ejemplos más comunes de implementaciones de Hadoop se encuentran en el espacio de las redes sociales, donde Hadoop puede administrar transacciones, brindar actualizaciones textuales, y desarrollar gráficos sociales entre millones de usuarios. Twitter y Facebook generan cantidades masivas de datos no estructurados y utilizan Hadoop y su ecosistema de herramientas para gestionar este gran volumen. Hadoop y su ecosistema se tratan en el Capítulo 10, "Análisis avanzado: tecnología y herramientas: MapReduce y Hadoop".

Por último, las redes sociales representan una gran oportunidad para aprovechar las interacciones sociales y profesionales para obtener nuevos conocimientos. LinkedIn ejemplifica una empresa en la que los datos en sí son el producto. Al principio, el fundador de LinkedIn, Reid Hoffman, vio la oportunidad de crear una red social para profesionales que trabajaban.

A partir de 2014, LinkedIn tiene más de 250 millones de cuentas de usuario y ha agregado muchas funciones adicionales y productos relacionados con los datos, como reclutamiento, herramientas de búsqueda de empleo, publicidad e InMaps, que muestran un gráfico social de la red profesional de un usuario. La figura 1-14 es un ejemplo de una visualización de InMap que permite

son contactos y entender cómo él



FIGURE 1-14 Visualización de datos de la red social de un usuario mediante InMaps

## Resumen

Big Data proviene de una mirada de fuentes, incluidas las redes sociales, los sensores, el Internet de las cosas, la videovigilancia y muchas fuentes de datos que pueden no haber sido considerados datos ni siquiera hace unos años. Mientras las empresas luchan por mantenerse al día con los requisitos cambiantes del mercado, algunas empresas están encontrando formas creativas de aplicar Big Data a sus crecientes necesidades comerciales y problemas cada vez más complejos. A medida que las organizaciones evolucionan sus procesos y ven las oportunidades que puede brindar BigData, intentan ir más allá de las actividades de BI tradicionales, como el uso de datos para completar informes y paneles, y avanzar hacia proyectos impulsados por la ciencia de datos que intentan responder más abiertos y complejos. preguntas.

Sin embargo, aprovechar las oportunidades que presenta Big Data requiere nuevas arquitecturas de datos, incluidos entornos de pruebas analíticos, nuevas formas de trabajar y personas con nuevos conjuntos de habilidades. Estos impulsores están haciendo que las organizaciones configuren entornos sandbox analíticos y creen equipos de ciencia de datos. Aunque algunas organizaciones tienen la suerte de tener científicos de datos, la mayoría no lo tienen, porque existe una brecha de talento cada vez mayor que dificulta encontrar y contratar científicos de datos de manera oportuna. Aún así, organizaciones como las de venta minorista en la web, la atención médica, la genómica, las nuevas infraestructuras de TI y las redes sociales están comenzando a aprovechar el Big Data y aplicarlo de formas creativas y novedosas.

## Ejercicios

1. ¿Cuáles son las tres características de Big Data y cuáles son las principales consideraciones en el procesamiento de Big Data?

¿Datos?

2. ¿Qué es una caja de arena analítica y por qué es importante?
3. Explique las diferencias entre BI y Data Science.
4. Describir los desafíos de la arquitectura analítica actual para los científicos de datos.
5. ¿Cuáles son los conjuntos de habilidades clave y las características de comportamiento de un científico de datos?

## Bibliografía

- [1] CBBB Manika, "Big Data: La próxima frontera para la innovación, la competencia y la productividad", McKinsey Global Institute, 2011.
- [2] DR John Gantz, "El universo digital en 2020: Big Data, sombras digitales más grandes y el mayor crecimiento en el Lejano Oriente", IDC, 2013.
- [3] <http://www.willisresilience.com/emc-datalab> [ En línea].
- [4] C. Duhigg, *El poder del hábito: por qué hacemos lo que hacemos en la vida y los negocios*, Nueva York: aleatorio Casa, 2012.
- [5] K. Hill, "Cómo Target descubrió que una adolescente estaba embarazada antes que su padre", Forbes, febrero 2012.
- [6] <http://hadoop.apache.org> [ En línea].

# 2

## Ciclo de vida de análisis de datos

*Conceptos clave*

*Descubrimiento*

*Preparación de datos*

*Planificación de modelos*

*Ejecución del modelo*

*Comunicar resultados*

*Operacionalizar*

Los proyectos de ciencia de datos difieren de la mayoría de los proyectos tradicionales de inteligencia empresarial y muchos proyectos de análisis de datos en que los proyectos de ciencia de datos son de naturaleza más exploratoria. Por esta razón, es fundamental tener un proceso que los gobiernos y garantizar que los participantes sean minuciosos y rigurosos en su enfoque, pero no tan rígidos que el proceso impida la exploración.

Muchos problemas que parecen enormes y abrumadores al principio se pueden dividir en partes más pequeñas o en fases procesables que pueden abordarse más fácilmente. Tener un buen proceso asegura un método completo y repetible para realizar análisis. Además, ayuda a concentrar el tiempo y la energía en las primeras etapas del proceso para comprender claramente el problema comercial que se debe resolver.

Un error común que se comete en los proyectos de ciencia de datos es apresurarse en la recopilación y el análisis de datos, lo que impide dedicar suficiente tiempo a planificar y determinar la cantidad de trabajo involucrado, comprender los requisitos o incluso enmarcar correctamente el problema del negocio. En consecuencia, los participantes pueden descubrir a mitad de camino que los patrocinadores del proyecto en realidad están tratando de lograr un objetivo que puede no coincidir con los datos disponibles, o están tratando de abordar un interés que difiere de lo que se ha comunicado explícitamente. Cuando esto sucede, es posible que el proyecto deba volver a las fases iniciales del proceso para una fase de descubrimiento adecuada, o el proyecto puede cancelarse.

Crear y documentar un proceso ayuda a demostrar rigor, lo que proporciona credibilidad adicional al proyecto cuando el equipo de ciencia de datos comparte sus hallazgos. Un proceso bien definido también ofrece un marco común para que otros lo adopten, de modo que los métodos y el análisis se puedan repetir en el futuro o cuando nuevos miembros se unan a un equipo.

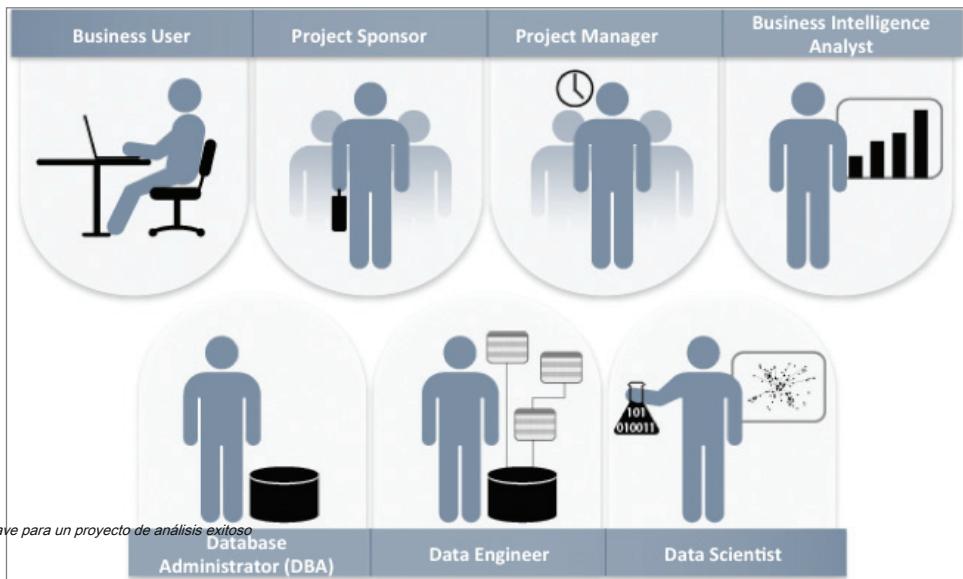
## 2.1 Descripción general del ciclo de vida del análisis de datos

El ciclo de vida de análisis de datos está diseñado específicamente para problemas de Big Data y proyectos de ciencia de datos. El ciclo de vida tiene seis fases y el trabajo del proyecto puede ocurrir en varias fases a la vez. Para la mayoría de las fases del ciclo de vida, el movimiento puede ser hacia adelante o hacia atrás. Esta descripción iterativa del ciclo de vida está destinada a representar más de cerca un proyecto real, en el que los aspectos del proyecto avanzan y pueden volver a etapas anteriores a medida que se descubre nueva información y los miembros del equipo aprenden más sobre las diversas etapas del proyecto. Esto permite a los participantes moverse iterativamente a través del proceso y avanzar hacia la operacionalización del trabajo del proyecto.

### 2.1.1 Funciones clave para un proyecto de análisis exitoso

En los últimos años, se ha prestado mucha atención al papel emergente del científico de datos. En octubre de 2012, Harvard Business Review publicó un artículo titulado "Científico de datos: el trabajo más sexy del siglo XXI" [1], en el que los expertos DJ Patil y Tom Davenport describieron el nuevo rol y cómo encontrar y contratar científicos de datos. Cada año se celebran más conferencias centradas en la innovación en las áreas de Data Science y temas relacionados con BigData. A pesar de este fuerte enfoque en el papel emergente del científico de datos específicamente, en realidad hay siete roles clave que deben cumplirse para que un equipo de ciencia de datos de alto funcionamiento ejecute proyectos analíticos con éxito.

La figura 2-1 muestra los distintos roles y las partes interesadas clave de un proyecto de análisis. Cada uno juega un papel fundamental en un proyecto de análisis exitoso. Aunque se enumeran siete roles, menos o más personas pueden realizar el trabajo según el alcance del proyecto, la estructura organizativa y las habilidades de los participantes. Por ejemplo, en un equipo pequeño y versátil, estos siete roles pueden ser cumplidos por solo 3 personas, pero un proyecto muy grande puede requerir 20 o más personas. Siguen los siete roles.



- **Usuario comercial:** Alguien que comprenda el área de dominio y generalmente se beneficie de los resultados. Esta persona puede consultar y asesorar al equipo del proyecto sobre el contexto del proyecto, el valor de los resultados y cómo se pondrán en práctica los productos. Por lo general, un analista de negocios, un gerente de línea o un experto en temas profundos en el dominio del proyecto cumple esta función.
- **Patrocinador de proyecto:** Responsable de la génesis del proyecto. Proporciona el ímpetu y los requisitos para el proyecto y define el problema empresarial central. Generalmente proporciona la financiación y mide el grado de valor de los resultados finales del equipo de trabajo. Esta persona establece las prioridades del proyecto y aclara los resultados deseados.
- **Gerente de proyecto:** Asegura que los hitos clave y los objetivos se cumplan a tiempo y con la calidad esperada.
- **Analista de inteligencia empresarial:** Proporciona experiencia en el dominio empresarial basada en una comprensión profunda de los datos, los indicadores clave de rendimiento (KPI), las métricas clave y la inteligencia empresarial desde una perspectiva de informes. Los analistas de inteligencia empresarial generalmente crean paneles e informes y tienen conocimiento de las fuentes y fuentes de datos.
- **Administrador de base de datos (DBA):** Provista y configura el entorno de la base de datos para respaldar las necesidades analíticas del equipo de trabajo. Estas responsabilidades pueden incluir proporcionar acceso a bases de datos o tablas clave y garantizar que existan los niveles de seguridad adecuados relacionados con los repositorios de datos.
- **Ingeniero de datos:** Aprovecha las habilidades técnicas profundas para ayudar a ajustar las consultas SQL para la gestión de datos y la extracción de datos, y brinda soporte para la ingestión de datos en la caja de arena analítica, que

se discutió en el Capítulo 1, "Introducción a BigData Analytics". Mientras que el DBA configura y configura las bases de datos que se utilizarán, el ingeniero de datos ejecuta las extracciones de datos reales y realiza una manipulación sustancial de datos para facilitar el análisis. El ingeniero de datos trabaja en estrecha colaboración con el científico de datos para ayudar a dar forma a los datos de la manera correcta para los análisis.

- **Científico de datos:** Proporciona experiencia en la materia para técnicas analíticas, modelado de datos y aplicación de técnicas analíticas válidas a problemas comerciales determinados. Garantiza que se cumplan los objetivos generales de análisis. Diseña y ejecuta métodos y enfoques analíticos con los datos disponibles para el proyecto.

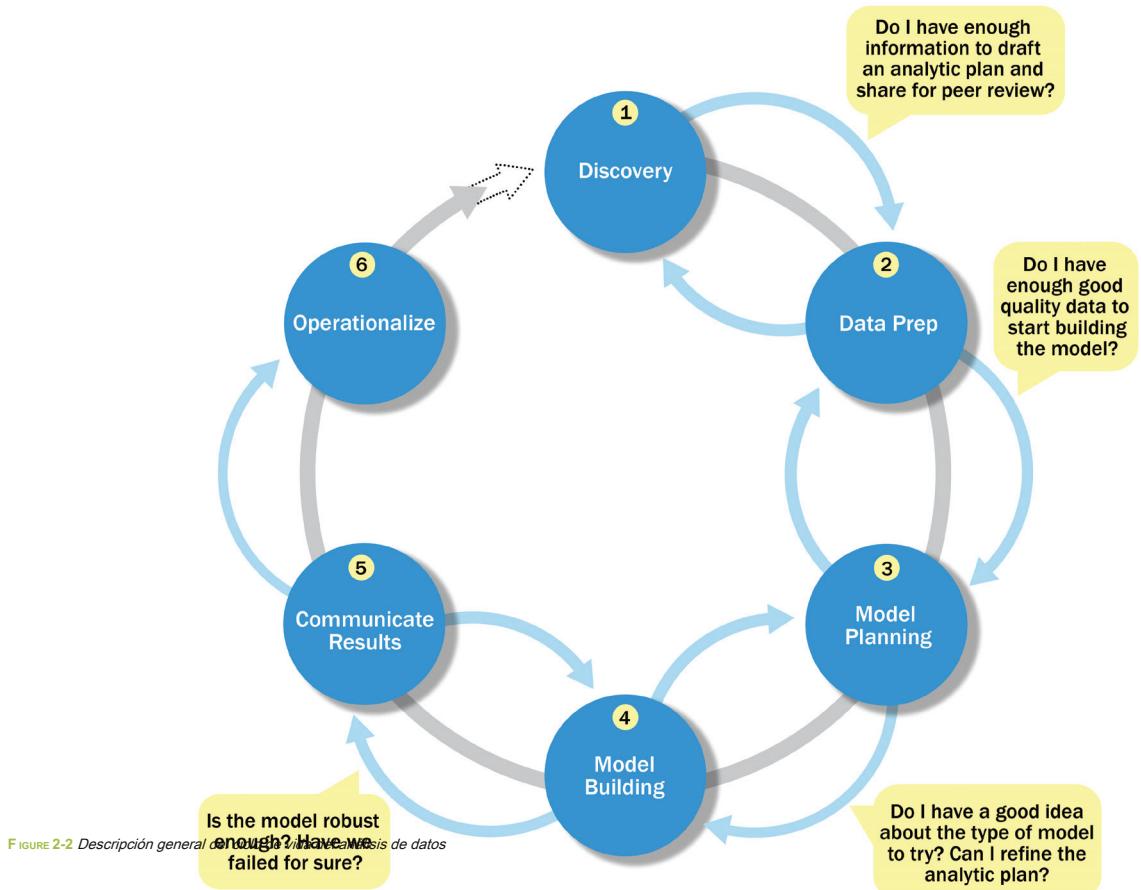
Aunque la mayoría de estos roles no son nuevos, los dos últimos roles, ingeniero de datos y científico de datos, se han vuelto populares y tienen una gran demanda [2] a medida que ha crecido el interés en Big Data.

## 2.1.2 Antecedentes y descripción general del ciclo de vida del análisis de datos

El ciclo de vida de análisis de datos define las mejores prácticas del proceso de análisis que van desde el descubrimiento hasta la finalización del proyecto. El ciclo de vida se basa en métodos establecidos en el ámbito del análisis de datos y la ciencia de decisiones. Esta síntesis se desarrolló después de recopilar aportes de científicos de datos y consultar enfoques establecidos que proporcionaron aportes sobre partes del proceso. Varios de los procesos que se consultaron incluyen estos:

- **Método científico** [ 3], en uso durante siglos, todavía proporciona un marco sólido para pensar y deconstruir problemas en sus partes principales. Una de las ideas más valiosas del método científico se relaciona con la formación de hipótesis y la búsqueda de formas de probar ideas.
- **CRISP-DM** [ 4] proporciona información útil para enmarcar problemas de análisis y es un enfoque popular para la minería de datos.
- TomDavenport's **DELTA** marco [5]: El marco DELTA ofrece un enfoque para proyectos de análisis de datos, incluido el contexto de las habilidades de la organización, los conjuntos de datos y el compromiso de liderazgo.
- DougHubbard's **Economía de la información aplicada (AIE)** enfoque [6]: AIE proporciona un marco para medir intangibles y proporciona orientación sobre el desarrollo de modelos de decisión, la calibración de estimaciones de expertos y la derivación del valor esperado de la información.
- "**Habilidades locas**" por Cohen et al. [7] ofrece información para varias de las técnicas mencionadas en las Fases 2-4 que se enfocan en la planificación, ejecución y hallazgos clave del modelo.

La Figura 2-2 presenta una descripción general del ciclo de vida del análisis de datos que incluye seis fases. Los equipos comúnmente aprenden cosas nuevas en una fase que los hace retroceder y refinar el trabajo realizado en fases anteriores en función de nuevos conocimientos e información que se han descubierto. Por esta razón, la Figura 2-2 se muestra como un ciclo. Las flechas circulares transmiten un movimiento iterativo entre fases hasta que los miembros del equipo tienen suficiente información para pasar a la siguiente fase. Las llamadas incluyen preguntas de muestra para ayudar a guiar si cada uno de los miembros del equipo tiene suficiente información y ha avanzado lo suficiente para pasar a la siguiente fase del proceso. Tenga en cuenta que estas fases no representan puertas de escenario formales; más bien, sirven como criterios para ayudar a probar si tiene sentido permanecer en la fase actual o pasar a la siguiente.



A continuación, se ofrece una breve descripción general de las principales fases del ciclo de vida del análisis de datos:

- **Fase 1 — Descubrimiento:** En la Fase 1, el equipo aprende el dominio empresarial, incluido el historial relevante, como si la organización o la unidad de negocio ha intentado proyectos similares en el pasado de los que pueden aprender. El equipo reúne los recursos disponibles para apoyar el proyecto en términos de personas, tecnología, tiempo y datos. Las actividades importantes en esta fase incluyen enmarcar los problemas comerciales, un desafío analítico que se puede abordar en fases posteriores y formular hipótesis iniciales (IH) para probar y comenzar a aprender los datos.
- **Fase 2: preparación de datos:** La fase 2 requiere la presencia de una caja de arena analítica, en la que el equipo puede trabajar con datos y análisis de rendimiento durante la duración del proyecto. El equipo necesita ejecutar extraer, cargar y transformar (ETL) o extraer, transformar y cargar (ETL) para obtener datos en la caja de arena. El ETL y ETL a veces se abrevian ETLT. Los datos deben transformarse en el proceso ETLT para que el equipo pueda trabajar con ellos y analizarlos. En esta fase, el equipo también debe familiarizarse a fondo con los datos y tomar medidas para condicionar los datos (Sección 2.3.4).

- **Fase 3: planificación del modelo:** La fase 3 es la planificación del modelo, en la que el equipo determina los métodos, las técnicas y el flujo de trabajo que pretende seguir para la siguiente fase de construcción del modelo. El equipo explora los datos para conocer las relaciones entre las variables y posteriormente selecciona las variables clave y los modelos más adecuados.
- **Fase 4: construcción de modelos:** En la Fase 4, el equipo desarrolla conjuntos de datos con fines de prueba, capacitación y producción. Además, en esta fase el equipo construye y ejecuta modelos basados en el trabajo realizado en la fase de planificación del modelo. El equipo también considera si sus herramientas existentes serán suficientes para ejecutar los modelos o si necesitará un entorno más robusto para ejecutar modelos y flujos de trabajo (por ejemplo, hardware rápido y procesamiento paralelo, si corresponde).
- **Fase 5: comunicar los resultados:** En la Fase 5, el equipo, en colaboración con las principales partes interesadas, determina si los resultados del proyecto son un éxito o un fracaso basándose en los criterios desarrollados en la Fase 1. El equipo debe identificar los hallazgos clave, cuantificar el valor comercial y desarrollar una narrativa para resumir y transmitir los hallazgos a las partes interesadas.
- **Fase 6 — Poner en funcionamiento:** En la Fase 6, el equipo entrega informes finales, resúmenes, código y documentos técnicos. Además, el equipo puede ejecutar un proyecto piloto para implementar los modelos en un entorno de producción.

Una vez que los miembros del equipo han ejecutado modelos y producido hallazgos, es fundamental enmarcar estos resultados de una manera que se adapte a la audiencia que involucró al equipo. Además, es fundamental enmarcar los resultados del trabajo de una manera que demuestre un valor claro. Si el equipo realiza un análisis técnicamente preciso pero no logra traducir los resultados a un lenguaje que resuene con la audiencia, la gente no verá el valor y se habrá desperdiciado gran parte del tiempo y esfuerzo en el proyecto.

El resto del capítulo está organizado como sigue. Las secciones 2.2 a 2.7 discuten en detalle cómo funciona cada una de las seis fases, y la sección 2.8 muestra un caso de estudio de la incorporación del ciclo de vida de análisis de datos en un proyecto de ciencia de datos del mundo real.

## 2.2 Fase 1: Descubrimiento

La primera fase del ciclo de vida del análisis de datos implica el descubrimiento (Figura 2-3). En esta fase, el equipo de ciencia de datos debe aprender e investigar el problema, desarrollar el contexto y la comprensión, y aprender sobre las fuentes de datos necesarias y disponibles para el proyecto. Además, el equipo formula hipótesis iniciales que luego pueden probarse con datos.

### 2.2.1 Aprendizaje del dominio empresarial

Comprender el área de dominio del problema es esencial. En muchos casos, los científicos de datos tendrán un profundo conocimiento computacional y cuantitativo que se puede aplicar ampliamente en muchas disciplinas. Un ejemplo de este rol sería alguien con un título avanzado en matemáticas aplicadas o estadística.

Estos científicos de datos tienen un conocimiento profundo de los métodos, técnicas y formas de aplicar la heurística a una variedad de problemas comerciales y conceptuales. Otros en esta área pueden tener un conocimiento profundo de un área de dominio, junto con experiencia cuantitativa. Un ejemplo de esto sería alguien con un Ph.D. en ciencias de la vida. Esta persona tendría un conocimiento profundo de un campo de estudio, como la oceanografía, la biología o la genética, con cierto conocimiento cuantitativo.

En esta etapa temprana del proceso, el equipo debe determinar cuánto conocimiento de dominio o de negocios necesita el científico de datos para desarrollar modelos en las Fases 3 y 4. Cuanto antes el equipo pueda realizar esta evaluación

mejor beca

el equipo tiene la plataforma

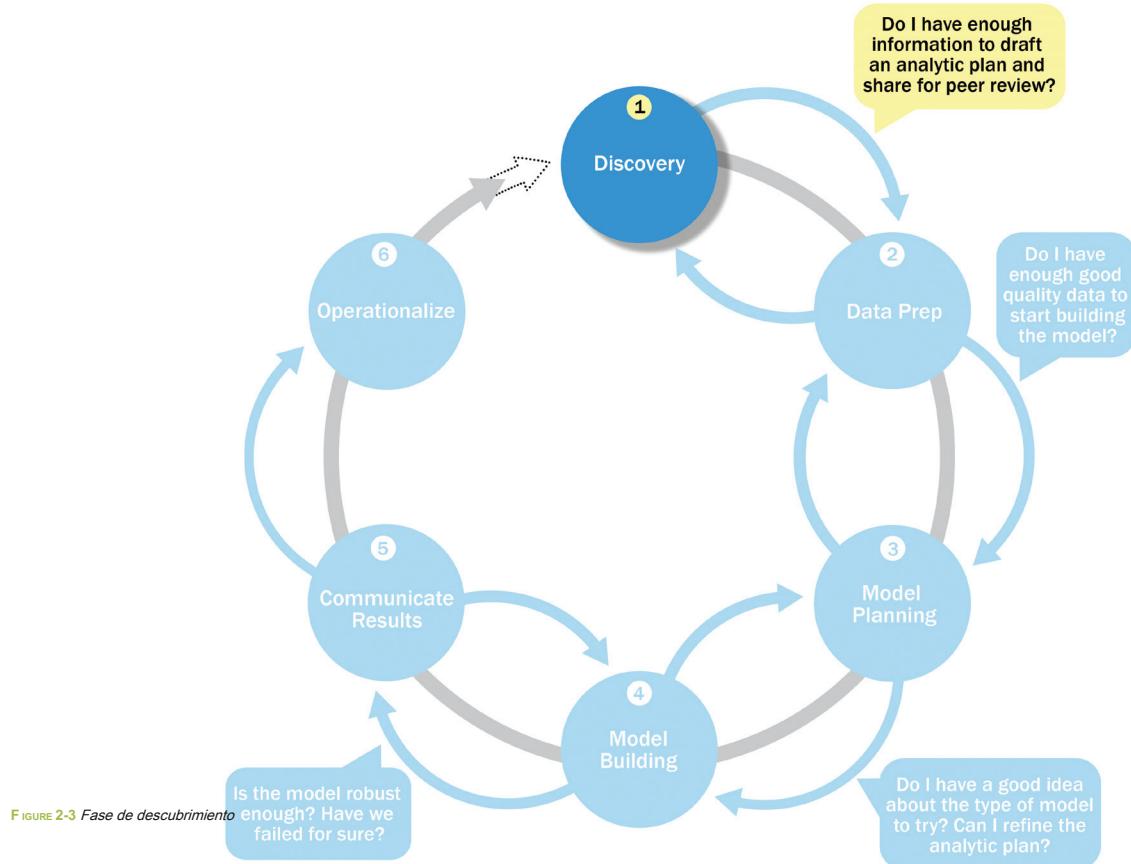


FIGURE 2-3 Fase de descubrimiento

Como parte de la fase de descubrimiento, el equipo debe evaluar los recursos disponibles para respaldar el proyecto. En este contexto, los recursos incluyen tecnología, herramientas, sistemas, datos y personas.

Durante esta determinación del alcance, considere las herramientas y la tecnología disponibles que el equipo utilizará y los tipos de sistemas necesarios para que las fases posteriores pongan en funcionamiento los modelos. Además, trate de evaluar el nivel de sofisticación analítica dentro de la organización y las brechas que puedan existir relacionadas con herramientas, tecnología y habilidades. Por ejemplo, para que el modelo que se está desarrollando tenga longevidad en una organización, considere qué tipos de habilidades y roles se requerirán que pueden no existir en la actualidad.

Para que el proyecto tenga éxito a largo plazo,

¿Qué tipo de habilidades y roles serán necesarios para los destinatarios del modelo que se está desarrollando? ¿Existe el nivel requerido de experiencia dentro de la organización hoy en día, o será necesario cultivarlo? Responder a estas preguntas influirá en las técnicas que el equipo seleccione y el tipo de implementación que el equipo elija seguir en las siguientes fases del ciclo de vida de análisis de datos.

Además de las habilidades y los recursos informáticos, es recomendable hacer un inventario de los tipos de datos disponibles para el equipo para el proyecto. Considere si los datos disponibles son suficientes para respaldar los objetivos del proyecto. El equipo deberá determinar si debe recopilar datos adicionales, comprarlos de fuentes externas o transformar los datos existentes. A menudo, los proyectos se inician mirando solo los datos disponibles. Cuando los datos son menores de lo esperado, el tamaño y alcance del proyecto se reduce para trabajar dentro de las limitaciones de los datos existentes.

Un enfoque alternativo es considerar los objetivos a largo plazo de este tipo de proyecto, sin estar limitado por los datos actuales. Luego, el equipo puede considerar qué datos se necesitan para alcanzar los objetivos a largo plazo y qué partes de este viaje de varios pasos se pueden lograr hoy con los datos existentes. La consideración de objetivos a más largo plazo junto con objetivos a corto plazo permite a los equipos perseguir proyectos más ambiciosos y tratar un proyecto como el primer paso de una iniciativa más estratégica, en lugar de como una iniciativa independiente. Es fundamental ver los proyectos como parte de un viaje a más largo plazo, especialmente si se ejecutan proyectos en una organización que es nueva en Data Science y es posible que no se haya embarcado en los conjuntos de datos óptimos para respaldar análisis sólidos hasta este momento.

Asegúrese de que el equipo del proyecto tenga la combinación adecuada de expertos en el dominio, clientes, talento analítico y gestión de proyectos para que sea eficaz. Además, evalúe cuánto tiempo se necesita y si el equipo tiene la amplitud y profundidad adecuadas de habilidades.

Después de hacer un inventario de las herramientas, la tecnología, los datos y las personas, considere si el equipo tiene recursos suficientes para tener éxito en este proyecto o si se necesitan recursos adicionales. Negociar los recursos al inicio del proyecto, mientras se definen las metas, los objetivos y la viabilidad, es generalmente más útil que más adelante en el proceso y garantiza el tiempo suficiente para ejecutarlo correctamente. Los gerentes de proyecto y las partes interesadas clave tienen más éxito en la negociación de los recursos adecuados en esta etapa que después una vez que el proyecto está en marcha.

### 2.2.3 Enmarcando el problema

Enmarcar bien el problema es fundamental para el éxito del proyecto. **Enmarcado** es el proceso de plantear el problema analítico que se va a resolver. En este punto, es una buena práctica escribir el enunciado del problema y compartirlo con las partes interesadas clave. Cada miembro del equipo puede escuchar cosas ligeramente diferentes relacionadas con las necesidades y el problema y tener ideas algo diferentes de posibles soluciones. Por estas razones, es fundamental establecer el problema de la analítica, así como por qué y para quién es importante. Esencialmente, el equipo necesita articular claramente la situación actual y sus principales desafíos.

Como parte de esta actividad, es importante identificar los principales objetivos del proyecto, identificar lo que se debe lograr en términos comerciales e identificar lo que se debe hacer para satisfacer las necesidades. Además, considere los objetivos y los criterios de éxito del proyecto. ¿Qué está tratando de lograr el equipo al realizar el proyecto y qué se considerará "suficientemente bueno" como resultado del proyecto? Este es un documento crítico y compartido con el equipo del proyecto y las partes interesadas clave. Es una buena práctica compartir la declaración de objetivos y criterios de éxito con el equipo y la comunicación con las expectativas del patrocinador del proyecto.

Quizás igualmente importante sea establecer criterios de falla. La mayoría de las personas que realizan proyectos prefieren pensar únicamente en los criterios de éxito y en cómo se verán las condiciones cuando los participantes tengan éxito. Sin embargo, esto es casi un enfoque en el mejor de los casos, asumiendo que todo procederá según lo planeado.

y el equipo del proyecto alcanzará sus metas. Sin embargo, independientemente de lo bien planificado que sea, es casi imposible planificar todo lo que surgirá en un proyecto. Los criterios de falla guiarán al equipo a comprender cuándo es mejor dejar de intentarlo o conformarse con los resultados que se han obtenido de los datos. Muchas veces la gente continuará realizando análisis más allá del punto en el que se pueda extraer información significativa de los datos. Establecer criterios tanto para el éxito como para el fracaso ayuda a los participantes a evitar esfuerzos improductivos y a permanecer alineados con los patrocinadores del proyecto.

#### **2.2.4 Identificación de las partes interesadas clave**

Otro paso importante es identificar a las partes interesadas clave y sus intereses en el proyecto. Durante estas discusiones, el equipo puede identificar los criterios de éxito, los riesgos clave y las partes interesadas, que deben incluir a cualquiera que se beneficie del proyecto o que se vea afectado de manera significativa por el proyecto. Al entrevistar a las partes interesadas, aprenda sobre el área de dominio y cualquier historial relevante de proyectos analíticos similares. Por ejemplo, el equipo puede identificar los resultados que cada interesado desea del proyecto y los criterios que utilizará para juzgar el éxito del proyecto.

Tenga en cuenta que el proyecto de análisis se está iniciando por una razón. Es fundamental articular los puntos débiles de la manera más clara posible para abordarlos y ser consciente de las áreas a perseguir o evitar a medida que el equipo avanza en el proceso analítico.

Dependiendo de la cantidad de partes interesadas y participantes, el equipo puede considerar describir el tipo de actividad y participación que se espera de cada parte interesada y participante. Esto establecerá expectativas claras con los participantes y evitara retrasos más adelante cuando, por ejemplo, el equipo pueda sentir que necesita esperar la aprobación de alguien que se ve a sí mismo como un asesor en lugar de un aprobador del producto de trabajo.

#### **2.2.5 Entrevistar al patrocinador de análisis**

El equipo debe planear colaborar con las partes interesadas para aclarar y enmarcar el problema de análisis. Al principio, los patrocinadores del proyecto pueden tener una solución predeterminada que no necesariamente logra el resultado deseado. En estos casos, el equipo debe utilizar su conocimiento y experiencia para identificar el verdadero problema subyacente y la solución adecuada.

Por ejemplo, supongamos que en la fase inicial de un proyecto, se le pide al equipo que cree un sistema de recomendación para el negocio y que la forma de hacerlo es hablando con tres personas e integrando el recomendador de productos en un sistema corporativo heredado. Aunque este puede ser un enfoque válido, es importante probar los supuestos y desarrollar una comprensión clara del problema. Por lo general, el equipo de ciencia de datos puede tener una comprensión más objetiva del conjunto de problemas que las partes interesadas, que pueden estar sugiriendo soluciones a un problema determinado. Por lo tanto, el equipo puede profundizar en el contexto y el dominio para definir claramente el problema y proponer posibles caminos desde el problema hasta el resultado deseado. En esencia, el equipo de ciencia de datos puede adoptar un enfoque más objetivo, ya que las partes interesadas pueden haber desarrollado sesgos a lo largo del tiempo, según su experiencia. Además, lo que pudo haber sido cierto en el pasado puede que ya no sea una suposición de trabajo válida. Una forma posible de eludir este problema es que el patrocinador del proyecto se centre en definir claramente los requisitos, mientras que los otros miembros del equipo de ciencia de datos se centran en los métodos necesarios para lograr los objetivos.

Al entrevistar a las principales partes interesadas, el equipo debe tomarse un tiempo para entrevistar a fondo al patrocinador del proyecto, que suele ser quien financia el proyecto o proporciona los requisitos de alto nivel. Esta persona comprende el problema y generalmente tiene una idea de una posible solución de trabajo. Esto es crítico

comprender a fondo la perspectiva del patrocinador para guiar al equipo en la puesta en marcha del proyecto. A continuación, se ofrecen algunos consejos para entrevistar a los patrocinadores de proyectos:

- Prepárese para la entrevista; redactar preguntas y revisarlas con colegas.
- Utilice preguntas abiertas; evite hacer preguntas capciosas.
- Investigue los detalles y plantee preguntas de seguimiento.
- Evite llenar cada silencio en la conversación; Dé tiempo a la otra persona para pensar.
- Deje que los patrocinadores expresen sus ideas y hagan preguntas aclaratorias, como “¿Por qué? ¿Es eso correcto? ¿Está esta idea en el objetivo? ¿Hay algo más?”
- Utilice técnicas de escucha activa; Repita lo que se escuchó para asegurarse de que el equipo lo escuchó correctamente, o reformule lo que se dijo.
- Trate de evitar expresar las opiniones del equipo, que pueden introducir sesgos; en cambio, concéntrese en escuchar.
- Tenga en cuenta el lenguaje corporal de los entrevistadores y las partes interesadas; use el contacto visual cuando sea apropiado y esté atento.
- Minimiza las distracciones.
- Documente lo que escuchó el equipo y reviselo con los patrocinadores.

A continuación se incluye una breve lista de preguntas comunes que resultan útiles durante la fase de descubrimiento al entrevistar al patrocinador del proyecto. Las respuestas comenzarán a dar forma al alcance del proyecto y le darán al equipo una idea de las metas y objetivos del proyecto.

- ¿Qué problema empresarial está intentando resolver el equipo?
- ¿Cuál es el resultado deseado del proyecto?
- ¿Qué fuentes de datos están disponibles?
- ¿Qué problemas de la industria pueden afectar el análisis?
- ¿Qué plazos deben tenerse en cuenta?
- ¿Quién podría proporcionar información sobre el proyecto?
- ¿Quién tiene la autoridad final para tomar decisiones sobre el proyecto?
- ¿Cómo cambiará el enfoque y el alcance del problema si cambian las siguientes dimensiones:
  - **Hora:** ¿Analizando datos de 1 año o 10 años?
  - **Personas:** Evaluar el impacto de los cambios en los recursos en el cronograma del proyecto.
  - **Riesgo:** De conservador a agresivo
  - **Recursos:** Ninguno a ilimitado (herramientas, tecnología, sistemas)
  - **Tamaño y atributos de los datos:** Incluyendo fuentes de datos internas y externas

## 2.2.6 Desarrollo de hipótesis iniciales

El desarrollo de un conjunto de IH es una faceta clave de la fase de descubrimiento. Este paso implica formar ideas que el equipo pueda probar con datos. Generalmente, es mejor proponer algunas hipótesis primarias para probar y luego ser creativo al desarrollar varias más. Estos IH forman la base de las pruebas analíticas que el equipo utilizará en fases posteriores y servirán como base para los hallazgos de la Fase 5. La prueba de hipótesis desde una perspectiva estadística se cubre con mayor detalle en el Capítulo 3, "Revisión de métodos analíticos de datos básicos utilizando R."

De esta manera, el equipo puede comparar sus respuestas con el resultado de un experimento o prueba para generar posibles soluciones adicionales a los problemas. Como resultado, el equipo tendrá un conjunto mucho más rico de observaciones para elegir y más opciones para acordar las conclusiones más impactantes de un proyecto.

Otra parte de este proceso implica recopilar y evaluar hipótesis de las partes interesadas y los expertos en el dominio que pueden tener su propia perspectiva sobre cuál es el problema, cuál debería ser la solución y cómo llegar a una solución. Estos interesados conocerían bien el área de dominio y podrían ofrecer sugerencias sobre ideas para probar mientras el equipo formula hipótesis durante esta fase. Es probable que el equipo recopile muchas ideas que puedan iluminar los supuestos operativos de las partes interesadas. Estas ideas también brindarán al equipo la oportunidad de expandir el alcance del proyecto a espacios adyacentes donde tenga sentido o diseñar experimentos de manera significativa para abordar los intereses más importantes de las partes interesadas. Como parte de este ejercicio, puede ser útil obtener y explorar algunos datos iniciales para informar las discusiones con las partes interesadas durante la etapa de formación de hipótesis.

## 2.2.7 Identificación de posibles fuentes de datos

Como parte de la fase de descubrimiento, identifique los tipos de datos que el equipo necesitará para resolver el problema. Considere el volumen, tipo y lapso de tiempo de los datos necesarios para probar las hipótesis. Asegúrese de que el equipo pueda acceder a más que simplemente datos agregados. En la mayoría de los casos, el equipo necesitará los datos brutos para evitar introducir sesgos en el análisis descendente. Recordando las características de Big Data del Capítulo 1, evalúe las características principales de los datos, con respecto a su volumen, variedad y velocidad de cambio. Un diagnóstico completo de la situación de los datos influirá en los tipos de herramientas y técnicas que se utilizarán en las Fases 2-4 del ciclo de vida del análisis de datos. Además, realizar la exploración de datos en esta fase ayudará al equipo a determinar la cantidad de datos necesarios, como la cantidad de datos históricos para extraer de los sistemas existentes y la estructura de datos.

El equipo debe realizar cinco actividades principales durante este paso de la fase de descubrimiento:

- **Identificar fuentes de datos:** Haga una lista de fuentes de datos candidatas que el equipo puede necesitar para probar las hipótesis iniciales descritas en esta fase. Haga un inventario de los conjuntos de datos disponibles actualmente y de los que se pueden comprar o adquirir para las pruebas que el equipo desea realizar.
- **Capture fuentes de datos agregadas:** Esto es para obtener una vista previa de los datos y proporcionar una comprensión de alto nivel. Permite al equipo obtener una descripción general rápida de los datos y realizar una exploración adicional en áreas específicas. También indica al equipo posibles áreas de interés dentro de los datos.
- **Revise los datos sin procesar:** Obtenga datos preliminares de las fuentes de datos iniciales. Comience a comprender las interdependencias entre los atributos de los datos y familiarícese con el contenido de los datos, su calidad y sus limitaciones.

- **Evalué las estructuras de datos y las herramientas necesarias:** El tipo y la estructura de los datos dictan qué herramientas puede utilizar el equipo para analizar los datos. Esta evaluación hace que el equipo piense qué tecnologías pueden ser buenas candidatas para el proyecto y cómo empezar a acceder a estas herramientas.
- **Alcance el tipo de infraestructura de datos necesaria para este tipo de problema:** Además de las herramientas necesarias, los datos influyen en el tipo de infraestructura necesaria, como el almacenamiento en disco y la capacidad de la red.

A diferencia de muchos procesos tradicionales de etapa-puerta, en los que el equipo puede avanzar solo cuando se cumplen criterios específicos, el ciclo de vida de DataAnalytics está diseñado para adaptarse a más ambigüedad. Esto refleja más de cerca cómo funcionan los proyectos de ciencia de datos en situaciones de la vida real. Para cada fase del proceso, se recomienda pasar ciertos puntos de control como una forma de medir si el equipo está listo para pasar a la siguiente fase del ciclo de vida de análisis de datos.

El equipo puede pasar a la siguiente fase cuando tenga suficiente información para redactar un plan de análisis y compartirlo para su revisión por pares. Aunque es posible que el proyecto no requiera una revisión por pares del plan, la creación del plan es una buena prueba de la comprensión del equipo del problema comercial y del enfoque del equipo para abordarlo. La creación del plan analítico también requiere una comprensión clara del área de dominio, el problema que se debe resolver y el alcance de las fuentes de datos que se utilizarán. El desarrollo de criterios de éxito al principio del proyecto aclara la definición del problema y ayuda al equipo cuando llega el momento de tomar decisiones sobre los métodos analíticos que se utilizan en las fases posteriores.

## 2.3 Fase 2: Preparación de datos

La segunda fase del ciclo de vida de análisis de datos implica la preparación de datos, que incluye los pasos para explorar, preprocesar y acondicionar los datos antes de modelarlos y analizarlos. En esta fase, el equipo necesita crear un entorno robusto en el que pueda explorar los datos que estén separados del entorno de producción. Por lo general, esto se hace preparando un entorno limitado de análisis. Para introducir los datos en la zona de pruebas, el equipo debe realizar ETLT mediante una combinación de extracción, transformación y carga de datos en la zona de pruebas. Una vez que los datos están en la caja de arena, el equipo necesita aprender sobre los datos y familiarizarse con ellos. Comprender los datos en detalle es fundamental para el éxito del proyecto. El equipo también debe decidir cómo condicionar y transformar los datos para obtener un formato que facilite el análisis posterior. El equipo puede realizar visualizaciones de datos para ayudar a los miembros del equipo a comprender los datos, incluidas sus tendencias, valores atípicos y relaciones entre las variables de datos. Cada uno de estos pasos de la fase de preparación de datos se analiza a lo largo de esta sección.

La preparación de datos tiende a ser el paso más laborioso en el ciclo de vida del análisis. De hecho, es común que los equipos dediquen al menos el 50% del tiempo de un proyecto de ciencia de datos en esta fase crítica. Si el equipo no puede obtener suficientes datos de calidad suficiente, es posible que no pueda realizar los pasos posteriores en el proceso del ciclo de vida.

La figura 2-4 muestra una descripción general del ciclo de vida de análisis de datos para la Fase 2. La fase de preparación de datos es generalmente la más iterativa y la que los equipos tienden a subestimar con mayor frecuencia. Esto se debe a que la mayoría de los equipos y líderes están ansiosos por comenzar a analizar los datos, probar hipótesis y obtener respuestas a algunas de las preguntas planteadas en la Fase 1. Muchos tienden a saltar a la Fase 3 o Fase 4 para comenzar a desarrollar rápidamente modelos y algoritmos sin perder el tiempo para prepararse. Los datos para modelar. En consecuencia, los equipos se dan cuenta de que los datos con los que están trabajando no les permiten ejecutar los modelos que desean y, de todos modos, terminan de nuevo en la Fase 2.

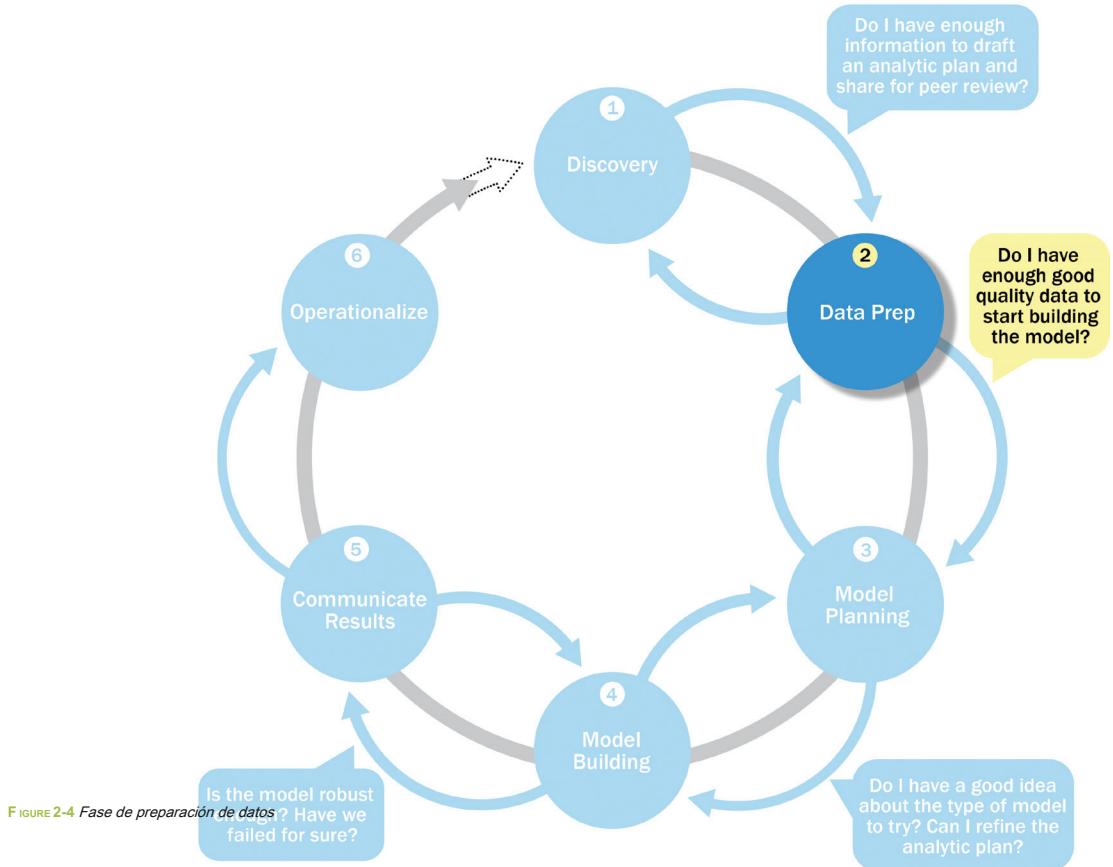


FIGURE 2-4

### 2.3.1 Preparación del entorno de pruebas analítico

La primera subfase de la preparación de datos requiere que el equipo obtenga una caja de arena analítica (también conocida como *espacio de trabajo*), en el que el equipo puede explorar los datos sin interferir con las bases de datos de producción en vivo. Considere un ejemplo en el que el equipo necesita trabajar con los datos financieros de una empresa. El equipo debe acceder a una copia de los datos financieros del entorno de pruebas analítico en lugar de interactuar con la versión de producción de la base de datos principal de la organización, porque estará estrictamente controlada y será necesaria para la presentación de informes financieros.

Al desarrollar la caja de arena analítica, es una buena práctica recopilar todo tipo de datos allí, ya que los miembros del equipo necesitan acceso a grandes volúmenes y variedades de datos para un proyecto de análisis de BigData. Esto puede incluir

todo, desde datos agregados a nivel de resumen, datos estructurados, feeds de datos sin procesar y datos de texto no estructurados de registros de llamadas o registros web, según el tipo de análisis que el equipo planea realizar.

Este enfoque expansivo para atraer datos de todo tipo difiere considerablemente del enfoque defendido por muchas organizaciones de tecnología de la información (TI). Muchos grupos de TI brindan acceso solo a un subsegmento particular de los datos para un propósito específico. A menudo, la mentalidad del grupo de TI es proporcionar la cantidad mínima de datos necesarios para permitir que el equipo logre sus objetivos. Por el contrario, el equipo de ciencia de datos quiere tener acceso a todo. Desde su perspectiva, más datos es mejor, ya que a menudo los proyectos de ciencia de datos son una mezcla de análisis impulsados por un propósito y enfoques experimentales para probar una variedad de ideas. En este contexto, puede ser un desafío para un equipo de ciencia de datos si tiene que solicitar acceso a todos y cada uno de los conjuntos de datos y atributos uno a la vez. Debido a estas diferentes opiniones sobre el acceso y uso de datos,

Durante estas discusiones, el equipo de ciencia de datos debe darle a TI una justificación para desarrollar un espacio aislado de análisis, que está separado de los almacenes de datos tradicionales gobernados por TI dentro de una organización. Equilibrar de manera amigable y exitosa las necesidades tanto del equipo de ciencia de datos como de TI requiere una relación de trabajo positiva entre múltiples grupos y propietarios de datos. El pago ~~ff~~ es genial. El sandbox analítico permite a las organizaciones emprender proyectos de ciencia de datos más ambiciosos y avanzar más allá del análisis de datos tradicional e inteligencia empresarial para realizar análisis predictivos más sólidos y avanzados.

Espere que la caja de arena sea grande. Puede contener datos sin procesar, datos agregados y otros tipos de datos que se utilizan con menos frecuencia en las organizaciones. El tamaño de la zona de pruebas puede variar mucho según el proyecto. Una buena regla es planificar que el sandbox tenga al menos 5-10 veces el tamaño de los conjuntos de datos originales, en parte porque se pueden crear copias de los datos que sirvan como tablas o almacenes de datos específicos para tipos específicos de análisis en el proyecto.

Aunque el concepto de una caja de arena analítica es relativamente nuevo, las empresas están progresando en esta área y están encontrando formas de ofrecer cajas de arena y espacios de trabajo donde los equipos pueden acceder a conjuntos de datos y trabajar de una manera que sea aceptable tanto para los equipos de ciencia de datos como para los grupos de TI.

### 2.3.2 Realización de ETLT

A medida que el equipo busca comenzar las transformaciones de datos, asegúrese de que la zona de pruebas de análisis tenga un ancho de banda amplio y conexiones de red confiables a las fuentes de datos subyacentes para permitir la lectura y escritura ininterrumpidas. En ETL, los usuarios realizan procesos de extracción, transformación y carga para extraer datos de un almacén de datos, realizar transformaciones de datos y volver a cargar los datos en el almacén de datos. Sin embargo, el enfoque de la caja de arena analítica difiere ligeramente; aboga por extraer, cargar y luego transformar. En este caso, los datos se extraen en su formato sin procesar y se cargan en el almacén de datos, donde los analistas pueden optar por transformar los datos en un nuevo estado o dejarlos en su estado original sin procesar. La razón de este enfoque es que existe un valor significativo en preservar los datos sin procesar e incluirlos en la caja de arena antes de que ocurra cualquier transformación.

Por ejemplo, considere un análisis para la detección de fraudes en el uso de tarjetas de crédito. Muchas veces, los valores atípicos en esta población de datos pueden representar transacciones de mayor riesgo que pueden ser indicativas de actividad fraudulenta de tarjetas de crédito. Con ETL, estos valores atípicos pueden filtrarse o transformarse y limpiarse inadvertidamente antes de cargarse en el almacén de datos. En este caso, los mismos datos que serían necesarios para evaluar instancias de actividad fraudulenta se limpiarían inadvertidamente, evitando el tipo de análisis que un equipo desearía hacer.

Seguir el enfoque ELT le da al equipo acceso a datos limpios para analizar después de que los datos se hayan cargado en la base de datos y le da acceso a los datos en su forma original para encontrar matices ocultos en los datos. Este enfoque es parte de la razón por la que la zona de pruebas analítica puede crecer rápidamente. El equipo puede querer datos limpios y datos agregados y puede necesitar guardar una copia de los datos originales para compararlos o compararlos.

busque patrones ocultos que puedan haber existido en los datos antes de la etapa de limpieza. Este proceso se puede resumir como ETLT para reflejar el hecho de que un equipo puede optar por realizar ETL en un caso y ELT en otro.

Dependiendo del tamaño y la cantidad de fuentes de datos, es posible que el equipo deba considerar cómo paralelizar el movimiento de los conjuntos de datos en la caja de arena. Para este propósito, el movimiento de grandes cantidades de datos a veces se denomina Big ETL. El movimiento de datos se puede paralelizar mediante tecnologías como Hadoop o MapReduce, que se explicarán con más detalle en el Capítulo 10, "Análisis avanzado: tecnología y herramientas: MapReduce y Hadoop". En este punto, tenga en cuenta que estas tecnologías se pueden utilizar para realizar ingestas de datos en paralelo e introducir una gran cantidad de archivos o conjuntos de datos en paralelo en un período de tiempo muy corto. Hadoop puede ser útil para la carga de datos, así como para el análisis de datos en fases posteriores.

Antes de mover los datos al espacio aislado analítico, determine las transformaciones que deben realizarse en los datos. Parte de esta fase implica evaluar la calidad de los datos y estructurar los conjuntos de datos correctamente para que puedan utilizarse para un análisis sólido en fases posteriores. Además, es importante considerar a qué datos tendrá acceso el equipo y qué nuevos atributos de datos deberán derivarse en los datos para permitir el análisis.

Como parte del paso ETLT, es recomendable hacer un inventario de los datos y comparar los datos actualmente disponibles con los conjuntos de datos que necesita el equipo. La realización de este tipo de análisis de brechas proporciona un marco para comprender qué conjuntos de datos puede aprovechar el equipo hoy y dónde el equipo necesita iniciar proyectos para la recopilación de datos o el acceso a nuevos conjuntos de datos que actualmente no están disponibles. Un componente de esta subfase implica extraer datos de las fuentes disponibles y determinar las conexiones de datos para datos sin procesar, bases de datos de procesamiento de transacciones en línea (OLTP), cubos de procesamiento analítico en línea (OLAP) u otras fuentes de datos.

La interfaz de programación de aplicaciones (API) es una forma cada vez más popular de acceder a una fuente de datos [8]. Muchos sitios web y aplicaciones de redes sociales ahora proporcionan API que ofrecen acceso a datos para respaldar un proyecto o complementar los conjuntos de datos con los que trabaja un equipo. Por ejemplo, conectarse a la API de Twitter puede permitir que un equipo descargue millones de tweets para realizar un proyecto de análisis de sentimientos sobre un producto, una empresa o una idea. Gran parte de los datos de Twitter están disponibles públicamente y pueden aumentar otros conjuntos de datos utilizados en el proyecto.

### 2.3.3 Aprendiendo sobre los datos

Un aspecto crítico de un proyecto de ciencia de datos es familiarizarse con los datos en sí. Dedicar tiempo a aprender los matices de los conjuntos de datos proporciona un contexto para comprender qué constituye un valor razonable y un resultado esperado frente a lo que es un hallazgo sorprendente. Además, es importante catalogar las fuentes de datos a las que el equipo tiene acceso e identificar fuentes de datos adicionales que el equipo puede aprovechar pero a las que quizás no tenga acceso hoy. Algunas de las actividades de este paso pueden superponerse con la investigación inicial de los conjuntos de datos que ocurren en la fase de descubrimiento. Al realizar esta actividad, se logran varios objetivos.

- Aclara los datos a los que tiene acceso el equipo de ciencia de datos al inicio del proyecto.
- Resalta las lagunas identificando conjuntos de datos dentro de una organización que el equipo puede encontrar útiles pero que pueden no ser accesibles para el equipo en la actualidad. Como consecuencia, esta actividad puede desencadenar un proyecto para comenzar a construir relaciones con los propietarios de los datos y encontrar formas de compartir los datos de manera apropiada. Además, esta actividad puede proporcionar un impulso para comenzar a recopilar nuevos datos que beneficien a la organización o un proyecto específico a largo plazo.
- Identifica conjuntos de datos fuera de la organización que pueden ser útiles para obtener, a través de openAPI, intercambio de datos o compra de datos para complementar conjuntos de datos ya existentes.

La Tabla 2-1 muestra una forma de organizar este tipo de inventario de datos.

**T PODER 2-1 Muestra de inventario de conjunto de datos**

Tipos de datos	Accesible	Inaccesible	Recoger	Fuentes de fiesta
Productos enviados	•			
Finanzas de producto		•		
Centro de llamadas de productos		•		
Datos				
Producto vivo			•	
Encuestas de retroalimentación				
Sentimiento del producto				•
de Redes sociales				

### 2.3.4 Acondicionamiento de datos

**Acondicionamiento de datos** se refiere al proceso de limpieza de datos, normalización de conjuntos de datos y realización de transformaciones en los datos. Un paso crítico dentro del ciclo de vida del análisis de datos, el acondicionamiento de datos puede implicar muchos pasos complejos para unir o fusionar conjuntos de datos o, de lo contrario, llevarlos a un estado que permita el análisis en fases posteriores. El acondicionamiento de datos a menudo se ve como un paso previo al procesamiento del análisis de datos porque involucra muchas operaciones en el conjunto de datos antes de desarrollar modelos para procesar o analizar los datos. Esto implica que el paso de acondicionamiento de datos lo realiza solo TI, los propietarios de los datos, un DBA o un ingeniero de datos. Sin embargo, también es importante involucrar al científico de datos en este paso porque muchas decisiones se toman en la fase de acondicionamiento de datos que afectan el análisis posterior. Parte de esta fase implica decidir qué aspectos de conjuntos de datos particulares serán útiles para analizar en pasos posteriores. Debido a que los equipos comienzan a formar ideas en esta fase sobre qué datos conservar y qué datos transformar o descartar, es importante involucrar a varios miembros del equipo en estas decisiones. Dejar tales decisiones en manos de una sola persona puede hacer que los equipos regresen a esta fase para recuperar datos que pueden haber sido descartados.

Al igual que con el ejemplo anterior sobre cómo decidir qué datos conservar en relación con la detección de fraudes en el uso de tarjetas de crédito, es fundamental reflexionar sobre qué datos elige conservar el equipo y qué datos se descartarán. Esto puede tener consecuencias de gran alcance que harán que el equipo vuelva sobre los pasos anteriores si el equipo descarta demasiados datos en un punto demasiado temprano de este proceso. Por lo general, los equipos de ciencia de datos prefieren conservar más datos que muy pocos para el análisis. Las preguntas y consideraciones adicionales para el paso de acondicionamiento de datos incluyen estas.

- ¿Cuáles son las fuentes de datos? ¿Cuáles son los campos de destino (por ejemplo, columnas de las tablas)?
- ¿Qué tan limpios están los datos?

- ¿Cuán consistentes son los contenidos y archivos? Determine hasta qué punto los datos contienen valores faltantes o inconsistentes y si los datos contienen valores que se desvian de lo normal.
- Evalúe la coherencia de los tipos de datos. Por ejemplo, si el equipo espera que ciertos datos sean numéricos, con firmeza que sean numéricos o si es una mezcla de cadenas alfanuméricas y texto.
- Revise el contenido de las columnas de datos u otras entradas y verifique que tengan sentido. Por ejemplo, si el proyecto implica analizar los niveles de ingresos, obtenga una vista previa de los datos para confirmar que los valores de los ingresos son positivos o si es aceptable tener ceros o valores negativos.
- Busque cualquier evidencia de error sistemático. Los ejemplos incluyen fuentes de datos de sensores u otras fuentes de datos que se rompen sin que nadie se dé cuenta, lo que provoca valores de datos no válidos, incorrectos o faltantes. Además, revise los datos para medir si la definición de los datos es la misma en todas las mediciones. En algunos casos, una columna de datos se reutiliza o la columna deja de llenarse sin que se anote este cambio o sin que se notifiquen otros.

### 2.3.5 Examinar y visualizar

Una vez que el equipo ha recopilado y obtenido al menos algunos de los conjuntos de datos necesarios para el análisis posterior, un paso útil es aprovechar las herramientas de visualización de datos para obtener una visión general de los datos. Ver patrones de alto nivel en los datos permite comprender las características de los datos muy rápidamente. Un ejemplo es el uso de la visualización de datos para examinar la calidad de los datos, por ejemplo, si los datos contienen muchos valores inesperados u otros indicadores de datos sucios. (Los datos sucios se analizarán con más detalle en el capítulo 3.) Otro ejemplo es la asimetría, como si la mayoría de los datos se desplazaran en gran medida hacia un valor o el final de un continuo.

Shneiderman [9] es bien conocido por su mantra para el análisis de datos visuales de "descripción general primero, zoom y filtro, luego detalles bajo demanda". Este es un enfoque pragmático para el análisis de datos visuales. Permite al usuario encontrar áreas de interés, hacer zoom y filtrar para encontrar información más detallada sobre un área particular de los datos, y luego encontrar los datos detallados detrás de un área particular. Este enfoque proporciona una vista de alto nivel de los datos y una gran cantidad de información sobre un conjunto de datos determinado en un período de tiempo relativamente corto.

Al aplicar este enfoque con una herramienta de visualización de datos o un paquete estadístico, se recomiendan las siguientes pautas y consideraciones.

- Revise los datos para asegurarse de que los cálculos se mantengan consistentes dentro de las columnas o entre las tablas para un campo de datos determinado. Por ejemplo, ¿cambió el valor de por vida del cliente en algún momento en la mitad de la recolección de datos? O, si trabajaba con finanzas, ¿cambió el cálculo del interés de simple a compuesto al final del año?
- ¿La distribución de datos se mantiene constante en todos los datos? Si no es así, ¿qué tipo de acciones deberían tomarse para abordar este problema?
- Evalúe la granularidad de los datos, el rango de valores y el nivel de agregación de los datos.
- ¿Los datos representan la población de interés? En el caso de los datos de marketing, si el proyecto se centra en los clientes en edad de crianza, ¿los datos lo representan o están llenos de personas mayores y adolescentes?
- Para las variables relacionadas con el tiempo, ¿las mediciones son diarias, semanales, mensuales? ¿Eso es lo suficientemente bueno? ¿El tiempo se mide en segundos en todas partes? ¿O es milisegundos en algunos lugares? Determine el nivel de granularidad de los datos necesarios para el análisis y evalúe si el nivel actual de marcas de tiempo en las reuniones de datos que necesitan.

- ¿Los datos están estandarizados / normalizados? ¿Son las escalas consistentes? Si no es así, ¿qué tan consistentes o irregulares son los datos?
- Para los conjuntos de datos geoespaciales, ¿las abreviaturas de los estados o países son consistentes en los datos? ¿Están normalizados los nombres personales? ¿Unidades inglesas? ¿Unidades métricas?

Estas son consideraciones típicas que deben ser parte del proceso de pensamiento mientras el equipo evalúa los conjuntos de datos que se obtienen para el proyecto. Tener un conocimiento profundo de los datos será fundamental cuando llegue el momento de construir y ejecutar modelos más adelante en el proceso.

### 2.3.6 Herramientas comunes para la fase de preparación de datos

Varias herramientas se utilizan comúnmente para esta fase:

- **Hadoop** [ 10] puede realizar ingesta masiva paralela y análisis personalizado para análisis de tráfico web, análisis de ubicación GPS, análisis genómico y combinación de alimentaciones masivas de datos no estructurados de múltiples fuentes.
- **AlpineMiner** [ 11] proporciona una interfaz gráfica de usuario (GUI) para crear flujos de trabajo analítico, incluidas manipulaciones de datos y una serie de eventos analíticos, como técnicas de extracción de datos por etapas (por ejemplo, primero seleccione los 100 clientes principales y luego ejecute estadísticas descriptivas y agrupación en clústeres) en Postgres. SQL y otras fuentes de BigData.
- **OpenRefine** ( anteriormente llamado Google Refinar) [12] es "una herramienta potente, gratuita y de código abierto para trabajar con datos confusos". Es una popular herramienta basada en GUI para realizar transformaciones de datos, y es una de las herramientas gratuitas más sólidas disponibles actualmente.
- Similar a OpenRefine, **DataWrangler** [ 13] es una herramienta interactiva para la limpieza y transformación de datos. Wrangler se desarrolló en la Universidad de Stanford y se puede utilizar para realizar muchas transformaciones en un conjunto de datos determinado. Además, las salidas de transformación de datos se pueden colocar en Java o Python. La ventaja de esta función es que un subconjunto de datos puede manipularse en Wrangler a través de su GUI, y luego las mismas operaciones pueden escribirse como código Java o Python para ejecutarse en el conjunto de datos completo y más grande en un entorno de pruebas analítico local.

Para la Fase 2, el equipo necesita asistencia de TI, DBA o quien controle EnterpriseDataWarehouse (EDW) para las fuentes de datos que el equipo de ciencia de datos quisiera usar.

## 2.4 Fase 3: Planificación del modelo

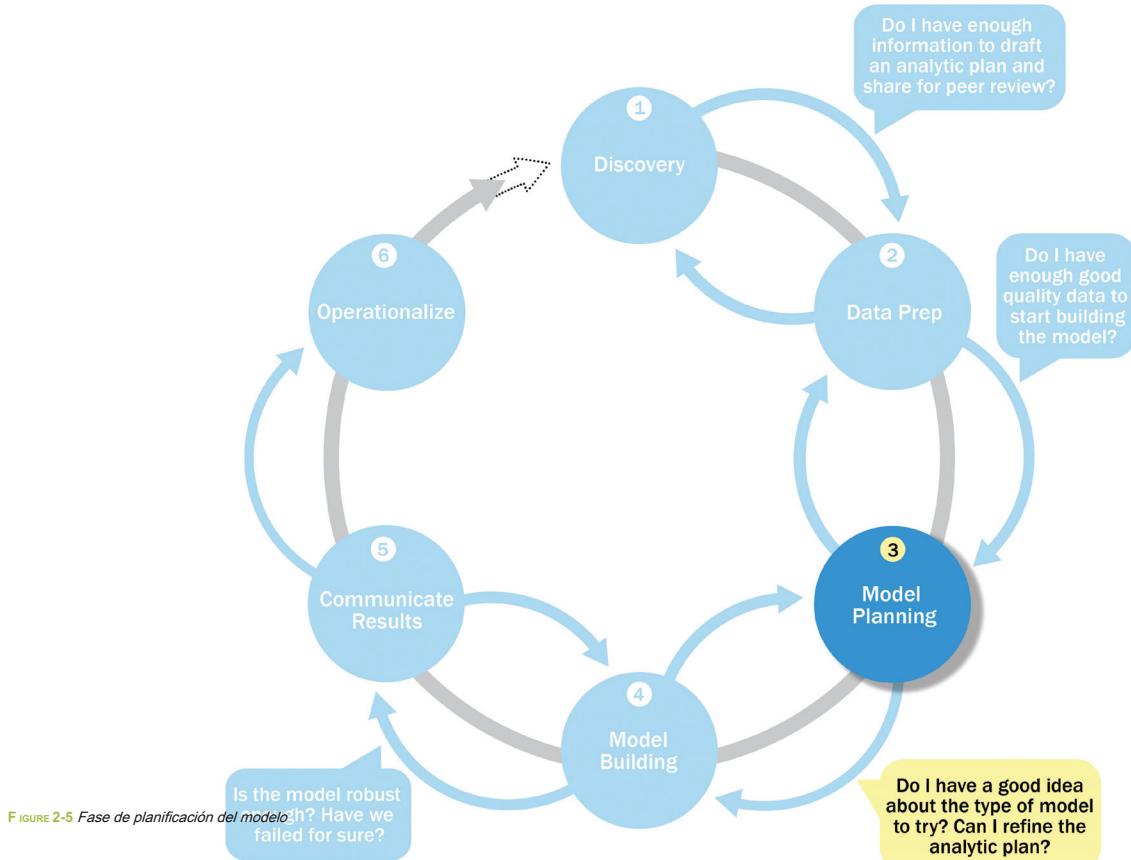
En la Fase 3, el equipo de ciencia de datos identifica modelos candidatos para aplicar a los datos para agrupar, clasificar o encontrar relaciones en los datos según el objetivo del proyecto, como se muestra en la Figura 2-5. Es durante esta fase que el equipo hace referencia a las hipótesis desarrolladas en la Fase 1, cuando por primera vez se familiarizan con los datos y comprenden los problemas de negocio o el área de dominio. Estas hipótesis ayudan al equipo a enmarcar las analíticas que se ejecutarán en la Fase 4 y a seleccionar los métodos adecuados para lograr sus objetivos.

Algunas de las actividades a considerar en esta fase incluyen las siguientes:

- Evaluar la estructura de los conjuntos de datos. La estructura de los conjuntos de datos es un factor que dicta las herramientas y técnicas analíticas para la siguiente fase. Dependiendo de si los planes de equipo para analizar datos textuales o transaccionales, por ejemplo, se requieren diferentes herramientas y enfoques.
- Asegúrese de que las técnicas analíticas permitan al equipo alcanzar los objetivos comerciales y aceptar o rechazar las hipótesis de trabajo.

- Determine si la situación amerita un modelo único o una serie de técnicas como parte de un flujo de trabajo analítico más amplio. Algunos modelos de ejemplo incluyen reglas de asociación (Capítulo 5, "Teoría y métodos analíticos avanzados: Reglas de asociación") y regresión logística (Capítulo 6, "Teoría y métodos analíticos avanzados: Regresión"). Otras herramientas, como AlpineMiner, permiten a los usuarios configurar una serie de pasos y análisis

PostgreSQL



Además de las consideraciones que acabamos de enumerar, es útil investigar y comprender cómo otros analistas generalmente abordan un tipo específico de problema. Dado el tipo de datos y recursos disponibles, evalúe si los enfoques existentes similares funcionarán o si el equipo necesitará crear algo nuevo. Muchas veces, los equipos pueden obtener ideas de problemas análogos que otras personas han resuelto en diferentes verticales de la industria o áreas de dominio. La Tabla 2-2 resume los resultados de un ejercicio de este tipo, que involucra varias áreas de dominio y los tipos de modelos previamente utilizados en un tipo de clasificación de problema después de realizar una investigación sobre modelos de abandono en múltiples verticales de la industria. Realizar este tipo de diligencia le da al equipo

ideas de cómo otros han resuelto problemas similares y presenta al equipo una lista de modelos candidatos para probar como parte de la fase de planificación del modelo.

**T PODER 2-2** *Investigación sobre planificación de modelos en verticales de la industria*

Empaquetado para el consumidor	Regresión lineal múltiple, determinación automática de relevancia (ARD) y árbol de decisión
Bienes	
Banca minorista	Regresión múltiple
Negocio al por menor	Regresión logística, ARD, árbol de decisión
Telecomunicaciones inalámbricas	Red neuronal, árbol de decisión, sistemas neurofuzzy jerárquicos, evolución de reglas, regresión logística

#### 2.4.1 Exploración de datos y selección de variables

Aunque parte de la exploración de datos se lleva a cabo en la fase de preparación de datos, esas actividades se centran principalmente en la higiene de los datos y en evaluar la calidad de los datos en sí. En la Fase 3, el objetivo de la exploración de datos es comprender las relaciones entre las variables para informar la selección de las variables y métodos y comprender el dominio del problema. Al igual que con las fases anteriores del ciclo de vida del análisis de datos, es importante dedicar tiempo y centrar la atención en este trabajo preparatorio para que las fases posteriores de selección y ejecución del modelo sean más fáciles y eficientes. Una forma habitual de realizar este paso implica el uso de herramientas para realizar visualizaciones de datos. Acerarse a la exploración de datos de esta manera ayuda al equipo a obtener una vista previa de los datos y evaluar las relaciones entre las variables a un alto nivel.

En muchos casos, las partes interesadas y los expertos en la materia tienen instintos y corazonadas sobre lo que el equipo de ciencia de datos debería considerar y analizar. Probablemente, este grupo tuvo alguna hipótesis que condujo a la génesis del proyecto. A menudo, los interesados tienen una buena comprensión del problema y el dominio, aunque es posible que no sean conscientes de las sutilezas dentro de los datos o el modelo necesario para aceptar o rechazar una hipótesis. Otras veces, las partes interesadas pueden tener razón, pero por razones equivocadas (por ejemplo, pueden tener razón sobre una correlación que existe pero inferir una razón incorrecta para la correlación). Mientras tanto, los científicos de datos tienen que abordar los problemas con una mentalidad imparcial y estar preparados para cuestionar todas las suposiciones.

A medida que el equipo comienza a cuestionar las suposiciones entrantes y a probar las ideas iniciales de los patrocinadores del proyecto y las partes interesadas, debe considerar las aportaciones y los datos que se necesitarán, y luego debe examinar si estas aportaciones están realmente correlacionadas con los resultados que el plan de equipo pretende predecir o analizar. Algunos métodos y tipos de modelos manejarán las variables correlacionadas mejor que otros. Dependiendo de lo que el equipo esté intentando resolver, es posible que deba considerar un método alternativo, reducir el número de entradas de datos o transformar las entradas para permitir que el equipo utilice el mejor método para un problema empresarial determinado. Algunas de estas técnicas se explorarán más a fondo en el Capítulo 3 y el Capítulo 6.

La clave de este enfoque es apuntar a capturar los predictores y las variables más esenciales en lugar de considerar todas las variables posibles que la gente cree que puede influir en el resultado. Abordar el problema de esta manera requiere iteraciones y pruebas para identificar las variables más esenciales para los análisis previstos. El equipo debe planear probar una variedad de variables para incluir en el modelo y luego enfocarse en las variables más importantes e influyentes.

Si el equipo planea ejecutar análisis de regresión, identifique los predictores candidatos y las variables de resultado del modelo. Planifique la creación de variables que determinen los resultados pero que demuestren una fuerte relación con el resultado en lugar de con las otras variables de entrada. Esto incluye permanecer atento a problemas como la correlación en serie, la multicolinealidad y otros desafíos típicos de modelado de datos que interfieren con la validez de estos modelos. A veces, estos problemas pueden evitarse simplemente buscando formas de replantear un problema determinado. Además, a veces lo único que se necesita es determinar la correlación (“predicción de caja negra”) y, en otros casos, el objetivo del proyecto es comprender mejor la relación causal. En este último caso,

#### **2.4.2 Selección de modelo**

En la subfase de selección del modelo, el objetivo principal del equipo es elegir una técnica analítica, o una lista corta de técnicas candidatas, en función del objetivo final del proyecto. Para el contexto de este libro, una *modelo* se discute en términos generales. En este caso, modelo simplemente se refiere a una abstracción de la realidad. Uno observa los eventos que ocurren en una situación del mundo real o con datos en vivo e intenta construir modelos que emulen este comportamiento con un conjunto de reglas y condiciones. En el caso del aprendizaje automático y la minería de datos, estas reglas y condiciones se agrupan en varios conjuntos generales de técnicas, como clasificación, reglas de asociación y agrupamiento. Al revisar esta lista de tipos de modelos potenciales, el equipo puede bajar la lista a varios modelos viables para tratar de abordar un problema dado. En el Capítulo 3 y el Capítulo 4, “Teoría y métodos analíticos avanzados: agrupación”, se proporcionan más detalles sobre cómo hacer coincidir los modelos correctos con los tipos comunes de problemas comerciales.

Una consideración adicional en esta área para tratar con Big Data implica determinar si el equipo utilizará las técnicas más adecuadas para datos estructurados, datos no estructurados o un enfoque híbrido. Por ejemplo, el equipo puede aprovechar MapReduce para analizar datos no estructurados, como se destaca en el Capítulo 10. Por último, el equipo debe tener cuidado de identificar y documentar las suposiciones de modelado que está haciendo al elegir y construir modelos preliminares.

Por lo general, los equipos crean los modelos iniciales mediante un paquete de software estadístico como R, SAS o Matlab. Aunque estas herramientas están diseñadas para algoritmos de minería de datos y aprendizaje automático, pueden tener limitaciones al aplicar los modelos a conjuntos de datos muy grandes, como es común con Big Data. Como tal, el equipo puede considerar rediseñar estos algoritmos para que se ejecuten en la propia base de datos durante la fase piloto mencionada en la Fase 6.

El equipo puede pasar a la fase de construcción del modelo una vez que tenga una buena idea sobre el tipo de modelo a probar y el equipo haya adquirido el conocimiento suficiente para perfeccionar el plan de análisis. Avanzar desde esta fase requiere una metodología general para el modelo analítico, una sólida comprensión de las variables y técnicas a utilizar y una descripción o diagrama del flujo de trabajo analítico.

#### **2.4.3 Herramientas comunes para la fase de planificación del modelo**

Hay muchas herramientas disponibles para ayudar en esta fase. A continuación, se muestran algunos de los más comunes:

- **R** [ 14] tiene un conjunto completo de capacidades de modelado y proporciona un buen entorno para construir modelos interpretativos con código de alta calidad. Además, tiene la capacidad de interactuar con bases de datos a través de una conexión ODBC y ejecutar pruebas y análisis estadísticos contra BigData a través de una conexión de código abierto. Estos dos factores hacen que Rwell sea adecuado para realizar pruebas estadísticas y análisis en BigData. En el momento de escribir este artículo, R contiene casi 5,000 paquetes para análisis de datos y representación gráfica. Los nuevos paquetes se publican con frecuencia y muchas empresas ofrecen valor agregado.

servicios para R (como capacitación, instrucción y mejores prácticas), así como empaquetarlo de manera que sea más fácil de usar y más robusto. Este fenómeno es similar a lo que sucedió con Linux a fines de la década de 1980 y principios de la de 1990, cuando las empresas parecían empaquetar y hacer que Linux fuera más fácil de consumir e implementar. Utilice extractos de archivos Rwith para análisis o ffl ine y un rendimiento óptimo, y utilice conexiones RODBC para consultas dinámicas y un desarrollo más rápido.

- **Servicios de SQLAnalysis** [ 15] puede realizar análisis en la base de datos de funciones comunes de minería de datos, agregaciones involucradas y modelos predictivos básicos.
- **SAS / ACCESO** [ 16] proporciona integración entre SAS y el entorno de pruebas de análisis a través de varios conectores de datos como ODBC, JDBC y OLEDB. El propio SAS se usa generalmente en extractos de archivos, pero con SAS / ACCESS, los usuarios pueden conectarse a bases de datos relacionales (como Oracle o Teradata) y dispositivos de almacenamiento de datos (como Greenplum or Aster), archivos y aplicaciones empresariales (como SAP y Salesforce. com).

## 2.5 Fase 4: Construcción de modelos

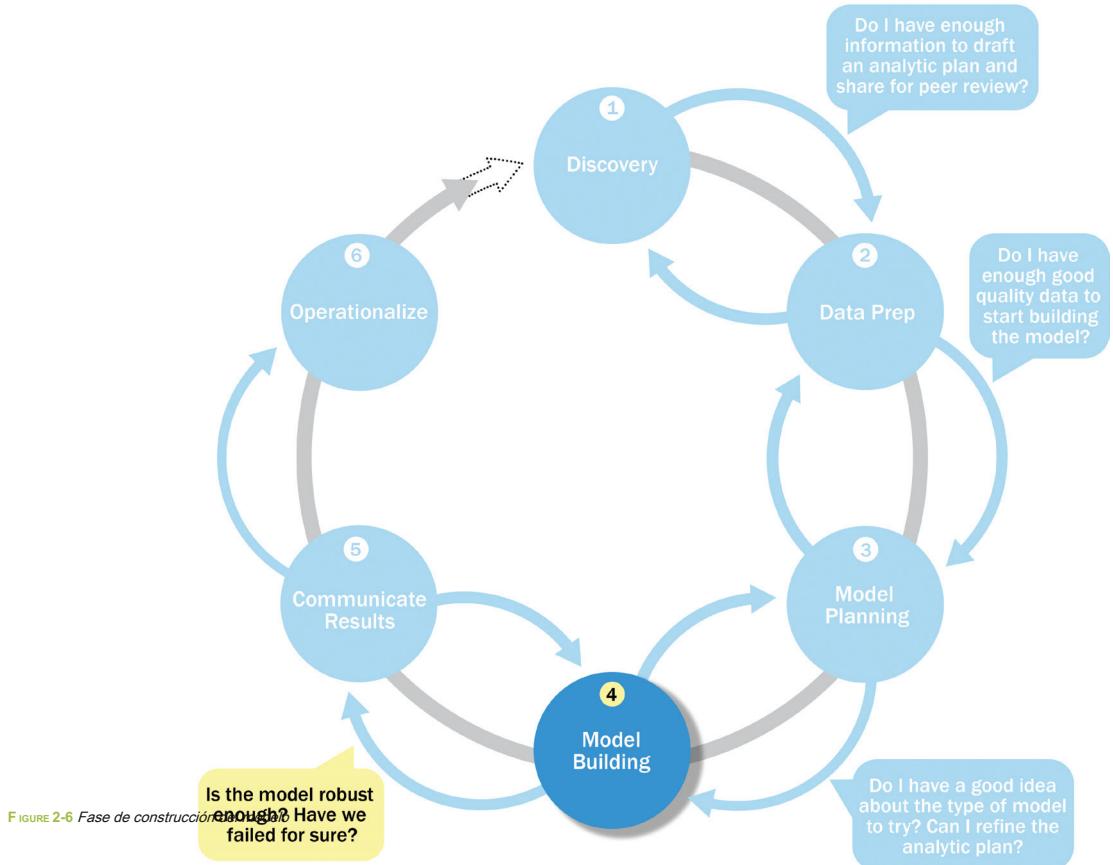
En la Fase 4, el equipo de ciencia de datos debe desarrollar conjuntos de datos con fines de capacitación, pruebas y producción. Estos conjuntos de datos permiten que el científico de datos desarrolle el modelo analítico y lo entrene ("datos de entrenamiento"), mientras mantiene a un lado algunos de los datos ("datos de reserva" o "datos de prueba") para probar el modelo. (Estos temas se tratan con más detalle en el Capítulo 3.) Durante este proceso, es fundamental asegurarse de que los conjuntos de datos de entrenamiento y prueba sean lo suficientemente robustos para el modelo y las técnicas analíticas. Una forma sencilla de pensar en estos conjuntos de datos es ver el conjunto de datos de entrenamiento para realizar los experimentos iniciales y los conjuntos de prueba para validar un enfoque una vez que se han ejecutado los experimentos y modelos iniciales.

En la fase de construcción del modelo, que se muestra en la Figura 2-6, se desarrolla un modelo analítico que se ajusta a los datos de entrenamiento y se evalúa (puntúa) contra los datos de prueba. Las fases de planificación y construcción del modelo pueden superponerse bastante y, en la práctica, uno puede ir y venir entre las dos fases durante un tiempo antes de decidirse por un modelo final.

Aunque las técnicas de modelado y la lógica necesarias para desarrollar modelos pueden ser muy complejas, la duración real de esta fase puede ser corta en comparación con el tiempo empleado en preparar los datos y definir los enfoques. En general, planifique dedicar más tiempo a preparar y aprender los datos (Fases 1–2) y elaborar una presentación de los hallazgos (Fase 5). Las fases 3 y 4 tienden a evolucionar más rápidamente, aunque son más complejas desde un punto de vista conceptual.

Como parte de esta fase, el equipo de ciencia de datos debe ejecutar los modelos definidos en la Fase 3.

Durante esta fase, los usuarios ejecutan modelos a partir de paquetes de software analíticos, como R o SAS, en extractos de archivos y pequeños conjuntos de datos con fines de prueba. A pequeña escala, evalúe la validez del modelo y sus resultados. Por ejemplo, determine si el modelo representa la mayoría de los datos y tiene un poder predictivo sólido. En este punto, refine los modelos para optimizar los resultados, por ejemplo, modificando las entradas de las variables o reduciendo las variables correlacionadas cuando sea apropiado. En la Fase 3, el equipo puede haber tenido algún conocimiento de las variables correlacionadas o los atributos de datos problemáticos, que se confirmarán o negarán una vez que los modelos se ejecuten realmente. Cuando se sumerge en los detalles de la construcción de modelos y la transformación de datos, a menudo se toman muchas decisiones pequeñas sobre los datos y el enfoque para el modelado. Estos detalles se pueden olvidar fácilmente una vez que se completa el proyecto. Por tanto, es vital registrar los resultados y la lógica del modelo durante esta fase. Además, uno debe tener cuidado de registrar cualquier supuesto operativo que se hizo en el proceso de modelado con respecto a los datos o al contexto.



La creación de modelos sólidos que sean adecuados para una situación específica requiere una consideración cuidadosa para garantizar que los modelos que se están desarrollando finalmente cumplan con los objetivos descritos en la Fase 1. Las preguntas a considerar incluyen las siguientes:

- ¿El modelo parece válido y preciso en los datos de prueba?
- ¿Tiene sentido el resultado / comportamiento del modelo para los expertos en el dominio? Es decir, ¿parece que el modelo está dando respuestas que tienen sentido en este contexto?
- ¿Tienen sentido los valores de los parámetros del modelo ajustado en el contexto del dominio?
- ¿Es el modelo lo suficientemente preciso para alcanzar el objetivo?
- ¿El modelo evita errores intolerables? Según el contexto, los falsos positivos pueden ser más graves o menos graves que los falsos negativos, por ejemplo. (Los falsos positivos y los falsos negativos se tratan con más detalle en el Capítulo 3 y el Capítulo 7, "Teoría y métodos analíticos avanzados: clasificación").

- ¿Se necesitan más datos o más entradas? ¿Es necesario transformar o eliminar alguno de los insumos?
- ¿El tipo de modelo elegido será compatible con los requisitos de tiempo de ejecución?
- ¿Se requiere una forma diferente del modelo para abordar el problema empresarial? Si es así, vuelva a la fase de planificación del modelo y revise el enfoque del modelo.

Una vez que el equipo de ciencia de datos puede evaluar si el modelo es lo suficientemente robusto para resolver el problema o si el equipo ha fallado, puede pasar a la siguiente fase del ciclo de vida de análisis de datos.

### 2.5.1 Herramientas comunes para la fase de construcción del modelo

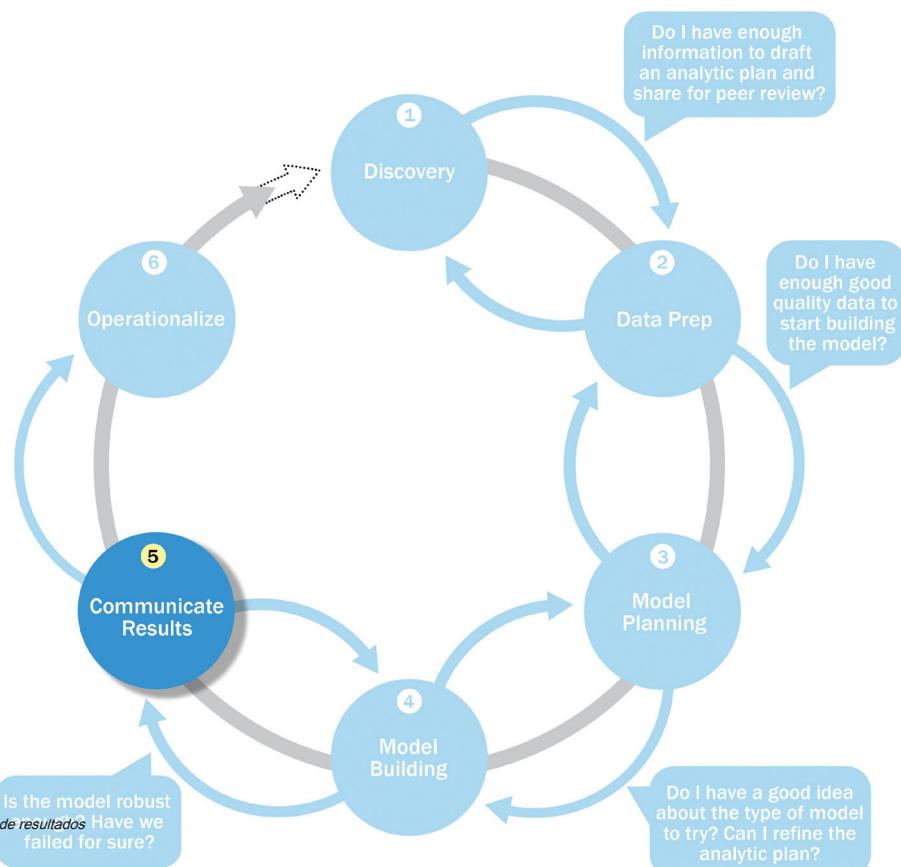
Hay muchas herramientas disponibles para ayudar en esta fase, enfocadas principalmente en análisis estadístico o software de minería de datos. Las herramientas comunes en este espacio incluyen, pero no se limitan a, las siguientes:

- Herramientas comerciales:
  - **SAS Enterprise Miner** [ 17] permite a los usuarios ejecutar modelos predictivos y descriptivos basados en grandes volúmenes de datos de toda la empresa. Interopera con otros grandes almacenes de datos, tiene muchas asociaciones y está diseñado para la informática y el análisis de nivel empresarial.
  - **Modelador de SPSS** [ 18] (proporcionado por IBM y ahora llamado IBM SPSS Modeler) ofrece métodos para explorar y analizar datos a través de una GUI.
  - **Matlab** [ 19] proporciona un lenguaje de alto nivel para realizar una variedad de análisis de datos, algoritmos ritmos y exploración de datos.
  - **Minero alpino** [ 11] proporciona una interfaz gráfica de usuario para que los usuarios desarrollen flujos de trabajo analíticos e interactúen con herramientas y plataformas de Big Data en el back-end.
  - **ESTADÍSTICA** [ 20] y **Mathematica** [ 21] también son minería de datos populares y bien considerados y herramientas analíticas.
- Herramientas gratuitas o de código abierto:
  - **R y PL / R** [ 14] R se describió anteriormente en la fase de planificación del modelo, y PL / R es un procedimiento lenguaje para PostgreSQL con R. Usar este enfoque significa que los comandos de R se pueden ejecutar en la base de datos. Esta técnica proporciona un mayor rendimiento y es más escalable que ejecutar R en la memoria.
  - **Octava** [ 22], un lenguaje de programación de software libre para modelado computacional, tiene algunos de la funcionalidad de Matlab. Debido a que está disponible gratuitamente, Octave se utiliza en las principales universidades para enseñar aprendizaje automático.
  - **WEKA** [ 23] es un paquete de software de minería de datos gratuito con un banco de trabajo analítico. Las funciones creadas en WEKA se pueden ejecutar dentro del código Java.
  - **Pitón** es un lenguaje de programación que proporciona conjuntos de herramientas para el análisis y el aprendizaje automático, como scikit-learn, numpy, scipy, pandas y visualización de datos relacionados con matplotlib.
  - **SQL implementaciones** en la base de datos, como **MADlib** [ 24], proporcionan una alternativa a la memoria herramientas analíticas de escritorio. MADlib proporciona una biblioteca de algoritmos de aprendizaje automático de código abierto que se pueden ejecutar en la base de datos, para PostgreSQL o Greenplum.

## 2.6 Fase 5: Comunicar resultados

Después de ejecutar el modelo, el equipo debe comparar los resultados del modelo con los criterios establecidos para el éxito y el fracaso. En la Fase 5, que se muestra en la Figura 2-7, el equipo considera la mejor manera de articular los hallazgos y resultados a los diversos miembros del equipo y partes interesadas, teniendo en cuenta las advertencias, los supuestos y las limitaciones de los resultados. Debido a que la presentación a menudo circula dentro de una organización,

priate para la au



Como parte de la Fase 5, el equipo debe determinar si tuvo éxito o fracasó en sus objetivos. Muchas veces la gente no quiere admitir que ha fallado, pero en este caso el fracaso no debe considerarse como un verdadero fracaso, sino más bien como una falla de los datos para aceptar o rechazar una hipótesis dada de manera adecuada. Este concepto puede ser contrario a la intuición para aquellos a quienes se les ha dicho que no fracsasen en toda su carrera. Sin embargo, la clave es

recordar que el equipo debe ser lo suficientemente riguroso con los datos para determinar si probará o refutará las hipótesis descritas en la Fase 1 (descubrimiento). A veces, los equipos solo han realizado un análisis superficial, que no es lo suficientemente sólido como para aceptar o rechazar una hipótesis. Otras veces, los equipos realizan análisis muy sólidos y buscan formas de mostrar resultados, incluso cuando los resultados pueden no estar ahí. Es importante lograr un equilibrio entre estos dos extremos cuando se trata de analizar datos y ser pragmático en términos de mostrar resultados del mundo real.

Al realizar esta evaluación, determine si los resultados son estadísticamente significativos y válidos. Si es así, identifique los aspectos de los resultados que se destacan y pueden proporcionar hallazgos destacados cuando llegue el momento de comunicarlos. Si los resultados no son válidos, piense en los ajustes que se pueden hacer para refinarse e iterar en el modelo para que sea válido. Durante este paso, evalúe los resultados e identifique qué puntos de datos pueden haber sido sorprendentes y cuáles estaban en línea con las hipótesis que se desarrollaron en la Fase 1. La comparación de los resultados reales con las ideas formuladas desde el principio produce ideas y conocimientos adicionales que se habrían perdido, si el equipo no se hubiera tomado el tiempo para formular hipótesis iniciales al principio del proceso.

Para entonces, el equipo debería haber determinado qué modelo o modelos abordan el desafío analítico de la manera más adecuada. Además, el equipo debe tener ideas de algunos de los hallazgos como resultado del proyecto. La mejor práctica en esta fase es registrar todos los hallazgos y luego seleccionar los tres más significativos que se pueden compartir con las partes interesadas. Además, el equipo debe reflexionar sobre las implicaciones de estos hallazgos y medir el valor comercial. Dependiendo de lo que surgió como resultado del modelo, el equipo puede necesitar dedicar tiempo a cuantificar el impacto comercial de los resultados para ayudar a prepararse para la presentación y demostrar el valor de los hallazgos. El trabajo de Doug Hubbard [6] ofrece ideas sobre cómo evaluar intangibles en los negocios y cuantificar el valor de cosas aparentemente incommensurables.

Ahora que el equipo ha ejecutado el modelo, ha completado una fase de descubrimiento exhaustiva y ha aprendido mucho sobre los conjuntos de datos, reflexione sobre el proyecto y considere qué obstáculos había en el proyecto y qué se puede mejorar en el futuro. Haga recomendaciones para el trabajo futuro o las mejoras a los procesos existentes y considere lo que cada uno de los miembros del equipo y las partes interesadas necesita para cumplir con sus responsabilidades. Por ejemplo, los patrocinadores deben defender el proyecto. Las partes interesadas deben comprender cómo el modelo afecta sus procesos. (Por ejemplo, si el equipo ha creado un modelo para predecir la rotación de clientes, el equipo de marketing debe comprender cómo utilizar las predicciones del modelo de rotación para planificar sus intervenciones). Los ingenieros de producción deben poner en práctica el trabajo que se ha realizado. Adicionalmente,

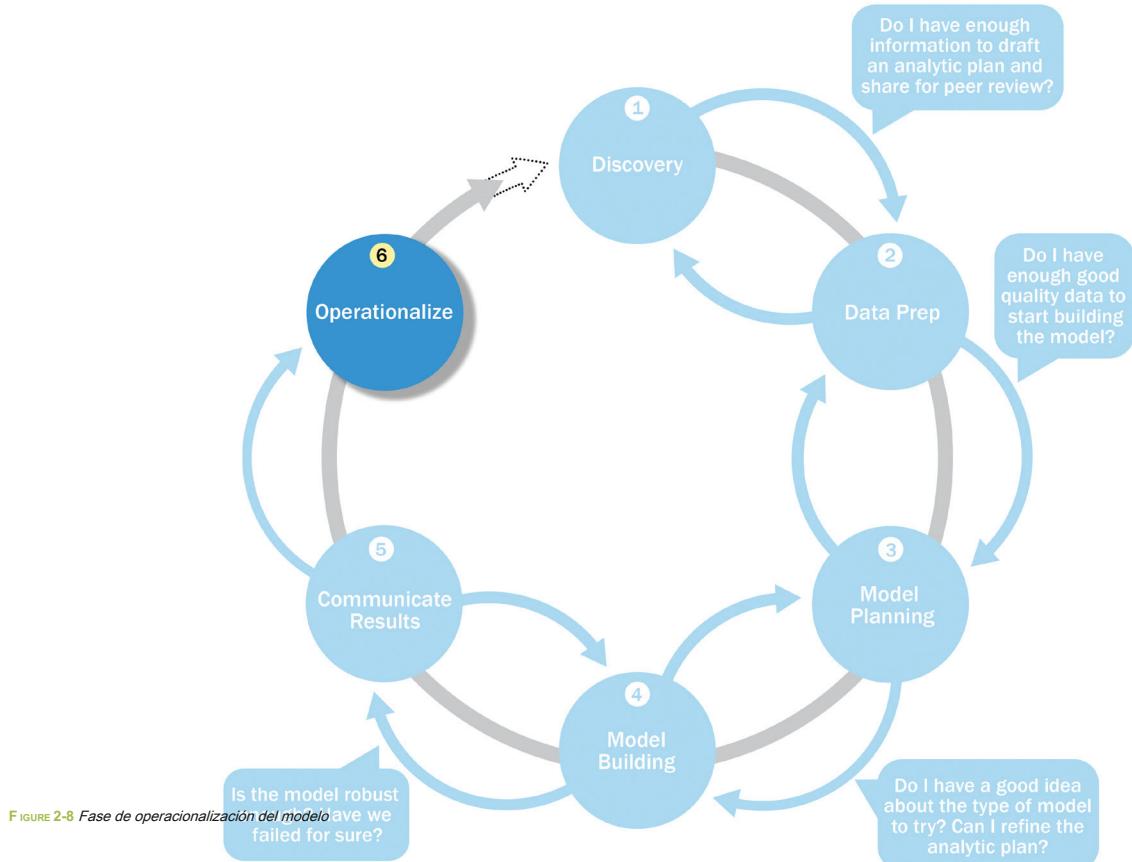
Como resultado de esta fase, el equipo habrá documentado los hallazgos clave y los principales conocimientos derivados del análisis. El producto final de esta fase será la parte más visible del proceso para las partes interesadas y los patrocinadores externos, por lo que debe tener cuidado de articular claramente los resultados, la metodología y el valor comercial de los hallazgos. Se proporcionarán más detalles sobre las herramientas de visualización de datos y las referencias en el Capítulo 12, "El final del juego, o ponerlo todo junto".

## 2.7 Fase 6: Operacionalización

En la fase final, el equipo comunica los beneficios del proyecto de manera más amplia y establece un proyecto piloto para implementar el trabajo de forma controlada antes de ampliar el trabajo a una empresa completa o ecosistema de usuarios. En la Fase 4, el equipo calificó el modelo en el entorno limitado de análisis. La fase 6, que se muestra en la Figura 2-8, representa la primera vez que la mayoría de los equipos de análisis abordan la implementación de los nuevos métodos o modelos analíticos en un entorno de producción. En lugar de implementar estos modelos inmediatamente a gran escala

base, el riesgo se puede gestionar de manera más eficaz y el equipo puede aprender mediante la realización de una implementación piloto de pequeño alcance antes de una implementación a gran escala. Este enfoque permite al equipo aprender sobre el rendimiento y las limitaciones relacionadas del modelo en un entorno de producción a pequeña escala y realizar ajustes antes de una implementación completa. Durante el proyecto piloto, es posible que el equipo deba considerar la ejecución del algoritmo en la base de datos.

más e ffi ciente t



Al determinar el alcance del esfuerzo involucrado en la realización de un proyecto piloto, considere ejecutar el modelo en un entorno de producción para un conjunto discreto de productos o una sola línea de negocios, que prueba el modelo en un entorno en vivo. Esto permite que el equipo aprenda de la implementación y realice los ajustes necesarios antes de lanzar el modelo en toda la empresa. Tenga en cuenta que esta fase puede traer un nuevo grupo de miembros del equipo, generalmente los ingenieros responsables del entorno de producción que tienen un nuevo grupo de problemas e inquietudes más allá de los del equipo central del proyecto. Este grupo técnico debe asegurarse de que

la ejecución del modelo se adapta sin problemas al entorno de producción y que el modelo se puede integrar en los procesos comerciales relacionados.

Parte de la fase de puesta en funcionamiento incluye la creación de un mecanismo para realizar un seguimiento continuo de la precisión del modelo y, si la precisión se degrada, encontrar formas de reentrenar el modelo. Si es posible, diseñe alertas para cuando el modelo esté operando "fuera de los límites". Esto incluye situaciones en las que las entradas están más allá del rango en el que se entrenó el modelo, lo que puede provocar que las salidas del modelo sean inexactas o no válidas. Si esto comienza a suceder con regularidad, el modelo debe volver a capacitarse con nuevos datos.

A menudo, los proyectos analíticos producen nuevos conocimientos sobre una empresa, un problema o una idea que la gente puede haber tomado al pie de la letra o que pensaba que era imposible de explorar. Se pueden crear cuatro entregables principales para satisfacer las necesidades de la mayoría de las partes interesadas. Este enfoque para desarrollar los cuatro entregables se analiza en mayor detalle en el Capítulo 12.

La Figura 2-9 muestra los resultados clave para cada uno de los principales interesados de un proyecto de análisis y lo que normalmente esperan al final de un proyecto.

- **Usuario comercial** normalmente trata de determinar los beneficios y las implicaciones de los hallazgos para la empresa.
- **Patrocinador de proyecto** Por lo general, hace preguntas relacionadas con el impacto comercial del proyecto, los riesgos y el retorno de la inversión (ROI) y la forma en que se puede evangelizar el proyecto dentro de la organización (y más allá).
- **Gerente de proyecto** necesita determinar si el proyecto se completó a tiempo y dentro del presupuesto y qué tan bien se cumplieron las metas.
- **Analista de inteligencia empresarial** necesita saber si los informes y paneles que administra se verán afectados y deben cambiar.
- **Ingeniero de datos y Administrador de base de datos (DBA)** normalmente necesitan compartir su código del proyecto de análisis y crear un documento técnico sobre cómo implementarlo.
- **Científico de datos** necesita compartir el código y explicar el modelo a sus compañeros, gerentes y otras partes interesadas.

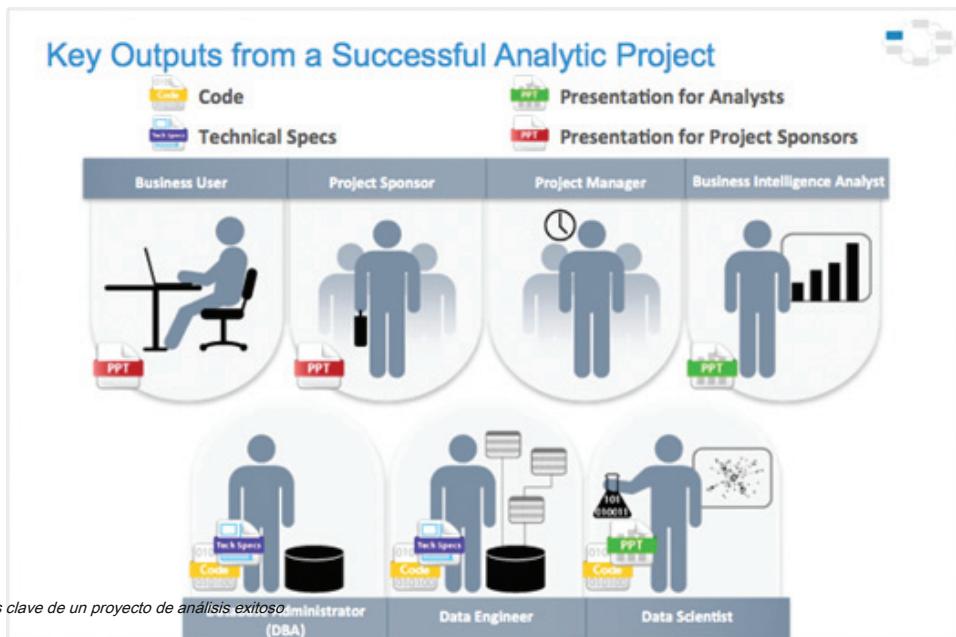
Aunque estos siete roles representan muchos intereses dentro de un proyecto, estos intereses generalmente se superponen y la mayoría de ellos se pueden cumplir con cuatro entregables principales.

- Presentación para los patrocinadores del proyecto: contiene conclusiones de alto nivel para las partes interesadas del nivel ejecutivo, con algunos mensajes clave para ayudar en el proceso de toma de decisiones. Concéntrese en imágenes claras y fáciles para que el presentador las explique y el espectador las comprenda.
- Presentación para analistas, que describe cambios en los procesos de negocio e informes de cambios. Los compañeros científicos de datos querrán conocer los detalles y se sentirán cómodos con los gráficos técnicos (como las curvas de Característica de funcionamiento del receptor [ROC], diagramas de densidad e histogramas que se muestran en el Capítulo 3 y el Capítulo 7).
- Código para personal técnico.
- Especificaciones técnicas de implementación del código.

Como regla general, cuanto más ejecutiva sea la audiencia, más sucinta debe ser la presentación. La mayoría de los patrocinadores ejecutivos asisten a muchas reuniones informativas en el transcurso de un día o una semana. Asegúrese de que la presentación vaya al grano rápidamente y enmarque los resultados en términos de valor para la organización del patrocinador. Por ejemplo, si el equipo está trabajando con un banco para analizar casos de fraude con tarjetas de crédito, resalte la frecuencia del fraude, el número de casos en el último mes o año y el impacto en los costos o ingresos para el banco.

(o enfóquese en lo contrario: cuántos más ingresos podría obtener el banco si aborda el problema del fraude). Esto demuestra el impacto empresarial mejor que una inmersión profunda en la metodología. La presentación debe incluir información de apoyo sobre la metodología analítica y las fuentes de datos, pero generalmente solo como su

analizar th



**FIGURE 2-9** Resultados clave de un proyecto de análisis exitoso

Cuando se presente a otras audiencias con antecedentes más cuantitativos, dedique más tiempo a la metodología y los hallazgos. En estos casos, el equipo puede ser más expansivo al describir los resultados, la metodología y el experimento analítico con un grupo de pares. Esta audiencia estará más interesada en las técnicas, especialmente si el equipo desarrolló una nueva forma de procesar o analizar datos que pueda reutilizarse en el futuro o aplicarse a problemas similares. Además, utilice imágenes o visualización de datos cuando sea posible. Aunque puede llevar más tiempo desarrollar imágenes, la gente tiende a recordar imágenes para demostrar un punto más que largas listas de viñetas [25]. La visualización y las presentaciones de datos se analizan con más detalle en el Capítulo 12.

## 2.8 Estudio de caso: Red y análisis de innovación global (GINA)

El equipo Global Innovation Network and Analytics (GINA) de EMC es un grupo de tecnólogos senior ubicados en centros de excelencia (COE) de todo el mundo. El estatuto de este equipo es involucrar a los empleados de los COE globales para impulsar la innovación, la investigación y las asociaciones universitarias. En 2012, un director recién contratado quiso

mejorar estas actividades y proporcionar un mecanismo para rastrear y analizar la información relacionada. Además, este equipo deseaba crear mecanismos más sólidos para capturar los resultados de sus conversaciones informales con otros líderes de opinión dentro de EMC, en el ámbito académico o en otras organizaciones, que luego podrían ser buscadas para obtener información.

El equipo de GINA pensó que su enfoque proporcionaría un medio para compartir ideas a nivel mundial y aumentar el intercambio de conocimientos entre los miembros de GINA que pueden estar separados geográficamente. Planeó crear un repositorio de datos que contenga datos estructurados y no estructurados para lograr tres objetivos principales.

- Almacene datos formales e informales.
- Seguimiento de la investigación de tecnólogos globales.
- Extraiga los datos en busca de patrones y conocimientos para mejorar las operaciones y la estrategia del equipo.

El caso de estudio de GINA proporciona un ejemplo de cómo un equipo aplicó el ciclo de vida de análisis de datos para analizar datos de innovación en EMC. La innovación suele ser un concepto difícil de medir, y este equipo quería buscar formas de utilizar métodos analíticos avanzados para identificar innovadores clave dentro de la empresa.

### 2.8.1 Fase 1: Descubrimiento

En la fase de descubrimiento del proyecto GINA, el equipo comenzó a identificar fuentes de datos. Aunque GINA era un grupo de tecnólogos capacitados en muchos aspectos diferentes de la ingeniería, tenía algunos datos e ideas sobre lo que quería explorar, pero carecía de un equipo formal que pudiera realizar estos análisis. Después de consultar con varios expertos, incluido Tom Davenport, un destacado experto en análisis de Babson College, y Peter Gloor, un experto en inteligencia colectiva y creador de CoIN (Redes de innovación colaborativa) en el MIT, el equipo decidió colaborar con el trabajo buscando voluntarios dentro de EMC.

A continuación se muestra una lista de cómo se cumplieron los distintos roles del equipo de trabajo.

- **Usuario comercial, patrocinador del proyecto, ProjectManager:** Vicepresidente de la Oficina de la CTO
- **Analista de inteligencia empresarial:** Representantes de TI
- **Ingeniero de datos y Administrador de base de datos (DBA):** Representantes de TI
- **Científico de datos:** Ingeniero distinguido, que también desarrolló los gráficos sociales que se muestran en el estudio de caso de GINA

El enfoque del patrocinador del proyecto fue aprovechar las redes sociales y los blogs [26] para acelerar la recopilación de datos de investigación e innovación en todo el mundo y motivar a los equipos de científicos de datos "voluntarios" en ubicaciones de todo el mundo. Dado que carecía de un equipo formal, necesitaba ser ingenioso para encontrar personas capaces y dispuestas a ofrecer voluntariamente su tiempo para trabajar en problemas interesantes. Los científicos de datos tienden a ser apasionados por los datos, y el patrocinador del proyecto pudo aprovechar esta pasión de personas altamente talentosas para realizar un trabajo desafianta de una manera creativa.

Los datos del proyecto se dividieron en dos categorías principales. La primera categoría representó cinco años de presentación de ideas de los concursos de innovación internos de EMC, conocidos como InnovationRoadmap (anteriormente llamado Innovation Showcase). InnovationRoadmap es un proceso de innovación formal y orgánico mediante el cual empleados de todo el mundo envían ideas que luego son examinadas y evaluadas. Se seleccionan las mejores ideas para su posterior incubación. Como resultado, los datos son una mezcla de datos estructurados, como recuentos de ideas, fechas de envío, nombres de inventores y contenido no estructurado, como las descripciones textuales de las propias ideas.

La segunda categoría de datos abarcó actas y notas que representan la actividad de innovación e investigación de todo el mundo. Esto también representó una combinación de datos estructurados y no estructurados. Los datos estructurados incluían atributos como fechas, nombres y ubicaciones geográficas. Los documentos no estructurados contenían información de "quién, qué, cuándo y dónde" que representa datos valiosos sobre el crecimiento y la transferencia del conocimiento dentro de la empresa. Este tipo de información a menudo se almacena en silos comerciales que tienen poca o ninguna visibilidad entre equipos de investigación dispares.

Los 10 principales IH que desarrolló el equipo de GINA fueron los siguientes:

- **IH1:** La actividad de innovación en diferentes regiones geográficas se puede asignar a direcciones estratégicas corporativas.
- **IH2:** El tiempo que lleva entregar ideas disminuye cuando la transferencia de conocimiento global ocurre como parte del proceso de entrega de ideas.
- **IH3:** Los innovadores que participan en la transferencia global de conocimientos aportan ideas con mayor rapidez que los que no lo hacen.
- **IH4:** La presentación de una idea se puede analizar y evaluar para determinar la probabilidad de recibir financiación.
- **IH5:** El descubrimiento y el crecimiento del conocimiento para un tema en particular se pueden medir y comparar entre regiones geográficas.
- **IH6:** La actividad de transferencia de conocimiento puede identificar traspasos de fronteras específicos de investigación en regiones dispares.
- **IH7:** Los temas corporativos estratégicos se pueden asignar a regiones geográficas.
- **IH8:** Los eventos frecuentes de expansión y transferencia de conocimiento reducen el tiempo que lleva generar un activo corporativo a partir de una idea.
- **IH9:** Los mapas de línea pueden revelar cuándo la expansión y transferencia de conocimiento no resultó (o no lo ha hecho) en un activo corporativo.
- **IH10:** Los temas de investigación emergentes pueden clasificarse y asignarse a ideadores, innovadores, traspasadores de fronteras y activos específicos.

Los GINA (IH) se pueden agrupar en dos categorías:

- Análisis descriptivo de lo que está sucediendo actualmente para generar más creatividad, colaboración y generación de activos.
- Análisis predictivo para asesorar a la dirección ejecutiva sobre dónde debería invertir en el futuro.

## 2.8.2 Fase 2: Preparación de datos

El equipo se asoció con su departamento de TI para configurar una nueva caja de arena de análisis para almacenar y experimentar con los datos.

Durante el ejercicio de exploración de datos, los científicos e ingenieros de datos comenzaron a notar que ciertos datos necesitaban acondicionamiento y normalización. Además, el equipo se dio cuenta de que varios conjuntos de datos faltantes eran fundamentales para probar algunas de las hipótesis analíticas.

A medida que el equipo exploraba los datos, rápidamente se dio cuenta de que si no tenía datos de calidad suficiente o no podía obtener datos de buena calidad, no podría realizar los pasos subsiguientes en el proceso del ciclo de vida. Como resultado, era importante determinar qué nivel de calidad y limpieza de los datos era suficiente para el

proyecto en curso. En el caso de la GINA, el equipo descubrió que muchos de los nombres de los investigadores y de las personas que interactúan con las universidades estaban mal escritos o tenían espacios iniciales y finales en el almacén de datos. Problemas aparentemente pequeños como estos en los datos tuvieron que abordarse en esta fase para permitir un mejor análisis y agregación de datos en las fases posteriores.

### 2.8.3 Fase 3: Planificación del modelo

En el proyecto GINA, para gran parte del conjunto de datos, parecía factible utilizar técnicas de análisis de redes sociales para observar las redes de innovadores dentro de EMC. En otros casos, fue difícil encontrar formas apropiadas de probar hipótesis debido a la falta de datos. En un caso (IH9), el equipo tomó la decisión de iniciar un estudio longitudinal para comenzar a rastrear puntos de datos a lo largo del tiempo sobre personas que desarrollan nueva propiedad intelectual. Esta recopilación de datos permitiría al equipo probar las siguientes dos ideas en el futuro:

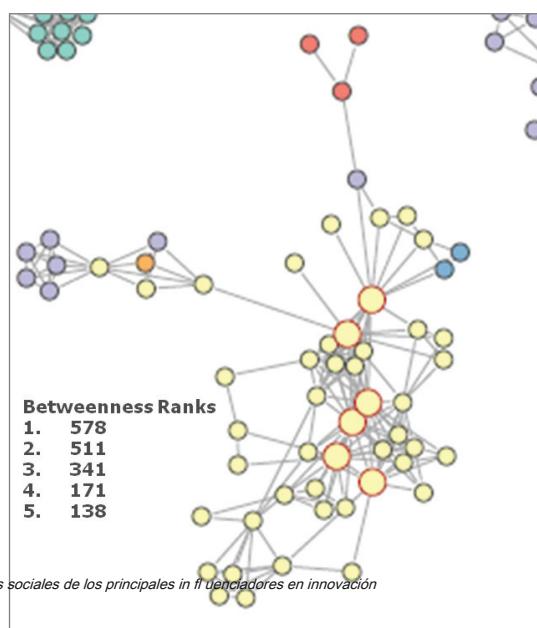
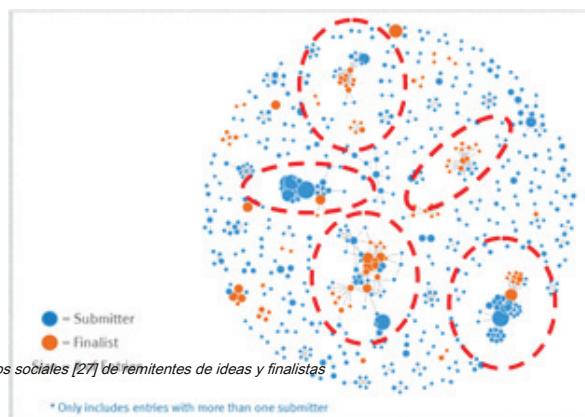
- **IH8:** Los eventos frecuentes de expansión y transferencia de conocimiento reducen la cantidad de tiempo que lleva generar un activo corporativo a partir de una idea.
- **IH9:** Los mapas de línea pueden revelar cuándo la expansión y la transferencia de conocimiento no dieron como resultado (o no) un activo corporativo.

Para el estudio longitudinal propuesto, el equipo necesitaba establecer criterios de objetivos para el estudio. Específicamente, necesitaba determinar el objetivo final de una idea exitosa que había atravesado todo el viaje. Los parámetros relacionados con el alcance del estudio incluyeron las siguientes consideraciones:

- Identificar los hitos correctos para lograr este objetivo.
- Rastree cómo las personas mueven ideas desde cada milla hacia la meta.
- Una vez hecho esto, rastree las ideas que mueren y rastree otras que alcancen la meta. Compare los viajes de las ideas que lo hacen y los que no.
- Compare los tiempos y los resultados utilizando algunos métodos diferentes (dependiendo de cómo se recopilen y recopilen los datos). Estos podrían ser tan simples como pruebas o quizás involucrar diferentes tipos de algoritmos de clasificación.

### 2.8.4 Fase 4: Construcción de modelos

En la Fase 4, el equipo de GINA empleó varios métodos analíticos. Esto incluyó el trabajo del científico de datos que utilizó técnicas de procesamiento del lenguaje natural (PNL) en las descripciones textuales de las ideas de la hoja de ruta de innovación. Además, realizó análisis de redes sociales usando R y RStudio, y luego desarrolló gráficos sociales y visualizaciones de la red de comunicaciones relacionadas con la innovación usando R's. ggplot2 paquete. En las Figuras 2-10 y 2-11 se muestran ejemplos de este trabajo.



La figura 2-10 muestra gráficos sociales que muestran las relaciones entre los emisores de ideas dentro de GINA. Cada color representa a un innovador de un país diferente. Los puntos grandes con círculos rojos alrededor de ellos representan ejes. UN **cubo** representa una persona con alta conectividad y una alta puntuación de "intermediación". El grupo de la Figura 2-11 contiene variedad geográfica, que es fundamental para probar la hipótesis sobre los traspasos de límites geográficos. Una persona en este gráfico tiene una puntuación inusualmente alta en comparación con el resto de los nodos del gráfico. El científico de datos identificó a esta persona y realizó una consulta con su nombre dentro de la caja de arena analítica. Estas acciones arrojaron la siguiente información sobre este científico investigador (del gráfico social), que ilustró cuán influyente era dentro de su unidad de negocio y en muchas otras áreas de la empresa en todo el mundo:

- En 2011, asistió a la conferencia ACM SIGMOD, que es una conferencia de primer nivel sobre problemas de gestión de datos y bases de datos a gran escala.
- Visitó a empleados en Francia que forman parte de la unidad de negocios de los equipos de administración de contenido de EMC dentro de Documentum (ahora parte del Information Intelligence Group, o IIG).
- Presentó sus pensamientos sobre la conferencia SIGMOD en una sesión virtual a la que asistieron tres empleados en Rusia, un empleado en El Cairo, un empleado en Irlanda, un empleado en India, tres empleados en los Estados Unidos y un empleado en Israel.
- En 2012, asistió a la conferencia SDM2012 en California.
- En el mismo viaje, visitó a innovadores e investigadores de empresas federadas de EMC, Pivotal y VMware.
- Más tarde, en ese viaje, se presentó ante un consejo interno de líderes tecnológicos y presentó a dos de sus investigadores a decenas de innovadores e investigadores corporativos.

Este hallazgo sugiere que al menos parte de la hipótesis inicial es correcta; los datos pueden identificar a los innovadores que abarcan distintas geografías y unidades de negocio. El equipo usó el software Tableau para la visualización y exploración de datos y utilizó la base de datos Pivotal Greenplum como el repositorio principal y el motor de análisis.

## **2.8.5 Fase 5: Comunicar resultados**

En la Fase 5, el equipo encontró varias formas de seleccionar los resultados del análisis e identificar los hallazgos más impactantes y relevantes. Este proyecto se consideró exitoso en la identificación de personas que traspasaron fronteras e innovadores ocultos. Como resultado, el CTO de EMC lanzó estudios longitudinales para comenzar los esfuerzos de recolección de datos y rastrear los resultados de la innovación durante períodos de tiempo más largos. El proyecto GINA promovió el intercambio de conocimientos relacionados con la innovación y los investigadores que abarcan múltiples áreas dentro y fuera de la empresa. GINA también permitió a EMC cultivar propiedad intelectual adicional que condujo a temas de investigación adicionales y brindó oportunidades para forjar relaciones con universidades para la investigación académica conjunta en los campos de Data Science y Big Data. Además, el proyecto se logró con un presupuesto limitado,

Uno de los hallazgos clave del proyecto es que había una densidad desproporcionadamente alta de innovadores en Cork, Irlanda. Cada año, EMC organiza un concurso de innovación, abierto a los empleados para que presenten ideas de innovación que generen un nuevo valor para la empresa. Al observar los datos de 2011, el 15% de los finalistas y el 15% de los ganadores eran de Irlanda. Estos son números inusualmente altos, dado el tamaño relativo del COE de Cork en comparación con otros centros más grandes en otras partes del mundo. Después de una mayor investigación, se supo que el COE en Cork, Irlanda había recibido capacitación enfocada en innovación por parte de un consultor externo, que

estaba resultando eficaz. Al Cork COE se le ocurrieron más ideas de innovación y mejoras que en el pasado, y estaba haciendo mayores contribuciones a la innovación en EMC. Habría sido difícil, si no imposible, identificar este grupo de innovadores a través de métodos tradicionales o incluso comentarios anecdóticos de boca en boca. La aplicación del análisis de redes sociales permitió al equipo encontrar un grupo de personas dentro de EMC que estaban haciendo contribuciones desproporcionadamente fuertes. Estos hallazgos se compartieron internamente a través de presentaciones y conferencias y se promovieron a través de redes sociales y blogs.

### 2.8.6 Fase 6: Operacionalización

La ejecución de análisis en un espacio aislado lleno de notas, actas y presentaciones de las actividades de innovación arrojó grandes conocimientos sobre la cultura de innovación de EMC. Los hallazgos clave del proyecto incluyen los siguientes:

- El CTOo ffi ce y GINA necesitan más datos en el futuro, incluida una iniciativa de marketing para convencer a las personas de que informen a la comunidad global sobre sus actividades de innovación / investigación.
- Algunos de los datos son confidenciales y el equipo debe considerar la seguridad y la privacidad relacionadas con los datos, como quién puede ejecutar los modelos y ver los resultados.
- Además de ejecutar modelos, es necesario crear una iniciativa paralela para mejorar las actividades básicas de inteligencia empresarial, como paneles, informes y consultas sobre actividades de investigación en todo el mundo.
- Se necesita un mecanismo para reevaluar continuamente el modelo después del despliegue. Evaluar los beneficios es uno de los principales objetivos de esta etapa, así como definir un proceso para volver a capacitar al modelo según sea necesario.

Además de las acciones y los hallazgos enumerados, el equipo demostró cómo la analítica puede generar nuevos conocimientos en proyectos que tradicionalmente son difíciles de medir y cuantificar. Este proyecto informó las decisiones de inversión en proyectos de investigación universitaria de la oficina de CTO e identificó innovadores ocultos de alto valor. Además, la oficina del CTO desarrolló herramientas para ayudar a los remitentes a mejorar las ideas utilizando modelos de temas como parte de los nuevos sistemas de recomendación para ayudar a los remitentes de ideas a encontrar ideas similares y refinar sus propuestas de nueva propiedad intelectual.

La Tabla 2-3 describe un plan de análisis para el ejemplo del estudio de caso de GINA. Aunque este proyecto muestra sólo tres hallazgos, hubo muchos más. Por ejemplo, quizás el mayor resultado global de este proyecto es que demostró, de manera concreta, que la analítica puede generar nuevos conocimientos en proyectos que tratan temas que pueden parecer difíciles de medir, como la innovación.

#### T PODER 2-3 Plan analítico del proyecto EMC GINA

<b>Descubrimiento de Negocio</b>	Seguimiento del crecimiento del conocimiento global, asegurando una transferencia de conocimiento efectiva y convirtiéndolo rápidamente en activos corporativos. La ejecución de estos tres elementos debería acelerar la innovación.
<b>Problema enunciado</b>	
<b>Hipótesis iniciales</b>	Un aumento en la transferencia de conocimiento geográfico mejora la velocidad de entrega de ideas.
<b>Datos</b>	Cinco años de presentación e historia de ideas de innovación; Seis meses de notas textuales de actividades de investigación e innovación global.

( continúa)

**T PODER 2-3 Plan analítico del proyecto EMC GINA (continuación)**

Planificación	Técnica analítica
<b>Resultado y hallazgos clave</b>	<p>1. Identificó a innovadores ocultos de gran valor y formas de compartir sus conocimientos.</p> <p>2. Decisiones de inversión fundamentadas en proyectos de investigación universitaria</p> <p>3. Creación de herramientas para ayudar a los remitentes a mejorar las ideas con sistemas de recomendación de ideas.</p>

La innovación es una idea que toda empresa desea promover, pero puede ser difícil medir la innovación o identificar formas de incrementar la innovación. ~~Este proyecto exploró este tema desde el punto de vista de la evaluación de las redes sociales informales para identificar personas que traspasan fronteras y personas influyentes dentro de las subredes de innovación. En esencia, este proyecto tomó un problema aparentemente nebuloso y aplicó métodos analíticos avanzados para encontrar respuestas utilizando un enfoque objetivo y basado en hechos.~~

Otro resultado del proyecto incluyó la necesidad de complementar la analítica con un almacén de datos separado para informes de Business Intelligence, accesible para buscar iniciativas de innovación / investigación. Además de apoyar la toma de decisiones, esto proporcionará un mecanismo para estar informado sobre las discusiones y las investigaciones que se llevan a cabo en todo el mundo entre los miembros del equipo en ubicaciones dispares. Finalmente, destacó el valor que se puede extraer a través de los datos y análisis posteriores. Por lo tanto, se identificó la necesidad de iniciar programas formales de marketing para convencer a las personas de que presenten (o informen) a la comunidad global sobre sus actividades de innovación / investigación. El intercambio de conocimientos fue fundamental. Sin él, GINA no habría podido realizar el análisis e identificar a los innovadores ocultos dentro de la empresa.

## Resumen

Este capítulo describe el ciclo de vida de análisis de datos, que es un enfoque para administrar y ejecutar proyectos analíticos. Este enfoque describe el proceso en seis fases.

1. Descubrimiento
2. Preparación de datos
3. Planificación de modelos
4. Construcción del modelo
5. Comunicar resultados
6. Operacionalizar

A través de estos pasos, los equipos de ciencia de datos pueden identificar problemas y realizar una investigación rigurosa de los conjuntos de datos necesarios para un análisis en profundidad. Como se indica en el capítulo, aunque se ha escrito mucho sobre los métodos analíticos, la mayor parte del tiempo dedicado a este tipo de proyectos se dedica a la preparación, es decir,

en las Fases 1 y 2 (descubrimiento y preparación de datos). Además, este capítulo analizó los siete roles necesarios para un equipo de ciencia de datos. Es fundamental que las organizaciones reconozcan que la ciencia de datos es un equipo y que se necesita un equilibrio de habilidades para tener éxito al abordar proyectos de Big Data y otros proyectos complejos que involucran análisis de datos.

## Ejercicios

1. ¿En qué fase esperaría el equipo invertir la mayor parte del tiempo del proyecto? ¿Por qué? ¿Dónde estaría el equipo esperando pasar el menor tiempo posible?
2. ¿Cuáles son los beneficios de realizar un programa piloto antes de la implementación a gran escala de un nuevo método analítico? Discuta esto en el contexto del mini estudio de caso.
3. ¿Qué tipos de herramientas se utilizarían en las siguientes fases y para qué tipos de escenarios de uso?
  - a. Fase 2: preparación de datos

**segundo.** Fase 4: construcción de modelos

## Bibliografía

- [1] TH Davenport y DJ Patil, "Científico de datos: el trabajo más sexy del siglo XXI", *Harvard Revision del negocio*, Octubre 2012.
- [2] J. Manyika, M. Chiu, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh y AH Byers, "Big Data: The Next Frontier for Innovation, Competition, and Productivity", McKinsey Global Institute, 2011.
- [3] "Método científico" [en línea]. Disponible: [http://en.wikipedia.org/wiki/M%C3%A9todo\\_cient%C3%ADfico](http://en.wikipedia.org/wiki/M%C3%A9todo_cient%C3%ADfico).
- [4] "CRISP-DM" [en línea]. Disponible: [http://en.wikipedia.org/wiki/Cross\\_Industry\\_Standard\\_Process\\_for\\_Data\\_Mining](http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining).
- [5] TH Davenport, JG Harris y R. Morison, *Analytics en el trabajo: decisiones más inteligentes, mejores resultados*, 2010, Harvard Business Review Press.
- [6] DW Hubbard, *Cómo medir cualquier cosa: encontrar el valor de los intangibles en los negocios*, 2010, Hoboken, Nueva Jersey: John Wiley & Sons.
- [7] J. Cohen, B. Dolan, M. Dunlap, JM Hellerstein y C. Welton, *Habilidades MAD: nuevas prácticas de análisis para Big Data*, Watertown, MA 2009.
- [8] "Lista de API" [en línea]. Disponible: <http://www.programmableweb.com/apis>.
- [9] B. Shneiderman [en línea]. Disponible: <http://www.ifp.illinois.edu/nabchs/abstracts/shneiderman.html>.
- [10] "Hadoop" [en línea]. Disponible: <http://hadoop.apache.org>.
- [11] "Alpine Miner" [en línea]. Disponible: <http://alpinenow.com>.
- [12] "OpenRefine" [en línea]. Disponible: <http://openrefine.org>.
- [13] "Data Wrangler" [en línea]. Disponible: <http://vis.stanford.edu/wrangler/>.
- [14] "CRAN" [en línea]. Disponible: <http://cran.us.r-project.org>.
- [15] "SQL" [en línea]. Disponible: <http://en.wikipedia.org/wiki/SQL>.
- [diecisésis] "SAS / ACCESS" [en línea]. Disponible: [http://www.sas.com/en\\_us/software/data-management/access.htm](http://www.sas.com/en_us/software/data-management/access.htm).

- [17] "SAS Enterprise Miner" [en línea]. Disponible: [http://www.sas.com/en\\_us/software/analytics/enterprise-miner.html](http://www.sas.com/en_us/software/analytics/enterprise-miner.html).
- [18] "SPSS Modeler" [en línea]. Disponible: <http://www-03.ibm.com/software/products/es/categoría/analítica-empresarial>.
- [19] "Matlab" [en línea]. Disponible: <http://www.mathworks.com/products/matlab/>.
- [20] "Statistica" [en línea]. Disponible: <https://www.statsoft.com>.
- [21] "Mathematica" [en línea]. Disponible: <http://www.wolfram.com/mathematica/>.
- [22] "Octave" [en línea]. Disponible: <https://www.gnu.org/software/octave/>.
- [23] "WEKA" [en línea]. Disponible: <http://www.cs.waikato.ac.nz/ml/weka/>.
- [24] "MADlib" [en línea]. Disponible: <http://madlib.net>.
- [25] KL Higbee, *Su memoria: cómo funciona y cómo mejorarla*, Nueva York: Marlowe y Compañía, 1996.
- [26] S. Todd, "Currículo de ciencia de datos y macrodatos" [en línea]. Disponible: [http://stevetodd.typepad.com/my\\_weblog/data-science-and-big-data-curriculum/](http://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/).
- [27] T. H Davenport y DJ Patil, "Científico de datos: el trabajo más sexy del siglo XXI", *Harvard Revision del negocio*, Octubre 2012.

# 3

## Revisión de datos básicos

## Métodos analíticos que utilizan R

*Conceptos clave*

*Características básicas de R*

*Exploración y análisis de datos con  
métodos estadísticos R para evaluación*

El capítulo anterior presentó las seis fases del ciclo de vida del análisis de datos.

- Fase 1: descubrimiento
- Fase 2: preparación de datos
- Fase 3: Planificación del modelo
- Fase 4: Construcción de modelos
- Fase 5: Comunicar resultados
- Fase 6: Operacionalización

Las primeras tres fases involucran varios aspectos de la exploración de datos. En general, el éxito de un proyecto de análisis de datos requiere una comprensión profunda de los datos. También requiere una caja de herramientas para extraer y presentar los datos. Estas actividades incluyen el estudio de los datos en términos de medidas estadísticas básicas y la creación de gráficos y diagramas para visualizar e identificar relaciones y patrones. Hay varias herramientas gratuitas o comerciales disponibles para explorar, acondicionar, modelar y presentar datos. Debido a su popularidad y versatilidad, el lenguaje de programación de código abierto R se utiliza para ilustrar muchas de las tareas analíticas y modelos presentados en este libro.

Este capítulo presenta la funcionalidad básica del lenguaje y entorno de programación R. La primera sección ofrece una descripción general de cómo usar R para adquirir, analizar y filtrar los datos, así como también cómo obtener algunas estadísticas descriptivas básicas en un conjunto de datos. La segunda sección examina el uso de R para realizar tareas de análisis de datos exploratorios mediante la visualización. La sección final se centra en la inferencia estadística, como la prueba de hipótesis y el análisis de varianza en R.

## 3.1 Introducción a R

R es un lenguaje de programación y un marco de software para análisis estadístico y gráficos. Disponible para su uso bajo la Licencia Pública General GNU [1], el software R y las instrucciones de instalación pueden obtenerse a través de Comprehensive RArchive and Network [2]. Esta sección proporciona una descripción general de la funcionalidad básica de R. En capítulos posteriores, esta base en R se utiliza para demostrar muchas de las técnicas analíticas presentadas.

Antes de profundizar en las operaciones y funciones específicas de R más adelante en este capítulo, es importante comprender el flujo de un script R básico para abordar un problema analítico. El siguiente código R ilustra una situación analítica típica en la que se importa un conjunto de datos, se examina el contenido del conjunto de datos y se ejecutan algunas tareas de creación de modelos. Aunque es posible que el lector aún no esté familiarizado con la sintaxis de R, el código puede seguirse leyendo los comentarios incrustados, indicados con #. En el siguiente escenario, las ventas anuales en dólares estadounidenses para 10,000 clientes minoristas se han proporcionado en forma de un archivo de valores separados por comas (CSV). La función `read.csv()` se utiliza para importar el archivo CSV. Este conjunto de datos se almacena en la variable R `ventas` utilizando el operador de asignación `<-`.

```
# importar un archivo CSV de las ventas anuales totales de cada cliente
ventas <- read.csv ("c:/data/yearly_sales.csv")
```

```
# examinar el conjunto de datos importado
cabeza (ventas)
```

## resumen (ventas)

```
# plot num_of_orders vs. ventas
plot(ventas $ num_of_orders, ventas $ sales_total,
     main = "Número de pedidos frente a ventas")

# realizar un análisis estadístico (ajustar un modelo de regresión lineal)
resultados <- lm (ventas $ ventas_total ~ ventas $ num_of_orders) resumen
(resultados)

# realizar algunos diagnósticos en el modelo ajustado
# trazar histograma de los residuos
hist (resultados $ residuales, rupturas = 800)
```

En este ejemplo, el archivo de datos se importa utilizando el `read.csv()` función. Una vez que se ha importado el archivo, es útil examinar el contenido para asegurarse de que los datos se hayan cargado correctamente y familiarizarse con los datos. En el ejemplo, el `cabeza()` función, por defecto, muestra los primeros seis registros de `ventas`.

```
# examinar el conjunto de datos importado
```

```
cabeza (ventas)
```

	<code>cust_id</code>	<code>sales_total</code>	<code>num_of_orders</code>	<code>género</code>
1		800,64	3	F
2	100002	217,53	3	F
3	100003	74,58	2	METRO
4	100004	498,60	3	METRO
5	100005	723,11	4	F
6	100006	69,43	2	F

los `resumen()` La función proporciona algunas estadísticas descriptivas, como la media y la mediana, para cada columna de datos. Además, se proporcionan los valores mínimo y máximo, así como el primer y tercer cuartiles. Porque el `género` La columna contiene dos caracteres posibles, una "F" (mujer) o "M" (hombre), la `resumen()` La función proporciona el recuento de la aparición de cada carácter.

## resumen (ventas)

<code>cust_id</code>	<code>sales_total</code>	<code>num_of_orders</code>	<code>género</code>
Min. : 100001	Min. : 30.02	Min. : 1.000	F: 5035
1er Qu.:102501	1er Qu.: 80.29	1er Qu.: 2.000	M: 4965
Mediana: 105001	Mediana: 151,65	Mediana: 2.000	
Media : 105001	Media : 249,46	Media : 2.428	
3er Qu.:107500	3er Qu.: 295.50	3er Qu.: 3.000	
Max. : 110000	Max. : 7606,09	Max. : 22.000	

Trazar el contenido de un conjunto de datos puede proporcionar información sobre las relaciones entre las distintas columnas. En este ejemplo, el `trama()` La función genera un diagrama de dispersión del número de órdenes (`ventas $ num_of_orders`) contra las ventas anuales (`ventas $ sales_total`). El `$` se usa para hacer referencia a una columna específica en el conjunto de datos `ventas`. La gráfica resultante se muestra en la Figura 3-1.

```
# plot num_of_orders vs. ventas
plot(ventas $ num_of_orders, ventas $ sales_total,
     main = "Número de pedidos frente a ventas")
```



FIGURE 3-1 Examinar gráficamente los datos

Cada punto corresponde al número de pedidos y al total de ventas de cada cliente. El gráfico indica que las ventas anuales son proporcionales al número de pedidos realizados. Aunque la relación observada entre estas dos variables no es puramente lineal, el analista decidió aplicar la regresión lineal utilizando la `lm()` función como un primer paso en el proceso de modelado.

```
resultados <- lm(ventas $ ventas_total ~ ventas $ num_of_orders) resultados
```

Llamada:

```
lm (fórmula = ventas $ ventas_total ~ ventas $ num_of_orders)
```

Coeficientes:

(Intercepción) ventas \$ num_of_orders	- 154,1	166,2
--	---------	-------

Los valores de intersección y pendiente resultantes son -154,1 y 166,2, respectivamente, para la ecuación lineal ajustada. Sin embargo, `resultados` almacena considerablemente más información que se puede examinar con el resumen(). Detalles sobre el contenido de `resultados` se examinan aplicando el atributos () función. Debido a que el análisis de regresión se presenta con más detalle más adelante en el libro, el lector no debe concentrarse demasiado en interpretar el siguiente resultado.

resumen (resultados)

Llamada:

```
lm (fórmula = ventas $ ventas_total ~ ventas $ num_of_orders)
```

Derechos residuales de autor:

Min	Mediana 1T	3T	Max
- 666,5	-125,5	-26,7	86,6
			4103,4

Coeficientes:

	Estimar Std. Valor t de error Pr (>   t  )		
(Interceptar)	- 154,128	4.129 -37.33	<2e-16 ***
ventas \$ num_of_orders	166.221	1.462 113.66	<2e-16 ***

---

Signif. códigos: 0 \*\*\*\* 0.001 \*\*\* 0.01 \*\* 0.05 .' 0,1 pulg. 1

Error estándar residual: 210,8 en 9998 grados de libertad R cuadrado múltiple: 0,5637,

R cuadrado ajustado: 0,5637

Estadístico F: 1.292e + 04 en 1 y 9998 DF, valor p: <2.2e-16

los resumen() función es un ejemplo de una función genérica. UN **función genérica** es un grupo de funciones que comparten el mismo nombre pero se comportan de manera diferente según el número y el tipo de argumentos que reciben. Utilizado anteriormente, trama() es otro ejemplo de función genérica; la trama está determinada por las variables pasadas. Las funciones genéricas se utilizan a lo largo de este capítulo y del libro. En la parte final del ejemplo, el siguiente código R usa la función genérica hist () para generar un histograma (Figura 3-2) de los residuos almacenados en **resultados**. La llamada a la función ilustra que se pueden pasar valores de parámetros opcionales. En este caso, el número de **rompe** se especifica para observar los grandes residuos.

```
# realizar algunos diagnósticos en el modelo ajustado
# plot histogr
hist (resultados $ r
```

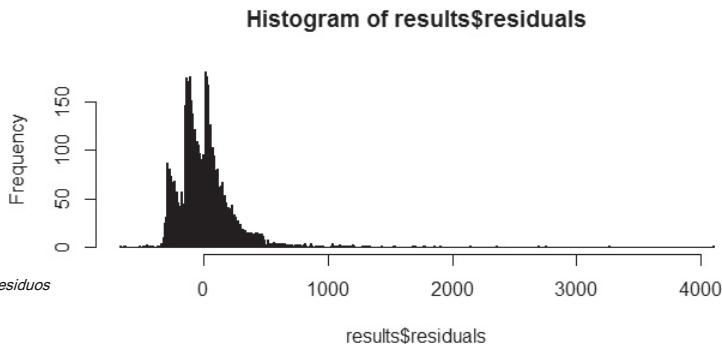


FIGURE 3-2 Evidencia de grandes residuos

Este sencillo ejemplo ilustra algunas de las tareas básicas de planificación y creación de modelos que pueden ocurrir en las Fases 3 y 4 del ciclo de vida de análisis de datos. A lo largo de este capítulo, es útil visualizar cómo se utilizará la funcionalidad R presentada en un análisis más completo.

### 3.1.1 Interfaces gráficas de usuario R

El software R usa una interfaz de línea de comandos (CLI) que es similar al shell BASH en Linux o las versiones interactivas de lenguajes de secuencias de comandos como Python. Los usuarios de UNIX y Linux pueden ingresar el comando R en el indicador de la terminal para usar la CLI. Para las instalaciones de Windows, R viene con RGui.exe, que proporciona una interfaz gráfica de usuario (GUI) básica. Sin embargo, para mejorar la facilidad de escritura, ejecución y depuración del código R, se han escrito varias GUI adicionales para R. Las GUI más populares incluyen R commander [3], Rattle [4] y RStudio [5]. Esta sección presenta una breve descripción general de RStudio, que se utilizó para construir los ejemplos de R en este libro. La Figura 3-3 proporciona una captura de pantalla del ejemplo de código R anterior ejecutado en RStudio.

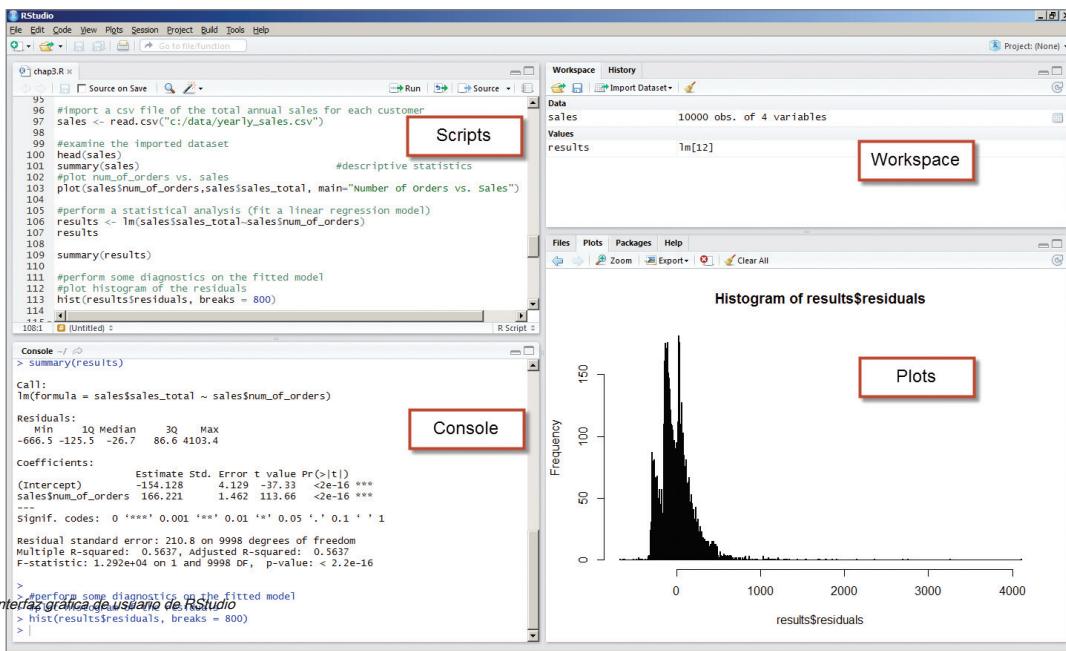


FIGURE 3-3 Interfaz gráfica del usuario de RStudio

A continuación, aparecen los cuatro paneles de ventana resaltados.

- **Guiones**: Sirve como área para escribir y guardar el código R
- **Espacio de trabajo**: Enumera los conjuntos de datos y las variables en el entorno R
- **Parcelas**: Muestra los gráficos generados por el código R y proporciona un mecanismo sencillo para exportar los gráficos.
- **Consola**: Proporciona un historial del código R ejecutado y la salida

Además, el panel de la consola se puede usar para obtener información de ayuda sobre R. La figura 3-4 lo ilustra ingresando `?lm` en el indicador de la consola, los detalles de ayuda del `lm()` función se proporcionan a la derecha. Alternativamente, `help(lm)` podría haberse ingresado en el indicador de la consola.

Funciones como `editar()` y `reparar()` Permitir al usuario actualizar el contenido de una variable R. Alternativamente, tales cambios se pueden implementar con RStudio seleccionando la variable apropiada en el panel del espacio de trabajo.

R permite guardar el entorno del espacio de trabajo, incluidas las variables y las bibliotecas cargadas, en un `.Rdata` archivo usando el `guardar_imagen()` función. Un existente `.Rdata` El archivo se puede cargar usando el `cargar_imagen()` función. Herramientas como RStudio preguntan al usuario si el desarrollador desea guardar las conexiones del espacio de trabajo antes de salir de la GUI.

Se anima al lector a instalar R y una GUI preferida para probar los ejemplos de R proporcionados en el libro y utilizar la función de ayuda para acceder a más detalles sobre los temas tratados.