Figure 1.1 What's driving the data deluge The rate of data creation is accelerating, driven by many of the items in Figure 1.1 .Analytic Sandbox (workspaces) Data assets gathered from multiple sources and technologies for analysis Enables flexible, high-performance analysis in a nonproduction environment; can leverage in-database processing Reduces costs and risks associated with data replication into "shadow" file systems "Analyst owned" rather than "DBA owned" There are several things to consider with Big Data Analytics projects to ensure the approach fits with the desired goals. Due to the characteristics of Big Data, these projects lend themselves to decision support for high-value, strategic decision making with high processing complexity. The analytic techniques used in this context need to be iterative and flexible, due to the high volume of data and its complexity. Performing rapid and complex analysis requires high throughput network connections and a consideration for the acceptable amount of latency. For instance, developing a real-time product recommender for a website imposes greater system demands than developing a near-real time recommender, which may still provide acceptable performance, have slightly greater latency, and may be cheaper to deploy. These considerations require a different approach to thinking about analytics challenges, which will be explored further in the next section. 1.2 State of the Practice in Analytics Current business problems provide many opportunities for organizations to become more analytical and data driven, as shown in Table 1.2 . Table 1.2 Business Drivers for Advanced Analytics Business Driver Optimize business operations Examples Sales, pricing, profitability, efficiency Identify business risk Predict new business opportunities Customer churn, fraud, default Upsell, cross-sell, best new customer prospects Comply with laws or regulatory requirements Table 1.2 Anti-Money Laundering, Fair Lending, Basel II-III, Sarbanes-Oxley (SOX) outlines four categories of common business problems that organizations contend with where they have an opportunity to leverage advanced analytics to create competitive advantage. Rather than only performing standard reporting on these areas, organizations can apply advanced analytical techniques to optimize processes and derive more value from these common tasks. The first three examples do not represent new problems. Organizations have been trying to reduce customer churn, increase sales, and cross-sell customers for many years. What is new is the opportunity to fuse advanced analytical techniques with Big Data to produce more impactful analyses for these traditional problems. The last example portrays emerging regulatory requirements. Many compliance and regulatory laws have been in existence for decades, but additional requirements are added every year, which represent additional complexity and data requirements for organizations. Laws related to anti-money laundering (AML) and fraud prevention require advanced analytical techniques to comply with and manage properly. 1.2.1 BI Versus Data Science The four business drivers shown in Table 1.2 require a variety of analytical techniques to address them properly. Although much is written generally about analytics, it is important to distinguish between BI and Data Science. As shown in Figure 1.8 ways to compare these groups of analytical techniques. , there are several Figure 1.8 Comparing BI with Data Science One way to evaluate the type of analysis being performed is to examine the time horizon and the kind of analytical approaches being used. BI tends to provide reports, dashboards, and queries on business questions for the current period or in the past. BI systems make it easy to answer questions related to quarter-to-date revenue, progress toward quarterly targets, and understand how much of a given product was sold in a prior

quarter or year. These questions tend to be closed-ended and explain current or past behavior, typically by aggregating historical data and grouping it in some way. BI provides hindsight and some insight and generally answers questions related to "when" and "where" events occurred.In-database processing for deep analytics enables faster turnaround time for developing and executing new analytic models, while reducing, though not eliminating, the cost associated with data stored in local, "shadow" file systems. In addition, rather than the typical structured data in the EDW, analytic sandboxes can house a greater variety of data, such as raw data, textual data, and other kinds of unstructured data, without interfering with critical production databases. Table 1.1 summarizes the characteristics of the data repositories mentioned in this section. Table 1.1 Types of Data Repositories, from an Analyst Perspective Data Repository Characteristics Spreadsheets and data marts ("spreadmarts") Spreadsheets and low-volume databases for recordkeeping Analyst depends on data extracts.Rather than aggregating historical data to look at how many of a given product sold in the previous quarter, a team may employ Data Science techniques such as time series analysis, further discussed in Chapter 8, "Advanced Analytical Theory and Methods: Time Series Analysis," to forecast future product sales and revenue more accurately than extending a simple trend line. In addition, Data Science tends to be more exploratory in nature and may use scenario optimization to deal with more open-ended questions. This approach provides insight into current activity and foresight into future events, while generally focusing on questions related to "how" and "why" events occur.Data Warehouses Centralized data containers in a purpose-built space Supports BI and reporting, but restricts robust analyses Analyst dependent on IT and DBAs for data access and schema changes Analysts must spend significant time to get aggregated and disaggregated data extracts from multiple sources.Figure 1.2 Examples of what can be learned through genotyping, from 23andme.com As illustrated by the examples of social media and genetic sequencing, individuals and organizations both derive benefits from analysis of ever-larger and more complex datasets that require . increasingly powerful analytical capabilities.See Figure 1.4 .See Figure 1.5