Chapter 1

Today's Data Systems Professional

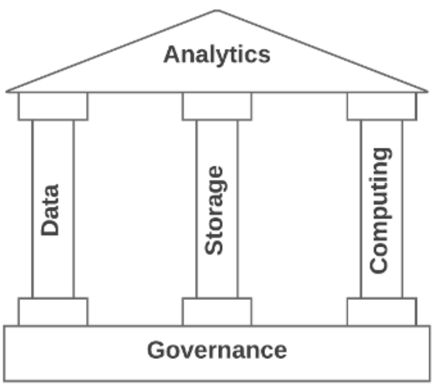
Data drives the modern business. Virtually every organization collects large quantities of data about its customers, products, employees, and service offerings. Managers naturally seek to analyze that data and harness the information it contains to improve the efficiency, effectiveness, and profitability of their work.

Data systems provide the ability to store, process, and analyze that data in an efficient and effective manner. Teams of dedicated data systems professionals design and manage these systems to improve their organization's ability to compete in today's marketplace. They are able to select appropriate data tools to deploy, manage, and maintain databases. They also understand the importance of maintaining the security of data and databases, as well as the essential nature of ensuring that business continuity programs protect data from loss. These skills allow them to better serve their colleagues throughout the business.

Data Drives the Modern Business

We are fortunate to live in the golden age of analytics. Businesses around the world recognize the vital nature of data to their work and are investing heavily in analytics programs designed to give them a competitive advantage. Organizations have been collecting this data for years, and many of the statistical tools and techniques used in analytics work date back decades. But if that's the case, why are we just now in the early years of this golden age? Figure 1.1 shows the three major pillars that have come together at this moment to allow analytics programs to thrive: data, storage, and computing power.

FIGURE 1.1 Analytics is made possible by modern data, storage, and computing capabilities.



Data

The amount of data the modern world generates on a daily basis is staggering. From the organized tables of spreadsheets to the storage of photos, video, and audio recordings, modern businesses create an almost overwhelming avalanche of data that is ripe for use in analytics programs.

Let's try to quantify the amount of data that exists in the world. We'll begin with an estimate made by Google's then-CEO Eric Schmidt in 2010. At a technology conference, Schmidt estimated that the sum total of all of the stored knowledge created by the world at that point in time was approximately 5 exabytes. To give that a little perspective, the file containing the text of this chapter is around 100 kilobytes. So, Schmidt's estimate is that the world in 2010 had total knowledge that is about the size of 50,000,000,000,000 (that's 50 trillion!) copies of this book chapter. That's a staggering number, but it's only the beginning of our journey.

Now fast-forward just two years to 2012. In that year, researchers estimated that the total amount of stored data in the world had grown to 1,000 exabytes (or one zettabyte). Remember, Schmidt's estimate of 5 exabytes was made only two years earlier. In just two years, the total amount of stored data in the world grew by a factor of 200! But we're still not finished!

In the year 2022, IDC estimates that the world created 94 zettabytes (or 94,000 exabytes) of new information. Compare that to Schmidt's estimate of the world having a total of 5 exabytes of stored information in 2010. If you do the math, you'll discover that on any given day in the modern era, the world generates an amount of brand-new data that is approximately 32 times the sum total of all information created from the dawn of civilization until 2010! Now, *that* is a staggering amount of data!

From an analytics perspective, this trove of data is a gold mine of untapped potential.

Storage

The second key trend driving the growth of analytics programs is the increased availability of storage at rapidly decreasing costs. Table 1.1 shows the cost of storing a gigabyte of data in different years using magnetic hard drives.

TABLE 1.1 Gigabyte Storage Costs over Time

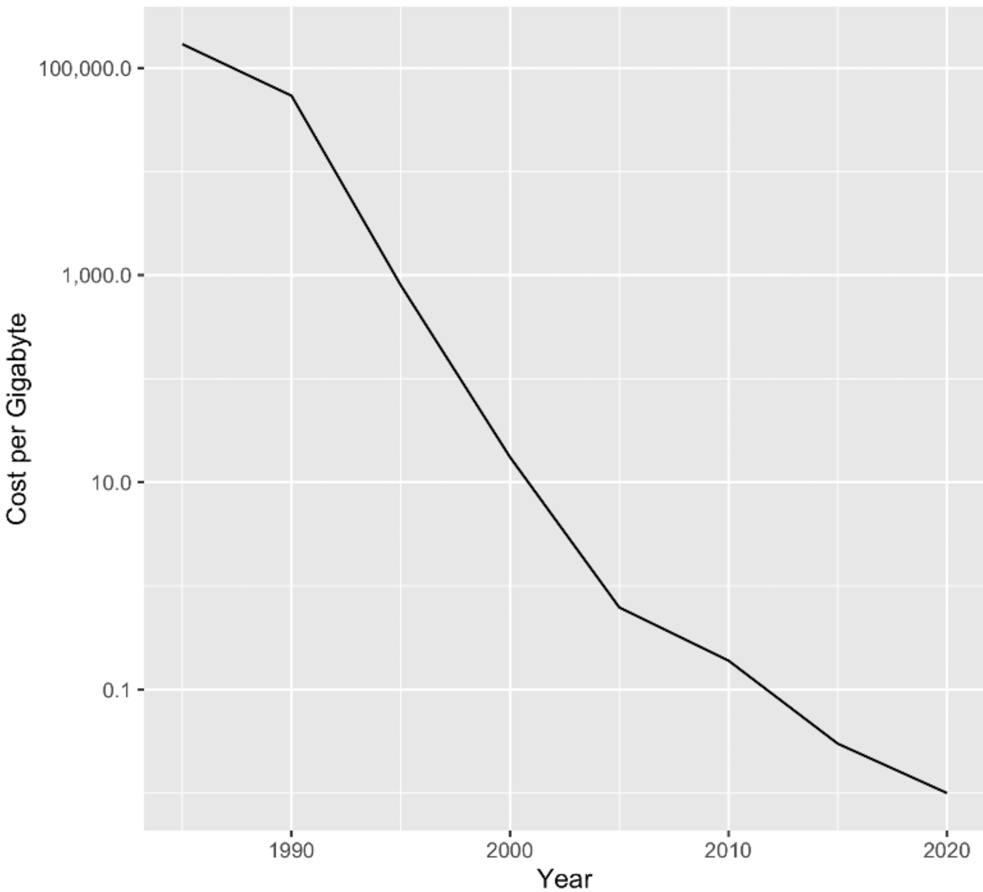
Year	Cost per GB
1985	$169,900
1990	$53,940
1995	$799
2000	$17.50
2005	$0.62

TABLE 1.1 Gigabyte Storage Costs over Time *(continued)*

Year	Cost per GB
2010	$0.19
2015	$0.03
2020	$0.01

Figure 1.2 shows the same data plotted as a line graph on a logarithmic scale. This visualization clearly demonstrates that storage costs have plummeted to the point where storage is almost free, and businesses can afford to retain data for analysis in ways that they never have before.

FIGURE 1.2 Storage costs have decreased over time.



Computing Power

In 1975, Gordon Moore, one of the cofounders of Intel Corporation, made a prediction that computing technology would continue to advance so quickly that manufacturers would be able to double the number of components placed on an integrated circuit every two years.

Commonly referred to as *Moore's law*, this prediction is often loosely interpreted to mean that we will double the amount of computing power on a single device every two years. That trend has benefited many different technology-enabled fields, among them the world of analytics.

In the early days of analytics, computing power was costly and difficult to come by. Organizations with advanced analytics needs purchased massive supercomputers to analyze their data, but those supercomputers were scarce. Analysts fortunate enough to work in an organization that possessed a supercomputer had to justify their requests for small slices of time when they could use the powerful machines.

Today, the effects of Moore's law have democratized computing. Most employees in an organization now have enough computing power sitting on their desks to perform a wide variety of analytic tasks. If they require more powerful computing resources, cloud services allow them to rent massive banks of computers at very low cost. Even better, those resources are charged at hourly rates and analysts pay only for the computing time they actually use.

These three trends—the massive volume of data generated by our businesses on a daily basis, the availability of inexpensive storage to retain that data, and the cloud's promise of virtually infinite computing power—come together to create fertile ground for data analytics.

Data Systems

Let's begin our exploration of data systems by looking at the material covered by each one of the domains on the CompTIA DataSys+ Exam. CompTIA organizes the exam content into five domains, and we have organized this book to follow those domains as well. The five domains and their weights on the exam are as follows:

- Database Fundamentals (24%)
- Database Deployment (16%)
- Database Management and Maintenance (25%)
- Data and Database Security (23%)
- Business Continuity (12%)

You'll find the material for each domain in either one or two chapters of the book. Those chapters are in order, so you will find, for example, complete coverage of the first domain in Chapters 2 and 3, followed by coverage of the second domain in Chapter 4.

The remainder of this chapter will summarize the information covered by each one of these domains and point you to where you will find full coverage.

Database Fundamentals

Databases are the most common way that organizations store their data. *Relational databases* organize data into tables of related information and then maintain relationships between those tables to preserve data integrity. These relational databases are the primary systems used by most organizations to store data, but *nonrelational databases* are growing in popularity. These databases, also called *NoSQL* databases, move beyond the relational model and store data in other formats, such as key-value stores and graphs.

In Chapter 2, "Database Fundamentals," you'll learn how data systems professionals work with databases. That chapter covers two of the major objectives found on the CompTIA DataSys+ Exam.

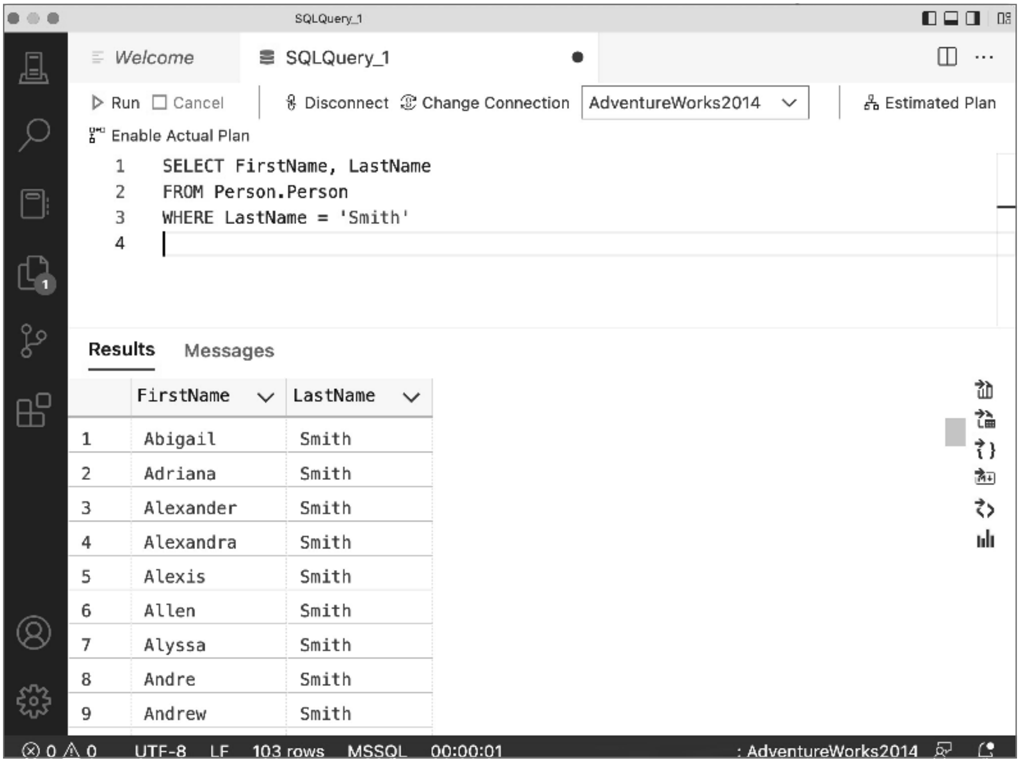
- 1.1 Compare and Contrast Database Structure Types.
- 1.4 Explain the Impact of Programming on Database Operations.

Chapter 3, "SQL and Scripting," explores the major tools that data systems professionals use to interact with databases. The primary tool is the *Structured Query Language (SQL)*, which provides a way to issue commands to relational databases. SQL includes two major sublanguages: the *Data Definition Language (DDL)* and the *Data Manipulation Language (DML)*. Data systems professionals use DDL to define and alter the structure of databases and then use DML to insert, update, delete, and retrieve data from databases. Figure 1.3 shows an example of a SQL command being executed against a Microsoft SQL Server database using the Azure Data Studio tool. The top portion of the interface shows the SQL query.

```
SELECT FirstName, LastName
FROM Person.Person
WHERE LastName = 'Smith'
```

The bottom portion of the interface shows the data retrieved from the database in response to that query. If you don't understand this syntax yet, don't worry. You'll learn all about it in Chapter 2!

FIGURE 1.3 SQLQuery



Data systems professionals also use programming and scripting languages to automate their work. You'll learn about the use of PowerShell and Python to create scripts on both Windows and Linux systems. Chapter 3 covers two of the CompTIA DataSys+ Exam objectives.

- 1.2 Given a scenario, develop, modify, and run SQL code.
- 1.3 Compare and contrast scripting methods and scripting environments.

Database Deployment

Data systems professionals must be able to examine a business need, identify the appropriate database tools to use to address that need, and then deploy a solution that fulfills the requirement.

This process begins with requirements gathering, where data professionals collect information from stakeholders to determine the basic parameters of the system. These include details on how many users will access the database, how they will interact with the database, and the type/quantity of storage needed to support the database. With this information in hand, data systems professionals may then design the architecture of the database that will meet those business requirements and document their design.

With a completed design in hand, the work then turns to the building and deployment of the database solution. Engineers will install and configure the database, import data, ensure that it has appropriate network connectivity, and then test and validate their solution before releasing it for use.

You'll find all of the material for this domain covered in Chapter 4, "Database Deployment." That chapter has two major objectives.

- 2.1 Compare and contrast aspects of database planning and design.
- 2.2 Explain database implementation, testing, and deployment phases.

Database Management and Maintenance

Once a database is deployed to end users, the work of data systems professionals has only just begun! Databases require ongoing monitoring and management to ensure that they are operating efficiently and meeting user needs.

Database monitoring is a continual task where administrators review any alerts generated by the database that may contain errors, warnings, or information regarding database performance. This monitoring may identify that the database is under- or over-resourced, requiring adjustments to the amount of memory, CPU, or storage available to the system. Monitoring activities also include periodic reviews of the logs generated by databases as well as the inspection of those logs in response to user trouble reports.

Modern databases do a remarkable job of interpreting user requests and developing efficient ways to answer queries, but data systems professionals have a set of tools at their disposal to improve database performance based upon user needs. For example, the use of *indexes* can dramatically speed up common user queries. Data professionals will also find themselves creating documentation for the database and performing data management tasks to ensure that the database remains useful to the organization.

You'll find all of the material for this domain covered in Chapter 5, "Database Management and Maintenance." That chapter has four major objectives.

- 3.1 Explain the purpose of monitoring and reporting for database management and performance.
- 3.2 Explain common database maintenance processes.
- 3.3 Given a scenario, produce documentation and use relevant tools.
- 3.4 Given a scenario, implement data management tasks.

Data and Database Security

Databases contain extremely sensitive information, including Social Security numbers, credit card numbers, and other personal information. Data systems professionals must be fluent in data security issues to protect their organizations against threats that might disrupt the confidentiality, integrity, or availability of that data.

Fortunately, there are many technical and administrative tools that we can use to protect our data. These include the use of *encryption* and *data masking* to obscure the meaning of sensitive data and secure data destruction techniques to remove data that is no longer needed. Data professionals must be familiar with these techniques and the implementation of auditing practices to ensure that they are properly followed.

Data professionals also play an important role in ensuring that organizations remain compliant with laws and regulations governing the use of personal information. These include the safeguarding of *personally identifiable information (PII)*, *protected health information (PHI)*, and *payment card information (PCI)*.

Chapter 6, "Governance, Security, and Compliance," introduces you to the important security concepts that data systems professionals must understand. It covers three of the objectives from the CompTIA DataSys+ Exam.

- 4.1 Explain data security concepts.
- 4.2 Explain the purpose of governance and regulatory compliance.
- 4.3 Given a scenario, implement policies and best practices related to authentication and authorization.

In addition to understanding the general security concepts described in Chapter 6, data systems professionals must understand the use of specific technologies to protect the data stored in databases. This includes configuring identity and access management systems to ensure appropriate authentication and authorization of user activities. Databases must be protected with logical security controls, including firewalls, perimeter protection, and port security controls. This domain also includes implementing physical security controls, such as surveillance, biometric authentication, fire suppression, and cooling systems.

Finally, data systems professionals must understand the threats posed to their systems by potential attackers in order to better defend against those attacks. This includes understanding SQL injection attacks, denial-of-service (DoS) attacks, on-path attacks, brute-force attacks, phishing attacks, and malware.

Chapter 7, "Database Security," dives into these specifics of securing databases. It covers two of the exam objectives:

- 4.4 Explain the purpose of database infrastructure security.
- 4.5 Describe types of attacks and their effects on data systems.

Business Continuity

Business continuity efforts are a collection of activities designed to keep a business running in the face of adversity. This may come in the form of a small-scale incident, such as a single system failure, or a catastrophic incident, such as an earthquake or a tornado. Business continuity plans may also be activated by human-made disasters, such as a terrorist attack or hacker intrusion.

Disaster recovery is a subset of business continuity activities designed to restore a business to normal operations as quickly as possible following a disruption. The disaster recovery plan may include immediate measures that get operations working again temporarily, but the disaster recovery effort is not finished until the organization is completely back to normal.

Data systems professionals must understand the role of business continuity and disaster recovery plans as well as the technical mechanisms used to achieve these goals in database systems. These include the use of database backups, database replication, and high availability controls to protect database systems.

You'll find all of the material for this domain covered in Chapter 8, "Business Continuity." That chapter has two objectives:

- 5.1 Explain the importance of disaster recovery and relevant techniques.
- 5.2 Explain backup and restore best practices and processes.

Careers in Data Systems

As businesses try to keep up with these trends, hiring managers find themselves struggling to identify, recruit, and retain talented data systems professionals. This presents a supply-and-demand situation that is problematic for businesses but excellent news for job candidates seeking to break into the field.

In a 2023 survey of business leaders, analysts at Gartner found that 84 percent believe that analytics is crucial to achieving their organizational goals. An earlier study from MicroStrategy found that 65 percent of firms planned to increase their analytics investment in the coming year. This will inevitably lead to increased demand for hiring data professionals, a fact that was confirmed by the World Economic Forum in their 2020 Future of Jobs Report. That study listed 10 occupations with the highest demand for professionals. The results, shown in Table 1.2, found that data analysts and data scientists are the most in-demand of any career field, closely followed by several other analytics-related fields.

TABLE 1.2 Highest-Demand Occupations

Rank	Occupation
1	Data analysts and scientists
2	AI and machine learning (ML) specialists
3	Big Data specialists
4	Digital marketing and strategy specialists
5	Process automation specialists

Rank	Occupation
6	Business development professionals
7	Digital transformation specialists
8	Information security analyst
9	Software and applications developers
10	Internet of Things (IoT) specialists

The future is bright. There's no reason to anticipate a reduction in this demand any time soon. It's the right time to enter the exciting field of data systems!

Summary

Analytics programs allow businesses to access the untapped value locked within their data. Today, many organizations recognize the potential value of this work but are still in the early stages of developing their analytics programs. These programs, driven by the unprecedented availability of data, the rapidly decreasing cost of storage, and the maturation of cloud computing, promise to create significant opportunities for businesses and, in turn, for data professionals skilled in the tools and techniques of data systems.

Data systems professionals perform a variety of important tasks within their organizations. These include the design and deployment of new database systems and the ongoing monitoring and management of deployed databases. They must also understand and support their organizations' security and business continuity programs to protect the sensitive and critical information stored within databases.