1

In the Beginning…

Data surrounds us. Every cell of every life form holds data of some sort. From the moment we are born, our senses send signals (data) to our brain that we continually process, creating yet more data as part of our thought processes. It's a never-ending cycle. To provide a context, it may be helpful to briefly explore how modern data processing evolved.

Since the very beginning of earliest civilizations, humankind has sought to share its thoughts and experiences with others through symbols such as drawings on cave walls, hieroglyphs in ancient tombs, ancient scrolls, papers, and books. As a species, our passion to learn and progress led to the desire and need to capture all this data, to store and share it for posterity and to pass our collective knowledge on to others as a means of building a civilization. The establishment of education delivered through scholastic programs and institutions helped formalize what we learn and how we learn. Educational, government, medical, public, and other organizations established their own libraries (the earliest forms dating back to 2600 BC), holding vast quantities of information, accessible for reference or for lending to patrons. Catalogs of this data have helped provide an indexed virtual representation of what is available, how it is stored, and where to find it, as well as often providing expert assistance from librarians or library technicians.

Early Data Storage and Management

In recent decades, analog recordings of audio, photos, and videos presented new dimensions of capturing data. Punched cards for gathering and processing early census data using tabulating machines appeared.

Recorded music on 78 rpm platters and "wire recorders" became a mainstay of radio. Magnetic tape emerged from the laboratory.

Information storage in most people's minds at the end of the World War II era meant books, filing cabinets or, to those at the leading edge of data processing technology, paper punch cards. Reels of tape, tape cartridges, and programmable computers were the stuff of science fiction. But in 1952 IBM announced the IBM 726, its first magnetic-tape unit, as shown in Figure 1.1. It shipped with the IBM 701 Defense Calculator. This innovation was significant because it was the first IBM large-scale electronic computer manufactured in quantity and was:

- IBM's first commercially available scientific computer
- The first IBM machine in which programs were stored in an internal, addressable, electronic memory
- Developed and produced in record time (less than two years from "first pencil on paper" to installation)
- Key to IBM's transition from punched-card machines to electronic computers with tape storage

Figure 1.1: An IBM 700 Series

IBM 701 Electronic Data Processing System included the IBM 701 electronic analytical control unit, IBM 706 electrostatic storage unit, IBM 711 punched-card reader, IBM 716 printer, IBM 721 punched-card recorder, IBM 726 magnetic-tape reader/recorder, IBM 727 magnetic-tape unit, IBM 731 magnetic-drum reader/recorder, IBM 736 power frame #1, IBM 737 magnetic-core storage unit, IBM 740 cathode-ray-tube output recorder, IBM 741 power frame #2, IBM 746 power distribution unit, and IBM 753 magnetic-tape control unit.

What followed was the advent of digital disk storage, which enabled organizations to collect and process more data faster than ever. In 1968, IBM launched the world's first commercial database-management system, called Information Control System and Data Language/Interface (ICS/DL/I). In 1969, it was renamed as Information Management System (IMS).

IBM's Database 2 (then abbreviated as DB2) traces its roots back to the beginning of the 1970s when Edgar F. Codd, a researcher working for IBM, described the theory of relational databases and in June 1970 published the model for data manipulation.

In 1974, the IBM San Jose Research center developed a relational Database Management System (DBMS), called System R, to implement Codd's concepts. A key development of the System R project was Structured Query Language (SQL). To apply the relational model, Codd needed a relational-database language he named DSL/Alpha. When IBM released its first relational-database product, it wanted to have a commercial-quality sublanguage as well, so it overhauled SEQUEL, and renamed the revised language Structured Query Language, a data-management language for relational databases still in use today.

The name "DB2" was first given to the DBMS in 1983 when IBM released DB2 on its MVS mainframe platform. (Source: https://en.wikipedia.org/wiki/IBM_Db2)

IBM and many other vendors continue to invest in relational and other forms of databases as they are one of the key technologies in online transactional processing (OLTP). IBM Db2, as it is known today, is also used for transaction analytics processing.

Relational databases have become the core technology for data warehouses and Master Data Management (MDM) systems (MDM systems are described below). In parallel to relational databases, other forms of data stores appeared, such as object-oriented, NoSQL, key value, wide-column store, and graph databases, to name but a few.

From Centralized to Distributed

For many years, data storage and processing were centralized. People had to take their work to the computer or access it through "dumb" terminals. With the advent of more affordable computers, processing and data became decentralized, putting computing power in the hands of individuals. However, this led to the problem of data being replicated in an uncontrolled manner.

With data being created, stored, and processed across many personal devices, it became increasingly difficult to control the sprawl of versions of data sets and apply standards of quality, security, and other controls. It didn't take long for individual departments in various enterprises to start organizing and storing just the data they needed. However, this gradually resulted in the problem of creating many data silos that usually didn't communicate with each other across the organization.

Master Data Management (MDM) is the discipline by which business and information technology work together to ensure the uniformity, accuracy, stewardship, semantic consistency, and accountability of the enterprise's official shared master data assets. Combined with data warehousing, MDM helps provide a 360-degree view of an entity, such as a person or product. (The reference to 360-degree view implies users should be able to look at an entity from many different perspectives to form a more complete understanding of that entity.) In a sense, MDM's creation was an attempt to recentralize some of the key data that was being held in disparate silos so it could be used across the whole organization as a trusted source of data—a single version of the truth, if you will. However, it still left the problem that there were often prime copies and distributed secondary copies of the data that needed to be kept synchronized to provide a source of truly trusted data.

Databases, OLTP, OLAP, Warehouses, Master Data Management, Marts, Lakes, Lakehouses, Hadoop

Numerous solutions appeared for managing and integrating data to enable reporting, analysis, and discovery of insights as data volumes grew. All of them were data stores given names such as database, OLTP, OLAP, data warehouses, MDM systems, data marts, data lakes, data lakehouses, and Hadoop. These terms tend to be used somewhat interchangeably at times. While the terms are similar, important differences exist that are explained in Appendix A. Each provides certain capabilities and values to different groups of users, but none was a panacea for all data management challenges, as originally hoped for when each was created. However, technology follows a maturity curve or cycle and these technologies eventually found their own niches as they matured.

Many forms of data stores and data servers are being used across the enterprise today. More variations of these, and new paradigms, will evolve in the future. Technology is constantly advancing. The authors of this book perceive the data fabric approach as offering enough longevity and flexibility to be able to integrate an organization's data assets and enable them for AI.