The financial industry has always been driven by data. Today, Big Data is prevalent at various levels of this field, ranging from the financial services sector to capital markets. The availability of Big Data in this domain has opened up new avenues for innovation and has offered immense opportunities for growth and sustainability. At the same time, it has presented several new challenges that must be overcome to gain the maximum value out of it. This chapter considers the impact and applications of Big Data in the financial domain. It examines some of the key advancements and transformations driven by Big Data in this field. The chapter also highlights important Big Data challenges that remain to be addressed in the financial domain.
17.1 INTRODUCTION

In recent years, the financial industry has seen an upsurge of interest in Big Data. This comes as no surprise to finance experts, who understand the potential value of data in this field and are aware that no industry can benefit more from Big Data than the financial services industry. After all, the industry not only is driven by data but also thrives on data. Today, the data, characterized by the four Vs, which refer to volume, variety, velocity, and veracity, are prevalent at various levels of this field, ranging from capital markets to the financial services industry. In recent years, capital markets have gone through an unprecedented change, resulting in the generation of massive amounts of high-velocity and heterogeneous data. For instance, about 70% of the US equity trades today are generated by high-frequency trades (HFTs) and are machine driven [1]. The prevalence of electronic trading has spurred up growth in trading activity and HFT, which, among other factors, have led to the availability of very large-scale ultrahigh-frequency data (UHFD). These high-speed data are already having a huge impact in the field in several areas ranging from risk assessment and management to business intelligence (BI). For example, the availability of UHFD is forcing the market participants to rethink the traditional ways of risk assessment and bringing up attention to more accurate, short-term risk assessment measures. Similar trends can be observed in the financial services sector, where Big Data is increasingly becoming the most significant, promising, and differentiating asset for the financial services companies. For instance, today, customers expect more personalized banking services, and to remain competitive as well as comply with the increased regulatory surveillance, the banking services sector is under tremendous pressure to best utilize the breadth and depth of the available data. In recent years, firms have already started using the information obtained from the vast oceans of available data to gain customer knowledge, anticipate market conditions, and better gauge customer preferences and behavior ahead of time, so as to offer highly personalized customer-centric products and services to their customers, such as sentiment analysis–enabled brand strategy management and real-time location-based product offerings as opposed to the historically offered product-centric services. Moreover, events like the credit crisis of 2008 have further shifted the focus of such financial entities towards Big Data as a strategic imperative for dealing with the acute stresses of renewed economic uncertainty, systemic monitoring, increasing regulatory pressure, and banking sector reforms. Unarguably, similar developments can be seen in other areas like asset management and insurance.

Clearly, such examples are indicative of the transformations ensuing in the finance sector, whereby more and more financial institutions are resorting to Big Data to strategize their business decisions based on reliable factual insights supported by real data rather than just intuition. Additionally, Big Data is now playing a critical role in several areas like investment analysis, econometrics, risk assessment, fraud detection, trading, customer interactions analysis, and behavior modeling.

In this digital era, we create approximately 2.5 quintillion bytes of data every day, and 90% of the data in the world today have been created in the last 2 years alone. The Big Data market is estimated to be at $5.1 billion this year and is expected to grow to $32.1 billion
by 2015 and to $53.4 billion by the year 2017 [2]. Today, almost all sectors of the financial field are inundated with data generated from a myriad number of heterogeneous sources, such as hundreds of millions of transactions conducted daily, ultrahigh-frequency trading activities, news, social media, and logs. A recent survey shows that around 62% of companies recognize the ability of Big Data to gain competitive edge [2], and there is no doubt that the prevalent Big Data offers immense potential and opportunity in the finance sector. However, the enormously large financial data volumes, high generation speeds, and heterogeneity associated with the relevant financial domain data, along with its susceptibility to errors, make the ingestion, processing, and timely analysis of such vast volumes of often heterogeneous data very challenging.

There is no clear consensus among and within financial institutions today on the best strategies to harvest and leverage the available Big Data to actionable knowledge. This can be attributed to the fact that a single solution is unlikely to cater to the growing needs of different businesses within the financial domain spectrum. Today, many financial organizations are exploring and adopting customized Big Data solutions that are specifically tailored for their domain-specific needs. Section 17.2 presents an in-depth view of the financial domain with details on the historical trends in this sector and innovations in the field as a result of Big Data. The section will also cover the three key elements involved in financial domain dynamics, namely, Big Data sources in finance, information flow, and data analytics.

### 17.2 FINANCIAL DOMAIN DYNAMICS

#### 17.2.1 Historical Landscape versus Emerging Trends

From the past several years, data warehouse systems primarily based on relational database management systems (RDBMSs) have been serving as the front-runners in providing access to the business community with necessary intelligence in the field of finance. These systems are mostly constructed out of quantitative data from operational systems, and BI tools are used to access the mostly well-understood operational data in the data warehouses [3]. Such systems are still widely used when it comes to simple analytical jobs or tasks like online analytical processing (OLAP), but their usage is restricted to small-scaled, well-defined, and structured data sets. However, the well-defined boundaries that once existed between the operational and decision-making tasks handled by such systems are increasingly becoming fuzzy in the financial world today. Although these systems were the norm in the early 90s, they are slowly losing their precedence in the area of BI and, hence, decision making due to their limited data handling and analytical capabilities.

In recent years, financial organizations have started to rethink and restructure the way they do business. This change is driven by the confluence of several factors like escalating regulatory pressures, ever-increasing compliance requirements, regulatory oversight, global economic instability, increasing competition in the global markets, growing business demands, need to optimize capital and liquidity, need to improve product and customer relationship margins, and so forth. For instance, regulations such as the European Market Infrastructure Regulation (EMIR) and the Financial Stability Oversight council’s...
Frank–Dodd Act have by themselves added hundreds of new rules affecting banking and securities industries. These directives for greater transparency are leading to enormous increases in the data volumes across such industries and forcing them to redesign their services infrastructure to cater to the new demands. Similar data growth trends can be seen in other parts of the financial domain. For example, there was a tenfold increase in market data volumes between 2008 and 2011, and the data volumes are growing stronger in all areas of the financial domain; for example, some of the top European insurers reported a six-fold increase in the amount of data and analytic reporting required by just the first pillar of the Solvency II insurance reform regulation [4]. The New York Stock Exchange (NYSE) by itself creates about several terabytes of market and reference data per day covering the use and exchange of financial instruments, whereas Twitter feeds, often analyzed for sentiment analysis in the financial domain, generate about 8 terabytes of data per day of social interactions [4]. There are around 10,000 payment card transactions executed per second across the globe; there were about 210 billion electronic payments generated worldwide in 2010, and the number is expected to double by the end of the decade [4]. Various other developments in the financial system are also contributing enormously to the overall volume of the data in the system. One such example of this shift can be explained by the emergence of the originate-to-distribute model that has broken down the traditional origination process into a sequence of highly specialized transactions and has led to an increase in the volume of the data in this domain. In this model, financial products like mortgages are systematically securitized and then structured, repackaged, and distributed again, so the loan details that traditionally might have been recorded only by the original lender and the borrower are now shared across multiple, diverse entities such as the originating bank, borrower, loan servicer, securitization trust and bondholders, as well as buyers and sellers of credit protection derivatives [5]. Besides contributing to the data volume, each new entity in the system adds to the complexity of the involved data. The digital universe is expected to grow nearly 20-fold, to approximately 35 zettabytes of data, by the year 2020 [6].

Traditional data management practices in finance can no longer effectively cope with the ever-increasing, huge, and rapid influx of heterogeneous (structured, semistructured, unstructured) data originating from a wide range of internal processes and external sources, including social media, blogs, audio, video, and so forth. Conventional data management technologies are destined to fail with such growing data volumes, which far exceed the storage and analysis capabilities of many traditional systems, and they have in many instances. For instance, with regard to the volume and complexity of created data, back offices of trading firms have failed to keep up with their own front office processes as well as the emerging data management practices adopted in other industries to handle growing data volumes and a multitude of diverse data types [5]. Moreover, traditional systems are not equipped to handle the wide variety of data, especially unstructured data, from social media, like news, Twitter, blogs, videos, and so forth, that is needed to gain insights about businesses processes (e.g., risk analysis, trading predictions) and keep up with the evolving needs of the customer in the financial services industry. Such systems often fail when it comes to integration of heterogeneous data or even real-time processing of structured data. In fact, this is becoming a bottleneck for many top-tier global banking systems since the
introduction of new regulations that require the banks to provide a complete horizontal view of risk within their trading arms. This task entails the integration of data from different schemas unique to each of the trade capture systems into a central repository for positions, counterparty information, and trades. Extraction, transformation, cleansing, and integration of such data via traditional extract, transform, load (ETL)-based approaches, coupled with samplings, often span several days and are not very accurate in the scenarios where only a sample of the data is used for analysis. New regulations, however, demand that this entire pipeline be executed several times a day, a feat clearly infeasible using the conventional approach. These regulations are similarly applicable to the capital markets where the regulations necessitate an accurate view of different risk exposures across asset classes and lines of businesses and firms to better estimate and manage systemic interplays. These tasks require simulations of a variety of risk situations, which can result in the generation of terabytes of additional data per day [7].

In recent years, it is becoming increasingly important for financial firms to adopt a data-centric perspective to handle the mounting regulatory pressures and succeed in today’s digital, global marketplace. In the past, financial organizations collected large amounts of data. However, these institutions depended primarily on the conventional ETL framework and lacked the ability to process the data and produce actionable knowledge out of it within realistic time frames. This approach prevented them from gaining a full perspective of their business insights and made it difficult for them to anticipate and respond to changing market conditions, business needs, and emerging opportunities, a few must-haves essential to thrive in today’s dynamic business environment. As a result, the firms relying on the traditional schemes have started to address the limitations inherent in their conventional systems. Today, a growing number of financial institutions are exploring new ways of unlocking the potential of available data to gain insights that can help them improve their performance and gain competitive advantage through factual and actionable knowledge, timely decisions, risk management and mitigation, and efficient operations in highly complex and often volatile business environments. Figure 17.1 highlights the importance and applicability of Big Data in the financial domain. The figure illustrates the key sectors of finance in which the power of Big Data is being harnessed to address critical business needs ranging from product innovation and fact-driven strategic decision making to the development of novel and intelligent business solutions.

The financial industry has always been one of the most data-driven industries. In the past few years, the prevalence of Big Data has opened up new horizons in the financial fields. Several industries have already started exploiting the value out of Big Data for information discovery in areas like predictive analytics based on social behavior mining, deep analytics, fraud detection, and risk management. For example, most of the credit card companies mine several thousands to millions of records, aggregated from customer transactional data (structured), call records (unstructured), e-mails (semistructured), and claims data (unstructured), to proactively anticipate future risks, accurately predict customer card switching behavior, and devise measures to improve customer relationships based on such behavioral modeling. Likewise, several firms involved in financial risk management perform risk assessment by integrating large volumes of transactional data with
Big Data applications in key financial domain sectors.

17.3.1 Big Data Origins

Nowadays, the financial markets are getting inundated with a flurry of data that is increasing in complexity as well as size. Today, the stock markets encompass a wide variety of data from diverse sources such as market data, which include orders and trades; reference data comprising data related to ticker symbols, exchanges, security descriptions, corporate actions, and so forth; and fundamental data, including corporate financials, analyst reports, filings, news (including earnings reports, economic news, etc.), and social media data (comprising blogs, Twitter feeds, etc.) [8,9]. These data sources provide a variety of information in different, structured, semistructured, and unstructured, formats, thereby
adding to the heterogeneity of the data. Moreover, the data generated and utilized by these markets are highly voluminous and complex. The reason for the complexity in financial data is that agents acting in markets trade increasingly faster and in more numerous and complex financial instruments, and have better information-acquiring tools than ever before. The reason for heterogeneity stems not just from the fact that agents and regulations across the globe are themselves heterogeneous but also from the more mundane fact that people report information in ways that are not standardized. For example, investment managers reporting holdings to the Securities and Exchange Commission (SEC) often err regarding the proper unit of measurement (thousands or units); performance data are often “dressed” to appear more attractive (most often, smoother); and so on.

Recent projections by the Options Pricing Reporting Authority for the years 2014–2015 estimate a total of 26.9 to 28.7 billion messages per day, 17.7 to 19 million messages per second, and a maximum output rate of at least 1 million messages per second [10]. For example, the data covering the quotes and transactions from the major US exchanges (trade and quote database [TAQ]) grow exponentially, now at a rate of hundreds of terabytes per year. Analysis of the operational structures of the underlying data generation entities partially explains the enormity and complexity of the massive data sets. For instance, unlike the traditional days of specialists and natural price discovery, today, the US stock market structure comprises an aggregation of different exchanges, broker-sponsored execution venues, and alternative trading systems, each of which contributes differently to the market data and volumes [11]. Specifically, around 14 exchanges, approximately 50 dark pools, and more than 200 international platforms or venues contribute around 66%, 13%, and 21% volume, respectively [8]. Orders are submitted through more than 2000 broker deals, and the system is governed by various regulatory agencies including SECs, SROs, and so forth [8]. The estimated average trading volumes for the market include about $50- to $100-billion-value trades, at least 2 billion order submissions, and 5 billion share trades [8,12]. The dynamic system environment comprising complex trade work flows (e.g., billions of trades or order submissions, price matches, executions, rejections, modifications, acknowledgments, etc.) and the changing market trading practices such as HFT further contribute to the volume and complexity of the generated data. Figure 17.2 portrays an example showing the high-level view of a typical automated electronic trading system. The key blocks shown in the figure are representative of extremely intricate models, strategies, and data, among other factors, which add to the complexity of the overall system. In the United States, the HFTs were estimated to account for more than 70% of equity trades in the year 2010 [11]. Besides adding to the volume, such evolving trading techniques are resulting in very high data generation speeds. For example, order matching and subsequent trade execution can now be accomplished in less than 100 $\mu$s via a colocated server in the exchange, and algorithmic trading can now be done within microseconds [11].

17.3.2 Information Flow

Financial markets offer comprehensive platforms that facilitate complex interplays among different market participants; support large-scale information ingestion and aggregation for price discovery; and provide liquidity for uninformed, liquidity-seeking
order flows [13]. The data generated through the individual market platforms are disseminated to the market participants through different channels. High-speed market data are directly delivered to some entities through the principal electronic communication networks (ECNs) such as INET, whereas the data are delivered to a majority of other entities through distribution channels like the National Association of Securities Dealers Automated Quotations (NASDAQs) dissemination or the Consolidated Trade System (CTS), which directly or indirectly collects all US trades, and the Consolidated Quote System (CQS) [14]. The latter dissemination channel, however, collects data at a much slower pace compared to the speed at which the data are generated [14]. Different market participants often require and utilize data with different price granularities, and hence precision, depending upon their diverse trading objectives. For instance, high-frequency traders and market makers like the NYSE dedicated market makers (DMMs) generally utilize UHFD (tick data), are highly sensitive to small price changes, and deal with several thousands of orders per day. In contrast, investors like the pension funds investors normally base their investment decisions on low-frequency or aggregated data, are not too sensitive to small price changes (e.g., at the intraday level), and usually deal with no more than a few hundreds of orders per day. Unregulated investors like the hedge funds and other speculators like the day traders, on the other hand, generally fall somewhere in between the investors and market makers. Figure 17.3 further exemplifies such market dynamics that exist among different market participants. All these different players and the inherent intricacies of the underlying processes complicate the order or other information flows. For example, about one-third of price discovery nowadays occurs in dark
pools, and even today, these pools remain largely unobservable. Such factors add to the complexity of the market structures and make it difficult to understand the transformations of orders into trades and how they drive the price discovery process.

Market participants usually deploy different trading strategies in line with their trading objectives. Typically, the raw market data, collected by these market agents, are refined, aggregated, and analyzed. Today, many market participants are resorting to novel ways of information discovery and incorporating additional information in their trading strategies, which involve integration of traditional data (e.g., orders and trades) with nontraditional data (e.g., sentiment data from social networks, trending over time, news, exploratory and deep analysis of the available data through efficient interactive or ad hoc queries). There is no doubt that the vast amounts of information that is generated by such trading systems along with the information that exists in the complex networks of legal and business relationships that define the modern financial system hold all the answers required to understand and accurately predict unexpected market events as well as address the most demanding questions that plague the financial systems and, hence, the regulators.

In recent years, much of the focus has shifted to finding ways to extract all the necessary answers from these ever-growing financial data within realistic time frames. However, to date, the problem remains challenging for many in this field not only because of the Big Data constraints but also due to factors like the lack of transparency; absence of standardized communication protocols; and ill-defined work flows among different data generation systems, processes, and organizations.
17.3.3 Data Analytics

As demonstrated in Sections 17.1 and 17.2, Big Data is gaining abundance in almost all levels of the financial spectrum, and financial institutions have already started leveraging Big Data to remain competitive; cater to the demands enforced by growing regulatory pressures, highly dynamic environments, evolving customer needs; seize market opportunities; and efficiently handle risk, to name a few. However, Big Data by itself does not hold much value, and not all of it may be useful at all times. It is necessary to transform the available data into actionable knowledge to realize the true potential of Big Data and extract valuable information or factual strategic insights from it. Figure 17.4 depicts the key Big Data drivers in the financial sector and shows the essential elements of a Big Data pipeline. As shown in the figure, every Big Data pipeline usually involves heterogeneous sources of data that could be intrinsic, extrinsic, or some combination of the two. The data collected from such sources subsequently go through a data ingestion and integration step. Traditional ways of data integration primarily followed an ETL data approach. As discussed in Section 17.2, the ETL approach is not always appropriate and can have many limitations, especially when it comes to Big Data. Several advancements in the last couple of years have led to the development of novel solutions and architectures that

FIGURE 17.4  Big Data ecosystem in finance.
address some of the limitations inherent in the traditional systems and offer unique capabilities to efficiently handle today’s Big Data needs. The ultimate goal of a Big Data pipeline is to facilitate analytics on the available data. Big Data analytics provides the ability to infer actionable insights from massive amounts of data and can assist with information discovery during the process. It has become a core component that is being deployed and used by entities operating at various spheres of the financial field. For example, predictive analytics tools are increasingly being deployed by the banks to predict and prevent fraud in real time. Predictive analytics applies techniques from data mining, data modeling, and statistics to identify relevant factors or interactions and predict future outcomes based on such interactions. Predictive analytics tools are also increasingly being used by market participants for tasks like decision making, improving trading strategies, and maximizing return on equities.

Big Data analytics is also being used in the capital markets for governance-based activities such as detection of illegal trading patterns and risk management. For example, NYSE Euronext has deployed a market surveillance platform that employs Big Data analytics to efficiently analyze billions of trades to detect new patterns of illegal trading within realistic time frames. The analytics platform allows Euronext to process approximately 2 terabytes of data volume everyday, and this volume is expected to exceed 10 petabytes a day by 2015. The deployed infrastructure has been reported to decrease the time required to run market surveillance algorithms by more than 99% and improve the ability of regulators or compliance personnel to detect suspicious or illegal patterns in trading activities, allowing them to take proactive investigative action to mitigate risks [15]. Similarly, another market information data analytics system (MIDAS) went online at the SEC in January 2013. MIDAS focuses on business data sources in the financial domain, with particular emphasis on the filings periodically required by the companies to be made with the SEC and Federal Deposit Insurance Corporation (FDIC). The system provides valuable insights about financial institutions at systemic and individual company levels by harnessing the value out of market data as well as data archived by the SEC and FDIC. It processes about 1 terabyte of stocks, options, and futures data per day and millions of messages per second. The analytics system is being utilized by subscribers to analyze mini flash crashes, assess impacts of rule changes, and detect abnormal patterns in the captured data [8].

17.4 EMERGING BIG DATA LANDSCAPE IN FINANCE

As discussed in Sections 17.1–17.3, Big Data has already initiated several innovations in the field of finance and has started to reshape some of the traditional ways of doing business. For instance, in recent years, capital markets have evolved from low-frequency trading involving simple trading strategies, like 1980s-paired models, to ultrahigh-frequency trading and the intricate gaming strategies of today. An ever-increasing need to efficiently and effectively exploit Big Data is driving the development of a new generation of technologies and architectures that allow real-time information extraction and discovery by facilitating high-speed data ingestion, data integration, and analysis of massively large and heterogeneous data sets within realistic time frames. Nonetheless, the development and
adoption of Big Data–centric technologies and architectures remain challenging. Section 17.4.1 discusses some of these challenges in the financial domain.

17.4.1 Challenges
Despite being aware of the significant promise that Big Data holds, many companies still have not started to make any investments to reap its benefits. A lot of companies today waste more than half of the data they already hold. If we assess the value of these data based on the Pareto principle, that is, 80% of the value comes from 20% of the data, then clearly, a lot of value is getting lost [6]. There is no doubt that the Big Data domain is rapidly evolving, but the domain is still premature in the financial field, with several factors slowing its growth and adoption in this domain. Data management and sharing has been a difficult problem for capital market firms for decades. The recent financial collapse and the mortgage/credit crisis of 2008 have uncovered some of the bottlenecks and inadequacies inherent in the information structure of the US financial systems. Factors like the lack of standardized data communication protocols, lack of transparency, complex interactions, and data quality gaps across different financial units in the system make it extremely difficult to unravel and connect different systems, processes, and organizations for any kind of analysis within realistic time frames. Many of these limiting factors represent an evolutionary outcome of the years of mergers, internal fragmentations, and realignments within the financial institutions and have been made worse by the business silos and inflexible Information Technology (IT) infrastructures. To date, the work flows in the system largely remain ill defined, and data reside in unconnected databases and/or spreadsheets, resulting in multiple formats and inconsistent definitions across different organizational units. Data integration remains point to point and occurs tactically in response to emergencies [9]. Convergence of such factors makes it almost infeasible to ingest, integrate, and analyze large-scale, heterogeneous data efficiently across different financial entities within a financial organization. The lack of best practices, data sharing procedures, quality metrics, mathematical modeling, and fact-based reasoning has left even the federal regulatory agencies unable to ingest market information in a timely manner, permit a proactive response, or even determine what information might be missing [9].

Financial institutions have historically spent vast sums on gathering, organizing, storing, analyzing, and reporting data through traditional data management approaches. As demonstrated in Section 17.2, the conventional data management structures are no longer sufficient to handle the massively large, high-velocity, heterogeneous financial data. Today, it is critical to deploy new supporting infrastructure components, data ingestion and integration platforms, as well as Big Data analytics and reporting tools, to handle extremely large-scale, often real-time, heterogeneous data sets. However, due to the varied data and business requirements of different organizational units, it is not realistic to expect a single solution to fit or even be equally applicable to all the financial entities. Moreover, due to the lack of well-defined solutions, an exploratory, incremental, and possibly iterative approach would be needed to devise customized and efficient Big Data solutions. Therefore, it is important for the financial organizations to devise their Big Data investment strategies with focus on their business-specific goals, like what is needed for risk management, product innovation, risk and market intelligence, cost reduction, services, operations, and so
forth. For instance, risk analytics and reporting requirements would necessitate a platform that could help consolidate risk measure and provide powerful risk management and reporting capabilities. A market data management system, on the other hand, would require a highly scalable platform that could store, process, and analyze massive amounts of heterogeneous data sets in real time.

17.4.2 New Models of Computation and Novel Architectures

Real-time Big Data platforms entail a multitiered architecture mainly comprising data ingestion, integration, analytics, visualization/reporting, and decision-making components. A lot of technological transformations are underway towards the development of such platforms in the financial sector to efficiently manage and utilize Big Data. For instance, to store and manage extremely large and growing volumes of data, the focus has shifted from the scale-up storage solutions to the scale-out storage solutions wherein large shared pools of storage devices are used to generate more capacity while reducing operational costs. Similarly, new data-intensive supercomputing (DISC) appliances are increasingly being deployed to overcome the limitations inherent in the traditional financial data management systems and perform complex analytics on the massively large data sets. Specifically, NASDAQ, NYSE Euronext, and Lucera have started to utilize systems like Netezza [16], Greenplum [17], and Scalable Informatics’ Cadence and Resonance [18] for their Big Data and analytics needs. DISC appliances like Netezza represent a new paradigm in data-intensive computing wherein the processing is done at the location of the data. Technologies like complex event processors (CEPs), in-memory databases, and data grids are also being used for real-time analytics in the financial industry. CEP engines enable analytics on streaming data and are used by front offices for algorithmic trading and so forth. They are used by the services industry to detect real-time events such as generation and monitoring of payment transactions or other core banking transactions [7].

Data grids facilitate distributed processing and storage of data in memory and are often used to complement CEP engines. To handle unstructured data, solutions based on open-source frameworks like Hadoop and MapReduce are commonly used in the industry today. These technologies offer high scalability and performance via parallelism and data locality-aware processing [7]. Besides Hadoop, other NoSQL (interpreted as Not only SQL) databases such as graph databases and document databases are also in use today for handling nontraditional, semistructured and unstructured, financial data sets. These Big Data management systems are primarily deployed based on the Brewer’s theorem, also known as CAP theorem, which states that the systems have to pick two out of the three properties, consistency, availability, and partition tolerance. Most of the solutions developed using these technologies are tailored as basically available, soft-state, eventually consistent (BASE) systems, unlike the strictly atomicity, consistency, isolation, durability (ACID) compliant traditional data management systems. Big Data technologies like Hadoop and NoSQL working in conjunction with event-stream processing systems like Storm are already playing a key role in financial market operations [19]. Many financial organizations have already started to reap the benefits of their Big Data forays in the field using a combination of existing and new Big Data technologies. More innovations and technological improvements in this field are expected to follow in the near future as more financial units begin to incorporate Big Data solutions.
17.5 IMPACT ON FINANCIAL RESEARCH AND EMERGING RESEARCH LANDSCAPE

17.5.1 Background

Financial markets are complex systems driven by the influx of exogenous information coming from sources external to the system and regulated by an internal dynamics [20]. In recent years, the availability of UHFD has opened up new horizons in the field. UHFD refer to a financial market data set that comprises of all the transactions or tick-level activity [21,22]. Nowadays, millions of data points, representing the tick- or transaction-level activity, stream out of the global financial markets, driving decision strategies of the various market participants. The data hold high practical relevance because a rising number of market participants today execute trades based on high-frequency strategies and are thus exposed to high-frequency market risks [23]. UHFD serve as the basic building blocks for such analysis. Consequently, analysis and modeling of UHFD has gained a lot of momentum in the academic sector and has become one of the key focal points of research in the fields of finance, financial econometrics, and statistics. Most of the financial studies in these areas are primarily concentrated to two fields, time-series analysis of fixed-resolution UHFD and inferences based on diffusion processes [24]. However, active research in these fields spans several areas like trading, microstructure theory, option pricing, risk management, and trading strategies like statistical arbitrage [25].

In the past several years, much of the research in the financial field has focused on measuring and forecasting volatility, and even today, volatility of asset returns remains one of the most important elements in finance. In the financial jargon, volatility is commonly defined as the dispersion in asset price movements over a certain time period. Measuring and forecasting volatility of financial asset returns plays a crucial role in several areas, including risk management, asset allocation, options, derivatives pricing, investment analysis, and portfolio management. In addition, financial market volatility has a direct impact on policy making, facilitated primarily due to its ability to gauge market and economic vulnerability, and plays a vital role in seizing profitable investment opportunities with optimal to high returns on risk trade-offs [26,27]. For instance, different market participants usually have different market expectations in terms of expected returns on investments and risks. Their investment decisions mainly revolve around the strategies that maximize their returns on investments while minimizing the associated risks. Among other market participants, equity and derivative traders use several volatility measures and forecasts, usually based on historical data, as proxies to assess risks associated with the different asset classes [28]. Today, volatility unarguably serves as one of the key measures of financial risk and drives the hedging and pricing of options or other securities and construction of optimal portfolios [27]. Correct estimation or assessment of financial risks helps reduce the probability of failures during extended periods of financial distress and is critical for the viability and stability of the financial system [29]. Accurate financial market risk assessments are much more important during periods of financial turmoil, and hence high volatility, due to the extensive risk of global financial instability [30]. During recent crisis events in the financial markets, such as the US subprime mortgage crisis and
Europe’s ongoing sovereign debt fiasco, a large number of financial organizations did not successfully enforce the Basel Committee of Banking Supervision mandates with respect to their VaR estimations [31]. These adverse financial episodes further underscore the significance of extreme asset price movements, and hence, accurate volatility forecasts, for efficient risk mitigation through appropriate risk measurement and management measures that can adapt in accordance with the changing market environments. Due to these reasons, it is not surprising that volatility estimation and inference have drawn a lot of attention in recent years. Also, another reason that makes volatility a more popular measure comes from the fact that unlike daily returns, which offer very little explanatory power and hence are difficult to predict, volatility of daily returns, due to their relatively high persistence and conditional dependence, is nonstationary and predictable [32,33]. Volatility clustering effects were first reported by Mandelbrot [34], who observed that periods of low volatility were followed by periods of low volatility and vice versa. These factors explain the large number of contributions and research efforts dedicated to the measurement and prediction of volatility.

Despite several advancements, accurate measurement of ex post volatility remains nontrivial largely due to the fact that volatility is unobservable and cannot be directly observed from the data [27]. In the past, several models have been developed to forecast volatility. The very first popular volatility model was the autoregressive conditional heteroskedasticity (ARCH) model introduced by Engle [35]. Subsequently, Bollerslev [36] proposed a more generalized representation of the ARCH model, namely, a generalized autoregressive conditional heteroskedasticity (GARCH) model. These models have been followed by a large number of models based on different variations of the original ARCH model. These include parametric models like the Glosten–Jagannathan–Runkle GARCH (GJR-GARCH), fractionally integrated GARCH (FIGARCH), component GARCH (CGARCH), stochastic volatility (SV), and other models [37–43]. The ARCH class of models incorporates time variation in the conditional distribution mainly through the conditional variance and is geared towards capturing the heavy-tails and long-memory effects in volatility. The stochastic models usually based on the assumption that the repeated low-frequency observations of assets return patterns are generated by an underlying but unknown stochastic process [44]. These models have been successful in explaining several empirical features of the financial return series, such as heavy-tails and long-memory effects in volatility. Since their introduction, an extensive literature has been developed for modeling the conditional distribution of stock prices, interest rates, and exchange rates [45]. This class of models has been extensively used in the literature to capture the dynamics of the volatility process. Until recent years, the ARCH model and its variants had been used by many in the field to model asset return volatility dynamics for daily, weekly, or higher-interval data across multiple asset classes and institutional settings [37,46]; however, despite their provably good in-sample forecasting performance, their out-of-sample forecasting performance remains questionable. Many studies in the past have reported insignificant forecasting ability of this class of models [47–50] based on low correlation coefficients in the assessments. These models usually forecast volatility based on low-frequency asset return data at daily or longer time
horizons. The use of low-frequency observations in these models results in a significant amount of information loss, which, for some asset classes like equities, can account for more than 99% of the available data [51]. Also, such models usually use daily squared returns as the highest order of granularity to compute estimates of “true volatility.” Since daily squared returns are calculated from closing prices, they fail to capture the intraday price fluctuations or interdaily volatility movements [52]. This results in less accurate volatility estimates due to reduced statistical efficiency. Moreover, models based on low-frequency data are often severely impacted by structural breaks and hence cannot appropriately adjust to the changing drifts in financial markets [53]. Most models in the past had been developed and tested by employing data sets that represented only a small subset of the available trading data [51]. In addition, today, many market participants, such as the high-frequency traders, noise traders, and speculators in future markets, deploy trading strategies that are often very sensitive to even the slightest changes in the asset prices. For these traders, even a minor intraday price fluctuation can result in significant trading volumes. Clearly, low-frequency asset returns–based volatility or risk models lack a tremendous amount of information central to the trading strategies of these market participants.

17.5.2 UHFD (Big Data)–Driven Research

In recent years, the increased availability of ultrahigh-frequency trading data has spurred strong growth in the development of techniques that exploit tick-level intraday price data to better estimate volatility forecasts and overcome at least some of the limitations inherent in the traditional low-frequency–based systems. UHFD carry a lot more information compared to their low-frequency counterpart and are not only useful for measuring volatility but also vital for model estimation and forecast evaluation [54]. Availability of high-frequency data coupled with recent technological advancements has made analysis of large-scale trading data more accessible to market participants [25]. A lot of research efforts today are dedicated to the use of high-frequency data for measuring volatility [44,55]. Access to UHFD within academia has led to significant recent progress in the field of econometrics of financial volatility [56]. Recently, much of the interest has shifted from low-frequency data–based volatility modeling to UHFD-based volatility models and even model-free approaches. Among other things, UHFD provide an appropriate basis for empirical tests of market microstructure theories, improve statistical efficiency, and have been shown to be extremely significant in the meaningful ex post evaluation of daily volatility forecasts [57,58].

In the last few years, modeling of financial data observed at high frequencies has become one of the key research areas in the field of financial econometrics. Consequently, several volatility measures and models have been proposed that exploit the power of UHFD for improved volatility forecasts. The most popular measure in this group is the realized volatility estimator [53,59,60] that utilizes intraday returns to measure the true volatility, thereby allowing the measurement of a latent variance process. Realized volatility is defined as the sum of squared intraday returns. It has been shown theoretically that realized volatility converges to the true integrated variance at ultrahigh frequencies,
that is, as the length of the intraday intervals approaches zero [61]. Specifically, it has
been shown empirically that the sum of squared high-frequency intraday returns of an
asset can be used as an approximation to the daily volatility [52,53,61–67]. This quadratic
variation is known as the estimator of the daily integrated volatility. Moreover, the fore-
casting performance of this estimator has been shown to be superior compared to the
performance of standard ARCH-type models [52]. Based on the theoretical and empirical
properties of realized volatility, several other studies also confirm that precise volatility
forecasts can be obtained using UHFD [61,63,68,69]. However, the experimental results
do not exactly match with the empirical or theoretical justifications, which mainly rely
on the limit theory and suggest that by increasing the observation frequency of asset
returns and representing the true integrated volatility of the underlying returns process
via realized volatility, one can obtain more efficient, less noisy estimates in comparison
to the estimate obtained using low-frequency data, such as the daily data [52,64,70–72].
This discrepancy can be primarily attributed to the presence of market microstructure
noise. Market microstructure noise collectively refers to the vast array of frictions inher-
ent in the trading process. These imperfections in the trading process arise due to several
factors, such as the bid–ask bounce, infrequent trading, discreteness of price changes,
varied informational content of price changes, and so forth. For example, changes in the
market prices occur in discrete units. Specifically, the prices usually fluctuate between
the bid and the ask prices (the bid–ask spread) and multiple prices may be quoted simul-
taneously by competitive market participants due to the heterogeneous market hypoth-
esis [73], thereby resulting in market microstructure frictions. At ultrahigh frequencies,
the problem gets worse because the volatility of the true price process shrinks with the
time interval, while the volatility of the noise components remains largely unchanged
[74]. Consequently, at extremely high frequencies, the observed market prices reflect val-
ues contributed largely by the noise component compared to the unobserved true price
component. As a result, the realized volatility measures become biased and thus are not
robust at ultrahigh frequencies when the price is contaminated with noise and the bias is
not accounted for during subsequent evaluations [52,75,76]. Market microstructure noise
effects and their impact in the high-frequency scenarios have been discussed and ana-
yzed in several studies [74–81].

Since the inception of the high-frequency–based realized volatility measure, several
alternative volatility estimation measures have evolved in an attempt to improve upon the
basic high-frequency–based realized volatility measure and cater to the growing demands
imposed by the availability of trading data at ultrahigh frequencies. For instance, in recent
years, many new methods have been developed to estimate spot or instantaneous volatility,
for example, volatility per time unit or per transaction, using high-frequency data. Spot
volatility is particularly useful to high-frequency traders, who often strategize using the
finest granularity of available tick data and to whose operations immediate revelation of
sudden price movements is critical. Various high-frequency data–based methods for spot
volatility estimation have been proposed in the literature, ranging from nonlinear state
space–based models to nonlinear market microstructure noise–based and particle filter–
based models [82–88].
Similarly, in an attempt to overcome the limitations inherent in the high-frequency data–based realized volatility measure, several high-frequency–based alternative measures have been proposed.

Many recent studies model the daily volatility of the underlying returns process using parametric models incorporating the realized volatility measure [89–92]. Different high-frequency variants of the realized volatility measure have been used to address a number of other problems. For example, the variants have been shown to be useful for reduced-form volatility forecasting [62,93,94] and formulation of highly informative and directly assessable distributional implications for asset returns [95]. Recently, a heterogeneous autoregressive model (HAR) was proposed [96]. This model estimates volatility by autoregressing volatilities realized over varying interval sizes (daily, weekly, monthly). Since its introduction, several variants of the base model have been developed, for example, HAR realized volatility model with jumps (HARRVJ) and HAR realized volatility model with discrete and continuous jump components (HARRVCJ). The latter variations of the model are based on the indications that relatively frequent jumps could be significant in the evolution of the price process and the volatility originating from jumps in the price level is less persistent than that generated by the continuous component of the realized volatility measure. The high-frequency–based volatility (HEAVY) method represents yet another example in this category. This method incorporates momentum and mean reversion effects of volatility and can adapt to structural breaks in the volatility process. It focuses on two main measures, namely, close-to-close conditional variance and conditional expectation of the open-to-close variation. It estimates both of these measures using the Gaussian quasi-maximum likelihood approach [97]. A realized range–based estimator is another example of HEAVY estimators. It is similar to the realized volatility estimator but uses squared intraperiod high–low price ranges instead to mitigate the effects of bid–ask bias.

Notably, in the last couple of years, due to the ill effects caused by the market microstructure noise at ultrahigh frequencies, when the prices are contaminated by market microstructure noise, academic interest has largely shifted towards the development of robust realized volatility measures that are immune to such noise. A number of variants of the realized volatility measure have been proposed, ranging from optimized sampling frequency–based estimators [75,98], subsampling-based methods [76,77], and Fourier volatility–based methods to realized kernel–based estimators [78], wavelet realized volatility methods [99], and preaveraging methods [100,101]. Most of the methods can be coarsely classified into two main groups. The first group primarily focuses on determining the optimal sampling frequency for volatility estimation in the presence of noise, and the second group comprises bias correction methods that accommodate the microstructure noise effects and validate the use of higher sampling frequencies. Specifically, in the former approach, the microstructure noise effects are alleviated by sampling prices sparsely at coarser intervals. With high-frequency data, even sampling intervals of 5 min can result in a significant amount of potentially useful data points getting discarded, especially for highly traded assets. This results in discretization error and reduced statistical efficiency [65,102,103]. The latter group focuses on utilizing all the available high-frequency data. For example, the two- and multi-time-scale realized volatility estimators utilize two or
more time scales respectively, to estimate integrated volatility. These estimators have been shown to be robust to time-series-dependent noise [14] and better than the classic realized volatility measure. Other estimators like the kernel-based estimators have also been shown to capture important characteristics of market microstructure noise and outperform the classic realized volatility measure [80]. UHFD also allow learning about jumps in the price process. Jumps represent discontinuous variations or movements in the price process that are totally incompatible with the observed volatility [104]. In the last couple of years, several parametric and nonparametric methods have been developed that distinguish between jumps and the continuous price process to estimate time-varying volatility robustly to jumps [105–110]. To date, bipower variation (BPV) or its variants remain to be among the dominant methods used to model jumps. These methods are based on the sums of powers and products of powers of absolute returns and have been shown to be robust to rare jumps in the log-price process [56,109,110].

All the high-frequency data–based methods discussed are mostly applicable to the univariate or single-asset volatility estimation scenarios. Nowadays, volatility estimation of the multiple asset scenarios is becoming increasingly important. Also, precise estimation of the covariance matrix of multiple asset returns is central to many issues in finance such as portfolio risk assessment and asset pricing. The availability of UHFD has spurred its use in many recent covariance asset estimation methods. This is mainly because UHFD better reflect the underlying assets returns processes because of better statistical efficiency and can greatly improve the accuracy of the covariance matrix estimates. However, as discussed earlier in this section, the use of UHFD can introduce noise as a result of market microstructure frictions. Estimation methods in multiple asset scenarios face additional difficulties due to nonsynchronous trading issues. These issues arise because the transactions for different assets occur at different points in time, are random, and are thus nonsynchronous. Due to this mismatch in the time points of the recorded transactions, returns sampled at regular intervals in calendar time will correlate with the previous or successive returns on other assets even in the absence of any underlying correlation structure [111,112]. This, called the Epps effect, causes the covariance estimator to be biased towards zero with increasing sampling frequencies [113]. Two key approaches have generally been used in the past to address these issues. One attempts to reduce the microstructure noise effects through the use of lead and lag autocovariance terms in the realized covariance estimator based on synchronized returns, whereas the other produces unbiased estimates of the covariance matrix by using the cross-product of all fully and partially overlapping event-time returns [114]. In recent years, several high-frequency data–based volatility estimation methods have been proposed for multiple asset scenarios, starting with the realized covariance estimator that is basically the sum of cross-products of intraday returns [115]. Like the realized volatility measure, this measure also suffers from the impact of microstructure noise [114,116–120]. Since then, many methods have been developed to deal with the inherent problems in the multiple asset settings, like the market microstructure noise and bias due to nonsynchronicity [114,116,118,119,121–123]. Many market participants often need to estimate matrices comprising a large number of assets using high-frequency data. However, many existing estimators can only be used for a small number of asset classes
and become inconsistent as the size of the matrix becomes closer to or exceeds the sample size \([124]\). A few recent methods suggest different ways to estimate large volatility matrices, some of which are robust to the presence of microstructure noise and nonsynchronicity effects. They incorporate different schemes, ranging from the use of factor models and low-frequency dynamic models to pairwise and all-refresh time schemes \([112,124–126]\).

### 17.5.3 UHFD (Big Data) Implications

UHFD-based estimators have important implications in many areas of finance, especially risk assessment and management or single assets and portfolios. For example, several studies have utilized the realized volatility estimators based on UHFD for VaR forecasting \([127–131]\). Realized volatility–based estimators have also been shown to explain variations in the cross-sectional and temporal behavior of risk premiums, when used in conjunction with the implied volatility measures \([70]\). Forecasting performance of realized volatility–based measures and its variants, such as the realized range–based estimators, has been shown to be favorable in many other risk measurement studies involving real high-frequency data from the NYSE and the Standards and Poors (S&P) 500 stock index \([55,90,132,133]\). Other studies based on high-frequency measures, such as BPV, reveal the significance of jumps in the price process evolution of commonly held assets \([109,134]\). Evidently, the growing prevalence of HFT has necessitated the utilization of UHFD for information gains in areas like short-term risk assessments. Unarguably, compared to the low-frequency–based risk models, high-frequency data yield relatively more precise and more adaptive short-term risk models. Besides benefiting high-frequency traders, UHFD have also been shown to benefit low-frequency traders like the traditional low-frequency risk and portfolio managers, who can benefit by utilizing high-frequency dynamic factor exposure estimates to hedge short-term risk factor exposures \([135]\). The use of UHFD is also being increasingly explored in other areas like asset pricing and governance \([136–139]\).

### 17.5.4 UHFD (Big Data) Challenges

Clearly, the availability of UHFD has led to significant advancements in the field of finance and has made it possible for empirical researchers to address problems that cannot be handled using data collected at lower frequencies. However, various factors, such as the underlying market structure, market dynamics, internal process flows, trading frequency, and so forth, present several difficulties in the effective utilization of the vast amounts of UHFD available today. Moreover, the storage, analysis, and management of near-continuous data present significant new challenges. For instance, a highly traded stock can easily result in several millions of data points per year and require several hundreds of gigabytes of storage for a 12-month period \([139]\). Recorded trade data at ultrahigh frequencies generally have numerous data errors. Usually, the reason for the higher number of errors is attributed to the large volumes of trading. Data errors could originate from various sources. For example, HFT data could have isolated bad ticks, multiple bad ticks in succession, wrong ticks, decimal errors, transposition errors, typing errors, and reporting errors like duplicate trades or delayed trades \([25,140]\). These errors necessitate the need for accurate data filtering mechanisms that can identify and resolve such errors and can help convert
the data into a usable form [25,140]. Also, as previously discussed in Sections 17.5.1–17.5.4, the usage of UHFD generally requires a trade-off between precision of the estimation technique and bias induced by the market microstructure noise effects. The bias increases with the increase in the sampling frequencies, which renders the estimator less accurate at high sampling frequencies. Organizational structure and institutional evolution of the equity markets further exacerbate such errors in the UHFD [14]. These issues represent some of the challenges that are being faced by both the researchers and practitioners with regard to UHFD. Despite several advancements in this field, a lot more remains to be done to efficiently utilize the available UHFD and extract the best value out of them.

17.6 SUMMARY

The financial industry has always been a data-intensive industry. Recent technological advancements coupled with several other factors like changing customer preferences and changing business needs have led to the generation and consumption of prolific amounts of data. Several changes in the last couple of years, driven by the confluence of factors like the escalating regulatory pressures, ever-increasing compliance requirements, regulatory oversight, global economic instability, increasing competition in the global markets, growing business demands, growing pressures to optimize capital and liquidity, and so forth, are forcing the financial organizations to rethink and restructure the way they do business. Also, traditional data management practices prevalent in finance can no longer effectively cope with the ever-increasing, huge, and rapid influx of heterogeneous data originating from a wide range of internal processes and external sources, including social media, blogs, audio, and video. Consequently, a growing number of financial institutions are resorting to Big Data to strategize their business decisions based on reliable factual insights supported by real data rather than just intuition. Increasingly, Big Data is being utilized in several areas such as investment analysis, econometrics, risk assessment, fraud detection, trading, customer interactions analysis, and behavior modeling. Efficient utilization of Big Data has become essential to the progress and success of many in this data-driven industry. However, Big Data by itself does not hold much value, and not all of it may be useful at all times. To gain relevant insights from the data, it is very important to deploy efficient solutions that can help analyze, manage, and utilize data. Many solutions have already been deployed in the financial domain to manage relevant data and perform analytics on them. Despite such advancements, many in the industry still lack the ability to address their Big Data needs. This is most likely due to the fact that different organizational units usually have different domain-specific requirements and, hence, solution specifications. So, it is highly unlikely that a solution deployed by one unit would be equally useful to others. The chapter described the impact of Big Data on the financial industry and presented some of the key transformations being driven by the data today. The availability of UHFD has resulted in significant advancements in the field of finance and has made it possible for empirical researchers to address problems that could not be handled using data collected at lower frequencies. Besides showcasing the impact of Big Data in the industrial sector, the chapter also highlighted the Big Data–driven progress in research in the fields of finance, financial econometrics, and statistics. The chapter exemplified some of the key developments in this area with special focus on financial risk
measurement and management through the use of UHDF-based volatility metrics. In recent years, the increased availability of ultrahigh-frequency trading data has spurred strong growth in the development of techniques that exploit tick-level intraday price data to better estimate volatility forecasts and overcome some of the limitations inherent in the traditional low-frequency–based systems. Availability of high-frequency data coupled with recent technological advancements has made analysis of large-scale trading data more accessible to market participants. Ultrahigh-frequency–based estimators have important implications in many areas of finance, especially risk assessment and management or single assets and portfolios. Among other things, such data have been shown to be extremely useful in the meaningful ex post evaluation of daily volatility forecasts. However, various factors, such as the underlying market structure, market dynamics, internal process flows, trading frequency, and so forth, present several difficulties in the effective utilization of the vast amounts of high-frequency data available today. Particularly, the market microstructure noise effects, such as those due to bid–ask bounce and infrequent trading, introduce a significant bias in the estimation procedures based on high-frequency data. The chapter also touched upon the trade-off that is often required between the precision of the estimation technique and bias induced by the market microstructure noise effects. The availability of Big Data in the financial domain has opened up new avenues for innovation and presented immense opportunities for growth and sustainability. Despite significant progress in the development and adoption of Big Data–based solutions in the field, a lot more remains to be done to effectively utilize the relevant data available in this domain and extract insightful information out of it. Compared to the promise Big Data holds in this domain and its potential, the progress in this field is still in its nascent stages, and a lot more growth in this area remains to be seen in the coming years.

REFERENCES

2. Connors, S., Courbe, J., and Waishampayan, V., Where have you been all my life? How the financial services industry can unlock the value in Big Data. PwC FS Viewpoint, October 2013.


