A Comprehensive Study of Approximate Query Processing Techniques for Big Data Analytics

Nguyen Minh Quan
Department of Computer Science and Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam
nguyen.minh.quan@hust.edu.vn
Vo Hung Cuong
Department of Information Systems, Vietnam Korea University of Information and Communication Technology
vhcuongdn@gmail.com

Abstract

In an era of exponential data growth, the need for fast and scalable query processing algorithms has increased. Traditional approaches, once a mainstay of data analysis, increasingly fail to provide quick results when querying across large data sets. Then Approximate Query Processing (AQP) comes into play, a hope for big data analysis. AQP points to a strategic pivot that prioritizes speed of query response over absolute precision. This shift has significant implications for real-time and interactive analytics, where speed is of the utmost importance. Our AQP trip is a guided tour around the planet. It analyzes numerous AQP strategies, offers a structured overview of their inner workings, classifies them according to important characteristics and critically evaluates their strengths and weaknesses. Additionally, we bridge theory and practice by highlighting specific use cases where AQP has made a lasting impression. From e-commerce and healthcare to finance and science, AQP has revolutionized data-driven decision-making across multiple industries. The aim of this study is to provide researchers, data scientists and industry experts with the knowledge and insights required to realize the potential of AQP in the era of big data. It provides a glimpse into the future of data analytics by capturing the core of AQP's role in balancing data accuracy and query performance.

Keywords:
- Big Data Analytics
- Approximate Query Processing (AQP)
- Query Response Time
- Data Accuracy
- Scalable Query Processing
- Real-time Analysis
- Interactive Analysis

Excellence in Peer-Reviewed Publishing:
QuestSquare

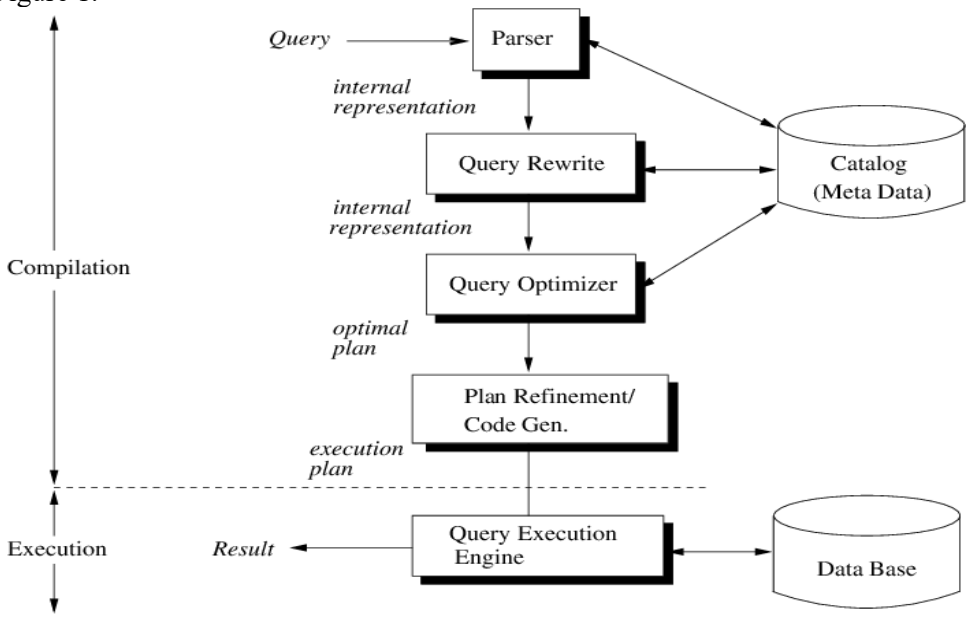
Introduction

Big data has brought about a fundamental shift in how we view and utilize information. In the current digital era, data is produced at an unprecedented velocity and from a vast array of sources, including social media interactions, sensor data, financial transactions, and scientific studies. This data explosion has resulted in an exponential growth in the volume, velocity, and variety of available information. Nonetheless, this data deluge presents a formidable challenge: the effective management and interpretation of massive data volumes. Traditional query processing

methods, which have served as the foundation of data analysis for decades, find themselves strained under the sheer weight of big data. These conventional techniques, designed for relatively modest data volumes, often falter when confronted with the immense scale of contemporary datasets [1]. The result is a significant lag in query response times, rendering them inadequate for real-time or interactive analysis requirements. As a consequence, there has been a pressing need to explore novel approaches that can balance the ever-increasing demand for rapid data analysis with the inherent limitations of traditional query processing.

Approximate query processing (AQP) has emerged as a promising and pragmatic solution to this conundrum. Rather than striving for precise but time-consuming query results, AQP techniques opt for a trade-off between accuracy and response time. By embracing approximations, AQP allows queries to be answered more swiftly, making it an ideal fit for scenarios where timely insights are paramount. In essence, AQP acknowledges that in many practical situations, the absolute precision of query results may not be essential, and that some degree of approximation is not only acceptable but also beneficial. In this article, we embark on an extensive exploration of the universe of approximate query processing techniques within the context of big data analytics [2]. We aim to provide a comprehensive study that delves into the intricacies of AQP, examining its various facets, applications, and implications. Our investigation is guided by a desire to unravel the potential of AQP as a fundamental tool for modern data analytics.

Figure 1.



The following sections will delve into the background of big data analytics and approximate query processing, elucidate the diverse approaches used in AQP, classify these techniques based on key attributes, and assess their advantages and limitations [3]. We will also present real-world use cases where AQP has proven to be transformative. Furthermore, this article will compare AQP with traditional query processing, assess its performance, and speculate on future trends and research directions in the field. By the end of our journey, it is our aspiration that readers will

have a comprehensive understanding of AQP and its role in addressing the challenges posed by big data analytics [4].

Big data analytics is a transformative field within data management and analysis, driven by the proliferation of vast and complex datasets that defy conventional processing methods. Big data is typically characterized by the three V's: volume, velocity, and variety.

Volume: Big data encompasses massive datasets that often range from terabytes to petabytes or even exabytes in size. This sheer volume of data necessitates novel approaches to storage and processing.
Velocity: Data is generated and updated at unprecedented speeds, with streaming data from various sources such as sensors, social media, and online transactions. The need for real-time or near-real-time analysis is paramount.
Variety: Big data comes in diverse formats, including structured data (e.g., databases), semi-structured data (e.g., XML, JSON), and unstructured data (e.g., text, multimedia). This heterogeneity poses significant challenges in terms of data integration and analysis.
Challenges associated with big data analytics are multifaceted. Processing and analyzing such data entail substantial computational resources and can lead to performance bottlenecks. Traditional database systems may struggle to provide timely responses to queries over these massive datasets, necessitating innovative solutions such as Approximate Query Processing (AQP).

Approximate Query Processing

Approximate Query Processing (AQP) is a pivotal technique born out of the need to strike a balance between query response time and accuracy when dealing with big data. The motivation behind AQP is to provide fast, yet reasonably accurate, query results, recognizing that exact solutions can be prohibitively slow for large datasets. Key attributes of AQP techniques include:
Trade-off Between Accuracy and Efficiency: AQP methods intentionally sacrifice query precision in favor of improved performance. By providing approximate answers, these techniques enable timely responses to queries that would otherwise be computationally infeasible.

Table 1: Classification of Approximate Query Processing Techniques

Technique Category	Accuracy vs. Efficiency Trade-off	Data and Query Type Support	Deployment Environment
Sampling-based Approaches	Emphasizes efficiency	Structured and unstructured	Cloud-based
Sketch-based Approaches	Balance between accuracy & efficiency	Structured data	On-premises
Wavelet-based Approaches	Emphasizes accuracy	Unstructured data	
Histogram-based Approaches			

Machine Learning-based Approaches			
Query Rewriting Approaches			

Sample-based Processing: Many AQP techniques rely on random or systematic sampling of data, allowing queries to be executed on a smaller subset of the dataset. This reduces the computational burden while still providing reasonably accurate results [5].

Summary Statistics: AQP often employs summary statistics or data synopses to represent the underlying data distribution. These synopses are used to estimate query results, enabling rapid responses.

Query Relaxation: In some cases, AQP techniques may relax the query conditions or constraints, further enhancing query performance. While this may lead to less precise results, it is often an acceptable trade-off in big data scenarios.

Approaches to Approximate Query Processing

Approximate Query Processing (AQP) encompasses a diverse set of techniques designed to strike a balance between query accuracy and response time, making them particularly well-suited for the challenges presented by big data analytics. In this section, we delve into various approaches to AQP, each with its unique set of strategies and methodologies.

Sampling-based Approaches: Sampling has long been a fundamental technique in statistics and data analysis. In the context of AQP, random sampling involves selecting a subset of data points randomly, which can provide a representative view of the entire dataset. Stratified sampling, on the other hand, divides the data into distinct strata or groups and samples from each stratum proportionally. Adaptive sampling techniques dynamically adjust the sampling rate based on query characteristics, ensuring that more accurate estimates are obtained for critical queries while optimizing for efficiency in less crucial cases.

Sketch-based Approaches: Sketches are compact data structures that offer an approximate representation of data, allowing for efficient estimation of query results. Count-min sketch, for instance, employs a matrix of counters to approximate frequency counts of elements in a dataset, proving useful in applications like frequency analysis and anomaly detection [6]. HyperLogLog is specifically designed for cardinality estimation, providing an efficient means to determine the distinct number of items in a dataset. Theta sketch extends the concept of sketches to capture intersections and unions of sets, making it valuable in scenarios involving set operations.

Wavelet-based Approaches: Wavelet-based AQP leverages wavelet transforms to analyze and approximate data. The Haar wavelet, a simple and efficient wavelet transform, can be used to decompose data into its high and low-frequency components, aiding in various signal processing and data analysis tasks. The Discrete Wavelet Transform (DWT) is a more advanced technique that decomposes data into different scales and orientations, enabling a multi-resolution analysis of the dataset. These approaches are particularly valuable when dealing with time-series data or images.

Histogram-based Approaches: Histograms are another well-established tool in data analysis. Equi-depth histograms divide data into equal-sized bins, summarizing data distribution and enabling approximate query processing for aggregate functions such as COUNT and SUM. Wavelet histograms combine the benefits of wavelet transformations with histogram techniques to provide a scalable and accurate representation of data distribution, making them suitable for a wide range of query types [7]. Machine Learning-based Approaches: Machine learning models have been increasingly integrated into AQP techniques. Regression models are employed to predict query results based on historical data, allowing for quick estimations of aggregations or other query functions. Classification models, on the other hand, categorize data into different classes or clusters, facilitating approximate query results by focusing on representative data points rather than the entire dataset [8].

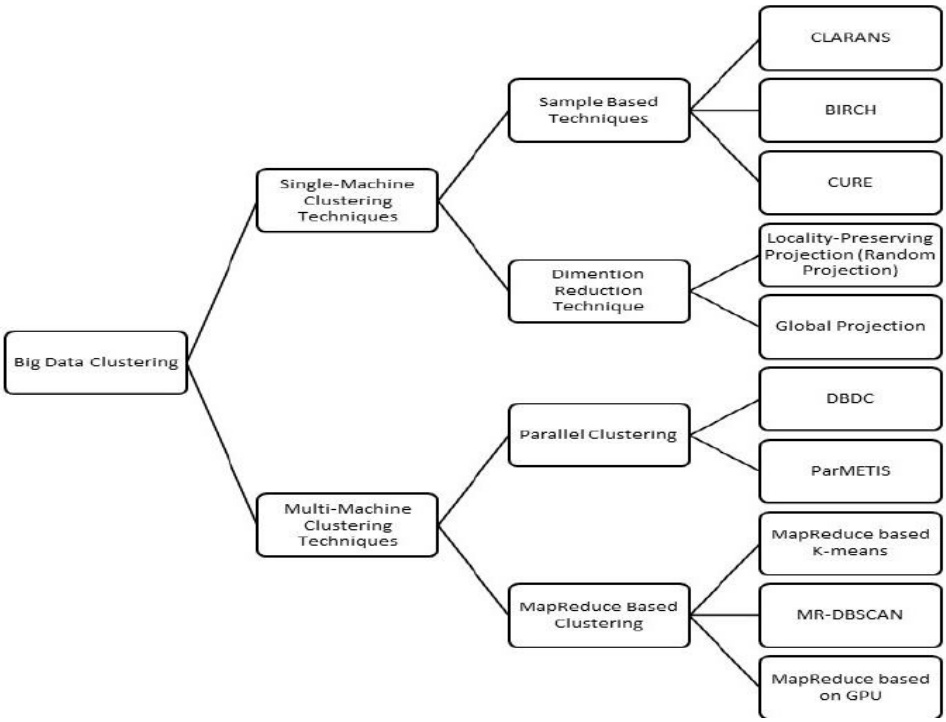
Query Rewriting Approaches: Query rewriting strategies aim to modify queries to leverage summary statistics or relax query constraints, thereby improving query response times while maintaining a certain level of accuracy [9]. These approaches are valuable in scenarios where stringent accuracy requirements can be relaxed without significantly impacting the usefulness of query results. Techniques for rewriting queries often involve altering aggregation functions, using materialized views, or employing probabilistic methods to approximate results.

Classification of AQP Techniques

Classification of Approximate Query Processing (AQP) Techniques involves several key dimensions that help categorize and understand the diverse landscape of AQP methodologies. These dimensions play a crucial role in determining which AQP approach is best suited for a specific use case.

Accuracy vs. Efficiency Trade-off: One fundamental classification criterion for AQP techniques revolves around the trade-off between query accuracy and processing efficiency. On one end of the spectrum, there are AQP techniques that prioritize accuracy. These methods aim to provide query results that closely resemble the exact results, ensuring a high level of fidelity in the approximate output. Such techniques are particularly useful in scenarios where data accuracy is paramount, such as scientific research or healthcare applications [10].

Figure 2.



On the other hand, there are AQP techniques that prioritize efficiency. These approaches optimize query processing speed and resource utilization, often at the expense of query result accuracy. They are well-suited for real-time analytics, where getting rapid insights from large datasets is crucial, and minor deviations in query results are tolerable, as seen in e-commerce recommendation systems or real-time financial market analysis.

Data Type and Query Type: Another crucial dimension for classifying AQP techniques revolves around the nature of the data being processed and the types of queries being executed. Firstly, AQP techniques need to account for the structure of data. Some are specialized in handling structured data, such as relational databases, while others excel at processing unstructured or semi-structured data, often found in text documents, sensor data, or social media content. The ability to adapt to the data structure is vital to ensure the effectiveness of AQP techniques [11]. Furthermore, AQP techniques differ in their support for various query types. Different queries, such as aggregations, joins, or range queries, have distinct characteristics and requirements. A comprehensive AQP system should be versatile enough to accommodate a wide array of query types efficiently. For example, in an e-commerce setting, supporting complex joins to analyze customer behaviors and purchase patterns can be just as important as handling simple aggregation queries.

Deployment Environments: The choice of where and how AQP techniques are deployed is another classification dimension. Modern data architectures offer various deployment options, and AQP solutions need to adapt accordingly. Cloud-based AQP solutions are designed to leverage the scalability and flexibility of cloud computing platforms, making them well-suited for organizations that rely on cloud infrastructure. These solutions can seamlessly scale resources up or down based on demand, accommodating fluctuating workloads and ensuring cost-efficiency [12]. On the other hand, on-premises AQP solutions are tailored for organizations that prefer to maintain

full control over their infrastructure. These solutions are typically deployed within an organization's own data centers or private cloud environments. They offer more extensive customization options and allow organizations to manage their hardware resources directly, which can be critical for industries with stringent security or compliance requirements, such as finance or healthcare.

Advantages and Limitations

Approximate Query Processing (AQP) techniques offer several distinct advantages in the realm of big data analytics. Firstly, they are known for their ability to significantly reduce query response times. In traditional query processing, complex queries over massive datasets can take an impractical amount of time to execute. AQP methods, by providing approximate answers, enable users to obtain rapid responses, making them particularly well-suited for applications where timely insights are crucial. Another key advantage of AQP techniques is their scalability to large datasets. As data continues to grow exponentially, the traditional processing approaches often face bottlenecks due to resource limitations. AQP methods, on the other hand, can efficiently handle vast amounts of data without a substantial increase in computational resources, making them highly adaptable to the challenges of big data analytics.

Table 2: Advantages and Limitations of Approximate Query Processing

Aspect	Advantages	Limitations
Query Response Time	Significantly reduced response times	Loss of query precision
Scalability	Scalable to large datasets	Lack of accuracy guarantees
Resource Efficiency	Efficient utilization of computational resources	Sensitivity to data distribution

Additionally, AQP techniques are known for their resource efficiency. By reducing the amount of computation required to answer queries, they can lead to substantial cost savings in terms of both hardware and energy consumption. This efficiency is of paramount importance in cloud computing environments, where minimizing resource usage translates to economic benefits. However, it is essential to recognize that AQP techniques come with their own set of limitations and challenges. One of the primary concerns is the potential loss of query precision. Since AQP methods provide approximate answers, there is a trade-off between accuracy and query response time. While these techniques are invaluable for applications where near-real-time responses are required, they may not be suitable for scenarios where precision is of utmost importance.

Another limitation is the lack of clear query accuracy guarantees. AQP methods often provide probabilistic answers, making it challenging to determine the exact level of accuracy one can expect. This uncertainty can be a hindrance in critical decision-making processes where precise information is indispensable. Furthermore, AQP techniques are sensitive to data distribution. The quality of approximate results can vary depending on how the data is distributed. In cases where data follows an irregular or skewed distribution, AQP methods may struggle to provide accurate approximations, leading to potential errors in analytical outcomes.

Real-world Use Cases

Real-world use cases of approximate query processing (AQP) techniques span a wide spectrum of industries and academic disciplines, showcasing the versatility and applicability of these methods.

Industry Applications

E-commerce: In the realm of e-commerce, where massive volumes of transactional and user interaction data are generated daily, AQP techniques are invaluable. Businesses employ AQP to swiftly analyze customer behavior, optimize inventory management, and make real-time pricing decisions. For instance, AQP can help e-commerce platforms quickly identify trending products, personalize recommendations, and enhance the overall shopping experience.

Healthcare: The healthcare industry faces the challenge of managing and analyzing vast amounts of patient data, including electronic health records, medical images, and genomic information. AQP plays a pivotal role in expediting the processing of medical data for clinical decision support, disease prediction, and drug discovery. By leveraging AQP, healthcare professionals can obtain rapid insights into patient histories and medical research can be accelerated [13].

Financial Services: Financial institutions deal with massive datasets in activities such as fraud detection, risk assessment, and algorithmic trading. AQP methods aid in the efficient analysis of financial data, allowing for faster detection of anomalies or potential fraud, as well as optimizing trading strategies in real time. Timely decision-making is paramount in this industry, making AQP techniques a critical component [14].

Research and Academic Applications: Scientific Data Analysis: Researchers in various scientific fields, including astronomy, genomics, and climate science, are confronted with astronomical volumes of data. AQP enables them to perform quick exploratory data analysis and gain insights into complex phenomena. For instance, astronomers can use AQP to process large-scale sky studys, while genomics researchers can analyze genomic sequences efficiently. Social Network Analysis: With the explosive growth of social media platforms, social network analysis has gained prominence in understanding human behavior, influence dynamics, and information diffusion. AQP techniques help researchers analyze massive social graphs, identifying influential nodes, tracking viral content, and studying online communities. These insights are invaluable for marketing, public opinion analysis, and social sciences research.

Recommendation Systems: Recommendation systems are pervasive in today's digital landscape, from e-commerce platforms to content streaming services. AQP can accelerate recommendation algorithms, allowing platforms to serve personalized content and product suggestions to users in real time. This enhances user engagement and drives revenue for businesses while also improving the user experience.

Comparison with Traditional Query Processing

When considering the realm of approximate query processing (AQP) in contrast to traditional query processing, a critical aspect that comes into focus is performance evaluation. Performance evaluation serves as the litmus test for the effectiveness and

efficiency of any query processing approach, and in the context of AQP, it plays a pivotal role in determining the suitability of these techniques for specific applications [15].

Table 3: Real-world Use Cases of Approximate Query Processing

Industry/Application	Specific Use Case
E-commerce	Personalized product recommendations
Healthcare	Real-time patient monitoring and analysis
Financial Services	Fraud detection and prevention
Scientific Data Analysis	Climate modeling and simulation
Social Network Analysis	Influence propagation analysis
Recommendation Systems	Movie or product recommendations

Benchmarks and Metrics: To gauge the performance of AQP techniques accurately, researchers and practitioners have developed a range of benchmarks and metrics. These benchmarks are designed to mimic real-world scenarios, ensuring that AQP methods are tested under conditions that align with the challenges presented by big data analytics. Common benchmarks include TPC-H for decision support systems and TPC-C for transactional workloads. Metrics encompass a wide array of parameters such as query response time, query accuracy, resource utilization, and scalability. Researchers employ these benchmarks and metrics to assess how AQP techniques compare against traditional query processing methods in terms of speed and resource efficiency.

Comparative Analysis: Comparative analysis between AQP and traditional query processing is crucial for understanding the trade-offs involved. While traditional query processing guarantees precise and accurate results, it often falls short when it comes to handling vast volumes of data. AQP, on the other hand, sacrifices some degree of accuracy for significant gains in query response time and scalability. Comparative studies delve deep into these trade-offs, shedding light on when and where AQP is advantageous and when the precision of traditional query processing cannot be compromised [16]. These analyses inform decision-makers about the applicability of AQP in specific use cases, helping them make informed choices based on their requirements.

Future Trends and Research Directions

In the ever-evolving landscape of big data analytics, the future holds exciting prospects and challenging research directions for approximate query processing (AQP) techniques. To stay at the forefront of this field, researchers and practitioners are actively exploring emerging AQP technologies that promise to address critical aspects of data analysis.

Emerging AQP Technologies

Integration with Stream Processing: As real-time data streams become increasingly prevalent, integrating AQP with stream processing is a natural progression. AQP techniques need to adapt to the dynamic and ever-changing nature of streaming data, where traditional batch processing is inadequate. This integration aims to provide timely approximations of query results over continuous data streams, opening new possibilities for applications in fields like IoT, fraud detection, and network monitoring.

AQP for Machine Learning Workloads: The convergence of AQP and machine learning (ML) is an exciting development. AQP can enhance ML workflows by

providing approximate insights into large datasets, reducing the computational burden of training complex models [17]. Researchers are exploring ways to incorporate AQP methods into various stages of the ML pipeline, from data preprocessing to model evaluation. This fusion can lead to more efficient and scalable ML solutions [12].

Privacy-preserving AQP: In an era of heightened concerns about data privacy and security, preserving the confidentiality of sensitive information during query processing is paramount. Privacy-preserving AQP techniques are under active investigation. These methods aim to provide approximate query results while ensuring that the underlying data remains encrypted or anonymized, protecting individual privacy and complying with stringent data protection regulations.

Open Research Challenges

Quality-aware AQP: A significant challenge lies in developing AQP techniques that are not only efficient but also capable of providing reliable quality guarantees [18]. Striking the right balance between query response time and result accuracy is essential. Researchers are actively exploring methods to quantify and control the error introduced by AQP techniques, enabling users to make informed decisions about the trade-offs between speed and precision.

AQP in Distributed and Federated Settings: As data sources become distributed across various platforms and organizations, AQP in distributed and federated environments is gaining prominence. Enabling approximate query processing across multiple, potentially decentralized data repositories requires novel solutions for coordination, synchronization, and data exchange. This research direction aims to make AQP feasible and efficient in the context of modern data ecosystems.

AQP in the Era of Edge Computing: With the proliferation of edge computing devices and the demand for real-time decision-making at the edge, AQP faces new challenges. Edge environments are resource-constrained and often subject to intermittent connectivity. Developing AQP techniques that can operate effectively in these conditions while delivering timely insights is a pressing research concern. These solutions must also account for the heterogeneous nature of edge devices and data sources [19].

Conclusion

With the advent and use of approximation query processing (AQP) techniques, the field of big data analytics has seen a significant revolution. According to the findings of this exhaustive investigation, AQP has become a vital instrument for efficiently managing the enormous data quantities that characterize the modern data landscape. The trip through the numerous dimensions of AQP has illuminated its significance, adaptability, and capacity to meet the persisting issues provided by big data. This study has highlighted the vital importance of AQP in the field of big data analytics [20]. Traditional query processing approaches have struggled to keep up with the ever-increasing amount, velocity, and variety of data being generated and stored [21]. AQP is a realistic solution that enables enterprises to extract valuable insights from their data at a rate commensurate with the requirements of the current fast-paced business climate. It achieves this by sacrificing some query precision for faster query response times.

Within the field of AQP, a variety of techniques have been investigated. It has been demonstrated that sampling-based approaches, such as random sampling and adaptive

sampling, can reduce query processing times while maintaining a fair degree of precision. Sketch-based methods, including Count-Min Sketch and HyperLogLog, provide compact data structures for estimating query outcomes with low memory overhead. In order to summarize and approximate data distributions, wavelet-based approaches, such as Haar wavelet and Discrete Wavelet Transform, offer sophisticated solutions. For quick query processing, histogram-based techniques, such as Equi-depth and Wavelet histograms, provide robust data representations [22]. AQP techniques based on machine learning leverage the power of predictive models to estimate query results, whereas query rewriting techniques aim to adapt queries to work on summarized data, hence enhancing efficiency.

This taxonomy of AQP approaches has shown the broad set of alternatives available to practitioners and data scientists. One can select a method that corresponds to particular needs, whether it prioritizes accuracy, efficiency, or a combination of the two. In addition, the choice of AQP technique can be adapted to the nature of the data being analyzed, whether it structured or unstructured, as well as the type of queries being conducted, such as aggregations, joins, or sophisticated analytical processes. AQP approaches are also not limited to a single deployment environment. They may be readily integrated into cloud-based systems, allowing businesses to take advantage of the scalability and adaptability of cloud computing resources [23]. On the other hand, on-premises solutions are also accessible for companies who demand greater data and infrastructure management [24]. The benefits of AQP approaches are obvious. They provide significant reductions in query response times, enabling businesses to make data-driven choices in real-time or near-real-time. AQP approaches can manage data volumes that would make standard query processing problematic, making scalability to huge datasets a major advantage. In addition, the resource efficiency of AQP approaches is a key benefit, as they often demand less memory and processing power than their conventional counterparts. However, it is essential to recognize the limitations and difficulties of AQP. In exchange for faster query processing, query precision is diminished. While AQP approaches try for moderate precision, exact outcomes cannot be guaranteed. Depending on the chosen approach and parameters, the degree of precision varies, necessitating careful evaluation of the trade-offs for each given use case. Another difficulty is sensitivity to data distribution, as AQP approaches may perform differently depending on the nature of the data [25]. A poorly selected AQP strategy can result in erroneous results or unexpected behavior.

Through our investigation of real-world use cases, we've discovered that AQP approaches have applications in a variety of industries. In the e-commerce sector, AQP offers individualized product recommendations, as well as real-time inventory management. AQP is utilized by healthcare organizations for clinical decision support, patient data analysis, and optimal resource allocation. AQP benefits financial services in the areas of fraud detection, risk evaluation, and algorithmic trading. AQP is also utilized in the research and academic fields for scientific data analysis, social network analysis, and recommendation systems. These use cases demonstrate the adaptability and applicability of AQP across several fields [26]. AQP represents a paradigm shift

in data analysis when compared to conventional query processing approaches. Evaluations of performance reveal that AQP can greatly surpass conventional query response times, particularly when working with huge datasets. AQP-specific benchmarks and indicators enable enterprises to make informed decisions regarding the applicability of AQP methodologies for their particular analytical needs. This comparison demonstrates the revolutionary potential of AQP in the realm of big data analytics [27].

AQP is positioned for ongoing growth and change in the future. Emerging technology and research initiatives are expected to increase the effectiveness of AQP approaches. AQP's integration with stream processing will enable real-time analysis of data streams, creating new opportunities for Internet of Things, social media monitoring, and other applications. The use of AQP for machine learning workloads will enable data scientists to train models on approximation data, hence expediting the development of machine learning solutions [28]. In the era of data protection rules, privacy-preserving AQP approaches will become increasingly vital, allowing enterprises to preserve sensitive data while still gaining insights. However, there are still outstanding research obstacles to address. Quality-aware AQP strives to provide users greater control over the trade-off between precision and efficiency, enabling fine-grained modifications to query precision. As enterprises rely more on decentralized data storage and processing architectures, AQP in distributed and federated environments will become essential. As data processing moves closer to the data source, AQP in the era of edge computing presents distinct difficulties and opportunities, necessitating resource efficient AQP solutions customized to edge contexts.

References

[1] C. K. S. Leung, “Big data analysis and mining,” *architecture, mobile computing, and data analytics*, 2019.

[2] M. van Rijmenam, T. Erekhinskaya, J. Schweitzer, and M.-A. Williams, “Avoid being the Turkey: How big data analytics changes the game of strategy in times of ambiguity and uncertainty,” *Long Range Plann.*, vol. 52, no. 5, p. 101841, Oct. 2019.

[3] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, “Context-aware query performance optimization for big data analytics in healthcare,” in *2019 IEEE High Performance Extreme Computing Conference (HPEC-2019)*, 2019, pp. 1–7.

[4] C. K.-S. Leung and Y. Hayduk, “Mining Frequent Patterns from Uncertain Data with MapReduce for Big Data Analytics,” in *Database Systems for Advanced Applications*, 2013, pp. 440–455.

[5] P. Braun, A. Cuzzocrea, F. Jiang, C. K.-S. Leung, and A. G. M. Pazdor, “MapReduce-Based Complex Big Data Analytics over Uncertain and Imprecise Social Networks,” in *Big Data Analytics and Knowledge Discovery*, 2017, pp. 130–145.

[6] W. Shi, A. Zhang, X. Zhou, and M. Zhang, “Challenges and Prospects of Uncertainties in Spatial Big Data Analytics,” *Ann. Assoc. Am. Geogr.*, vol. 108, no. 6, pp. 1513–1520, Nov. 2018.

[7] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, “Federated query processing for big data in data science,” in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 6145–6147.

[8] A. Nassar and M. Kamal, "Ethical Dilemmas in AI-Powered Decision-Making: A Deep Dive into Big Data-Driven Ethical Considerations," *IJRAI*, vol. 11, no. 8, pp. 1–11, Aug. 2021.
[9] C. Shang and F. You, "Data Analytics and Machine Learning for Smart Process Manufacturing: Recent Advances and Perspectives in the Big Data Era," *Proc. Est. Acad. Sci. Eng.*, vol. 5, no. 6, pp. 1010–1016, Dec. 2019.
[10] M. Fahmideh and G. Beydoun, "Big data analytics architecture design—An application in manufacturing systems," *Comput. Ind. Eng.*, vol. 128, pp. 948–963, Feb. 2019.
[11] B. Chin-Yee and R. Upshur, "Clinical judgement in the era of big data and predictive analytics," *J. Eval. Clin. Pract.*, vol. 24, no. 3, pp. 638–645, Jun. 2018.
[12] M. Kamal and T. A. Bablu, "Machine Learning Models for Predicting Click-through Rates on social media: Factors and Performance Analysis," *IJAMCA*, vol. 12, no. 4, pp. 1–14, Apr. 2022.
[13] S. Chaudhuri, B. Ding, and S. Kandula, "Approximate Query Processing: No Silver Bullet," in *Proceedings of the 2017 ACM International Conference on Management of Data*, Chicago, Illinois, USA, 2017, pp. 511–519.
[14] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Approximate query processing for big data in heterogeneous databases," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 5765–5767.
[15] M. Garofalakis and P. B. Gibbons, "Approximate query processing: Taming the TeraBytes!," 2001. [Online]. Available: http://www.vldb.org/conf/2001/tut4.pdf.
[16] K. Chakrabarti, M. Garofalakis, R. Rastogi, and K. Shim, "Approximate query processing using wavelets," *VLDB J.*, vol. 10, no. 2, pp. 199–223, Sep. 2001.
[17] B. Babcock, S. Chaudhuri, and G. Das, "Dynamic sample selection for approximate query processing," in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, San Diego, California, 2003, pp. 539–550.
[18] Y. Park, B. Mozafari, J. Sorenson, and J. Wang, "VerdictDB: Universalizing Approximate Query Processing," in *Proceedings of the 2018 International Conference on Management of Data*, Houston, TX, USA, 2018, pp. 1461–1476.
[19] A. Galakatos, A. Crotty, E. Zgraggen, and C. Binnig, "Revisiting reuse for approximate query processing," *Proceedings of the*, 2017.
[20] M. Muniswamaiah, T. Agerwala, and C. Tappert, "Data virtualization for analytics and business intelligence in big data," in *CS & IT Conference Proceedings*, 2019, vol. 9.
[21] S. Agarwal *et al.*, "Knowing when you're wrong: building fast and reliable approximate query processing systems," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, Snowbird, Utah, USA, 2014, pp. 481–492.
[22] S. Chaudhuri, G. Das, and V. Narasayya, "Optimized stratified sampling for approximate query processing," *ACM Trans. Database Syst.*, vol. 32, no. 2, pp. 9-es, Jun. 2007.
[23] M. Mohamed Nazief Haggag Kotb Kholaif, M. Xiao, and X. Tang, "Covid-19′s fear-uncertainty effect on renewable energy supply chain management and ecological sustainability performance; the moderate effect of big-data

analytics," *Sustain. Energy Technol. Assessments*, vol. 53, no. 102622, p. 102622, Oct. 2022.
[24] C. Jermaine, S. Arumugam, A. Pol, and A. Dobra, "Scalable approximate query processing with the DBO engine," *ACM Trans. Database Syst.*, vol. 33, no. 4, pp. 1–54, Dec. 2008.
[25] S. Fosso Wamba, A. Gunasekaran, and R. Dubey, "Big data analytics in operations and supply chain management," *Ann. Oper. Res.*, 2018.
[26] R. Ak and R. Bhinge, "Data analytics and uncertainty quantification for energy prediction in manufacturing," *Conference on Big Data (Big Data)*, 2015.
[27] K. Li and G. Li, "Approximate query processing: What is new and where to go?," *Data Sci. Eng.*, vol. 3, no. 4, pp. 379–397, Dec. 2018.
[28] C. Maraveas, D. Piromalis, K. G. Arvanitis, T. Bartzanas, and D. Loukatos, "Applications of IoT for optimized greenhouse environment and resources management," *Comput. Electron. Agric.*, vol. 198, p. 106993, Jul. 2022.