IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111251

End-to-End Cloud-Scale Data Platforms for Real-Time AI Insights

Phanish Lakkarasu

Staff Data Engineer, phanishlakarasu@gmail.com, ORCID ID: 0009-0003-6095-7840

Abstract: Cloud service providers are adopting AI-based systems to efficiently offer data services. A data platform combines data management systems, serving storage, and compute infrastructure. A data service is the result of the careful orchestration of these services, with a set of users on the right-hand side and their data and knowledge requests on the left. The data service is typically composed of microservices with data as an object. Each microservice then offers APIs and SDKs for users to submit tasks and queries to the data platforms. The cloud-scale data platforms and the data service architecting and planning should be automated so that users can focus on describing their workloads without worrying about the underlying architectures. Data service design is extremely challenging. The workloads are broad and horizontally scaling. All the architectural components are stateful, dynamic, and performance-sensitive. The architectural complexity is huge due to the vast design space and requirement sets. The trade-offs on diverse metrics and concerns are crucial. The aforementioned challenges are further magnified in cloud-scale systems. A simulation-based framework is built to facilitate performance- and power-accuracy exploration across heterogeneous hardware implementations. The framework is employed to explore the design space of big data analytics written in a high-level domain-specific language for reconfigurable systems. An architecture transformation framework is presented to transform plain applications into efficient hardware blocks. The framework performs automatically instruction-level optimization on a petascale simulation kernel, achieving speedup over state-of-the-art toolchains and domain-specific compilers.

Keywords: Real-Time Data Processing, Cloud-Native Architecture, Scalable Data Pipelines, AI-Driven Insights, End-to-End Data Integration, Data Lake House Architecture, Streaming Analytics, Machine Learning at Scale, Event-Driven Architecture, Unified Data Platform, Low-Latency AI Inference, Big Data Orchestration, Cloud Data Warehousing, Predictive Analytics in Real Time, Automated Data Engineering.

I. INTRODUCTION

Artificial Intelligence (AI) and data science's rapid adoption has transformed numerous industries. AI technologies—including machine learning, deep learning, and natural language processing—extract insights from vast amounts of data. This fuels a data-driven economy where Data Science helps companies make informed business and operational decisions. The foundation of many successful AI/DS business use cases is a cloud-scale data platform capable of handling data volume, velocity, and variety efficiently and effectively.

High levels of automation are required to ensure that a cloud-scale data platform is robust, reliable, and managed effectively 24/7 in a cost-efficient manner. Few technology companies have invested heavily to automatically manage their internal data platforms. This naturally raises the question of whether this effort is possible in an open-source field within a few years. The increasing interest in strong cyber-attack protection treatment has led many major service providers to apply large amounts of engineering effort into such a system to manage the data platform.

The design of a data platform at cloud scale involves many different technologies. For cost-effective bulk storage and fast processing, a normal data platform utilizes a hybrid storage engine consisting of high-latency hard disks, fast random-access solid-state drives, and memory. In the cloud, log-structured storage systems with a cloud-scale nature are a reliable option. In the cloud, open-source alternatives with a guarantee of high availability are usually a better option for stream processing. This new design requires many technologies to be integrated more deeply to achieve better overall performance and efficiency.

More than 60% of the Fortune 100 companies are routinely using multiple data stores to run their missions' most critical applications, but the increasing heterogeneity and distribution of the resulting systems are posing more difficult challenges for City and event data management. Nevertheless, there are still fundamental problems in data consistency, availability, and partition tolerance. Considerable research has gone into resolving such problems, producing, for example, the proof of the CAP theorem, which states that no distributed data store can simultaneously provide all three guarantees.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111251

1.1. Background and Significance

Applications that require continuous monitoring and analysis of rapidly evolving event flows have become prevalent across various sectors. This has led to growing interest in developing frameworks and systems for processing continuous data streams. Several initiatives have centered on building software systems for real-time streaming data analytics. Research has concentrated on devising scalable, high-throughput architectures and algorithms as efficient stream processing and data mining engines to achieve a cluster-centric framework. The complexity and size of many applications make it difficult to design a single, monolithic real-time processing or learning system to accommodate all the envisioned functionalities. During runtime, different resources dynamically evolve in terms of volume, velocity, or complexity. These developments could be a result of emergent needs of higher-level tasks as well as noise, anomalies, or a change of characterization in the underlying data sources. The notable rise of ubiquitous, interconnected services has quickly broadened the range of potential applications. Frequent spikes in the frequency of transactions and their execution volumes are usually met with technological advancements in high-performance data stream processing. It is equally important that caution is taken to scrutinize the validity and integrity of key input data, service parameters, and models utilized in these automatic processes when it comes to safety- or reputation-critical business-critical tasks. Knowledge extracted from high-frequency time series or event streams may need to be interpreted differently due to long-lasting, gradual changes in behavior. Another fundamental requirement is to produce interpretable, trustworthy insights instead of "black boxes" for continuous processes, which would otherwise only be utilized by a few experts.

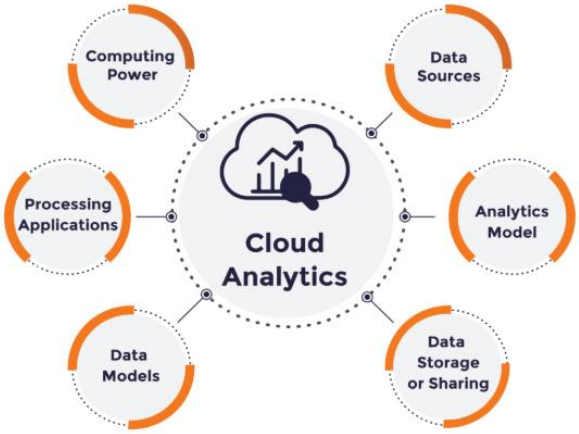

Fig 1: Cloud-Scale Data Platforms for Real-Time AI.

II. UNDERSTANDING CLOUD-SCALE DATA PLATFORMS

Cloud-scale data platforms are responsible for managing mission-critical data of organizations of all sizes and shapes. To succeed in the new dynamic landscape, cloud service providers are introducing a new generation of cloud-scale data platforms that are autonomous, intelligent and real-time. They want to realize the vision of cloud-scale data engineering and analytics, where cloud services can fully automate themselves and enable two orders of magnitude faster analyses of fresh data. However, this is a prohibitively challenging problem. While it is possible to build either cloud-scale data services or autonomous data services based on existing technologies, the overall solution is much harder. This problem is viewed as an extreme optimization problem in the distributed and asynchronous data platform ecosystem.

Several advancements are outlined to understand the challenges in attaining closed-loop optimization on cloud-scale data platforms and the recent developments towards an autonomous data services architecture. While it is a long and hard journey ahead, it is equally exciting as it has the potential of bringing autonomous data services across the world ten years earlier. For a cloud service, especially at scale, the service can be a very complex system with multiple closely interacting components, each of which may consist of multiple complex sub-parts, and may rely on other technology stacks or off-the-shelf systems. It's impractical to create a massive optimization problem that simultaneously optimizes all components that involve too many considerations across multiple dimensions and layers.

A challenging problem is that in real-world systems, there are interactions between co-designed components that have common data, goals, and business KPIs. However, to address the problem, this is to focus on optimizing a selection of related components that work together in a coordinated way.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111251

On one hand, it is not possible to minimize the whole system as one single optimization problem due to the scale and complexity. On the other hand, improving the joint optimization of these selected, tightly coupled components, which are captured and modeled in the respective solvable, smaller optimization problem, can indeed lead to improvement in overall performance. However, it is equally important to expose only a small and discriminating set of tunable parameters to the customer at the end. There is a responsibility to ensure that the solutions provided to customers, both online and offline, are insightful and safe for them.

Equ 1: Real-Time Data Ingestion Throughput.

$$T_{ingest} = R_d \times N_s$$

T_{ingest} = Total data ingestion throughput (e.g., MB/s)

R_d = Data rate per source (MB/s)

N_s = Number of data sources

2.1. Definition and Key Features

The autonomous orchestration of cloud-scale data services requires resolving configuration choices, controller parameters, and the mapping of the services to compute and storage capacity. Any substantial cloud service that ingests, processes, and serves out data requires multiple components to work together. For example, a data ingestion service logging raw data to files can consist of a data stream messaging system, worker nodes capable of reading and labeling those files, and databases storing indexed events.

The architecture is further complicated by cloud applications being continuously deployed and updated. Existing cloud services run on multiple clusters in multiple cloud regions for availability, durability, bursting capacity, disaster recovery, capacity for increasingly engorged data, and the further distillation of products. A cloud service must not only operate well in isolation but also interact with its many neighbors. Adding a shard of a database must not degrade performance beyond tolerable timeouts. Migrations, crushes, failovers, failbacks, and performance fluctuations can frequently disrupt a multi-service cloud or data platform.

Multi-cloud service tiers or products within a cloud have different operational requirements. Vertical multi-cloud stacks must ensure consistent and coherently good QoS levels for all customers. The high performance of one service independently of the others may not equate to overall high performance or a good customer SLA.

2.2. Architecture Overview

The architecture of the proposed cloud-based data platforms consists of an Oan T-Capacity Layer that supports massively-parallel processing of multi-dimensional parallel spatial RDDs on the cloud with performance isolation guarantees; Log-Capacity Layer for compressing, indexing, and caching indexed logs to augment the originally-source logs; Querying Layer with scalable data access decomposers to fetch and merge the needed compressed and indexed logs; Processing Layer that provides advanced stream and batch graph operators that are optimized to leverage the underlying OT-Capacity Layer; and an automated query optimizer that provides effort-based strategy selection for jobs that involve generic stream/batch operators. The cloud platform is built upon an RDD-based streaming and batch graph processing framework on a cloud cluster that provides performance isolation at the OT level and supports executing multi-dimensional parallel RDD workloads.

The system model of the cloud platform is illustrated. Each application submits a set of jobs to the Application Scheduler, which prioritizes each job according to the Completion Time Estimator. In the framework, each query operates on partitioned RDDs of logs and converts them to partitioned continuous data streams of history logs, and the logs are incrementally augmented to one or more OT-RDDs. The Log-Indexed OT-RDDs convert the OT-RDDs into OT-RDDs of indexes for the underlying compressed logs. The output of the Querying Layer can be indexes and batches of logs to be processed with one or more stream or batch processing query, and the processed results can be archived into compressed, indexed logs or output as analysis results. The Processing Layer is executed by a final processing engine according to the type of the output data from the Querying Layers.

An online test with 80 cases of anomaly episodes in 5 weeks of 1-minute logs was conducted. During the test, a human operator is called to execute the faults or issues. The daily results of false alarms and missed alarms were gathered. The Recall/Precision is the ratio between the number of cases correctly detected and the number of total holes. New cases or falsely detected cases in the test data but not in the explanations are false positives.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022
DOI: 10.17148/IJARCCE.2022.111251

2.3. Benefits of Cloud-Scale Solutions

The world's largest cloud services providers have deployed over 10 million servers worldwide and new servers are continuously added to support the explosive growth in the mobile-first AI-first world. The cloud services democratize access to computation, storage, and diverse services. Millions of developers and enterprises leverage the cloud to build rapidly innovative applications like social networks, e-commerce platforms, intelligent education systems, transportation networks, or healthcare solutions. To assist data scientists and software engineers in building production-ready AI solutions freighted with big data coming from the cloud services, it is important to automatically provision needed cloud service resources and construct out-of-the-box production-ready pipelines to serve big data for real-time insights.

The enormous weight of big data introduces a huge latency budget for real-time AI inference pipelines. Typical online services adhere to a latency budget of a few milliseconds at the service provision end to ensure that insights are timely. There is a huge engineering overhead to ensure the low latency of big AI insights because of the complex structure of the AI models, large volumes of data, and diverse cloud services adopted for pipeline construction. Despite the rapid adoption of model-centric and data-centric AI methods, the leveraging of automated tools for effective and efficient pipeline provision remains under investigation.

Creating production-ready AI services is a core aspect of the AI paradigm and a tough problem to address. Surprisingly little work is focused on the cloud service procurement and part of the infrastructure establishment over the cloud consumptions. This omission is a glaring oversight in the cloud-centric services of multi-cloud background and diverse edge services from large cloud service providers. The platforms in current cloud providers offer the option of defining users' requirements by manually setting up constitutive components for part of the service. However, to procure a new service rapidly in the multi-cloud choice, there is an urgent need to automatically end-to-end provision data pipelines from the data ingestion on the edge with cloud services provisioned at the edge to batch or stream processing of the data at a cloud region and to visualize results.

III. REAL-TIME DATA PROCESSING

Big data is generated in a variety of dimensions, including space, time, and semantics. Aggregating the data in the three dimensions provides an alternative way for the decision-makers to comprehend and analyze the big data. The incremental computation of the cube is quite expensive when it comes to real-time data. Many applications query for just a few metrics simultaneously, which suggests the polynomial decomposition of the metrics. This paper proposes a few solutions for incremental algorithms, revealing the fine-tuning and challenging tasks on the other dimensions of big data.

The last decade has witnessed an increasing amount of interest in big data among researchers, standard bodies, and related organizations such as cloud vendors. Big data mining focuses on discovering valuable knowledge from big data. The big data is stored in a variety of systems such as relational systems, key-value stores and HDFS. Recently, an emerging research area is flow data processing, in which data is generated continuously in a potentially unbounded stream. The diverse sources and stringent requirements of big data pose many challenges to traditional data mining models, which inspires the research of big data processing, mining, and analysis. This paper surveys the existing frameworks, systems, algorithms, and recent advances in the above aspects. Enhancing the dimensionality of big data (also known as high-dimensional big data) has also been an active research topic.

A key assumption made by the majority of existing works on dimensionality reduction is two-fold: (1) the data to be embedded into the lower-dimensional space is well represented by the original high-dimensional space, and (2) they are in a typical stream model. The one with the increasing number of dimensions is significantly harder to analyze and process, and it has become an increasingly popular research topic due to the great significance of big data and many practical applications across many domains including search engines, text analysis systems, and machine learning. These applications raise new challenges, and opportunities for scalable, efficient, reliable, and accurate algorithms and systems for processing real-time intelligent big data, which have not been adequately studied and have from the prerequisite understanding of the characteristics.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111251

Fig 2: Real-Time Processing.

3.1. Stream Processing vs Batch Processing

Stream Processing is regarded as one of the upcoming technologies with high market potential. The power of SP and SP systems can be leveraged to provide solutions to emerging classes of Big Data problems. The ideal application is data-centric systems involving large-scale and complex, dynamic data that need to be analyzed and transformed in real-time. SP technologies such as Storm, Spark Streaming, and S4 are implementations of these ideas. SP technologies were also introduced as commercial products running in a cloud environment, which offered a combination of some desired features. It is important to understand some of the fundamental research issues in massively distributed stream processing that cannot be addressed by existing tools and products.

The need for continuous processing of real-time data has resulted in the emergence of a new group of data processing systems referred to as stream processing (SP). The ongoing massive growth of data motivated a switch from traditional data processing solutions toward systems that can continuously process a stream of real-time data. SP systems can provide solutions to a variety of emerging classes of problems in areas that include social networks, infrastructure monitoring, stock market predictions, fraud detection, logs analysis, transportation and traffic monitoring, meteorological monitoring, sensor networks, and telecommunications. Such systems have a broader scope than earlier event-processing systems. These stream processing solutions must scale by an order of magnitude and involve massive amounts of computations and data.

Existing research on SP systems has predominantly focused on performance and scalability. A massive SP system requested the design of architectures, data flow models, and principles of system resilience to hardware and software failures. Performance issues include latency, queueing, complexity of optimization, inter-node communication, and node-level scheduling issues. These topics are of high theoretical merit. However, the cloud is a reality, and thus, it is time for measurements of real-life SP systems that include. Research into performance issues must also include Internet-scale measurements of internal nodes. A potential research direction would be a quantitative investigation into architecture and software trade-offs on the performance metrics stated above.

3.2. Technologies for Real-Time Processing

Real-time streaming data is characterized by the need to respond to the data converging on a system while it is still arriving. Stream data is inherently different from traditional data, as it is continuous and the volume is usually exponentially larger than that of conventional data. For this streaming data, the first question that comes to people's minds is querying. Stream data itself does not hold any value if there is no level of intelligence embedded in the same. To extract knowledge from such a diversified and huge amount of streaming data, the need for smart analytics is on the rise. Stream data has diversity in multiple forms, and capturing heterogeneity is a new challenge in the data analysis field. Consequently, stream analytics itself should have different layers of intelligence. The major progress of stream processing (SP) is to provide sweet tools for data scientists to craft knowledge from raw streaming feeds. Thereby, a stream computing (SC) system to extract structures from diverse data is described.

DOI: 10.17148/IJARCCE.2022.111251

The key component in SC systems is stream processing engines (SPEs), which have heavily focused on rules or patterns to filter streaming data. To fill this gap, the DataAna system is designed to denotation modeling choice sets with a context-sensitive pattern mentioned in a division. It records the behavior state of data in a tree structure and exploits the powerful operation of choice sets over this compact structure. Simultaneously, streaming processing based on an entire graph is acknowledged to forget cognition and get rid of temporal requirements. Profile-Stream, a scalable and low-latency platform based on auto-regressive integrated moving average for context modeling over time. The stream graph is structured by node-batch and edge-batch to handle graph evolution, and sketches are integrated to efficiently and robustly resolve the graph capture participative nodes.

3.3. Use Cases for Real-Time Data

Timely information processing provides various benefits across diverse domains, e.g., in economic areas, timely detection of fraudulent transactions prevents a huge amount of money loss; in social domains, timely detection of fire tweets may save many lives; in mobile computing environments, timely monitoring of users' context changing may better serve them. Given a growing number of applications that process real-time data streams at a large scale, cost-effective cloud computing is becoming important for real-time data processing. A cloud stream processing framework should provide technologies from servers, databases, and streaming systems, to data management models. This enables users to quickly deploy their cloud-scale real-time data processing applications.

Cloud-based stream processing solutions have proliferated recently, driven by the market interest of IT companies. From the platform view, Google Cloud Dataflow allows dynamic resource allocation and global auto-scaling. Amazon Kinesis provides a suite of services for ingesting, storing, and processing streaming data. Microsoft Azure Stream Analytics focuses more on connecting data sources. Open-source distributed stream processing systems also attract lots of attention, e.g., Apache Flink, Apache Storm, Apache Samza, Apache Spark Streaming, etc. These systems are appreciated for better flexibility and wider compatibility with diverse tendering frameworks and message queuing platforms.

Real-time data analytics is the "online" analysis of data that were previously captured in "offline" processing. With the advent of modern applications that generate various data streams ranging from fine-grained digital trading transactions, GPS location updates from smart vehicles, social network status updates, and snapshots of sensor data in IoT systems, data analytics has been shifting from "offline" batch processing, which focuses on mining usually static relational datasets, to "online" real-time analytics on data streams that usually arrive in incompleteness and continuous nature. Real-time insights delivered in a timely fashion in time meters are expected, which significantly differ from traditional offline batch processing systems that focus on revealing deeper yet retrospective insights but at greater times (in minutes to hours or even days). As a result, streaming data platforms differ in architecture, query execution model, underlying programming framework, and real-time analytics functions adopted.

IV. AI AND MACHINE LEARNING INTEGRATION

Data platforms are expected to scale to 1000s of concurrent data consumers like mobile applications, business dashboards, AI and ML pipelines, etc. Data systems at this scale are expected to ingest petabytes of data daily and process terabytes of data per second. Such data platforms also need to ensure the quality of data and computations, and they should be accessible to end users as user-friendly SQL interfaces.

With recent advancements in modulations and algorithms, AI and ML are now critical workloads for businesses. There's a growing need for a technology stack that can easily integrate data pipelines and AI and ML pipelines on the same platform for better insight generation.

In this paper, the design of a data platform powered by end-to-end cloud-scalability, flexibility with open architecture, and real-time data processing pipeline streaming, batch, AI and ML pipelines is proposed. With the data platform providing the necessary data readiness for insightful generation, the user-friendly environment with SQL interfaces and accessibility for diverse groups of end users to generate insights are built.

The technical details for building end-to-end cloud-scale data platform on data ingestion and storage, data pipeline on real-time stream processing and batch processing, AI and ML cherished by friendly SQL interface, and orchestration on deployment of these pipelines are all covered. Data engineers can define complex data ingestion jobs without writing any pipelines or meta-programming. On the data consumption side, end users, such as data analysts, data scientists, and business interpretation officers, can directly use SQL to query on the dashboard and BI tools and to deploy data analysis or AI and ML pipelines as well.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111251

Equ 2: Latency of Real-Time AI Pipeline.

$$L_{total} = L_{ingest} + L_{prep} + L_{model} + L_{serve}$$

- L_{total} = End-to-end latency (in milliseconds or seconds)
- L_{ingest} = Latency of data ingestion
- L_{prep} = Latency of data preprocessing
- L_{model} = AI inference latency
- L_{serve} = Latency of serving results

4.1. AI Models and Data Requirements

The AI community has addressed a broad range of problems, and AI models have been developed in numerous areas based on many machine learning and deep learning algorithms, which have enabled unprecedented advancements and commercial adoption. AI models for CAPTCHAs, financial forecasting, traffic analytics, photo labeling, auto-captions, personalized ad recommendation, speech recognition, and many others have been developed and made available to a broader user community through cloud services. The current solutions regard these models as stand-alone components. A considerable amount of effort is still needed to adapt the data to the model, including model selection, hyper-parameter tuning, and full-fledged parameter training before it can be served and used to make online or near-real-time predictions on a new data/feed. Furthermore, these broader applications require one or more of the following kinds of data processing tasks to be performed before the model invocation to enable full-fledged end-to-end, real-time use at scale.

The data preprocessing tasks include extraction, cleaning, normalization, filtering, and transforming the incoming input data from various sources before being served into the model running components, which can require real-time processing of large volumes of streaming data. In addition, feature engineering is usually needed on the original data features/schema to derive additional features with enhanced information content or redundancy so that the model can make more accurate predictions, which is usually an offline batch task over the new historical data. To conduct this end-to-end online prediction service, a cloud-streaming architecture is often adopted. There is typically a data collecting service to obtain data from the original venues and various inter-process communications services to publish this data to end-user clients and/or application servers. With the current big question in AI of how to identify and argue the “success” of an insight, data, and insight authorship and provenance are set to become a key focus in both big data, knowledge, and privacy arenas, and new standards are needed here.

4.2. Real-Time AI Insights Generation

In recent years, more and more enterprises have deployed artificial intelligence (AI) systems for business intelligence (BI). In particular, AI-based real-time BI systems have attracted much attention, and are capable of providing a competitive advantage for modern enterprises. AI-enabled real-time BI systems gather enormous streams of rapidly changing data and generate actionable insights through a chain of data processing with real-time stream data technology and AI models such as machine learning. Existing systems, including platforms, solutions, and applications make progress in either real-time data processing or AI insights generation.

Such two aspects of the production line are independently conducted and often disconnect, failing existing systems to deliver AI-based real-time BI to meet the growing demand. Meanwhile, a high convergence of real-time data processing and AI insights generation emerges:

Real-Time Data as Input: The data stream for AI insights generation in real-time. The policing time of data input at the monitoring points is much less than the time the AI model runs on accumulated data.

Real-Time Models as Output: Real-time AI insights are presented in the form of a model set. The model set contains a decision model for targeted data mining and a report model for visualized presentation. The former is often trained on data batches and with some latent period, therefore it is not fully real-time. Instead, only the coefficients of trained models are updated with the data stream. These models are called real-time AI models. However, the latter must drive from real-time data output, that is, they change with the chances of important data. Finally, a new end-to-end cloud-scale data platform for AI-based real-time BI, termed AI Real-time Insight Platform (AI-RIP), is proposed to overcome the difficulties and challenges in existing engineering solution and meet both the requirement and convergence of AI-based real-time BI, finally cope with the massive stream data and the growing insight demand.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111251

4.3. Challenges in AI Integration

There are exciting opportunities for many companies across a variety of industries emerging from the rapid advances in artificial intelligence (AI) technologies. Adding intelligence where data and optimization are concerned is an effective path to effectiveness, efficiencies, differentiation, and profitability. It is imperative first though to access the data at a scale and quality necessary to support effective and scalable training and optimization of the machine learning algorithms (MLAs) that provide these capabilities. Additionally, data coming from a variety of sources is complicated and complex. The assurance of understanding, value, interconnectivity, and usability of these multi-tenancy data sources (and the management of the value it creates) becomes an equally important aspect of this intelligent opportunity space.

This solution involves holistic data management, where raw (unstructured) and raw tagged (structured and time-stamped) data, ML input-output results, machine learning ideas, business outputs, display results, and end-user interaction, are finally managed like configuration, model, output, and transactional data. Learning, inferencing, configuration, model building, and transformation are all automatically applied and separated from raw data management because these core capabilities can create added (higher order and higher risk) value (vectors). They have an understandable data life cycle that starts with the raw data and ends when they and their added value go inactive or unavailable (deleted). This contrasted with raw data which can be tagged, and transformed, but not managed by configurations, models, and rules.

V. DATA INGESTION STRATEGIES

Ingesting live data streams for trending analysis or stock prices for predictions is the foundation of building end-to-end, real-time AI pipelines. Ingestion and storage environments can utilize either commercial cloud solutions or open-source systems. A continuous stream of tweets arriving from various sources is collected, pre-processed, and transported to a storage system, where it will be stored in raw format for both real-time predictions and offline training. Data modeling preparation in different formats is needed for each analytics framework. A standardized linguistic and structural model can streamline the subsequent services by making it easier to be reused. The ingestion side also includes the implementations of producers for streaming data sources and consumers for archiving both raw and processed data on storage. The streaming modeling could include GUI-based programming and visual-oriented programming approach for analysts with limited expertise in programming.

Almost all emerging types of streaming data require some form of pre-processing. Filtering irrelevant information, and enriching data using reference information is critical for the efficacy of machine learning models. The current options offer limited choices for enrichment. The general approaches include native functions by distributed stream processing systems, connectors to professional data enrichment services and custom, external enrichment jobs. The rich pipeline options should also help with the maintenance, updates and composition of analyses. If any single pre-processing step required is too complicated to be user-configurable, an enrichment service that can be directly labelled on the pipeline regarding what pre-processing steps were conducted on what data is essential. Another interesting property is pipeline aggregation, i.e., compositional analyses where the enriched data is further processed. Intensive, batch and iterative processing are two additional analytical paradigms to be studied. The mature academic benchmarks used in the batch and offline processing.

There has been great interest in cloud-based AI pipelines, which involve distributed processing on massively parallel clusters. Like the cloud, distributed model training and processing frameworks are being re-implemented. The analysis languages being studied also include SQL-like languages for stream and big data and domain specific languages on top of model processing frameworks. The hot topic of AI-based consumption demand prediction of videos is being looked at.

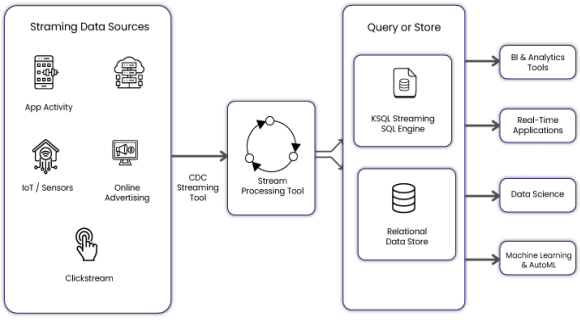

Fig 3: Data Ingestion Strategies.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022
DOI: 10.17148/IJARCCE.2022.111251

5.1. Types of Data Ingestion

Today's data sources are diverse: sensor and Internet of Things (IoT) devices, enterprise applications, social media, and e-commerce websites, to name a few. Such data is often far from the form required for analytics and needs to be ingested and enriched based on other static or reference data, dynamic metadata, or global models to support complex analytical queries. For instance, given a collection of recent tweets, enrichment operations may need to obtain their sentiment scores, identify the named entities in related news articles, determine where and when the tweets were broadcasted, or cluster/graphically map the tweet locations in real-time.

The ingestion pipeline, which connects disparate data sources to both online and offline data stores, is one important component of data management systems. Real-time ingestion pipelines are expected to keep up with continuously changing data while producing a correct and up-to-date producer view of the sources. However, satisfying the above requirements is inherently challenging. For one, there is often a large volume and velocity of data to be ingested. For another, data changes in a variety of forms: news tweets being reposted, price data being uploaded or updated in a batch, or stock trending becoming available in a time-ordered stream. Moreover, frequently, based on the use case, enrichment operations can be compiled code, declarative queries in SQL, arithmetic expressions in Scala, or complex machine learning models represented in Dask or TensorFlow. These facilities may have completely different system environments and resource slots, as well as data formats. Lastly, furthermore, in many cases, the consumed data sources are in the public domain while the enrichment models and reference data should be kept private and need to be gradually enriched offline.

To meet these steadily growing requirements, a new data ingestion framework called IDEA, which can ingest online data at scale and enrich it based on configured and adaptive enrichment models and reference data, is presented.

5.2. Tools and Frameworks

Deep Learning frameworks specialized for extremely large neural networks and intricating workloads. These frameworks can efficiently utilize a powerful computing environment by offering sophisticated execution scheduling algorithms, optimized communication schemes, and low-level resource-aware kernels. However, deploying DL pipelines running on these complex frameworks, especially in a multi-cloud environment, is a challenge. A unified abstraction is proposed to compose general resource-intensive workloads on cloud computing and an easy-to-implement execution framework for a wide range of workloads is introduced. Hyper-unified frameworks can employ multiple levels of data parallelism. A software framework is implemented to unify computation distribution and manage hyperparameters. Both sensitiveness to early stopping states and execution strategy are investigated. The performance study containing two large models on three machines demonstrates that 5.6x and 8.2x improvements in speedup are observed compared to widely adopted frameworks with consistent prediction accuracies. Nowadays, networks with huge computing and memory costs are surfacing, which would lead to a speed limit on DL frameworks using static task scheduling decisions defined in the compiled runtimes. Cloud frameworks empowered by computing resources could afford this heavy overhead, while data transport and tool incompatibility are challenges during the transformation. Advanced programming languages are proposed to unify and collaborate heterogeneous computing resources. In addition, based on new designs for the programming languages, live automation of resource allocation is studied. Compared with edge-cloud computing patterns, they can achieve a lower overall cost and guarantee the rapid growth of production results. Cloud workloads can consist of various scheduling patterns, from parallel map-based computing manufacturers with mandatory performance guarantees to a loose style of standalone models with soft consistency constraints. Hence, the computation-device association granularity in cloud computing frameworks is essential for efficient execution. Initially in cloud architecture, computation modules specified fine sub-computation tasks in a unified framework, while these frameworks might not efficiently accommodate the granularity of scalable nodes, leading to data partitioning and transmittance overheads. High-performance white-box libraries for intermediate computations that can handle incrustation model architectures are employed in a fine-grained pipeline. These workloads can incorporate a numerical computing library, stateless complex operators, or on-demand data queries, which benefit the two-level scheduling run pipelines as both dynamic task partitioning and decentralized compiling in HPCs. The off-peak hyperparameter tuning status over execution history can also be leveraged to shrink the burden in hyperparameter tuning.

5.3. Best Practices for Ingestion

To allow users to control data streams during ingestion, some ingestion systems offer the concept of "data stream management". Apache Pulsar is designed from the ground up to provide end-to-end Quality of Service guarantees, with storage as part of the technology stack. This allows Pulsar to act as a highly scalable persistent event log, supporting very high throughput applications where data delivery is crucial. Furthermore, as part of the answer to the "data is not enough" maxim, platforms like Apache Ignite and Apache Druid seek to accelerate data ingestion while enabling indexes over the data.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111251

Although typically a part of the earlier W1 pipeline, the process of capturing data from applications and transforming them into a more usable format is equally important in the W2 search and exploration part of data platforms. Stream data capture via Change Data Capture (CDC) technology is gaining momentum for multiple reasons. For example, besides being (almost) generic in its input and output formats, it allows data to be obtained from previously unreachable data sources with minimal changes to the source applications, and it provides capabilities for near real-time analytics with ambiguity and containment of failures. The CDC technology stack comprises message brokers such as Apache Kafka or strongly typed streams such as Apache Beam that allow the composition of processing pipelines. On top of them, one can find transformation wheels implemented via custom programs servicing data consumers at production quality with failures recovery, data duplicity checks, etc.

Concerning data transformation, there is a spectrum of systems catering to use cases of varying complexity. Event buses operating at exceptionally low latencies and offering guarantees of delivery are often foundational components in a data engineering stack. Atop these stream-based systems, SQL-oriented tools processing data with declarative SQL-like languages enable enriching the incoming event streams with selected metadata by foreign key lookups. At the highest level of the abstraction foreground (read compute-intensive), streaming applications can be programmed to process the event streams independently of the rest of the system where the code is tightly coupled with the application that authored the data.

VI. DATA STORAGE SOLUTIONS

The modern data architecture is undisputedly cloud-first, being developed and operated by today's cloud data teams on hybrid and multi-cloud technologies. Partly to take advantage of their scale, and partly because many analytical algorithmic techniques developed are fundamentally accumulation based rather than point-update based. However, real-time business intelligence is critically important to many enterprises doing online businesses. This insight led to the pioneering of an alternative cloud data architecture also capable of near real-time fast data processing, meeting immediate business intelligence needs.

Big data analytics has to switch from pull to push — from batch querying to streaming data assembly, as evidenced by the emergence of streaming data platforms and specialized stream analytics engines. Real-time business intelligence demands radical shifts from batch querying on cloud data lakes backed by high-throughput object stores to the critical use of streaming data, aggregate materialized views or event stream features, and fast data processing. Distributed stream processing engines have become an indispensable part of modern analytical architectures. Cloud-streaming-based, on-the-fly event-sourced data lakes and continuous analytics service solutions now seamlessly deliver end-to-end scalable data platforms.

Moreover, cloud data processing, at minimum, demands three essential needs: massively parallel data and resource scalability, heterogeneous cloud services interoperability, and complete processing cycle — from data acquisition to workload allocation, filtering, batching, assimilating, processing, storage, and publishing. Existing cloud data platforms either provide highly integrated services feeding each other or are flexible with plug-and-play technologies. But very few tie up, integrating a total data processing cycle, catering for the compelling data cohesion and inter-operation across cloud services, and the extremely fragmented configuration complexity.

6.1. Choosing the Right Storage Type

In modern AI architecture, a solid foundation layer is required to collect raw data from diverse storage sources, clean, transform the data, and ingest it into a persistent storage repository so that it can be used for data training or batch computing. The foundation data systems are often referred to in AI architecture as the "Data Lake" or the "Data Warehouse" repository. In a data lake, streaming, quick, and event-based sensor data in original and semi-structured formats are ingested, queried, and analyzable. Batch data or historical data from traditional relational databases, CSV files, and enterprise applications are also converted, cleaned, and trained in this type of repository. Data Lake is the storage repository of both real-time and batch sensor data. Some data can be used for real-time machine learning (ML) and AI insight on the same query. For batch data, other methods are often used, such as Extract Transformation Load (ETL) approaches with offline analysis.

In cloud-based data platforms, a myriad of structured, semi-structured, and unstructured data should be collected from different data sources including NoSQL databases, relational databases, flat files, message queues, APIs, IoT devices, and data reservoirs. As the volume of data is massive, it is important to first compare different storage backends after collecting it. In general, cloud storage should be intelligent and seamlessly integrated with data pipelines. It should be able to take proactive actions and re-configure the architecture while system components might fail.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111251

Furthermore, storage backends would change depending on the types of data, their flow, and the workloads utilized to consume this data. Finally, network and performance metrics should be monitored in such data systems. In similar scenarios where cloud-based data platforms collect unprocessed raw data from different data sources. Either a seamless data movement strategy is required or a complex data pipeline not only transfers data but also captures the type of data.

Equ 3: Scalability of Processing with Parallelism.

$$P_{eff} = \frac{W_{ideal}}{W_{actual}} = \frac{N \cdot T_{single}}{T_{parallel}}$$

- P_{eff} = Parallel processing efficiency
- W_{ideal} = Ideal workload time (linear scaling)
- W_{actual} = Actual processing time
- N = Number of parallel compute nodes
- T_{single} = Time to process data on a single node
- $T_{parallel}$ = Time using N nodes

6.2. Data Lakes vs Data Warehouses

The terms data lakes and data warehouses are commonly used today in the field of big data systems. Stakeholders can be easily confused by their different and overlapping meanings. For instantiating a cloud storage system, the choice might be between some cloud file system or a data warehouse. A careful analysis will reveal that the cloud storage systems provided by all these vendors share some common aspects. First, one can store and query data in a cloud data lake without knowing how the use cases will evolve in the future. Ones from the data warehouse family require prior knowledge of the types and formats of data and queries, and they perform best when those aspects are stable. The data lake is flexible. However, performance is low, and run-time evaluation is required to retrieve the data. Second, one can scale storage space and computation power dynamically on a cloud data lake. The former can be done almost instantaneously, and the latter in minutes. This is not the case for on-premises solutions. On-premises data lakes are more expensive since they require investment in hardware and computation in advance. The prices of resources for the former are also more economical. However, relying on a cloud platform also leads to serious risks and challenges, including data security, data provenance of pipelines, and fault tolerance of pipelines. In the cloud big data systems that are based on a data lake architecture, the term data lake describes an architecture, that consists of the following components: a cloud storage system to hold raw data in the cloud; a query language to implement pipelines; a computation framework to execute those queries; and connectors and clients for external experience. This architecture has been implemented as a collection of systems. One or several cloud storage systems are deployed using the cloud storage APIs of different vendors. A great number of early cloud data warehouse solutions can be used. Open-source frameworks are preferable for pipelines, but cloud-based services are also available. For languages and jobs, cloud protocols are used to access the systems. The big data network architecture can be seen as comprising three layers (i.e., data storage systems, applications, and data analysis systems), with the cloud integrated into the bottom two layers.

6.3. Scalability Considerations

End-to-End Cloud-Scale Data Platforms for Real-Time AI Insights. 6.3. Scalability Considerations. Data volume is a direct consequence of data quality in data platforms. Big-data systems (BD) and cloud systems (CS) ensure that high-data volumes are effectively ingested, cleaned, processed, extracted, and served on demand for analytics workloads. In contrast, a multitude of future data-and-analytic-intensive workloads suggests an explosion of data volumes, requiring petabytes or more storage, processing capabilities, and cost-aware retention strategies. The Cloud-Scale Data Platform is responsible for these tuning considerations and comprises major scalable sub-systems on data management (i.e., batch processing, streaming processing, and serving) solutions, supplemented by cloud infrastructures. To ensure the scalability of the proposed system architecture, three aspects are elaborated: Hardware scalability, software scalability, and out-of-scope scalability. At the hardware level, scalability concerning node scale and communication load is addressed. Therefore, sharing hardware resources among operators is avoided and specific operator-to-node scheduling is treated on-the-fly or based on preconfigured configurations. To prevent communication hot spots, operators on different nodes are assigned based on monitoring information on skewed workload and data volume after a static partitioning and deployment stage. For the software level, several software engineering solutions are followed to enable software scalability. In particular, almost all components of the system architecture can be horizontally scaled, Leave-One-Out (LOO) mechanisms are designed to maintain the high availability of the system, and a storage-efficient throughput-awareness scheduling mechanism is passed across components to achieve load-balancing. The mechanism intelligently

DOI: 10.17148/IJARCCE.2022.111251

controls task granularities and the pre-specified number of executions for each component. Out-of-scope scalability encourages other research endeavors. The extent of scalability and possible aspects of future research are addressed. It discusses tuning in regards to elasticity, reactiveness of future workloads, and upgrading/overhauling strategies of components.

VII. DATA GOVERNANCE AND SECURITY

Privacy breaches of data sources target sensitive details on persons, usually depending on which sectors they belong to. In this work, the users can consume data knowing that they remain entirely anonymous at the data sources or production level. The aim is to manage privacy breaches during queries targeting reidentification, correlation, compromise, and similar attacks. Therefore, all usual queries can be run and cannot be retrained to a specific user. The most promising approach is Relational Randomization, where all data is transformed by creating and working on encrypted tables. An example of this paradigm is the Perturbated Table, which is a single table with the same schema as its original and obfuscated by a nulling cascade.

Data integrity is a vital aspect of this storage. Table contents must be checked for data manipulations, as well as for hardware component damage that could lead to lost data. For this, main paradigm here is MACs combined with hash values. In this logic, every data manipulation is run through a secure, isolated module that checks the integrity of source data and creates signed MAC and hash values for the new data.

In this way, the process calms a flow of events, checks them at every point, and adds new test statements on the new objects added to the process. They are stored along with the table and checked every time a flow is queried. Therefore, corrupted or missing keys can be spotted, and the damage can be identified. Aside from the MAC, the resulting hash value is also stored along with the MAC to allow a basic check with minimal overhead when querying a table.

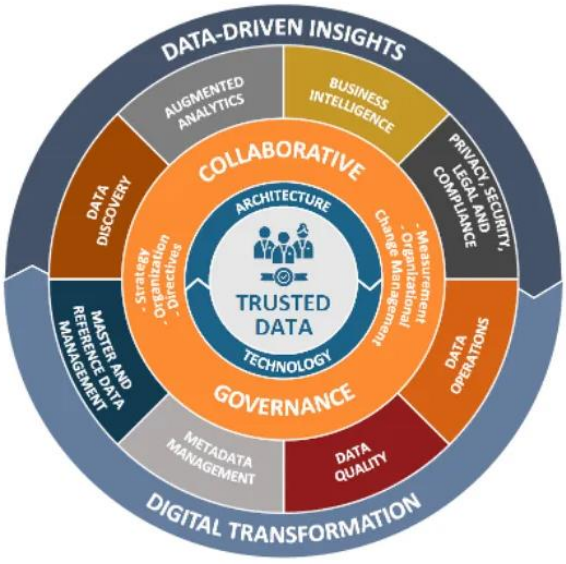

Fig 4: Data Governance and Security.

7.1. Importance of Data Governance

Data is no longer just a synonym for Big. Data is diverse. Different data sources provide different types of datasets with diverse semantics. This data deluge can be both challenging and enriching for organizations. From a governance point of view, the research on smart and sustainable solutions to extract value from 'All Data X' is driven by business needs. The motivation to extract value from all company data is at the heart of data governance (DG) initiatives. Demand-side DG is concerned with identifying types of data, knowing what data is available, and making decisions on how to define, manage, protect, use, and share data across the enterprise. Data governance is a broad area covering several main aspects such as data definition, data management, and data accessibility.

Before aspects of data cleansing, data integration, data exploration, data mining, data visualization, data retrieval, and data exchange can be tackled, data needs to be systematically identified with an adequate level of granularity and

DOI: 10.17148/IJARCCE.2022.111251

characterized. This information is scattered across different types of users, processes, data formats, and systems. The challenge is to provide a global, complete, and trustworthy view of which data exists and how to create queries to retrieve only the types of data desired, taking into account that all this information may be wrong or out of date. In addition to the above challenges, many organizations are also dealing with excessive production of data that goes unused after delivery. This can be due to a lack of understanding of the data's content, size, or quality. In due course, this accumulation can lead to data islanding, silent failure liability, data quality issues, or increased censorship and governance costs. Data governance for this data is low.

A broad use case around the provision of adequate answers to the business question 'What does the Credit Facilities data provided by Bank X to Bank Y mean?' is formulated. Using this question, the need for a business view of the Credit Facilities data is underlined which makes the answer understandable to business stakeholders. This business view contains a meta-representation of the Credit Facilities data in terms of business concepts. Furthermore, it allows for an easy understanding of the entities that this data source contains, the purpose of its creation, its owners, the last change made to it, and whether it is sufficiently governed.

7.2. Security Best Practices

Artificial Intelligence (AI) based applications raise concerns for the underlying cloud or fog services. A successful attack against a cloud service, such as a successful authentication attack against an access control service, can lead to the complete impairment of other AI applications. Because most service attacks, and passive and active attacks against data storage services, can also apply in the cloud, these services must be secured. However, there are services where the requirements can be considered differently, such as for access control. Here requirements are introduced for this service, including the use of the cloud or physics-based methods, such as multimedia security. Similar models for ensuring privacy in AI environments are shown. Other attacks could be thought of against data itself due to faulty cloud or fog services. These include manipulative adversaries, an often neglected domain, especially in cloud applications. The manipulation can be successfully performed using AI-based approaches or by inexperienced users.

Network-based services are an important part of AI-friendly applications, just as for any cloud agent. Denial-of-service attacks are not necessarily more clever and could either be indiscriminate network flooding or specific attacks against known machine learning algorithms. Similar to data-based adversaries, AI-based solutions could also be employed here against floods, which could learn the normal data rate and take action accordingly. Instead of preventing attacks, they could at least make them known faster and thus enable countermeasures. Fog computing as an extension of cloud applications also brings new attacks. The responsibility for security can be divided between more parties. With it, responsibility for security and auditing is spread across many parties. Security deficiencies at a lower level could affect other areas like social cloud, where shifts in user data could be problematic.

For a network of private devices with a limited security model, physical access poses a different threat. Networks could be manipulated simply by taking control of devices or through data transmission. It is crucial to restrict access to a minimum. Automated fault detection mechanisms capable of detecting unusual activity are also conceivable, or devices from the same provider. Data encryption can be used when machines do not require tightly coupled data access. Tracking means should also be put in place to enable faster reactions. Since moving to the fog portion of a network may be a critical point in the outflow of private data, secondary user accounts can also be created.

7.3. Compliance and Regulatory Considerations

Recent transformations in technology developed various machine learning (ML) models to carry out different tasks with increasing complexity, involving natural language processing or computer vision. Processing vast amounts of data with massive computing has made these models grow by orders of magnitude. Large Language Models, with tens of billions or even trillions of parameters, can generate human-like text, summarizing documents, and even programming code. Compliance and regulation are crucial in AI-enabled decision-making systems due to the risk of systematic discrimination and biases in technology. As machine-made decisions become more prevalent across various domains, the lack of explainability and transparency in algorithms can endanger fundamental rights and democratic values. However, compliance with these legal obligations is a challenge due to the intrinsically opaque nature of ML models. To handle accountability and compliance in AI systems, organizations submit algorithms to a trusted third party with sufficient domain knowledge and technical expertise, such as a special unit within a regulatory authority or an auditing company, to monitor their behavior on behalf of public authorities. By assigning decisions on who is supervised to semi-trusted entities that collect and store technical reviews and compliance evidence, only a symbolic representation of compliance in the form of digital signatures is publicly accessible. On the other hand, the specialized units of authorities have no access to the sensitive non-public data on which algorithms make decisions. Providing compliance evidence to authorities is complicated by both internal and external factors.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022
DOI: 10.17148/IJARCCE.2022.111251

Compliance evidence is often non-standardized and consists of various artifacts that differ in format, language, or encoding. This heterogeneous nature of compliance evidence hampered the construction of a unified proof submission framework. Additionally, ML is often deployed as non-fungible model weights in docker containers, leading to difficult access to the model architecture, data processing, and training details. Before converting models to a translucent format to share proof submission with auditors, the absence of required infrastructure may prohibit compliance evidence sharing. Reputational risks, cybersecurity threats, and business losses are concerns constituents have in sharing compliance evidence.

VIII. PERFORMANCE OPTIMIZATION TECHNIQUES

In the past decade, there has been a significant increase in interest in designing end-to-end, cloud-scale, multi-tenant data platforms for providing real-time, AI-powered insights using large volumetric data. Nowadays, enterprises are collecting a massive amount of data, including video, audio, text, images, and more, at an unprecedented scale and in real time. At the same time, there has been a massive increase in the number and amount of compute- and storage-based services provided by the cloud. There are already hundreds of solutions and tools that can address different stages of the cloud-scale processing pipeline, from data collection to big data processing frameworks, OLAP databases, web serving tools, dashboards, and MLOps systems. Therefore, this project will design an end-to-end architecture that brings together some of the most popular, state-of-the-art solutions into a fully automated streaming-based ETL and CD pipeline for providing real-time, AI-powered insights at a cloud scale.

Today, data ingestion has become a priority for many organizations, regardless of shape or size. However, the ingestion process can be time-consuming, especially when datasets are very large, and estimates are needed on how long that will take. There is a rich history of developing algorithms that can efficiently estimate how long the ingestion process will take on traditional on-premise infrastructures, but the cloud contributes more additional and more diverse cloud services, which are not available in the traditional environment.

There are systematic techniques that can provide ingestion time estimates on the cloud by writing queries on the main underlying cloud services, mainly changing their geographic locations. It thus allows cloud scaling before ingestion and will be evaluated on a popular cloud-based data platform. It opens the door for wider adoption of data ingestion and more innovative designs for approaching data analysis.

Cloud scale has made agnostic bare-metal platforms and OS, bringing more scalable and cheaper cloud-based computation and storage engines popular. Systems originally designed for on-premise infrastructure, such as distributed pixel engines, graph engines, and multi-node databases, may require excessive adaptations to be able to run on cloud infrastructures, not only due to the lack of suitably flexible abstractions and interfaces on the cloud but also due to limitations in the design space of existing engines.

8.1. Monitoring and Tuning Performance

Cloud applications built on modern data platforms need to conform to stringent customer performance expectations that include low service latencies, high throughput, and reliability. The performance of cloud-scale data platforms is conditioned by many interacting software and hardware layers. Customer-facing services operate with queues of incoming requests that need to be processed and responded to. Backfill services are needed to ingest external data and perform warm starts. Common performance problems include service queues growing too long, backfilling services ingesting data slower than it is written, and hot parts of data decreasing the performance of cloud services. This section describes how to monitor and tune the performance of cloud-scale data platforms for real-time AI insights.

Modern data platforms monitor and visualize a multitude of performance metrics in a real-time UI console. Service queues of front-end servers are visualized alongside system topology diagrams. Performance alarms are triggered when the queue length of any front-end service is too long. Custom responses include executing built-in dashboards that provide real-time views of streaming ingestion latency and load-balancing history across ingestion servers.

Monitoring alone does not deliver high performance. Scaling cloud services up and out can improve service performance but often leads to higher operational cost. For instance, adding more ingest pipelines can leave some ingest servers under-utilized. Various monitoring and tuning techniques are needed to analyze ongoing performance problems and incrementally tune system-wide performance. Streamlining cloud customer jobs by re-distributing region graphs over servers can improve service throughput by 2-3X. Sophisticated stream trimming thresholds in data retention tuning can help improve ingestion performance by 30%.

DOI: 10.17148/IJARCCE.2022.111251

8.2. Cost Optimization Strategies

Enterprises engaged in critical cloud-based workloads tackle the dual challenge of keeping their workloads error-free and their cloud costs manageable. The ownership of cloud infrastructure inevitably affects the Cost of Goods Sold (COGS) of an enterprise. The COGS for cloud infrastructure is mainly influenced by cloud spending which has two parts – the Compute Cost and the Storage Cost. This section emphasizes various factors impacting cloud storage cost and discusses possible optimization strategies to minimize it. On a broad level, cloud providers offer a tiered form of storage with different expenditures. The pricing difference is due to the fundamental trade-off in the cloud between the level of redundancy and the accessibility of data. Cloud providers also charge extra costs for storing data having certain characteristics – large, out of tier, or rarely accessed. All these charges lead to substantial expenditures forcing enterprises to keep a closer watch on their storage requirements.

Preserving historic customer orientations, service records, live product details, etc., helps in better decision-making in organizations but results in a huge amount of data stored on cloud platforms. As a result, the volume of data is increasing massively and so is the number of datasets cloud enterprises are working with. At one extreme, data generation is moving on to the cloud, and consequently, cloud storage providers are offering services such as storing data blobs, tables, and files. At the other extreme, with massive data growth, a significant number of companies withdraw from the cloud, since the costs of storing and later accessing data are getting enormous. Cloud providers charge extra costs for cold datasets based on usage patterns and this sudden increase in Cost of Goods Sold (COGS) forces enterprises to keep a close watch on their datasets. Correcting costs of cloud storage can be achieved via an appropriate selection of restrictions – Tier Selection Restrictions, Cost Restrictions, Usage Restrictions, and Time Restrictions that are mathematically framed.

8.3. Scaling Solutions

In practice, AI applications of increasing size and complexity are often developed iteratively in stages. Most AI projects start with a Python notebook running on a single laptop without driving access to additional resources. After fine-tuning and validating a model, a data scientist may require more data but need quicker predictions. Such a need for speed may arise as a result of a new product rollout, data arrival, or the beginning of the holiday season, for example. In such scenarios, data preparation needs to be ramped up to daily or real-time processing. Raw data may have to be fetched from different sources, filtered and cleaned, combined, aggregated, or reshaped for analysis. On a larger scale, data pipelines need to be built to convert raw data into structured data pipelines analyzed, modeled, and optimized by a data science team.

The cost of building such a data pipeline from scratch can be excessively high. Part of the pain is that previous AI applications written in Python are not compatible with other languages or with other cloud services and tools commonly used for ETL operations. AI applications require new data sources, modified probability distributions, new models or changed parameters, or different toolkits or frameworks applied. Turning an AI model into a valid production system requires a dedicated engineering effort. Meanwhile, there is also a need to augment current data, such as historical data. Before modeling, data must be cleaned by removing duplicate data entries, free text parsing, addressing missing values, or filtering out outlier points. Analysis data must be aggregated, and graphic functions considering data granularity are applied. Subsequently, an initial model is built to provide insights into the data. Performance metrics, regularization, hyperparameters, and ensemble methods are tuned by the data science team. Insights into the data may differ among data scientists due to different understandings of the data or groups of data. When deploying the model in a production system, the pipeline must be designed and implemented. When the results derived from the model must be serialized and displayed, new questions or further analysis may arise that require feedback from the pipeline.

Unfortunately, building pipelines is a time-consuming and compositional task for data engineers due to the limited combinational capability of existing operators. Large-scale applications span data extraction, transformation, and load processes involving multiple distributed systems. Each system introduces heterogeneity and different programming paradigms. Current data-intensive workflows are designed for ad-hoc analysis of small, manageable data. Thus, they lead to low productivity and suboptimal resource usage. Distributed and cloud-resident implementations are difficult to environment engineer knowledge, resolve routing information, or efficiently control execution resources. For a machine learning engineer, the diffusion of different languages, platforms, and models requires more time to find fit solutions and less focus on modeling.

IX. CASE STUDIES

Data analysis applications have become more data-intensive and compute-intensive, driven by the unprecedented growth in data, the exponential increase in computing resources provided by Cloud / Cluster, and the rapidly evolving compute models from traditional Batch to interactive, streaming, and mixed. Multitude of choices for computation frameworks

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022
DOI: 10.17148/IJARCCE.2022.111251

with significantly different APIs and semantics to process the data, such as SQL, Workflow, MapReduce, Streaming, and AI/ML on Graphs, etc. However, the huge performance boost due to the fast growth in computing resources or the rapid evolution of computing frameworks could not be fully reaped by their users. Most computing frameworks are capable of hitting 80% or more performance potential in ideal computing environments for some simple compute workloads, however, it may take several months and a variety of skill sets or expertise for a data stream processing system to be configured to fully harvest the performance boost in real-world applications. C-Pipe with Smart Discovery and Intelligent Re-Optimizations can on-the-fly tune the data analytics applications. C-Pipe has been deployed and integrated as a drop-in module to the MTA, using the system auto-recommendation and self-reconfiguration, the systems can be reconfigured automatically and on-the-fly to draw 10-fold overall performance. To address the performance gap in building, deploying, and maintaining end-to-end AI pipelines on top of Cloud-scale Data Platforms, it proposes a unified, deep, automatic end-to-end toolkit, Complexo, of which three key components and AI design principles are introduced. Key Components: A low-code Application Interface to build end-to-end AI pipelines through a programming-in-the-linguistic-action approach. It enables Data Scientists to easily configure or customize big data computation operations to deploy native applications or third-party workloads, A high-level Automatic Optimization Interface, on the other hand, intuitively optimizes the whole AI pipeline through a programming-in-the-interaction-action approach. The optimization is on the wide range of optimization models, including the "use" optimization strategy such as custom integration & tuning AI ops, configuration tuning, etc, on the one hand, and the "structure" restructuring, displaying, and even modifying AI ops, on the other. A sub-symbolic Deep Optimization Engine, taking the AI and AO as the input and generating an optimal distributed execution plan is the last key component. It enables a new paradigm in end-to-end automatic optimization toolkits by pursuing the discovery-exploration optimization philosophy and satisfying the usual deep learning requirements of robustness, generalization, and online learnability. The end-to-end optimization generative learning model, techniques, and implementation are introduced.

9.1. Industry-Specific Implementations

A comprehensive approach for building end-to-end cloud-scale data platforms has been outlined. Many industries are deploying cloud-scale data solutions on various platforms. Based on unique requirements, industries are proposing these solutions in their ways, which are unique to themselves.

According to the needs of the service-based industry, architecture for cloud-scale data platforms is structured. In this effort, the need for cloud-scale data platforms is pointed out. A demand framework for cloud-scale data platforms is defined. Based on the demand framework, a cloud-scale data platform architecture is proposed and elaborated step-by-step. In the end, use case examples are provided to validate the practicality and efficacy of the suggested architecture.

Cloud-scale data platforms represent a comprehensive solution stack for cloud-scale data solutions. Such a solution stack provides a topology and set of working components for building cloud-scale data solutions. Cloud-scale data platforms can be structured by satisfying the individual needs of each particular domain. Tools and frameworks for each domain are of vital importance for building cloud-scale data platforms. As such, the need for framework-based cloud-scale data platforms comes up in many industries. In response, a well-structured approach is proposed to build cloud-scale data platforms by accommodating different requirements from various domains, which induces the idea of industry-specific implementations. Focus has been laid on the case study of the service-based industry, as an example of industry-specific implementations. Constructing the architecture of cloud-scale data platforms for the service-based industry is a nontrivial and challenging task inherently. This architecture has been investigated, defined, structured, and elaborated step-by-step. Cloud-scale data platforms have become the de facto solution for cloud-scale data analysis and management. By collecting, storing, managing, processing, analyzing, and visualizing cloud-scale data of various formats, they provide an environment for deriving insights and values from the raw data. In the past few years, many platforms for building cloud-scale data solutions have been developed, forming a solution stack that contains tools covering most of the phases of cloud-scale data analysis and management. All these platforms, which have unique and custom-specific implementations, are in the same format as plug-ins. As such, the development of a framework for organizing, integrating, managing, and orchestrating these tools is important for decreasing the cost and effort of adopting cloud-scale data platforms.

9.2. Success Stories

E-commerce Recommendation Engine: Mastercard engineers built an end-to-end AI platform called Ace that easily integrates big data analysis and deep learning AI applications in a single environment. It includes data ingestion, preparation, training, model debugging, monitoring, and serving features. This platform, built on open-source technologies, scales from local laptops to distributed cluster data centers. It is first used in Lambda architecture to analyze transactions in near real-time and generate alerting events for potentially fraudulent transactions. As the weekly data volume grows to over 400TB, the multidimensional analysis demands on big data increase. These led to using Spark for

DOI: 10.17148/IJARCCE.2022.111251

batch analysis of historical data for 30+ days and using streaming analytics engines such as Storm and Flink for real-time processing of transactions. Daily architectures now run on new TPC-DS data, but the interrogation can take days or prevent it from running entirely due to unexpected data distributions.

Biological Activity Prediction for Drug Discovery: Deep learning for structured data problems is extensively applied in cheminformatics, bioinformatics, and in predicting certain biological activity through complex chemical and biological structures. With a passion for rapid drug discovery and development in the race to medicine, the goal of the project was to minimize the time and costs of drug discovery and development while maximizing safety and efficacy. This is accomplished through the application of deep learning to data-driven drug science with a focus on quantitative structure-activity relationship models for predicting drug activity. This includes applying statistical machine learning methods for drug discovery together with chemoinformatics and bioinformatics domain knowledge by pharmaceutical scientists. Attempts to broaden the applicability of chemoinformatic representation and use deep learning methods are also made. This is completed through open-source development efforts in data science libraries and platforms that have already been well undertaken in industrial communities.

9.3. Lessons Learned

Deep learning workloads can heavily use GPUs due to their acceleration for matrix multiplication and convolutions. Tasks can contain heterogeneous workloads including pre-processing steps, distributed training, hyper-parameter search, and batch inference, which are often implemented in various languages and frameworks. A computing environment can combine on-premise infrastructure and multiple clouds providing different resources with distinct pricing models. Managed solutions can provide ease of use and setup. However, they add complexity to dealing with multiple environments. Specifically, different cloud providers offer a variety of storage and compute resources priced differently and upon different usage conditions. A computing environment can combine on-premise infrastructure and multiple clouds providing different resources and distinct pricing models. Moreover, typically pricing costs stay the same, while on-demand resources are usually the most expensive and can change pricing models significantly. Preemptable or spot instances are significantly cheaper but less stable and can cancel a task at any point in time.
Even when using a managed solution, there appears a problem with how to distribute tasks to different resources. Tasks can contain complex machine learning workloads with heterogeneous operations: pre-processing and batching inference utilizing CPUs while training, hyper-parameter search, and model retraining being executed on GPUs. Each workload can be cross-implementable across different languages and deep learning frameworks. Some computations hint at language, some are language agnostic while requiring an ML framework-specific compiler, and some are hybrid requiring both an ML compiler and deployment engine. Each workload requires specific resources and compute clusters.

Cloud costs impose stronger restrictions, multiple clouds provide different resources and pricing models with additional charges for accessing remote resources. Exploit cheaper preemptible instances that cancel tasks or models can simply drop down in accuracy requiring remediation and fine-tuning. A remarkable approach can be using a narrow band with only one or two neighbors of each model, instead of a yield limit all neighbors can contribute to reducing overfitting. Each cloud provider can be weighed by many different parameters having a different impact depending on scale by market capital per cloud provider. This situation can be complex and too dramatic. Human management alone simply cannot deal with it.

X. FUTURE TRENDS IN DATA PLATFORMS

Real-time AI has become ubiquitous, generating insights in milliseconds within cloud-scale data platforms ingesting millions of events per second. Much progress has been made in building these real-time AI engines end-to-end, with a focus on enabling advanced mission-critical AI use cases while delivering cloud-scale performance and high availability. Modern cloud-scale data platforms comprise multiple components within each layer and span multiple layers of the architecture. Integration, scalability, performance, and operationalization are key focus areas that haven't received sufficient attention in the literature. In the future, these challenges will be ever more prominent due to a dramatically growing range of use cases and a race to cloud and automation across industries. Via their unique perspectives, this covers forward-looking trends that will gather momentum over the next few years across both technological and business model fronts.

Big players such as Meta, Microsoft, Google, Amazon, and Alibaba have been investing heavily in AI for many years with recent success. The big cloud providers have built a comprehensive set of end-to-end services over the years, covering every piece of the cloud. However, no one component is designed to work perfectly as a building blocks for scalable and affordable enterprise AI. There are many potential opportunities to evolve tools such as observability insights, data lineage, and automated data preparation powered by smarter, task- and resource-aware models.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111251

New and improved service/functionality interfaces enabling efficient building and deployment of end-to-end data pipelines must also be developed. Open-sourcing cloud scale solutions/tooling, cooperation across companies, and machine learning efficacy enabled by extensive workloads will gradually normalize the offerings of new products and services. However, many entrenched technologies have become too sophisticated and normalized to ever host a new competition, such as graphics processors, tensor accelerators, and modern databases.

A paradigm shift in how organizations view their data will take place over the next decade. Besides business intelligence, analytics shall significantly impact enterprise credibility and interactions with the world. Organizations will become smarter and build web insights and tools to simplify and automate interactions with customers, competitors, supply chains, and regulatory bodies. Cloud computing shall dramatically change the cost structure on ownership and usage of IT and data/insights pipelines. This shall incentivize the rapid emergence/exponential growth of new cloud-based firms and services embracing the multi-tenant model. Organizations historically large in size will either become bankrupt or gradually crushed by the cloud revolution, leading to disassembly to a confederation of cloud upstarts, and/or reinvention and rebuilding in the cloud.

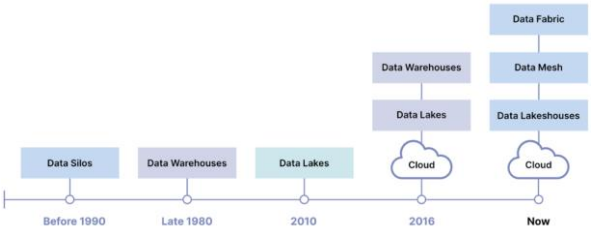

Fig 5: Future Trends in Data Platforms.

10.1. Emerging Technologies

The ever-growing amounts of data generated, collected, and stored result in forms of data that challenge the current database management systems. Such voluminous data is commonly referred to as “Big Data”, and no single vanilla DBMS vendor can battle the climate of data complexity and its demands alone. To collect, store, and analyze rapidly increasing amounts of mostly distributed and heterogeneous data continuously is a huge, exciting, and challenging opportunity that needs to be addressed with new creative approaches. Such “Cloud-Scale” is achieved via joint multi-faceted, multi-layered, and multi-dimensional optimization of diverse components of these multi-component data platforms. Work in this area will continue to emphasize the interplay between technologies to find the best design, scheduling, and configuration of these flexible components given various SLAs and diverse data workloads.

Large Cloud Service Providers operate various forms of multi-component data platforms that integrate various kinds of data services. Such cloud-scale data platforms pose new challenges and exciting opportunities for research and innovation. Paramount research challenges include designing robust and flexible ambient microservices for the multi-data services platforms, optimizing their designs under various multi-faceted, multi-layered, and multi-dimensional cost objectives, and regulating various aspects of operational practices of the platform and evolving and co-evolving their components in real-time.

10.2. Predictions for AI and Data Integration

Significant progress has been made toward developing and deploying systems that embody these recommendations over the past few years. It is predicted that organizations will leverage it to build integrated systems that provide rapid evaluation and integration of alternative products based on complex data. As a result, organizations will be able to obtain higher-quality predictions through domain-guided search in any domains for which suitable models can be deployed. This integrated capability can be deployed in a data basin to protect the data collection processes from service interruptions, data corruption, and data loss. AI and data integration technologies will provide organizations with early signs of emerging inflection points as determined by sudden shifts in trends, changes in key features measured across large data sets, and changes in data quality or integrity. AI will be able to automate the majority of required product development tasks and suggest detailed reusable models that match the requirements of the new problem. A significant breakthrough in AI service quality and widespread adoption in many industries are expected.

AI and data integration technologies will expedite the design of tailored data generation mechanisms that will replace deep learning models for a wide variety of data types. Noise will be automatically incorporated based on model and task needs. End-to-end systems for generating arbitrarily large data that comply with complex nonuniform constraints will be

DOI: 10.17148/IJARCCE.2022.111251

developed. More image detail, wider video scene coverage, and new video product types will be provided. Text length and vocabulary types that were previously unthinkable for typical dialogues will be generated while adhering to the rules of natural human interaction. In addition, AI will design entirely new physical manufacturing processes to implement the above products and data generation screening and buying agents will be designed that understand target user requirements. AI and data integration technologies will enable real-time awareness of production environments, production stream characterization based on historical and real-time data, rapid identification of outliers and slow-moving streams, and development of refined models for finer quality predictions. Additionally, AI and data integration technologies will assist in automating product stream and production equipment management via expert and active learning, respectively.

10.3. Impact of Quantum Computing

Research in the field of quantum algorithms, machines, and simulators is likely to carry on quickly and gain significantly more recognition over the foreseeable future. Despite an often inherent hype, quantum computing has the potential to substantially impact the development of such applications and their prototype implementations on laser-chip or superconducting quantum hardware. The situation is similar to the first presentation of classical computers after World War II. A wealth of algorithms for direct applications such as crypto-breaking or realistic tasks like error-corrected simulation, large optimization problems or AI datasets have been presented, but the noise, limited connectivity, and connectivity mappings of such early classical machines pose enormous challenges to sizable implementations.

Thus, several research avenues and still unsolved theoretical questions are open for discussion and exploration, for instance, the question of how to evaluate the efficacy of non-exponential speedup claims and to give guarantees on how quantum or classical solvers would perform on a given problem instance beforehand. These questions are especially pertinent for complex real-world simulation or optimization problems, which are often approximated or abstracted from large complex sets based on a reduced view. However, as can be seen from the PRIMES project, quite reasonable and often easily computable approximations or abstractions may greatly condense extremely large complex problems while staying NP-hard, and finding appropriate strategies may be as difficult as the original problem.

Such efficiency bounds and complexity considerations would assist in understanding the advantages and limitations of classical and quantum solvers clearly. In addition, the exploration of other quantum-inspired schemes that endow classical algorithms with replenished coolness or other avenues to speed up classical algorithms or important and similar complexities for learning data pools are pointed to promoting discussions on the potential impact of quantum computing.

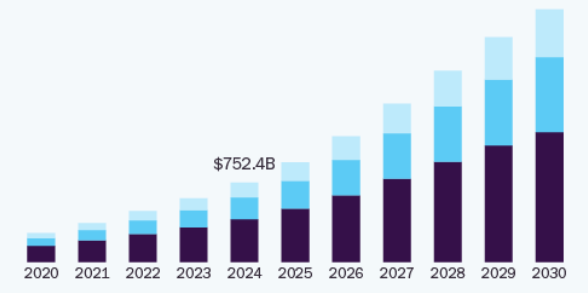

Fig 6: Cloud-Scale Data Platforms for Real-Time AI Insights.

XI. CONCLUSION

This paper describes the latest work on building end-to-end cloud-scale data platforms for real-time AI insights. In the past few years, efforts have focused on cloud infrastructure and building data management services from the ground up on the cloud, with the experience of building and running machines with petabytes of storage to build general-purpose cloud database services and designing real-time analytics systems to analyze hundreds of terabytes of web data every day, as well as streaming data management platforms capable of ingesting and processing tens of millions of records daily. The idea of turning telemetry monitoring and data measurement into customer insights has also emerged. As data continues to grow in volume, variety, and velocity, there is renewed interest in building systems to extract useful insights from it. To address this challenge, an AI-augmented approach has been developed to build futuristic data service systems augmented by ML techniques that help manage data, speed up DB design, and enhance AI-assisted developer experience. These services allow intelligent data design, accurate choice of data engines, structured data programming, explanatory query diagnosis against time, and good input sampling selection among others. Scientific discoveries and engineering

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 11, Issue 12, December 2022

DOI: 10.17148/IJARCCE.2022.111251

breakthroughs have driven progress in many domains, including artificial intelligence, microbiology, robotics, observatory networks, and sensor paradigms. The advancement in computer or data-powered techniques has led to the state of the art machine learning or data management approaches. Advances in AI and machine learning have further allowed breakthroughs in domains such as drug discovery, healthcare, cell biology, image processing, and pandemic prevention. Such developments have led to more data being collected and exploited, introducing new challenges and opportunities to the data management community.

REFERENCES

[1] Kommaragiri, V. B., Preethish Nanan, B., Annapareddy, V. N., Gadi, A. L., & Kalisetty, S. (2022). Emerging Technologies in Smart Computing, Sustainable Energy, and Next-Generation Mobility: Enhancing Digital Infrastructure, Secure Networks, and Intelligent Manufacturing. Venkata Narasareddy and Gadi, Anil Lokesh and Kalisetty, Srinivas.

[2] Pamisetty, V., Dodda, A., Singireddy, J., & Challa, K. (2022). Optimizing Digital Finance and Regulatory Systems Through Intelligent Automation, Secure Data Architectures, and Advanced Analytical Technologies. Jeevani and Challa, Kishore, Optimizing Digital Finance and Regulatory Systems Through Intelligent Automation, Secure Data Architectures, and Advanced Analytical Technologies (December 10, 2022).

[3] Paleti, S. (2022). The Role of Artificial Intelligence in Strengthening Risk Compliance and Driving Financial Innovation in Banking. International Journal of Science and Research (IJSR), 11(12), 1424–1440. https://doi.org/10.21275/sr22123165037

[4] Komaragiri, V. B. (2022). Expanding Telecom Network Range using Intelligent Routing and Cloud-Enabled Infrastructure. International Journal of Scientific Research and Modern Technology, 120–137. https://doi.org/10.38124/ijsrmt.v1i12.490

[5] Pamisetty, A., Sriram, H. K., Malempati, M., Challa, S. R., & Mashetty, S. (2022). AI-Driven Optimization of Intelligent Supply Chains and Payment Systems: Enhancing Security, Tax Compliance, and Audit Efficiency in Financial Operations. Tax Compliance, and Audit Efficiency in Financial Operations (December 15, 2022).

[6] Mashetty, S. (2022). Innovations In Mortgage-Backed Security Analytics: A Patent-Based Technology Review. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3826

[7] Kurdish Studies. (n.d.). Green Publication. https://doi.org/10.53555/ks.v10i2.3785

[8] Motamary, S. (2022). Enabling Zero-Touch Operations in Telecom: The Convergence of Agentic AI and Advanced DevOps for OSS/BSS Ecosystems. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3833

[9] Kannan, S. (2022). AI-Powered Agricultural Equipment: Enhancing Precision Farming Through Big Data and Cloud Computing. Available at SSRN 5244931.

[10] Suura, S. R. (2022). Advancing Reproductive and Organ Health Management through cell-free DNA Testing and Machine Learning. International Journal of Scientific Research and Modern Technology, 43–58. https://doi.org/10.38124/ijsrmt.v1i12.454

[11] Nuka, S. T., Annapareddy, V. N., Koppolu, H. K. R., & Kannan, S. (2021). Advancements in Smart Medical and Industrial Devices: Enhancing Efficiency and Connectivity with High-Speed Telecom Networks. Open Journal of Medical Sciences, 1(1), 55-72.

[12] Meda, R. (2022). Integrating IoT and Big Data Analytics for Smart Paint Manufacturing Facilities. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3842

[13] Annapareddy, V. N., Preethish Nanan, B., Kommaragiri, V. B., Gadi, A. L., & Kalisetty, S. (2022). Emerging Technologies in Smart Computing, Sustainable Energy, and Next-Generation Mobility: Enhancing Digital Infrastructure, Secure Networks, and Intelligent Manufacturing. Venkata Bhardwaj and Gadi, Anil Lokesh and Kalisetty, Srinivas, Emerging Technologies in Smart Computing, Sustainable Energy, and Next-Generation Mobility: Enhancing Digital Infrastructure, Secure Networks, and Intelligent Manufacturing (December 15, 2022).

[14] Phanish Lakkarasu. (2022). AI-Driven Data Engineering: Automating Data Quality, Lineage, And Transformation In Cloud-Scale Platforms. Migration Letters, 19(S8), 2046–2068. Retrieved from https://migrationletters.com/index.php/ml/article/view/11875

[15] Kaulwar, P. K. (2022). Securing The Neural Ledger: Deep Learning Approaches For Fraud Detection And Data Integrity In Tax Advisory Systems. Migration Letters, 19, 1987-2008.

[16] Malempati, M. (2022). Transforming Payment Ecosystems Through The Synergy Of Artificial Intelligence, Big Data Technologies, And Predictive Financial Modeling. Big Data Technologies, And Predictive Financial Modeling (November 07, 2022).

[17] Recharla, M., & Chitta, S. (2022). Cloud-Based Data Integration and Machine Learning Applications in Biopharmaceutical Supply Chain Optimization.

[18] Lahari Pandiri. (2022). Advanced Umbrella Insurance Risk Aggregation Using Machine Learning. Migration Letters, 19(S8), 2069–2083. Retrieved from https://migrationletters.com/index.php/ml/article/view/11881

DOI: 10.17148/IJARCCE.2022.111251

[19] Paleti, S., Burugulla, J. K. R., Pandiri, L., Pamisetty, V., & Challa, K. (2022). Optimizing Digital Payment Ecosystems: Ai-Enabled Risk Management, Regulatory Compliance, And Innovation In Financial Services. Regulatory Compliance, And Innovation In Financial Services (June 15, 2022).

[20] Singireddy, J. (2022). Leveraging Artificial Intelligence and Machine Learning for Enhancing Automated Financial Advisory Systems: A Study on AIDriven Personalized Financial Planning and Credit Monitoring. Mathematical Statistician and Engineering Applications, 71 (4), 16711–16728.

[21] Paleti, S., Singireddy, J., Dodda, A., Burugulla, J. K. R., & Challa, K. (2021). Innovative Financial Technologies: Strengthening Compliance, Secure Transactions, and Intelligent Advisory Systems Through AI-Driven Automation and Scalable Data Architectures. Secure Transactions, and Intelligent Advisory Systems Through AI-Driven Automation and Scalable Data Architectures (December 27, 2021).

[22] Sriram, H. K. (2022). Integrating generative AI into financial reporting systems for automated insights and decision support. Available at SSRN 5232395.

[23] Koppolu, H. K. R. (2021). Leveraging 5G Services for Next-Generation Telecom and Media Innovation. International Journal of Scientific Research and Modern Technology, 89–106. https://doi.org/10.38124/ijsrmt.v1i12.472

[24] End-to-End Traceability and Defect Prediction in Automotive Production Using Blockchain and Machine Learning. (2022). International Journal of Engineering and Computer Science, 11(12), 25711-25732. https://doi.org/10.18535/ijecs.v11i12.4746

[25] Chaitran Chakilam. (2022). AI-Driven Insights In Disease Prediction And Prevention: The Role Of Cloud Computing In Scalable Healthcare Delivery. Migration Letters, 19(S8), 2105–2123. Retrieved from https://migrationletters.com/index.php/ml/article/view/11883

[26] Sriram, H. K., ADUSUPALLI, B., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks.

[27] Avinash Pamisetty. (2021). A comparative study of cloud platforms for scalable infrastructure in food distribution supply chains. Journal of International Crisis and Risk Communication Research , 68–86. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/2980

[28] Gadi, A. L., Kannan, S., Nanan, B. P., Komaragiri, V. B., & Singireddy, S. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization. Universal Journal of Finance and Economics, 1(1), 87-100.

[29] Dodda, A. (2022). The Role of Generative AI in Enhancing Customer Experience and Risk Management in Credit Card Services. International Journal of Scientific Research and Modern Technology, 138–154. https://doi.org/10.38124/ijsrmt.v1i12.491

[30] Gadi, A. L. (2022). Connected Financial Services in the Automotive Industry: AI-Powered Risk Assessment and Fraud Prevention. Journal of International Crisis and Risk Communication Research, 11-28.

[31] Pamisetty, A. (2022). A Comparative Study of AWS, Azure, and GCP for Scalable Big Data Solutions in Wholesale Product Distribution. International Journal of Scientific Research and Modern Technology, 71–88. https://doi.org/10.38124/ijsrmt.v1i12.466

[32] Adusupalli, B. (2021). Multi-Agent Advisory Networks: Redefining Insurance Consulting with Collaborative Agentic AI Systems. Journal of International Crisis and Risk Communication Research, 45-67.

[33] Dwaraka Nath Kummari. (2022). Iot-Enabled Additive Manufacturing: Improving Prototyping Speed And Customization In The Automotive Sector . Migration Letters, 19(S8), 2084–2104. Retrieved from https://migrationletters.com/index.php/ml/article/view/11882

[34] Data-Driven Strategies for Optimizing Customer Journeys Across Telecom and Healthcare Industries. (2021). International Journal of Engineering and Computer Science, 10(12), 25552-25571. https://doi.org/10.18535/ijecs.v10i12.4662

[35] Adusupalli, B., Singireddy, S., Sriram, H. K., Kaulwar, P. K., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks. Universal Journal of Finance and Economics, 1(1), 101-122.

[36] AI-Based Financial Advisory Systems: Revolutionizing Personalized Investment Strategies. (2021). International Journal of Engineering and Computer Science, 10(12). https://doi.org/10.18535/ijecs.v10i12.4655

[37] Karthik Chava. (2022). Harnessing Artificial Intelligence and Big Data for Transformative Healthcare Delivery. International Journal on Recent and Innovation Trends in Computing and Communication, 10(12), 502–520. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11583

[38] Challa, K. (2022). The Future of Cashless Economies Through Big Data Analytics in Payment Systems. International Journal of Scientific Research and Modern Technology, 60–70. https://doi.org/10.38124/ijsrmt.v1i12.467

[39] Pamisetty, V., Pandiri, L., Annapareddy, V. N., & Sriram, H. K. (2022). Leveraging AI, Machine Learning, And Big Data For Enhancing Tax Compliance, Fraud Detection, And Predictive Analytics In Government Financial

DOI: 10.17148/IJARCCE.2022.111251

Management. Machine Learning, And Big Data For Enhancing Tax Compliance, Fraud Detection, And Predictive Analytics In Government Financial Management (June 15, 2022).

[40] Innovations in Spinal Muscular Atrophy: From Gene Therapy to Disease-Modifying Treatments. (2021). International Journal of Engineering and Computer Science, 10(12), 25531-25551. https://doi.org/10.18535/ijecs.v10i12.4659

[41] Kaulwar, P. K. (2022). Data-Engineered Intelligence: An AI-Driven Framework for Scalable and Compliant Tax Consulting Ecosystems. Kurdish Studies, 10 (2), 774–788.

[42] Operationalizing Intelligence: A Unified Approach to MLOps and Scalable AI Workflows in Hybrid Cloud Environments. (2022). International Journal of Engineering and Computer Science, 11(12), 25691-25710. https://doi.org/10.18535/ijecs.v11i12.4743

[43] Nandan, B. P., & Chitta, S. (2022). Advanced Optical Proximity Correction (OPC) Techniques in Computational Lithography: Addressing the Challenges of Pattern Fidelity and Edge Placement Error. Global Journal of Medical Case Reports, 2(1), 58-75.

[44] Raviteja Meda. (2021). Machine Learning-Based Color Recommendation Engines for Enhanced Customer Personalization. Journal of International Crisis and Risk Communication Research , 124–140. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3018

[45] Rao Suura, S. (2021). Personalized Health Care Decisions Powered By Big Data And Generative Artificial Intelligence In Genomic Diagnostics. Journal of Survey in Fisheries Sciences. https://doi.org/10.53555/sfs.v7i3.3558

[46] Implementing Infrastructure-as-Code for Telecom Networks: Challenges and Best Practices for Scalable Service Orchestration. (2021). International Journal of Engineering and Computer Science, 10(12), 25631-25650. https://doi.org/10.18535/ijecs.v10i12.4671

[47] Vamsee Pamisetty, Lahari Pandiri, Sneha Singireddy, Venkata Narasareddy Annapareddy, Harish Kumar Sriram. (2022). Leveraging AI, Machine Learning, And Big Data For Enhancing Tax Compliance, Fraud Detection, And Predictive Analytics In Government Financial Management. Migration Letters, 19(S5), 1770–1784. Retrieved from https://migrationletters.com/index.php/ml/article/view/11808

[48] Someshwar Mashetty. (2020). Affordable Housing Through Smart Mortgage Financing: Technology, Analytics, And Innovation. International Journal on Recent and Innovation Trends in Computing and Communication, 8(12), 99–110. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11581

[49] Srinivasa Rao Challa,. (2022). Cloud-Powered Financial Intelligence: Integrating AI and Big Data for Smarter Wealth Management Solutions. Mathematical Statistician and Engineering Applications, 71(4), 16842–16862. Retrieved from https://philstat.org/index.php/MSEA/article/view/2977

[50] Paleti, S. (2022). Fusion Bank: Integrating AI-Driven Financial Innovations with Risk-Aware Data Engineering in Modern Banking. Mathematical Statistician and Engineering Applications, 71(4), 16785-16800.

[51] Pamisetty, V. (2022). Transforming Fiscal Impact Analysis with AI, Big Data, and Cloud Computing: A Framework for Modern Public Sector Finance. Big Data, and Cloud Computing: A Framework for Modern Public Sector Finance (November 30, 2022).

[52] Kommaragiri, V. B., Gadi, A. L., Kannan, S., & Preethish Nanan, B. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization.

[53] Annapareddy, V. N. (2022). Integrating AI, Machine Learning, and Cloud Computing to Drive Innovation in Renewable Energy Systems and Education Technology Solutions. Available at SSRN 5240116.

[54] Transforming Renewable Energy and Educational Technologies Through AI, Machine Learning, Big Data Analytics, and Cloud-Based IT Integrations. (2021). International Journal of Engineering and Computer Science, 10(12), 25572-25585. https://doi.org/10.18535/ijecs.v10i12.4665

[55] Venkata Bhardwaj Komaragiri. (2021). Machine Learning Models for Predictive Maintenance and Performance Optimization in Telecom Infrastructure. Journal of International Crisis and Risk Communication Research , 141–167. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3019

[56] Paleti, S. (2021). Cognitive Core Banking: A Data-Engineered, AI-Infused Architecture for Proactive Risk Compliance Management. AI-Infused Architecture for Proactive Risk Compliance Management (December 21, 2021).

[57] Harish Kumar Sriram. (2022). AI-Driven Optimization of Intelligent Supply Chains and Payment Systems: Enhancing Security, Tax Compliance, and Audit Efficiency in Financial Operations. Mathematical Statistician and Engineering Applications, 71(4), 16729–16748. Retrieved from https://philstat.org/index.php/MSEA/article/view/2966

[58] Chava, K., Chakilam, C., Suura, S. R., & Recharla, M. (2021). Advancing Healthcare Innovation in 2021: Integrating AI, Digital Health Technologies, and Precision Medicine for Improved Patient Outcomes. Global Journal of Medical Case Reports, 1(1), 29-41.

DOI: 10.17148/IJARCCE.2022.111251

[59] Data Engineering Architectures for Real-Time Quality Monitoring in Paint Production Lines. (2020). International Journal of Engineering and Computer Science, 9(12), 25289-25303. https://doi.org/10.18535/ijecs.v9i12.4587

[60] Pallav Kumar Kaulwar. (2021). From Code to Counsel: Deep Learning and Data Engineering Synergy for Intelligent Tax Strategy Generation. Journal of International Crisis and Risk Communication Research , 1–20. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/2967

[61] Pandiri, L., & Chitta, S. (2022). Leveraging AI and Big Data for Real-Time Risk Profiling and Claims Processing: A Case Study on Usage-Based Auto Insurance. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3760

[62] Kummari, D. N. (2022). AI-Driven Predictive Maintenance for Industrial Robots in Automotive Manufacturing: A Case Study. International Journal of Scientific Research and Modern Technology, 107–119. https://doi.org/10.38124/ijsrmt.v1i12.489

[63] Gadi, A. L. (2022). Cloud-Native Data Governance for Next-Generation Automotive Manufacturing: Securing, Managing, and Optimizing Big Data in AI-Driven Production Systems. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3758

[64] Dodda, A. (2022). Secure and Ethical Deployment of AI in Digital Payments: A Framework for the Future of Fintech. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3834

[65] Gadi, A. L. (2021). The Future of Automotive Mobility: Integrating Cloud-Based Connected Services for Sustainable and Autonomous Transportation. International Journal on Recent and Innovation Trends in Computing and Communication, 9(12), 179-187.

[66] Dodda, A. (2022). Strategic Financial Intelligence: Using Machine Learning to Inform Partnership Driven Growth in Global Payment Networks. International Journal of Scientific Research and Modern Technology, 1(12), 10-25.

[67] Just-in-Time Inventory Management Using Reinforcement Learning in Automotive Supply Chains. (2021). International Journal of Engineering and Computer Science, 10(12), 25586-25605. https://doi.org/10.18535/ijecs.v10i12.4666

[68] Srinivasa Rao Challa. (2021). From Data to Decisions: Leveraging Machine Learning and Cloud Computing in Modern Wealth Management. Journal of International Crisis and Risk Communication Research , 102–123. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3017

[69] Kommaragiri, V. B. (2021). Enhancing Telecom Security Through Big Data Analytics and Cloud-Based Threat Intelligence. Available at SSRN 5240140.