International Journal of Scientific Research and Modern Technology (IJSRMT) *ijsrmt.com*
Volume 1, Issue 12, 2022
DOI: **https://doi.org/10.38124/ijsrmt.v1i12.466**

A Comparative Study of AWS, Azure, and GCP for Scalable Big Data Solutions in Wholesale Product Distribution

Avinash, Pamisetty[1]

[1] Mulesoft Developer, Farmers Insurance

Publication Date: 2022/12/28

Abstract

Due to the increasing volume of information generated worldwide, every day more companies are in search of new ways to process and gain business value from data. Part of this circumstance can be attributed to the constant growth of the Internet, which is a factor of considerable relevance for the creation of New Information and a binding constraint for the collection of a Big Set of Data. In this way, organizations that have data in enormous volume and variety have, at their disposal nowadays, new forms of treatment of this information that guarantee efficiency in the decision-making process. In this niche, Cloud Computing is inserted, breaking paradigms on the way that companies treat their data. Thus, with the democratization of Data Processing, an increasing number of companies are looking for Cloud Providers capable of putting Data Processing as a Service. In this context, the goal of this research is to perform a comparative study of the main providers of Cloud Computing services, analyzing the main features of their respective products, pros and cons, architectural frameworks, tools, set of features, and pricing. In addition, special attention was given to the services offered for the creation of Data Lakes, as well as the possibilities offered by the companies studied for the processing of Big Data in a way that allows obtaining business value in business solutions. The present study is exploratory and descriptive, employing a qualitative and quantitative approach. Finally, this research presented a consideration of a possible architecture for the available tools and models. The results pointed to a favorable scenography for the use of technologies by companies, given the new features available for their tools, as well as the lower cost and greater scalability, given the recent price reductions in the services offered. Finally, as an informal contribution to the community, we presented a set of points to be considered for the decision on a Database solution for Big Data.

Keywords: *Data Processing, Internet Growth, Big Data, Cloud Computing, Data Processing as a Service, Cloud Providers, Data Lakes, Architectural Frameworks, Comparative Study, Pros and Cons, Business Value, Decision Making, Scalable Solutions, Price Reductions, Qualitative Analysis, Quantitative Approach, Exploratory Research, New Information, Digital Transformation, Database Solutions.*

I. INTRODUCTION

Analyzing big data is becoming important in various domains such as healthcare, climatology, and marketing among others. The introduction of big data as a revolution in storage and analytical capabilities has doubled the amount of physical data created by 2612% since the first report on the subject. The growing demand for Collaboration, Communication, Sourcing and Supplier Management, Procurement Management, Inventory Management, Warehousing, Demand Management, Order Fulfillment, Logistics Management and Management have added to the volume data generated by the wholesale distribution industry. The volume, variety and velocity at which this data is created has expounded the challenge of analyzing wholesale product distribution data. Organizations must derive valuable insights from wholesale product distribution

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*, *1*(12), 71–88.
https://doi.org/10.38124/ijsrmt.v1i12.466

data with an increasing focus on performance improvement, reducing costs and increasing revenue.

The volume, variety and velocity of the data generated by the wholesale distribution industry has created the demand for a scalable big data solution. The shipping and receiving coupled with exchanges and sales has created a wide array of physical distribution data repositories such as the warehouse management repositories, order entry repositories, billing repositories, data marts, and data warehouses. A solution that is both scalable and is capable of analyzing the big data has to be cost-efficient and easy to use. The common set of five big data analytics tools, which are incorporated to analyze big data with high velocity are cross-platform, distributed, disk-based, in-memory and in-warehouse analytics tools. The cloud-based services offered by key players are becoming increasingly popular as organizations look to leverage the cloud computing capabilities. The objective of this study is to compare the big data services offered by the key players in the cloud-based analytics tool space. The services will be compared with regard to their Analytics Tool Type, Scalability, Ease of Use and Pricing.

➢ *Objectives and Significance of the Study*

Big Data is considered one of the most important technologies of the present and, possibly the future. Given the vast amounts of data generated by and within organizations every moment of every day, how organizations make use of the data they have at their disposal shapes how they survive in a competitive world. Enterprises that build scalable data solutions capable of analyzing Big Data can gain insights that improve decision making and drive product and service innovation. Indeed, the data-crunchers and analysts of the Big Data era are armed with the prowess to advise their organizations on when and where to strike, and what opportunities to detect, and those who do can stay ahead of the curve. The cloud, and the sheer amount of data storage it offers, is one area of focus, because companies that can solve the increasing costs of storing the explosive amounts of data can gain low-cost power for the future and a major leg-up on their competition. This has seen the growth of several cloud providers, each offering their own unique set of technology solutions.

Organizations increasingly prefer cloud for data storage due to their ability to easily scale capacity and pay only for their usage. These companies can therefore avoid the prohibitive costs of purchasing the physical data storage infrastructure, which needs to be considerably large and up to date, and be maintained by a dedicated and trained IT team. Public cloud providers also have the advantage of being trusted vendors with the capabilities to ensure consistent uptime and stringent security policies. However, it has become pertinent for organizations to know the strengths and limitations of each cloud provider before choosing the provider to host their services, especially critical management systems, to avoid negative impacts on performance and reliability. This is vital as providers have a differentiated set of tools for the described area. After reviewing these platforms, organizations may opt to move their data solutions to the cloud environment.

Storage Services	AWS	Azure	Google
Object Storage Service for Use Cases	Simple Storage Services (S3)	Storage (Block Blob)	Cloud Storage
Archive Storage	S3 Infrequent Access Glacier Data Archive	Storage (Cool) Storage (Archive)	Nearline Coldline
Hybrid Storage	Storage Gateway	StorSimple	Egnyte Sync
Automatic Protection and Disaster Recovery	Disaster Recovery	Site Recovery	
Bulk Data Transfer Solutions	Import/Export Disk Snowball Edge SnowMobile	Import/Export Azure Data Box	Storage Transfer Service
Backup	Object Storage Cold Archive Storage Storage Gateway	Backup	-

Fig 1 AWS Vs Azure Vs Google

➢ *Overview of Big Data in Wholesale Distribution*

Data is an important factor for the informed decision-making of wholesale distributors, allowing them to adjust their actions relative to changes in purchasing behaviors, inventory levels, partner conditions, and supplier strengths. Any data that a wholesale distributor needs for business

decisions and performance measurement touts the key assets of big data in wholesale distribution. For example, a distributor's proprietary data could consist of information from operations, customer and product data, logistics data, or purchasing data.

Omni-channel is a game-changer for wholesale distribution, and big data/integration is the game-changer technology. Wholesale distributors who are on top of their key performance metrics and leverage technology to have speed and ease of execution will have a significant advantage. In addition, analyzing data with the latest analytical tools will allow distributors to have visibility into their business and provide their customers with the best solution in the right channel at the right time. Predicting customer needs by leveraging availability data, product shelf life data, or data regarding usage differences by channel and party will allow wholesale distributors to fuel their 3Cs. Distributors' specialization in private label or e-commerce will play a role on the "how innovation" side.

Equation 1 Scalability Efficiency Metric

$$S_e = \frac{Q_p}{T_s \cdot C_u}$$

S_e = Scalability efficiency

Q_p = Quantity of processed product data (e.g., SKUs/hour)

T_s = Time taken to scale infrastructure

C_u = Cloud resource usage cost during scaling

(Lower T_s, C_u, and higher Q_p imply better scalability performance)

- *Key Trends and Innovations in Big Data for Wholesale Distribution*

An increasing majority of wholesale distribution companies consider data to be a game-changing tool. For many, comprehensive, analyzed data is the most important asset, far surpassing any products or service qualities. Wholesale distributors are hoping the use of big data can help them cut margins while standing out from their competitors. A key trend among big data usage in wholesale distribution companies is predictive analysis. By summoning traffic data, purchasing history, campaign performance, and the ever-elusive "wish list," big data technology can assist wholesalers in understanding what consumers want even before they make a purchase. Advanced modeling can also determine what days are the best for seasonal promotions, colors, and price points which will entice consumers.

Demand forecasting in wholesale distribution is particularly complicated, because it is not the final consumers who make purchases: Wholesalers sell to grocery stores, gas stations, car washes, drug stores, department stores, and any other business needing mass amounts of products. Wholesalers help stock their buyers' shopping shelves by offering a variety of product categories, such as grocery, discount variety, and seasonal merchandise. It is the job of the wholesaler to decide which products to stock and in what amounts, as well as which product promotions the buyer should take and when. Carefully analyzing big data collected across multiple retailers can help wholesalers make these decisions, taking into account holidays, consumer sentiment, competitive models, and weather variations, as well. Big data allows wholesalers to analyze their inventories and warehouse space in order to pinpoint and eliminate costly inefficiencies. Since technology has made tracking inventory through the supply chain easier than ever, a wholesaler can constantly analyze big data over time to help determine the necessary minimum product quantity in each stage of the supply chain. It can be helpful to include reorder points for different products in your big data program.

- *Cloud Computing Basics*

Cloud computing relies on a comparable abstract model for its operations, designed around the concepts of virtualization and distributed computing. With cloud computing, a customer can remotely access IT capabilities, such as software, storage, processes and applications, via the Internet from a public cloud provider or a personal corporate intranet private cloud provider. Cloud computing services are provisioned and billed on a self-service, pay-per-use, on-demand basis to the consumers. The cloud service providers allocate and pool massive data storage and processing resources to serve multiple consumers efficiently and at a low cost. Large numbers of conventional applications handle customer data and processes on a stand-alone basis one at a time. Cloud computing allows the implementation of functions that aggregate and analyze customer application data in parallel in order to identify patterns and determine outcomes; these can be applied broadly to the base of customer applications. In effect, cloud computing democratizes computing as a service, making it available to all sizes of enterprises and people. IT services usage expands from a capex-heavy model organized as a service and risk management center into a usage-based charge back, service-oriented IT center.

Cloud computing is mostly about virtualization. This allows a cloud provider to pool many servers with diverse capabilities behind a single administratively controllable interface, which abstracts the virtual characteristics of the servers behind it. Each of the many users of the storage and process capabilities defined by the cloud interfaces operates in isolation from all of the others operating through the same interface. Security and privacy controls allow the carriers of a cloud to promise the different users of the storage and processing services provided by the cloud that their data is protected and that their processes operate properly.

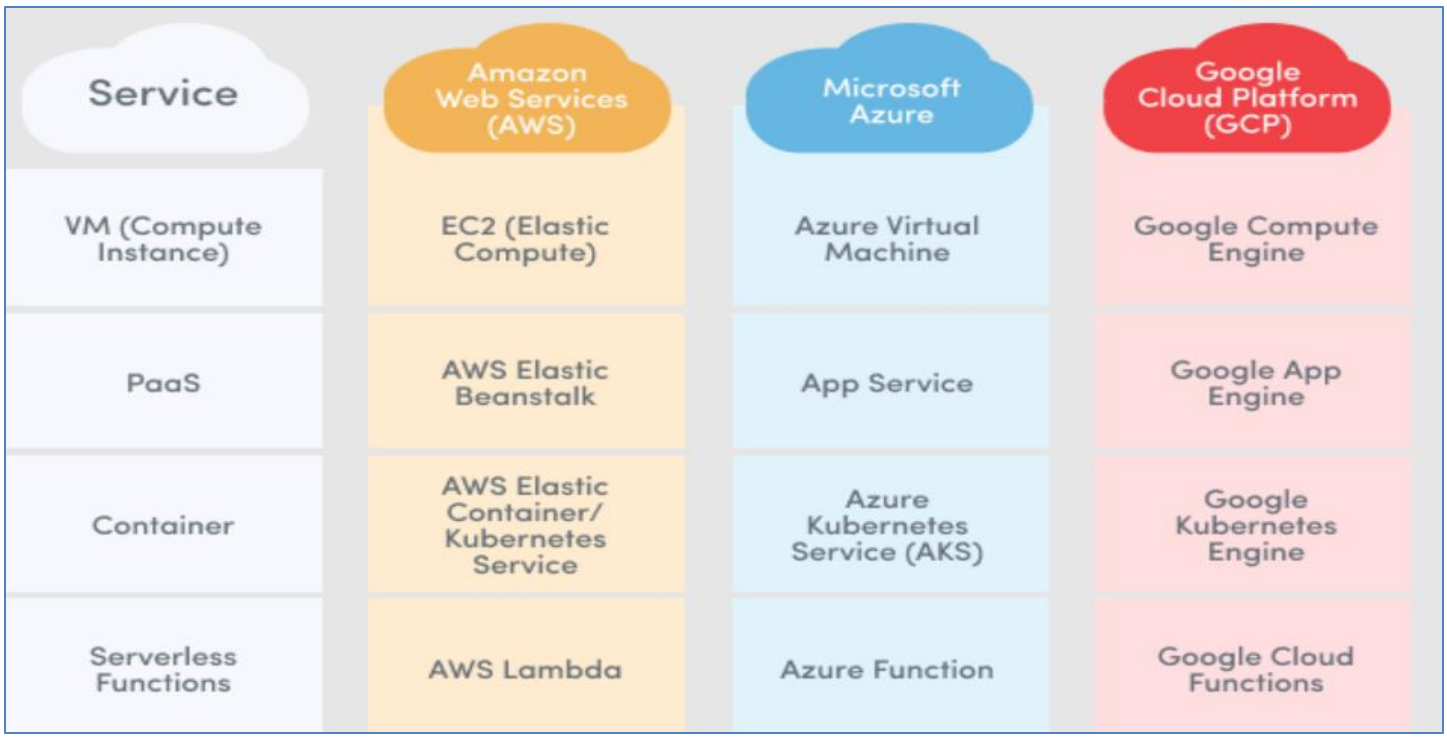

Fig 2 Comparing the Big 3 Cloud Platforms

- *Understanding Cloud Computing Fundamentals*

Cloud computing is a computing service where a central server provides various services via the Internet for on-demand use by the clients who use individual devices to access it. The underlying concept is akin to the public utility model, where the public utility company has a central supply system, and consumers tap into it to meet their individual needs. These days, many individuals and enterprises in diverse fields obtain computing resources from cloud computing service providers instead of maintaining private data centers and servers. This eliminates the need for local setup, and users may utilize the service for free with restrictions or pay according to usage pattern only for consumption above a free limit.

The cloud service provider has a centralized setup with considerable high-end resources and uses ubiquitous Internet connectivity to deliver and monitor the services. As on-demand service, dashboard tools are made available to the consumer to monitor and plan the service utilization. Features such as automatic provisioning and de-provisioning, scalability, availability, reliability, and guaranteed Quality of Service have customized cloud computing to suit various needs. At present, there are SaaS, PaaS, and IaaS services available in the cloud with packages available for various application areas such as storage, analytics, machine learning, artificial intelligence, Internet of Things, big data, and many more. Users may connect to the cloud using various functions-enabled devices and the services they use may be powered by heterogeneous deployment environments across the globe.

II. AMAZON WEB SERVICES (AWS)

➢ *Overview of AWS Services*

AWS is a collection of on-demand cloud computing services launched by Amazon in 2006. Since its inception, AWS has become the market leader in cloud services, providing a wide range of infrastructure software that enables companies to rent servers, storage, databases, networking, mobile development, tools and other services on a pay-as-you-go basis. Creative leaders in organizations of all sizes use AWS services to enhance agility, accelerate their businesses, and minimize costs. More than a million customers, including the fastest-growing startups, largest enterprises, and leading government agencies, are using AWS. AWS is creating the world's most secure, extensive, and cost-effective cloud computing environment. Customers trust the AWS cloud and a growing number of independent software vendors are developing or migrating their applications to the AWS cloud.

The AWS cloud gives users limitless capacity and allows for the rapid deployment of applications and services. In many cases, businesses need to scale quickly and using AWS services allows for rapid scaling; users can add resources in minutes and can purchase services through a credit card. Data can be accessed through a web service interface and the user pays for what is used on a monthly basis. Charging is done according to fixed, low prices for a wide range of cloud-computing components. Complementary services and offerings are enabling virtually all sales, marketing, product, finance, and business support functions to be hosted online via the AWS cloud service environment.

- *Scalability Features*

One of the key features of AWS that makes it a leader in cloud service is its scalability. Scalability can enable companies to improve their profit margins as business growth recovers from the global recession: They can both avoid heavy investments upfront and quickly scale the infrastructure, using AWS cloud services, in a matter of seconds or minutes when the sales increase. This rapid response capability can be a competitive advantage, particularly if the infrastructure can be readily scaled back during downturns to avoid excess capacity expense.

The AWS cloud is the most reliable, scalable, and inexpensive service available. The conclusion is based on the experience of growing more than 200 million members—delivered through the highest-availability infrastructure of its kind in the world, to customers throughout the global marketplace. The innovation and expertise that created Amazon's world-class, e-commerce infrastructure are now focused on providing developers and entrepreneurs with the tools needed to build and optimize the next generation of e-commerce services. Businesses have computing needs that vary widely. For the largest companies, especially those that conduct business across national borders, it is important to be able to rely on a scale of infrastructure that is not just global horizontally, but vertically as well.

- *Overview of AWS Services*

Amazon Web Services was one of the first major providers of on-demand hosting solutions. It began by offering cloud capacity for hosting websites and applications, but quickly expanded into storage, data processing, and database services. Over the years it has built an extensive portfolio of more than 200 services, including high performance computing tools. It is both the largest and best-known cloud provider today.

AWS offers a robust, highly reliable, and scalable environment for hosting applications and processing data. It has service locations in more global cities than either Azure or GCP. The AWS platform can be used by both small and large companies to create and host applications without having to procure hardware. Users can develop and run applications using a variety of frameworks on demand, and only pay for the resources consumed. AWS affords its users a high level of support for security, especially businesses that operate in highly regulated sectors, such as financial services and healthcare. Customers can encrypt their communications to and from AWS, as well as their stored data, using one-time or long-term keys. Databases hosted on AWS can be configured to prevent anyone from accessing them in the event of compromise. AWS has compliance solutions for healthcare, travel, payment systems, and media. It is also equipped with tools for securing machine learning operations. Security is easier with AWS, since excessive network security configuration issues have put many organizations at risk.

- *Scalability Features*

Amazon Web Services (AWS) encompasses a plethora of services equipped with inherent scalability and reliability features necessary for the execution of big data solutions. AWS services are designed to leverage available resources or budgets to comprise almost a limitless cloud platform capable of scaling to meet workload requirements at any level. Key components such as Amazon Elastic Cloud Computing, Amazon Simple Storage Service, Amazon Elastic MapReduce, and Auto Scaling make this possible. AWS core computations typically revolve around EC2 instances; however, EC2 alone would be impractical or unfathomable for a production-level big data solution without the aid of associated scalable web service products. The scalable web support services allow companies to focus on business functionality without having to expend resources developing, deploying, and managing the core functions of hosting a big data solution.

Elastic scaling is an important feature of most AWS services. EC2 Auto Scaling for compute powers is used to incrementally increase or decrease the number of EC2 instances being used in response to actual increases or decreases in workload levels. S3 storage scaling is performed manually and services such as EMR will automatically create and terminate the cluster nodes based on the configuration settings specified by the user for ETL processing routines. The elasticity feature of AWS continues downstream into other services involved in the processing of workloads. AWS Redshift, a managed data warehouse product, will allow users up to 12 data nodes that can range from small to large configurations. In addition, a major benefit to their cloud architecture is the ability to spin up small, micro, or spare capacity EC2 instances at a significantly lower cost than standard instances. While it would be unrealistic to utilize spot instances for year-round operations, they can be extremely cost-effective and result in substantial savings during data load windows that could last for days or weeks at a time.

- *Cost Analysis*

The cost of using any cloud technology is one of the most crucial factors to evaluate before starting a project. In this section, we are discussing the various services offered that we have used in our project and the cost associated with them.

All the services offered have the option of usage for Hourly basis like – instances, EMR, and storage. In our project, we have used instances for web scrapping as well as running the analysis. We have used EMR for creating the environment, running the data extraction, text processing and transformation, and data loading into storage. As the data was not continuous it was used for daily basis in the respective intervals, the resource was shut down and started during our process.

The services like instances, EMR usage are billed and have pricing based on the time and type of instances selected. The storage has different types of classes based on the time, how long the data will be stored for and charged accordingly based on the size consumed. It is a non-expensive solution for data storage if you are storing any medium to high volume data for a longer duration. The data access from storage on an ad hoc basis for small files is an expensive solution. Products have their own pricing.

It should be noted that the charges will vary based on the region and the data egress and ingress volume consumed. To reduce any incurred costs, it is good to check the instances of services offered in each region. As mentioned earlier, all the activities in the cloud are not permanent, resources can be scaled down or up based on the need. Services can be used for such visualization needs as it is charged for the user-based and can be designed on-demand. The table shows the incurred costs associated while running the different components as part of the project.

- *Use Cases in Wholesale Distribution*

Many companies in the wholesale product distribution sector have built enterprise-scale infrastructure deployments. The enterprise solutions help their customers become more efficient in managing their time and labor costs, with complex business workflows that can be customized to customer needs. One supplier and manufacturer of high-performance printing systems has been managing the distribution of physical products and support services for its customers in Latin America through a distributor network. In 2006, the company started the project to move its Latin America logistics to a single online platform. The company selected a cloud infrastructure provider for its distribution system, which offers an array of managed IT services combined with a rapid deployment model and on-demand pricing. The system now serves around 400 distributor accounts associated with more than 4,500 stores. The main tasks performed include fulfillment of customer orders and reporting of transaction data for control and credit payment.

Another service is provided by a distribution service company in Hong Kong that focuses on e-commerce fulfillment. In September 2012, the company adopted a caching solution to scale an e-commerce fulfillment management system running on cloud-based IT infrastructure in order to meet the growing demand from customers for business in Hong Kong, China, and Japan. After using cloud computing services for a year, the company found that cloud computing met the security and development speed requirements to resume the growth of their e-commerce fulfillment business. Their e-commerce fulfillment management system is easier to scale up or down according to the growing demand.

Equation 2 Cost-Performance Index Across Cloud Platforms:

CPI = Cost-performance index

P_r = Platform runtime performance (e.g., throughput or data processing rate)

C_t = Compute and storage cost

O_h = Operational overhead (e.g., management, setup, support)

$$CPI = \frac{P_r}{C_t + O_h}$$

III. MICROSOFT AZURE

Microsoft Azure started out as a cloud computing platform, initially releasing services from a data center in the United States. It has since grown into a portfolio of solutions that addresses most enterprise IT needs, from application development, hosting and support, to backup, storage, and disaster recovery. Recently, Microsoft Azure focused heavily on big data, analytics and IoT, rolling out many services in those areas.

- *Overview of Azure Services*

A broader look at the services offered by Azure reveals a huge extent of services in Machine Learning, analytics, and Data. In these domains, it provides the Azure Machine Learning Platform, a variety of Azure Analytics Services, and a host of Data services that include Azure SQL Database, Azure Cosmos DB, Azure Cache for Redis, Azure Database for MySQL, and Azure Database for PostgreSQL. The Azure AI is based on the research and innovations from Azure's Cognitive Services, Azure Databricks, and Azure Synapse Analytics.

- *Scalability Features*

Microsoft Azure provides Auto scaling capabilities to all its services. Some of them are available in Constant, Scheduled Scaling and custom scaling based on rules. In the cores and database capacity, it provides user-configured scaling. Virtual Azure Datapool allows the creation of SQL Datapool Resources and massively parallel clusters at scale, with on-demand as-a-service resource provisioning. It also supports real-time scaling of Azure Data Engineering.

- *Cost Analysis*

Unlike other platforms, their Interface does not provide a total-at-a-glance view across different types of services. The best information is found in the Compare Pricing. A spreadsheet is available to calculate estimated usage. I found the estimated pricing quite baffling. For example, an Azure Datapool with 4 small-size Utils is $1033 for light usage and $8240 for heavy usage. In contrast, similar capabilities on another platform with free storage are free for the same periods. However, with a multitude of services on different pricing and many other types of resources, providing pricing of Basic Services could be misleading.

➢ *Use Cases in Wholesale Distribution*

In using the Machine Learning services like Bot, Visual and Prediction, one could create process ordering based on previous experience. Using the Data Analysis or the Video Indexer, one can extract intelligence from video data logs in processing planning and prepare action scenarios in a warehouse that could reduce cycle time.

Azure Services can be ideal for Predictive Maintenance based on IoT data logs. Other examples of POCs for Wholesale Industry are Demand forecasting optimizing inventory and shipment schedules. The latter can utilize Azure Machine Learning Technology to leverage historical weather data and Machine Learning algorithms to understand the pattern of prediction.

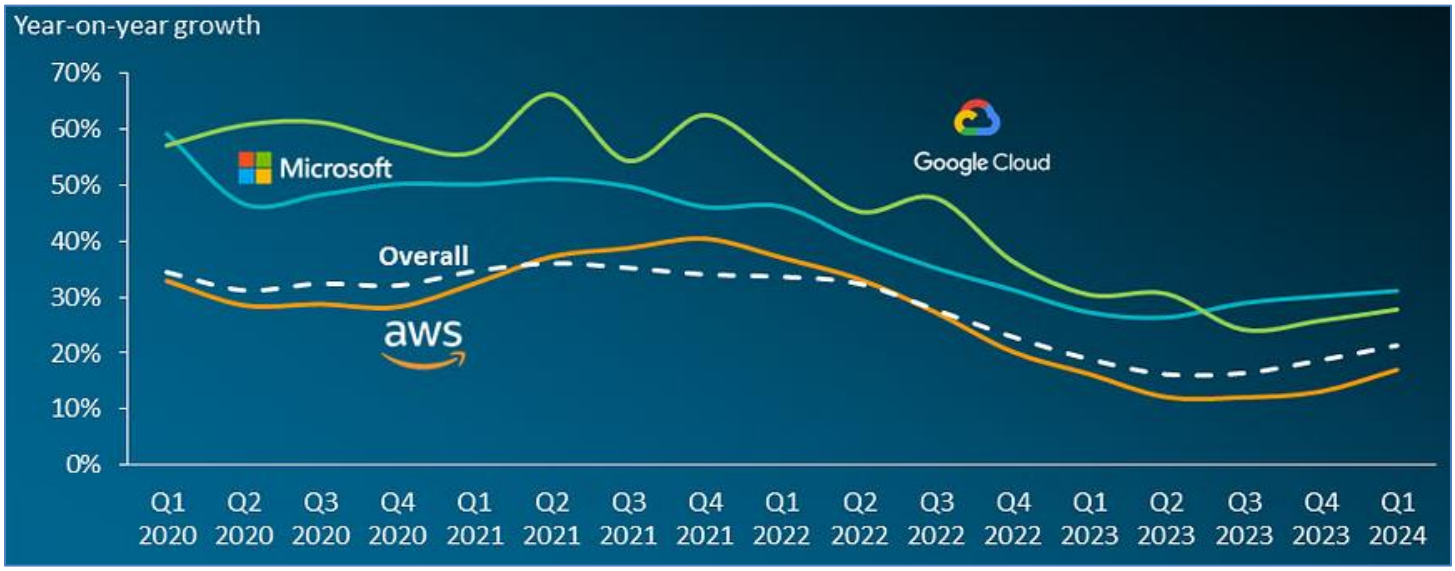

Fig 3 The Cloud War

➢ *Overview of Azure Services*

Microsoft Azure is a comprehensive set of cloud services designed by Microsoft that addresses the diversity and layering of enterprise business needs, offering Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) functionalities to support and scale legacy, on-premise enterprise solutions along with cloud scale development board solutions. Azure cloud services are designed and configured to scale easily with the business needs, protecting customer investment. Azure cloud services are designed to cut costs at the enterprise level in order to reduce the cost of ownership. The migration path to Azure services does not require a radical departure to the cloud for IaaS customers, making it easier to work with existing solutions by linking on-premises infrastructure with the Azure connected infrastructure. Security of sensitive data is popular in a fast-paced world that breaches sensitive customer data, so Microsoft Azure services are designed with numerous, well-published security methods with an eye toward regulatory compliance. Successful warehouse distribution systems house massive amounts of customer data from loyalty programs, retail Point-Of-Sale systems, mobile app sales, web-based E-commerce, shopping cart abandoners, and a wide variety of third-party digital sales partners through the supply chain.

On the 5G horizons, a partnership with Azure began by recommending services for IaaS solutions. Azure Virtual Machines support popular Microsoft server operating systems, but also support other operating systems like Linux and Unix. Azure also offers prepackaged PaaS solutions with Azure Kubernetes Service and Azure Logic App for developers to easily integrate into Azure app service for web-based solutions. Azure DevOps and GitHub Tools are also offered to ease the PaaS functional development of enterprise applications resident in the Azure cloud. Cosmos DB offers serverless and autoscale capability for low-cost database solutions for enterprise mobile and web SaaS solutions. Azure SQL offers popular serverless data solutions to ease financial burdens for enterprises needing to integrate their data into massive insights. Azure Synapse is Microsoft's fully-integrated cloud offering to compete with other Data Warehouse solutions - Microsoft's Synapse offering integrates data engineering, data warehousing, data lakes, and big data analytics. It integrates Azure services and other third-party services as well using servers against Data Lakes and Data Warehouses. Utilizing on-demand serverless architecture, it automatically spins up the servers to speed processing and shuts down to save costs when processing is complete or scheduled.

➢ *Scalability Features*

The scalability of Azure services is evident in multiple ways. First, it has numerous geographical locations worldwide where data centers and their services are at customers' disposal. Azure exclusively has regions in Antarctica and Austria for Azure Government Services and Switzerland, and Privacy Services designed to comply with specific data protection legislation, such as the new Swiss Data Protection Act, the EU General Data Protection

Regulation, or the Liechtenstein Data Protection Act. Other regional considerations involve where Azure is available to provide services to customers, called service availability, and published service-level agreements, which ensure that the availability of core services meets the strict agreements.

The scalability of data analytics and big data services enables more thorough data analysis. Azure Stream Analytics enables the creation of processing units that up- and down-scale to meet stream processing requirements based on event patterns. This service operates near real time and can customize alerts and take actions based on particular situations. Azure Synapse facilitates scalable multi-thread operations and parallel processing of large amounts of data from various sources. Azure automatically scales compute resources to meet the processing requirements. Other features that support scalability include Azure Machine Learning, which scales predictions, and Azure Databricks, which supports a scalable serverless architecture. There are also many different accelerators, including templates, for Azure Cognitive Services, with many different types of readily scalable machine learning and AI solutions. Such solutions include analytics on large amounts of unstructured data, which requires considerable amounts of virtual machine power. Azure Virtual Machines enables the creation of a variety of sizes based on particular business, MO, and mission workloads to scale up appropriately.

- . *Cost Analysis*

Cloud service pricing is complex and can vary greatly across service providers and usage scenarios. Our goal for the cost analysis is to give a high level estimate for 1 week of service for the three candidate architectures in our analytic experiments, utilizing representative datasets of publicly available service prices. The biggest component of cost is Storage – we will provision a Data Lake of 365 days of transaction and master data. The next biggest component of cost is Databricks – we will provision similar Databricks Spark environments and Spark job types as used in our analytic experiments. Azure Synapse and Azure Data Factory have a smaller time utilization but we will provision Azure Synapse Databases for the entire week.

Colocation of various applications enables us to minimize costs of our Data Lake for Resell, Transact, and Report use cases, and Databricks for the Transact Python, Report PySpark, and Cleanse, Dedupe & Conform Delta PySpark, but not for Raw set LTV Analytics. Therefore it is not clear whether there will be a cost advantage of our 4-Pipeline Model compared to the 3-Pipeline Model. While Databricks has a leader and makes work easy, offering support with Delta, considerable price surprises can arise due to its preemptive worker pricing, versus alternatives such as Azure Synapse. Consequently, it is wise to compare both options of Azure Databricks and Synapse through a testbed progress, learning by Doing policy. Similarly, instructions to execute custom Container queries through Azure Synapse can also be submitted to other registrars. Subsequently, the Azure Data Factory and Synapse Scheduler services used to request query execution along with associated checkpointing and result storage can also be explored in detail; however, for our focus, we will utilize them at high level.

- *Use Cases in Wholesale Distribution*

However, in this section, we show two use cases on the Microsoft Azure platform. The first case studied is a Traditional Analytics Data Warehouse Solution with vendor integration. The integrated solution consists of a combination of PaaS Azure services with the vendor data warehouse solution. Moreover, we show a second practical case study as a Modular Big Data Solution using cloud service modularity and elasticity core features. The advantage of these modular solutions is that business units inside wholesale distributor companies can consume needed resources in a Just-in-Time approach.

The first example studied is the Azure Data Warehouse native solution. Enterprise data warehouse solutions are the traditional selection in terms of architecture for big data. These solutions and architectural approaches are data integration hub architectures related to traditional data hub technologies. These technologies are extremely good for traditional business intelligence analysis and enterprise reporting as they are well designed relatively simple analytic operations. In the modern wholesale distribution context, a data warehouse universe is always needed in parallel to advanced analytics solutions that are maintained and updated. Additionally, the new solutions in modern analysis contexts needed by the traditional data warehouse are the online advanced analytics reporting with data in motion and the internal/external data streams collection. These solutions use micro-batch computing, server-less computing, orchestration, and personal data pipeline architecture.

IV. GOOGLE CLOUD PLATFORM (GCP)

GCP was launched in 2011 as a suite of cloud computing services that runs on the same infrastructure that Google uses internally for its end-user products, such as Google Search and YouTube. The GCP portfolio also includes units focused on offering business services such as Google Workspace, which combines cloud-based productivity applications formerly branded under Google Docs and Google Drive. Within the data analytics segment, GCP offers products such as BigQuery, Data Studio, Cloud Bigtable, and Cloud Dataproc, as well as Cloud Dataflow and Cloud Pub/Sub, which offer data streaming and batch processing services. In machine learning and artificial intelligence, GCP is the exclusive cloud partner for TensorFlow, an open source machine learning library created by Google. Google has extended its machine learning platform with AutoML, which essentially allows non-experts to build their own custom machine learning models.

GCP relies heavily on edge computing, which uses geo-distributed servers to handle much of the processing tasks close to where data is being generated. Although the most talked-about GCP products are geared toward data analytics, Google also offers services such as Google Kubernetes Engine (GKE), which allows customers to access the container orchestration and management capabilities of Kubernetes, and Compute Engine, which lets customers manage compute instances that run on Google's infrastructure. In addition to these specialized data analytics services, GCP also competes with other cloud service providers on general purpose cloud services such as Infrastructure as a Service (IaaS) and Platform as a Service (PaaS).

Fig 4 Google Cloud Platform (GCP)

➢ *Overview of GCP Services*

The Google Cloud Platform contains a series of modular cloud modules and services, which help customers to develop and scale new products, and operate workloads in the same environment as Google. These services run on the same infrastructure as Google's end-user products, such as Google Search, Gmail, file storage, and YouTube. Google Cloud provides a suite of services for computing, storage and application development that run on Google's hardware. GCP allows the customers to build, test and deploy applications on the same infrastructure that Google uses internally for its end-user products, such as Google Search and YouTube. GCP has a lot of services on storage, databases, compute, big-data and machine learning. Big Data services in GCP include BigQuery, DataFlow, DataLab, DataStore and CloudML. BigQuery provides high-speed user-friendly SQL-like query to an implementation of the Dremel system, based on columnar storage, cluster management and massively parallel processing. DataFlow, in contrast, provides general programming per-situation user interfaces over the Flume and Sawzall systems. DataLab is a very convenient, standard Jupyter notebook interface that is professionally developed and maintained. It provides flexibility and is easy to use with custom code, and the services are also extensive, providing great power to develop custom data pipelines and machine-learning models. DataStore is a NoSQL implementation designed for big-data applications, and CloudML provides simple services to support TensorFlow. Google built these tools to meet their own needs and used them extensively internally with operational data at extreme scale. They have now productized them and are now available publicly.

➢ *Scalability Features*

Part of the excitement over GCP stems from its speed and scale, as well as its focus on analytics. The unrivaled real-time querying capabilities of its BigQuery service has made it a frequent choice of organizations needing immediate insights on huge data stores. Google Cloud has many low-cost options. Customers first buy the underlying hardware that supports its cloud. Then they pay only for the cloud services they use, at prices often cheaper than its key competitors. GCP's virtual computing service, Compute Engine, was built from the ground up to offer enhanced performance for big data and large-scale workloads, including a revolutionizing container-first technology. The service is touted as the global leader for Big-Data Processing, specifically for Hadoop and Batch Processing workloads. Here, GCE has a 63.25 percent market share. A key Google partner that optimizes the operation of Apache's Hadoop distribution on Google's cloud for customers also claims that GCE is the most optimal cloud for running Spark or Hadoop workloads. The underlying GCE infrastructure may be especially attractive for organizations needing to quickly-scale Big Data-Processing capabilities since GCE recently upgraded its compute and storage infrastructure by becoming the first cloud service to offer Intel's Xeon KNL processors designed for high-performance computing applications.

➢ *Cost Analysis*

When performing a cost analysis, we have to take many points into consideration: type of technology, data volumes, time of execution, technology used, and people hiring, to name a few. So, we developed some comparisons below, trying to cover basic situations. Next, we analyze some costs

involved in big data solutions. In general, some low-level services have a concrete cost for their use. For example, service costing is calculated by the hour, minute, GBs processed, and duration of requests or even requests for data transfer. So, any special prediction regarding service level isn't taken into consideration.

Next, we have predictions for both services, since they don't have a constant behavior during the day – for example, there might be execution peaks at a specific moment of the day, but the overall report might take a longer time to be executed. This report, for example, is run only during the night. The ETL, on the other hand, is run only when there are no business hours. Because of this, we have the need to hire some extra hours of high-level services to take care of the other services' peaks. For this study, we are using different VMs depending on service statistics. The workers are based on a pipeline and its static workers (one each for both startup and shutdown) are the same according to predictions, which generally indicates the last 24-hour period and autocreates capacities for the other periods of day and week. Due to the characteristics of our solution, we have a lower worker time at ETL than at the report, so we believe that a lower number of workers is a good solution for our business.

- *Use Cases in Wholesale Distribution*

Google BigQuery has several use cases in wholesale product distribution. We worked with a 12-month data set from a medium-sized European wholesaler that offered a product assortment of around 100,000 articles which sold to approximately 1,000 customers in the different European countries. The data was stored in a 14-table database. The wholesaler wanted to analyze different marketing strategies focusing on the item assortment relevant for the customers' sales from the wholesaler, as well as optimizing assortments of customer groups to create customer segments which could be addressed with specific marketing messages.

To analyze customer assortment relevance and optimal customer segments for marketing addresses, it was necessary to replicate the database in order to create a working table which should be populated by queries with several JOINs. A working table would then enable queries for the analysis of customer assortment relevance and customer segments with the best item assortments which could be sold and used for marketing addresses of customers.

The first SQL script was responsible for loading sales data of customers for a specific monthly interval into a storage bucket. The steps of this SQL script were repeated for all monthly sales data stored in the database. The final sales data set was then ingested into Google BigQuery. In BigQuery, an appended table was created which was populated for the preceding 12 months with data for the analysis of customer segments and customer assortment significance. With the final sales data set, a query was executed for customer sales data which was then used for analyzing customer segment characteristics and for optimizing the best customer segments for marketing addresses.

V. COMPARATIVE ANALYSIS

This chapter presents a comparative analysis of the discussed cloud services concerning the requirements outlined in the prior chapters. The analysis provides the basis for determining the optimal solution for distributing big data in the case study. While the three considered cloud service models have their advantages and disadvantages, our intent is not to determine a best and a worst option, but rather to select the best solution for the case presented at hand.

- *Performance Metrics*

For performance metrics, the metrics of choice are the response time for increasing loads, service uptime, and network latency. As presented in the previous chapters, GCP appears to be the leader in performance-based metrics, in particular regarding service uptime and network latency.

- *Cost Efficiency*

Cost efficiency is measured in terms of the monthly cost of a solution, the pay-as-you-go option used on a typical case study basis, and the price prediction ability of each service in terms of continuity. Based on the available data, GCP continues to offer the best base amount for the least amount of services used.

- *Ease of Use*

Ease of use is measured by the learning curves for each solution and the availability of templates and how intuitive each service is. Based on reports on ease of use, AWS is the leader in a steep learning curve for less technical persons. Microsoft Azure offers a better option for enterprises already based in the Microsoft ecosystem.

- *Integration Capabilities*

Integration is defined in terms of the base service and third-party possibilities for connectors and tools outside each offered service. It continues to be that AWS leads in this category due to the breadth of possibilities available. Based on user reports, GCP has been adding many more possibilities and edge cases for use.

- *Performance Metrics*

Data warehouses are typically architectured such that data is loaded in bulk on a pre-defined schedule, which is often nightly, and the results from analytical queries are reported, or predictions made. While this bulk ETL (extract, transform and load) process may take a significant amount of time and system resources, resources are typically constrained only to the execution time of the ETL. Query response time is a factor for consideration, especially if users are running analytical queries in an ad-hoc fashion. With cloud computing, the focus needs to change, to consider the total cost of ownership.

Data warehouses are also typically architectured for storing structured data, which is in rows and columns. With cloud computing, it is desirable to bring in unstructured data, which does not fit a pre-defined schema, such as text data from social media, or event data from sensors or known as the Internet of Things. To that end, solutions must support data from these various sources and of various formats, and to host different types of data in different cloud services, and the supported transformations must also be diverse. Creative ways to merge the various types of data, using batch or real-time processing. Finally, data warehouses are also typically architected for a relatively smaller and more fixed size of data, usually from 1–10 Tera Bytes. With cloud computing, these limits need to be increased to the hundreds of TBs, and scaling up to Petabytes as required. Additionally, the ability to auto-scale, such that the cost incurred is related to the time the system was in a state of auto-scaling is a desirable trait.

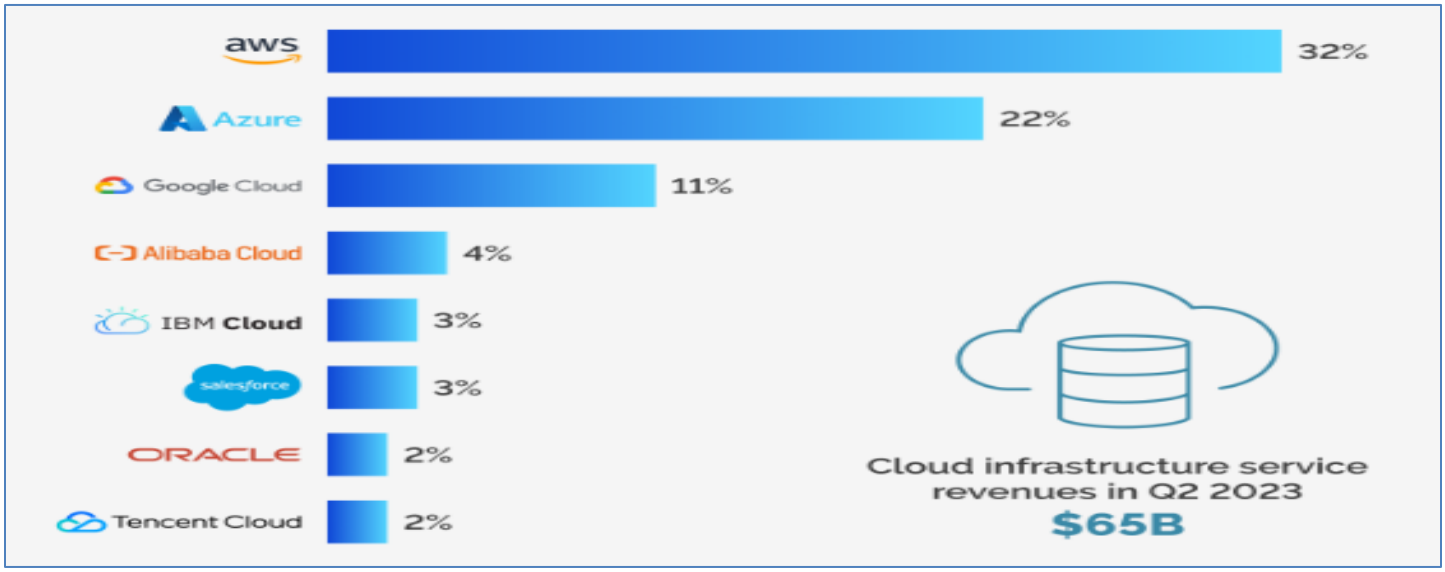

Fig 5 Lead in The Cloud Market

➢ *Cost Efficiency*

For contemporary organisations, cloud computing has become important not only for performance and ease of use but also for cost. Because big data processing software requires many resources, as well as putting more load on the cloud computing infrastructures, creating a big data solution in a cloud environment usually comes with high cost. Such characteristics of big data have led to results where conventional cloud computing economic models no longer seem to apply. In addition to its notable features, the time to completion is also an important aspect for businesses, as any delays take away the cost-genic effects of cloud computing.

In the cloud landscape, there are various services that offer lower prices, though the factors for direct comparison of prices are quite different. Some services provide standard pricing, while others have an interesting pricing mix model that uses an Hour-based Pricing with a minimum of 5 hours, after which users are charged with a second pricing option, the per minute pricing with cluster versioning. The hourly minimum is especially useful for heavy processing tasks, while the per minute pricing is meant for small batch jobs that do not incur high costs. Also of note is how for some services, the image differences imply small configuration costs per image.

Cloud costs comprise of many parameters, which make up this parameter Internal Cost. The first of these costs is the data size, or InputSize, which is the Input Dataset Size. Considered to be a put cost, Upload Cost is incurred on the input dataset. Data Throughput is the Data Transfer Rate from Cloud to Local or Local to Cloud. A second type of cost that makes up Internal Cost is performed on the cloud services, a use cost called Task Cost, made up of the Task Running cost and the Task Number cost. Partitioning and large number of tasks also add Task Number Usage Cost.

➢ *Ease of Use*

Among the criteria to evaluate the three defined cloud providers are their skill to implement scalable Big Data solutions in their platform. After the study of the construction of warehouses in the specific platforms, it is observable that Big Data solutions in AWS are much easier to implement than in the other two. Azure is the second cloud provider that is easier to implement a scalable Big Data solution. For Google Cloud Platform, it was not possible to create a Data Warehouse. Therefore, it is taken out of the consideration to implement a scalable Big Data solution with DWH. As such, after a careful inspection and experiments performed directly on the data collection and preparation steps by using each cloud provider, the modeling, query and dashboard creation were prepared by using Windows 11. In that OS, dashboards upload and do all the transformations that are defined by each SQL code before running it in the cloud.

The platform's implementation streams had some inherent differences in their approach to their tasks. Certainly, there is an observable mean difference of one week in deploying the specific scalable Big Data solutions in each of the three cloud providers. Therefore, for the case of AWS, it was defined the following number of dependency tasks to implement one Big Data solution. Furthermore, the easier to implement cloud provider was AWS. There are several reasons for that, but mainly it is due to the UX and good deployments that can be seen throughout the projects made directly on the platform. This way, it was possible to integrate and connect each service in an easy way. Moreover, it has good and trustworthy services, visible through analysis and reliable publication references.

- *Integration Capabilities*

While the previous sections focused on the IaaS offerings of AWS, Azure, and GCP, we also need to consider the other services that are needed in a typical big data processing workflow. In particular, the distributed computing frameworks, the backend data warehouses, and the integrated development environments. While big data processing on the cloud has its unique challenges, the greatest advantage of a cloud infrastructure lies in its ability to easily integrate new services as needed.

All three providers support all the popular big data frameworks to a large extent. However, AWS contributes more code to Spark and indicates that more third-party Spark connectors are available for AWS than for Azure and GCP. Google's BigQuery service does allow one to use BigQuery in their Spark workflows but involves some work. There is also the possibility of submitting a Hive job to Amazon EMR accessing Hive Tables stored in Google Cloud Storage, though that wouldn't be possible the other way around without some workarounds.

The other major integrated service for Big Data is data warehousing. While Amazon Redshift, Azure Synapse and Google BigQuery are designed to integrate seamlessly with their respective cloud frameworks, there are options available for using them with external environments. AWS Redshift can integrate with any external ecosystem using the new capability of Redshift Spectrum and multiple opening connectors for ETL customers. Azure Synapse has a flexible, pipeline-centered architecture that allows a user to orchestrate and visualize their ETL process into Synapse from an external environment using Data Factory, which is the PaaS service for data integration in Azure. While Google BigQuery is primarily meant to be accessed from GCP's other Big Data solutions, it is also possible to load and extract data from BigQuery using third party ETL tools or using custom-developed pipelines utilizing third party cloud-native integration clusters.

VI. SECURITY CONSIDERATIONS

Security is a major consideration when using commercial clouds to store big-data in a multi-tenancy model. Clouds provide highly available, fault-tolerant services that, with proper implementation, are cost-effective, scalable, and become competitive with on-premises solutions. But is your data and its processing secure when stored in a commercial cloud? For organizations subject to stringent regulations and operational governance in sectors such as finance, insurance, health care, and defense that mandate the protection of users and users' data, a move to the cloud—especially as a basic, inexpensive storage—entails a careful review of the potential security risks. Security is the greatest barrier cited by companies with reluctance to leverage resources from outside the barricade of their organization. In order for companies to consider the cloud an extension of the enterprise and to ensure that applications such as data analytics are adopted, commercial cloud providers have poured resources into making their offerings compelling from a security perspective. A look at the security services offered by the three leading cloud providers illustrates the progress that has been made to erase security-related concerns.

Security suffusion architecture built to solve all of the problems inherent in secure data access in the cloud. It relies on a service that relies on a modified device to create strong, physical authentication for the user accessing the data, even in a multi-user, multi-tenant environment. By tightly binding the security keys to the service, it becomes the weak link in the process. This is a straightforward method to provision, authorize, and manage the granting, revocation, and re-granting of security credentials.

- *Data Protection*

Data protection has become more critical than ever for enterprises. As technology and the number of transistors on a single device increase, spontaneity and error are inevitable. Enterprises must not only work within the realms of what is legal and ethical but also within the trust and integrity of the customers. No one wants to create goodwill with their customers only to be exposed for shady practices of data handling later on. Simple solutions such as encryption, masking, redaction, pseudonymization, and classifying must be used and used often to remain compliant with local government and international laws. Managing and monitoring access privileges and testing security must be a constant chore. Paper-based solutions, such as non-disclosure agreements, insurance, employee training, and physical security, cannot be ignored when dealing with data protection and security. Security of all internal and third party vendors who handle enterprise data must be consistently reviewed and tested. Encryption must be used for data at rest as well as during transmission. Detection, validation, remediation, and recovery also must not be neglected by enterprises. Data loss can happen for multiple reasons, and it is crucial to have a back-up plan so that the

full brunt of the impact is not felt. Care must also be taken on how data is deleted and not just left in the system. Caches can reveal information that is often overlooked. Attempts must also be made to find and mitigate insider threats, real or imagined, which can come from employees as well as higher-level officials. Simply trusting employees and officials is not enough; the proper monitoring of their access and activity is required.

- *Compliance Standards*

Compliance standards refer to the mandatory rules imposed by government regulations on business activities. Security compliance standards specify technical safety measures and properly documented business policies for the safe processing, storage, transfer, and deletion of information containers. Security compliance is mandatory. Companies must bill customers, send documents to regulators, and otherwise share sensitive information with third parties who must trust that the business is sufficiently secure.

Key compliance standards include those that protect cardholder data for card-not-present debit and credit card transactions; those that protect patient health information; and those that protect unclassified information in government computer systems. These government regulations cover the most obvious cases, but there are many compliance standards for a variety of business areas: those for consumer data; those for student data; those for financial data; those for financial institution data; and those for Federal information.

One of today's foremost concerns is the serious threats to personal data privacy posed by hacking and other illegal activities. For a variety of reasons, government regulations have been slow to react, resulting in gaps in data privacy protection. Recent legislation, however, reflects the growing realization that personal data privacy must be safeguarded by compliance standards. These compliance standards are not developed by government agencies but are imposed on companies voluntarily, since failure will result in loss of trust from customers—who will take their business elsewhere—and incalculable fines from government regulators.

- *Risk Management*

Protecting sensitive data is the most important security requirement for applications processing big data. Increased cost of storing data leads organizations to use less expensive cloud storage services to preserve its sensitive data while relying on PaaS or cloud-based SaaS applications to process it. In doing so, organizations run the risk of utilizing an unsecured environment and leaving sensitive data exposed to unauthorized access, disclosure, and theft or corruption. Organizations running big data applications must deploy additional layers of security both while processing data on servers and during subsequent storage in order to meet data protection requirements.

Enterprises have understood the need to secure the pipelines accessing critical sensitive data by implementing key management policies to authorize and restrict who has permission when accessing what data. Caution is imposed on the application developers by providing guides that enforce data security by providing security controls at sensitive data access points, such as application runtime environment, API calls, or during application back-end storage. Other important data protection services adopted by organizations deploying cloud-based SaaS big data applications are encryption services, physical and data access control, identity and certificate management, and monitoring and logging services. All cloud vendors provide a set of data protection, threat intelligence, and detection, monitoring and management services to track, manage, and monitor sensitive data and respond to any unauthorized access or disclosure attempts, both virtual and physical. Cloud administrators are expected to use the optimal combination of the data protection services offered by cloud vendors as hyperscalers have different capabilities when it comes to fulfilling data protection regulatory compliance.

VII. FUTURE TRENDS IN BIG DATA SOLUTIONS

The trends and Innovations in the arena of Big Data Solutions to further enhance its utility and effectiveness towards fulfilling its goal of extracting more and valuable information from the data at hand through the developed Data Pipeline has been discussed in this chapter.

- *Emerging Technologies*

There are several emerging technologies that form the base for Future Big Data Solutions aiming to solve the existing shortcomings of the present products or introduce entirely novel machine learning algorithms that open new paths for new requirements for a growing set of industries. Further these emerging technologies employ innovative means to discover new insights from the data at hand. These technologies can be perceived as the Next Generation technologies and can be generally classified into the following categories. Advanced solutions enabling Large Scale Deep Learning, Transfer Learning and Few Shot Learning for the diverse and large set of unstructured data in novel products. Out of the box Predictive Maintenance Solutions, Advanced Predictive Analytics. Advanced Large Scale Augmented, Virtual Reality Solutions and its Services and New Area of Research Outputs and Products from the NLP, NLU, NLG, Visual Computers in the Large and huge set of unstructured text data available.

- *Market Predictions*

The global Predictive Maintenance Market is expected to grow from USD 3.27 billion in 2019 to USD 10.96 billion by 2024, at a Compound Annual Growth Rate (CAGR) of 27.87% during the forecast period. The global Predictive Analytics Market size is expected to grow from USD 10.95 billion in 2019 to USD 27.69 billion by 2024. The big data infrastructure market is projected to grow from USD 25.83

billion in 2020 to USD 60.36 billion by 2025, at a CAGR of 18.49% during the forecast period.

➢ *Emerging Technologies*

It is expected that the demand for Big Data technologies will grow stably in the coming years. This anticipates Cloud Computing, Machine Learning and Artificial Intelligence, as well as the Internet of Things as the major trends for the development of Big Data solutions until 2025. Cloud Computing allows enterprises to host and leverage their Big Data solutions without having to invest in their on-premise infrastructure. Machine learning and Artificial Intelligence technology will allow employing more efficient algorithms to support Big Data tools. However, the production of high-quality data that are capable of leveraging such algorithms and tools still represents a challenge for multiple industries and companies. Finally, the number of devices connected to the Internet will significantly increase in the coming years, and each of such devices will produce tremendous amounts of data, which will increase the amount of data available worldwide, consequently stimulating the Big Data industry.

Generative AI Technologies are defined as having perhaps the highest visibility and expectations in the AI space – Developments such as certain technologies are certainly creating a tremendous amount of interest. However, is this interest sustainable? – Additionally, Blockchain becomes a company's main security concern, at a time when security is gradually becoming the most important issue for IT segments. Nevertheless, in terms of expectations for IT investments, the Blockchaining Technology is still the lowest, although there is some optimism about the return of investments in this space of low investments and, as a consequence, not very high risks.

Equation 3: Data Throughput Consistency Score (Cloud-Native Pipelines):

$$T_c = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{T_i - \bar{T}}{\bar{T}} \right|$$

where:

- T_c = Data throughput consistency (lower is better)
- T_i = Throughput at time interval i
- $\bar{T}$ = Average throughput over n intervals
- n = Number of time intervals evaluated

➢ *Market Predictions*

A recent report claims big data use is starting its 3rd stage. During the first stage, data was hard to get so only the most strategic use cases were implemented, but currently companies are realizing they can use more data to support more of their operations. We are in the second stage. However, in the near future, data will be treated as a strategic asset shared with other companies to optimize a broad range of business opportunities, not just on the company itself. This will mark the 3rd stage. While Stage 1 and 2 are clearly identified and already taking place, little has been discussed on the Stage 3 predictions. We have both a pragmatic and an imaginative vision on Stage 3, so we decided to share our insights here with you. Pragmatically, companies have begun to recognize that they do not own all the data relevant to their operations, and to grow they need to look outward, sourcing and monetizing more data from other companies. This is the basis for what is called Data Collaboration. We are currently in Stage 2, with a small fraction of organizations adopting strategies and technologies for Data Collaboration. This stage will transition to Stage 3, the Shared Context Stage, which will see the whole data economy transformed. Our bold prediction is that companies will adopt the similar level collaborative data strategies that they have had regarding their finance or supply chain in their business decisions, utilizing an ecosystem of partners providing Market Context Data that enhances visibility and insight.

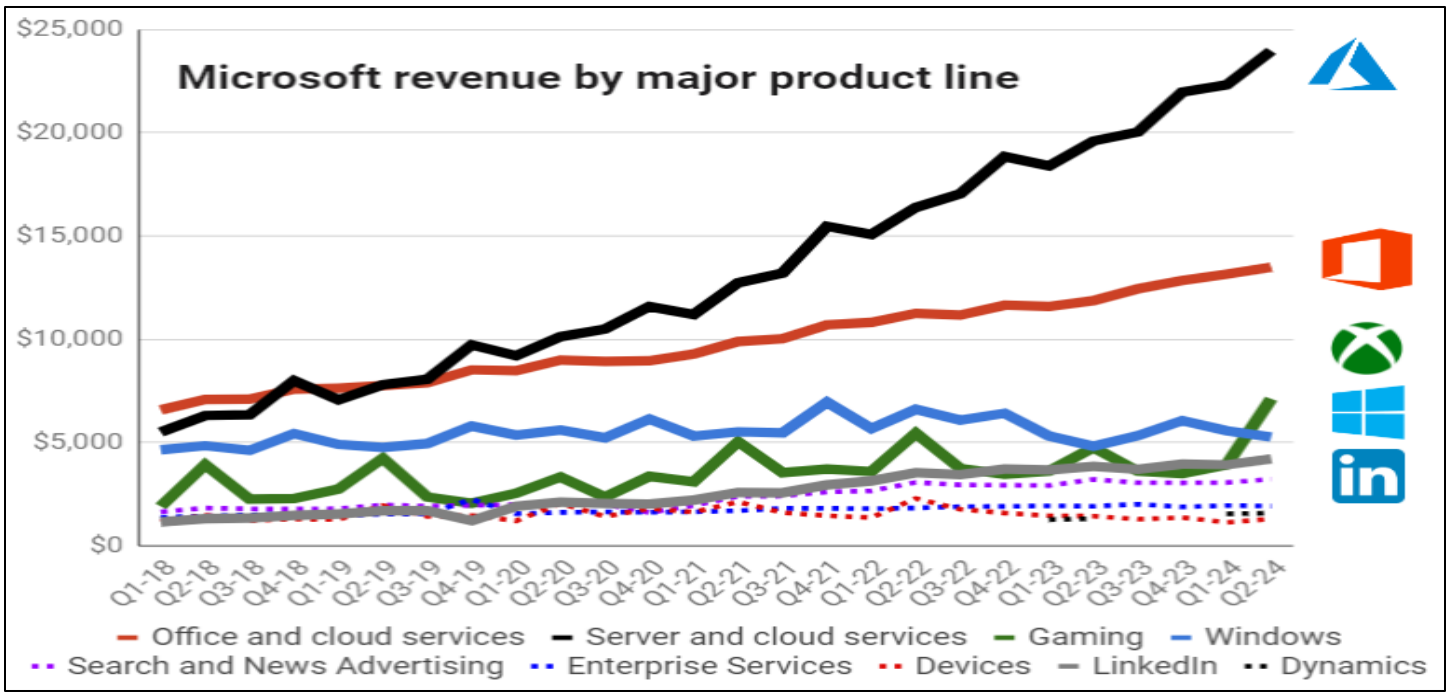

Fig 6 Cloud Market Share

VIII. CONCLUSION

The significant conclusion to be drawn from the findings of this research is that an existing success story of big data in "resource-constrained" industries, such as product distribution, performing close to the demand in a pull-setting may not necessarily be true for other so-called "sophisticated" industries. The reasons are manifold but technological readiness is certainly one of these. The findings of the cloud platform study indicate that current offers are rather limited when it comes to supporting the implementation of scalable and cost-efficient big data pipelines that are key for success in these distinctive industries. If one believes in the limitations are only of a temporary nature as cloud providers will continuously enhance their offerings to domain-specific chip layouts.

This research contributes to the ongoing discourse about the role of technological readiness and current gaps in technology infrastructure provision by cloud service providers, with the aim of stressing the need for further investment in the development of new, domain-specific chip layouts and technological building blocks that are conducive to support success in all clients of the cloud, irrespective of their own technology stature. As of now, the entry barrier to scaling the collection, storage, management, and analysis of big data is still high for many sophisticated industries whose technology focus rests heavily on software development and where physical world interactions are secondary. The findings of the cloud study indicate that the current gaps in technology cloud infrastructure provision reside in the observable limitations of maturity relative to on-premises architecture on which a large share of the criticized legacy systems in sophisticated industries are based.

➢ *Final Reflections and Implications for the Industry*

This work evaluated and compared the leading three public cloud platforms, namely AWS, Azure, and GCP, for use in scalable Big Data solutions. The comparison perspectives were based on a conceptual technology framework that targets the analysis of traditional, in-house, Big Data platforms and tools as well as cloud-based equivalents. The results of our investigation were then summarized in an easily accessible comparative table. The work's primary motivation was to help Big Data solution architects and planners select the most suitable environment on which to host their company's Big Data ecosystem.

The work revealed that public cloud platforms have matured considerably over the last few years, and while none of the examined platforms cover the entire BD ecosystem quadrant, a couple of them come quite close. This situation is significant for Big Data practitioners as it removes a number of usability concerns that plagued their technological option of choice. Nevertheless, it was also observed that the correspondence still does not cover some segments of the BD ecosystem quadrant. This may potentially intimidate or stall organizations that wish to harness the power of Big Data. Clearly, some organizations feel more comfortable – meaning less risk-adverse – purchasing on-premises solutions that provide them with tools for managing their internal Big Data workloads. Indeed, the challenges and obstacles presented by the analysis of the procurement decision journey surrounding Big Data are still in place. Hence, the knowledge gaps described in the literature remain valid today.

REFERENCES

[1]. 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

[2]. Jeevani Singireddy,. (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. Retrieved from https://philstat.org/index.php/MSEA/article/view/2964

[3]. 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. https://doi.org/10.38124/ijsrmt.v1i12.436

[4]. Koppolu, H. K. R. (2022). Advancing Customer Experience Personalization with AI-Driven Data Engineering: Leveraging Deep Learning for Real-Time Customer Interaction. Kurdish Studies. Green Publication. https://doi. org/10.53555/ks. v10i2, 3736.

[5]. Pallav Kumar Kaulwar. (2022). Data-Engineered Intelligence: An AI-Driven Framework for Scalable and Compliant Tax Consulting Ecosystems. Kurdish Studies, 10(2), 774–788. https://doi.org/10.53555/ks.v10i2.3796

[6]. Srinivasarao Paleti. (2022). Adaptive AI In Banking Compliance: Leveraging Agentic AI For Real-Time KYC Verification, Anti-Money Laundering (AML) Detection, And Regulatory Intelligence. Migration Letters, 19(6), 1253–1267.

[7]. Balaji Adusupalli. (2022). Secure Data Engineering Pipelines For Federated Insurance AI: Balancing Privacy, Speed, And Intelligence. Migration Letters, 19(S8), 1969–1986. Retrieved from https://migrationletters.com/index.php/ml/article/view/11850

[8]. 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. Retrieved from https://www.scipublications.com/journal/index.php/gjmcr/article/view/1292

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

[10]. 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. In Kurdish Studies. Green Publication. https://doi.org/10.53555/ks.v10i2.3760

[11]. Someshwar Mashetty. (2022). Enhancing Financial Data Security And Business Resiliency In Housing Finance: Implementing AI-Powered Data Analytics, Deep Learning, And Cloud-Based Neural Networks For Cybersecurity And Risk Management. Migration Letters, 19(6), 1302–1818. Retrieved from https://migrationletters.com/index.php/ml/article/view/11741

[12]. Anil Lokesh Gadi. (2022). Transforming Automotive Sales And Marketing: The Impact Of Data Engineering And Machine Learning On Consumer Behavior. Migration Letters, 19(S8), 2009–2024. Retrieved from https://migrationletters.com/index.php/ml/article/view/11852

[13]. Pamisetty, A. (2022). Enhancing Cloud native Applications WITH Ai AND Ml: A Multicloud Strategy FOR Secure AND Scalable Business Operations. Migration Letters, 19(6), 1268-1284.

[14]. Burugulla, J. K. R. (2022). The Role of Cloud Computing in Revolutionizing Business Banking Services: A Case Study on American Express's Digital Financial Ecosystem. Kurdish Studies. Green Publication. https://doi. org/10.53555/ks. v10i2, 3720.

[15]. Nuka, S. T. (2022). The Role of AI Driven Clinical Research in Medical Device Development: A Data Driven Approach to Regulatory Compliance and Quality Assurance. Global Journal of Medical Case Reports, 2(1), 1275.

[16]. Challa, K. (2022). Generative AI-Powered Solutions for Sustainable Financial Ecosystems: A Neural Network Approach to Driving Social and Environmental Impact. Mathematical Statistician and Engineering.

[17]. Malempati, M. (2022). Machine Learning and Generative Neural Networks in Adaptive Risk Management: Pioneering Secure Financial Frameworks. Kurdish Studies. Green Publication. https://doi. org/10.53555/ks. v10i2, 3718.

[18]. Chakilam, C. (2022). Generative AI-Driven Frameworks for Streamlining Patient Education and Treatment Logistics in Complex Healthcare Ecosystems. Kurdish Studies. Green Publication. Kurdish Studies. Green Publication. https://doi. org/10.53555/ks. v10i2, 3719.

[19]. Komaragiri, V. B. (2022). AI-Driven Maintenance Algorithms For Intelligent Network Systems: Leveraging Neural Networks To Predict And

Optimize Performance In Dynamic Environments. Migration Letters, 19, 1949-1964.
[20]. Chava, K. (2022). Redefining Pharmaceutical Distribution With AI-Infused Neural Networks: Generative AI Applications In Predictive Compliance And Operational Efficiency. Migration Letters, 19(S8), 1905-1917.
[21]. 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
[22]. Kannan, S. (2022). The Role Of AI And Machine Learning In Financial Services: A Neural Networkbased Framework For Predictive Analytics And Customercentric Innovations. Migration Letters, 19(6), 985-1000.
[23]. Annapareddy, V. N. (2022). Innovative Aidriven Strategies For Seamless Integration Of Electric Vehicle Charging With Residential Solar Systems. Migration Letters, 19(6), 1221-1236.
[24]. Siramgari, D. (2022). Enhancing Telecom Customer Experience Through AI Driven Personalization - A Comprehensive Framework. Zenodo. https://doi.org/10.5281/ZENODO.14533387
[25]. Challa, S. R. (2022). Optimizing Retirement Planning Strategies: A Comparative Analysis of Traditional, Roth, and Rollover IRAs in LongTerm Wealth Management. Universal Journal of Finance and Economics, 2(1), 1276.
[26]. Daruvuri, R. (2022). An improved AI framework for automating data analysis. World Journal of Advanced Research and Reviews, 13(1), 863-866.
[27]. Ganesan, P. (2021). Advancing Application Development through Containerization: Enhancing Automation, Scalability, and Consistency. North American Journal of Engineering Research, 2(3).
[28]. 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
[29]. Sriram, H. K. (2022). AI Neural Networks In Credit Risk Assessment: Redefining Consumer Credit Monitoring And Fraud Protection Through Generative AI Techniques. Migration Letters, 19(6), 1017-1032.
[30]. 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. Retrieved from https://www.scipublications.com/journal/index.php/gjmcr/article/view/1294
[31]. Komaragiri, V. B., & Edward, A. (2022). AI-Driven Vulnerability Management and Automated Threat Mitigation. International Journal of Scientific Research and Management (IJSRM), 10(10), 981-998.
[32]. Chakilam, C. (2022). Integrating Generative AI Models And Machine Learning Algorithms For Optimizing Clinical Trial Matching And Accessibility In Precision Medicine. Migration Letters, 19, 1918-1933.
[33]. Malempati, M. (2022). AI Neural Network Architectures For Personalized Payment Systems: Exploring Machine Learning's Role In Real-Time Consumer Insights. Migration Letters, 19(S8), 1934-1948.
[34]. 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. Retrieved from https://www.scipublications.com/journal/index.php/ojms/article/view/1295
[35]. Kishore Challa, Jai Kiran Reddy Burugulla, Lahari Pandiri, Vamsee Pamisetty, Srinivasarao Paleti. (2022). Optimizing Digital Payment Ecosystems: Ai-Enabled Risk Management, Regulatory Compliance, And Innovation In Financial Services. Migration Letters, 19(S5), 1748–1769. Retrieved from https://migrationletters.com/index.php/ml/article/view/11807
[36]. Anil Lokesh Gadi. (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. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/2965
[37]. Botlagunta Preethish Nadan. (2022). Emerging Technologies in Smart Computing, Sustainable Energy, and Next-Generation Mobility: Enhancing Digital Infrastructure, Secure Networks, and Intelligent Manufacturing. Mathematical Statistician and Engineering Applications, 71(4), 16749–16773. Retrieved from https://philstat.org/index.php/MSEA/article/view/2967
[38]. 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. Retrieved from

https://www.scipublications.com/journal/index.php/ujfe/article/view/1297

[39]. Srinivasarao Paleti. (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.

[40]. Pallav Kumar Kaulwar. (2022). Securing The Neural Ledger: Deep Learning Approaches For Fraud Detection And Data Integrity In Tax Advisory Systems. Migration Letters, 19(S8), 1987–2008. Retrieved from https://migrationletters.com/index.php/ml/article/view/11851

[41]. Singireddy, J., Dodda, A., Burugulla, J. K. R., Paleti, S., & Challa, K. (2021). Innovative Financial Technologies: Strengthening Compliance, Secure Transactions, and Intelligent Advisory Systems Through AI-Driven Automation and Scalable Data Architectures. Universal Journal of Finance and Economics, 1(1), 123–143. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1298

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

[43]. Satyaveda Somepalli. (2022). Beyond the Pill: How Customizable SaaS is Transforming Pharma. The Pharmaceutical and Chemical Journal. https://doi.org/10.5281/ZENODO.14785060

[44]. Daruvuri, R. (2022). Harnessing vector databases: A comprehensive analysis of their role across industries. International Journal of Science and Research Archive, 7(2), 703-705.

[45]. Sikha, V. K., Siramgari, D., Ganesan, P., & Somepalli, S. (2021). December 30. Enhancing Energy Efficiency in Cloud Computing Operations Through Artificial Intelligence. Zenodo.

[46]. Somepalli, S. (2021). Dynamic Pricing and its Impact on the Utility Industry: Adoption and Benefits. Zenodo. https://doi.org/10.5281/ZENODO.14933981

[47]. Ganesan, P. (2021). Advanced Cloud Computing for Healthcare: Security Challenges and Solutions in Digital Transformation. International Journal of Science and Research (IJSR), 10(6), 1865-1872.

[48]. Satyaveda Somepalli. (2020). Modernizing Utility Metering Infrastructure: Exploring Cost-Effective Solutions for Enhanced Efficiency. European Journal of Advances in Engineering and Technology. https://doi.org/10.5281/ZENODO.13837482

[49]. Ganesan, P. (2021). Leveraging NLP and AI for Advanced Chatbot Automation in Mobile and Web Applications. European Journal of Advances in Engineering and Technology, 8(3), 80-83.

[50]. Kaulwar, P. K. (2022). The Role of Digital Transformation in Financial Audit and Assurance: Leveraging AI and Blockchain for Enhanced Transparency and Accuracy. Mathematical Statistician and Engineering Applications, 71 (4), 16679–16695.

[51]. Anil Lokesh Gadi. (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. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11557

[52]. Sondinti, L. R. K., & Yasmeen, Z. (2022). Analyzing Behavioral Trends in Credit Card Fraud Patterns: Leveraging Federated Learning and Privacy-Preserving Artificial Intelligence Frameworks.

[53]. Ganti, V. K. A. T., & Valiki, S. (2022). Leveraging Neural Networks for Real-Time Blood Analysis in Critical Care Units. KURDISH. Green Publication. https://doi. org/10.53555/ks. v10i2, 3642.

[54]. Kothapalli Sondinti, L. R., & Syed, S. (2022). The Impact of Instant Credit Card Issuance and Personalized Financial Solutions on Enhancing Customer Experience in the Digital Banking Era. Universal Journal of Finance and Economics, 1(1), 1223. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1223

[55]. Vankayalapati, R. K., & Pandugula, C. (2022). AI-Powered Self-Healing Cloud Infrastructures: A Paradigm For Autonomous Fault Recovery. Migration Letters, 19(6), 1173-1187.

[56]. Kalisetty, S., Vankayalapati, R. K., Reddy, L., Sondinti, K., & Valiki, S. (2022). AI-Native Cloud Platforms: Redefining Scalability and Flexibility in Artificial Intelligence Workflows. Linguistic and Philosophical Investigations, 21(1), 1-15.

[57]. 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. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1296