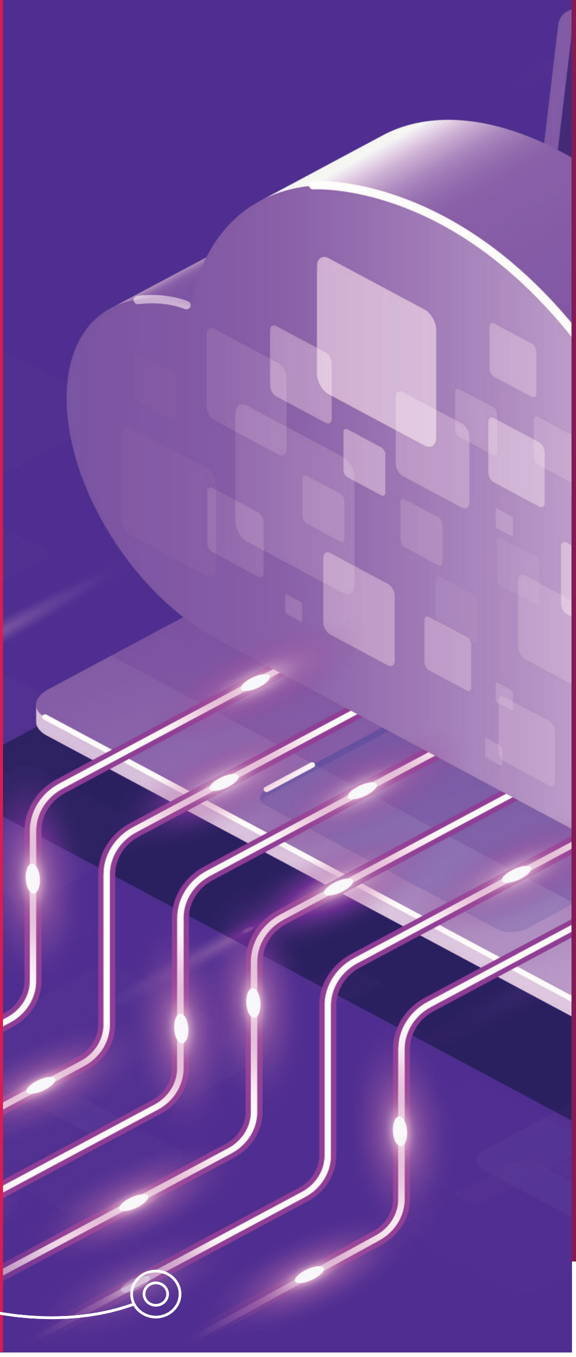
Mindtree
A Larsen & Toubro Group Company

ETL options across cloud

Simplifying the data preparation, data transformation & data exploration for augmented analytics x-clouds



Welcome to possible

A Mindtree White paper| October, 2020

Scope of this white paper

Would you like to revolutionize and uncover the unlimited power of data from various sources and productionalize your AI/ML models for amazing recommendations? Do you want to consolidate multiple formats & multiple data sources to have powerful big data (volume, variety, veracity and velocity) platform that has faster, easier to manage data lineage with repeatable advanced analytics processes, and receive billions of recommendations from PB scale data? Would you like to see the utmost bigdata features, functions, and 'augmented next-gen analytics' best practices for achieving data-driven rich, deeper insights in a very near-real-time (streaming) or batch model?

> ***"The building blocks for achieving that goal is to set up a flexible, insights & low-latency search infused Enterprise DataLake (DWH) or 'augmented analytics' platform that should include a data driven ETL & ELT- batch and streaming unified platform; with accelerated practices of data preparation, data enrichment, data transformation and data governance & exploration solutions".***

Many organizations that are trying to become data-driven or insights oriented organizations in the near future have started setting up the environment and culture needed for building and using the power of advanced analytics for their business to make swift recommendations and business decisions. Augmented analytics platform enhances the quality and availability of the services for growing the business footprints.

To be a harbinger and stay ahead in the current competitive world, there are massive requirements to have the capability for getting deeper insights, customer 360 degree preferences & recommendations, and integration of business systems across multi-channels (social media/,etc.) for seamless user onboarding/marketing.

On the analytics side, there are multiple products available for performing batch and streaming (for example- CloverETL, Domo, Talend, Pentaho, Informatica, IBM DataStage, etc.). But driving insights appropriately and quickly creating data pipelines of these products requires immense tool understanding, waiting for licenses and the need for a group of skilled ETL developers & DBA. If you are planning to integrate your organizations' security policy algorithm/code, adjusting data format & sources, data encryption/decryption & data governance process to these kinds of tools; it may require a lot of code and dependency on vendor partner, which will be time-consuming and not a cheap solution!

Using a cloud vendor (Azure-ADF, Google-Dataflow and AWS-Glue) ETL/ELT/streaming framework means establishing an easier-to-use, robust, and more scalable next-gen platform that overcomes the problems related to the lack of polyglot persistence, not having a single source of truth and waiting for months for an ETL tool procurement, months long search for a skilled tool SME, and lack of agility. With the help of these products, we can streamline the overall process and focus more on core business logic and values rather than consuming time for setup & maintenance of the tool.

Also, the unified framework with low code/no code approach of these Cloud ETL products yields to a unique way of data collection/loading at target as per defined velocity from a variety of sources, the fastest way of exploring & transforming in-flight data with data security, run-time code & config management with scale that will never be the problem anymore. Organizations can leverage the existing/Legacy ETL framework to migrate quickly and setup the next-gen analytics foundations with any of these tools.
To dive deep, the next-gen ETL frameworks reflect the predictability that will be perceived in the future as unified, innovative and standardized frameworks; leveraging the power of cloud, auto-scalability, serverless, chaos engineering and AI/ML ops IAM & data governance; for providing key business decisions in no-time.

Deep dive comparison study at a single pane of glass for Azure- ADF, GCP-Dataflow and AWS-Glue :-

	Azure Data Factory	Amazon Glue	Dataflow
Cloud Provider	Microsoft Azure	aws	Google Cloud
	Azure- Microsoft	Amazon Web Services (AWS)	Google Cloud Flow
Introduction	**ADF (Azure Data Factory)** is the Microsoft-Azure cloud data processing (data pipelines) & ETL product. ADF is used to create data-driven framework and pipelines for managing, orchestrating & tapping the data for analysis and recommendations for big data's 4 Vs (variety, volume, veracity and velocity). For designing **code free** data pipelines, ADF is a great choice for Extract, Transform Load (ETL) and Extract, Load and Transform (ELT). ADF uses more than 100+ exclusively built and no-ops connectors at no added cost. The user can plan & focus on data logic and framework—the powerful serverless Azure will do the rest (heavy lifting).	**AWS Glue** is a complete ETL (extract, transform, and load) solution of AWS that focuses on cost-effective services- catalog data, prep/clean data, transform/summarize and load/move into the target systems. **AWS Glue** has a central repository for having metadata. This metadata is known as the data catalog of AWS Glue and it is a core service or engine of ETL. At run time, this engine generates pipeline code (Scala/Python) framework, policy, batch scheduler layout and job monitoring. This makes it a great choice for data enrichment and movement for Analytics/Sagemaker (ML).	**Google Cloud Dataflow** is a great tool for **unified** (batch and streaming) platform. This has ETL, batch processing, streaming, real-time analytics for big data/ML use cases. The focus of Dataflow is to eliminate the latency & performance issues of MapReduce/Hadoop, while building complex Dataflow pipelines. **Google Cloud Dataflow** uses Apache Beam (opensource) for defining logic/framework of pipelines, whereas worker node/pipeline processing is managed perfectly within the Google Cloud Platform. **Apache beam, makes** Dataflow more robust for unified model for stream and batch data processing with Auto-scaling, no-ops (Serverless), interoperability and highly cost effective reusable solution at GCP.
Serverless	Yes	Yes	Yes

Features	1. No development skill (zero code) or no management required to implement ETL and ELT pipelines. 2. Auto scalable, serverless, cost-efficient and fully managed solution for analytics and ML use cases. 3. For integrating **on-premises data sources, cloud-based, and SaaS** (software-as-a-service) apps, leveraging the power of Azure platform and security. 4. SSIS integration capability. 5. Data migration with cleaning & enrichment at low cost and at low efforts.	1. We can use AWS Glue when we run serverless queries against our Amazon S3 data lake. 2. It's easier to have data-driven and event-driven ETL pipelines. Moreover, it can be used to understand the data lineage and catalog. 3. Integrated data catalog. 4. Automatic data discovery. 5. Automated code/script generation in Scala/python to run at Apache Spark. 6. Clean and de-duplicate data. 7. Developer endpoints to use IDE to edit code/script 8. Flexible job scheduler. 9. Serverless streaming ETL.	1. Automated resource management and dynamic work rebalancing. 2. Horizontal autoscaling. 3. Elastic resource reservation/allocation for batch processing. 4. Streaming engine. 5. Dataflow SQL. 6. Dataflow templates. 7. Notebooks integration. 8. Inline monitoring. 9. Customer managed encryption keys. 10. Dataflow VPC service control- additional security. 11. Private Ips.
Use cases (where to use)	1. Code free (no-code) ETL as a service. 2. Best fit for augmented & advanced analytics, complex pipelines uses cases because it has seamless integration with big data services such as Azure HD Insight Hadoop, Azure Databricks, and Azure SQL Database, ADLS gen2. 3. For creating DataMart/data stores for ML/AI models from raw data.	1. Discovers and catalogs metadata. 2. Has reusable modules or scripts for data enrichment, summary, aggregation, or transformation. 3. Change data capture or highlighting the schema change use cases. 5. Summarizing the runtime metrics of monitoring data warehouse or data lake (S3).	1. Stream analytics. 2. Real-time AI. 3. Sensor and log data processing. 4. Data science model predictions comparison via Dataflow pipelines. 5. Data security at transit- creating pipelines for converting plain text into encrypted data using GCP KMS before loading at target.

Benefits	1. ADF pipelines are executed within the Azure platform to leverage autoscaling and serverless capabilities of cloud. 2. Parallel workflow execution and pay per use utilization. 3. ADF has prebuilt connectors for data transfer with no additional cost to the users/business. 4. Fully managed and CICD.	1. Less hassle, integrated with a wide range of AWS services. 2. Cost effective and fully managed. 3. More automation power- with the automation of build, manage and run job.	1. Fully managed data processing service. 2. Automation infused provisioning, governance, and management of worker nodes of Dataflow. 3. Horizontal autoscaling of worker resources to maximize resource utilization. 4. OSS community-driven innovation with Apache Beam SDK. 5. Reliable and consistent exactly once processing.
Focus	ETL (Extract, Transform, Load)	ETL (Extract, Transform, Load)	Stream and batch processing
Connects to data warehouse?	Yes	Yes	Yes
Connects to Datalake?	Yes (BLOB/ADLS 2)	Yes (S3)	Yes (GCS)
Developer tools	Rest API, .Net/Python SDKs	AWS Glue is dependent on development endpoint for having reusable scripts. These runtime scripts can be created, modified, replaced, or deleted in development endpoints using the AWS Glue console or via API.	Cloud Dataflow API, SDK for Java and Python, Apache Beam
Compliance, Governance, and Security Certifications	ADF data compliance is certified by major security standards like HIPAA, CCPA, etc.	SOC, PCI, FedRAMP, HIPAA	HIPPA

Pricing	Pay only for what user uses, with no initial commitment. ADF engagement starts with cheaper- $1 for per thousand activities runs per month plans. The pricing for data pipeline is calculated based on: 1) Number of pipeline execution and how the orchestration is built on top of it. 2) How is data flowing and data enrichment time is taken with error tracking/debugging? 3) How many data factory operations are involved in ETL?	AWS Glue is charging the user Less than half $ per data processing unit for an ETL job of type Apache Spark or type Python shell, whereas other platform services will add up.	Google Cloud Dataflow jobs are charged back to the user based on each second's usage. This is calculated based on the usage of Dataflow batch or streaming workers/nodes.
SLA	99.9 %	99.90%	99.95%
Integrations /Vendor	100+ connectors of Azure, AWS-S3, NO-SQL databases, protocols, services, and apps (3rd party apps).	JDBC- Connectivity apps/services	Google Products (Big Query, GCS, Composer, etc.) Also has integration to other ETL tools. Talend, SnowPlow and Confluent.
Developer tools	Rest API, .Net/Python SDKs	AWS Glue is dependent on development endpoint for having reusable scripts. These runtime scripts can be created, modified, replaced, or deleted in development endpoints using the AWS Glue console or via API.	Cloud Dataflow API, SDK for Java and Python, Apache Beam

How ETL pipelines architectures look like x-clouds....

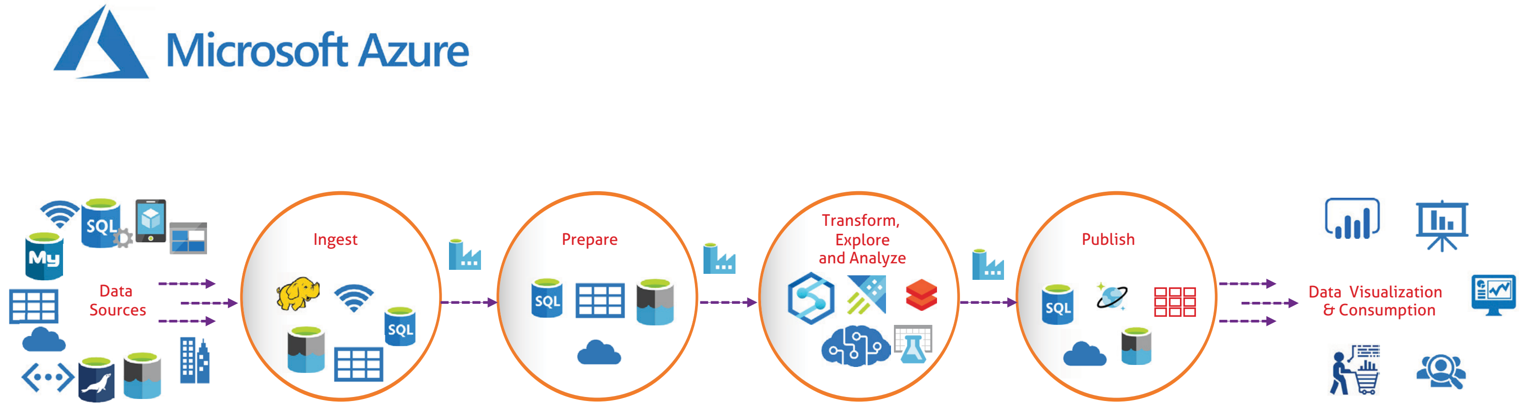

Microsoft Azure
Data Sources
Ingest
Prepare
Transform, Explore and Analyze
Publish
Data Visualization & Consumption

Fig-1: Azure-ADF data pipeline (courtesy- MS)

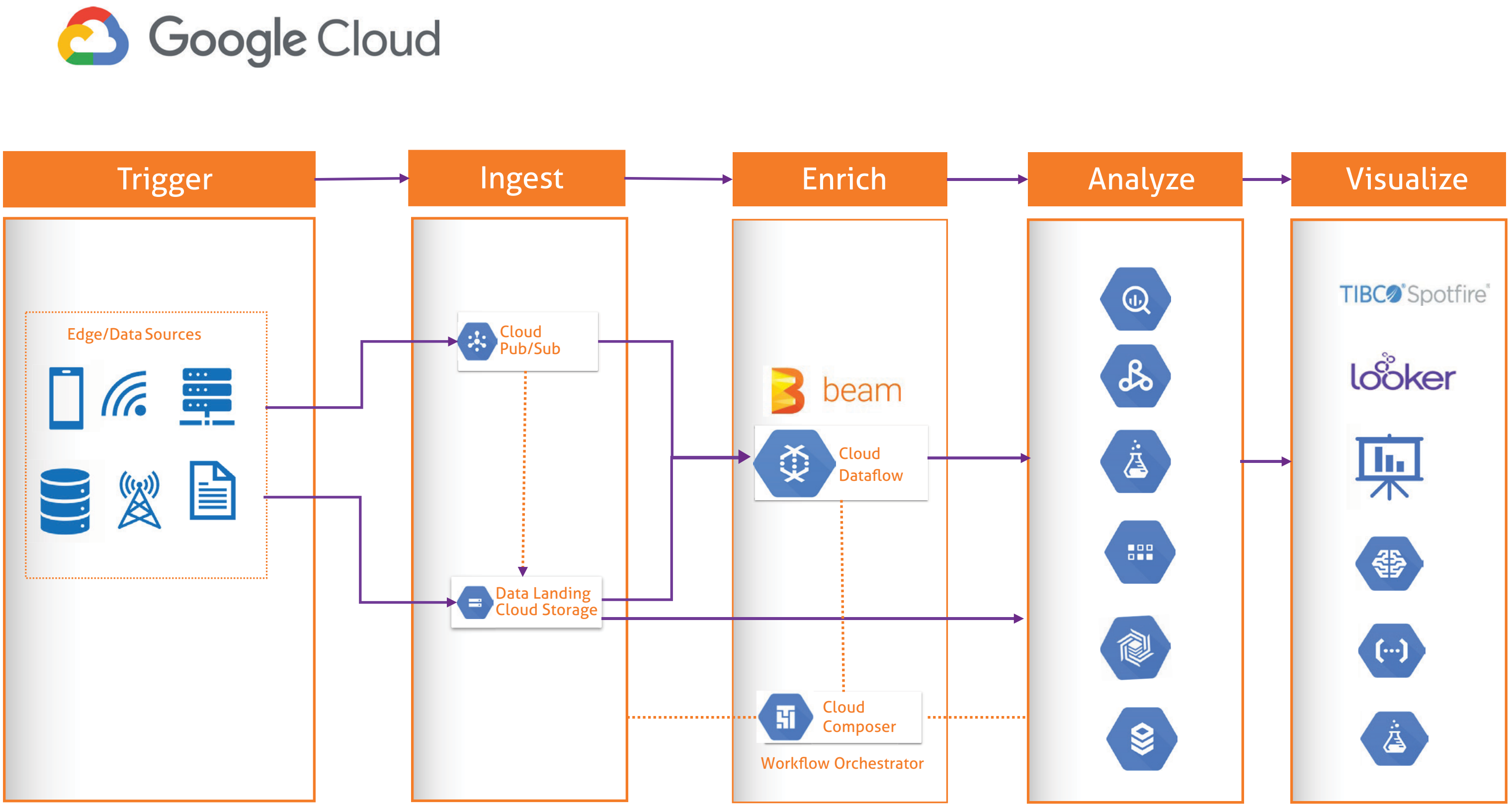


Fig-2: Google Cloud Platform-Dataflow/Apache Beam Data pipeline (courtesy-Google)

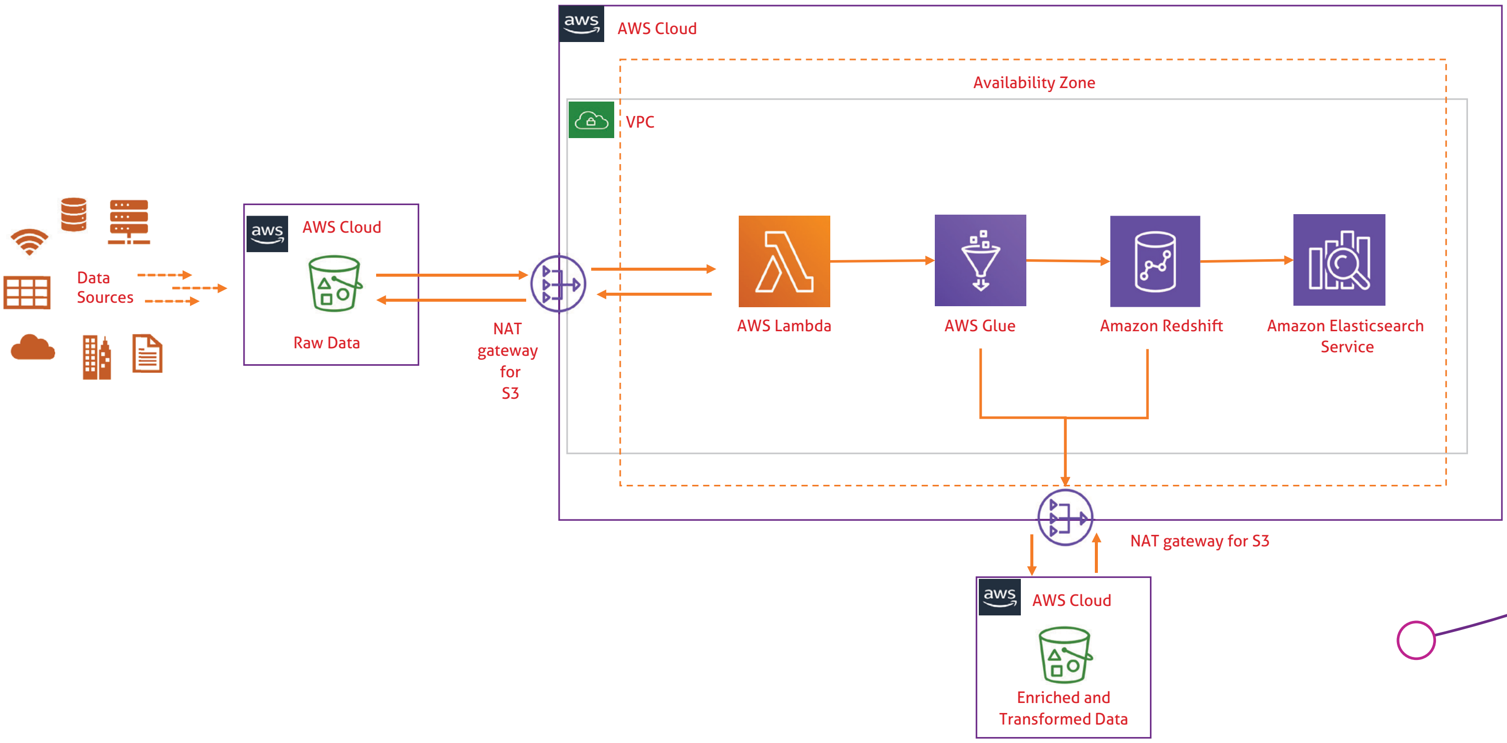


Fig-3: AWS-Glue Data pipeline (courtesy-AWS)

Conclusion:

Finally, the use cases, and organization decisions for public multi-cloud and business needs are the driving factors to choose the ETL tool. There are certain trade-offs for implementing each cloud ETL solution, whereas earlier limitations of connectivity with data sources, compute, storage, and integrations with DBs, data security, RBAC and data governance tools are eliminated by each cloud. At cloud, these ETL tools are easily scalable for streaming data pipelines from the traditional batch paradigm. This reduces our advanced analytics community's complexity & effort to drive insights, real-time, and intelligent-edge solutions. Also, very little learning curve is required for data analysts/data scientists to design and build an ETL framework for any cloud with no/low time.

References & documentations:

1. https://cloud.google.com/dataflow
2. https://azure.microsoft.com/en-us/services/data-factory/
3. https://aws.amazon.com/glue

Abbreviations:

ETL	Extract, Transform and Load
ELT	Extract, Load and then Transform
DBA	Database Administrator
ADF	Azure Data Factory (Microsoft)
AWS	Amazon Web Services
DWH	Data Ware House
SME	Subject Matter Expert
GCP	Google Cloud Platform
BLOB	Binary Large Object (Storage)
ADLS	Azure Datalake Storage
IDE	Integrated Development Environment (such as Eclipse)
CICD	Continuous integration and Continuous Delivery
HIPPA	Health Insurance Portability and Accountability Act (Compliance)
CCPA	California Consumer Privacy Act (Compliance)

Nagendra Sharma,
Technical Director-
Multi-Cloud & Data Science

Nagendra is part of the Mindtree consulting group, where he works as a chief architect for multi-cloud and data science practice. His expertise lies in building highly scalable & efficient architectures, and he is passionate about evangelize strategies for multi-cloud, big data, mar-tech, AR/VR, advanced data insights & analytics, augmentedAI, ML, deep learning & IoT for our global customers.

About Mindtree

Mindtree [NSE: MINDTREE] is a global technology consulting and services company, helping enterprises marry scale with agility to achieve competitive advantage. "Born digital," in 1999 and now a Larsen & Toubro Group Company, Mindtree applies its deep domain knowledge to 280+ enterprise client engagements to break down silos, make sense of digital complexity and bring new initiatives to market faster. We enable IT to move at the speed of business, leveraging emerging technologies and the efficiencies of Continuous Delivery to spur business innovation. Operating in more than 15 countries across the world, we're consistently regarded as one of the best places to work, embodied every day by our winning culture made up of over 21,800 entrepreneurial, collaborative and dedicated "Mindtree Minds".