Modern Data Platform for Consumer Goods Businesses

WHITEPAPER

While we witnessed shifts in the retail and consumer goods business ecosystem during the last quarter of 2020 the ebbing Omicron wave is forcing businesses to rethink back-to-office policies, differentiated channels to reach consumers, and continued acceleration in pace of adoption of digitization. Some estimates say that companies met three to four years of digitization targets within few months during the pandemic. What we have now is a new paradigm of retail and consumer goods businesses that is aided by technology. This paper describes how a modern data platform has become an imperative for consumer goods businesses while highlighting the major trends in the industry. Subsequently, it also covers the critical components that make a modern data platform.

Introduction

This paper is aimed at the technology leaders and CXO's who are trying to build a resilient, scalable, and future-ready data platform for their consumer goods businesses. We have seen tectonic shifts in methods of business operations for all kinds of businesses in the last three years. The necessity of a contactless world and the resulting push for digitization has put data at the centre of everything. According to an estimate by IDC, the spending on big data technology and services will grow with a five-year CAGR of 14.6% over 2020-2025.

While businesses have already realized the importance of capturing and analysing data generated in day-to-day operations, they have struggled to come up with methods, architecture, and the operational setup that would stand the test of time. The blistering pace of innovation in compute and storage technologies have also added to the woes of the leaders.

In a such a scenario, it becomes an imperative that businesses get the fundamentals of their data platform architecture right. Once businesses identify the key components and understand how those can put together to build systems that would last longer and would be able to cater to change in environment and business needs, they can get one-step closer to realizing the vision of becoming a data-driven business.

With detailed discussions around key business and technical trends, this paper aims to assists data and analytics leaders in consumer goods industry plan their technology and investment strategy better

Key Trends in Consumer Goods Industry

1. **Disintermediation and direct to consumer sales:** Given the availability of several alternative channels, many consumer goods businesses are now trying to reduce reliance on intermediaries and engage directly with consumers. Such engagement with consumers helps to gather a better understanding of needs, get better demand forecast, and increase. Direct engagements with consumers help the business capture first-party data from consumers that helps them develop a better understanding of consumption patterns and consumer behaviour.

2. **Everywhere commerce:** In past few years, the retail and commerce landscape has been disrupted with proliferation of new channels. The touchpoints with customers are no longer just restricted to mobile apps or websites. Consumers can shop using Voice Assistants, Smart Speakers, or even in livestreams. With increase in such channels, it becomes imperative for consumer goods businesses to gather and make sense of data from those channels.

3. **Increased focus on sustainability:** The millennials and Gen-Z have greater awareness of and appreciation for climate change. Hence, they demand the same from businesses they interact with. Recent technological advances such smart packaging and blockchain make it possible for consumer goods businesses to bring in more transparency in their interactions and adhere to local regulations.

4. **The near-instant delivery-based business models:** Last-mile delivery and within a short duration such as overnight shipment have been embraced by many retailers. In 2021, several Q-Commerce businesses have come up that promise to fulfil customer demands in 10 mins. While the viability of their business models can be debated, we cannot deny their impression on the retail landscape across the globe. It is now important for consumer goods businesses to either collaborate with them or think of similar business models to remain in consumer's considerations.

5. **Local and smaller competitors:** The market is replete with several new start-ups that are trying to create a niche for themselves with differentiated products and offerings. Moreover, ecommerce platforms such as Amazon have lowered the barriers of entry for such businesses. They have

an easy path to consumers and thus can compete with the established giants in specific niche segments.

A combined impact of all trends has created a situation where Consumer Goods businesses need to understand their customers better to cater to their ever-changing needs across the channels and keeping the several smaller competitors at bay. The situation combined with the stretched global supply chain has squeezed the margins for the businesses. Enterprises are forced to ameliorate the delicacy and speed of their opinions on fraudulent pitfalls. AI-powered warranty solutions should provide enterprises with innovative ways to reduce warranty cost and waste, potential fraudulent claims, and streamline claim processes thereby reducing overall warranty spend

.

4 C's Disrupting Consumer Goods Businesses

Consumer	Competitor	Channels	Cost Pressure
Extremely demanding millennials & Gen Z with multitude of options	Plethora of small brands that can reach customers directly and competition from local brands	Shopping has gone beyond brick-and-mortar & ecommerce; we have social and live commerce as well	Increasing raw material costs and supply chain disruptions are leaving little margin for businesses
Understand & serve them better	Respond to competition promptly	Track growth and presence across channels & unify the experience	Ensure growth while cutting unwanted cost by optimizing operations & supply chain

Establish a scalable and modern data & analytics infrastructure, solution, and value chain that brings data from maximum possible sources in real-time and generates insights with minimum possible latency

Figure 1: Trends in Consumer Goods Industry

To respond to all the challenges, the consumer goods businesses need a comprehensive data strategy that enables them to build a modern data platform with the capability to bring-in data from maximum possible touchpoints and generate actionable insights from that data with minimum possible latency.

Aligning to the needs of the businesses, we have identified the key themes in data, analytics, and artificial intelligence (AI). Moreover, we have also highlighted key considerations for the six major aspects of a modern data platform.

Analytics & Data Sciences

- Conversational Analytics
- IoT Analytics
- Machine Learning Deep Learning
- Interactive and Adhoc Query
- Interactive and collaborative Analytics Notebooks
- Information Explorer
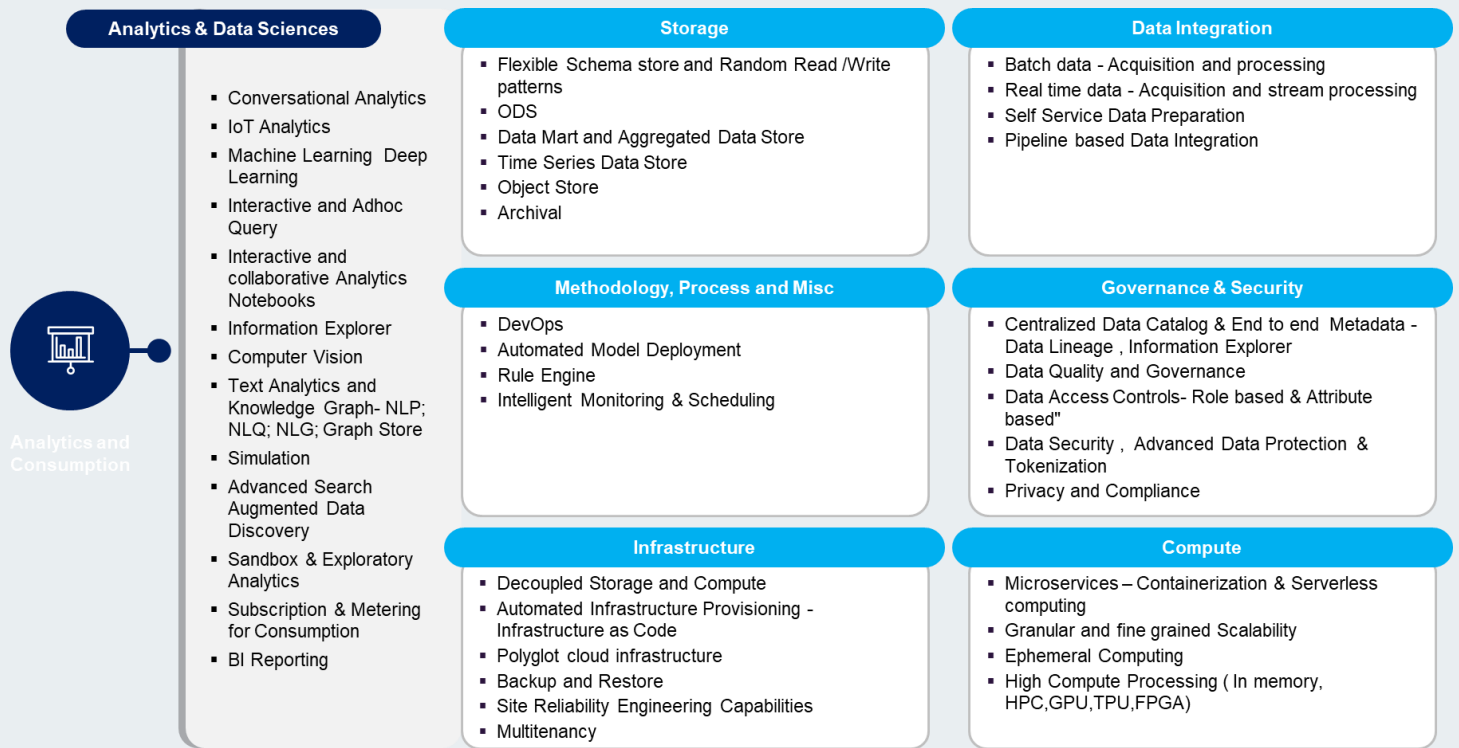- Computer Vision
- Text Analytics and Knowledge Graph- NLP; NLQ; NLG; Graph Store
- Simulation
- Advanced Search Augmented Data Discovery
- Sandbox & Exploratory Analytics
- Subscription & Metering for Consumption
- BI Reporting

Analytics and Consumption

Storage

- Flexible Schema store and Random Read /Write patterns
- ODS
- Data Mart and Aggregated Data Store
- Time Series Data Store
- Object Store
- Archival

Data Integration

- Batch data - Acquisition and processing
- Real time data - Acquisition and stream processing
- Self Service Data Preparation
- Pipeline based Data Integration

Methodology, Process and Misc

- DevOps
- Automated Model Deployment
- Rule Engine
- Intelligent Monitoring & Scheduling

Governance & Security

- Centralized Data Catalog & End to end Metadata - Data Lineage , Information Explorer
- Data Quality and Governance
- Data Access Controls- Role based & Attribute based"
- Data Security , Advanced Data Protection & Tokenization
- Privacy and Compliance

Infrastructure

- Decoupled Storage and Compute
- Automated Infrastructure Provisioning - Infrastructure as Code
- Polyglot cloud infrastructure
- Backup and Restore
- Site Reliability Engineering Capabilities
- Multitenancy

Compute

- Microservices – Containerization & Serverless computing
- Granular and fine grained Scalability
- Ephemeral Computing
- High Compute Processing (In memory, HPC,GPU,TPU,FPGA)

Figure 2: Major Themes in Data, Analytics & AI

Further, we have also produced a reference architecture for modern data platforms considering the key trends. In Figure 3, we have depicted the architecture and highlighted the pivotal components below. In the section that follows the architecture, we have delved deeper into each of the components to show how they can contribute to building a robust, agile, and new-age data platform.
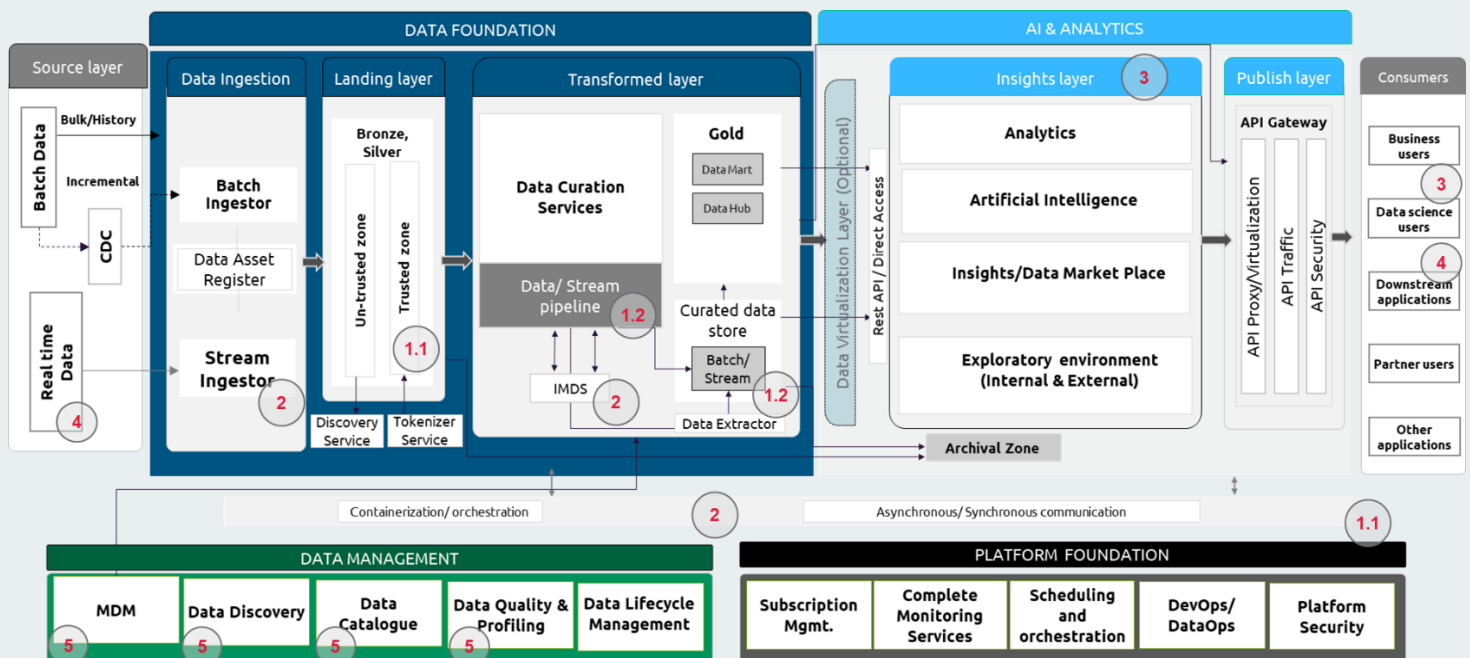

Figure 3: Modern Data Platform Architecture

The reference architecture has several components numbered according to the themes that they relate to.

1. Convergence of internet of things (IoT), cloud, and big data
   i. Big data on cloud to grow faster than on premise. Existing Big data implementations are migrating to hybrid (on-prem and cloud) data lake environment
   ii. Increased focus on cloud agnostic computing to leverage hybrid ecosystem and reduce risk on cloud providers dependency
2. Fast data is getting mainstream
3. A step towards factory model for data and AI products: MLOps
4. Emergence of customer data platform
5. Evolving data governance landscape

Convergence of Internet of Things (IoT), Cloud, and Big Data

Since businesses need to capture and analyse a greater variety and volume of data to respond to the changing market dynamics and to understand their consumers better, it is imperative for them to adopt a cloud-first strategy for their data, analytics, and AI needs. The strategy banks on the convergence of Cloud

Cloud-based analytics and AI workload allows businesses to not only ingest several types of data i.e., geospatial, audio files etc. but also continuously experiment with use-cases and services. Moreover, the focus on serverless compute and separation of compute and storage ensures that the cost of such experiments remains low.

The growing dependence on cloud-based data and analytics solutions has led to two different trends relating to infrastructure and architecture, respectively. The first one is several deployment options on cloud and the second is inclination towards a cloud-agnostic solution development.

- Big data on cloud to grow faster than on premise. Existing big data implementations are migrating to hybrid (on-prem and cloud) data lake environment

As businesses grow and expand with cloud-based analytics, we see a growing trend to a hybrid- and multi-cloud deployment. There are several factors driving this trend such as avoiding a vendor lock-in, mergers and acquisitions, and geographical expansion.
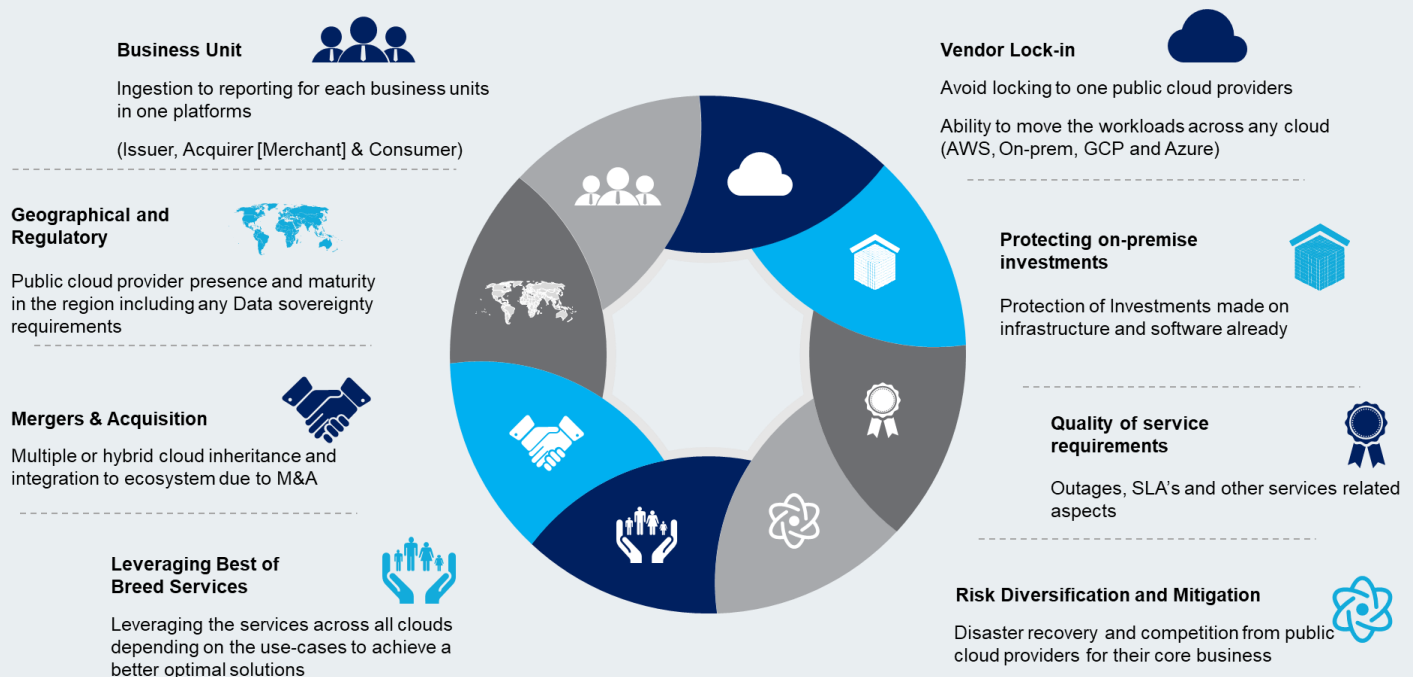
Business Unit

Ingestion to reporting for each business units in one platforms

(Issuer, Acquirer [Merchant] & Consumer)

Geographical and Regulatory

Public cloud provider presence and maturity in the region including any Data sovereignty requirements

Mergers & Acquisition

Multiple or hybrid cloud inheritance and integration to ecosystem due to M&A

Leveraging Best of Breed Services

Leveraging the services across all clouds depending on the use-cases to achieve a better optimal solutions

Vendor Lock-in

Avoid locking to one public cloud providers

Ability to move the workloads across any cloud (AWS, On-prem, GCP and Azure)

Protecting on-premise investments

Protection of Investments made on infrastructure and software already

Quality of service requirements

Outages, SLA's and other services related aspects

Risk Diversification and Mitigation

Disaster recovery and competition from public cloud providers for their core business

Figure 4: Drivers for multi/hybrid cloud adoption

Based on our experiences with several customers, following are the three different options to construct hybrid/multi cloud data platforms:

Resident Data + Integrated Compute	Integrated Data + Integrated Compute	Resident & Integrated Data + Integrated Compute
- All data layers in single platform (Resident) - Containerized Data Processing/ compute across platforms (Integrated) - Data movement across platforms causing network and egress costs	- Logical data platform with Data fabric (Integrated) with processed/limited data sharing across platforms - Containerized Data Processing/ compute across platforms (Integrated) - Data movement across platforms by using local data to reduce the egress costs	- Centralized raw and harmonized data with consumption being local (Resident & Integrated) - Containerized Data Processing/ compute across platforms (Integrated) - Drift in data models and KPI's on consumption losing the single version of truth.

- Increased focus on Cloud agnostic computing to leverage hybrid ecosystem and reduce risk on cloud providers dependency

Apart from the deployment options, there is also a need for building a cloud-agnostic data processing or computing framework. The following picture present Tech Mahindra's framework for the purpose:
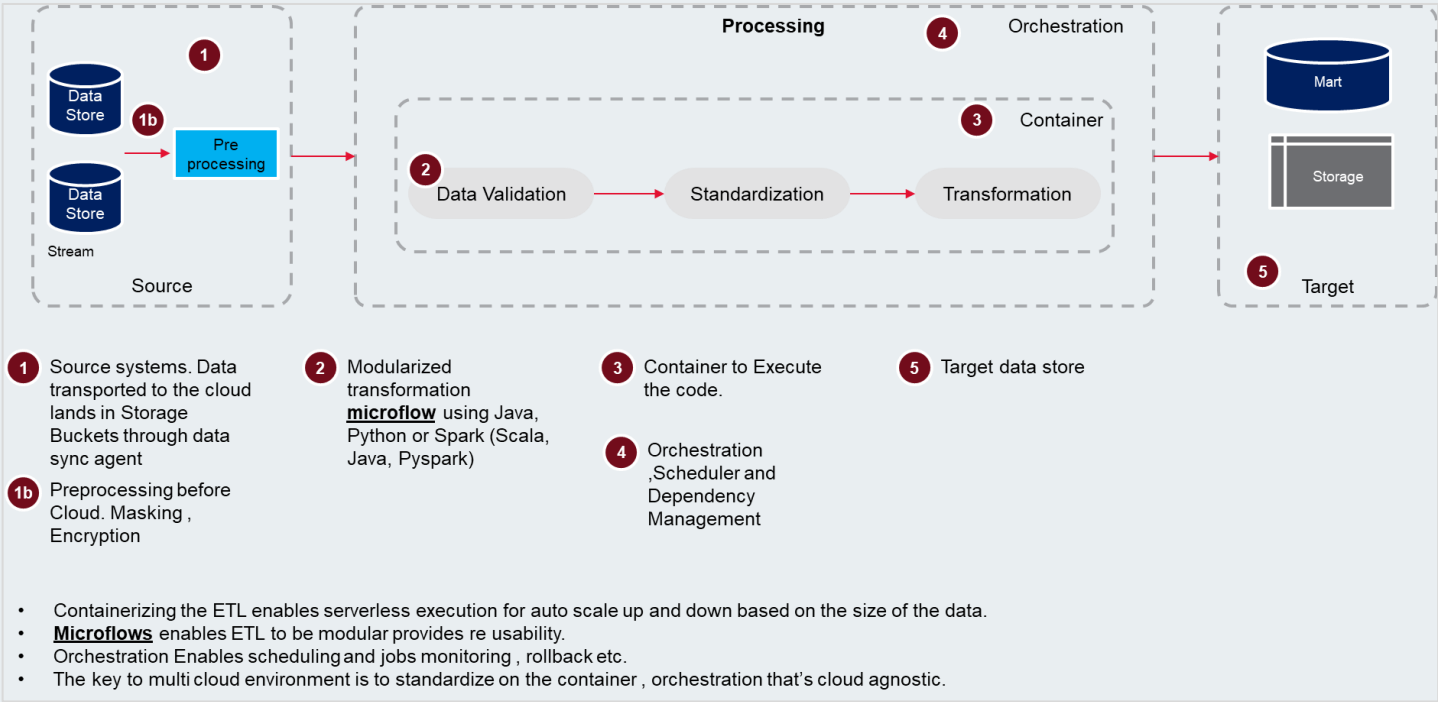

1. Source systems. Data transported to the cloud lands in Storage Buckets through data sync agent

1b. Preprocessing before Cloud. Masking , Encryption

2. Modularized transformation **microflow** using Java, Python or Spark (Scala, Java, Pyspark)

3. Container to Execute the code.

4. Orchestration ,Scheduler and Dependency Management

5. Target data store

- Containerizing the ETL enables serverless execution for auto scale up and down based on the size of the data.
- **Microflows** enables ETL to be modular provides re usability.
- Orchestration Enables scheduling and jobs monitoring , rollback etc.
- The key to multi cloud environment is to standardize on the container , orchestration that's cloud agnostic.

Figure 5: Cloud-agnostics Processing/Compute Framework

Fast Data is Getting Mainstream

Velocity side of Big Data

Manage continuous streaming of in-flight data that helps business derive value and relevance for business-critical decisions

Extremely time sensitive and often difficult to extract and analyze

Complicated relationship or event patterns

Monitoring, responding and gaining valuable insight in real time

Fast Data Requirements

Ingest
Ingest constant streams of data

Analyze
Analyze each event in the stream

Respond
Respond to each event or to outliers

Integrate
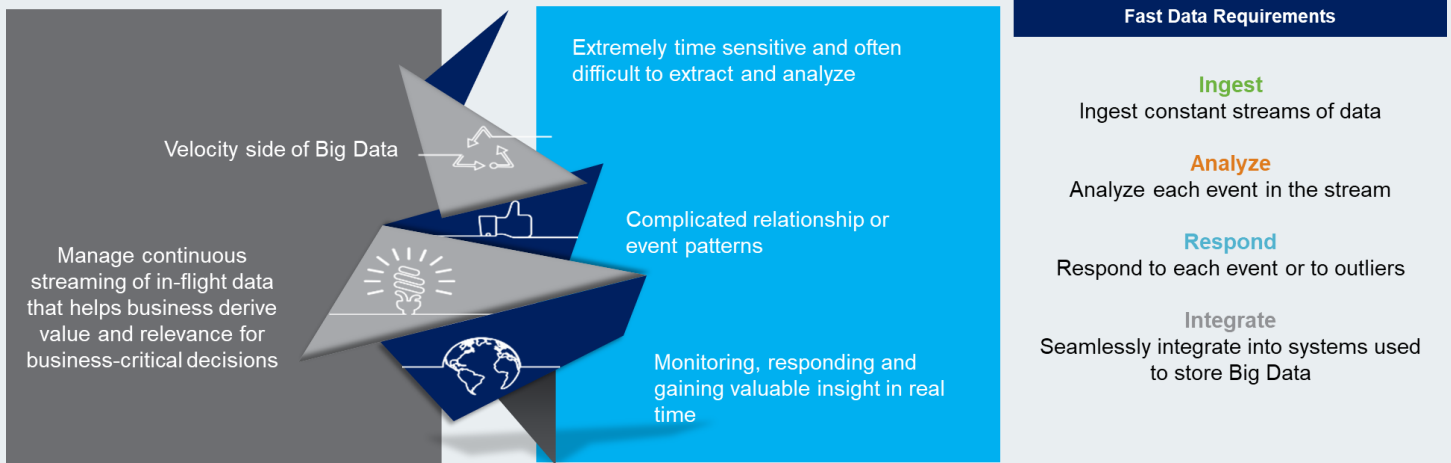Seamlessly integrate into systems used to store Big Data

Figure 6: Fast Data Processing is Key to Survival & Growth of Businesses

We highlighted the challenges that consumer goods businesses are facing on multiple fronts. Things get ever more complicated for them given the breakneck pace at which everything around us is changing. In the current situation, ingestion and processing of Fast Data has become a basic requirement for their survival and growth.

The capability refers to an integration of systems that enables businesses to identify and respond to risk and opportunities with minimum delay. Thereby making them more agile and capable of exploiting even short-term opportunities while avoiding losses by constantly assessing risks. Handling Fast Data requires considerable technical maturity and a modern set-up. Figure 7 depicts the conceptual architecture of a solution that can ingest, process, and analyse fast data at scale
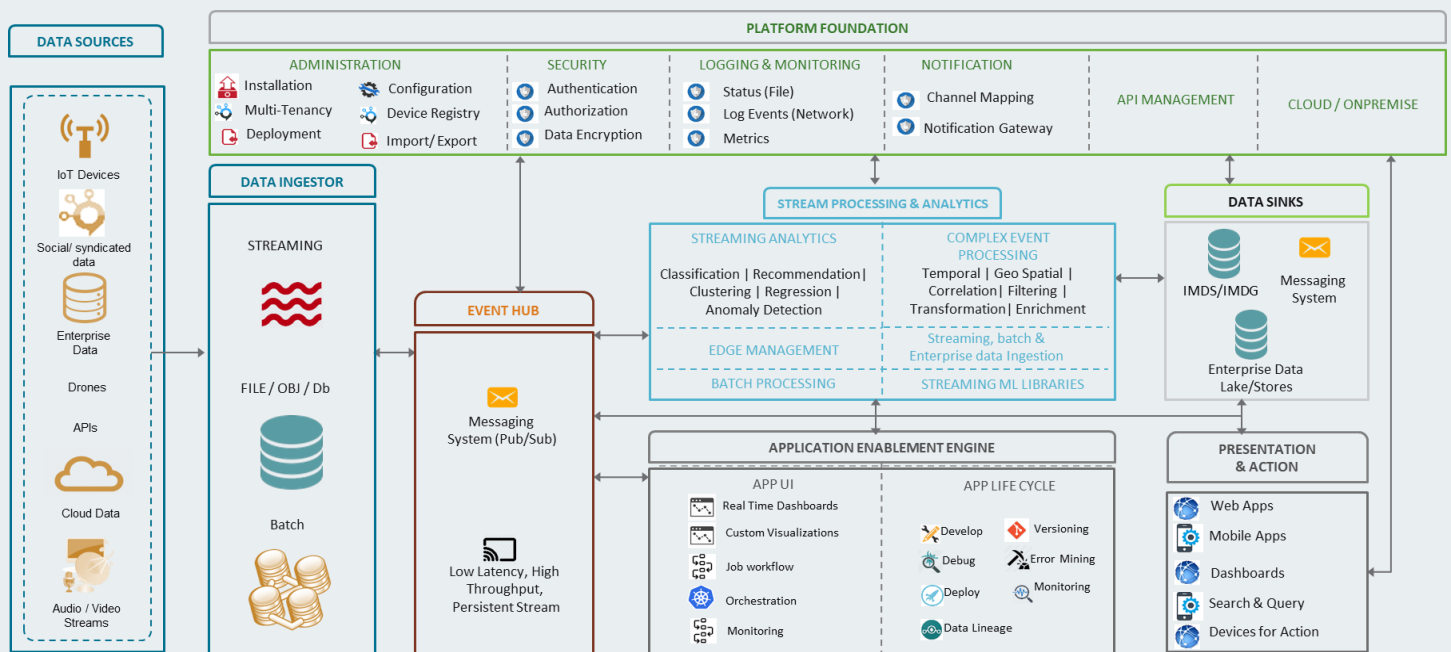

Figure 7: High-level Conceptual Architecture of a Fast Data Processing & Analysis system

Emergence of Customer Data Platforms

The increase in number of customer touchpoints and transactions across channels has necessitated for consumer goods businesses to have a better understanding of customers. Businesses have been striving for the customer 360 view for a long time now. The goal, however, has remained elusive for most of the businesses.

The emergence of customer data platforms (CDP) aims to solve that problem through its purpose-built design that not only integrates and unifies data from several sources but also allows enrichment of customer data by bringing third party data into the mix. Not just that, CDPs also come with AI/ ML models that help calculate metrics such as lifetime value (LTV) (churn propensity, etc.

On the implementation front, CDP presents the same challenge to businesses as other such strategic investments i.e. whether to build one or buy one. In the following diagram we have tried to provide the information required to make that decision.
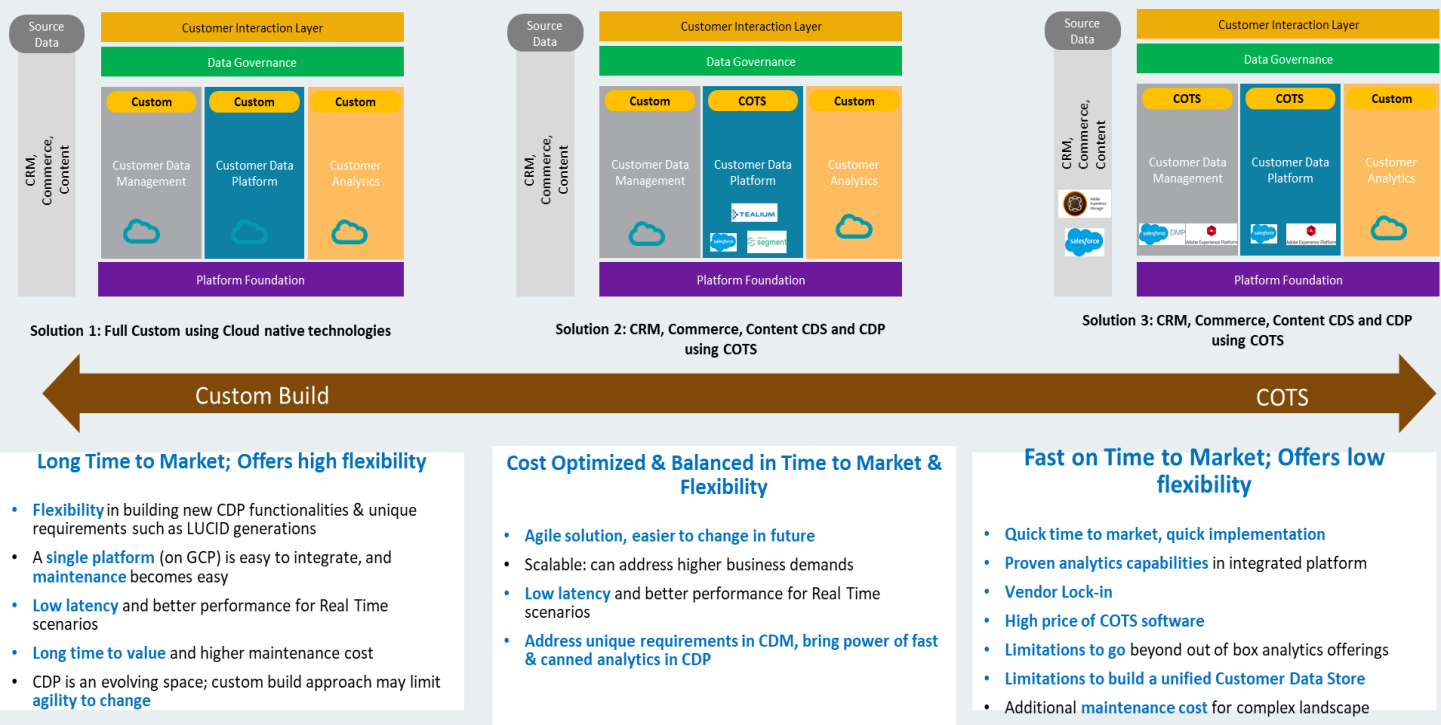

Solution 1: Full Custom using Cloud native technologies

Solution 2: CRM, Commerce, Content CDS and CDP using COTS

Solution 3: CRM, Commerce, Content CDS and CDP using COTS

Custom Build   COTS

Long Time to Market; Offers high flexibility

- **Flexibility** in building new CDP functionalities & unique requirements such as LUCID generations
- A **single platform** (on GCP) is easy to integrate, and **maintenance** becomes easy
- **Low latency** and better performance for Real Time scenarios
- **Long time to value** and higher maintenance cost
- CDP is an evolving space; custom build approach may limit **agility to change**

Cost Optimized & Balanced in Time to Market & Flexibility

- **Agile solution, easier to change in future**
- Scalable: can address higher business demands
- **Low latency** and better performance for Real Time scenarios
- **Address unique requirements in CDM, bring power of fast & canned analytics in CDP**

Fast on Time to Market; Offers low flexibility

- **Quick time to market, quick implementation**
- **Proven analytics capabilities** in integrated platform
- **Vendor Lock-in**
- **High price of COTS software**
- **Limitations to go** beyond out of box analytics offerings
- **Limitations to build a unified Customer Data Store**
- Additional **maintenance cost** for complex landscape

Figure 9: CDP: Build vs. Buy

Customer Data Platforms hold significant promise for consumer goods businesses. However, the planning and implementation for that needs to be done with extreme care keeping the business objectives and vision in mind.

Evolving Data Governance Landscape

In the past, data governance has often appeared as an afterthought for businesses in their roadmaps. The massive data landscapes with innumerable data & analytics assets and stringent regulatory frameworks have forced them to prioritize data governance and all the entailing activities i.e., metadata management, master data management (MDM), data quality mgmt., data catalog, lineage, etc.

On the supply side, a new breed of data governance applications and concepts have come up to solve the challenges for organizations. Following are some of the key happenings in the data governance space:

Active Metadata Management Backed by Knowledge Graphs: Organizations are leaning towards use of knowledge graphs to map the data entities across the organizations in a proactive and automated manner. The resulting setup enables them to build not only a comprehensive data lineage but also the foundation of an enterprise data fabric.

Data Catalogs: The usage of data catalogs has also increased in past couple of years with several new vendors producing relevant products and solutions. Data catalogs enabled greater visibility and availability of data assets across organizations leading to greater adoption of analytics and faster time-to-insight through better reusability of data assets.

Machine Learning Assisted Data Quality Management: Machine Learning (ML) methods are being used across all the data management and governance activities these days to automate and enrich those. Data quality management is an area where ML is being used extensively to detect anomalies and impute values for data enrichment where applicable. Following figure depicts our ML-assisted data quality measure methodology at a high level.
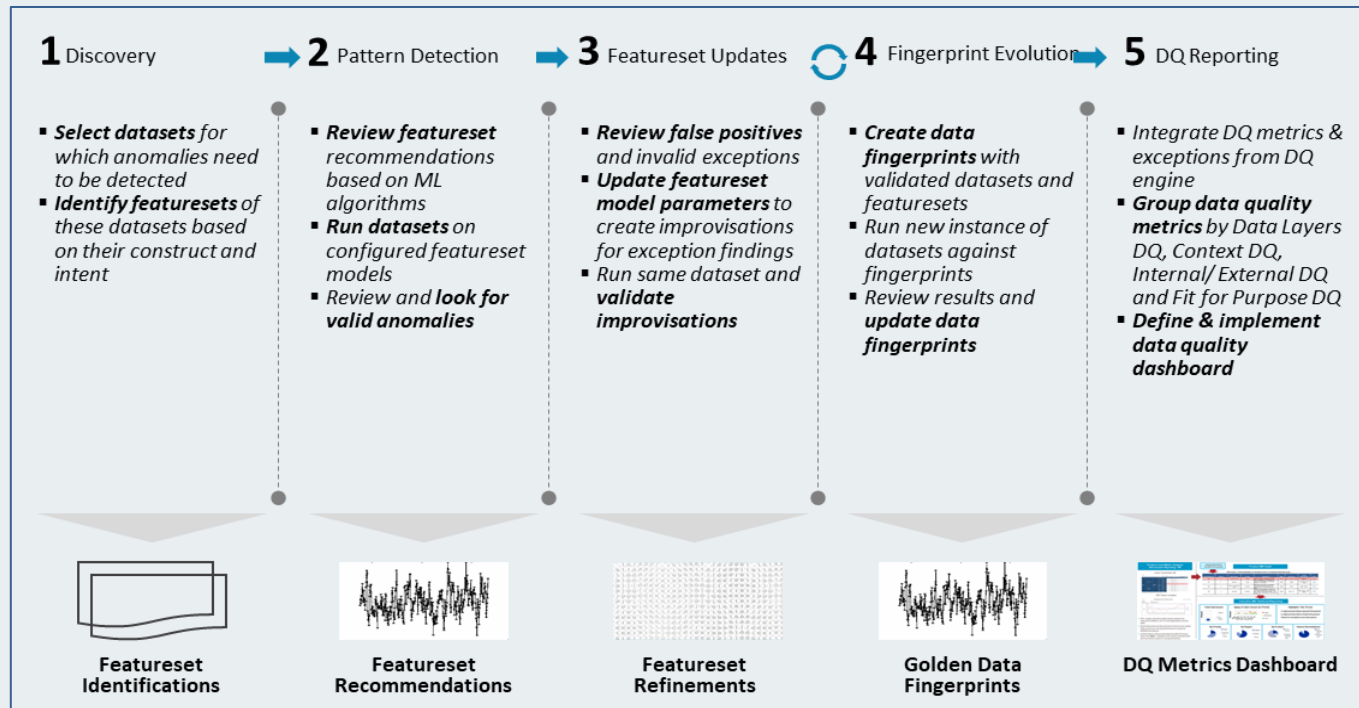

1 Discovery → **2** Pattern Detection → **3** Featureset Updates ↻ **4** Fingerprint Evolution → **5** DQ Reporting

- *Select datasets* for which anomalies need to be detected
- *Identify featuresets* of these datasets based on their construct and intent

- *Review featureset recommendations* based on ML algorithms
- *Run datasets* on configured featureset models
- Review and *look for valid anomalies*

- *Review false positives and invalid exceptions*
- *Update featureset model parameters* to create improvisations for exception findings
- Run same dataset and *validate improvisations*

- *Create data fingerprints* with validated datasets and featuresets
- Run new instance of datasets against fingerprints
- Review results and *update data fingerprints*

- Integrate DQ metrics & exceptions from DQ engine
- *Group data quality metrics* by Data Layers DQ, Context DQ, Internal/ External DQ and Fit for Purpose DQ
- *Define & implement data quality dashboard*

Featureset Identifications | **Featureset Recommendations** | **Featureset Refinements** | **Golden Data Fingerprints** | **DQ Metrics Dashboard**

Figure 10: ML-assisted Data Quality Framework

Intelligent Master Data Management: Getting the X360 view is still an elusive dream for majority of businesses since the traditional process of master data management is inherently slow with several manual interventions. Employing ML methods to build inherent intelligence into the master data management process can expedite that and improve its accuracy. Machine learning can be used in each step of MDM i.e., statistical sampling for sampling of data from various sources, natural language processing for standardization, matching and merging, etc.

While several new-age applications promise Intelligent MDM, businesses with high D&A maturity and skill can also build such a solution in-house.

Way Forward

In the next five years, the variety, volume, and velocity of the data generated by businesses will grow multi-fold. According to Statista research, more than 180 zettabytes (1021) of data will be created by 2025. A business's ability to make sense of such huge volumes of data in a timely, cost-effective, and flexible manner will be a key differentiator in the days to come. The architecture and concepts put forward in this paper is a step in that direction.

While we have laid out our approach to building a futureproof data platform, the initiative may not succeed if the softer aspects of implementation such as skill building, change management, and user training are not taken care of. There must be a robust data strategy and a roadmap in place that factors in the organization vision, structure, culture, and existing processes and practices in place.

References

1. Spending on big data, analytics to touch $2 billion in 2021: IDC. (2021, August 24). https://Www.Business-Standard.Com/Article/Companies/Spending-on-Big-Data-Analytics-to-Touch-2-Billion-in-2021-Idc-121082401341_1.Html. Retrieved April 1, 2022, from https://www.business-standard.com/article/companies/spending-on-big-data-analytics-to-touch-2-billion-in-2021-idc-121082401341_1.html

2. More than 180 zettabytes (1021) of data will be created by 2025: Statista. (2022, March 18). https://www.statista.com/statistics/871513/worldwide-data-created/

Authors:

Shivam Kumar and Venkatesh Thyagarajan

Tech Mahindra

www.youtube.com/user/techmahindra09
www.facebook.com/techmahindra
www.twitter.com/tech_mahindra
www.linkedin.com/company/tech-mahindra
www.techmahindra.com