Whitepaper

Data Warehouse 4.0

Reimagining Data Warehouse for Cloud & Big Data

By **Ranganath Ramakrishna**

Table of Contents

1 Paradigm Shift, This Time for Real

The technology industry has a questionable reputation for using the term paradigm shift. Every new feature, product version, or insignificant knob is marketed as a silver bullet, which would magically solve every challenge.

The perception is the same about the data industry, perhaps. Data Warehouses have been the workhorses for analytics systems - reporting and dashboarding for nearly four decades. At regular intervals, new technologies have emerged which were supposed to supersede or obviate Data Warehouses.

1.1 Perfect Storm

Today, Big Data and Cloud are impacting the data warehousing technology in a fundamental and irreversible way. Big data widens the cognitive bandwidth of an enterprise by providing capability to harness multi-structured and streaming data. Cloud significantly reduces entry barriers for leveraging data by reducing costs and complexity of procuring and setting up the data warehouse. By themselves, big data and cloud are individually potent, but together they are transformational. They offer unprecedented opportunity to enterprises to simplify yet amplify the potential of their data warehouse.

The purpose of this white paper is to examine the structural limitations of the traditional data warehouses and how cloud can liberate the traditional ones from these structural constrains. We call this new approach Data Warehouse 4.0. Data Warehouse 4.0 is not just about re-platforming or migrating existing on-premise Data Warehouses to the cloud, it is also about re-imagining the potential of data to business.

2 A New Vision for the Data Warehouse

Contrary to the common belief, fundamental technologies evolve at glacial speed. Relational database, a technology of 1970's is still going strong, the internet is at least three decades old. However, small innovations accumulate over time and every few decades reach a critical mass of disruption. They change not just the technology, but the underlying economics leading to new waves of adoption and business disruption. The market place gets leveled and new set of competitors emerge. Old players either re-invent themselves to meet the new normal or become irrelevant or obsolete.

To take advantage of the new technology, it necessary to start from first step and build solutions on the strengths of the new technology. Cloud and Big Data represent a fundamental discontinuity. Therefore, it is necessary to relook at every aspect of Data Warehouse technology, including customer experience, cost structures, use cases, data and application architectures, skill sets, and tool sets.

2.1 Limitations of Traditional Data Warehouse

The Data Warehouse, as a technology, suffered many debilitating problems due to the underlying infrastructure and architecture of core platforms. The architecture placed certain hard constraints, which made the Data Warehouse platforms costly, slow, complex, and effort-intensive. To mention few constraints:

1. **Complexity** – The entire life cycle, rght from procuring the hardware, software, racking and stacking, installation, setting up DR/HA/BCP, etc. was slow. It was possible only for the biggest enterprises with deep pockets.

2. **Lack of Agility & High Cost** – The projects on the legacy Data Warehouse platforms rolled on for months and years. Although few projects completed on time and on budget, most were never on value. By the time the project got delivered, the business had moved on. Businesses felt like they were feeding white elephants which guzzled up huge amounts of cash, but rarely delivered the concomitant value.

3. **Rigidity** – Projects needed upfront requirements, which once defined could not be changed without an elaborated process. Data models had to be defined and agreed upon upfront before any data could be ingested. Every change resulted in change requests, additional costs, and updated gnatt charts. This rigidity frustrated businesses, resulting in building shadow IT teams.

4. **Poor Asset Utilization** – Multi-million-dollar platforms were most of the time lying idle and depreciating. The peak sizing done to avoid capacity issues compelled the business, paid too much for too little usage most of the time. Also, when utilization exceeded the capacity, it led to poor customer experience.

5. **Operational Complexity** - Operational complexity was very high as there was a never-ending cycle of upgrades, EOL of hardware and software, patches, bug fixes, etc. Enterprises spent most of time managing servers and databases instead of managing data and customer experience.

6. **High End Skills** – Operating, architecting, and tuning many of the proprietary platforms was akin to black magic. It was reserved for a chosen few, driving up costs significantly.

Cloud lifts most of these constraints. It also adds new capabilities not available earlier. However, no technology, including cloud, can substitute rigorous thinking and hard work, so one still needs to architect and design with business outcomes and business requirements in mind.

2.2 Embracing Big Data and Cloud

2.2.1 Big Data

Big data expands traditional Data Warehouse horizontally by adding new types of data and expanding the universe of use cases. Big data enriches Data Warehouse by expanding the ambit of what is possible. By adding data, streaming and multi-structured data, the next-generation Data Warehouses are expected to augment core transactional data with the data generated from IoT devices, business applications and people. Data Warehouse can be integrated natively with machine learning frameworks like Spark to deliver use cases related to predictive, prescriptive, and causal analytics.

2.2.2 Cloud

Cloud re-platforms the Data Warehouse vertically by changing the core layers of the Data Warehouse stack. Object storage, decoupled compute, new dynamic query engines, connectors for machine learning, cluster less ETL, and streaming technologies provide an entirely new dimension of capability and price-point. Many Data Warehouse products are being built with cloud-native technology for Data Warehouse fully exploiting the benefits for the cloud.

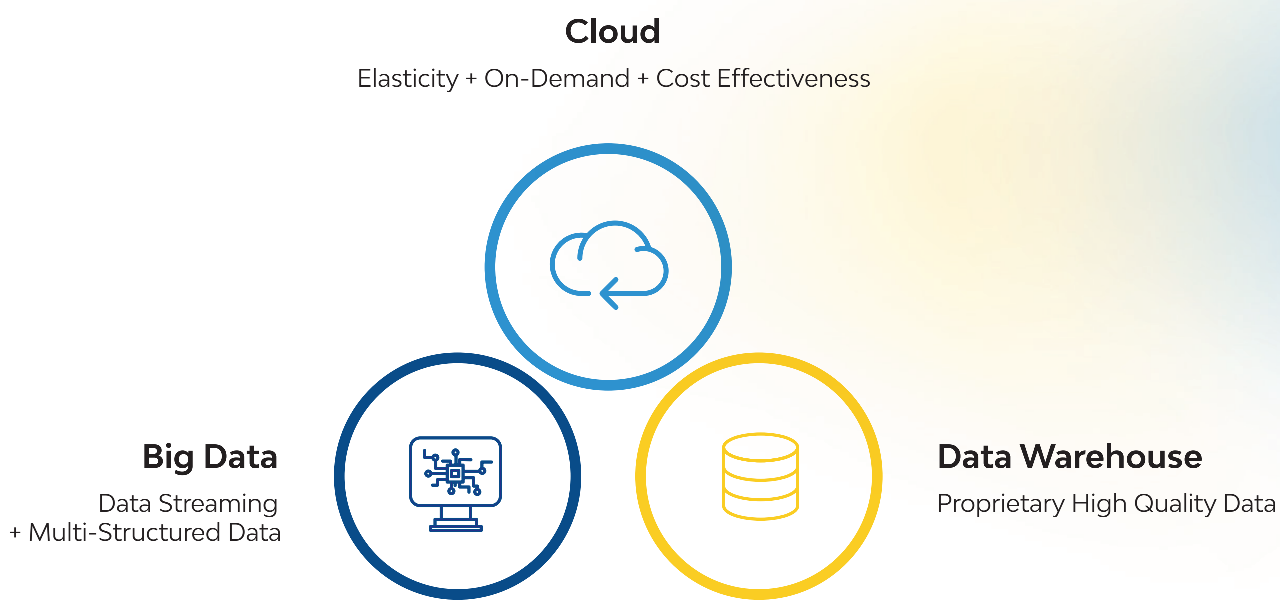

Figure 1:
Data Warehousing Amplified with Big Data and Cloud

2.3 Defining Data Warehouse 4.0

Data Warehouse is defined as a subject-oriented, integrated, time-variant, and non-volatile collection of data to support decision-making. The definition does not represent a product, it defines a set of capabilities inherently provided by an architecture. The diagram below maps the evolution of the Data Warehouse over the years across multiple dimensions.

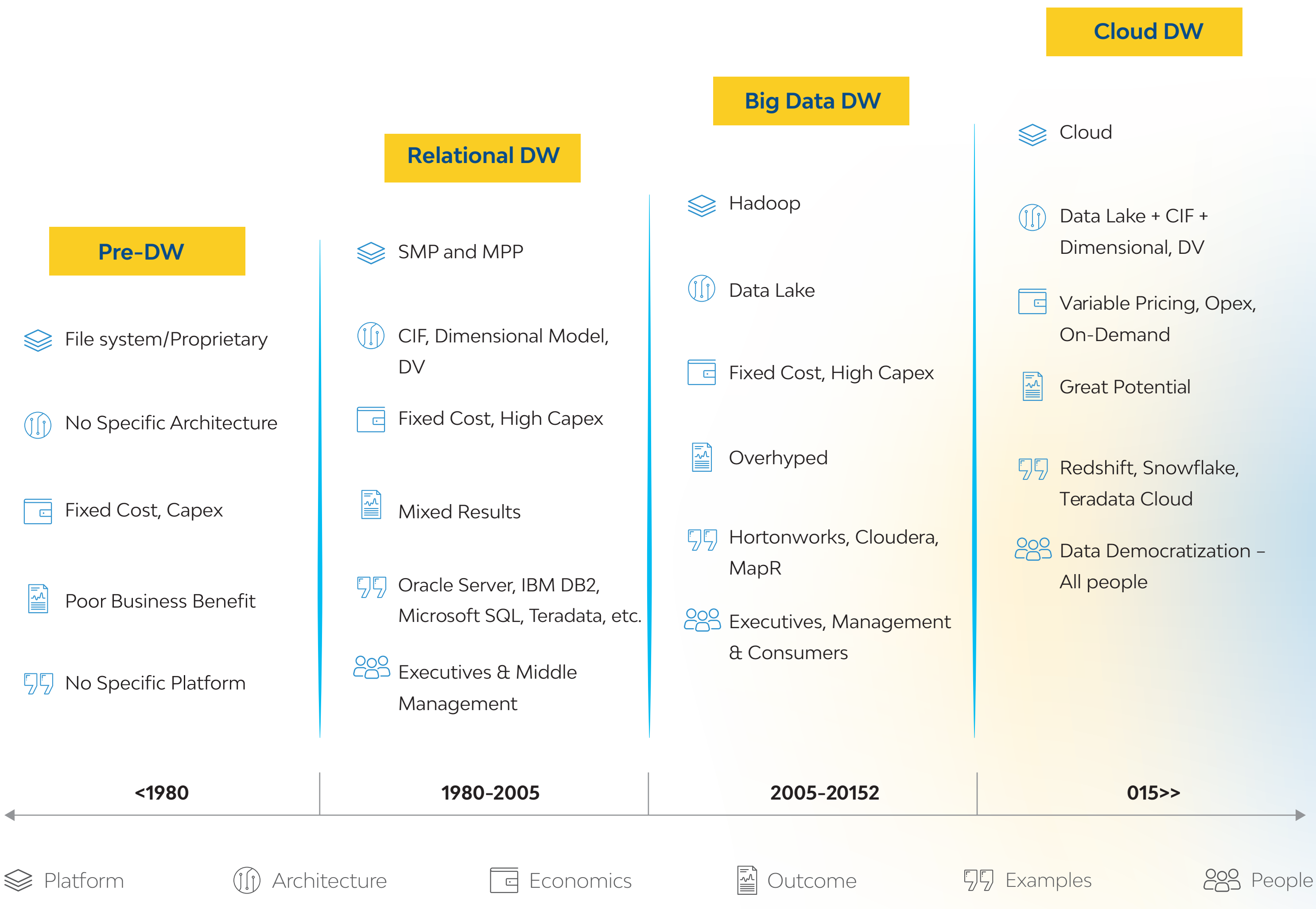

Figure 2: Evolution of Data Warehouse

However, the scope and context of Data Warehouse has expanded significantly with the arrival of big data and cloud. It is now necessary to relook at the definition, considering the new breakthrough technologies. We define Data Warehouse 4.0 by the fundamental capabilities it must natively provide. It is a combination of infrastructure and architecture which should provide a set of core capabilities relevant to digital age.

2.4 Core Capabilities of Data Warehouse 4.0

Figure 3: Core Capabilities of DW 4.0

The core vectors that define an enterprise Data Warehouse in the digital era are:

2.4.1 Decoupled Storage and Compute

One of the core principles of designing enterprise systems is to achieve loose coupling with high cohesion. It is desirable to have decoupled physical layers, but project a unified façade to the end user. This will enable to manage and scale the physical layers independently, without impacting end users. Traditional approaches to data warehousing coupled compute, storage, and query engines tightly. This necessitated scaling compute when more storage was required and vice versa. The coupled architecture forced enterprises to solve a heterogeneity of use cases with single monolith architecture, creating cost and scale inefficiencies.

Data Warehouse 4.0 should be essentially built leveraging the cloud object storage and virtual machines. This enables to scale and pay for compute and storage separately. An active archive use case might need more storage compared to modeling & simulation application, which needs more compute. Both the use cases can be now handled appropriately by scaling the desired layer.

2.4.2 Elastic Scalability

Traditional Data Warehouses suffer from the challenge of fixed capacity resources that are inflexible to scale up or down based on demand. This is an inherent structural limitation due to the very nature of on-premise infrastructure. This cannot be circumvented by using any amount of clever engineering.

Data Warehouses typically struggled to scale along the vectors of data, compute/queries, and users. The Cloud offers an ability to scale across all these vectors instantly and almost infinitely.

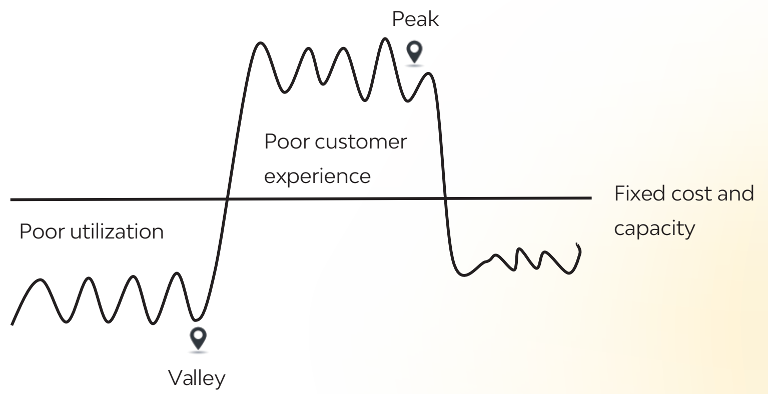

Figure 4: Problem of Fixed Capacity

2.4.3 Cost Effectiveness

Traditional Data Warehouses built on proprietary hardware and software incur significant cost. Cost of storing a terabyte varies anywhere between USD 10,000 to 30,000. This represented a significant entry barrier for more enterprises. On the top of platform, there are significant costs in terms of specialized skills required to design, build, and maintain these systems. Only the largest enterprises could afford such huge costs. The costs also escalate due to the need to setup multiple environments - DEV, TEST, QA, Production and have stood by instances for production to achieve high availability. The downside being that all this upfront investment at best could guarantee technical excellence, not business success. Many business cases did not deliver the expected value, but enterprises were left with white elephants gobbling up a lot of cash.

With the emergence of big data, the volume of data to be managed is seeing exponential growth. On-premise platforms cannot cost-effectively address big data. Also, with digital, every business is a data business so the core data platform needs to provide a low-cost entry point for all businesses.

2.4.4 Flexibility

The cognitive bandwidth of today's enterprises has expanded significantly, including structured, semi-structured and unstructured data. The traditional Data Warehouse predominantly got data in either fixed width, delimited files, or ODBC/JDBC connection to database. Today, the universe is enlarged to a vast variety of new semi-structured formats, like JSON, XML, AVRO, Parquet, ORC and unstructured formats like audio, video, and images. Data Warehouse 4.0 needs to support all the above myriad formats if the business must fully exploit data for business success.

Data Warehouse 4.0 also needs to provide schema on read capability as unlike the structured data sources, which have low frequency control changes, the format of semi-structured data can change frequently. Therefore, the technology should not only support storage of semi-structured and unstructured data, it also needs to automatically parse semi-structured data and provide mechanism to infer schema on read.

2.4.5 Insights

Data Warehouse 4.0 should enable a rich variety of use cases. Traditional Data Warehouses were focusing too much on KPI monitoring using reports and dashboards. This is not sufficient for the digital era. Data Warehouse must enable predictive, prescriptive, and causal analytics.

Also, velocity is one of the key dimensions of big data. IoT devices, instrumented applications and PDAs continuously emit data. Many new business models are directly co-related to harnessing streaming data to deliver differentiated customer experience for business outcomes. A Data Warehouse platform should natively support data streaming capability. As the volume for data can be extremely high in case of streaming, the cost of storage needs to be supportive. The data warehouse platform should provide either SQL or API-based mechanism to query/analyze the streaming data to generate business value out of the streams.

2.4.6 Enterprise class security

The Cloud provides industrial strength and security. Hardened security exists across all layers of infrastructure and application. It provides a variety of tools and mechanism for –

- Authentication
- Authorization
- Data Encryption
- Auditing
- Perimeter/Network/Firewall
- Key Management
- Configuration and others

Also, most of the security mechanisms that are simplified and standardized, offer better security at low effort.

2.4.7 Simplified Operations

Many on-premise Data Warehouses are like sophisticated engines requiring skilled craftsmen. There are hundreds of levers/parameters, which when tuned together, result in a combinatorial explosion. One requires years of experience and tribal knowledge on what works to manage these systems. Digital requires a simplified platform since enterprises neither have the luxury of time nor money to spend on costly technical skills. Most of the tasks related to data distribution, indexing, query tuning, data parsing, and high availability need to be autonomous. In most cases, the users should load the data into the Data Warehouse and run queries without going through a series of operational tasks to make the query run.

This, combined with schema on read paradigm, should eliminate the costly upfront data modeling activity and significantly simplify Data Warehouse development and operations, while cutting down on the operational cost.

2.4.8 Reliable

One of significant advantages of cloud is that it improves reliability and durability of data without a lot of operational work. Data is replicated across multiple availability zones automatically, and optionally across regions. A combination of autosensing, auto-scaling, and load balancing can ensure high availability, without the concomitant administrative and operations activities.

Data Warehouse 4.0 should leverage these capabilities to build on the strength of cloud to solve some of the key problems related to HA, BCP and DR related to on premise Data Warehouses.

2.4.9 Ambidextrous

It is obvious that large enterprises would, over a period of time, have a mixture of on-premise and multi-cloud infrastructure. There is data which is suited for on-premise, and data which naturally belongs to cloud - IoT, Social, and Mobile. At the same time, enterprises would not like to commit themselves to a lock-in relation with a single cloud provider, as this is akin to replicating the old on-premise model in cloud. Data sensitivities will dictate what data will move to cloud and risk mitigation will result in multiple cloud providers.

Data Warehouse 4.0 must support the ability to connect data across on-premise, cloud, and multi-cloud setup. At best, the querying should be transparent to users delivering low latency performance, at minimum an effective data sharing mechanism needs to be provided to easily share data across hybrid and multi-cloud setup.

3 Logical Architecture for Data Warehouse 4.0

The diagram below shows logical architecture of Data Warehouse 4.0. It combines the capabilities of Data Warehouse and data lake in one solitary architecture. The logical architecture can be instantiated on any of the public platforms.

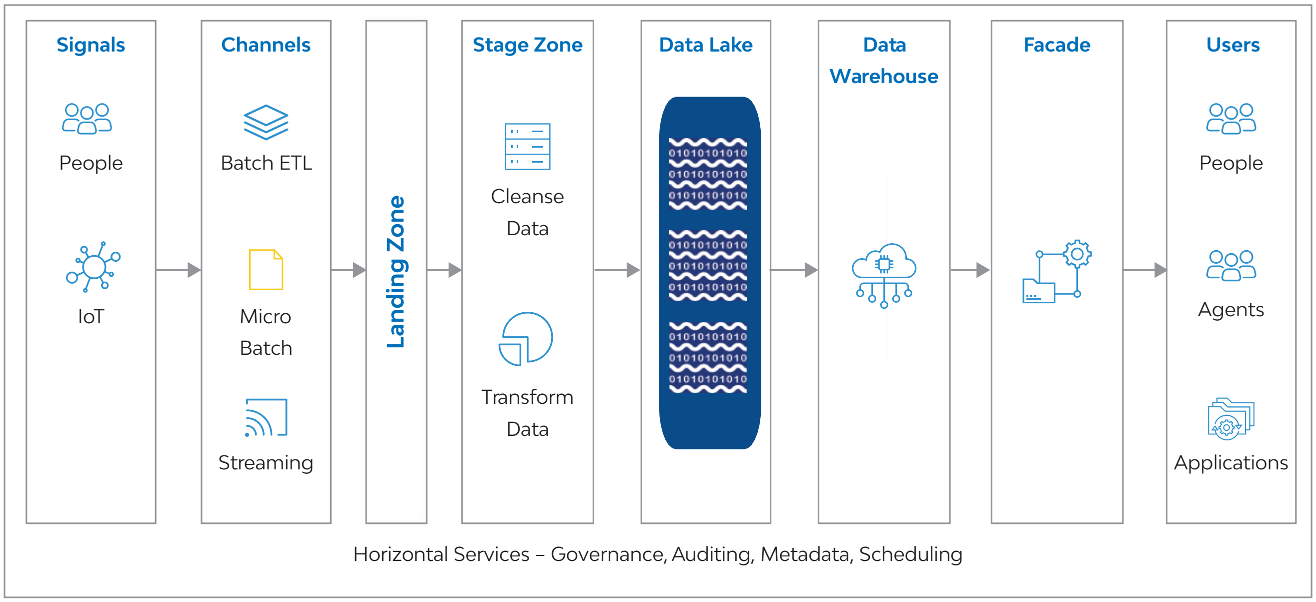

Figure 5: Logical Architecture of Data Warehouse 4.0

Signals: the input signals include structured, semi-structured, and unstructured data sources. Data Warehouse 4.0 needs to support multi-structured data. It also needs to support data streaming from IoT devices and applications. Together, the data sources increase the cognitive bandwidth of the enterprise.

Channels: Fata Warehouse architecture needs to support multiple data ingestion channels. The channel needs to support batch ETL, micro batch ETL, and streaming mode. The channel needs to be built on a distributed platform for delivering scale and performance for big data workloads.

Landing Zone: data from sources needs to be captured with high fidelity and with minimum touch so that sources are not adversely impacted. Business rules should not be applied at this layer to avoid exception/errors. This area provides entry point for further data processing.

Staging Zone: parsing of data and minimal data quality checks can be done in the zone. This is to ensure we are building a governed Data Warehouse and not a data dump. Data Warehouse 4.0 needs to provide certified data quality, so that consumer can leverage data with confidence.

Data Lake: it is the core repository of an enterprise data. It can be organized on the basis of a dominant business theme. The data is preferably stored in an object storage system with folder structures. This minimizes storage costs, provides light weight structure, and helps to implement governance. The data is stored in an application-agnostic pattern. This data is re-purposed based on the requirements of the end user applications. Data Lake layer converts data into an enterprise asset.

Data Warehouse: data is modeled in this layer based on the chosen Data Warehouse architecture. The data can be organized into subject areas and full history can be maintained for the transactions and dimensional data. The Data Warehouse can be built using Inmon, Kimball, or Data Vault methodology and modeling concepts.

Façade: this layer is required to hide the internal complexity and implementation details from the end users and consuming applications. Data must be modeled in a format so that it is easy for consumers to consume. The data in this layer can be instantiated either physically or logically based on end user needs.

Users: the users, who consume the data from the warehouse, may include business users, other applications, and agents. In addition to traditional reports and dashboards, an API layer should be levered to fully exploit the data available in the warehouse.

Horizontal Services: Data Warehouse needs to provide a set of horizontal services to support a data platform with industrial strength. Some of these services include:

1. Exception/Error Handling
2. Slowly Changing Dimension Handler
3. Fact Builder
4. Encryption/Masking
5. Audit and Lineage Tracking
6. Metadata Data Management
7. Log Handling
8. Notification

4 Changing Mindset, Not Just Toolset

Data Warehouse for digital requires a fundamental shift in the mindset and not just toolset. Shifting to a new toolset, keeping the prior archaic mental models will not result in any core business transformation. Practitioners need to escape the gravity of the past to tap into opportunities of present and future. Captured below are some of shifts one needs to make while building a Data Warehouse, which not only uses latest technology, but also reflects today's zeitgeist.

From	To
Data as a by Product	Data as First-Class Citizen of Enterprise
Data for Few	Data for All
Rigidity	Agility & Flexibility
KPI Monitoring	Data Driven Business

4.1 Data as First-Class Citizen in Scientific Enterprise

Traditional Thinking

The industrial era model of an enterprise is that it transforms inputs into outputs, using standardized layers of process, people, and technology. The focus is on standardization and economies of scale. This works if the environment is benign, stable, and the company has figured out the profit formula.

New Thinking

Digital era is the very antithesis of the above model. The environment is volatile, uncertain, complex, and causally ambiguous. It is not possible to create business models or cash cows which endure for years or decades. Digital business is more like running many experiments simultaneously in the market, getting feedback on what is working, scaling the ones which work, and pivot/eliminate the rest. This is more akin to a scientific research group running number of hypothesis, not a top down 10-year strategic plan. What this necessitates is the ability to instrument environment, collect signals rapidly, transform signals into insights instantly, and fully leverage insights to harvest business outcome.

4.2 From Data Oligarchies to Data Democratization

Traditional Thinking

Right Data to Right People at Right Time. This was one of the cardinal tenants of traditional Data Warehouse. The Data Warehouse was supposed to select the right data from multiple sources, clean and conform it, and ensure that this data is available for reporting at right time. This thinking essentially eliminated all the data that was not deemed useful. Right time basically meant before the opening of business next day. Right people meant a handful of executives.

New Thinking

All The Data to All The People All The Time- Data for few is flawed thinking in the digital economy, where data is a factor of production not a by-product. We need to harness all the data, all the time and enable all the people to leverage data to make decisions. Even more importantly, we need to monetize data directly, it is not good just to generate reports and dashboards to make decisions. Data must result in first order effects of business namely revenue and profit. Right time means real-time. Businesses need situational awareness on what is happening at this moment, what will happen next, and what they should be doing about it. Data Warehouse 4.0 needs to enable that.

4.3 From Rigidity to Strategic Flexibility

Traditional Thinking

Traditionally, Data Warehouses required months, if not years to design and develop. The aim was to build a perfect solution, which will deliver 100% on the requirements. However, in many cases, it resulted in Data Warehouses which arrived too late to be relevant or which never got fully built as requirements kept changing. The huge upfront effort required to harmonize the business rules, define the data model, and develop the data flows consumed huge amount of effort/money without enough concomitant benefits to show.

New Thinking

In a digital era, the volume of data is growing exponentially. It is neither viable nor possible to build perfectly stable structures into which data will be housed. We need to house the data in flexible structures, which can be distilled on demand. We still need to apply business rules, build the data model, and data flows, however, this need not be done upfront and may not be necessary for all the data. The data architecture for holding the loosely structured data as an enterprise asset can be called Data Lake. This data can be marshaled into more tightly structured data models to deliver superior application performance. In Data Warehouse 4.0, we embrace Data Lake as a legitimate component of the Data Warehouse. Data Lake is subsumed in the Data Warehouse.

4.4From KPI to Data Driven Business

Traditional Thinking

The traditional goal of Data Warehouse was to leverage data to measure and monitor Key Performance Indicators (KPIs) of the business. These KPIs would then guide the decision cycle in terms of taking corrective actions. Data was not considered critical and integral to every aspect of business. It was primarily used for top-down decision making on how to correct business processes.

New Thinking

All business is data business. Data is integral part of every business activity. It is necessary to exploit the full potential of data by becoming data-driven, which is a cultural change. It is about re-imagining how data can deliver a competitive advantage in every aspect of the business, whether it is marketing, sales, inventory management, or customer experience. It is about leveraging data at every touch point of business – customers, suppliers, partners, etc. for deriving value.

5 The Hadoop Distraction

Hadoop has been touted as the latest silver bullet, which would replace the Data Warehouse. Hadoop, at present, does not fully provide the architecture capabilities, that which can support Data Warehouse 4.0. Hadoop can be part of the solution, but not the full solution. A combination of distributed storage engine - HDFS and SQL engines like Hive and Spark SQL were supposed to replace decades old Data Warehouses. However, it never became a reality.

Many Data Warehouse workloads shifted from traditional platforms like MPP systems to Hadoop, which delivered poor end user experience, were notoriously difficult to implement, and hiring competent skills was perennial challenge. Though few use cases like active archive, and batch ETL offloading served customers well, Hadoop fell well short of customer expectations in terms of concurrency, fast query performance, SCD handling, and other critical Data Warehouse features. This unfortunately has created a situation where Data Warehouse practitioners have begun to see new technologies with a sense of suspicion.

Figure 6: Hadoop and SQL distraction

6 Framework for Success with Data Warehouse 4.0

In a digital world, technology and business are intricately interwoven. One can draw a straight line from a data model to a business model. Digital has completely changed the way data is leveraged by enterprises. Data today defines the customer experience, products, solutions, and ultimately, profit margins.

The success framework proposed here captures some of the strategic elements needed. Why comes before how in any transformation initiative. Use the framework as a guide rope for your own journey.

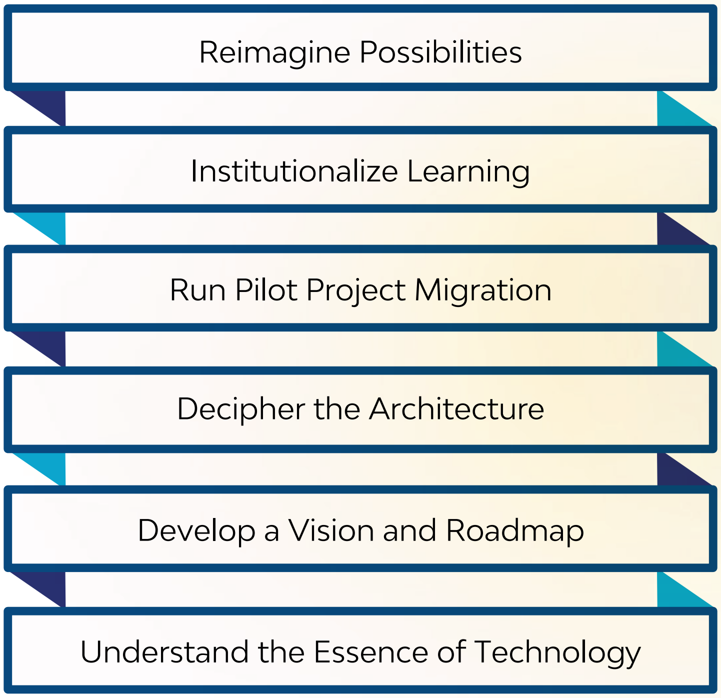

Figure 7: DataWarehouse Success Framework

6.1 Understand the Essence of Technology

The key to implement Data Warehouse 4.0 is to develop a shared mental model within the organization of what it means to the business and technology practioners. One of the fundamental difficulties and dangers when the paradigm shifts is one needs to breakup the existing mental model and reconfigure the pieces, synthesize to appreciate the possibilities, and understand risks of the new paradigm.

For example, developers with expertise in a structured programming learn to use the syntax of the object-oriented programming embracing. This effectively means they fail to leverage new capabilities offered by the new paradigm. A similar predicament awaits when developers move from object-oriented to functional programming paradigm.

6.2 Develop a Common Vision and Roadmap

Business and IT need a shared vision on how to fully exploit the capability of new technologies. They can span a continuum of strategic, tactical, or operational use cases. There are many pathways to value available, including but not limited to:

1. Re-platform to cloud to reduce cost
2. Develop new use cases which provide predictive, prescriptive, and causal insights
3. Democratize data by sharing it within and outside the enterprise
4. Improve customer experience by leveraging cloud elasticity
5. Archive data at a lower cost

Without a shared vision, the implementations can devolve into an exercise of technical puzzle solving without any significant business impact.

6.3 Decipher the Architecture, Don’t be Misled by Marketecture

One of the cardinal truths of the technology business is that one needs to look at the underlying architecture of a technology, not just the marketing of it. As stated earlier, many enterprises wrongly conceived a strategy of moving Data Warehouse to Hadoop without understanding the underlying technology. A file system which does not support updates, lack of indexes, materialized views, poor SQL support, java-based execution engine, high latency scheduling, poor caching, and other were obvious flaws, which were overlooked in the hype. Ultimately, the reality of architecture grounded many projects and careers.

6.4 Run Pilot Project to Learn the Technology Capabilities

It is necessary to run a true production grade use case to validate the technology. Pilots with sample data, simple reports, or few users don’t truly help to understand the capabilities and limitations of the technology. Pick a use case which has sufficient complexity and business value to validate new tools and technology.

6.5 Institutionalize Learning

A stable core of understanding is needed, especially when disruptions are happening all around. Learning what works and does not work is very essential in the era of emerging technologies. This can be a significant competitive advantage is era of high people turnover. Institutionalize knowledge into frameworks, best practices, and ultimately into automation. This will provide a base upon which constant innovation can be done.

6.6 Reimagine Possibilities

Do not implement just plain old use cases; reimagine the possibilities with Data Warehouse 4.0. It is necessary to run multiple experiments as the upside is huge and downside limited. Cloud offers a cost-effective and agile platform for experimentation. Aim for 10x improvements not a 10% gain.

Transition to Data Warehouse 4.0 should include cultural and business transformation aspects. Data Warehouse 4.0 should enable harvesting of enterprise data, exploiting full set of value embedded in data.

7 Summary

The global economy has transitioned from an industrial to a digital economy. Data Warehouses form a core component of the analytical infrastructure. They can provide significant competitive advantage if done well. Emergence of cloud and big data significantly expands the realm of possibilities for building a modern Data Warehouse. Cloud simplifies and makes Data Warehouse affordable. Big data adds multi-structured and streaming dimension.

However, new age data warehousing will not only be new, but will be fundamentally different from on-premise setup. Businesses need to re-imagine the possibilities of Data Warehouse data for business transformation. Data Warehouse 4.0 can help in delivering better customer experiences, operational efficiencies, and new capabilities. This can be accomplished keeping costs low and complexity in check, with imagination being the only limitation.