YOURSTORY

AIQRATE
A Bespoke AI Advisory & Consulting Firm

DATA ENGINEERING 4.0

Evolution, Emergence & Possibilities in the next decade

FEBRUARY 2020

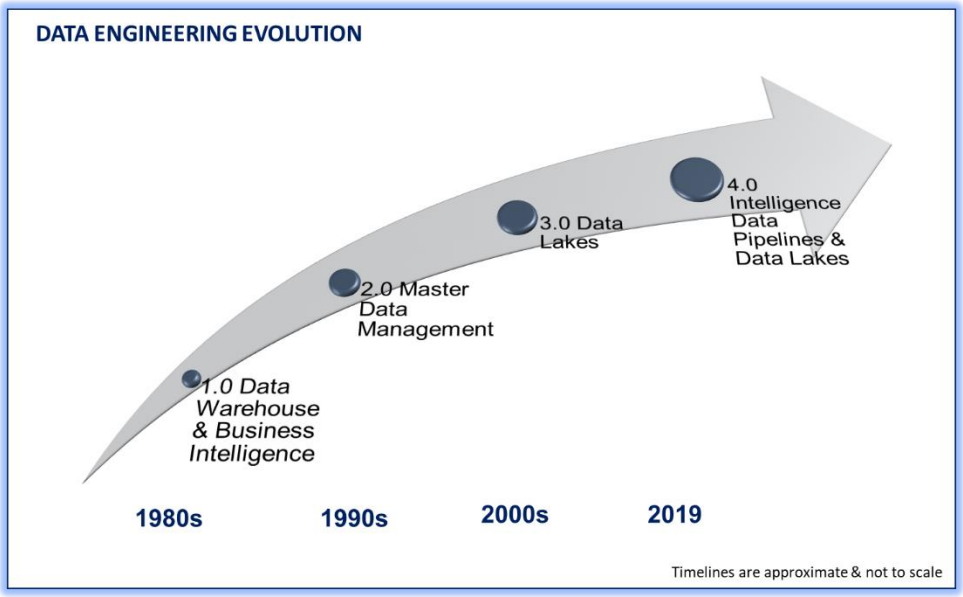
TABLE OF CONTENTS

DATA ENGINEERING 4.0

There have been significant exponential technology advancements in the past decade. Data Engineering is the most topical of them. Burgeoning data volume, data trajectory, data insertion points, data structure, demand for faster processing, new security & compliance needs and so on….Data Engineering is aligning itself to these advancements by far has gone unnoticed by many. Data Engineering has reinvented itself from being passive network and server management to a more active business friendly function. Though it has established an overlap with AI, its ability to enable jumpstart the AI function is not well recognized.

> As more unprecedented advancements are slated to occur in the next decade, there is a need for additional focus on Data Engineering. Unless there is attention to enhance Data Engineering, there will not be successful applications of technological advancements and these advancements' potential may not be leveraged efficiently.

If we try to look back at the trajectory of Data Engineering evolution, it is very clear that, Data Engineering has evolved always to its external ecosystem demands and advancements.



1.0: Data Warehousing & Business Intelligence

- Processing of historical data
- Purpose of reporting
- Unidirectional data
- Typically for business silos
- Relevant data only
- Usual consumer data

2.0: Master Data Management

- Processing of current data
- Purpose of having updated data
- Focus on entity
- Enterprise level data
- Bidirectional data of consumer and provider

3.0: Data Lakes

- Structured/semi-structured/unstructured data
- Schema-on-read
- Data store with no known purpose

- Data in its raw form
- Stream processing

Data Engineering 4.0 (DE 4.0): Intelligent Data Pipelines & Data Lakes

- **Numerous data generators:** With industries emerging out of silos and starting to interact with each other and with use cases cutting across; there will be interconnection of millions of devices, systems, applications and entities. All these will be generating data which needs to be managed by Data Engineering 4.0.
- **Zero lag turn-around:** With life-critical use cases becoming digital, the expectation is of processing of response with no latency.
- **Significant increase in data volume:** As per IDC forecast, there will be more than 41 billion connected Internet of Things devices generating zettabytes of data by 2025. As more and more IOT devices gets connected, the amount of data being generated will keep mounting.
- **Data generation faster than storage speed:** 5G is being pitched as 100x faster than 4G and billions of IOT devices connected to this network will generate large volumes of data in quick succession. However, the storage speed is not expected to increase in a similar proportion; making it imperative that the industry look for an alternative to address the situation.
- **Intelligent & distributed data lakes:** As industries/domains are going to shed their silos, data lakes are going to get interconnected. The efficiency and resilience of this interconnection would be ensured by data lakes becoming distributed/semantic and powered by AI for its management.
- **Mid-pipeline data for processing:** For time-critical use cases, any minute time lag in response will be unaffordable thus making even waiting for data to reach the edge cloud unaffordable. Mid-pipeline data processing would address these scenarios and, in some cases, even AI enabled devices will play their part.
- **Stringent legislations:** As the borders between industries and domains are blurring out, there is a need of stringent regulations and legislations to ensure ethical play across.

Emergence of DE 4.0: A Generation beyond DE3.0

- **New sources of data:** DE 4.0 is going to receive an increased quantum of new sources of data - Spatiotemporal data which encapsulates location and time of an event; Machine data which will be generated by special IOTs including machines in factories and connected vehicles; Genomics data carrying DNA information etc.
- **Polydirectional data:** Data in DE 4.0 will be flowing in multi-directions as entities will start to behave both as data generators and consumers. This will pose challenging questions on schema-on-read constructs widely popular in DE 3.0.
- **Intelligent data storage:** In DE 3.0 it was about store anything and everything into a data lake with no current idea on usage. With oceans of data going to be generated in DE 4.0 era, it will be impossible to continue to store everything. There will be intelligent algorithms which will be deciding what to store and what to ignore.
- **Multi-industry data:** DE 3.0 was siloed to an industry if not a domain. With these boundaries blurring out in DE 4.0, it will be about dealing with data that belongs to multiple industries & domains.
- **Stream processing 2.0:** DE 3.0 brought-in the stream processing mechanism. DE 4.0 is going to push stream processing beyond its potential specifically to manage time-critical data and Stream processing 2.0 will emerge here.

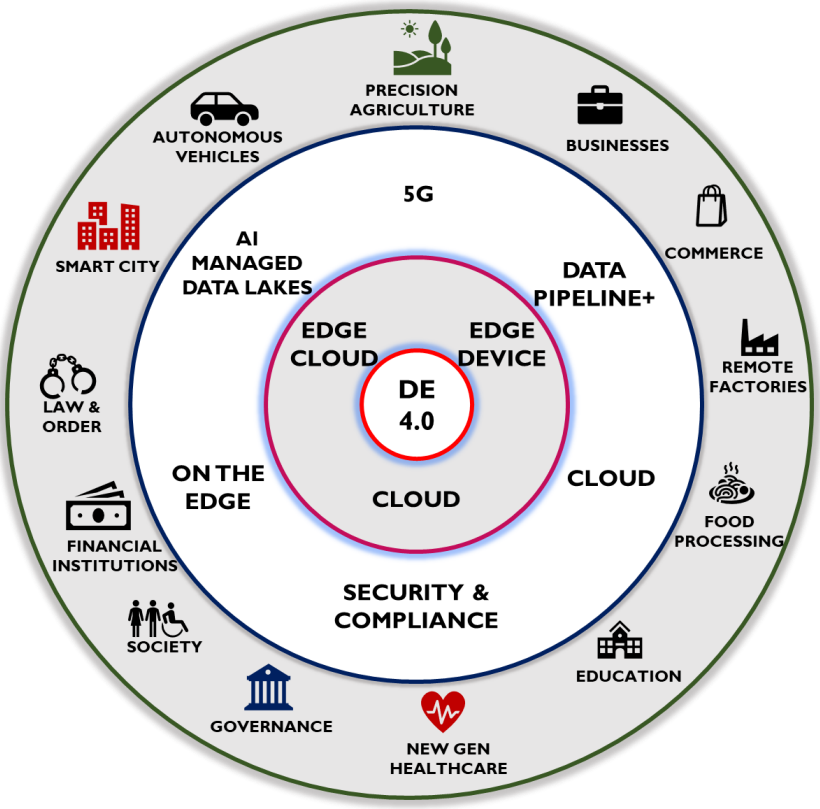
In the evolution of data engineering, the significance and complexity of data is consistently increasing. This is aided by multiple factors like new data generation sources, advancements in network technologies, affordability of storage, strategic insights for decision making, new business scenarios, lifestyle, and most importantly the unprecedented increase in the volume of connected people and devices.

Several disruptions occurred in the previous decade which eliminated multiple industries, brought in new ones and changed the way of life. The pace of industry's evolution is

accelerating multi-fold. The next decade is predicted to be one of massive transformation of a scale never envisioned before. The highly anticipated 5G rolls-out over the next decade will make many use cases possible which were just a figment of our imagination before.

DRIVERS OF DATA ENGINEERING 4.0

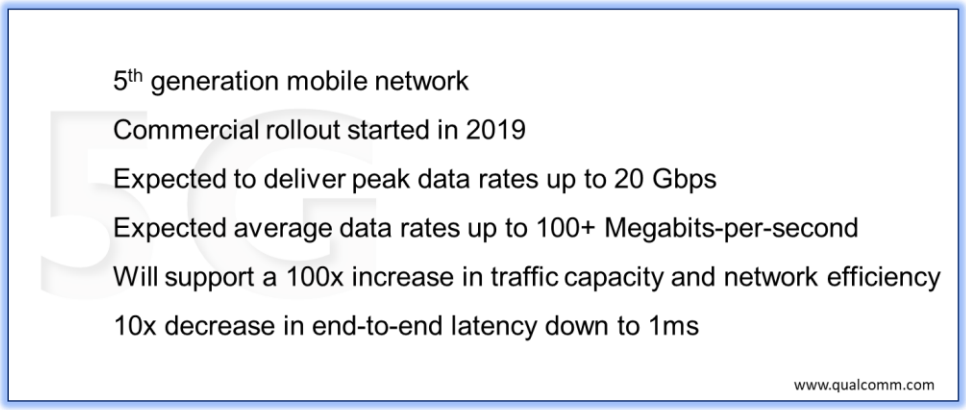
There are multiple technology and business advancements lining up to push data engineering against the wall in the coming decade. Data engineering may even wilt under these severe expectations, challenges and complexities. But, as it has done successfully before, data engineering is well equipped to play an integral part in the next age of exponential technology evolution. Some of the technology advancements are slated to have unprecedented impact on humanity, businesses and administration. Some of these are discussed here.



5G – NEXT GENERATION MOBILE BROADBAND

5G is going to bring an avalanche of disruptions with its potential levels of performance and ability to enable new user experiences. It will redefine a range of industries and even enable the creation of brand-new industries. 5G is slated as a transformative technology in the lines of electricity & automobile which changed the humanity & their world of living.

5G rollout is expected to continue through this decade both as a technology implementation and ongoing usage. A new ecosystem needs to be shaped up to full leverage the potential of 5G offerings. As it is going to redefine the data networks, data engineering world has been studying the 5G demands carefully.



In Data Engineering, it is not just the pipeline that needs to undergo overhaul, but the overall architecture must be revisited. It is anyway clear by now that, large part of network world will get disrupted by 5G across its ecosystem of devices, machines, networks, technologies, data centers and people. Although there may not be use cases existing yet which can fully utilize the 5G potential, but they are surely being cooked as we speak and aren't far away from being placed on the table.

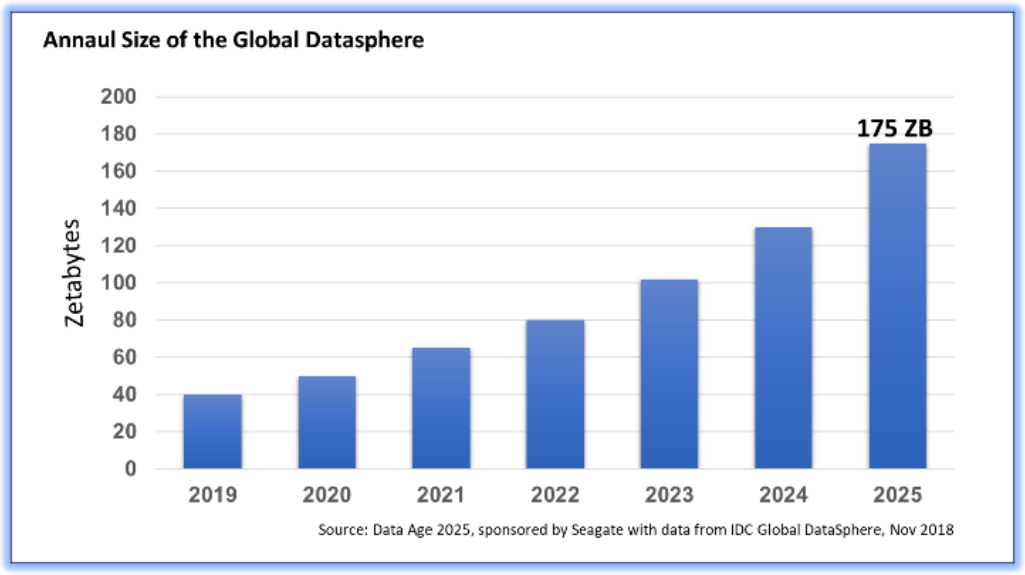
Salient features 5G will bring into Data Engineering:

1. Network speed: >100x speed compared to 4G
2. Near zero latency: Average latency at <10ms
3. More storage requirement: Zettabytes
4. Demand for faster/efficient processing: Real-time
5. Enhanced reliability: Near zero packets dropping rate
6. Next generation Edge Computing
7. Distributed AI: distributed processing by autonomous agents mainly to overcome large volume of data to be processed

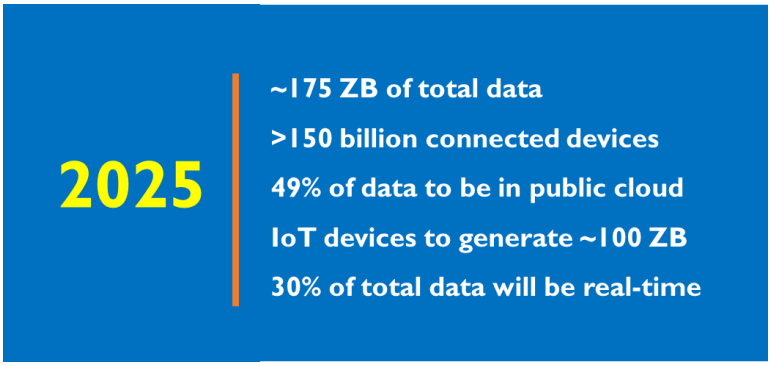
DATA PIPELINE+

IDC predicts that the Global Datasphere will grow from 33 Zettabytes in 2018 to 175 Zettabytes by 2025 and foresees 49% of world's stored data to reside in public cloud environments.

By 2025, 75% of world's population will be connected and each will do average one interaction every 18 seconds. IoT devices themselves are expected to generate near to 100 Zettabytes of data by 2025.



To get to 175 ZB in 2025 (and beyond), pipelines will need to accommodate data that continues to evolve and transform in terms of amount, velocity, purpose, trajectory, and format, to name just a few attributes.



IDC estimates more than 150 billion connected devices by 2025 and most of them generating real-time data. This would result in approximately 30% of total data by 2025 will be real-time data. Data engineering need to be in its efficient best to plumb the right type of data pipelines.

Data Pipelines have evolved from interacting with flat files, data bases, data warehouse, and data lakes. It has evolved from being a batch jobs ran at steady intervals to critical real-time streaming. With exponential explosion data in the horizon there is a need for data pipelines to become smarter, faster and larger. In addition, the data pipelines must be efficiently polydirectional covering thousands of devices, thousands of applications, hundreds of data lakes, multiple processing centers, and multiple business/agencies/entities. Hence it is being called as Data Pipeline+ denoting its next generation. Let's look at some of the new aspects data pipelines must deal with in the coming years.

Data Pipeline+ will be highly complex as they must be capable of auto-scaling, sharding and partition-tolerance with no human intervention. They should become self-healing and auto-configurable by being format agnostic. They should have intelligence to requeue the

> The newer world of Smart cities and remote factories are going to break the shackles of siloed applications and adopt more of a horizontal approach. Use cases will cut across multiple industries not just the applications and systems. Common protocols of sharing data, acting on received data, collaborative access to disparate systems and integrated business cases will be executing such use cases.

events accurately in case of misses and duplication. Most importantly, it should have the smartness to halt when at fault and scream for human intervention.

As the data pipeline is getting increasingly complex, it is now taking shape of a new business of **Data-Pipe-as-a-service**. Alooma offers such services through its Enterprise Data Pipeline, a data in motion platform. It comes with features to integrate with multiple business systems, adopt to ad hoc data & schema changes, ensure high availability, monitor in real-time and ensures highest level security with certifications by SOC_2, HIPAA and EU-US Privacy Shield.

AI MANAGED DATA LAKES

All data cannot be stored as the pace of data generation will outpace the storage speed. While dealing with so much data, it may not be feasible to choose data to be storage. So, chances of data lake turning into a swamp is very high. At this point, many organizations have this problem on ground and just have no clue on how to turn this swamp into efficient data.

Machine learning algorithms will keep a watch on the data getting stored to identify the data which has the potential to turn into swamp. Identifying potential swamp early enables data engineering team to work with respective business to act on the data in question. This will have not only impact on cost reduction, but it will also enhance the efficiency of data processing.

Though the data lake took center stage and overshadowed the data warehouse due to main differentiators like ability to store data in raw data, different types of data which data warehouses cannot handle and schema-on-read mechanism. But, has the data lake solved all the problems and replaced the data warehouse completely? Not really. Providing access to critical data has become difficult. As highly critical decisions are at

> Scanning through continuous streams of data to identify potential swamp data is humanly impossible and Machine learning algorithms will keep a check on this efficiently.

stake, there is a need for rethinking here. This again brightens the emergence of AI powered hybrid metadata-based data lakes that can pre-identify & label critical data to be served instantly.

ON THE EDGE

In the coming days the number of data generators would increase multi-fold and would generate an unprecedented quantum of data. Some of this data are critical and are required to be processed and acted upon instantly. Thus, there is no time for such a large amount of data to be transported to the cloud and back from a device. This bottleneck is giving rise to the need for Edge computing. Edge computing can be either on an Edge device or on an Edge cloud which is very near the data generator.

At its first stage, data computing used to be done on the edge before it was transferred to the cloud for further processing by AI and Machine Learning. Soon, even the edge would have full-fledged AI/ML modules to work on time-critical-data for instant responses. Coming are the days where Edge Device, Edge Cloud & Cloud would be powered by AI and will be working in tandem to offer huge room for several disruptions.

Intel's introduction of the Edge computing chip signals the advancements on the chart for coming days. Linux through its LF Edge project has already started to work on standardizing the open source around edge computing.

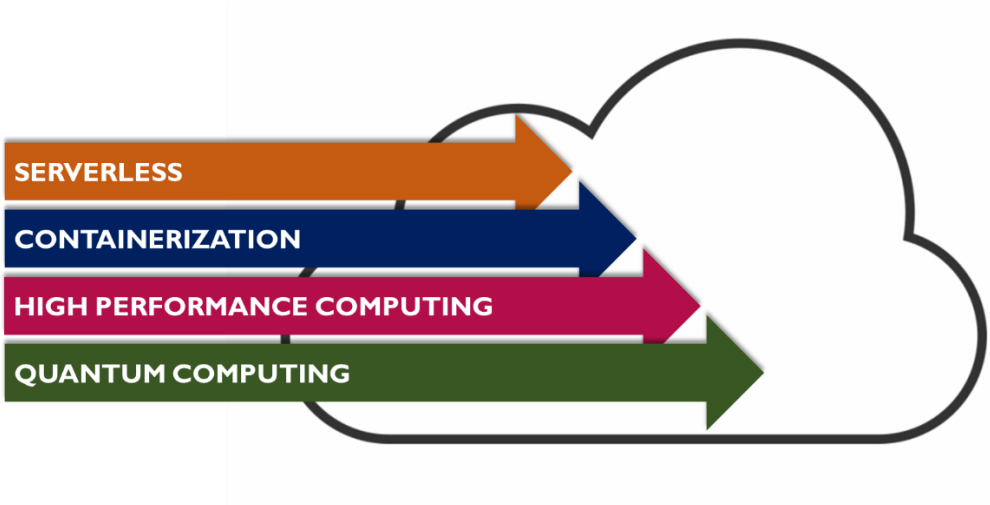
5G networks will fuel the advantage of Edge computing with its sheer speed, less latency and mega volume. 5G will not be leveraged to its potential without embracing edge computing.

> The challenge for data engineering would be to innovate efficient ways to setup intelligent clouds, intelligent edges and intelligent devices to work in tandem via newer ways of configuration and new connected applications.

CLOUD

With all the above-mentioned advancements, is Cloud computing becoming a trend of the past? Not really. There are huge expectations from cloud providers to increase their processing ability for the new world. So, cloud providers are reinventing to take the next leap.

Serverless or Function-as-a-Service is expected to develop in leaps and bounce in the coming years. Many organizations are already building the strategic approach required for this careful transition. Recently AWS unveiled Firecracker, a new open source virtualization technology which enables to pack thousands of micro-virtual machines on the same machine. Workloads of different customers can run on the same machine without any concern on security or efficiency. Firecracker also provides metadata service which can be configured using Firecracker API.



Containerization will continue grow in the coming years. As the trend of multi-cloud is taking added pace with portability between cloud providers, containerization is the best bet. Kubernetes has already captured the leading place in the container world with its pathbreaking features and will continue to bring out more advancements for new age computing.

More and more discussions and inventions are happening on Quantum Computing, High Performance Computing and Exponential technologies which will continue to leverage cloud computing forcing it to mature equally. So, cloud will continue to be a busy zone with frequent technological advancements in the coming years pushing all connected entities to change, improve and leverage its overall potential.

GOVERNANCE – SECURITY & COMPLIANCE

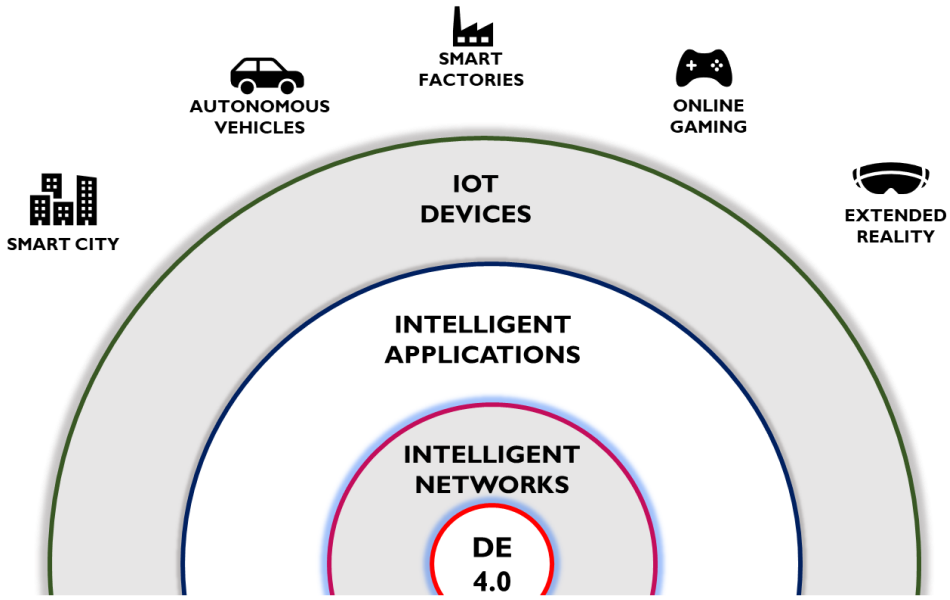
GDPR (General Data Protection Regulation) introduced new rights for individuals such as Right to be Forgotten and Right to Portability. Technically these kind of changes in

legislation puts data engineering teams under severe stress in implementing a suitable solution to be compliant. In the process it should ensure no impact on current applications, data, and pipelines. GDPR continues to be a landmark privacy law in the digital age.

With more and more edge devices going to capturing data in the coming years, there will be some of it falling into the privacy zone. Legislation will accordingly make changes to address these. These changes need to be swift not only for legislation compliance but to mitigate probable cybercrimes, data breaches and mishandling of sensitive data. Unless the data pipeline is made fool-proof, the impact on the businesses, agencies and people can be highly damaging.

INDUSTRY SEGMENTS APPLICATIONS

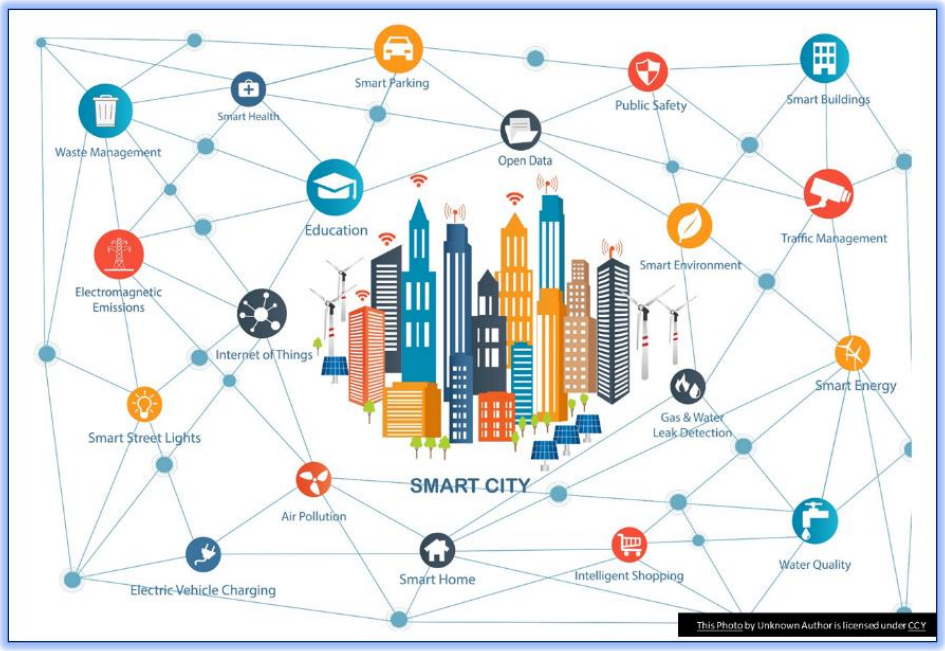
Several traditional industries will be disrupted or reset in the coming years. Each of these industries will have interconnections to serve pan-industry use cases. The change within industries will be to leverage the technology advantages to push the business boundaries. The four dimensions of technology advantages are IOT, Intelligent Applications, Intelligent Network and Data Engineering 4.0.



Some of these industries that are going to be growing leaps and bounds on the four dimensions of technology advantages in this decade:

Smart Cities

More than 50% of world's population lives in cities and growing consistently. Cities are expected to have 2.5 billion more people by 2050. Such popularity comes with challenges across infrastructure which directly influences the quality of living in cities.



Smart cities are going to be equipped with digital intelligence through the deployment and interconnection of smart technologies across city. Anything and everything are expected to get enabled with cognitive abilities to monitor, guide, report, act, deliver and perform tasks to enhance city living. Interconnected traffic signals, next generation health services, intelligent crime monitoring services, autonomous vehicles, automated public utilities, sophisticated emergency response services, active environment management, and many more. IoT is the biggest enabler of Smart cities integrated by new age intelligent data exchanges and processing installations. Here we are talking about the potential of generating oceans of data, most of these are critical real-time data to be integrated and analyzed in real-time and served immediately for decision making. So, in a nutshell, smart cities are going to be equipped with 3 layers of intelligent technologies, IoT sensors &

devices, super-fast & super-efficient communication networks and ultra-powerful processing centers & applications.

Autonomous Vehicles

Autonomous vehicles are already running on roads in test zones. These vehicles are equipped with technologies which keep processing data to take near real-time decisions to manage the vehicle. Though the amount of data received, processed and sent depends on number of mounted devices, at a minimum, an average test car is estimated to generate as much data as 7-8 thousand internet users in a day. Data generated/consumed by autonomous vehicles will increase exponentially when seen holistically along with its ecosystem of traffic signals, intelligent lanes, interaction with other vehicles, connection to authorities, insurance, etc. The exponential amount of data, types of data, crisscross data/insights exchanges, need of near-real-time response, etc. can be imagined understanding the complex environment that data engineering will have to handle in this decade.

Smart Factories

AI & 5G have started to disrupt the whole functioning of factories. Autonomous machines, autonomous vehicles, autonomous logistics and wireless backhaul backed by AI, 5G and a new generation computing are enabling significant increase in industry value creation. Remote control of heavy machineries and factory automation with real-time monitoring will reduce workers' exposure to risky & hazardous zones.

Online gaming

Popularity of video gaming has made it a multi-million-dollar industry expected to cross USD 300 billion mark by 2025.There are nearly 2.5 billion gamers around the world. Games have now become highly sophisticated with hi-tech content, presentation, player integration, space-age gaming gadgets, cloud gaming, etc. Continental and inter-continental remote participative tournaments are being organized that seeing huge participation from players and fans. These online games heavily rely on superfast network speed, near zero latency, and highly efficient data pipeline for data exchange between players and servers. With players sitting thousands of kilometers from each

other conveys the type of data engineering that needs to be in place to transport critical data back and forth almost with zero lag.

Extended Reality – XR

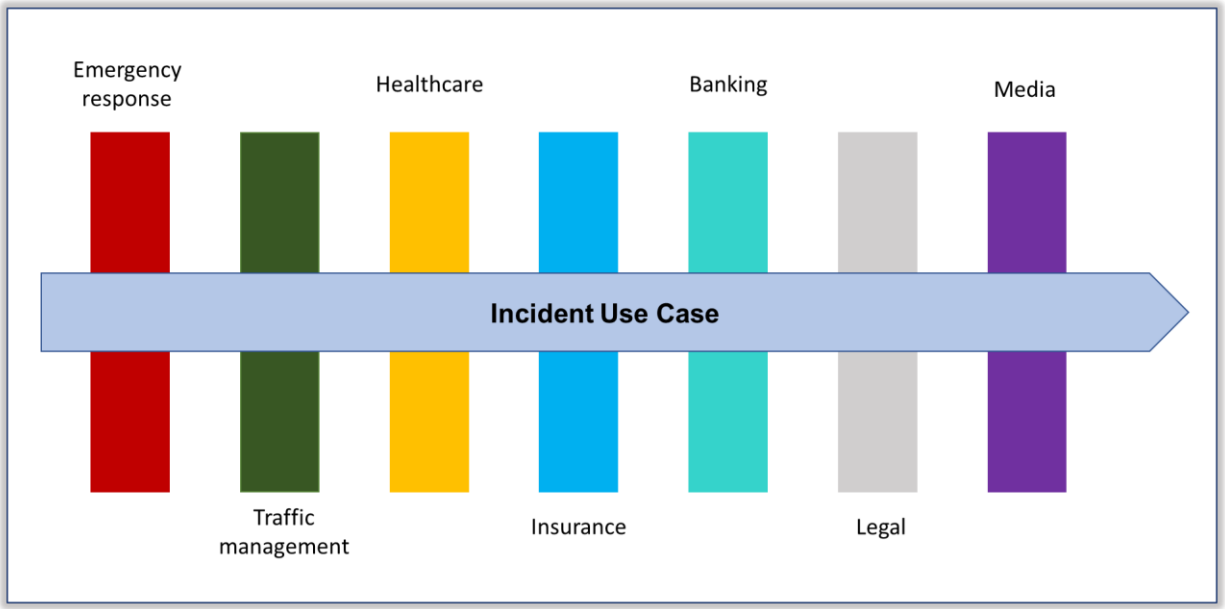
Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR) are the immersive technologies which merge the physical and virtual worlds to be experienced together. All these immersive technologies encapsulated together is known as Extended Reality (XR). XR market is going to be USD 200+ billion market in next 2 years. XR solutions leverages technologies like 3D graphics, AI, machine learning, cognitive security and 5G. Multiple industries like retail, education, entertainment and real estate are already using XR in limited form. Many more industries and use cases are waiting for XR to be at its best with the new age infrastructure to make it a mainstay for offering experiential services.

CROSS INDUSTRY DATA ENGINEERING COLLABORATION

The technology setups working within the industry as silos are going to open their platforms with other industries for collaboration.

Let’s take an unfortunate accident as an instance or subject of use case herein, which traverses across multiple industries with no or minimum human intervention.

1. The cameras or sensors at the streetlights or traffic signals or billboards would trigger a request for an emergency response with location information for all emergency response agencies to arrive at the earliest
2. All traffic signals around the designated radius will intelligently adjust the signal intervals to manage the slow traffic and information billboards will convey detour suggestions to vehicles in the impact radius
3. Healthcare function is informed immediately for swift action for probable emergency medications and other necessary medical accessories
4. The relevant insurance company is reported instantly about the incident with video feed, vehicle and passenger details
5. Relevant bank of the passengers involved are notified on the incident to enable priority processing of transactions.
6. Police and other legal authorities are reported on the incident.
7. Media is informed on the incident for news reporting



Traffic management, emergency response, health care, insurance, banking, legal and social media will have a single use case of an incident cutting across them. So, the data pipelines need to be architected/designed keeping this horizontal arrangement in mind. It is not difficult to imagine the volume and variety of IoT that will be connected in this scenario besides applications. So, the data pipeline dealing with this should be smart enough to be distributing the data to multiple edge computing centers and cloud computing centers and in some cases, it may need to apply intelligence within the pipeline itself to provide a quick response.

Though this kind of cross-industry collaboration looks technologically feasible, there will be lot more effort that goes into setting up suitable business & operating models. Multiple factors need to be investigated while building sustainable cross-industry business & operating models.

Identifying Interface Points: Functions within each industry to investigate their business & operating models to identify the interface points especially the entry and exit points of the cross-traversing use case. Besides tweaking of the functions, there could a need of redesigning or introducing new functions.

Identify Data to be Exchanged: This will be the most challenging task of all. Based on the trajectory of the cross-traversing use case all data points to be dealt with need to be

identified and the necessary processing and the return feed. This will have impact on other internal systems which may not be in direct line with the cross-traversing use cases. All these internal impacts need to be identified and addressed.

Standard Protocols: In any collaboration standardizing protocols is mandatory to establish dependable expectations across. These protocols to clear guidelines on when, what, how, where & why of the collaboration clearly defined.

Security & Compliance: Here respective agencies must work closely with industries to establish well thought and rigorously tested security & compliance guidelines for industries to comply with.

SLAs & OLAs for collaboration: As the guiding principles, there will be list of all the collaboration requirements identified along with their criticality & priority. Based on this, applicable SLAs & OLAs will be implemented, risks identified, mitigation mechanisms built and adhered to. It is important to be sensitive to the fact that, not adhering or inability to deliver to these SLAs will cause severe damages to society.

DATA ENGINEERING IN STARTUPS

DATA ENGINEERING IN STARTUPS

INTRODUCTION

Being insights-driven, data-driven and AI-first are the most important discussions happening in boardrooms when it comes to laying strategy. It is almost clear to the world that, to survive competitively, it is important to have these conversations in the strategy road mapping exercise. Startups are not insulated either. Startups would like to get into AI bandwagon as soon as possible to gain competitive advantage. It is also clear now that the mandatory ingredient to become one of these is data. Large, abundant data. So, the first and foremost task for a Startup will be to embrace Data Engineering practice and

> As the DE 4.0 ecosystem evolves, this scenario would change in the coming days. Startups will be able to plug into the mainstream data engineering 4.0 setup by subscribing to cloud-based pay-as-you-go offerings.

take concerted steps to scale as the business grows. There are misconceptions that, Data Engineering is only for the large organizations and is expensive, complex and time intensive There are 2 major areas of attention a Startup must focus on to build DE foundation for becoming data driven and insights rich.

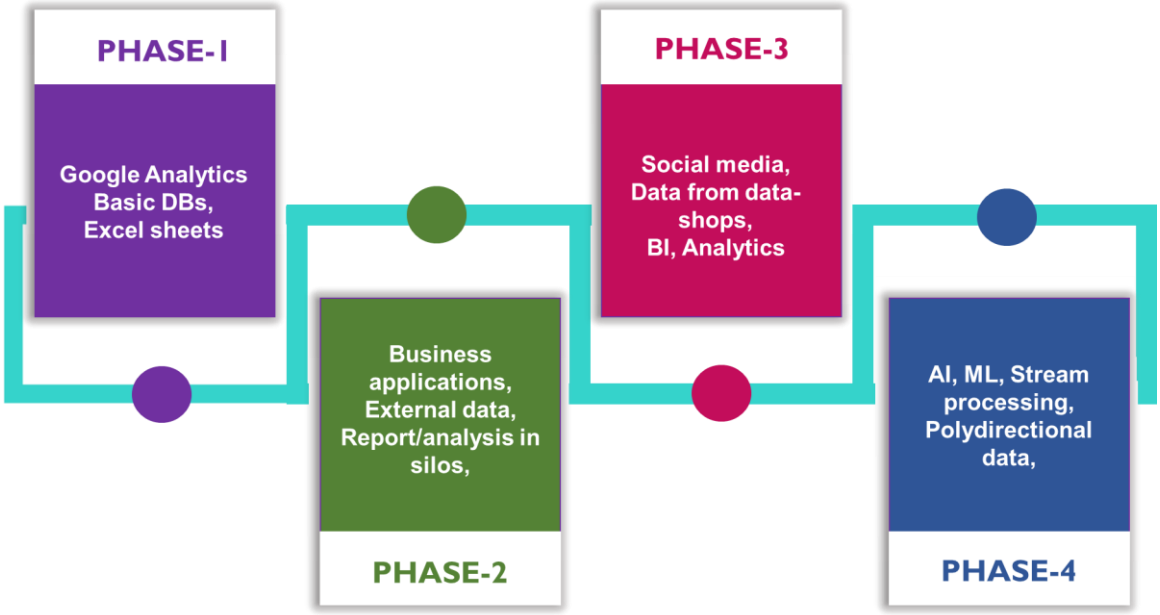
1. Identify Data from non-obvious sources
2. Build & manage robust Data Infrastructure platform

Identify Data from non-obvious sources

At the initial stage, startup will have website or portal generating some data. Other than this, there will be basic internal systems to manage the operations; in most cases, excel spreadsheets. It is important for the startups to retain this data as much as possible in its original form. Apart from the transactional data, startups would like to pull statistics from web pages and digital portals. So, setting up Google Analytics will be the first step to start

receiving this data. Most often, (potential) clients start demanding data or information and this will be a clue for startups to identify the data or data sources to capture it.

As businesses grow in a few months, systems will be used more, and web pages & portals will see more visitors. Businesses now will start looking for more strategic information about product or services most visited, number of visitors, their interest levels, type of visitors, etc. It may also be interested in knowing the performance of recent advertisements released or a campaign unfolded. It will ask for sales data either from internal systems or e-commerce aggregators.



Earlier due to lack of future visibility, these data silos would evolve isolated from each other. So, integrating data generated as well as managing the quality of the data was a herculean task. But, with all these best practices and learnings in place, these silos can be designed in advance to have commonness and inter-operability.

Startups are now realizing the fact that the information being extracted is not helping them in collective or overall decision making or strategy. So, they will start demanding for going over the fence to seek the information from other silos and try relating them together. By now, they would have started to accumulate large amount of data and stored separately.

Here comes the thought for having one place store of all these data. At the same time, awareness gathers on so much useful data being available in the external ecosystem that can spruce up data-driven decisions. Startups will go out and look for these data sources on social media channels and data on shelves for purchase. On more realization by now is about different types of data including semi-structured & unstructured data. Businesses are now keen to access data for analysis; it is impacting the production systems. It is time now for the data engineering team to evolve into the next phase of setting up advanced data infrastructure.

Build & Manage Robust Data Infrastructure

A data lake is a judgmental decision to be taken by data engineering based on its data landscape. If the emphasis is only on the structured data and then later evolving into semi-structured & unstructured data, then decide to deploy a Data warehouse. At this stage, it will be to look for cheaper, simple and scalable technology options for obvious spending related reasons. Same factors by now would have brought the cloud into the picture. Redshift is typically such a data warehouse solution and there are similar solutions from other vendors too.

Next comes the need of transporting data from all source systems into this data warehouse or data lake. For ages now, ETL has being preferred in here for structured data. Data lake is tweaking this around to ELT now-a-days to eliminate Transforming from impacting the data gathering performance. As micro-services are not so complex now, this is a preferred option for transporting data into storage especially the external data and unstructured data.

By now, Startups would have started to see TBs of data at its disposal for analytics and insights. Implementation of AI & ML will bring in more volume of data, types of data and the need for quicker processing. Reason is, data is no more unidirectional, but it is multi-directional. There will be newer use cases taking birth which will not have time to wait for data to reach the storage. Instead they demand the insights straight at the ingestion pipe itself leading to stream processing. This way, as business continues to grow and expands, newer use cases that deal with data keep taking birth to keep Data Engineering busy and to be innovative continuously.

Provisioning processed data, setting up analytics tools, deploying models, configuring visualizations, curating algorithms, etc., are not discussed here as these would be taken care by data engineering selectively. This is where the overlap of tasks and responsibilities starts with AI.

DATA DRIVEN COMPANIES OF INDIA

DATA DRIVEN COMPANIES OF INDIA

India has experienced very successful startups turned unicorns that are technology driven and data-centric with data being the integral part of their business model. These companies as they grew, through futuristic data engineering, kept scaling their data platform with newer technologies and inhouse development. Here are use cases of two high performing companies and their data engineering landscape.

OLA

Ola operates in more than 150 cities across India, UK, Australia & New Zealand and estimates to be serving more than 150 million users. Total rides completed per year is at around 1 billion. Ola connects more than 1.3 million driver partners and connects more than 10,000 routes.

Ola's data platform has horizontally scalable predominantly open source software capable of processing billions of rows of data coming in from different data sources – micro-services, logs, message queues, data bases, etc.

Ola employs more than 250 business analysts querying the platform for different data points across verticals, business lines, product lines and horizontal functions to derive business insights on an hourly and daily basis.

Platform has Hortonworks implementation of Hadoop framework. MySQL and PostgreSQL receive data from micro-services and pumps them into Apache Hive data warehouse hosted on Amazon S3 for scalability. Apache Hue and Presto (scans more than 100 million records every 15 mins for building visualizations) are used for querying and Apache Ranger for democratizing data with integration in LDAP for row-level filtering and column-level masking. Data protection and privacy is addressed through compliance to GDPR and data-privacy norms with role-based data access and data encryption.

Python/R being used by citizen data scientists to build and execute ML algorithms and MicroStrategy is the BI front-end tool for creating in-memory data cubes and visualizations.

Not to mention, Ola has a sophisticated geospatial representation rendered over Google maps across 50,000+ geohashes across the globe. In nutshell, Ola is equipped with highly efficient platform capable of operating at PB scale data.

FLIPKART

Thousands of micro-services power the user experiences at Flipkart. Be it product listing service, delivery date estimation, search service, or pricing engine, each micro-service maintains domain data store across MySQL, HBase, Redis, Elasticsearch and more. Flipkart Data Platform (FDP) enables their analytics, insights, data science and other teams to consume and act. The FDP comprises ingestion system, batch data processing system, real time processing system, visualization and query platforms as 5 main components across ingestion, batch processing, real-time processing, visualization and querying.

Every byte of data generated in Flipkart gets routed into FDP. FDP has the stack with HDFS, Hive, Yarn, Spark, Storm, MapReduce, API services and Kafka message queue supporting the data. FDP currently has more than 800 nodes Hadoop cluster and runs nearly 25,000 compute pipelines on Yarn cluster. Over 3 billion events are being ingested daily as per the estimates available. More than 30TB of data is ingested. HDFS is growing beyond 15 PBs.

Apache Jena with its integrations with Lucene and Elasticsearch is delivering quick and reliable search experience to its users. Jena stores all the metadata to create knowledge graph to support efficient search. The deep searches at average are taking approx. 500mx for 99 percentile which used to take an average of 10-15 seconds before.

SUMMARY

- Data Engineering has created its own niche through the multiple technology advancements in past decade by becoming equally sophisticated if not better.
- Data engineering has evolved with notable advances from Data warehouse, MDM, Data Lake and now to intelligent solutions
- Among the future trends, 5G is expected to lead the disruptions across human living ecosystem.
- Peak & average & uniform data rates, near-zero latency, massive capacity and most importantly Wireless.
- 5G will bring disruption galore across industries with ocean large playfield for innovative use cases.
- Smart cities, Autonomous vehicles, Smart factories, Online gaming, Extended reality are some of the areas to be boosted by 5G.
- Data Engineering's Data Pipelines to evolve into their next generation to meet the new age demands.
- Use cases cutting across industries and connected systems would unleash Data Pipeline+.
- AI managed Data Lakes to evolve into self-monitoring and optimizing Intelligent Data Lakes.

- Edge computing to evolve strongly powered by increasing demand for instant responses. Combination of Edge devices, Edge cloud and Cloud powered by AI will usher in numerous innovations.

- Cloud with its continuing popularity would reinvent itself through advanced offerings around containerization, high performance computing, serverless, etc.

- The legislation wing will be on its toes in tweaking existing regulations and in building new ones to ensure protection of individual rights, security and safety of the ecosystem.

- In the middle of these advancements, there will be new Startup companies who will take smaller steps in Data Engineering to sustain and scale.

- Data Engineering 4.0 will be an overall ultra-upgradation of the complete landscape across ingestion, processing, storage and serve areas.

REFERENCES

- www.yourstory.com
- www.qualcomm.com
- www.alooma.com
- www.aiven.io
- tech.flipkart.com
- www.forbes.com
- cio.economictimes.com
- www.cloudreach.com
- www.techrepublic.com

https://yourstory.com/

Yourstory Bengaluru Office
#259, 6th Cross Rd, 2nd Main Indiranagar, 1st Stage Bengaluru, Karnataka 560038

The YourStory Team brings you stories of entrepreneurs and change-makers, funding analyses, resource pieces and the first glimpse of emerging trends from India's entrepreneurial ecosystem, as well as profiles of great businesses and entrepreneurs from all over the world.

The YourStory team primarily works out of Bengaluru, Karnataka, but has a presence throughout India through its correspondents in English and 12 Indian languages, including Hindi, Kannada, Tamil, Telugu, Malayalam, Marathi, Gujarati, Punjabi, Urdu, Bengali, Oriya and Assamese.

YourStory has published close to 60,000 stories of entrepreneurs and change-makers and helped more than 50,000 entrepreneurs access networking and funding opportunities.

YourStory also works with some of the biggest brands worldwide, enhancing their visibility and engagement with a tech-savvy base of readers through high-quality content.

YourStory's mission has been to tell stories that matter, stories with heart, with drive, and that wouldn't be possible without the passion of our team – they are the heart and soul of YourStory. They each have something that drives them to write, build, design, shoot, edit and publish these stories.

AIQRATE Advisory & Consulting
consult@aiqrate.ai
www.aiqrate.ai
Bangalore | Delhi | Hyderabad

AIQRATE, A bespoke global AI advisory and consulting firm. A first in its genre, AIQRATE provides strategic AI advisory services and consulting offerings across multiple business segments to enable clients on their AI powered transformation & innovation journey and accentuate their decision making and business performance.

AIQRATE works closely with Boards, CXOs and Senior leaders advising them on navigating their Analytics to AI journey with the art of possible or making them jumpstart to AI culture with AI@scale approach followed by consulting them on embedding AI as core to business strategy within business functions and augmenting the decision-making process with AI. We have proven bespoke AI advisory services to enable CXO's and Senior Leaders to curate & design building blocks of AI strategy, embed AI@scale interventions and create AI powered organizations.

AIQRATE's path breaking 50+ AI consulting frameworks, assessments, primers, toolkits and playbooks enable Indian & global enterprises, GCCs, Startups, SMBs, VC/PE firms, and Academic Institutions enhance business performance and accelerate decision making.

Visit **www.aiqrate.ai** to experience our AI advisory services & consulting offerings