Larix Publications

Advanced Journal of Robotics

https://ajrjournal.com/

Vol. 1, Issue 1, 2020

Research Article

Building Autonomous Data for the Enterprise with Data Fulcrum

Niranjana Mangaleswaran*

** Specialist in Business Intelligence, Advanced Analytics & Technology Evangelist*

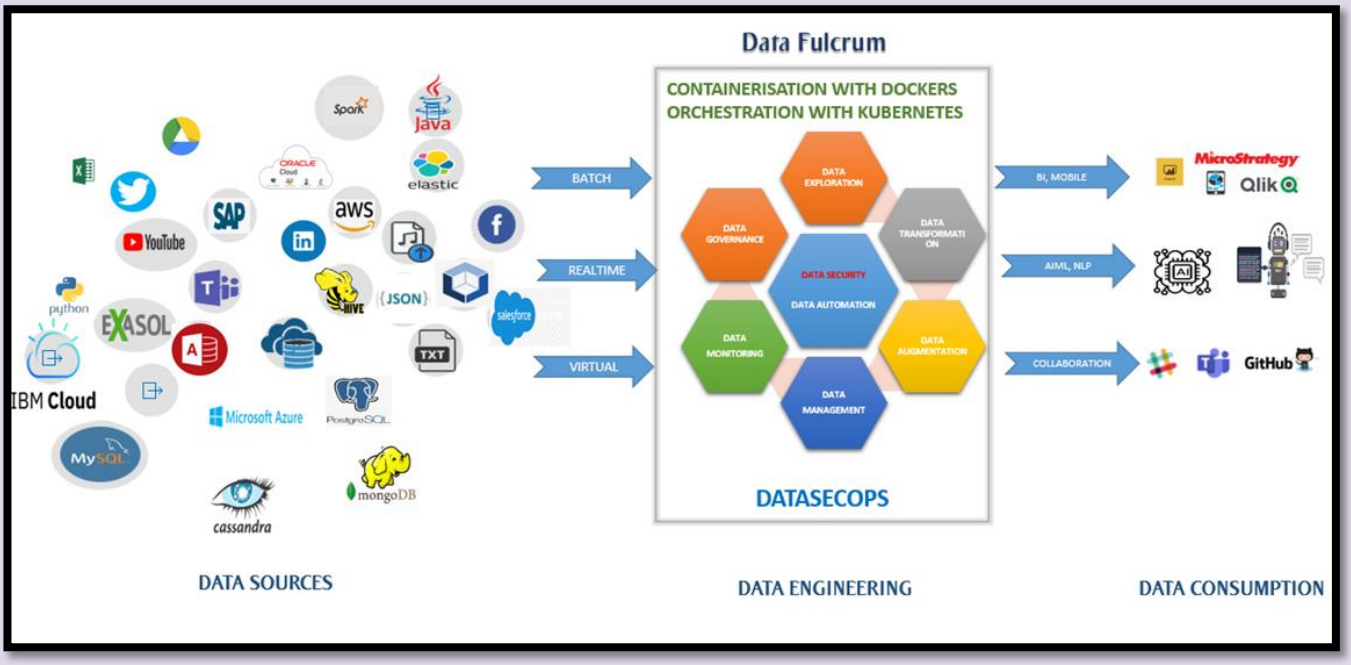
** contactniranjana@gmail.com | +61410892386*

Received on: 06-08-2020; Revised and Accepted on: 31-08-2020

ABSTRACT

M*any years ago, the Data architecture included structured normalized and denormalized data, in a semantic layer, where the data was queried by a BI layer on top of it, providing Descriptive Analytics and Diagnostic Analytics. Descriptive Analytics is a visualization of the past/current state of the business; Diagnostic Analytics is identifying the root cause of the business's past/current state. However, in recent years, the Enterprise's Data Landscape is evolving to become complex, and the different types of Data – structured, semi-structured, and unstructured – is increasing exponentially due to its explosion from various scattered, isolated, unrelated, heterogeneous data sources. This leads to the risk of information overload. Hence, it is essential to narrow down reliable, appropriate, and comprehensive data in order to derive meaningful information. This can be made possible by orchestrating end-to-end data flow of diversified, siloed data across various systems and technology applications. Furthermore, in this competitive world, most organizations are looking for Predictive Analytics and Prescriptive Analytics. Predictive Analytics is the prediction of a business's state based on the past/current factors; Prescriptive Analysis is the required approach to improve 'Revenue' and 'Time to Market'. The process requires agility to perform informed decisions. The data that is collected from disparate sources experiences 7 stages: it is cleansed, processed, transformed, enriched, monitored, governed, and secured. This data is then consumed by BI and AI applications to produce information with actionable insights. Interactive visualizations, NLP in AI, and chatbot assistants are examples of tools that will help in achieving the aforementioned result. Several Organizations use both On-Premise and Cloud Data solutions. So, it is essential to employ a Hybrid Cloud approach; the solution requires scalability, resilience, and effortless deployment. The best approach is to separate Data Storage from Data Processing layers. Further, the Data should be containerized using Dockers and orchestrated using Kubernetes, to provide elevated benefits.*

The main focus of the article is on ***Data Fulcrum.*** *Data Fulcrum is a concept designed to deliver robust and scalable data pipelines for AI/BI. It focuses on the future outlook of data and brings together the right set of tools, processes and people, for better data management (****Refer Figure 1****).*



Keywords: *DataOps, DataSecOps, Containerisation, Data Orchestration, Autonomous data, Data Engineering, Data Fulcrum, Data Hub, Data Pipeline, Data Architecture, Data Integration, Artificial Intelligence, Machine Learning, Data Governance, Data Management, Structured data, Unstructured data, GIT, Unit Testing, Digital Landscape, Business Intelligence, Kafka, Descriptive Analytics, Cognitive Analytics, Prescriptive Analytics, Predictive Analytics, Data Lineage.*

Fig 1: Building Autonomous Data for the Enterprise with Data Fulcrum

INTRODUCTION:

Every Organisation needs Data to be re-usable across all business units and available on demand to achieve razor-blade precision in decision making. This article will focus on the importance of data, how it should be processed so that the governance is not slipped at any part of its journey and lightning speed of data through automation techniques.

1. The Data

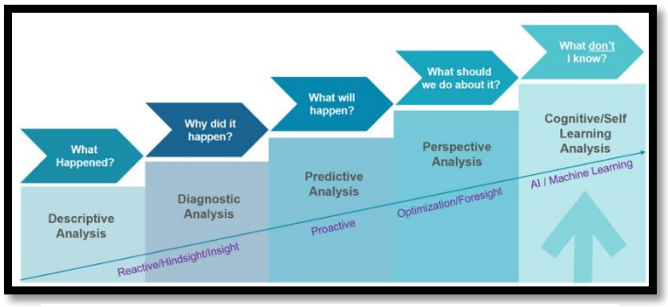


Fig 2: Evolution of data

***Corresponding Author**:

Niranjana Mangaleswaran, Specialist in Business Intelligence, Advanced Analytics & Technology Evangelist
DOI: https://doi.org/10.5281/zenodo.4021931

Over the past few years, we would have witnessed the evolution of the Digital Landscape, data complexities, distinct representations of data, innumerable technologies used in processing and orchestrating data, and disparate Business Intelligence tools used for the creation of visualizations and mobile applications. But the principal aim is to ensure that the data provides meaningful outcomes.

Business users look out for meaningful outcomes, not technical byproducts

The actionable insights that drive value is the only outcome that every business user would actually need. Any other technical outcome is just a byproduct. **Figure 2** explains the evolution of data and analytics ranging from Descriptive to Cognitive Analytics. The GenZ end users are not restricting themselves to Descriptive Analytics and Diagnostic Analytics anymore, but are focused on exploring Prescriptive Analytics. Prescriptive Analytics is widely used since it helps in understanding what went wrong, what could happen next, and what actions need to be taken in order to prevent future failures or even to optimize the systems based on foresight. For example, travel applications use huge datasets that include customers' frequently travelled locations, to prescribe optimized routes and sectors, thereby attracting customers with competitive pricing strategies that leads to an increase in revenue.

Data can go an extra mile…

Some users are more curious and are drawn towards Cognitive Analytics, to measure the unknown - the factors that they do

not know. This is achieved with the right set of algorithms using Machine Learning techniques, NLP (Natural Language Processing) and neural networks. The advantage of Cognitive Analytics is the ability to learn, reason and interact with human beings using natural language. For example, chatbots used in healthcare interacts with customers, processes their queries and offers personalized health recommendations. These chatbots help the healthcare providers with large amounts of data to understand the emotions of patients and performs sentiment analysis to serve them efficiently.

Prevent the disaster!

To make sure we foresee opportunities or disruptions, the Organisation should focus on listing out the Strategic Key Objectives that would drive the business values and decisions and help them with an upsurge of Revenue and Cost Optimization.

So, Data is directly related to Revenue Optimization!

2. TYPES OF DATA

According to Adrian Bridgewater, data is classified into 13 types as listed in **Figure 3.**

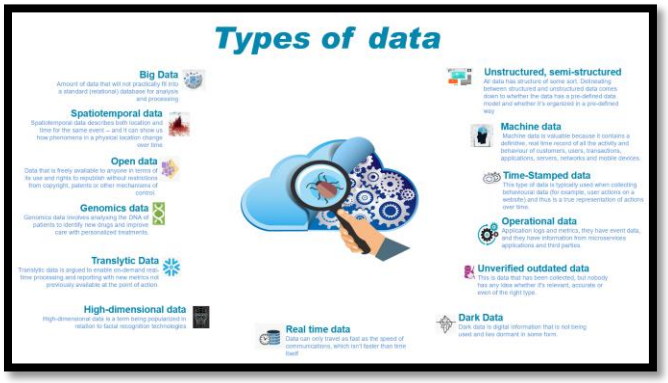


Fig 3: Types of data

3. CLASSIFICATION OF DATA

This article focuses on 4 major classifications: Transactional Data, History Data, Reference Data and Metadata (**Refer Figure 4**)

Transaction data is again categorized into 3 types – structured, semi-structured and unstructured data; **Structured data** includes relational & dimensional databases such as ERP and CRM. **Semi-structured data** includes XML files, tab delimited files, .csv files, JSON. **Unstructured data** includes Social Media data, Emails, photos, audios, videos and presentations.

History data is usually data from Data Warehouse, Data Lakes and Data marts.

Reference Data includes units of measurement, country codes, corporate codes, fixed conversion rates, calendar, geographical data and conformed dimensions.

Metadata includes business glossary, data dictionary, logical data models and data lineage. Data lineage provides information on the origin of data – what happens to data through its journey and where it moves over time.

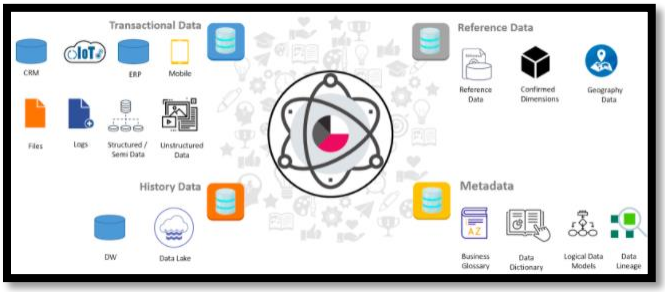


Fig 4: Four major classifications of data

4. CHALLENGING DIMENSIONS FOR DATA ENGINEERING

Simply converting business requirements into products may not meet the business needs. The product owner should analyze the business requirements, apply his/her own intelligence and experience, to convert the business requirements into business goals, by providing the best in class solutions. Businesses are generally not aware of technological advances. The product owner needs to look into the requirements from the user's perspective, learn about the platform where the product would be used and also should be able to compare with similar products that are already available in the market. He/she must also consider the pain points of the users, technology stack, in order to determine the value and benefits that can be derived from the new product, thus designing the vision to be more human and less abstract. Therefore, when business is happy, so are we!

Visionary thinking of the product owner nails the solution.

The industry is experiencing a tsunami of tools and technologies, that are changing every day, high employees' turnover, multiple vendors, humongous amounts of data that crush us every day, varying business streams, and the list goes on. From a Betamax cassette, we graduated to VHS; from VHS, we graduated to VCD; from VCD to DVD; then to Blu-ray and today, we have Netflix and Amazon Prime! Correspondingly, from paper to on-premise.

Now almost everything is in the Cloud!

The constantly changing dimensions can be compiled into 6 different categories: Business Requirements, Tools & Technologies, Employee Turnover, Third Party (Vendors), Business Streams and Data Volume.

Business Requirements are dynamic to the ever-changing economical, commercial, personal and political demands; **Tools & Technologies** keep evolving every day. **Employee Turnover** is fluctuating constantly. **Vendors** keep changing. **Varied Business streams** are being introduced frequently. **Data Volume** experiences exponential upsurge.

But data remains a constant element!

5. CLOUD SERVICES OFFERED BY VENDORS

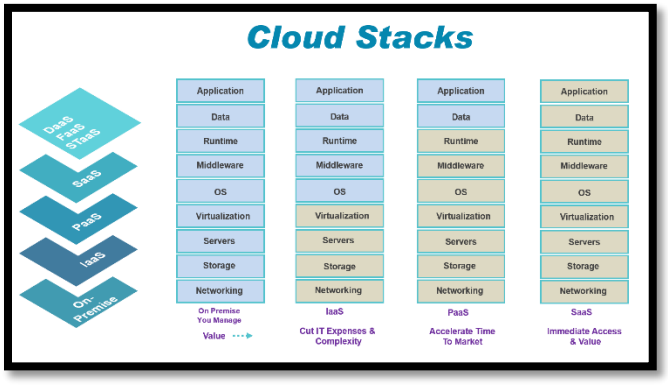


Fig 5: Cloud Solutions

Figure 5 represents an array of solutions that are available in the market to choose from. The colours on the stack signifies a range of the cloud services that are offered by vendors.

6. CLOUD SOLUTION BENIFITS

Multiple benefits are offered by Cloud that includes storage, flexibility, disaster recovery and business continuity. **Figure 6** explains the different benefits obtained by using Cloud solutions.

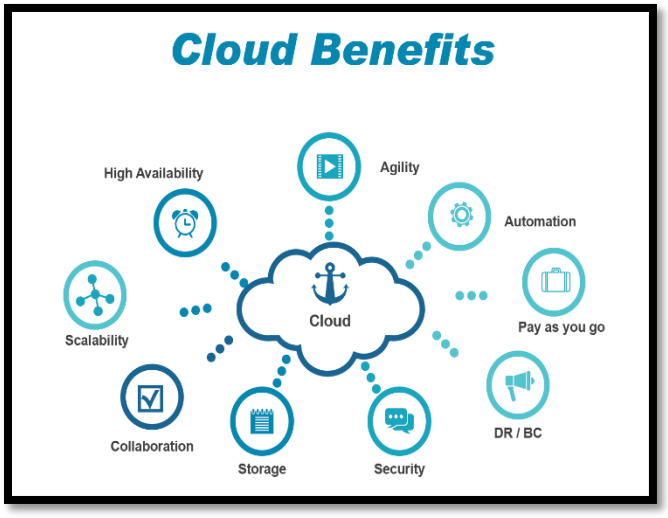


Fig 6: Cloud solution benefits

7. HYBRID APPROACH – MIX OF IAAS, PAAS, SAAS

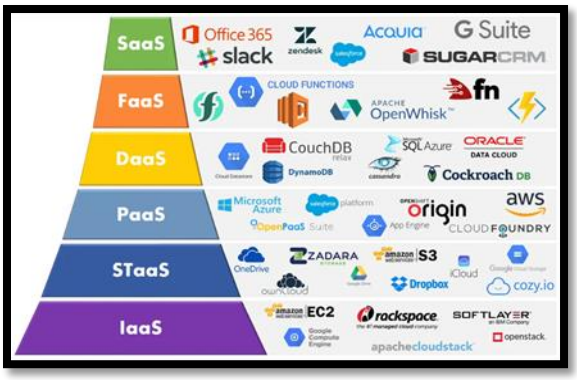


Fig 7: Examples of different Cloud services

If the organization operates on sensitive data – mission critical to business – it should opt for Private Cloud and implement **IaaS** (Infrastructure as a Service), to isolate the data on its own servers. If development is the priority, then **PaaS** (Platform as a Service) should be the option to consider. For other standard requirements, where organizations prefer to use off-the-shelf products, **SaaS** (Software as a Service) is recommended. **Figure 7** provides examples of **Iaas**, **Paas** and **SaaS** and the supported applications from different vendors.

Applying due diligence, we can cherry-pick the likely solution that serves the purpose. In the real world, companies prefer a hybrid approach.

8. Cloud providers and Consumers

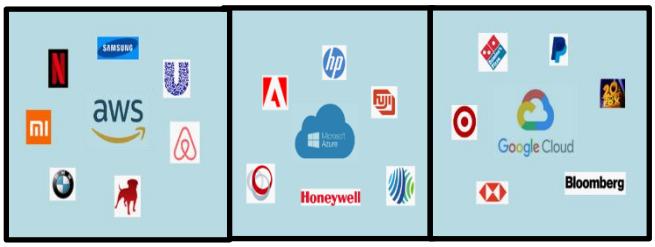


Fig 8: Top 3 Cloud services and their Customers

Listed above in **Figure 8**, are some of the top-end customers and their choices of cloud providers. **AWS cloud services** are used by Netflix, Airbnb, Unilever, BMW, Samsung, MI & Zynga. **Azure** services are used by Johnson Controls, Polycom, Adobe, HP, Fujifilm, and Honeywell. **Google Cloud solutions** are used by HSBC, Paypal, 20th Century Fox, Bloomberg, Target, and Dominos.

9. JOURNEY OF DATA

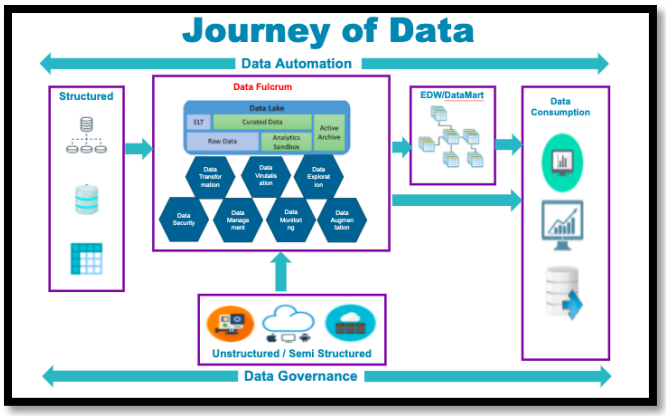


Fig 9: Journey of Data

Now that we understand the different data types, complexities and environments, we will now examine the journey of data. There are 4 major steps involved in the journey of data: Ingestion, Storage, Compute and Analysis.

Figure 9 represents the Journey of Data. The data begins its journey from the source systems that includes **structured**, **semi-structured** and **unstructured data**. It is then injected into the **data lake** in a raw format. Owing to the large volume of data at this stage, various tools and techniques can be applied to transfer the data efficiently. The advantage of **data lake** is the separation of Storage and Compute. This means, they both operate in separate layers and hence the Compute does not get affected by the Storage. Compute involves data processing by the application of business logic and converting the data into consumable formats that can be used by DW (Data Warehouse), BI & AI for advanced analysis. Data experiences security and governance throughout its journey, to keep all of us out of legal trouble.

10. DEFINITION OF DATA FULCRUM

Data Fulcrum is a concept that brings data to the spotlight with its own flavors of processes like data governance, data transformation, data security, data automation, data augmentation and data monitoring.

11. MASTER DATA GOVERNANCE

Figure 10 displays the different functions of data governance involved with respect to people, processes and technologies.

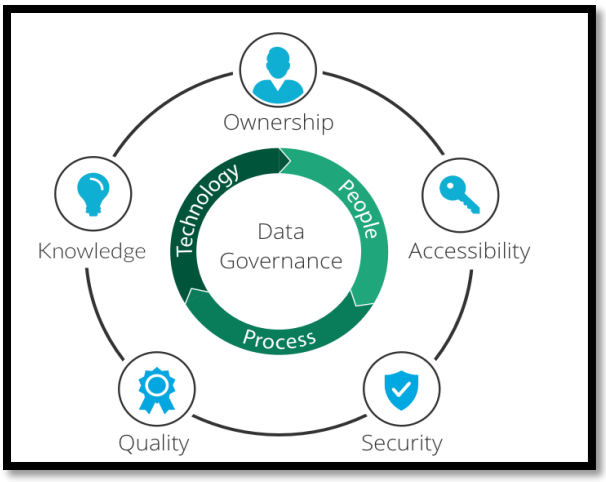


Fig 10: Data Governance

Key people involved in Data Governance are **Data Custodians** and **Data Stewards**. Data Custodians manage and govern the data. Data Stewards are the Subject Matter Experts, who are familiar with the data used by a specific business unit.

Master Data Management (MDM) is a subset of Data Governance that serves as a single point of reference for business-critical data. It includes definition of data types, different varieties of data and specifies the values.

Data Governance covers data quality, data availability, data usability, data integrity and data security. **General Data Protection Regulation (GDPR)** is an example of Data Governance framework, which was implied on the 25th May 2018, all over the world. GDPR placed guidelines for collecting and processing customers' personal information.

12. RESULT OF IMPLEMENTING GOVERNANCE

The result of governance framework implementation is phenomenal. The organisation complies with standards like SOX and GDPR. The value of data is amplified as the data can be re-used. Data-driven decision making is widely improved across the organisation; cost of data management and performance optimization is enhanced.

DATA AUTOMATION

Following are the tools and processes that helps in understanding data automation and optimization.

a) Data transfer and computation using Spark on Kubernetes.

Figure 11, explains at a high level, about a popular tool, "SPARK", that is used for data transfer and computation.

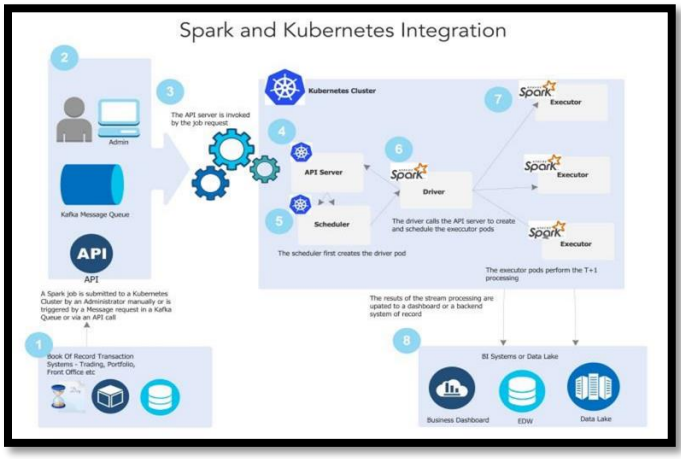


Fig 11: Automation of Data transfer and compute using Spark and Kubernetes

If the volume of data is huge, we use Spark on Kubernetes, to optimize the performance. There are 7 major steps involved in this process: The Spark job is first submitted to Kubernetes Cluster, by the administrator, either manually or triggered by **Kafka** or through **API**. The API Server gets invoked by the job request. The **Driver** & **Executor** nodes then get started. The driver node in the cluster distributes the workload to executor nodes. Executors are assigned to each partition so that the spark jobs are optimized. Executors will send the result back to driver after the execution is completed. The processed data is then sent to DW or exposed to AI & BI through APIs or returned to the Data lake, where the files get stored in a compressed format like the parquet format. Row-level and column level security filters are then applied to the data in DW, before being brought into BI applications for advanced analytics.

Note: If there is a lot of streaming data, then data is sent to the queue using **Kafka**.

There are various automation techniques that can be applied for unit testing, deployment to QA (Quality Assurance) and release to production.

b) Deployment Optimization using Jenkins and Third-Party Tools

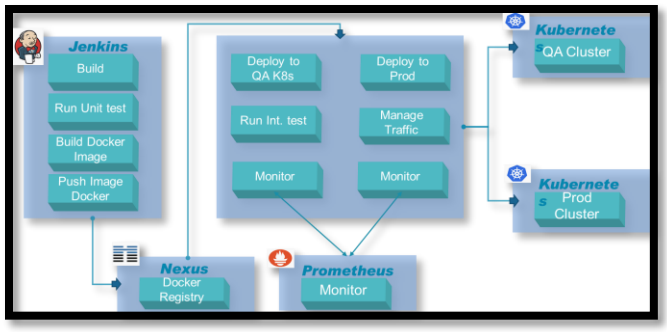


Fig 12: Automation of testing and deployment using Jenkins and third-party tools

To optimize testing, releases or deployment, we use CI-CD (Continuous Integration – Continuous Deployment) tools, to experience reduced 'Time to Market'.

Figure 12 displays the workflow of testing and deployment automation by integrating **Jenkins** with third party tools. Jenkins workflow will run the unit tests. The image gets built and sent to the Docker registry. Nexus is an example for Docker registry. The software release management tools are very useful for managing traffic and 'canary deployments', that are called 'blue-green deployments'.

For example, if some changes need to be added to the production environment, a new environment is spinned up in parallel and the changes are amended accordingly. The workload is then slowly shifted to the new environment; then the existing production environment will be deleted.

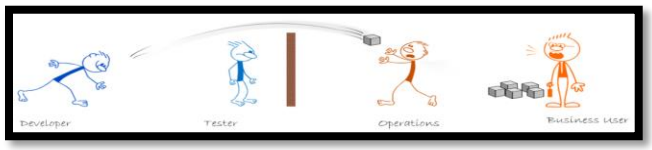
This way, the data never experiences downtime!

To control this kind of traffic, third party plugins or load balancers can be used. There are different open-source softwares available in the market, that will help with software release management between different environments.

Generally, if a workflow has to be written in Jenkins, it is a very tedious process, since there is a need to write tasks for controlling the workflows. However, in open-source tools, we can easily code in YAML files. Hence, we can opt for a third-party plugin or load balancer.

c) Linear Methodology of Production Release

During development and release to production…



And when data defects are found in production after the release!

Fig 13: Linear methodology of Production release

Figure 13 explains the issues involved in production release using linear methodology.

In linear methodology, there is always a wall between the development team and the operations team. The developer writes the code and passes it on to the QA, that finishes running the tests and passes it on to the operations to productionise it. Once the product is released, it ends up with bugs.

The second image shows the ownership is shifted to the developer: the business user points to the operations; the operations points to testing and the testing points to the development team.

The developer is held responsible in the end!

d) DataOps Transform the Culture of the Organization

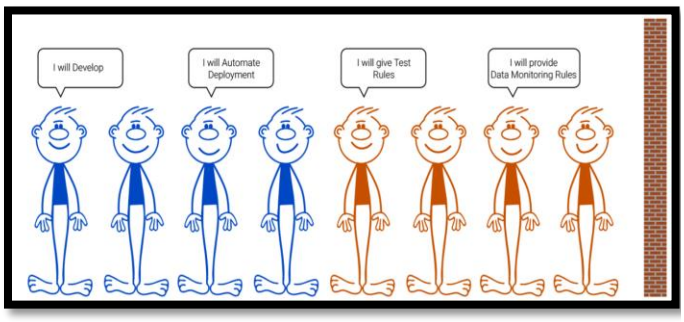


Fig 14: DataOps methodology of Production release

Figure 14 displays a happy autonomous team, where everyone owns the responsibility and shares the workload.

e) Comparison of Linear Vs DataOps

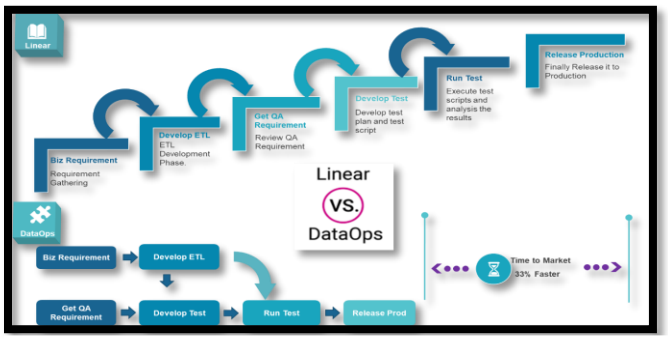


Fig 15: Linear vs DataOps

Figure 15 shows that Adoption of Data Ops results in 'Time to Market' optimization by 33%

f) Test Automation using GIT workflow

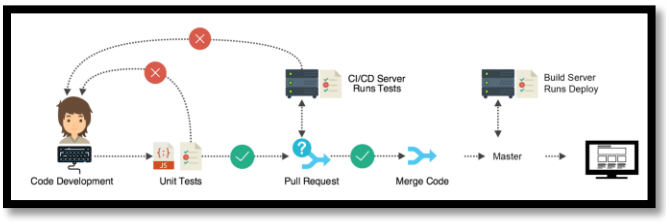


Fig 16: GIT workflow

Figure 16 displays a detailed workflow of GIT – distributed version control system. Every code that gets submitted to GIT needs to experience rigorous testing, followed by code review and approval, before submission of the code to the Master Branch. In the figure above, the developer submits the code and creates a pull request. The CI/CD server runs the tests to validate this code. The approver then reviews and approves the code. The code is finally submitted to the Master Branch and gets deployed to the Production.

And that is the journey of DATA!

REFERENCES

1. Bruno, Eric J. "How to Use Containers to Take on Hybrid-Cloud Data." TechBeacon, TechBeacon, 15Aug. 2018, techbeacon.com/enterprise-it/how-use-containers-take-hybrid-cloud-data. Accessed 4 Dec.2019.
2. Ayswarrya G. "From Data Oops to DataOps: 5 Things You Need to Know - Atlan | Humans of Data." Atlan | Humans of Data, 14 Nov. 2019, humansofdata.atlan.com/2019/11/what-is-dataops/. Accessed 5 Dec. 2019.
3. Bridgwater, Adrian. "The 13 Types Of Data." Forbes, Forbes Magazine, 7 July 2018, www.forbes.com/sites/adrianbridgwater/2018/07/05/the-13-types-of-data/.
4. "Discipline Is Freedom: How Data Governance Generates Alpha for Asset Managers." FinServ Consulting, 26 June 2020, www.finservconsulting.com/2020/02/data-governance-asset-management/.
5. "Main Page." Wikipedia, Wikimedia Foundation, 23 July 2020, en.wikipedia.org/wiki?curid=44783487.
6. "What Is Data Governance: Imperva." Learning Center, Imperva, 30 Dec. 2019, www.imperva.com/learn/data-security/data-governance/.
7. Melgrati, Iván. "Cloud Services Delivery Models. Which Can Help Your Business?" IMELGRAT.ME, 15 June 2018, imelgrat.me/cloud/cloud-services-models-help-business/.

Article Citation:
Authors Name. Niranjana Mangaleswaran. Building Autonomous Data for the Enterprise with Data Fulcrum. AJR 2020; 1(1): 29 - 36
DOI: https://doi.org/10.5281/zenodo.4021931