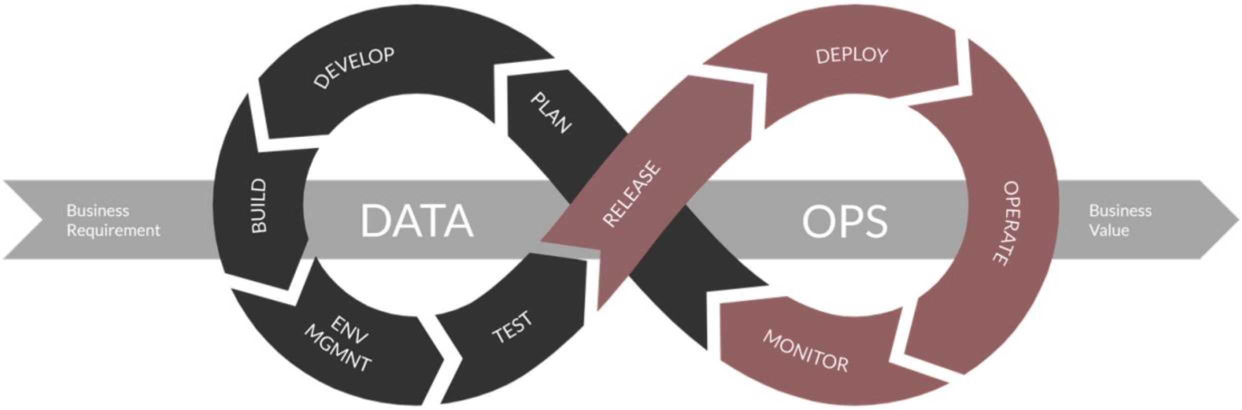
Now more than ever Data Governance is playing a key role in *Cloud* Data Analytics! Data-centric regulation such as

- General Data Protection Regulation (GDPR),
- California Consumer Privacy Act (CCPA) and,
- Health Insurance Portability & Accountability Act (HIPAA) …*to name a few*…

has made the focus on analytics to not only deliver innovation but to do it ***ethically*** and ***securely*** as well! With so much focus on regulation and security data-driven companies must still find the resources to deliver ***innovation*** and ***value***; one such data automation movement is **DataOps**!

What is DataOps?



DataOps extends DevOps[i]

Inspired by **Dev**Ops[ii] (for software engineering), **Data**Ops is a set of practices and processes that extends on the principles of DevOps by focusing on data analytics as repeatable, test driven and agile data pipelines. Easy hey? Well... what is **Dev**Ops?

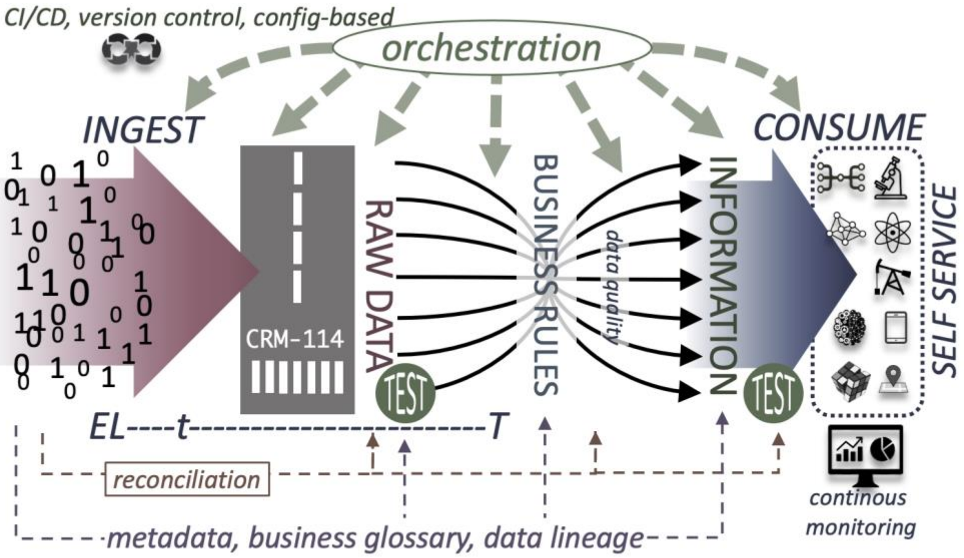
For over a decade software engineering has evolved and produced the practices and ***tools*** to turn software deployment from development and into production-ready code by leveraging

- ***version control*** into code ***repositories***,
- apply ***configuration*** over code***,***
- ***self-organizing*** teams and the use of ***collaboration*** software and,
- Continuous Integration /Continuous Deployment (**CD/CD**) tools to automate code changes through a pre-defined ***test*** suit …

…in order to churn out ***high quality*** code deployments rapidly into a production environment. Code is ***modularized*** and designed to ***scale***. Future changes should ***not*** introduce ***regression***.

DataOps applies this same rigor but to ***data pipelines*** that ***govern*** the flow of data from source (recorded *business events*) through to *consumption* (dashboards, reports, self-service, machine learning) by using data automation *patterns*. *Patterns* in code, *patterns* in automation, *patterns* in agile delivery all the while ***monitoring*** data pipelines to proactively detect ***data quality*** problems before they happen and therefore ***reduce waste***. These are themes also described in ***Lean*** and ***Total Quality Management*** (TQM).

True DataOps emphasizes ***ELT*** over ETL but elaborates on the little "t" in E**t**LT, what is it? The necessary step to deal with ***personally identifiable information*** (PII) by applying ***obfuscation*** on the column contents. As data is landed from a source platform it may appear as plane text, this is okay for data that cannot be used to ***identify*** a person ***uniquely*** (and perhaps impersonate a president and launch a nuclear attack on USSR! – *"Wing Attack Plan R"*). PII content is not restricted to a single column but may involve a number of columns that together can uniquely identify a person. The little "t" is the processing we do either *before* landing or in *staging* to obscure that data consistently by applying ***data masking*** or ***hashing*** and ***encryption*** methods before loading to the ***enterprise data warehouse***.



Modern analytics platform

Where is the Data Governance?

DG Steerco

*"Governance is often stated as ensuring that the **right people** use the **right data** for **the right business purpose** at the **right time** using the **right technology**"- DAMA*

At the center of the DAMA-wheel[iii] is data governance (DG), it is an inherent ***separation of duty*** between ***oversight*** and ***execution*** and it touches every other point of the DAMA framework. This means that the nominated ***data stewards*** (nominated in a federated manner) define and

- manage the ***business glossary***,
- document the ***business rules***, data ***standards***, ***data quality*** rules,

- manage data quality ***issues***,
- ensure that the DG ***policies*** are adhered to and
- the organization meets ***regulatory compliance***

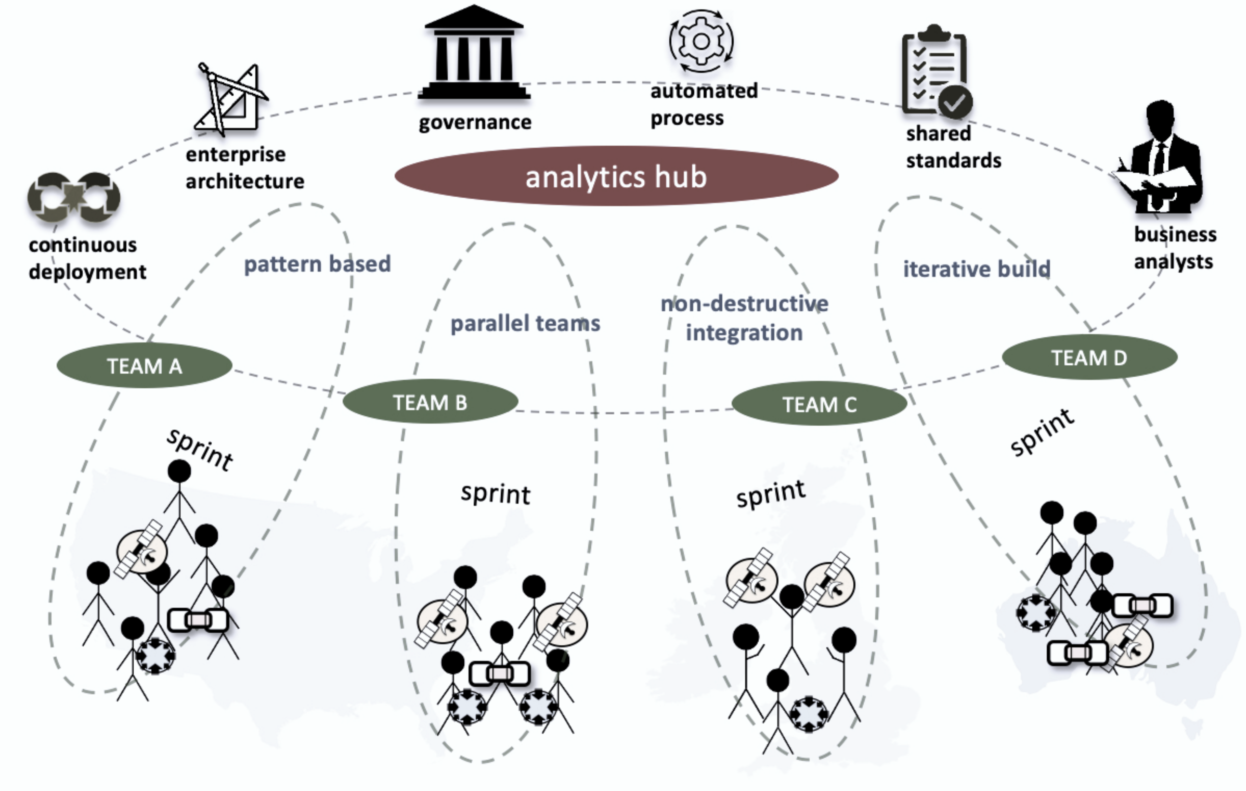
Within the context of DataOps, DG is engrossed into every component but not by sacrificing agility or auditability but instead it is automated through ***modularity*** and ***patterns*** … each ***component*** of the analytics platform can act independently and be ***versioned individually***, including the data governance.

A balance is struck between ***agility*** and governance… too much governance and it can ***stifle*** innovation, too little and it can lead to ***unnecessary exposure*** to risk such as ***data privacy*** and not being able to fulfil industry ***compliance*** requirements that can lead to ***fines*** and ***reputation risk***.

What does this have to do with Data Vault?

WHAT DOES THIS HAVE TO DO WITH DATA VAULT?
WORLD TARGETS IN MEGADEATHS

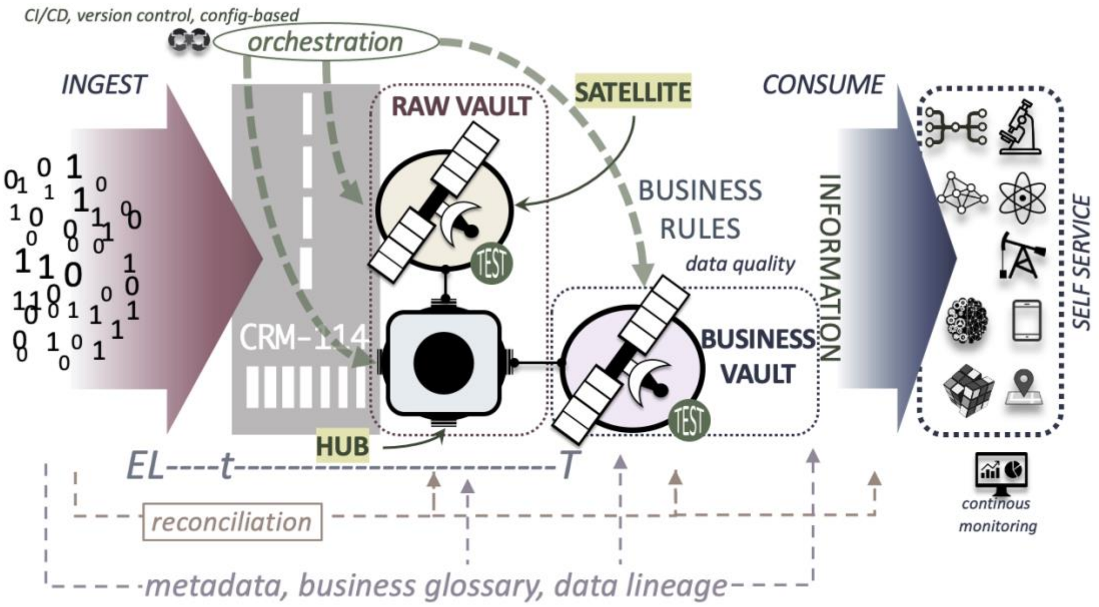
Data Vault 2.0 is the ***methodology*** that describes the best practices for data ***architecture***, ***agile*** delivery and data ***modelling***. It does this by focusing on the thing that is important to the business: the ***customer***. The idea of data vault 2.0 is to ***automate ingestion*** into the enterprise data warehouse as data arrives, this is supported by a repeatable loading framework (pattern) configured to load an enterprise data model as ***hub***, ***link*** and ***satellite*** tables. As the common data vault model delivers the historical reference to everything important to the business through a historized data model, making changes to the enterprise data model does not impact the existing data vault model and can be delivered through ***parallel*** running ***sprints***.



Parallel sprints working on an enterprise data vault model

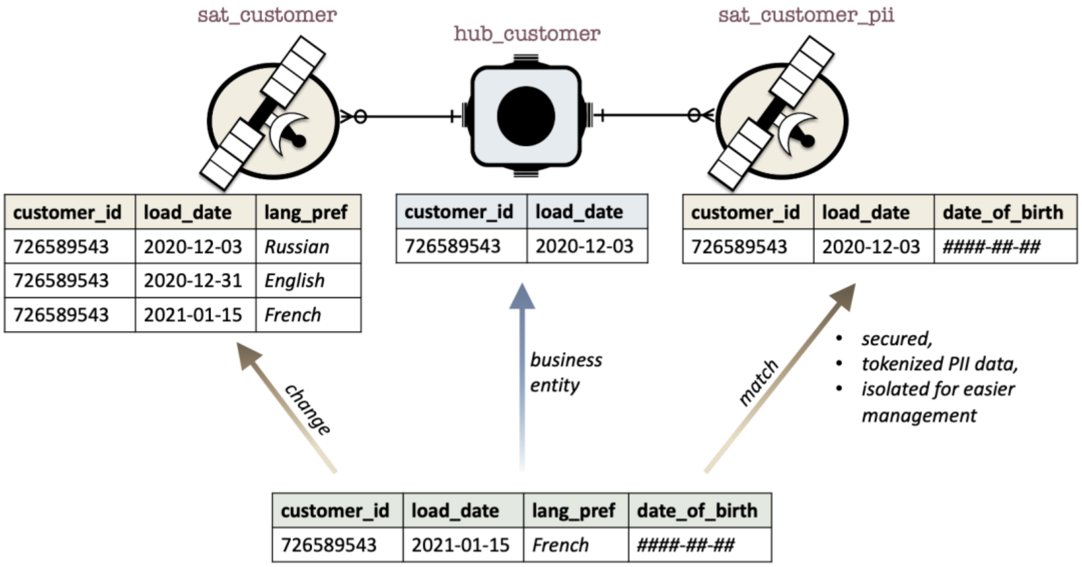
This means that the data vault model is a manifestation of the enterprise ***data governance*** practice and ***data pipeline automation*** as the hubs map to the ***enterprise ontology*** and the links represent the ***relationships*** between ***business entities*** that are often called ***business processes*** or ***unit of work***.

What about the automation you say? Well, there are only ***three loading patterns*** to data vault, and this is where the repeatable patterns through ***configuration*** come into play, we only load hubs (unique list of ***business entities***), links (***unique list of relationships***) and satellites (***change-tracked descriptive data*** about hubs and links) for ***raw vault*** and ***business vault***. As a data vault modeler, you need to ***identify*** what those business entities are, ***understand*** the relationships between them and ***track*** the historical data about those entities and relationships. Where the source application does not supply certain ***information***, we need we build them into business vault reusing the ***same*** but ***decoupled*** automation patterns. As either ***surrogate-hash key*** or ***natural key*** based data vault; all data vault artefacts can be loaded independently without having to implement a staggered loading paradigm ensuring data analytics is delivered as flexibly as possible and at any time portions of the overall data vault model is updated via platform ***orchestration***.



Raw + Business vault = Data Vault

Every data vault artefact includes the ***record source***, ***extract/applied date***, ***load date*** and ***task id*** metadata tags that gives you the data lineage right down to the record level. In addition, Data Vault provides the flexibility to deal with PII data by recognizing a very distinct attribute about PII data, they hardly (if ever) ***change***, observe… **satellite splitting**



customer_id	load_date	lang_pref
726589543	2020-12-03	*Russian*
726589543	2020-12-31	*English*
726589543	2021-01-15	*French*

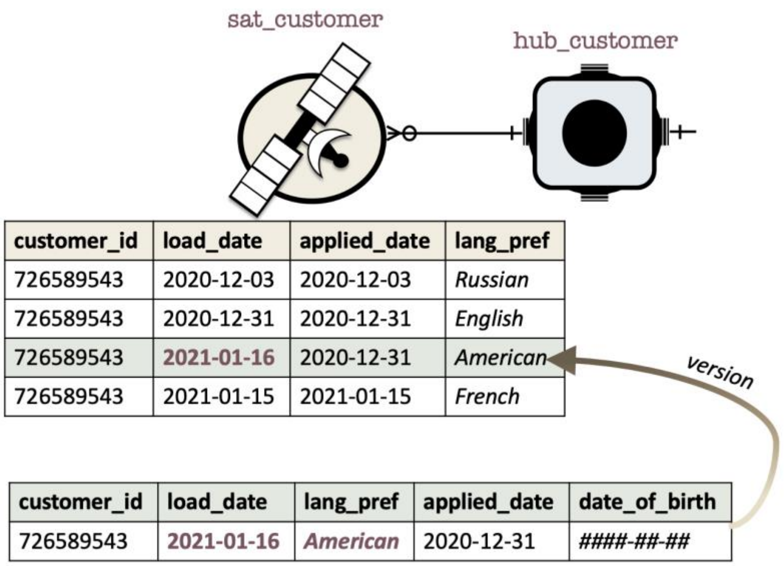
customer_id	load_date
726589543	2020-12-03

customer_id	load_date	date_of_birth
726589543	2020-12-03	*####-##-##*

customer_id	load_date	lang_pref	date_of_birth
726589543	2021-01-15	*French*	*####-##-##*

Improved data governance with data vault

With extract/applied date available this essentially means you can ***version*** records for any point in time. Why is this important? It means if corrections are made to a business entity's or relationship's timeline the update can be applied historically ***without*** removing the record the correction is updating. This ensures that the optimal level of auditability is achieved with data vault that even corrections can be explained by just querying the data.



customer_id	load_date	applied_date	lang_pref
726589543	2020-12-03	2020-12-03	*Russian*
726589543	2020-12-31	2020-12-31	*English*
726589543	2021-01-16	2020-12-31	*American*
726589543	2021-01-15	2021-01-15	*French*

customer_id	load_date	lang_pref	applied_date	date_of_birth
726589543	2021-01-16	*American*	2020-12-31	*####-##-##*

Complete record auditability

Data Vault 2.0 provides the necessary components for building your data analytics and business intelligence platform ***and*** because these are all repeatable and modular patterns, a test suite based on repeatable ***test patterns*** can be utilized to **build trust** in the data platform. The methodology supports change ***without regression*** and ***schema evolution*** without ***refactoring*** when the ***business processes evolve*** and ultimately because data vault artefacts reflects the ***corporate history*** of the enterprise ontology, data vault provides the ***agile*** means for integrating your ***entire*** analytics platform with data governance at the ***top of mind*** and incorporated by design… until next time…

https://bit.ly/379W7kQ

#thedatamustflow #datavault #dataops

[i] True DataOps, https://truedataops.org/
[ii] What is DevOps, https://www.atlassian.com/devops/what-is-devops
[iii] DMBOK2, https://www.dama.org/content/what-data-management