

PP02: Data Engineering Techniques (A PHUSE Educating for the future workgroup)

Mohit Juneja (Independent Consultant, Toronto)
Renu Shukla (Janssen Research & Development, Raritan NJ)

Introduction

Clinical development has made huge efficiencies over recent decades, (e-CRF and CDSIC standards to name a few). However, just as we're reaching the summit of traditional techniques we're facing with another step-change in our journey – the big data mountain.

The methodologies we're currently using cannot cater for these vast and varied data sources without a huge increase in resource, something which most biopharma companies haven't provisioned for.

Common data management and statistical programming techniques need to give way to centralized automated data engineering mindset – including transformative data consolidation, integration, and access – which goes beyond cost reductions, to enable a new phase of data-driven clinical discovery.

“Master data management (MDM)–A technology-enabled discipline wherein business and IT collaborate to ensure uniformity, accuracy, stewardship, semantic consistency and accountability of the enterprise's official shared master data assets (e.g. identifiers and extended attributes that describes the core entities of the enterprise such as customers, suppliers etc.)

Key Success Factors:

- Clear ownership of assets
- Robust Governance
- Ongoing 'cleaning'
- Strong support across enterprise

A Data Lake is a repository that can accept data in any format, hence data can be stored in their native format with no transformation/mapping required.

Data lakes provide following capabilities:

Storage of different types of data for instance – SQL, Text, Binary (images/pdfs), or any other type of data

Metadata Discovery – A centralized Metadata repository that can be used to identify/search data across the lake

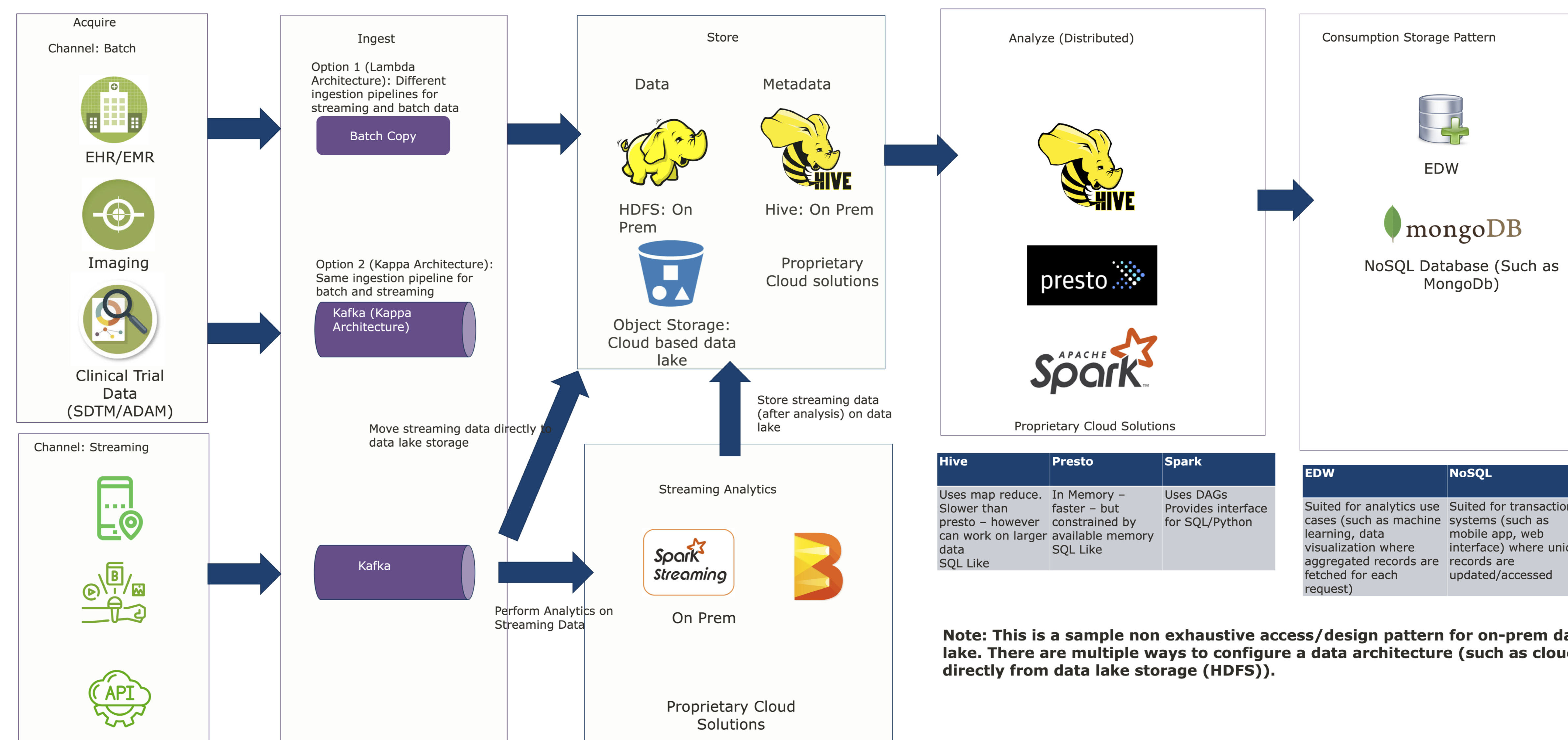
Analytics: Data Lakes provide capabilities to run analytics (such as ETL, data visualization, or machine learning) on the underlying data

Fine grained security model: For managing access control

Data Marketplace

Providers collect appropriate data for certain usage and allow access to users and researchers. Often, data available for free, though more “valuable” data command a fee.

The method for extracting this data is often via RESTful Application Programming Interfaces (APIs). APIs allow developers to call methods & functions (packages of code) to return specific information. APIs vary from provider to provider often based on the expected desire of the data consumer



Note: This is a sample non exhaustive access/design pattern for on-prem data lake. There are multiple ways to configure a data architecture (such as cloud, directly from data lake storage (HDFS)).

References

<https://www.informatica.com/services-and-training/glossary-of-terms/master-data-management-definition.html#fbid=9X0aV1w73Ob>
https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/master-data-mgmt-techniques-106161.pdf
https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/sas-data-governance-framework-107325.pdf
https://whitepapers.em360tech.com/wp-content/files_mf/white_paper/wp_iway_7steps.pdf
http://www.tcdii.com/PDF/Data_Governance_white_paper_-_April_2006.pdf
https://www.sas.com/en_us/insights/articles/data-management/data-lake-and-data-warehouse-know-the-difference.html
<https://www.forbes.com/sites/bernardmarr/2018/08/27/what-is-a-data-lake-a-super-simple-explanation-for-anyone/#63364cb176e0>
<https://tdwi.org/articles/2016/02/19/six-validation-techniques-data-quality.aspx>
https://ec.europa.eu/eurostat/cros/system/files/methodology_for_data_validation_v1.0_rev-2016-06_final.pdf
<https://resources.zaloni.com/blog/validating-data-in-the-data-lake-best-practices>
<https://www3.epa.gov/ttnamti1/files/ambient/pm25/qa/vol2sec17.pdf>
https://www.fws.gov/Planning/Documents/Validation/DOI_Performance_Data_V&V_Guide_Jan03.pdf
<https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/>
<https://www.idginsiderpro.com/article/3226393/what-is-a-data-fabric-and-why-should-you-care.html>

Data Fabric is a system that delivers seamless, real-time integration and access across the multiple data silos of a big data system (data warehouse, data lakes, social media, unstructured data etc.)

There are several goals a data fabric must address

Combine data from established systems – regardless of size and future requirement for scalability, and make that data available to applications

Provide speed, scale and reliability – meet business requirements for speed, scale and reliability across multiple computing environments

Support multiple locations – Allow access to data from systems residing at the edge of the network, in the enterprise data center, and in cloud computing environments

Create a unified data environment – create a global namespace making files easy to find and access, provide strong levels of security, provide compression to reduce overall storage requirements, provide an ability to take a snapshot of the data for backups and to support application development, as well as supporting multi-tenant (multiple company) computing environments

Provide high reliability and availability – the data fabric must provide a highly reliable environment that manages itself and heals itself when there is a problem and provides high availability to service mission critical needs.

Conclusion

This poster only provides an overview of selected established data engineering techniques and potential use cases that can be explored to apply these in the Biopharma industry. For data scientists working in clinical research, these techniques offer powerful approaches to storing, searching and using data for decision making. However, we recommend that before investing in any techniques, teams conduct a thorough cost benefit analyses and evaluate anticipated return on investment. For example, potential adopters of master data should consider governance and buy-in by key stakeholders. Teams should evaluate both short-term and long-term sustainability and applicability of these techniques along with flexibility to integrate with newer techniques in the future.

For more information, contact: Mohit Juneja at junemohit12@gmail.com; Renu Shukla at rshukla3@its.jnj.com