

DH03 – Educating for the Future: Data Engineering Techniques

Presentation:

Guy Garrett, *Achieve Intelligence*

Renu Shukla, *Janssen Research & Development*

Mohit Juneja, *LyfeScience*

Contribution:

Andy Richardson, Amy Gillespie, Beate Hientzsch, Berber Snoeijer, Beverly Hayes, Jatin Patel, Karnika Dalal, Mark Bynens, Mike Carniello, Parag Shiralkar, Paul Slagle, Ralf Goetzelmann, Rohit Banga, Sagar Jain, Susan Olson, Vijay Pasapula, Vince Marinelli, Xiaohui Wang

Special thanks:

Ian Fleming, James McDermott, Wendy Dobson

<https://education.phuse.eu/eftf/data-engineering/>

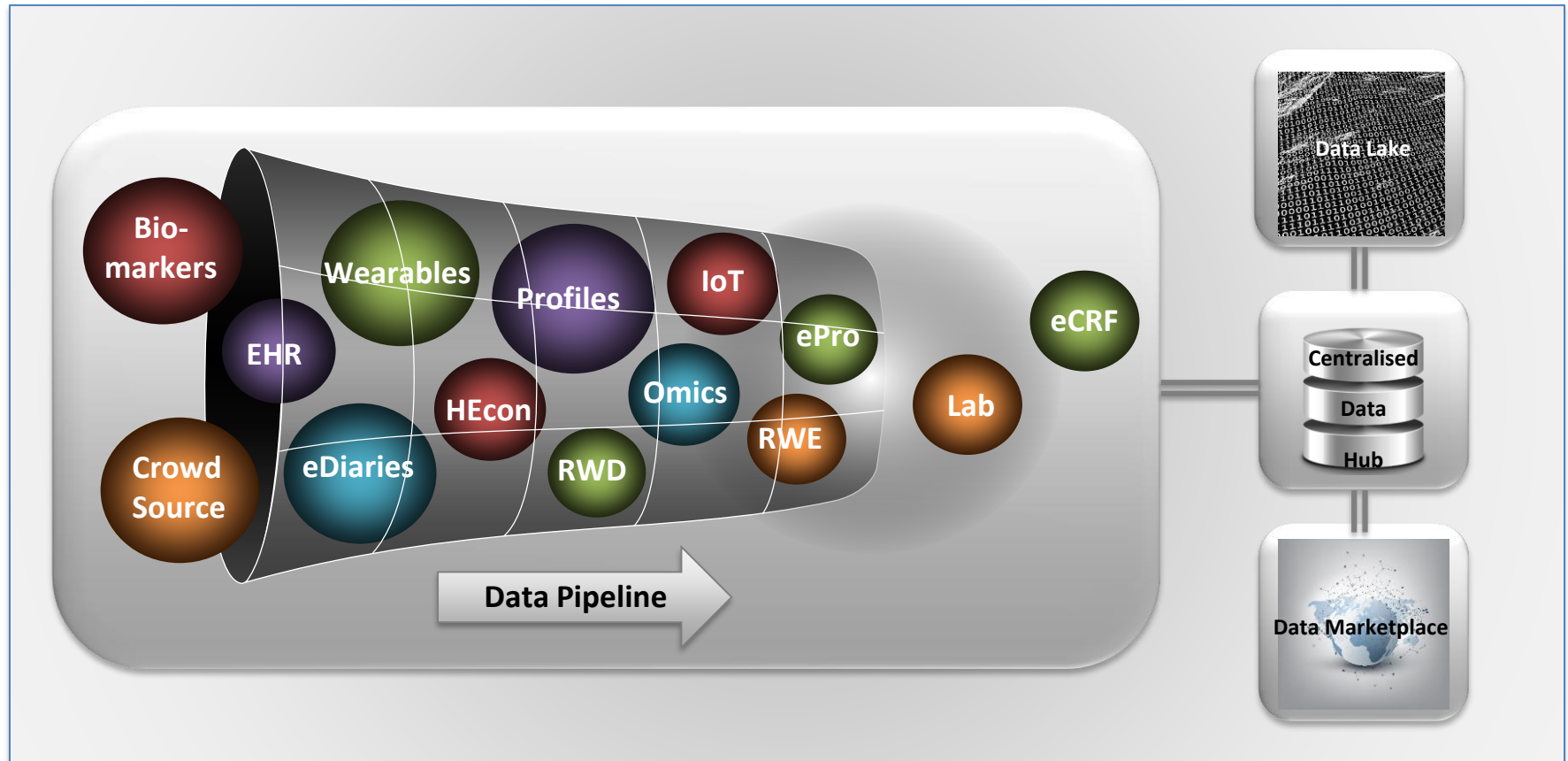
Agenda

- Introduction
- Master Data Management - Principle
- ETL Overview – “Transforming”
- Data Lake – “Trawling”
- Data Marketplace – “Targeting”
- Conclusion

Agenda

- Introduction
- Master Data Management - Principle
- ETL Overview – “Transforming”
- Data Lake – “Trawling”
- Data Marketplace – “Targeting”
- Conclusion

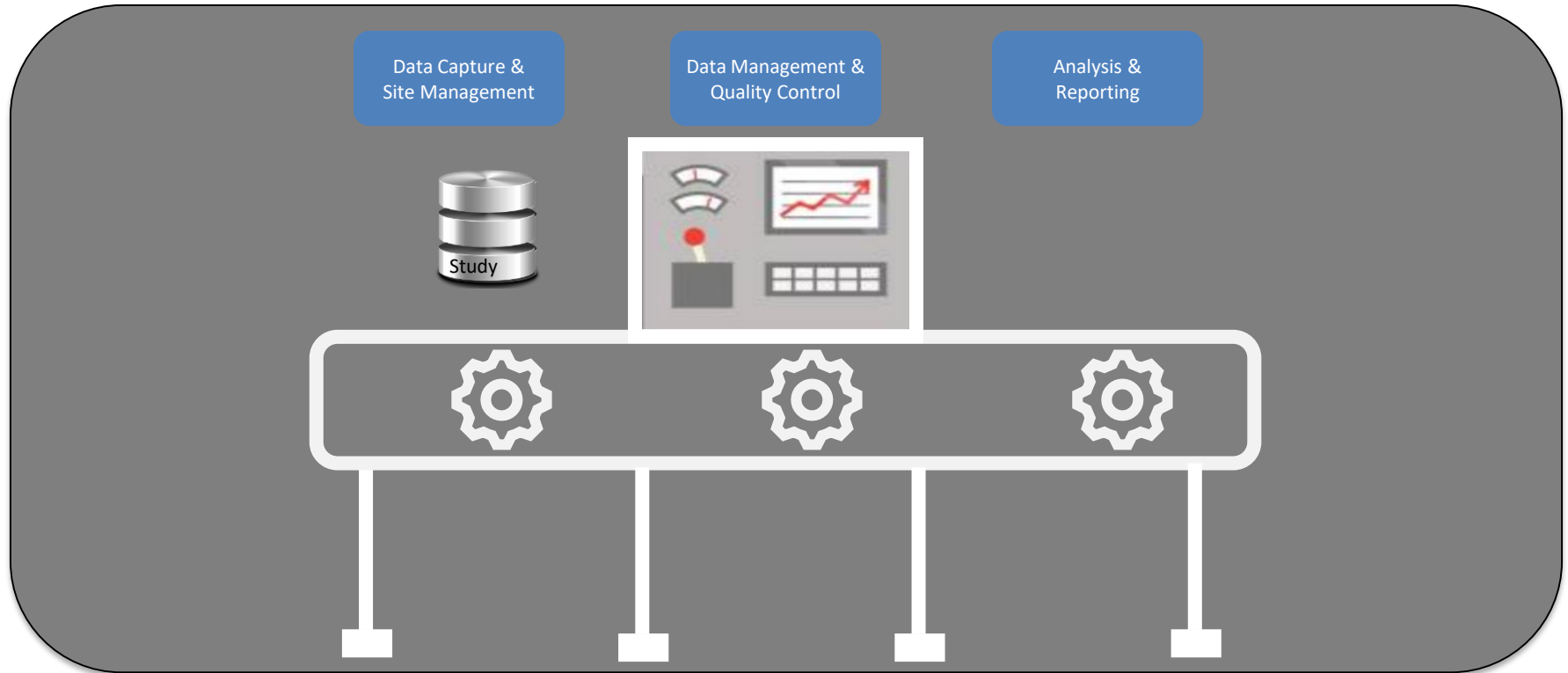
Data Engineering Landscape



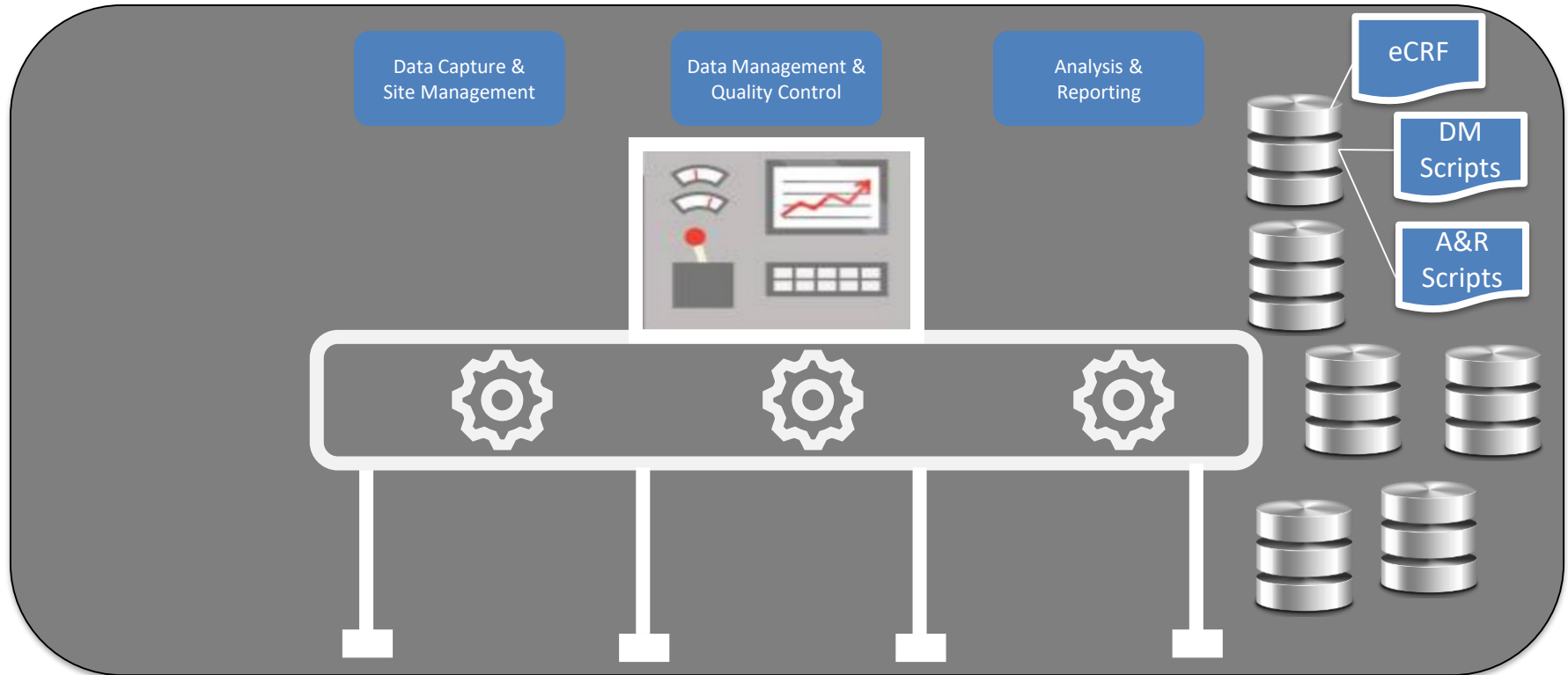
Agenda

- Introduction
- **Master Data Management - Principle**
- ETL Overview – “Transforming”
- Data Lake – “Trawling”
- Data Marketplace – “Targeting”
- Conclusion

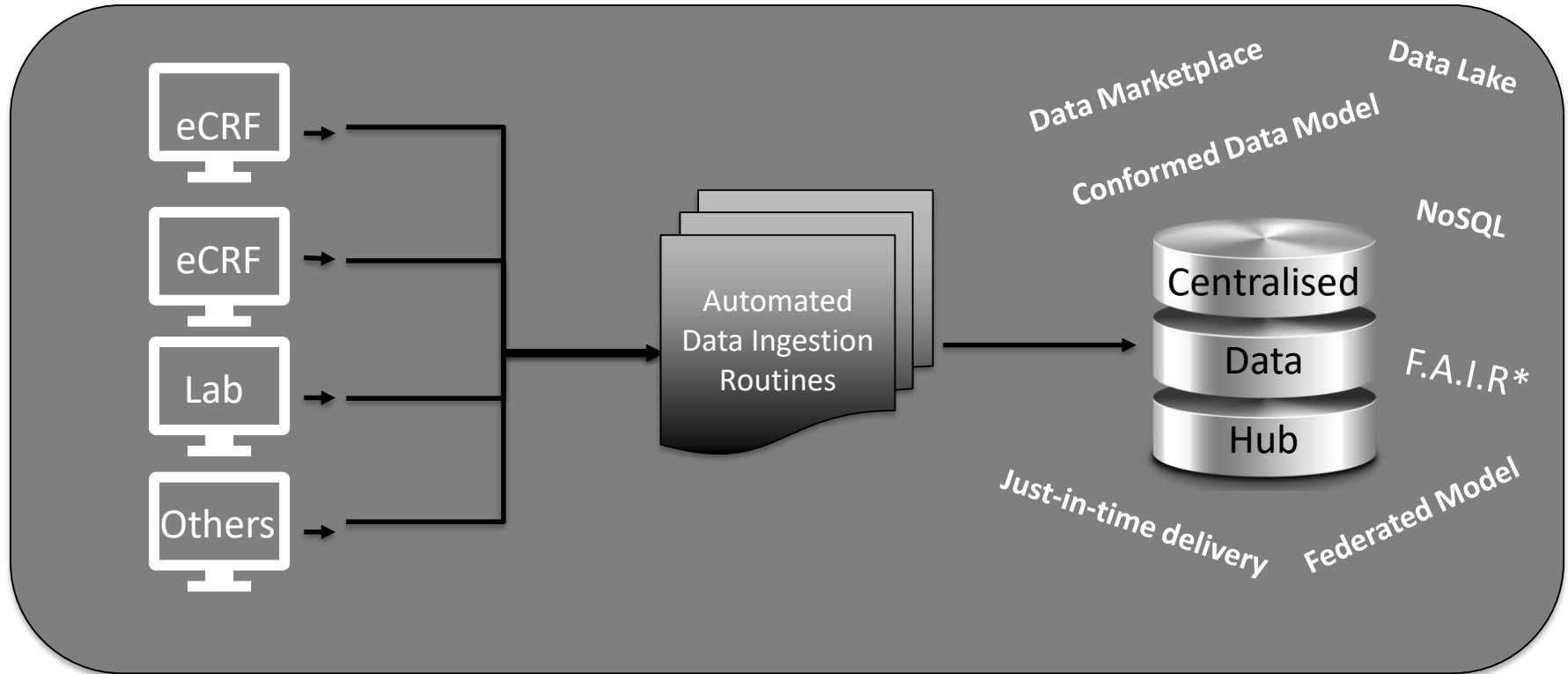
Master Data Management



Master Data Management



Master Data Management



* F.A.I.R. = Findability, Accessibility, Interoperability, Reusability

Master Data Management

- Principle

Approach	Investment	Return on Investment
Master Data Management	More up-front investment	Longer term savings
Siloed Data Processing	Less up-front investment	Most costly long-term

Master Data Management

- Principle

Approach	Investment	Return on Investment
Master Data Management	More up-front investment	Longer term savings
Siloed Data Processing	Less up-front investment	Most costly long-term



Data Engineering Project is looking for Use Cases to evidence this principle.

- 1) Measure current processes
- 2) Adopt MDM
- 3) Measure MDM processes
- 4) Analyse ROI

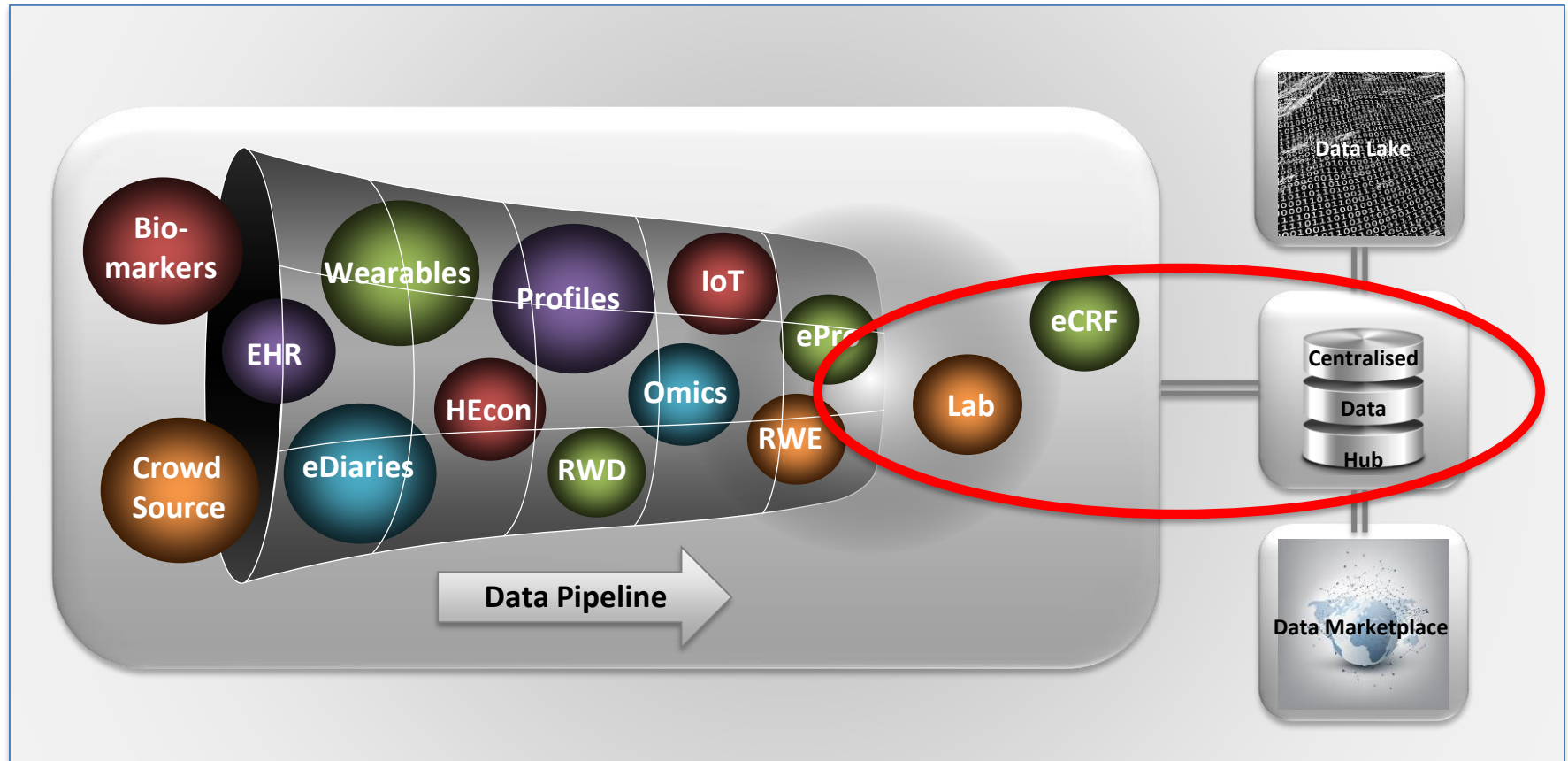
Agenda

- Introduction
- Master Data Management - Principle
- **ETL Overview – “Transforming”**
- Data Lake – “Trawling”
- Data Marketplace – “Targeting”
- Conclusion

Agenda

- Introduction
- Master Data Management - Principle
- ETL Overview – “Transforming”
 - Definition
 - Staging & Loading
 - Change Data Capture
 - Data Validation
 - Scheduling & Batch Processing
- Data Lake – “Trawling”
- Data Marketplace – “Targeting”
- Conclusion

ETL Overview – “Transforming”



ETL Overview – “Transforming”

- Wikipedia definition

“In computing, **extract, transform, load (ETL)** is the general procedure of copying data from one or more sources into a destination system which represents the data differently from the source(s) or in a different context than the source(s).”

ETL Overview – “Transforming”

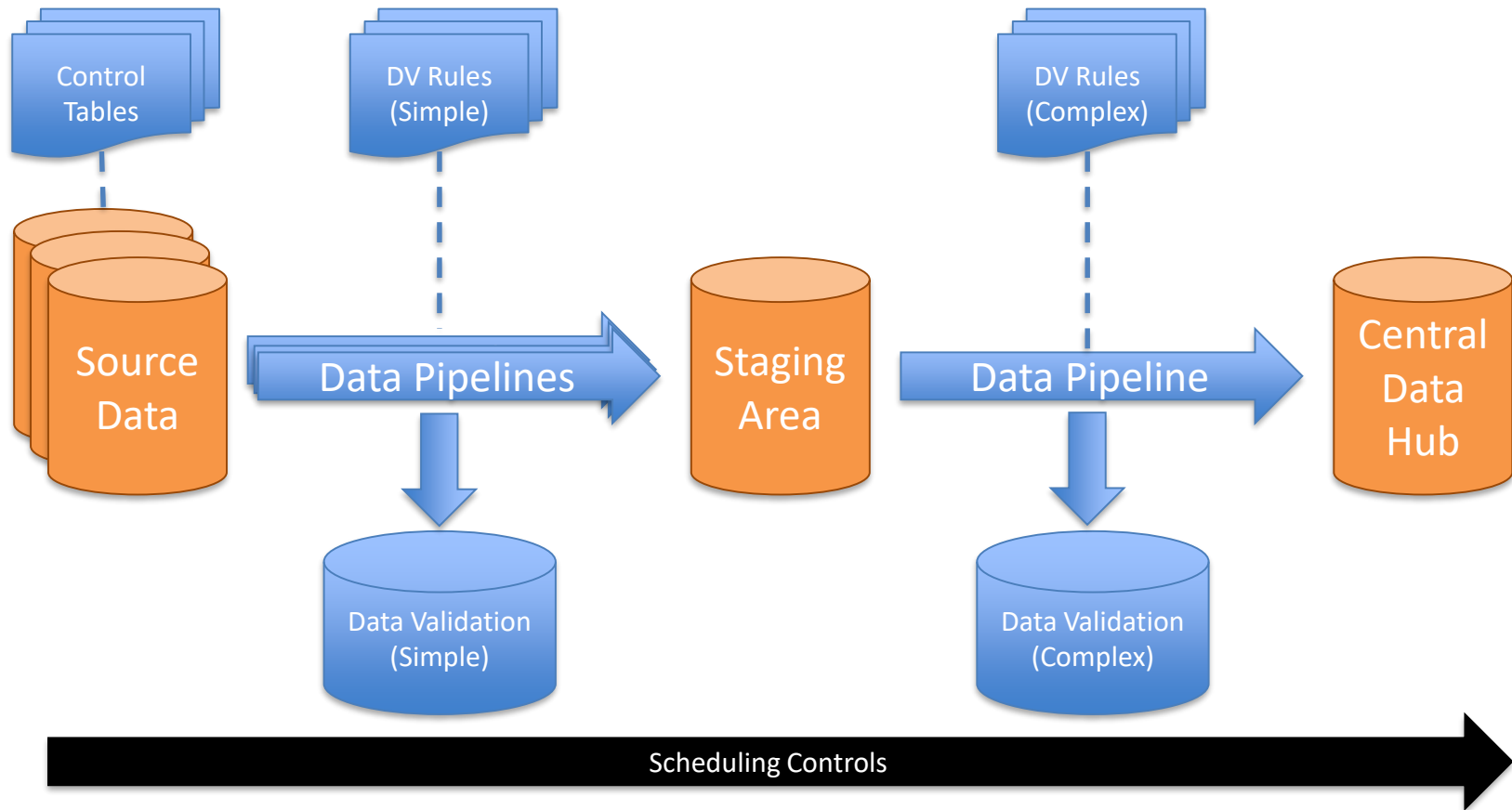
- Wikipedia definition

“In computing, **extract, transform, load (ETL)** is the general procedure of copying data from one or more sources into a destination system which represents the data differently from the source(s) or in a different context than the source(s).”

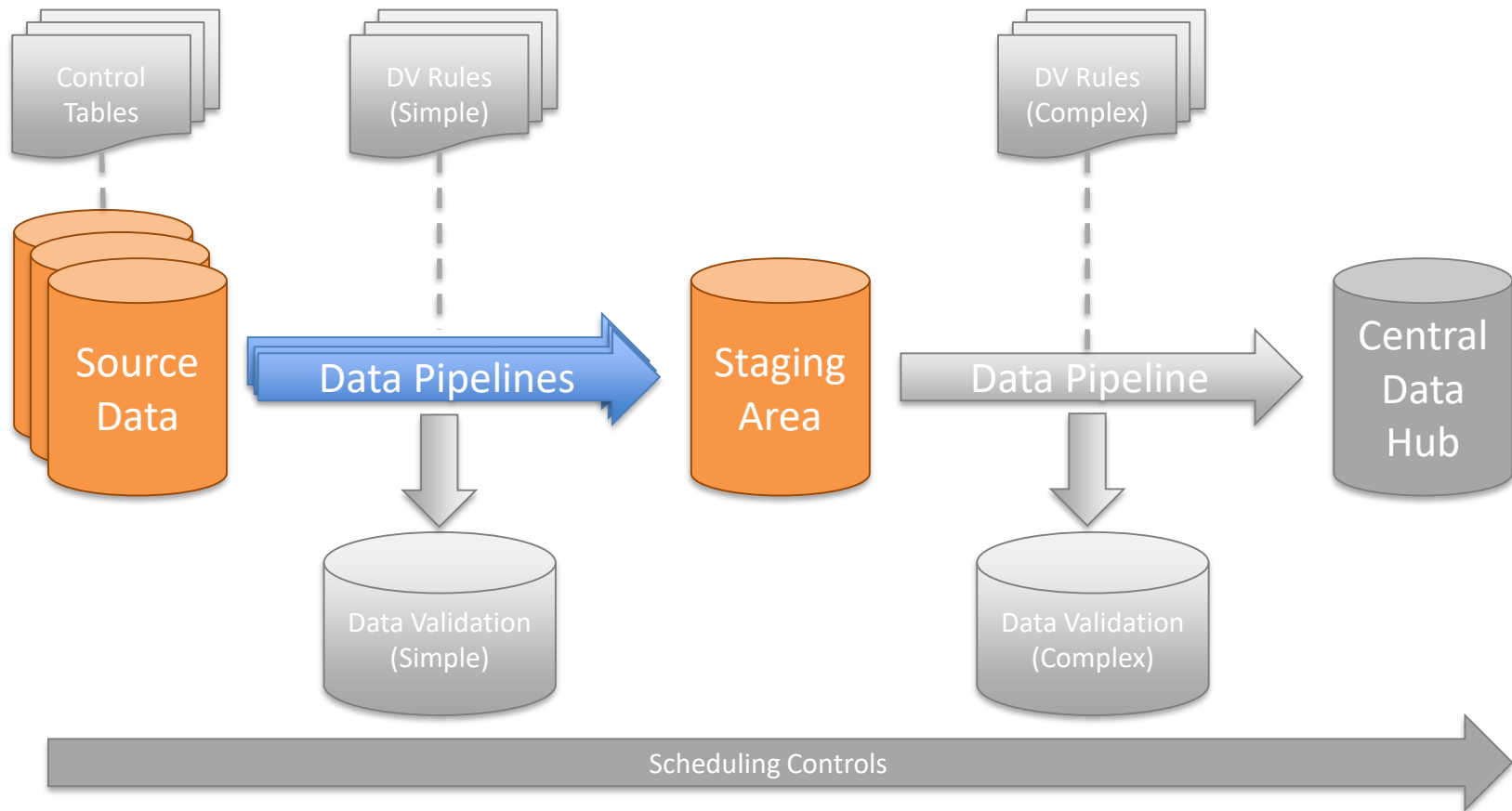
“...**data engineering**, is a software engineering approach to designing and developing information systems. “

i.e. It's about building data pipelines, rather than datasets.

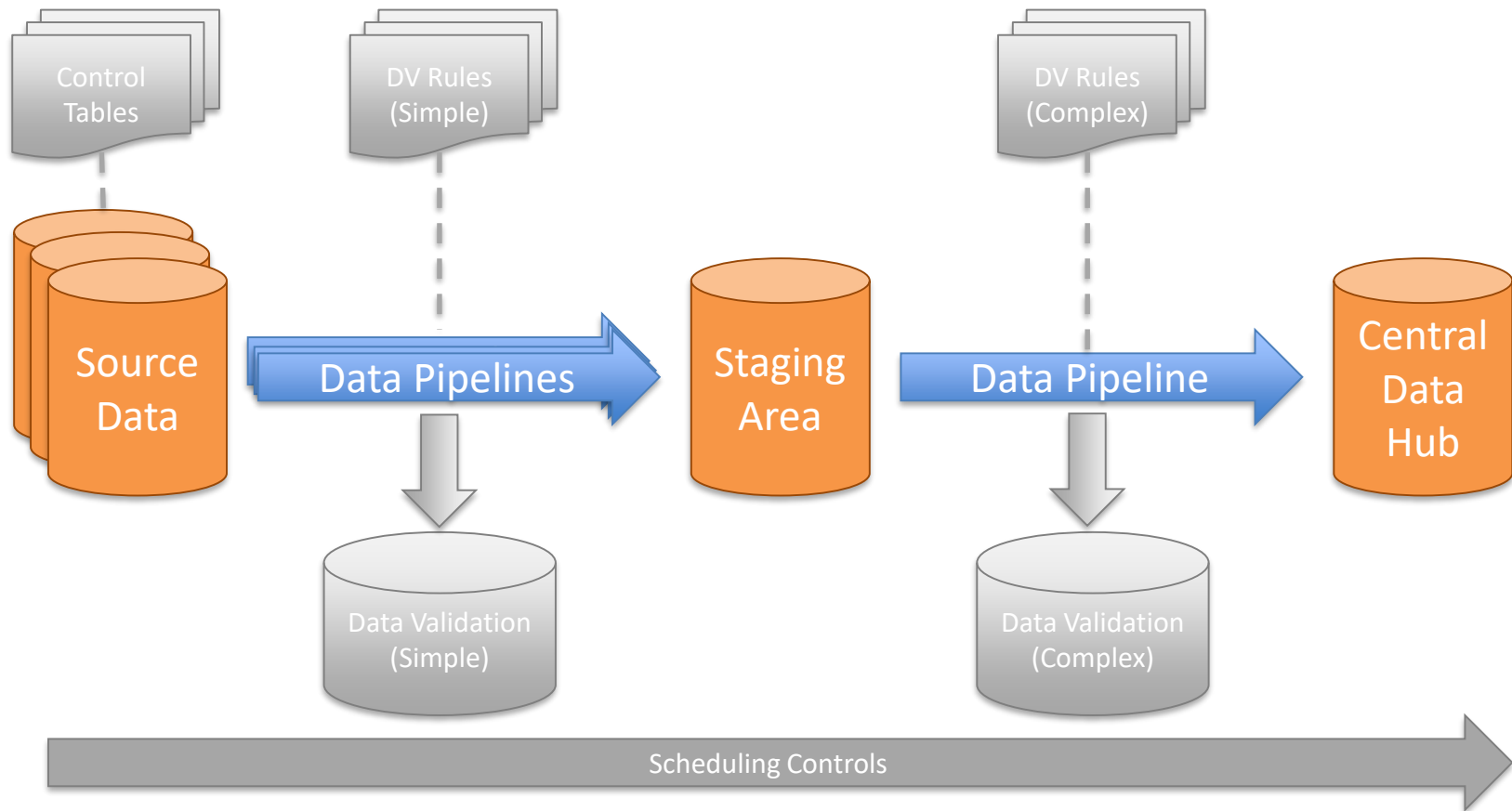
ETL Overview – “Transforming”



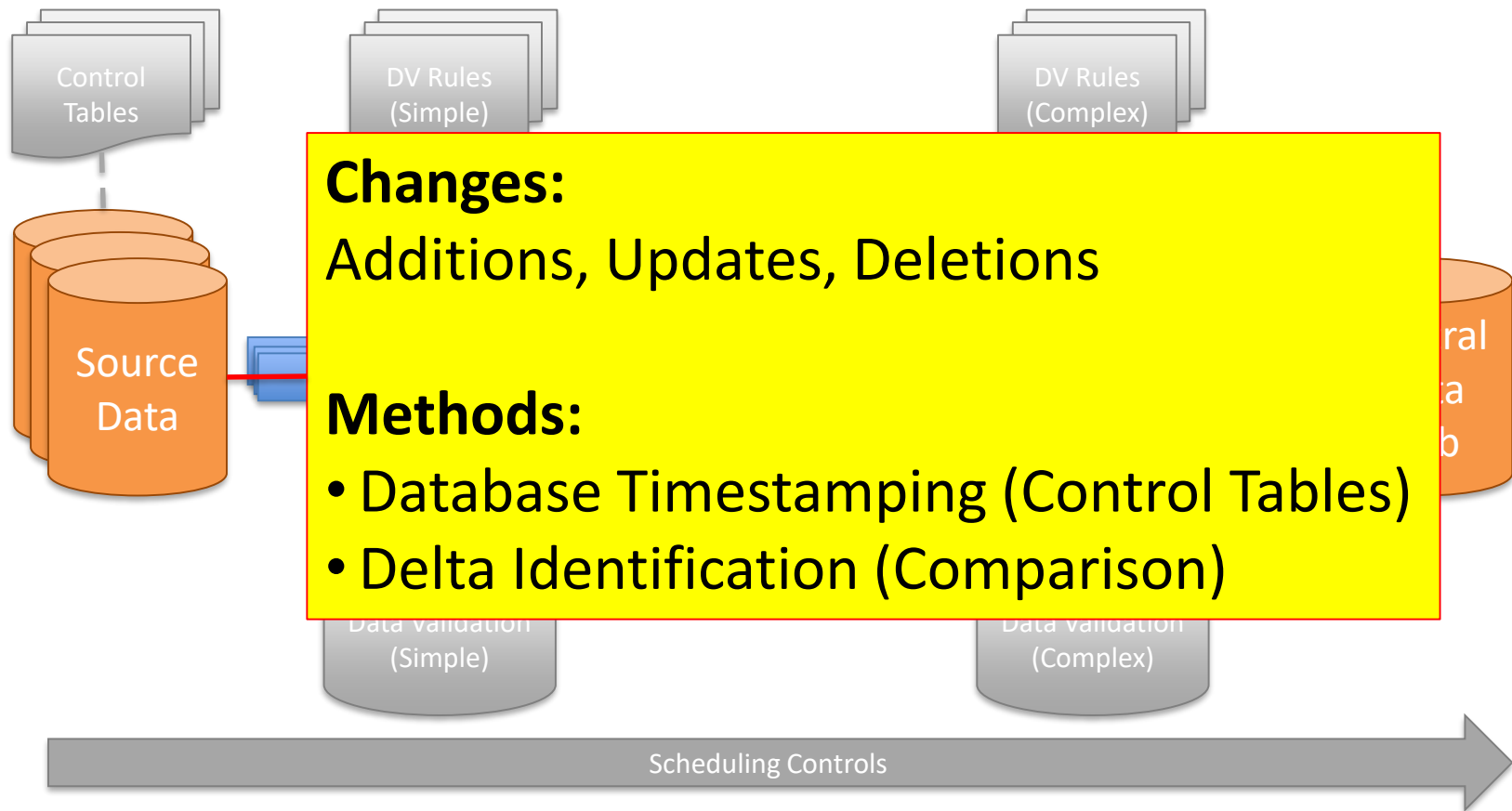
Staging



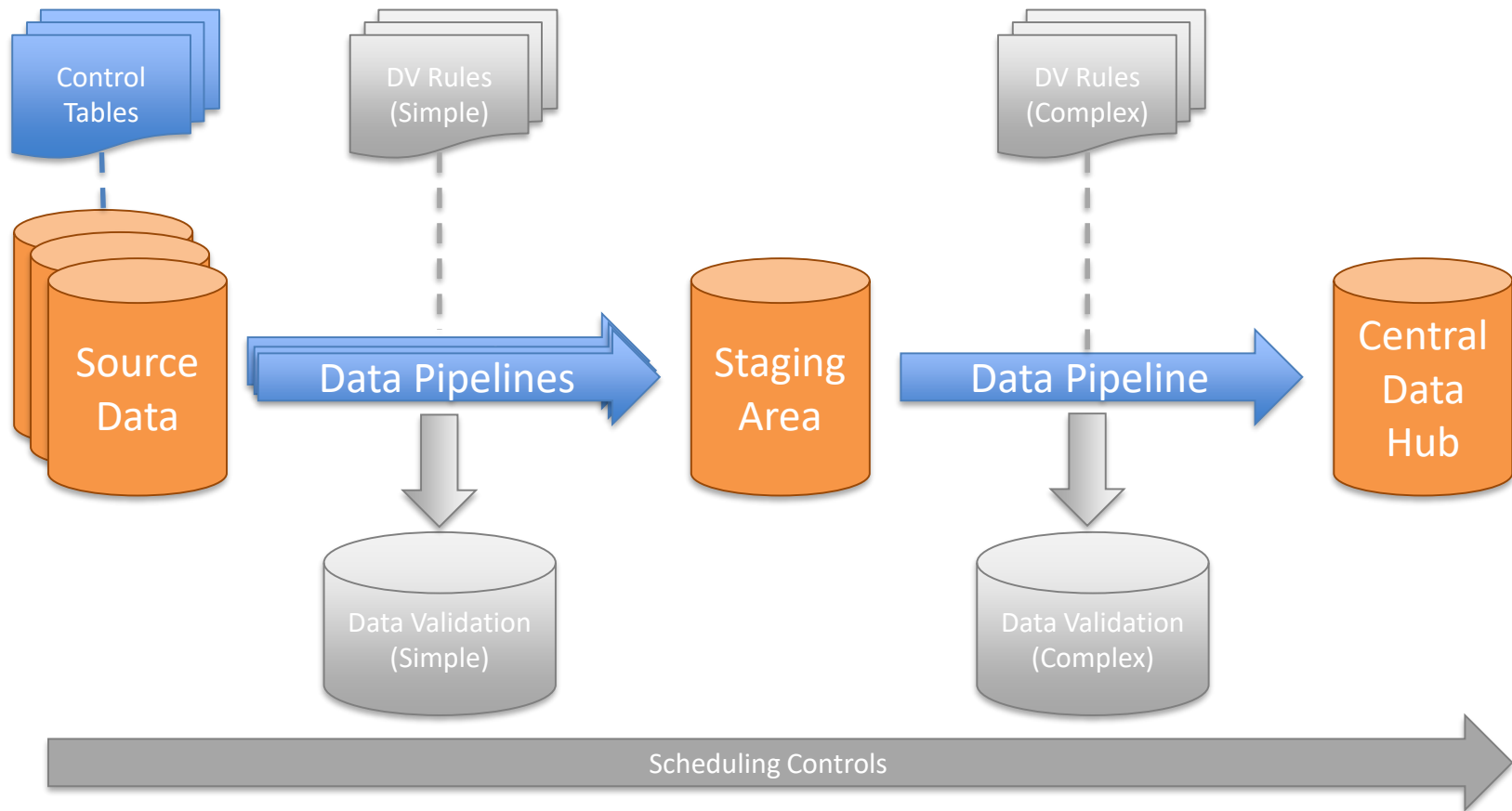
Loading



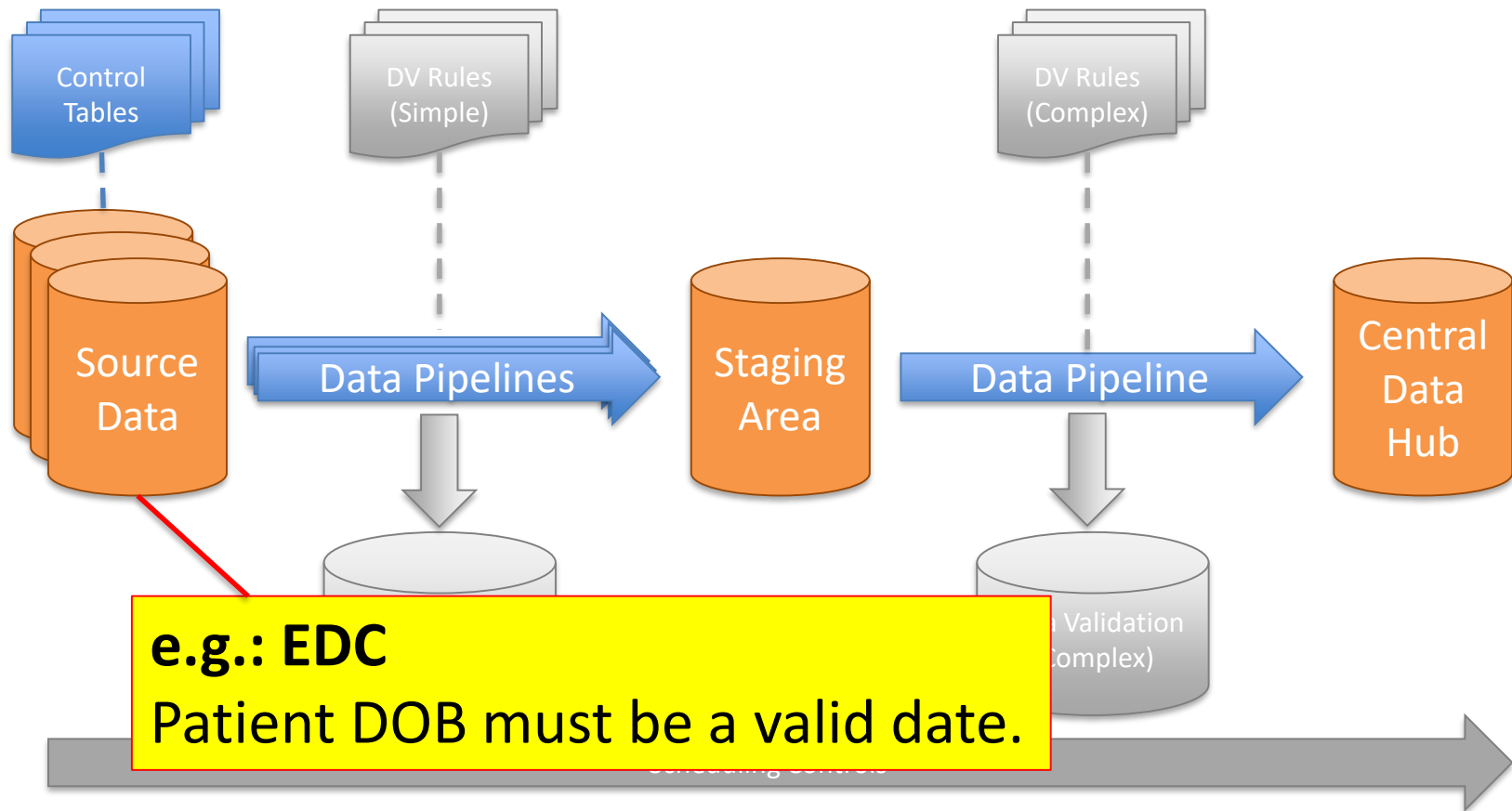
Change Data Capture



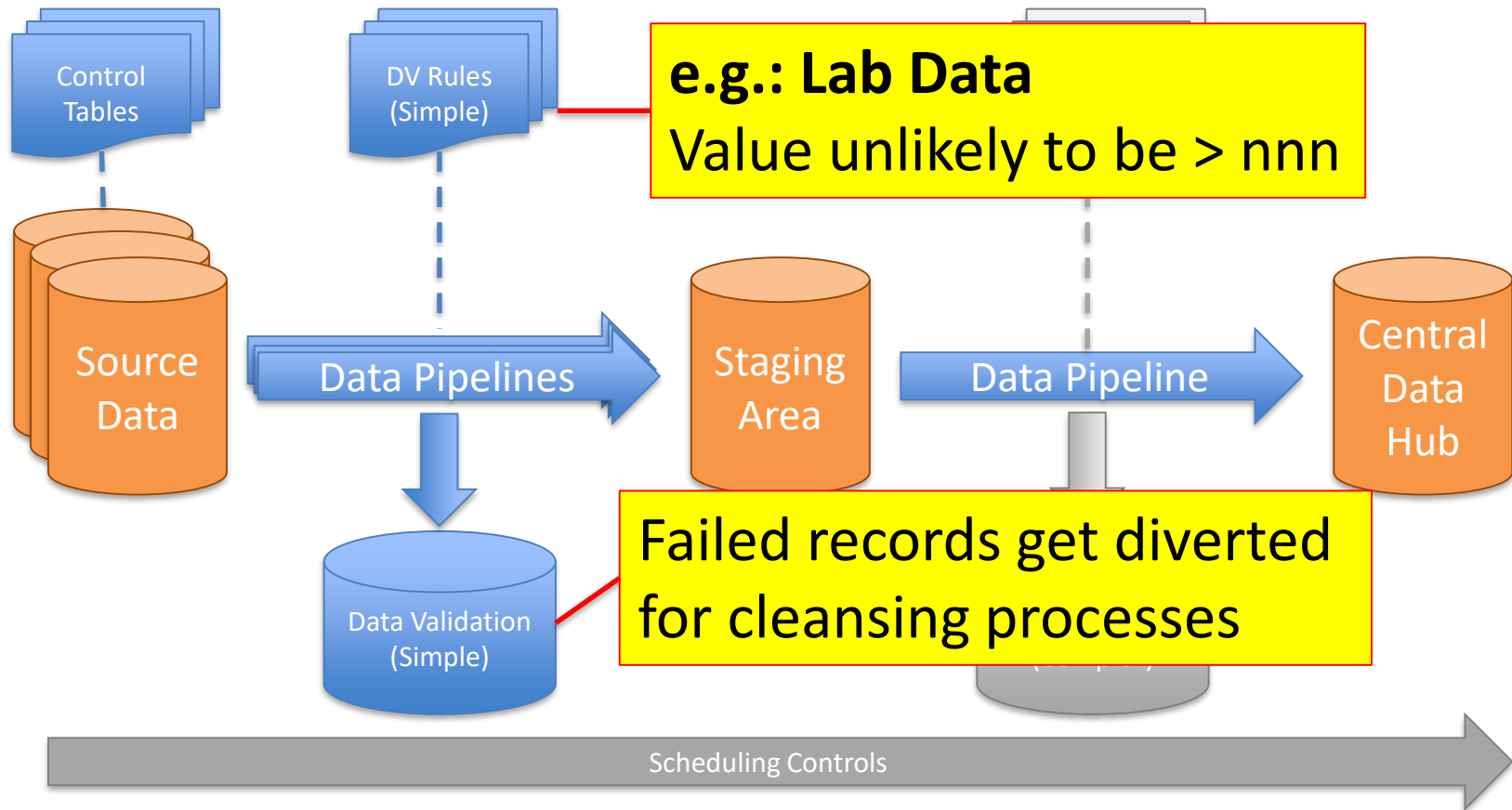
Change Data Capture



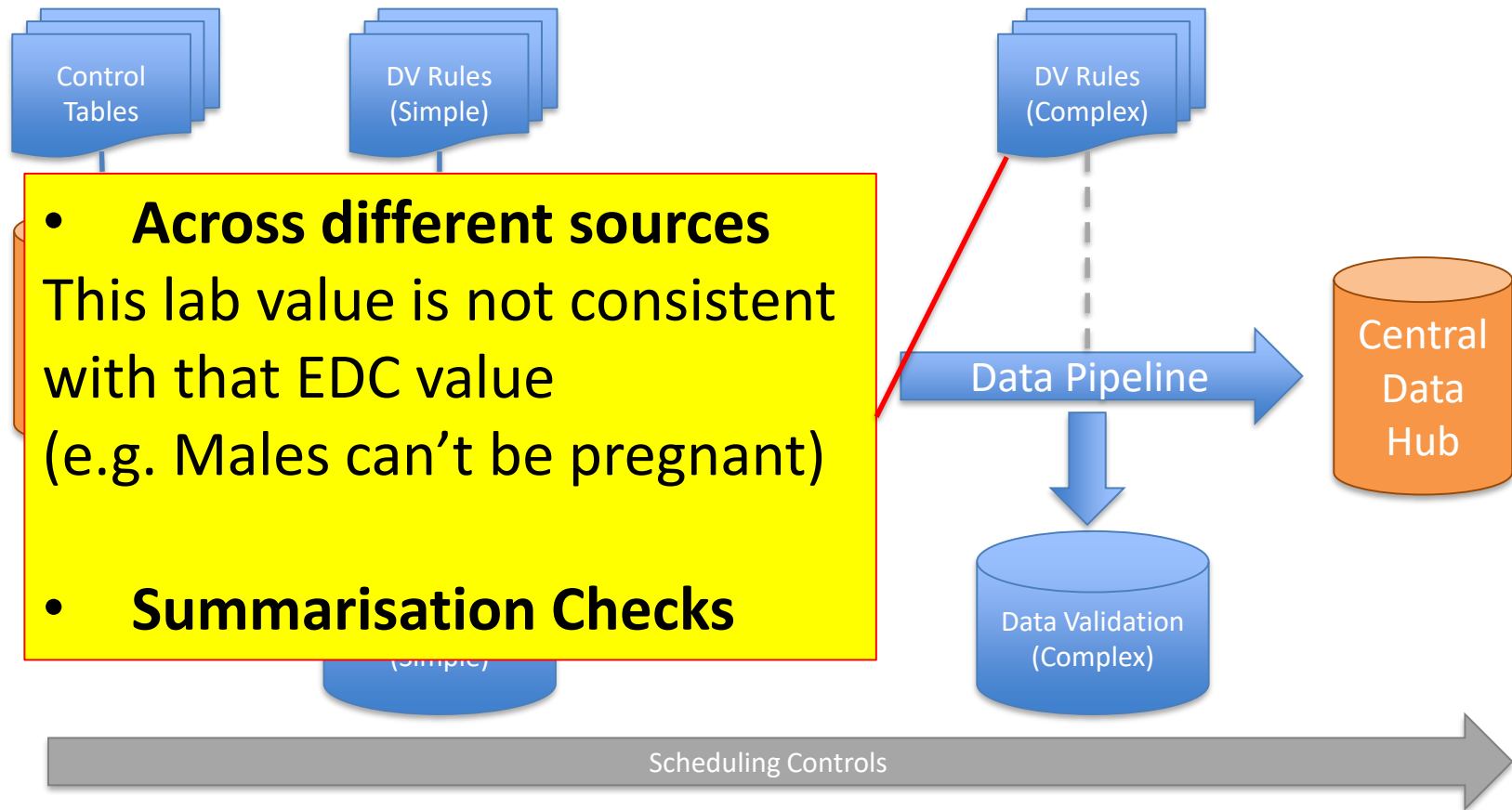
Data Validation (At Source)



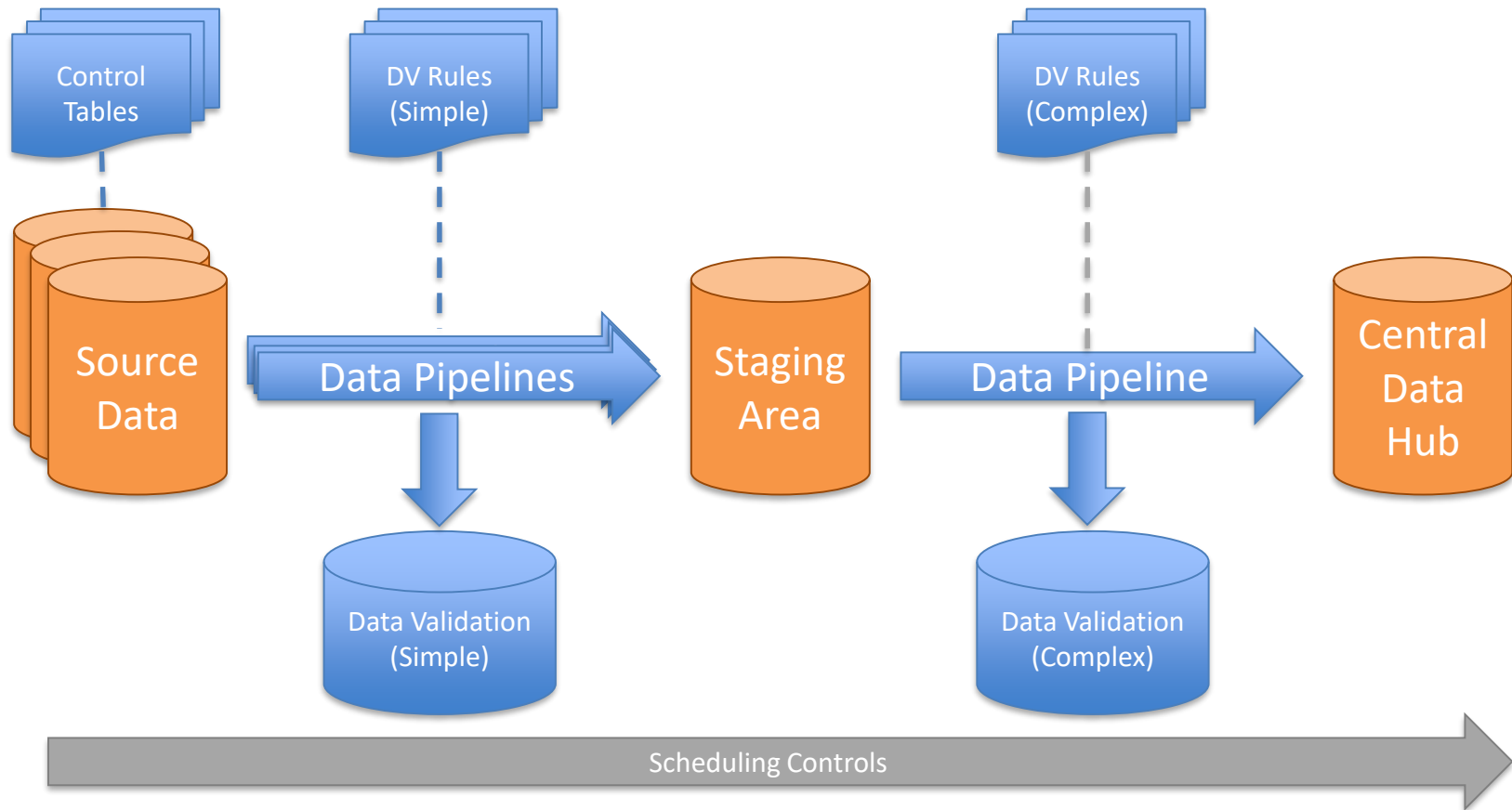
Data Validation (Simple)



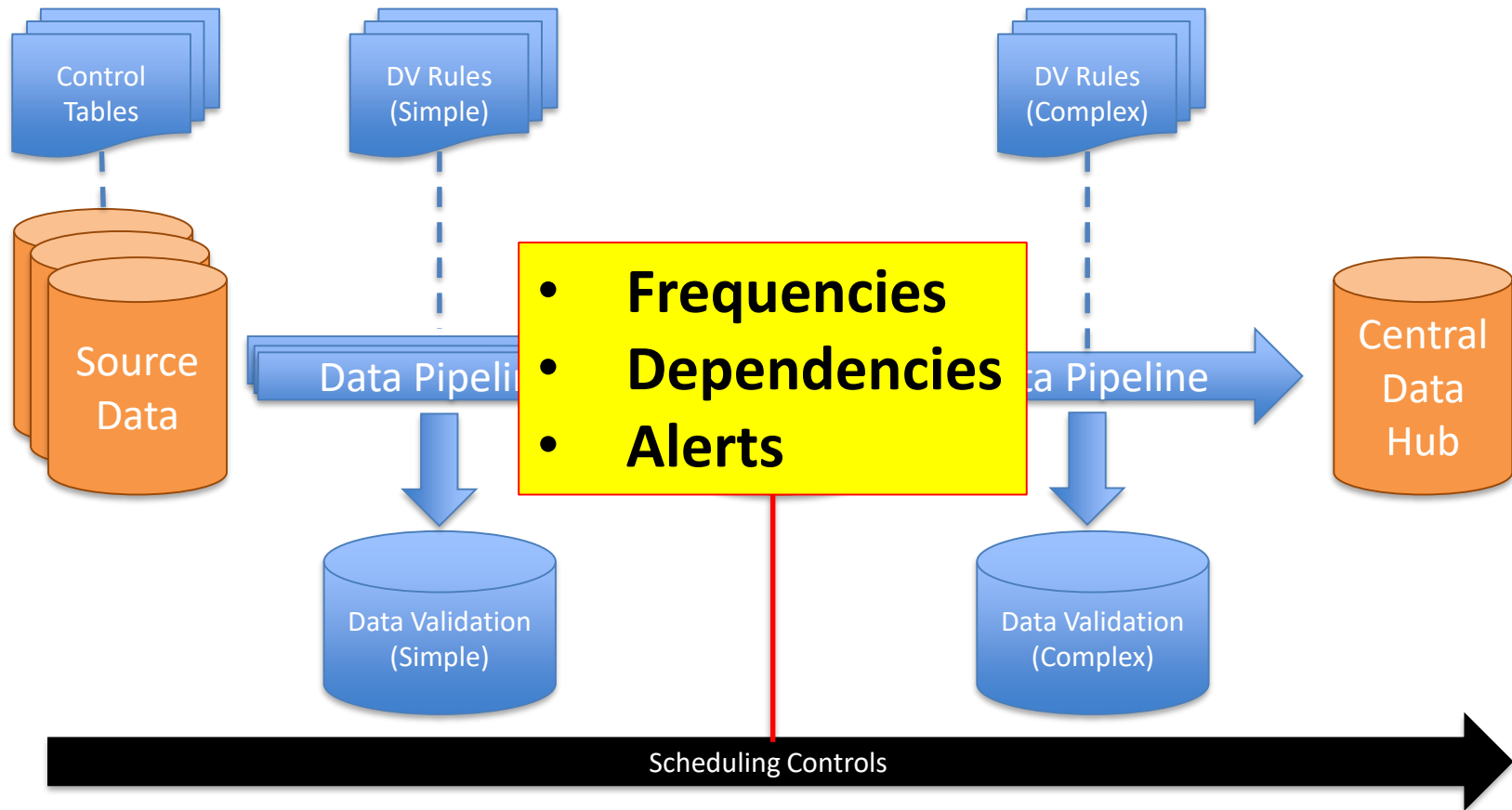
Data Validation (Complex)



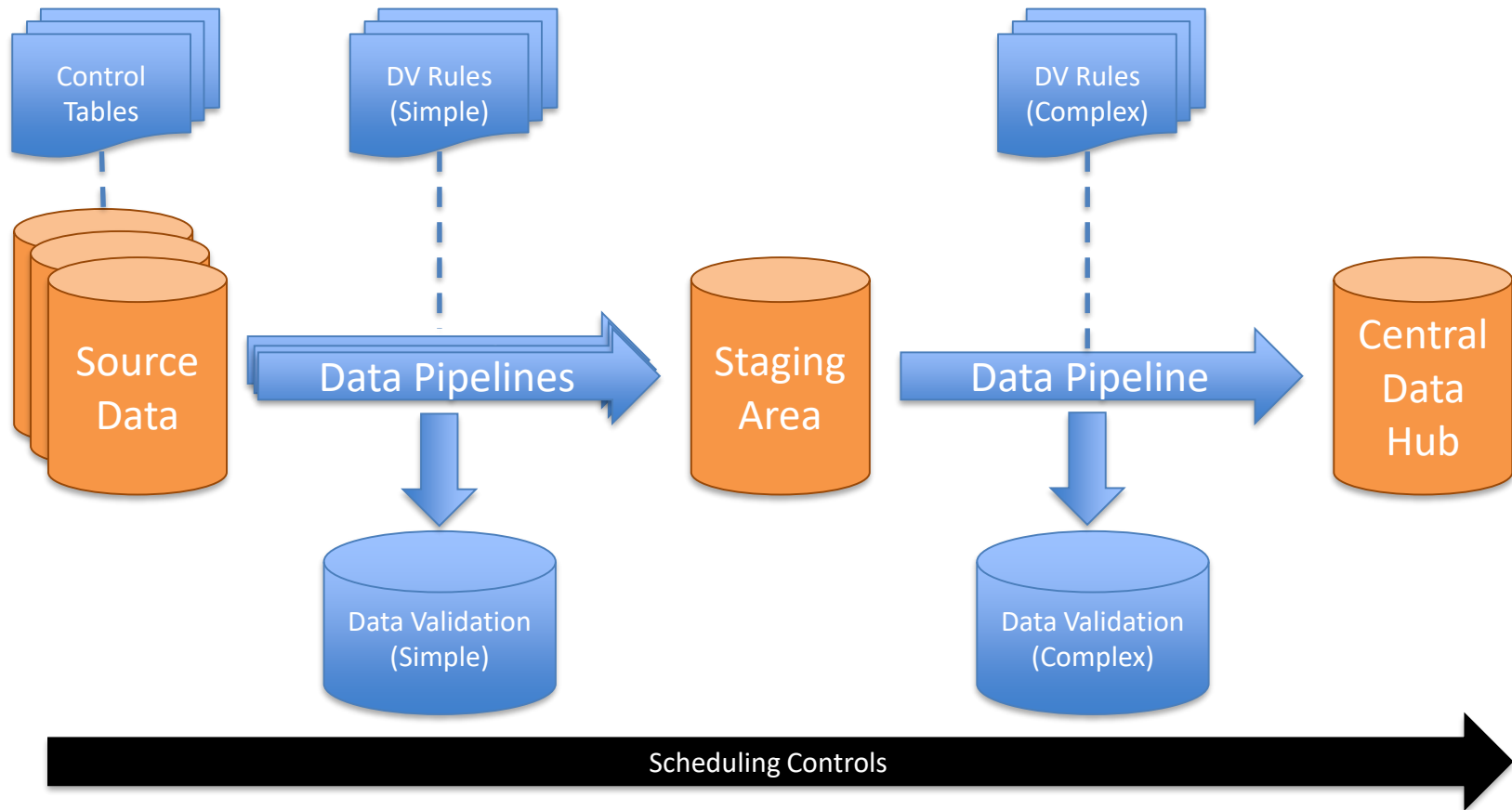
Data Validation (Complex)



Scheduling & Batch Processing



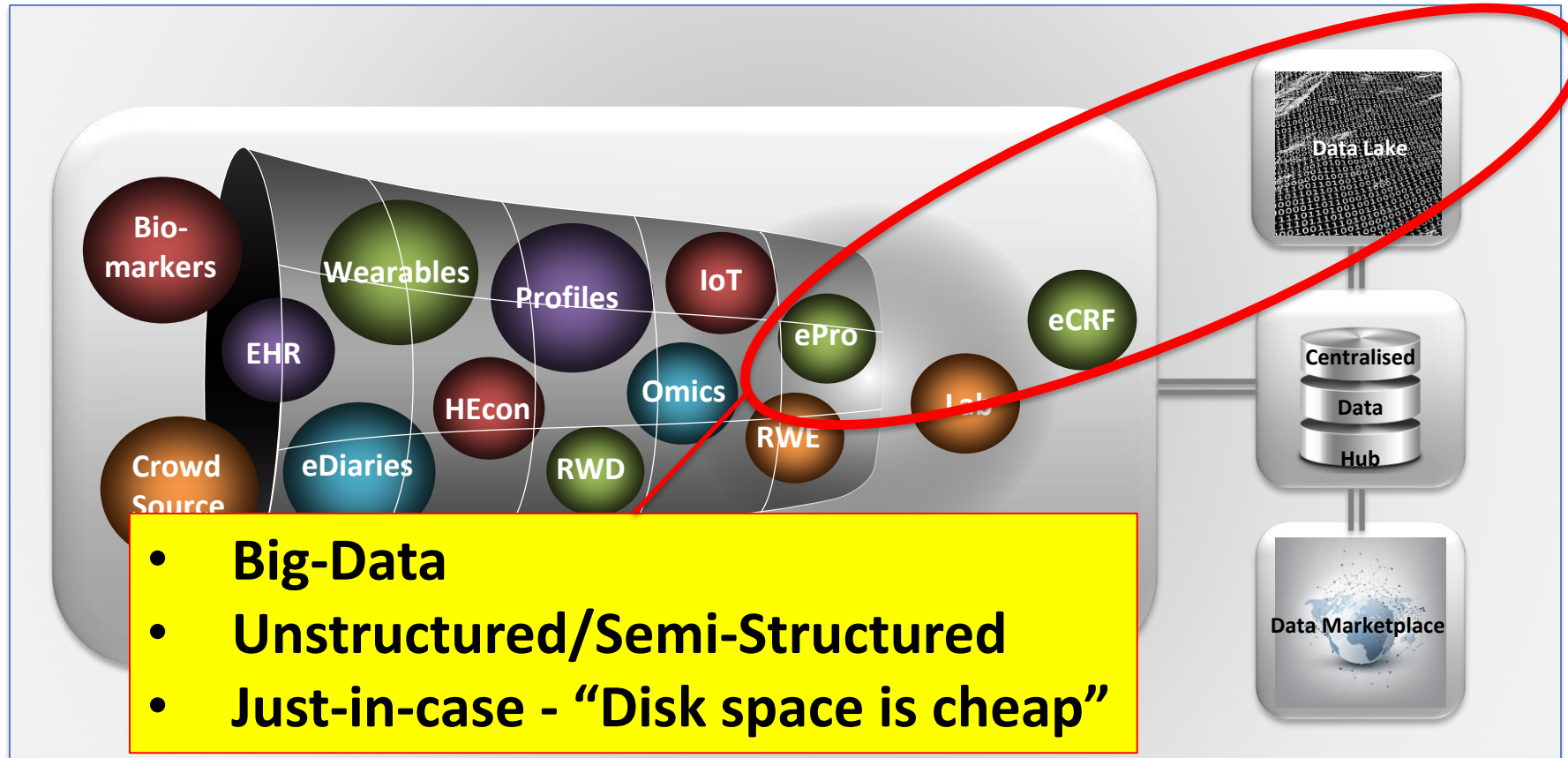
ETL Overview – “Transforming”



Agenda

- Introduction
- Master Data Management - Principle
- ETL Overview – “Transforming”
- **Data Lake – “Trawling”**
- Data Marketplace – “Targeting”
- Conclusion

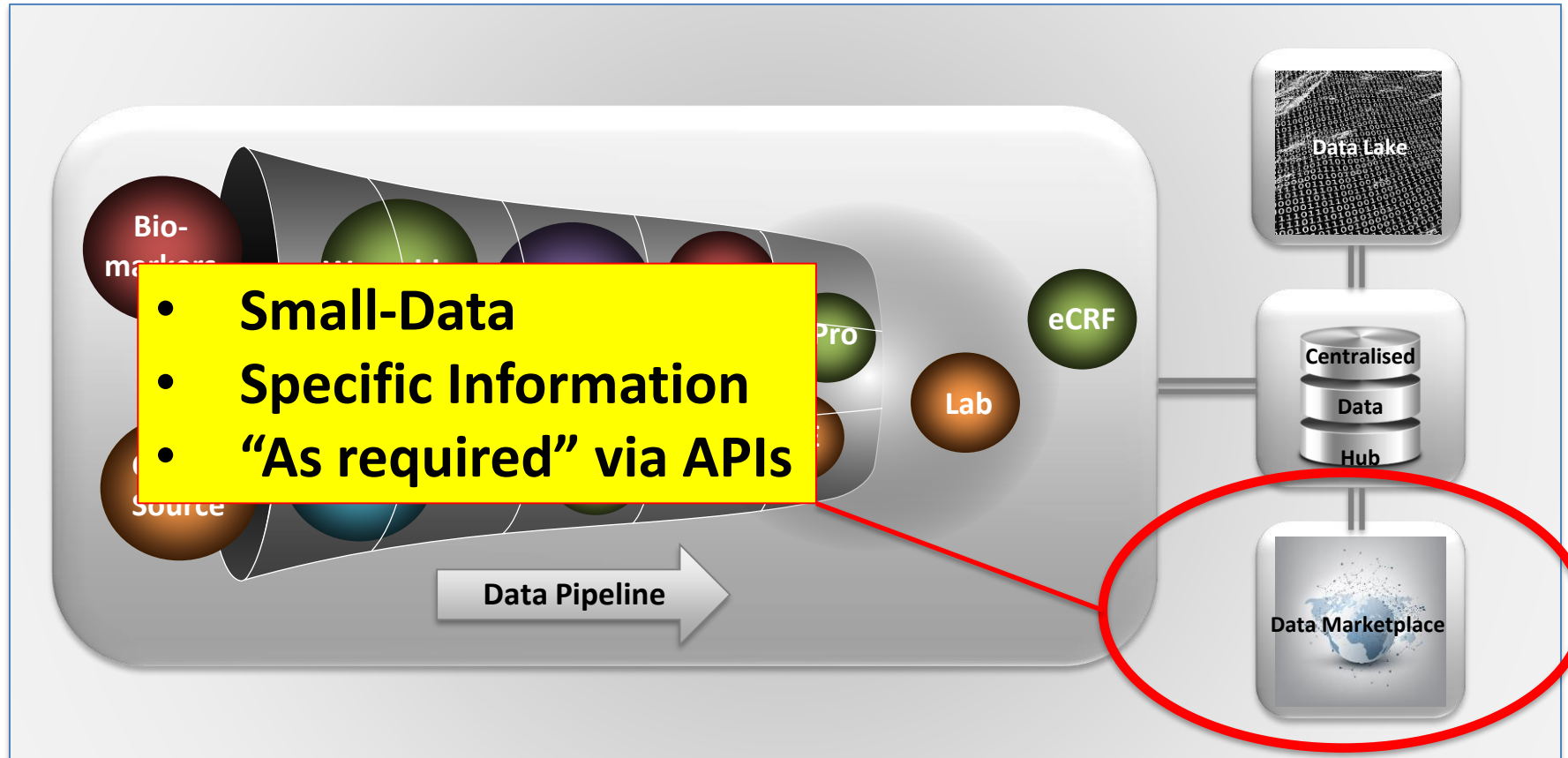
Data Lake – “Trawling”



Agenda

- Introduction
- Master Data Management - Principle
- ETL Overview – “Transforming”
- Data Lake – “Trawling”
- **Data Marketplace – “Targeting”**
- Conclusion

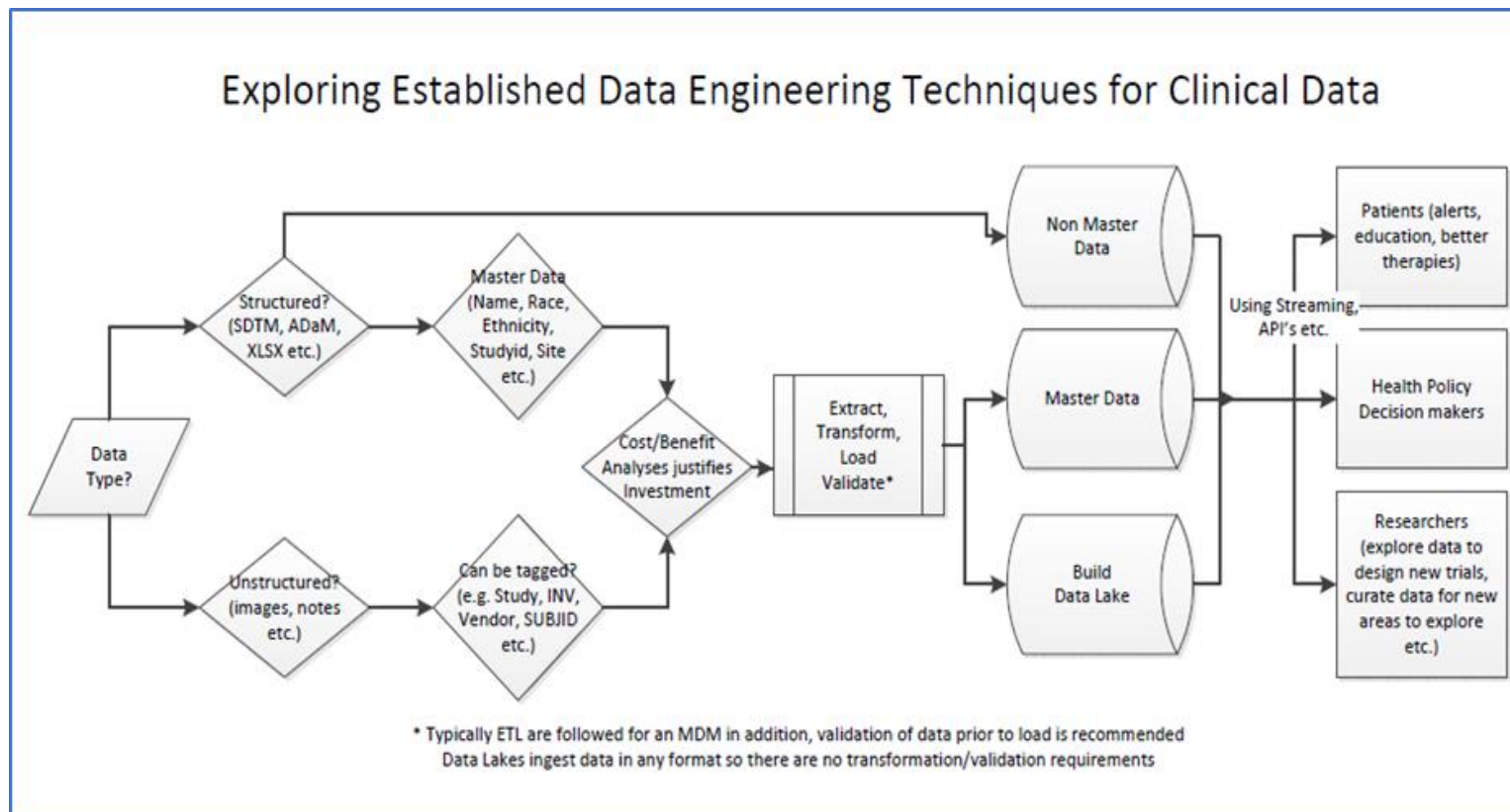
Data Marketplace – “Targeting”



Agenda

- Introduction
- Master Data Management - Principle
- ETL Overview – “Transforming”
- Data Lake – “Trawling”
- Data Marketplace – “Targeting”
- Conclusion

Conclusion



Data Engineering Project (Educating for the Future PHUSE Working Group)



BHayes2@its.jnj.com

Guy.Garrett@AchieveIntelligence.com

<https://education.phuse.eu/eftf/data-engineering/>

No Copyright infringement intended.