Topics in Applied Data Science for Social Scientists

Week 5

The Data Engineering Perspective

February 20th, 2019

Nana Yaw Essuman

Agenda

1. Introduction
2. Overview on Data Engineering
3. Accessibility of Data
4. Data Quality/Monitoring
5. Transforming Data
6. Model Outputs
7. Data - Model Lifecycle
8. Collaboration w/ Data Engineers
9. How to ask the right questions
10. Q&A

Over 7 years of Data Engineering Experience

Over 7 years of Data Engineering Experience

Over 7 years of Data Engineering Experience

Over 7 years of Data Engineering Experience

SIEMENS Healthineers

CSX

COMCAST

Over 7 years of Data Engineering Experience

SIEMENS Healthineers

CSX

COMCAST

UNIVERSAL © A COMCAST COMPANY

Over 7 years of Data Engineering Experience

SIEMENS Healthineers

CSX

COMCAST

UNIVERSAL
A COMCAST COMPANY

FANDANGO

Over 7 years of Data Engineering Experience

SIEMENS Healthineers

CSX

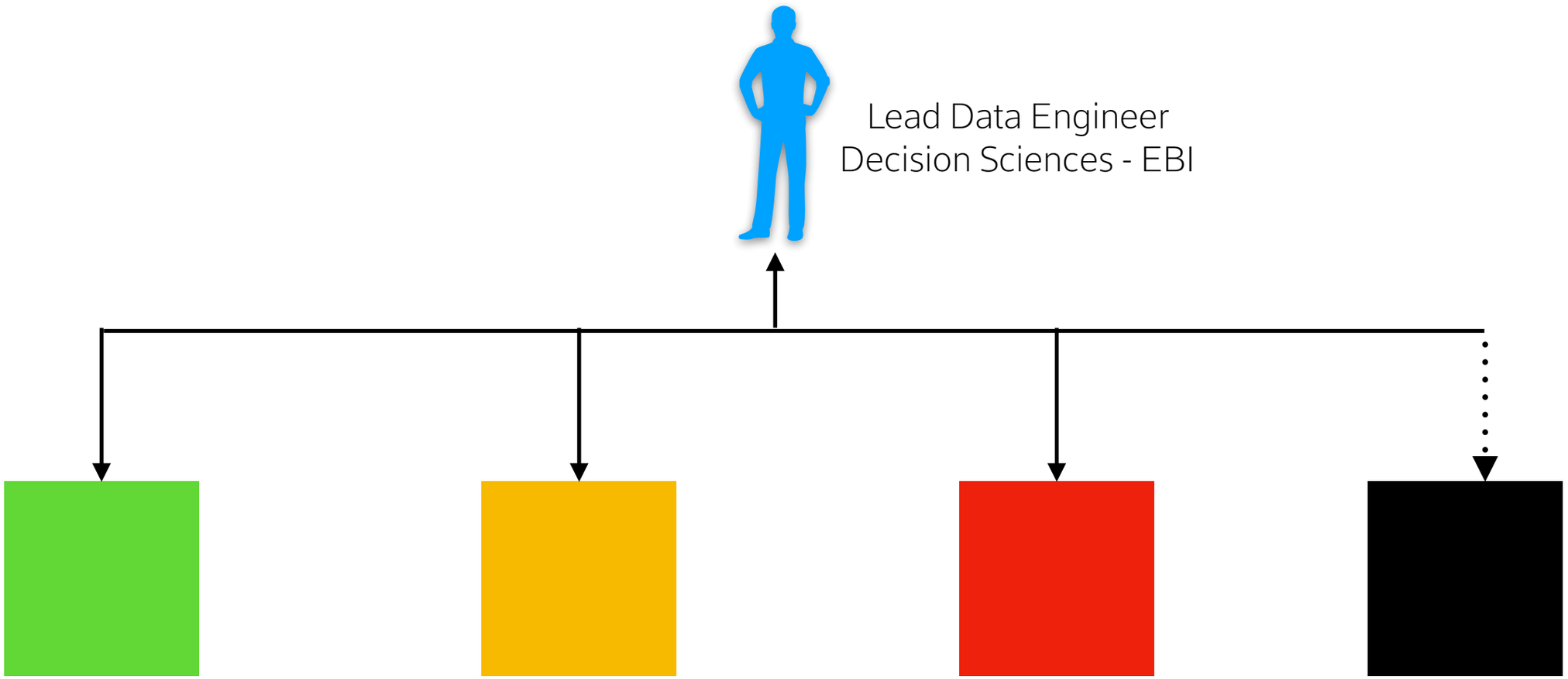
COMCAST

UNIVERSAL ©
A COMCAST COMPANY

FANDANGO

NBCUniversal

9am-5pm



Lead Data Engineer
Decision Sciences - EBI

Linear Forecasting
Data Science Team

Digital Forecasting
Data Science Team

Viewer Behavior & R&D
Data Science Team

Corporate Analytics
Center of Excellence

What is Data Engineering?

https://www.lynda.com/Apache-Spark-tutorials/What-data-engineering/578078/638978-4.html

What is Data Engineering?

Data engineering focuses on data, its curation, movement, storage, security, and processing. Data engineering is performed by data engineers. They work on building pipelines, applications, APIs, and systems that produce, process, and consume data to meet business needs. To compare; data engineering converts raw data into knowledge data. Data Scientists deals with using this data produced by data engineers to generate insights & to predict the future using data from the past.

https://www.lynda.com/Apache-Spark-tutorials/What-data-engineering/578078/638978-4.html

Accessibility of Data

Accessibility of Data

Question: A Data Scientist has been tasked to predict the likelihood Show A will perform during a particular time slot on a network compared to the current time slot proposed by TV Network Executives. What should be the Data Scientists first worry?

Accessibility of Data

Question: A Data Scientist has been tasked to predict the likelihood Show A will perform during a particular time slot on a network compared to the current time slot proposed by TV Network Executives. What should be the Data Scientists first worry?

1. Where do I get my data from?
2. Is the data structured or unstructured?
3. How do I pull the data? One time or Continuous
4. Where do I store the data?
5. What is the right way to store the data?
6. How do I create various subsets of the data?
7. How do I access the data for my models?

The Data

Structured

Unstructured

The Data

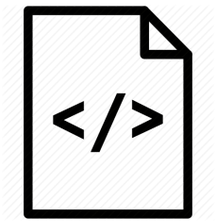
Structured

Unstructured

	Structured Data	Unstructured Data
Characteristics	• Pre-defined data models • Usually text only • Easy to search	• No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	• Relational databases • Data warehouses	• Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	• Airline reservation systems • Inventory control • CRM systems • ERP systems	• Word processing • Presentation software • Email clients • Tools for viewing or editing media
Examples	• Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information	• Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery

Extract-Transform-Load (ETL)

Extract

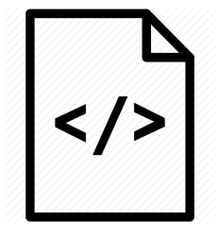


Transform

Load

Extract-Transform-Load (ETL)

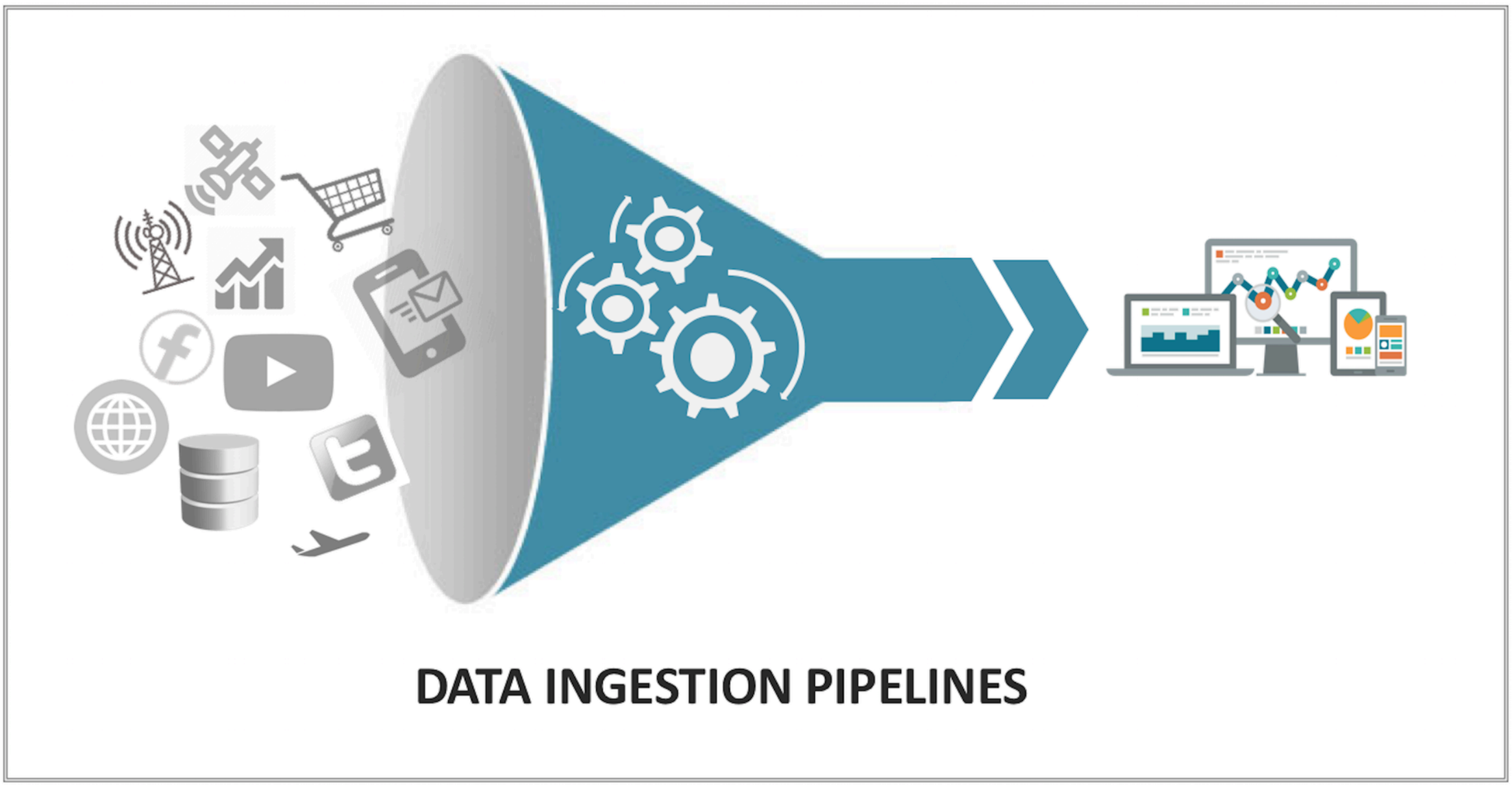
Extract



Transform

Load

Automate & Scale



https://community.siftdev.com/wp-content/uploads/2018/02/UIP.png

Data Quality & Monitoring

Effective data quality maintenance requires periodic data monitoring and cleaning. In general, data quality maintenance involves updating/standardizing data and deduplicating records to create a single data view.

https://www.techopedia.com/definition/14653/data-quality

Data Quality & Monitoring

Effective data quality maintenance requires periodic data monitoring and cleaning. In general, data quality maintenance involves updating/standardizing data and deduplicating records to create a single data view.

Key data quality components are as follows:

https://www.techopedia.com/definition/14653/data-quality

Data Quality & Monitoring

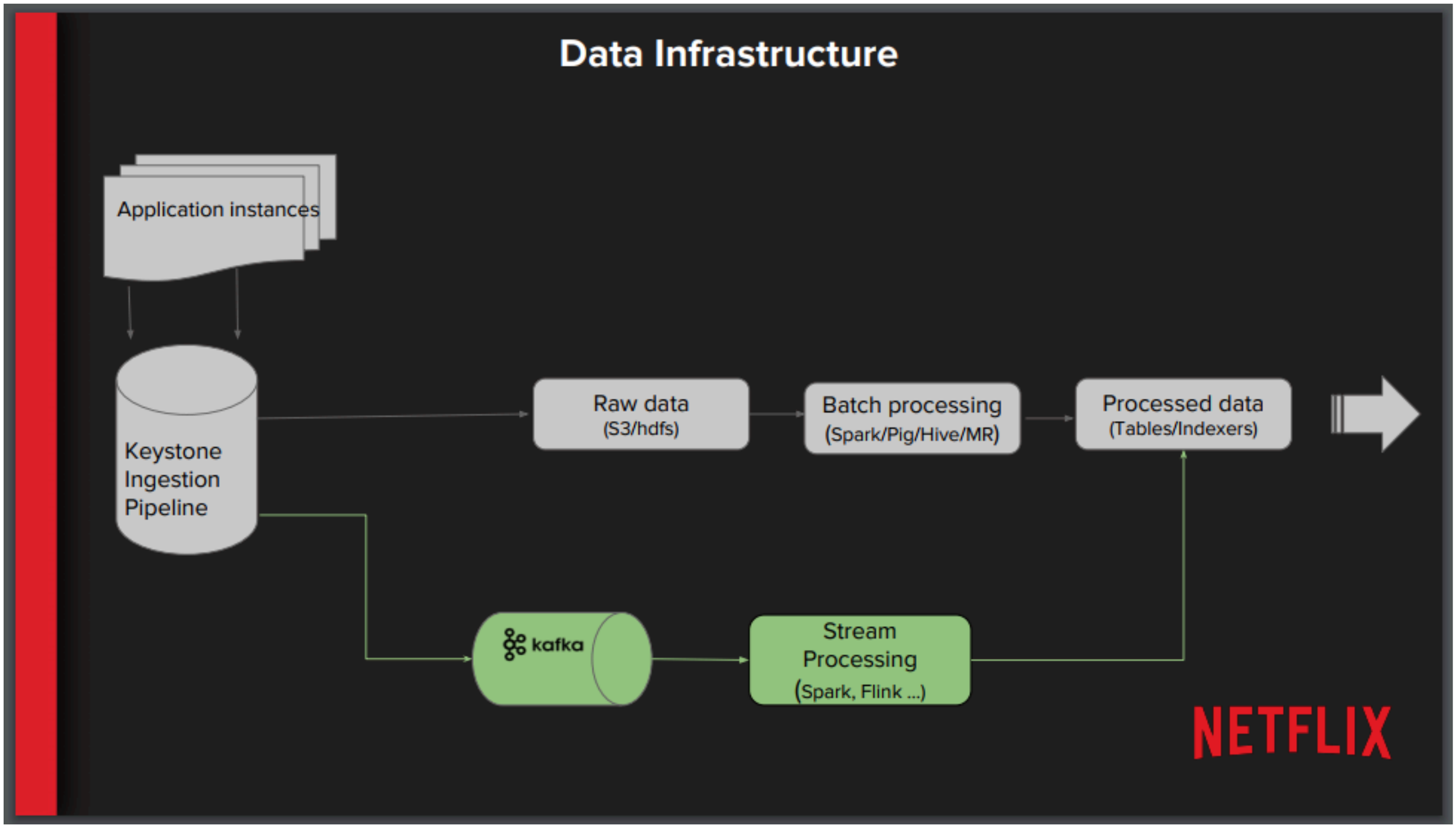
Effective data quality maintenance requires periodic data monitoring and cleaning. In general, data quality maintenance involves updating/standardizing data and deduplicating records to create a single data view.

Key data quality components are as follows:

- Completeness: Level at which desired data attributes are supplied. Data does not need to be 100 percent complete
- Accuracy: Represents data's real world status. May be calculated by using an automated method with the help of various lists and mapping
- Credibility: Extent to which data is considered credible and true. May differ from by source
- Timeliness (age of data): Extent to which data is adequately updated for a current venture
- Consistency: Assesses whether various dataset facts match
- Integrity: Assesses reference validity and accurate joining of various datasets

https://www.techopedia.com/definition/14653/data-quality

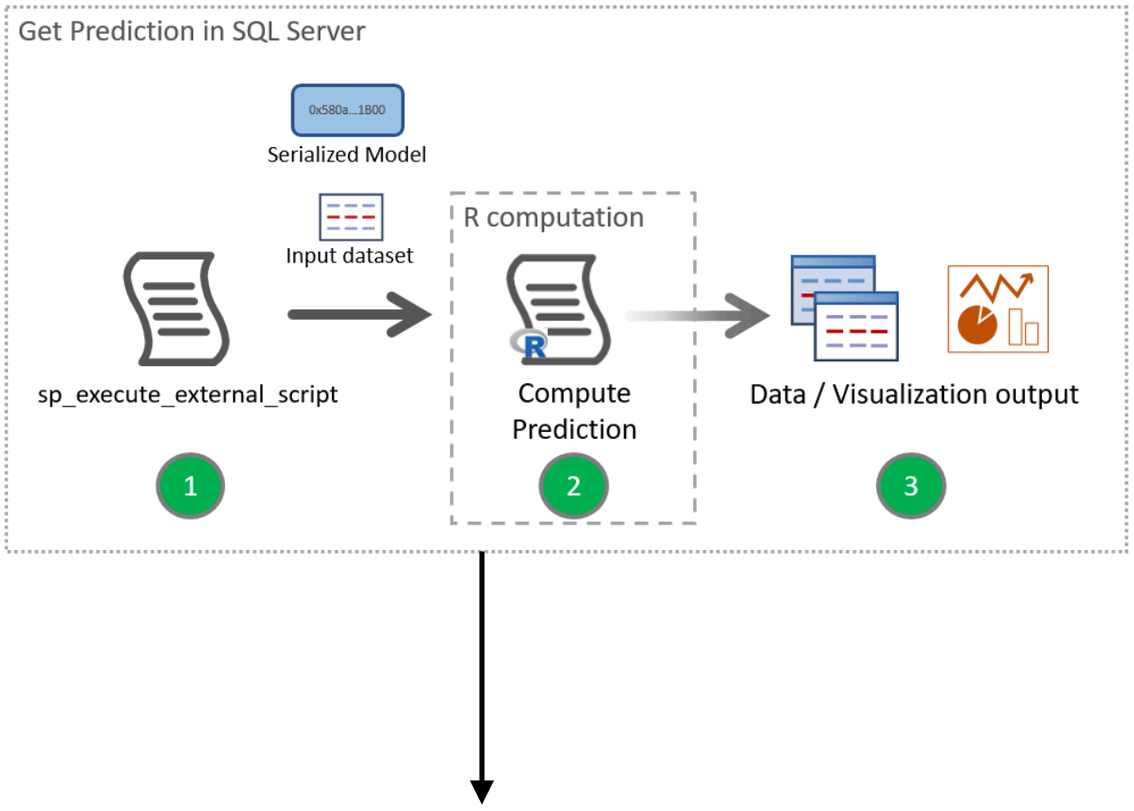
Transforming your Data



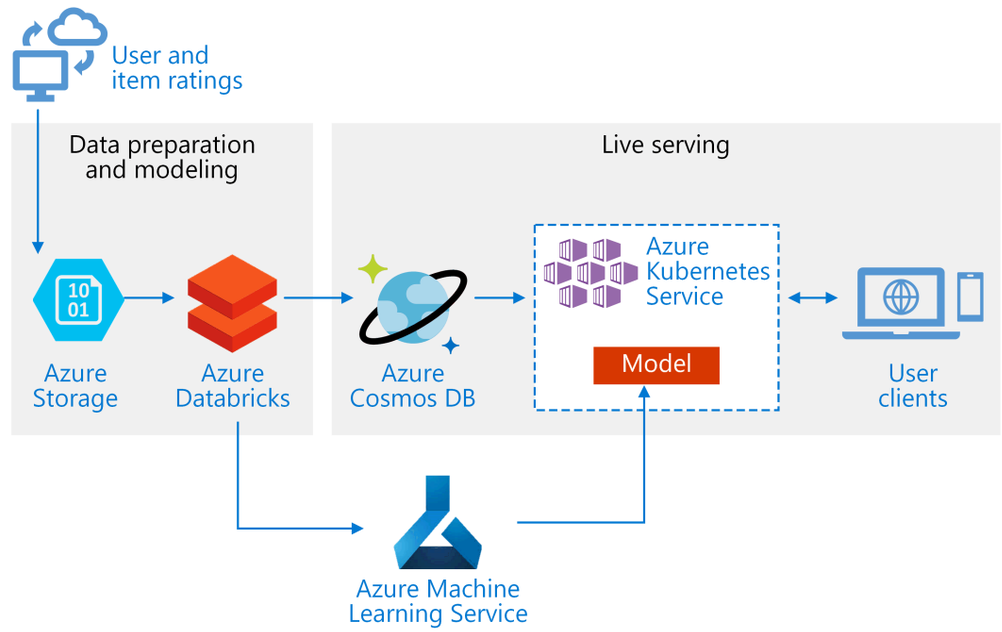
Source: Netflix Data Infrastructure Blog

Operationalize Model Outputs

Batch

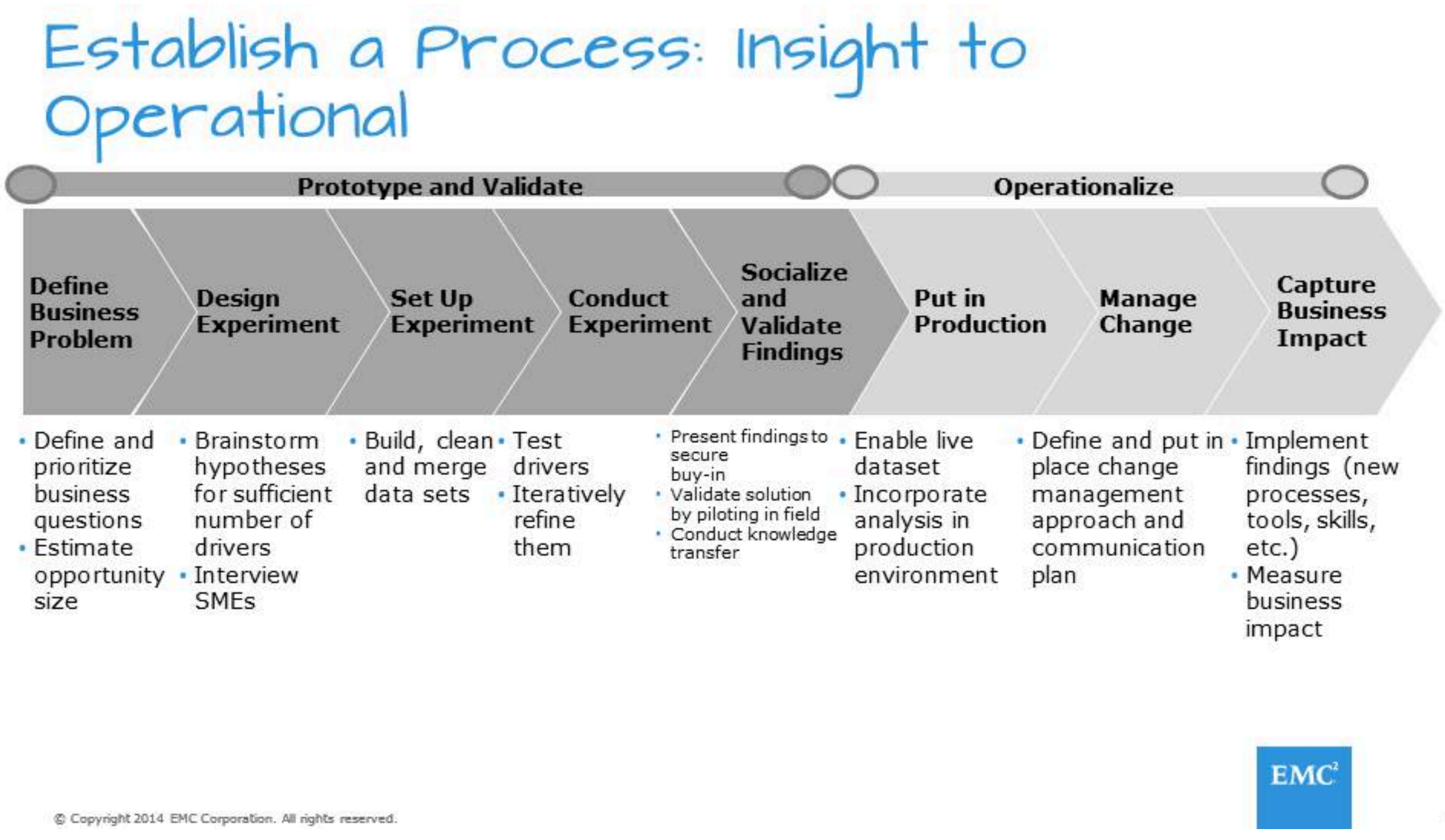


Real-Time



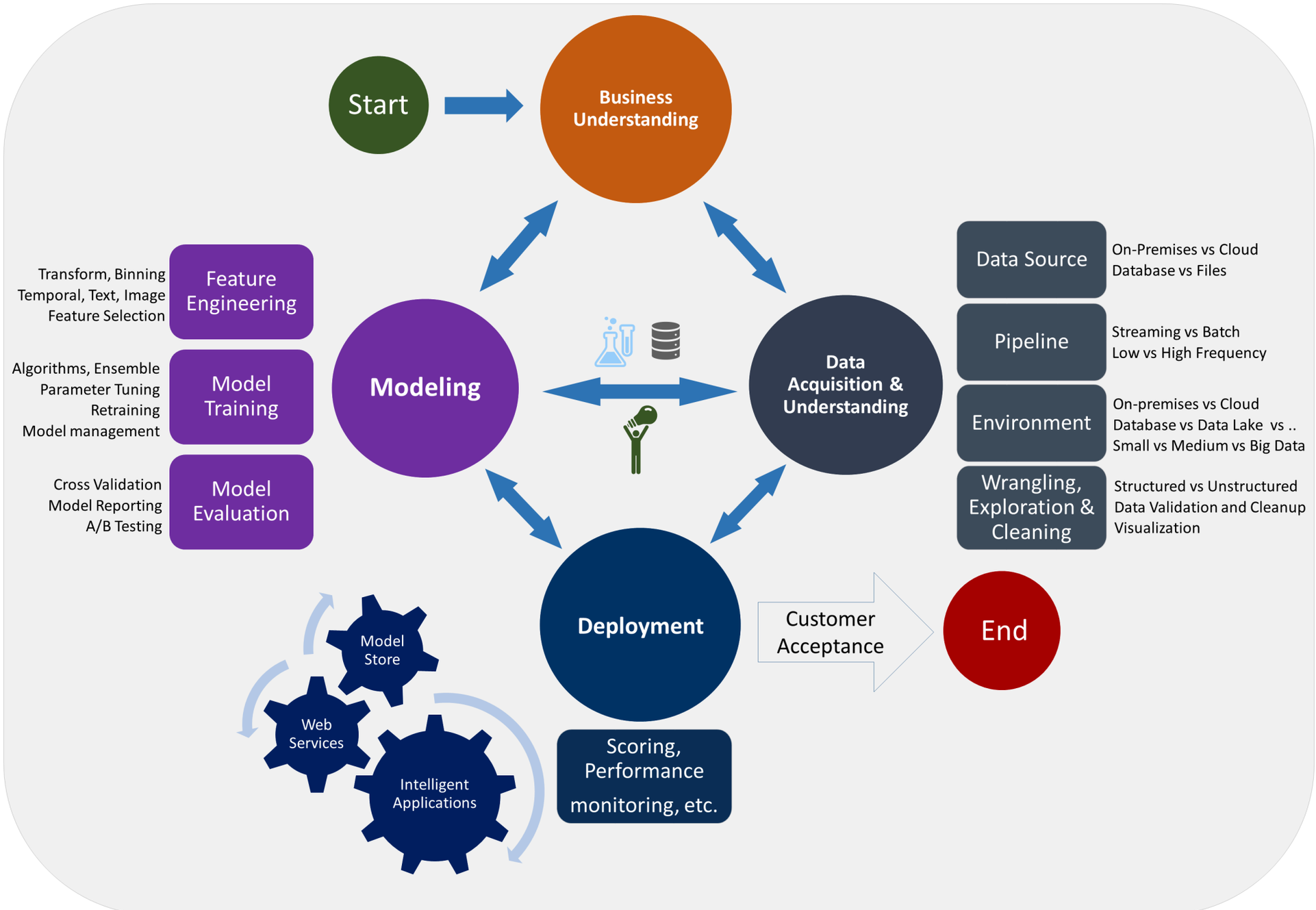
Source: Netflix Data Infrastructure Blog

Data - Model Lifecycle



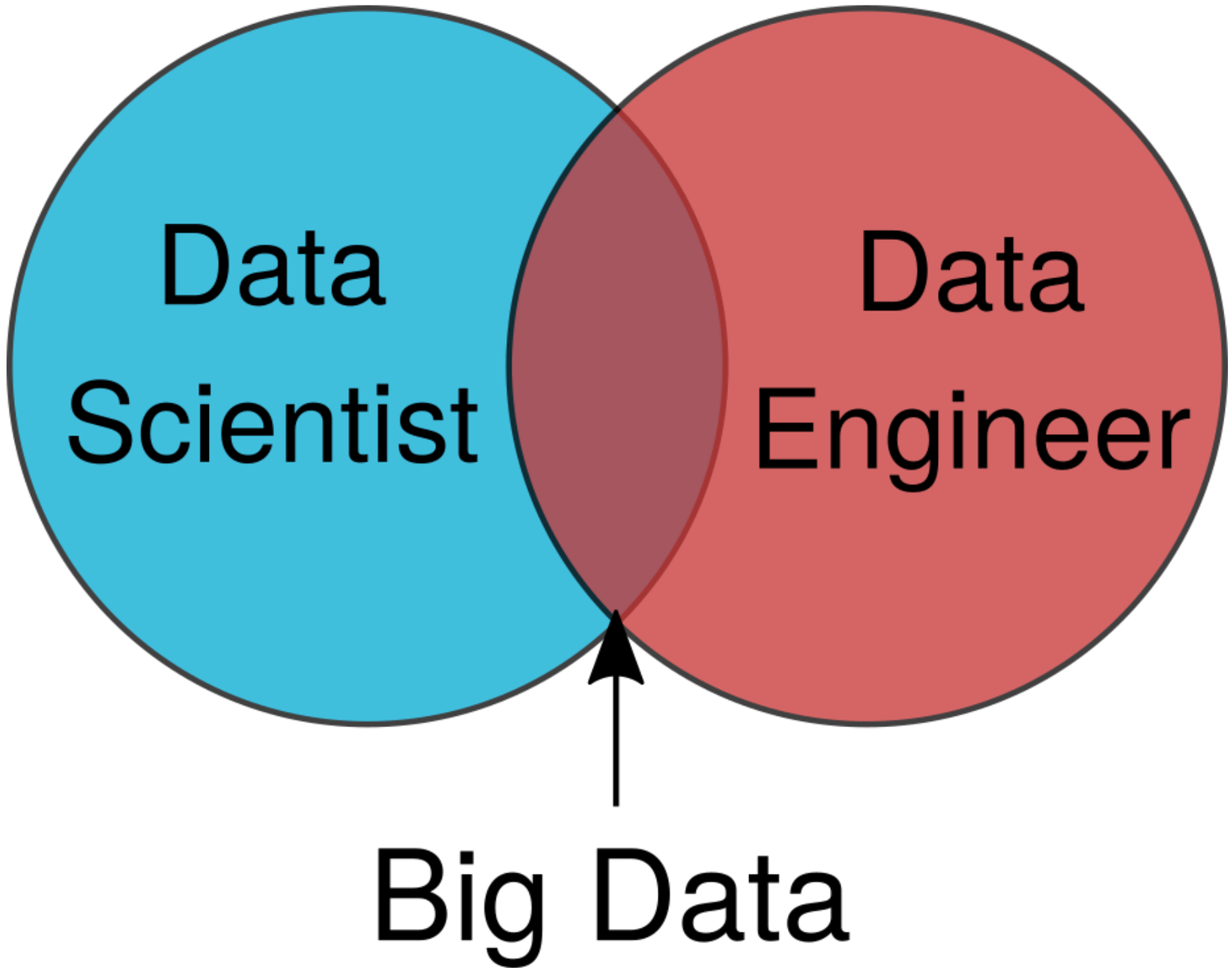
https://infocus.dellemc.com/wp-content/uploads/2014/12/Slide2.jpg

Data - Model Lifecycle



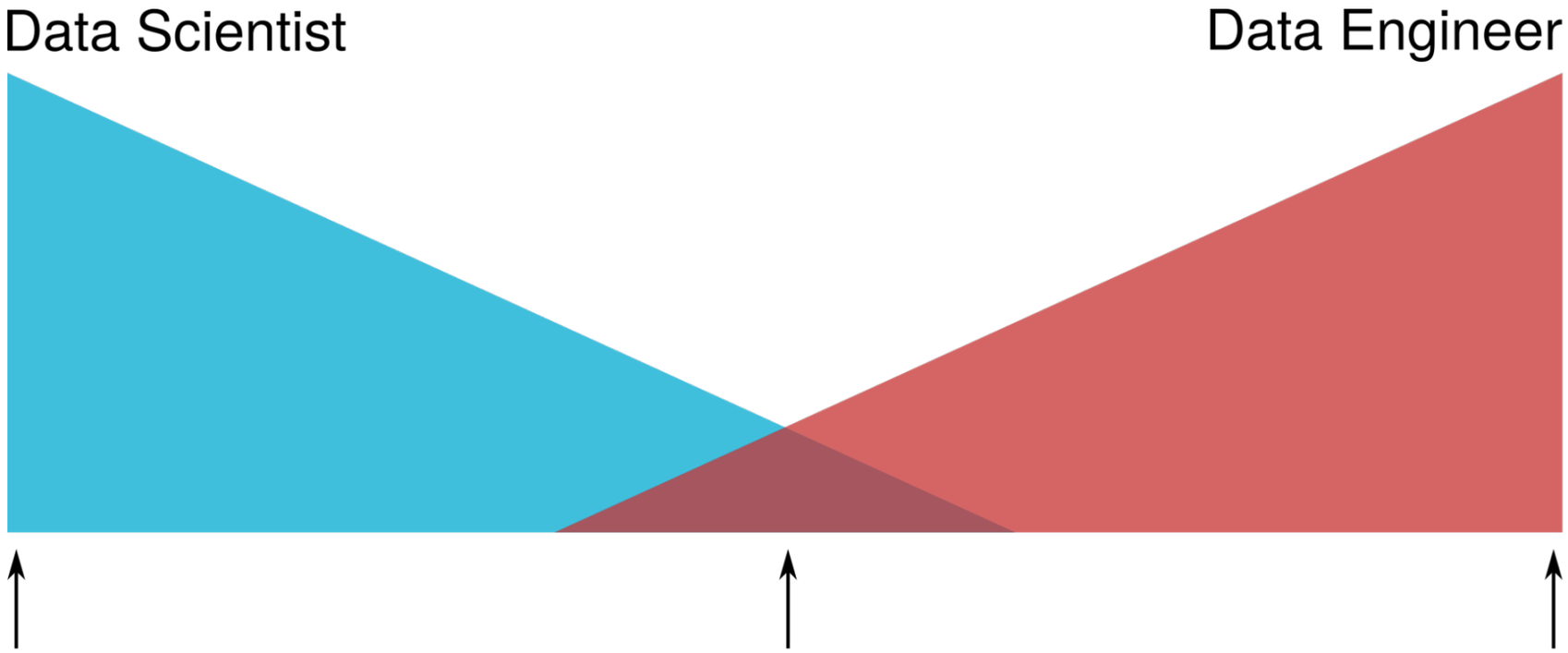
https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/media/lifecycle/tdsp-lifecycle2.png

Collaboration w/ Data Engineers



https://www.oreilly.com/ideas/data-engineers-vs-data-scientists

Collaboration w/ Data Engineers



Core Comptencies	Overlapping Skills	Core Comptencies
Adv. Math/Statistics	Analysis	Adv. Programming
ML/AI	Programming	Distributed Sys.
Adv. Analytics	Big Data	Data Pipelines

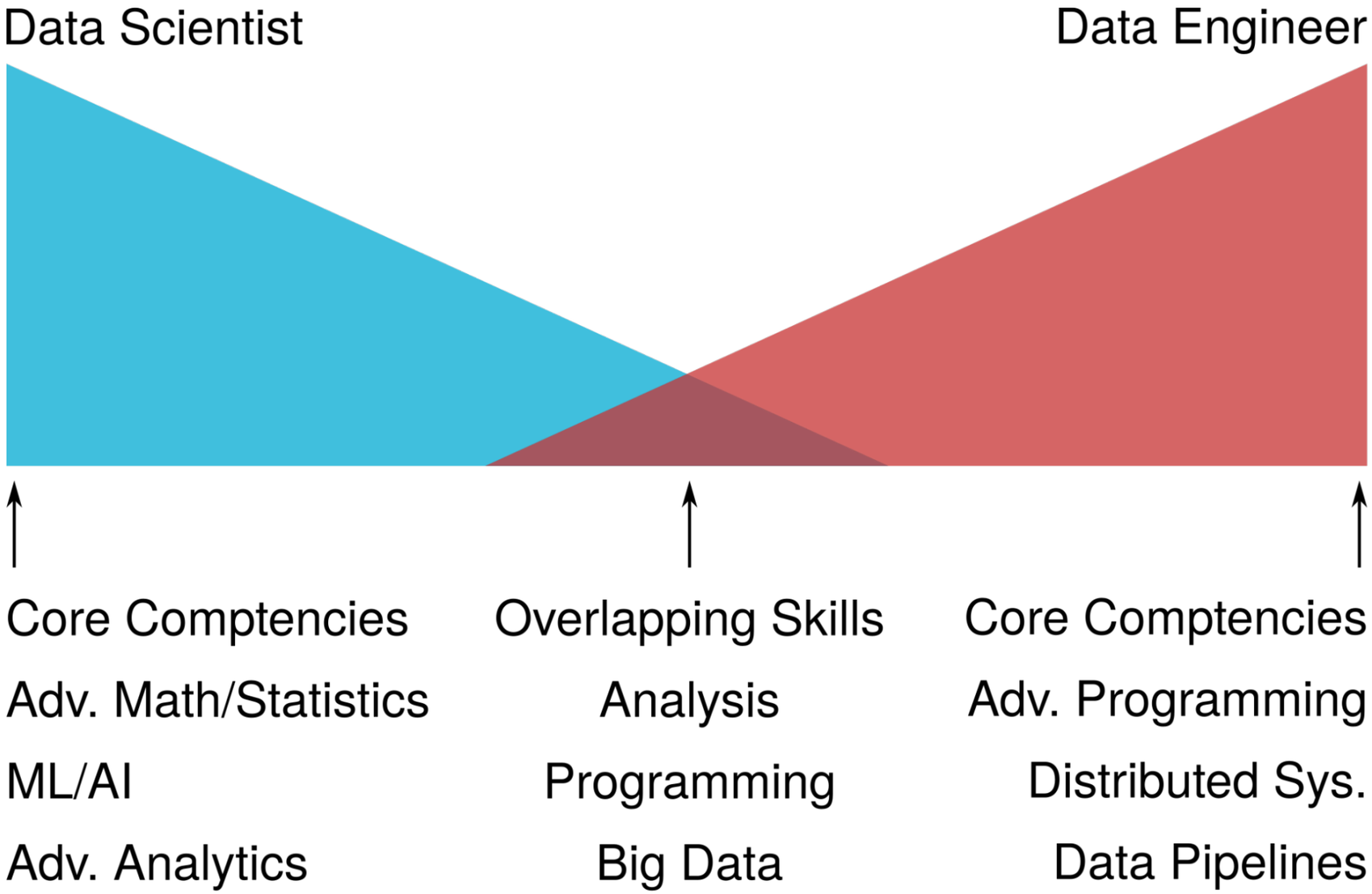
BDI

BIG DATA INSTITUTE

For more information go to http://bigdatainstitute.io

https://www.oreilly.com/ideas/data-engineers-vs-data-scientists

Collaboration w/ Data Engineers



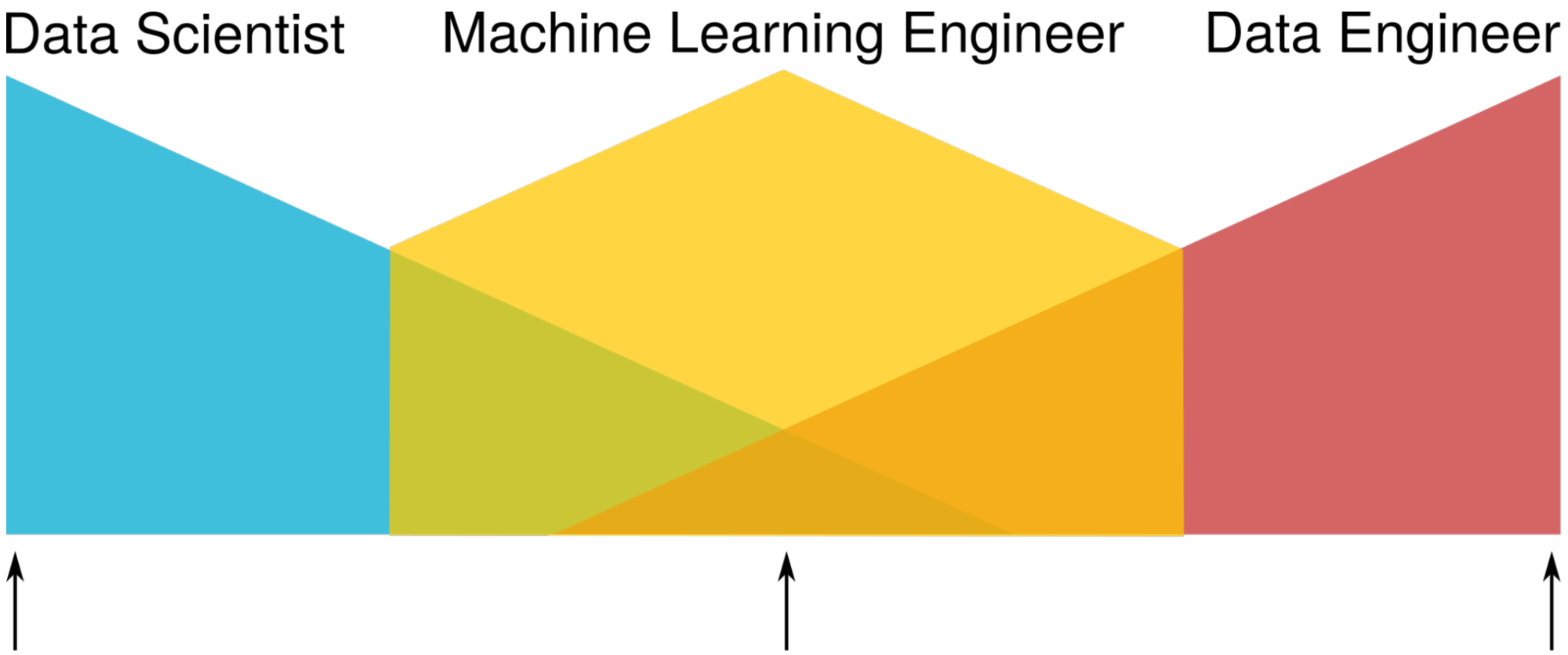
Varies based on size of organization

BDI
BIG DATA INSTITUTE

For more information go to http://bigdatainstitute.io

https://www.oreilly.com/ideas/data-engineers-vs-data-scientists

Collaboration w/ Data Engineers



BDI

BIG DATA INSTITUTE

For more information go to http://bigdatainstitute.io

https://www.oreilly.com/ideas/data-engineers-vs-data-scientists

What to ask your Data Engineer or yourself

https://www.oreilly.com/ideas/data-engineers-vs-data-scientists

What to ask your Data Engineer or yourself

1. Where do I get my data from?
2. Is the data structured or unstructured?
3. How do I pull the data? One time or Continuous
4. Where do I store the data?
5. What is the right way to store the data?
6. How do I create various subsets of the data?
7. How do I access the data for my models?
8. How do I deploy this model to production? Real-Time or Batch
9. How do I monitor my model metrics?

https://www.oreilly.com/ideas/data-engineers-vs-data-scientists

Resources

https://www.oreilly.com/ideas/data-engineers-vs-data-scientists

https://www.datamation.com/big-data/structured-vs-unstructured-data.html

https://www.infoworld.com/article/3250852/data-science/handing-off-models-from-data-science-to-it.html

https://www.datascience.com/blog/predictive-data-models-from-data-cleaning-to-model-deployment

Contact

nanayaw.essuman@nbcuni.com
nyessuman@gmail.com

https://www.oreilly.com/ideas/data-engineers-vs-data-scientists