DataCon 2019

Managing Data Effectively – Maintaining the Quality & Keeping the Lake Clean

Junior Muka
Data Architect / Data Management Educator
@juniormuka1

Disclaimer

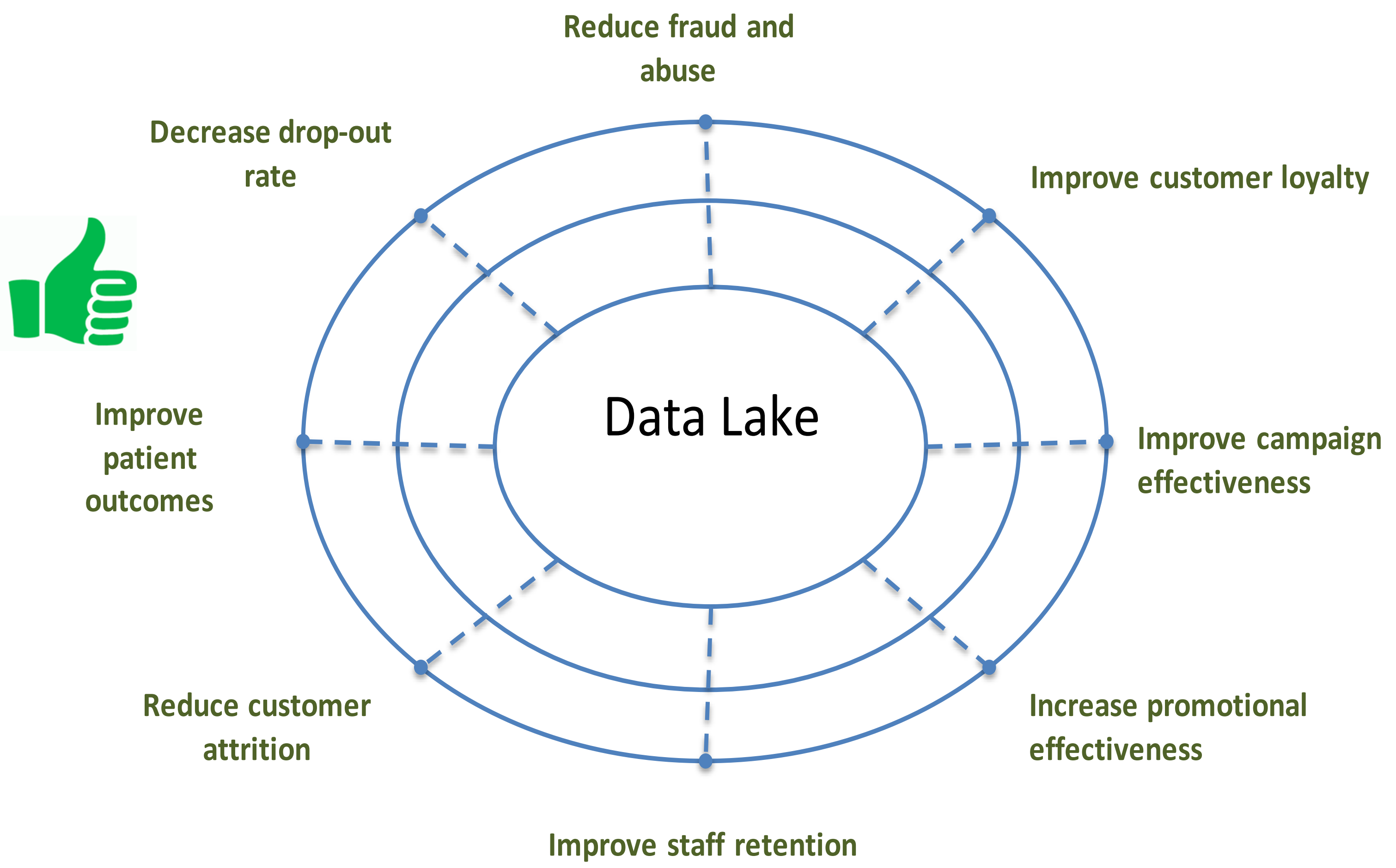
The views, thoughts and opinions expressed in this presentation are the author's own and do not necessarily reflect those of his employer.

Understanding Data Lakes

	Traditional Data warehouse	Data Lake
Content Variety	Designed to store and process structured content	Designed to store and process content in a wide variety of states (including multistructured, unstructured, and structured content)
Data Structure	Database structure is defined upfront. Physical database is modelled and defined prior to transforming and loading data into it (Schema-on-Write)	Users can access and structure data at comsunption time ('Schema on Read' or 'late-binding execution')
Data Quality	Usually extensive quality testing built in ETL process	Focus on fast ingestion means quality is tested at the time data is accessed for analysis
Effort	Significantly higher effort required due to upfront modelling; long period of time to design and build integration requirement	Significantly easier to implement due to deferment of data modelling until users need to analyse the data

Keeping the Lake clean

Tip 1: Ingest data in the lake, one use case / analytics outcome at a time.



Reduce fraud and abuse

Decrease drop-out rate

Improve customer loyalty

Data Lake

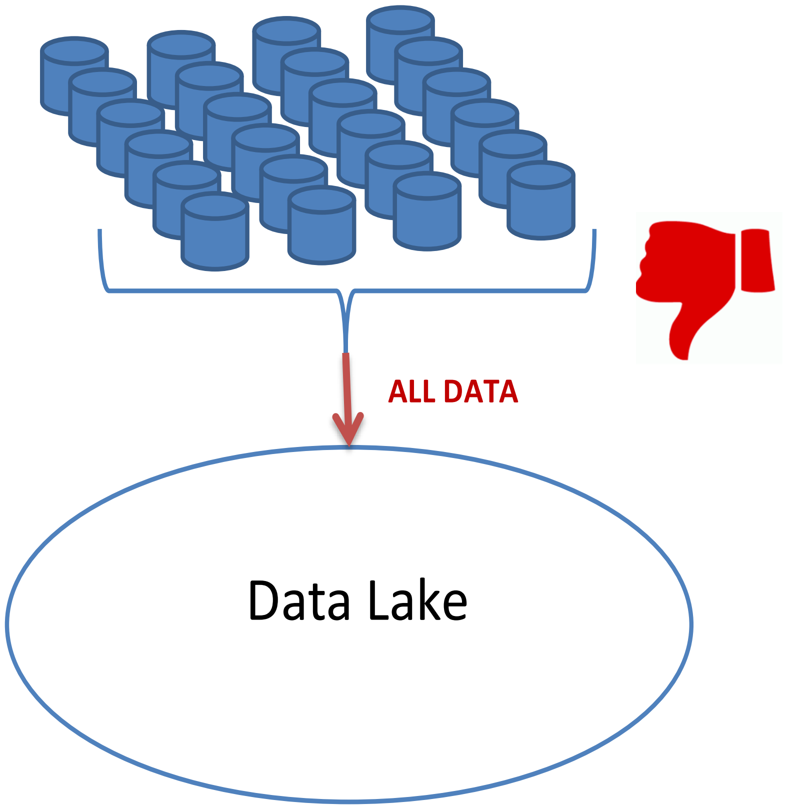
Improve patient outcomes

Improve campaign effectiveness

Reduce customer attrition

Increase promotional effectiveness

Improve staff retention



ALL DATA

Data Lake

@juniormuka1

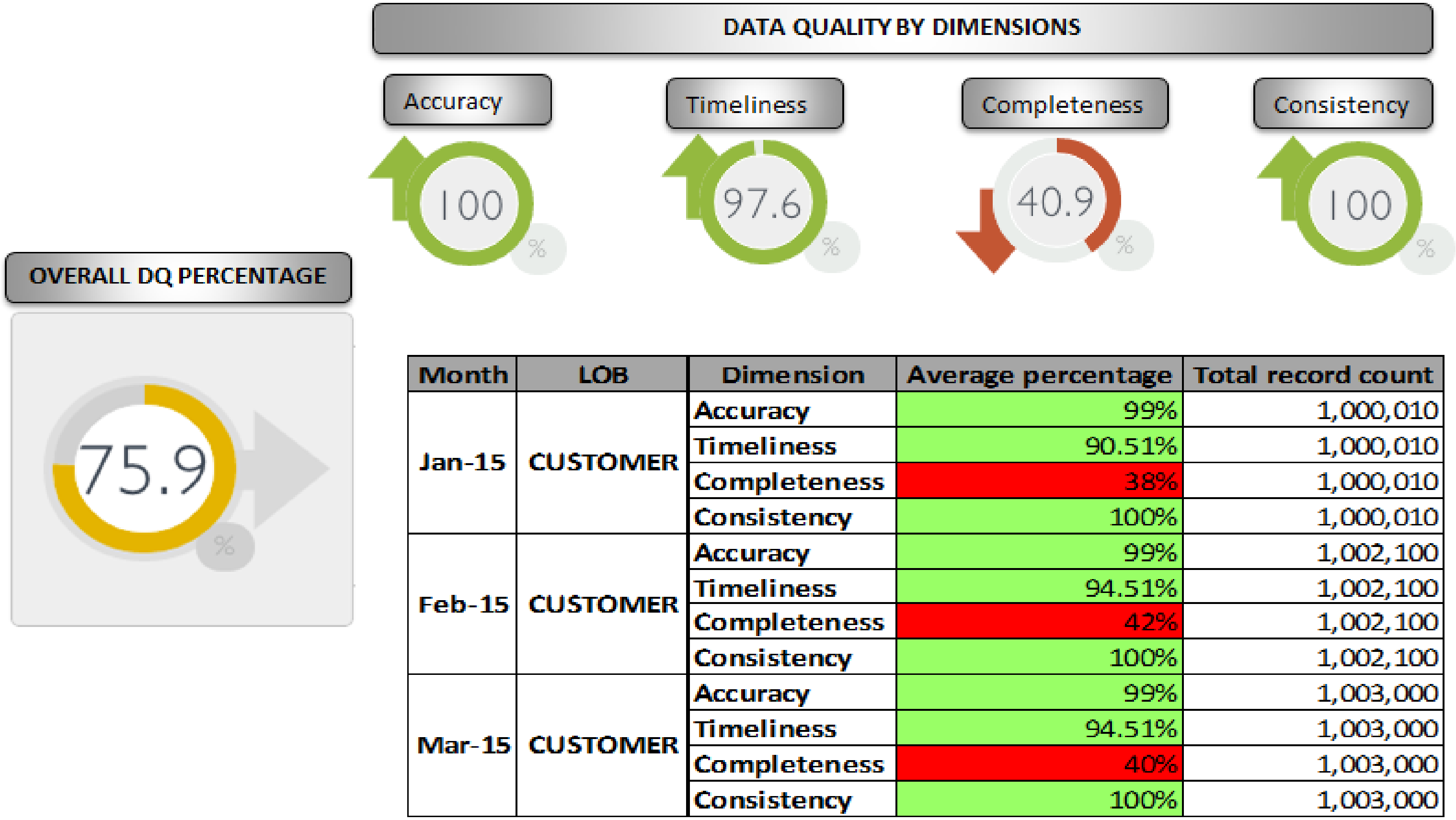
Keeping the Lake clean

Tip 2: Raw data is not always usable

Balancing act between Rawness and Usability

Keeping the Lake clean

Tip 3: Continuously monitor and report on the level of data quality score for key data elements



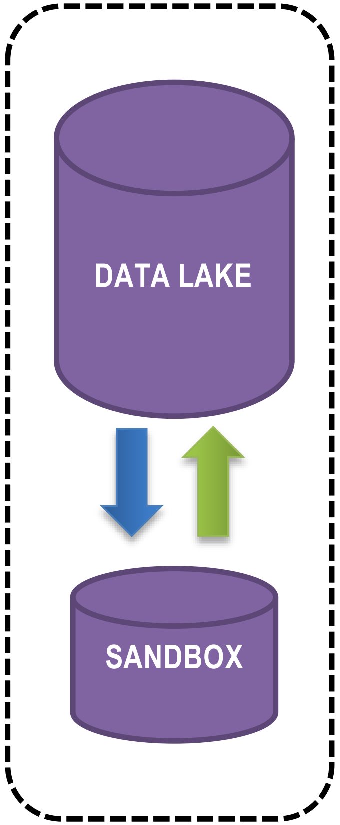
Month	LOB	Dimension	Average percentage	Total record count
Jan-15	CUSTOMER	Accuracy	99%	1,000,010
		Timeliness	90.51%	1,000,010
		Completeness	38%	1,000,010
		Consistency	100%	1,000,010
Feb-15	CUSTOMER	Accuracy	99%	1,002,100
		Timeliness	94.51%	1,002,100
		Completeness	42%	1,002,100
		Consistency	100%	1,002,100
Mar-15	CUSTOMER	Accuracy	99%	1,003,000
		Timeliness	94.51%	1,003,000
		Completeness	40%	1,003,000
		Consistency	100%	1,003,000

“To measure is to know “

@juniormuka1

Keeping the Lake clean

Tip 4: Cater for non-curated data by introducing the concept of Sandbox



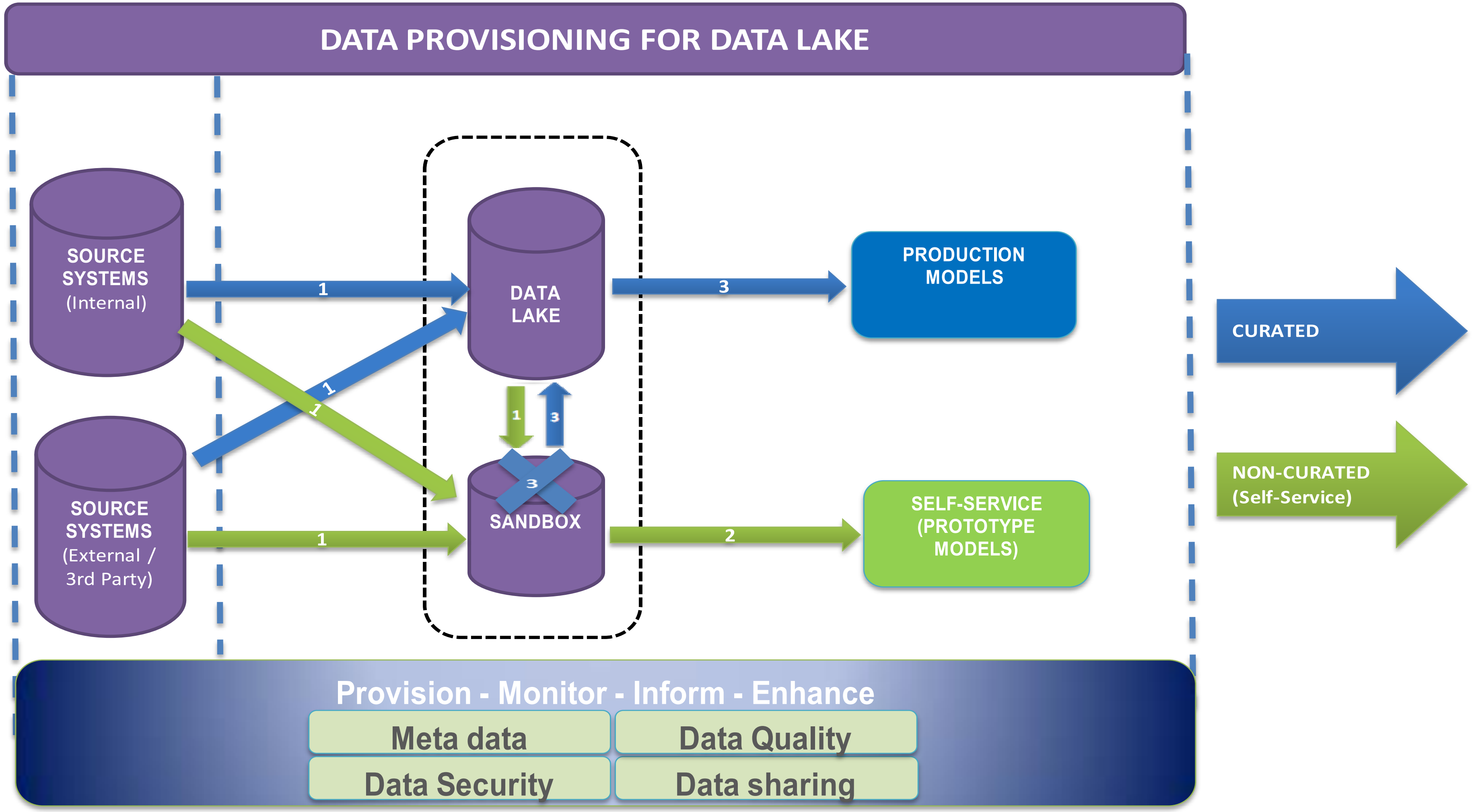
TRANSIENT

SMALL

NOT FOR EVERY USER

Keeping the Lake clean

Tip 5: Curate data provisioned through the data lake


DATA PROVISIONING FOR DATA LAKE
SOURCE SYSTEMS (Internal)
SOURCE SYSTEMS (External / 3rd Party)
DATA LAKE
SANDBOX
PRODUCTION MODELS
SELF-SERVICE (PROTOTYPE MODELS)
CURATED
NON-CURATED (Self-Service)
1
1
1
1
1
1
2
3
3
3
Provision - Monitor - Inform - Enhance
Meta data
Data Quality
Data Security
Data sharing

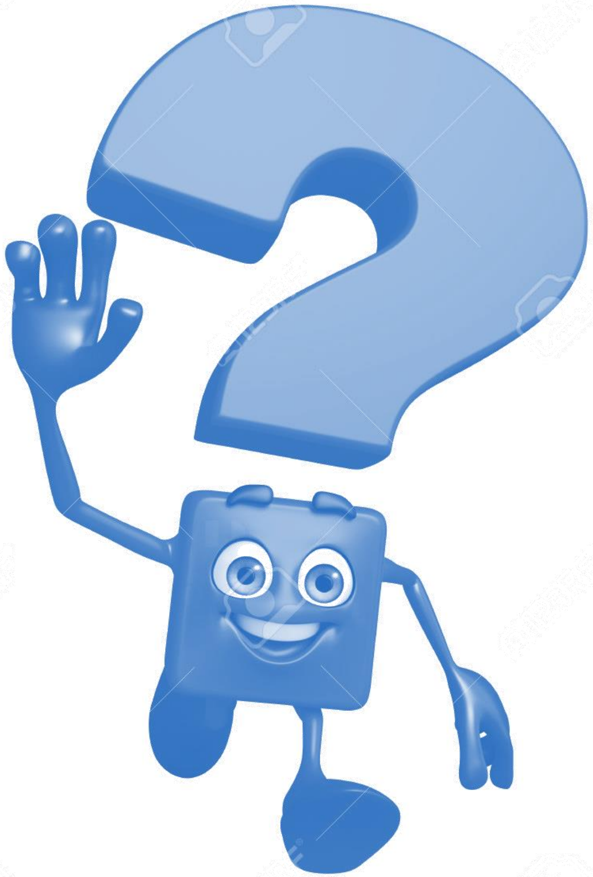
@juniormuka1

Keeping the Lake clean - Discussion

Q1: What have you done (or what would you do) to balance the need for fast ingestion vs. provisioning quality data in your data lake ?

Q2: What have you done (or what would you do) to ensure that your data lake is trusted ?

Questions



@juniormuka1