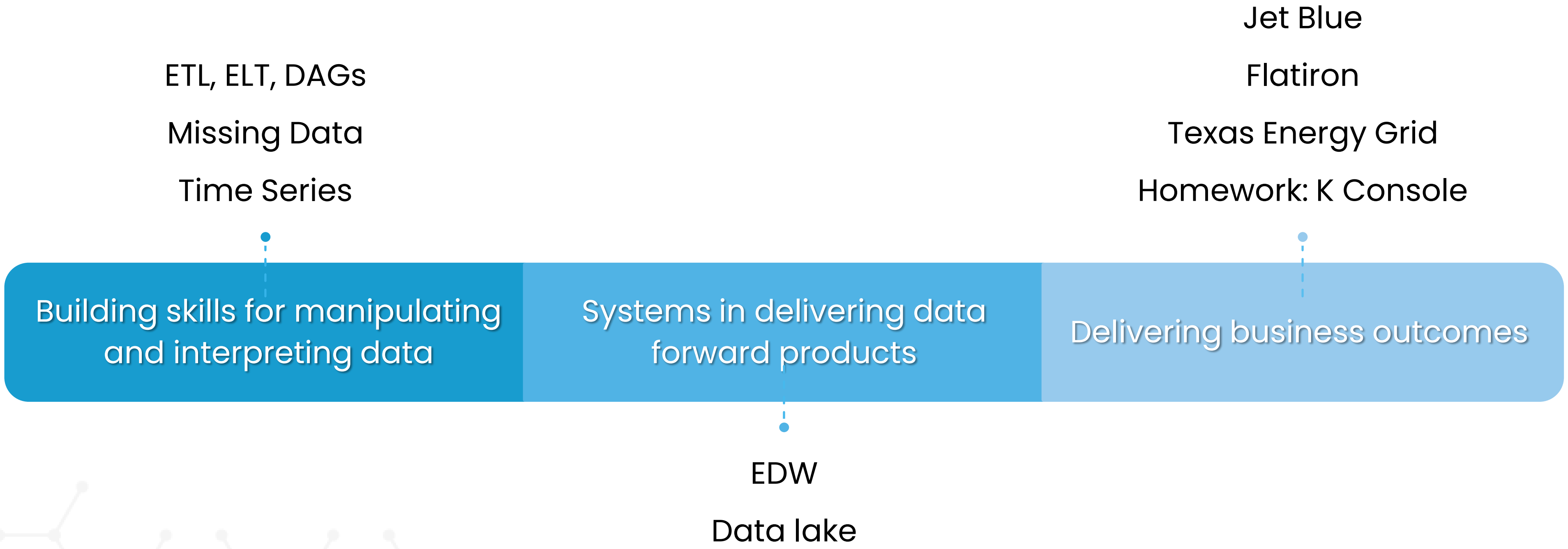
ENTERPRISE DATA & DATA QUALITY

OUR GOALS



ETL, ELT, DAGs

Missing Data

Time Series

Building skills for manipulating and interpreting data

Systems in delivering data forward products

EDW

Data lake

Delivering business outcomes

Jet Blue

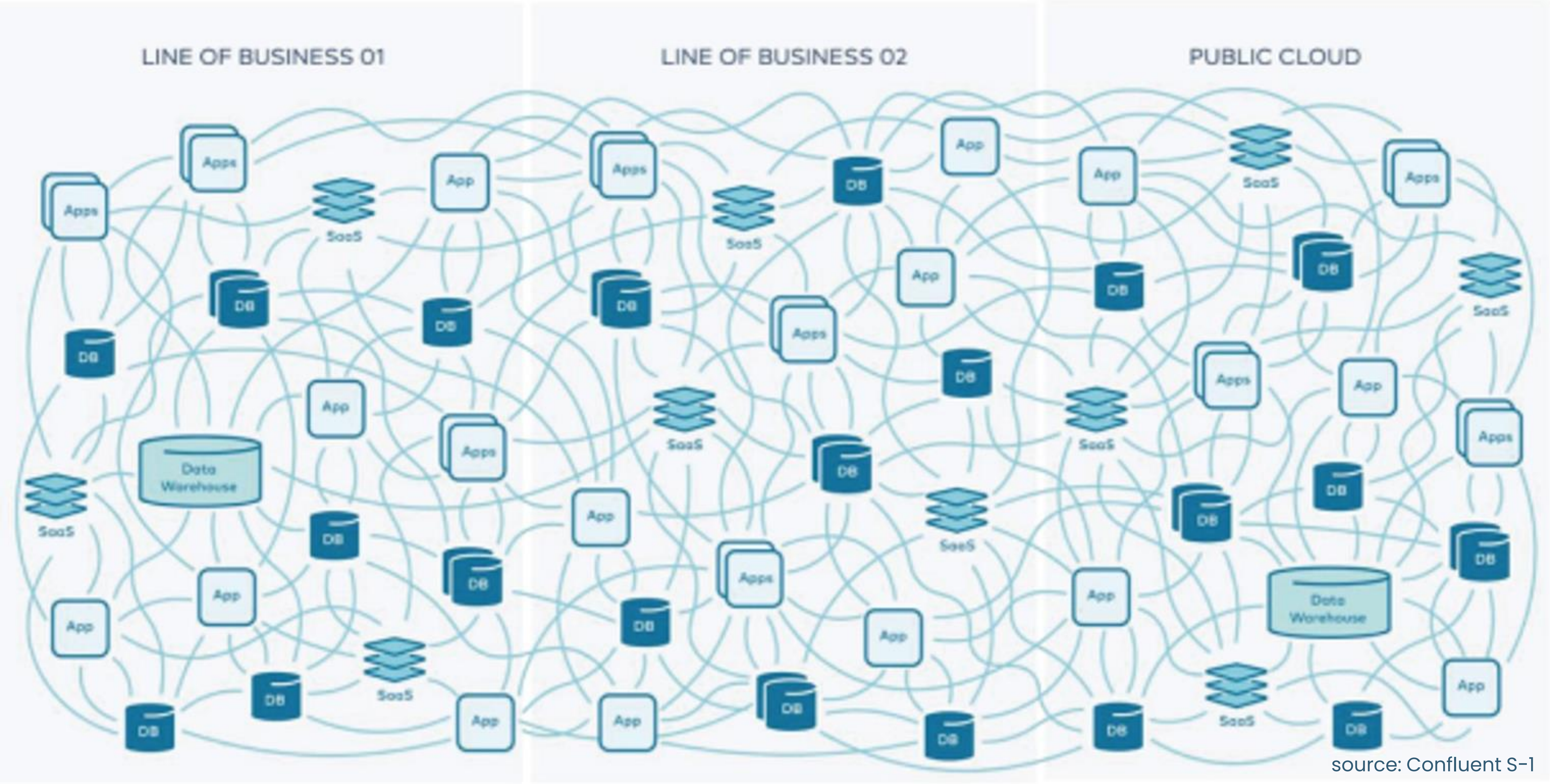
Flatiron

Texas Energy Grid

Homework: K Console

CLASS ROADMAP

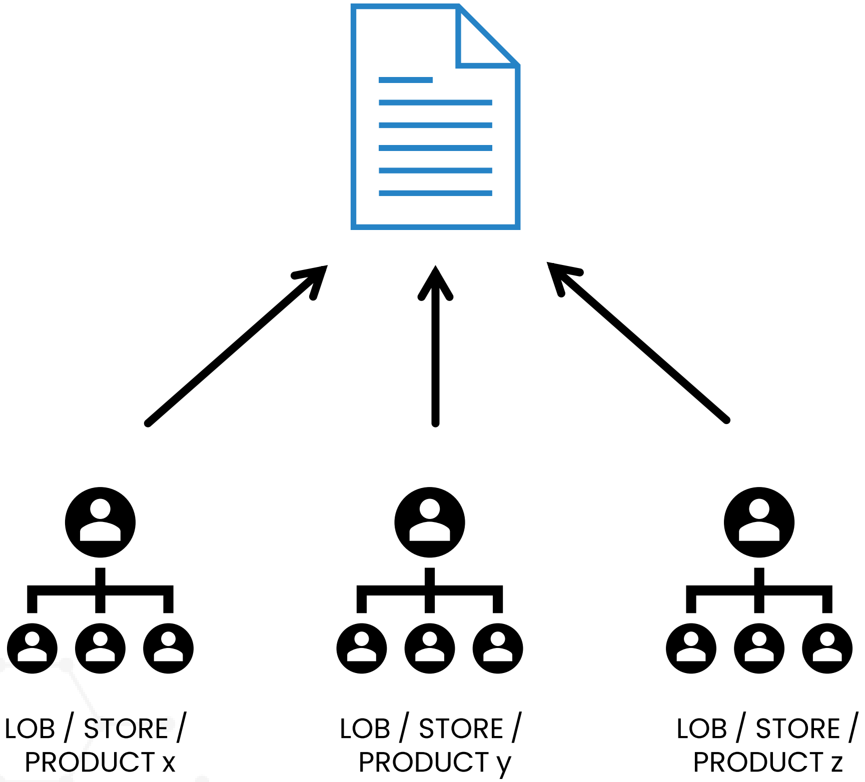
- Data Warehouse
- Jet Blue Breakout
- Data Pipelines
- Flatiron Breakout
- Data Quality
- Break
- Lab



source: Confluent S-1

ENTERPRISE DATA ARCHITECTURE

CALCULATING SALES IN THE PRECEDING PERIOD



THE PROBLEMS WITH DECENTRALIZED DATA

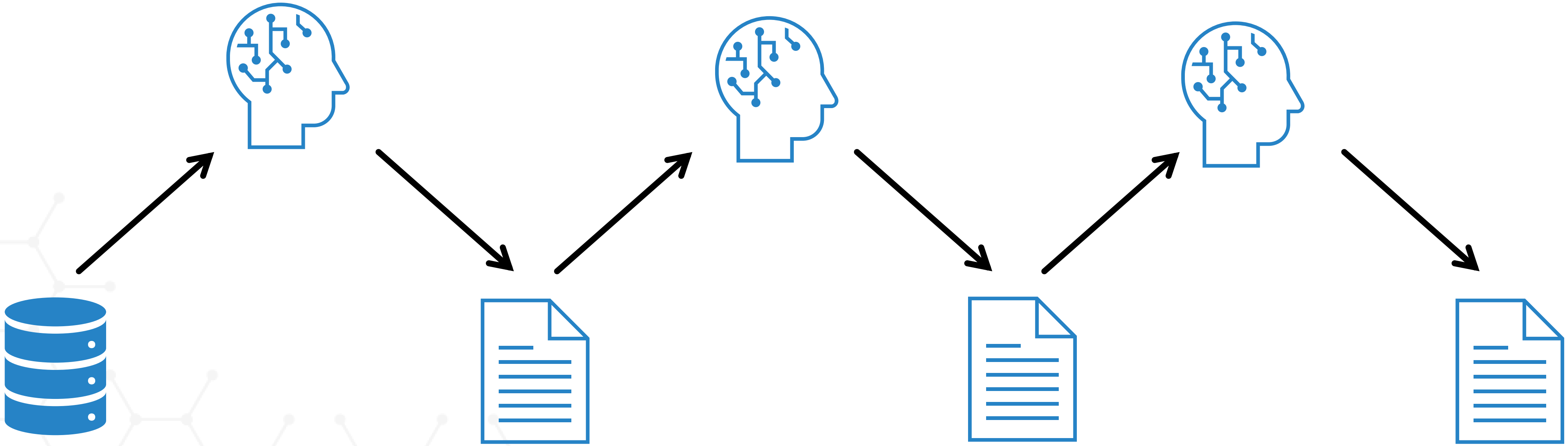
Time basis of data

Algorithmic differential of data (inconsistent business logic)

Summarization and aggregation discrepancies

CREDIBILITY

DELIVERING THE REPORT



PRODUCTIVITY

OPERATIONAL VS DERIVED DATA

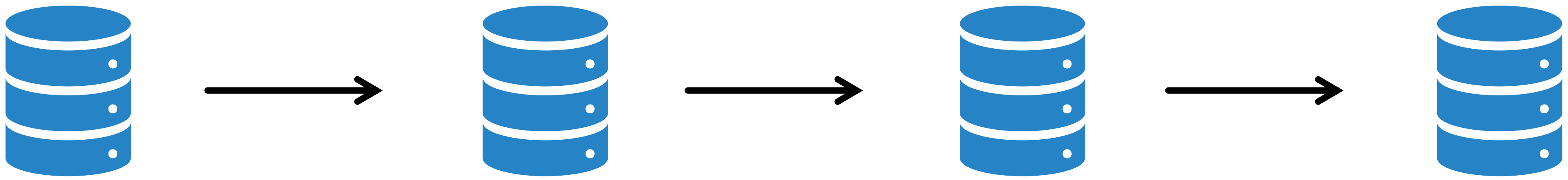
OPERATIONAL

- Dealing with day-to-day operations
  - HCM
  - POS
  - ERP
  - CRM
  - Financial
- Application oriented
- Transaction driven
- Relatively atomic
- Bounded history

DERIVED

- Management and reporting oriented
  - Attrition
  - Total sales
  - WIP Inventory
  - Total deal pipeline, win rates
  - P&L
- Subject oriented
- Analysis driven
- Connected
- Deep, consistent history

AN IDEALIZED ENVIRONMENT



OPERATIONAL

DATA WAREHOUSE

DEPARTMENTAL

INDIVIDUAL

(Enterprise) Data Warehouse: OLAP database where data from different systems is stored and modeled to support analysis and other activities related to answering questions with it. Data in a data warehouse is structured and optimized for reporting and analysis queries.

CLOUD DATA WAREHOUSE

- columnar database -> I/O efficiency, data compression
- highly scalable -> distribute data and queries across many nodes
- store & run bulk transforms
- extracting data and loading it into a data warehouse -> then perform the necessary transformations to complete the pipeline

BREAKOUT

- Groups of 4 or 5
- 10 mins
- Add a new bullet / text box or add a + to the end if you think another group has made the key point

JET BLUE

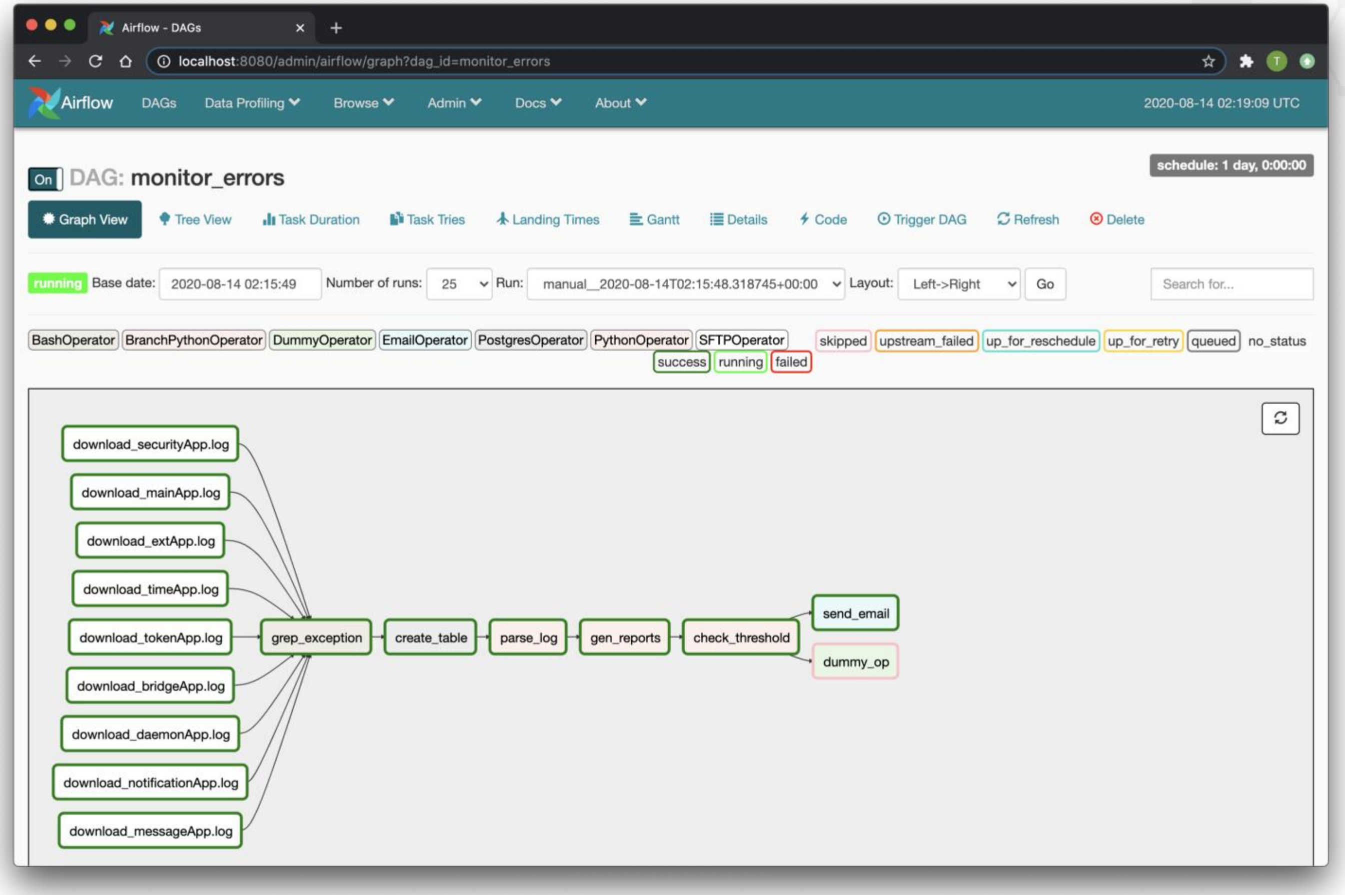
https://docs.google.com/presentation/d/1G8gN3N9_KjlFb9GltZ5g8B3Nzf1C3hkTYiP62P0rrJg/edit?usp=sharing

ETL, ELT, EtLT, REVERSE ETL –> DATA PIPELINES

Data pipelines: processes that move and transform data from various sources to a destination where new value can be derived.

DAGS:
DIRECTED ACYCLIC GRAPHS

- Directed – Data moves in a single direction
- Acyclic – no cycles



localhost:8080/admin/airflow/graph?dag_id=monitor_errors

DAG: monitor_errors

schedule: 1 day, 0:00:00

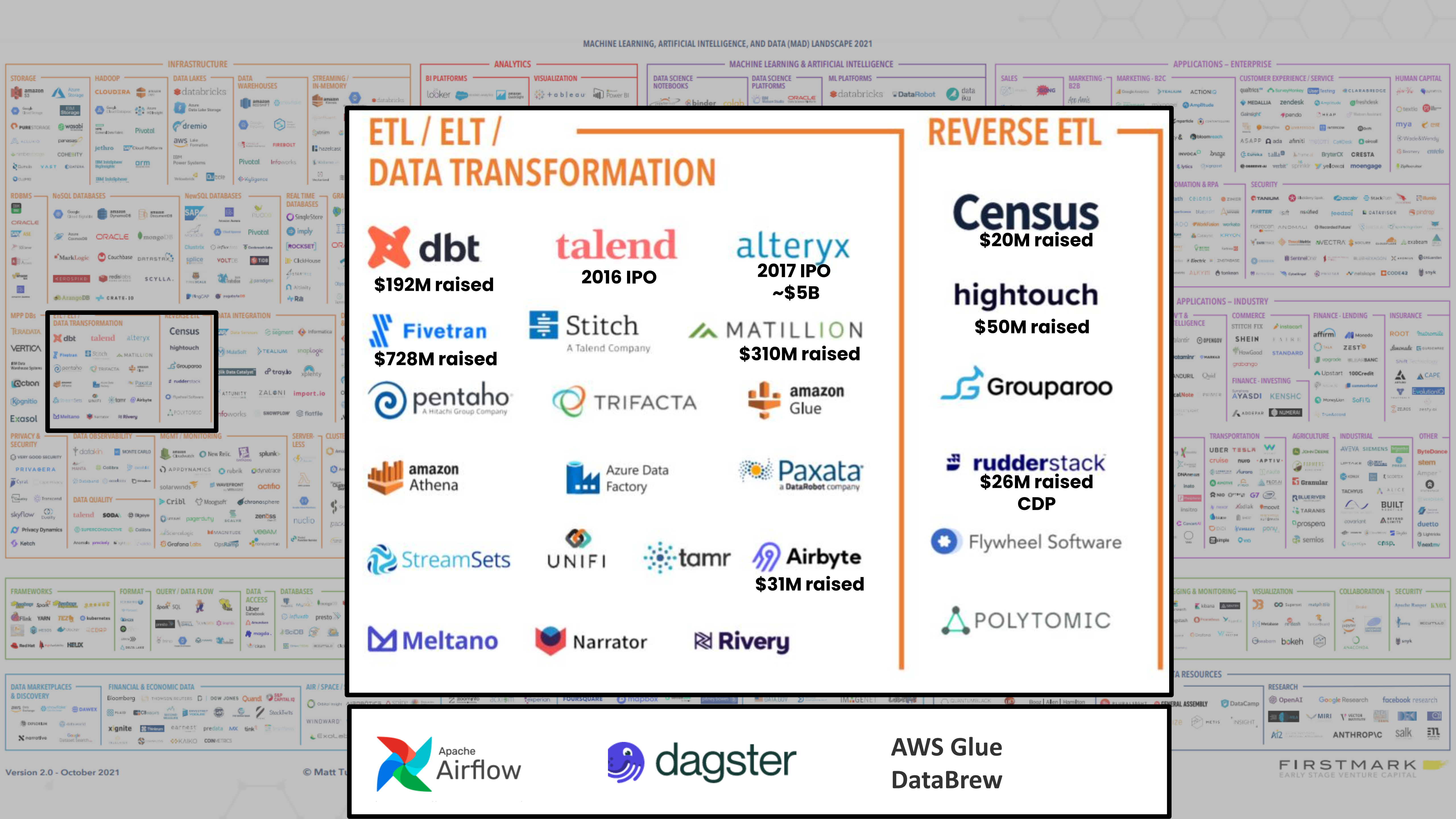
running Base date: 2020-08-14 02:15:49 Number of runs: 25 Run: manual__2020-08-14T02:15:48.318745+00:00 Layout: Left->Right

https://towardsdatascience.com/step-by-step-build-a-data-pipeline-with-airflow-4f96854f7466

DATA ENGINEER

The builder of data pipelines

- Loading data into a data warehouse, transforming data in prep to derive value
- Deliver a scalable production state
- SQL & Data Warehousing
- Python / Java
- Distributed and cloud computing


MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021
INFRASTRUCTURE
ANALYTICS
MACHINE LEARNING & ARTIFICIAL INTELLIGENCE
APPLICATIONS – ENTERPRISE
APPLICATIONS – INDUSTRY
Version 2.0 - October 2021
© Matt Tu
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

ETL / ELT / DATA TRANSFORMATION

dbt
$192M raised

talend
2016 IPO

alteryx
2017 IPO
~$5B

Fivetran
$728M raised

Stitch
A Talend Company

MATILLION
$310M raised

pentaho
A Hitachi Group Company

TRIFACTA

amazon Glue

amazon Athena

Azure Data Factory

Paxata
a DataRobot company

StreamSets

UNIFI

tamr

Airbyte
$31M raised

Meltano

Narrator

Rivery

REVERSE ETL

Census
$20M raised

hightouch
$50M raised

Grouparoo

rudderstack
$26M raised
CDP

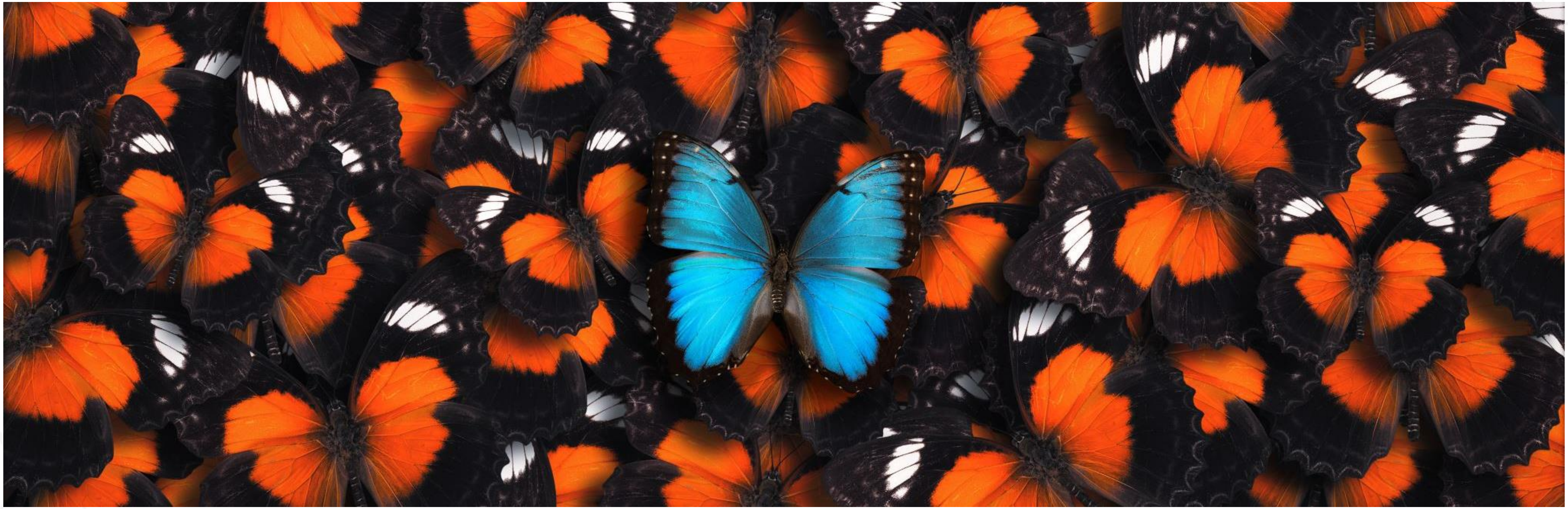
Flywheel Software

POLYTOMIC

Apache Airflow

dagster

AWS Glue DataBrew



DATA TRANSFORMATION

As simple as converting timestamps. Creating new metric from multiple source columns that are aggregated and filtered through some **business logic.**

BUSINESS LOGIC

Software: custom rules or algorithms that handle the exchange of information between a database and user interface

Data: + standards and idiosyncratic rules that are applied throughout the organization

- Recall: JetBlue passengers on a plane
  - Idiosyncratic (positively) to LOB / analysis needs
- What's a valid email address, customer name?
- How to divide the year into reporting periods?
- How are discounts applied?
  - Product, Invoices
  - Impacts on product P&Ls, sales commissions

DATA LAKE VERSION 1.0

COMPLEXITY AND GROWING DATA TYPES LED TO THE DEVELOPMENT OF DATA LAKES

ALL THE DATA DUMPED TO A SINGLE LOCATION.

SUCCESS:

GOT THE DATA TO A SINGLE LOCATION

CHALLENGES:

BUSINESS LOGIC

LACK OF META DATA

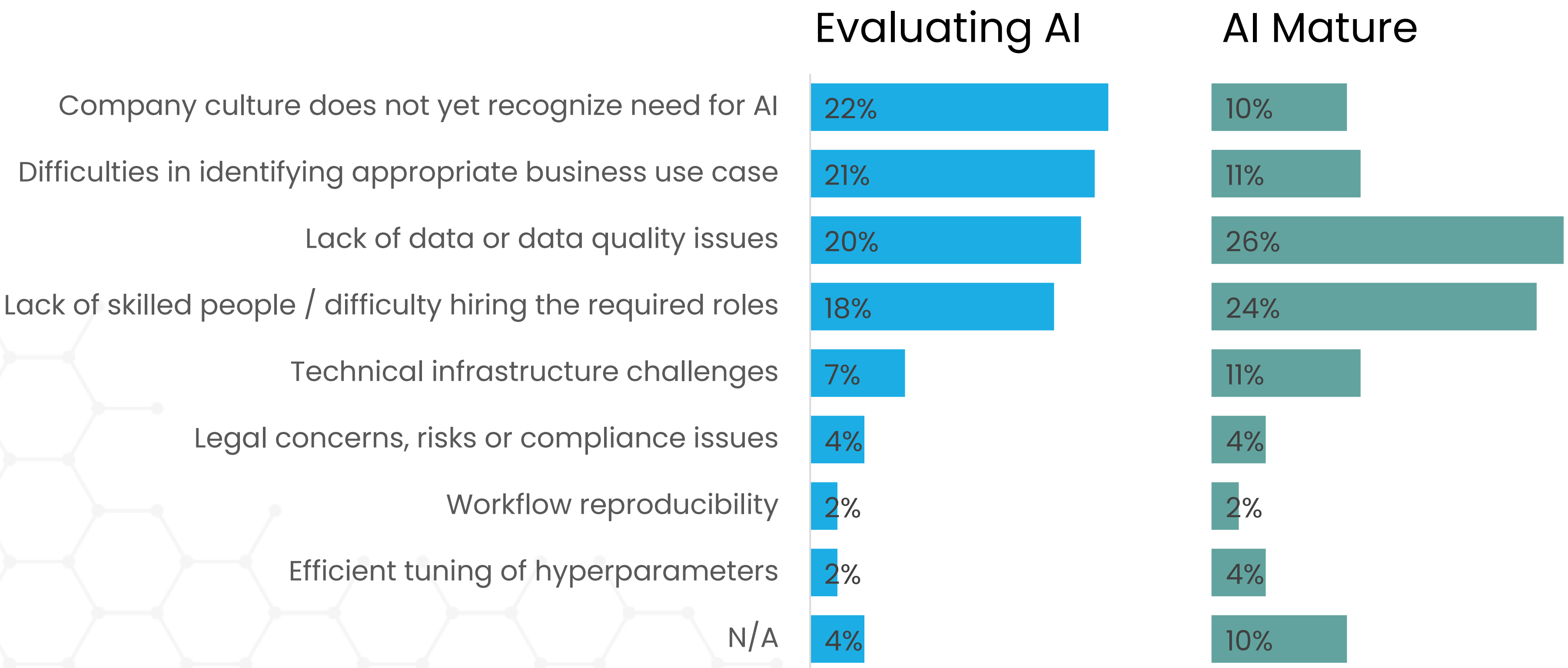
DATA QUALITY

flatiron

https://docs.google.com/presentation/d/1RMaSPhZoIQb-f0mieuiOX2QEHDwDa4BHDhJCDNQSiUs/edit?usp=sharing

FLATIRON HEALTH

TOP ISSUE IN AI DEPLOYMENT: DATA QUALITY



O’Reilly 2019 AI Adoption Survey

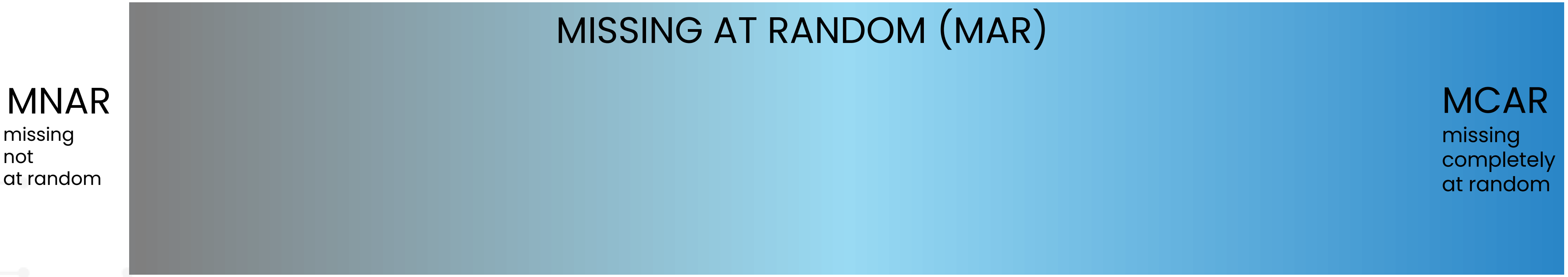
DATA QUALITY

ACCURATE	COMPLETE	TIMELY	CONSISTENT	UNIQUE
Does the data correctly represent the real world?	Is all of the data present?	Is the data available when needed?	Is the data consistent across datasets?	Is the data duplicated?
Negative ages, 867-5309 email address	Missing fields	Daily sales report only available to last week	CRM: $100, ERP: $300	Duplicated Brook Miller’s with same phone, SSN

DATA QUALITY

ACCURATE	COMPLETE	TIMELY	CONSISTENT	UNIQUE
Does the data correctly represent the real world?	Is all of the data present?	Is the data available when needed?	Is the data consistent across datasets?	Is the data duplicated?
Negative ages, 867-5309 email address	Missing fields	Daily sales report only available to last week	CRM: $100, ERP: $300	Duplicated Brook Miller’s with same phone, SSN

MISSING DATA



MISSING AT RANDOM (MAR)

MNAR
missing
not
at random

MCAR
missing
completely
at random

For more: MICE: Flexible Imputation of Missing Data

BREAK

TEXAS ELECTRICAL GRID

APPENDIX