CSU The California State University
OFFICE OF THE CHANCELLOR

Modern Data Engineering – Concepts, Best Practices & Applications

- Subash DSouza
- Director, Cloud Data Engineering
- California State University, Office of the Chancellor

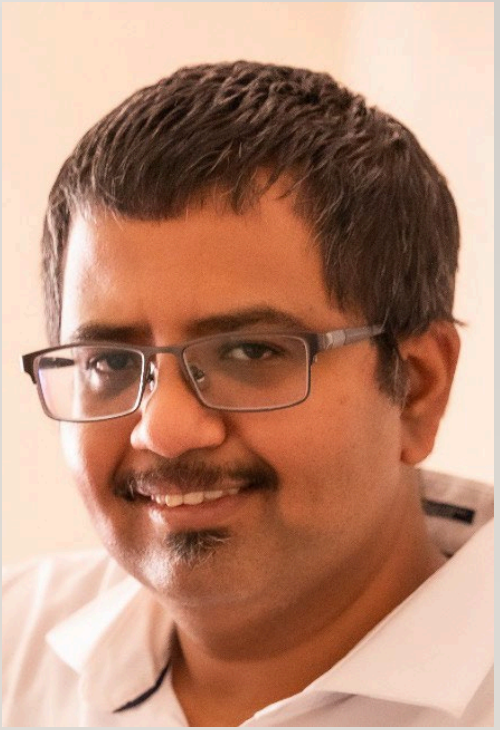
Agenda

CSU The California State University
OFFICE OF THE CHANCELLOR

- Who Am I?
- About CSU
- CSU Data Lake Architecture
- Agile, CI/CD & Testing: Critical components
- Digging Deeper: Data Processing
- Digging Deeper: Data Egress
- Enterprise Data Governance
- MDM(Master Data Management)
- Data Catalog/ Dictionary

WHO AM I?

- Director, Cloud Data Engineering, California State University, Office of the Chancellor
- Founder of Data Con LA, the largest data conference in the SoCal region
- Founder of Data 4 Good, a nonprofit using data for solving social challenges
- AWS Education Champion
- ACM and IEEE – Senior Members
- LinkedIn - https://www.linkedin.com/in/sawjd/
- Twitter - https://twitter.com/sawjd22
- Email – sdsouza@calstate.edu



CSU The California State University
OFFICE OF THE CHANCELLOR

About California State University

- Public University
- Largest 4-year degree program in the nation with nearly half a million students and 50000 faculty and staff across 23 campuses in California.
- Each year, the university across all 23 campuses awards nearly 100,000 bachelors, masters and doctoral degrees
- Interim Chancellor – Jolene Koester

Data trends

Growing exponentially

From new sources

Increasingly diverse

Used by many people

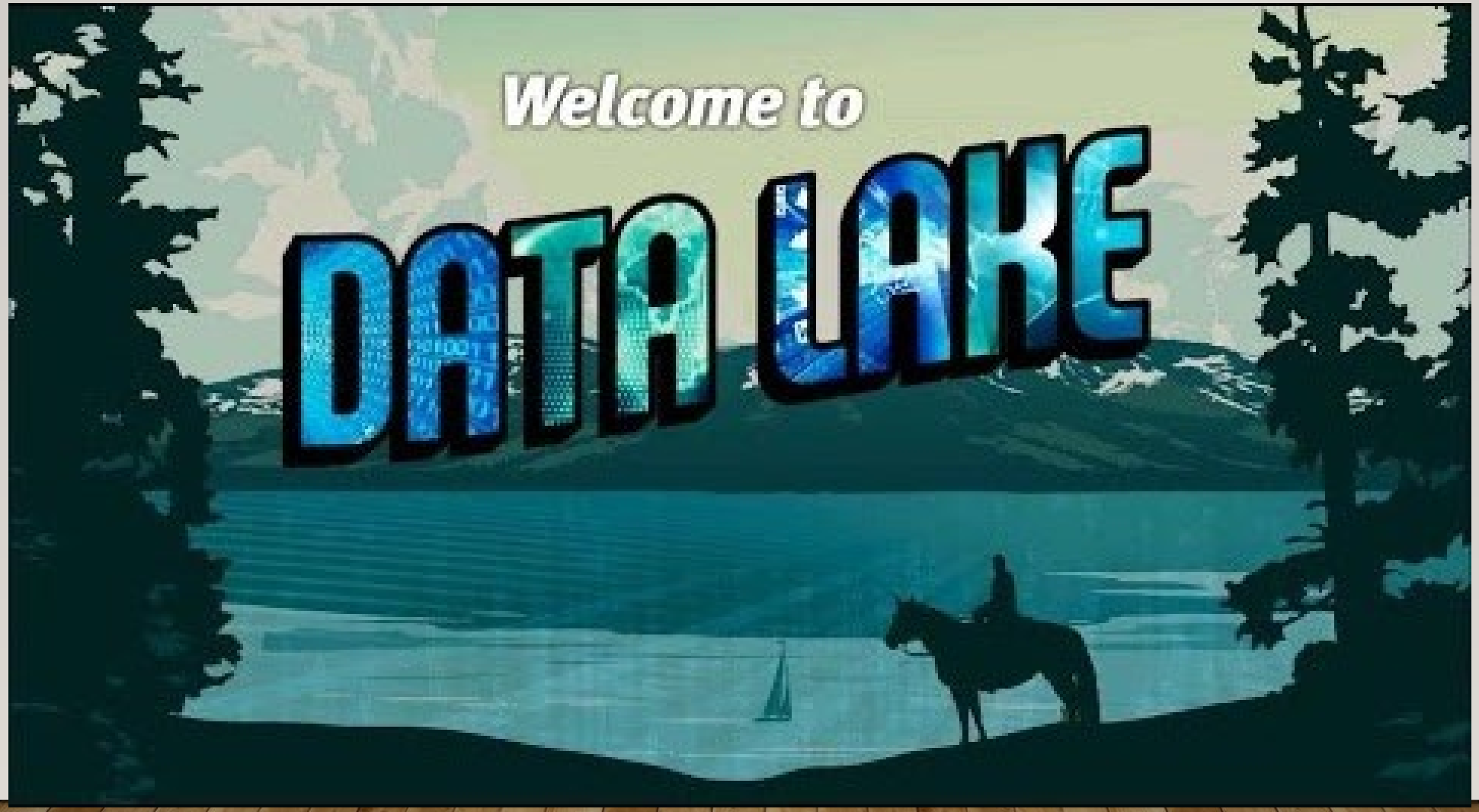
Analyzed by many applications

Companies moving to data lake architectures

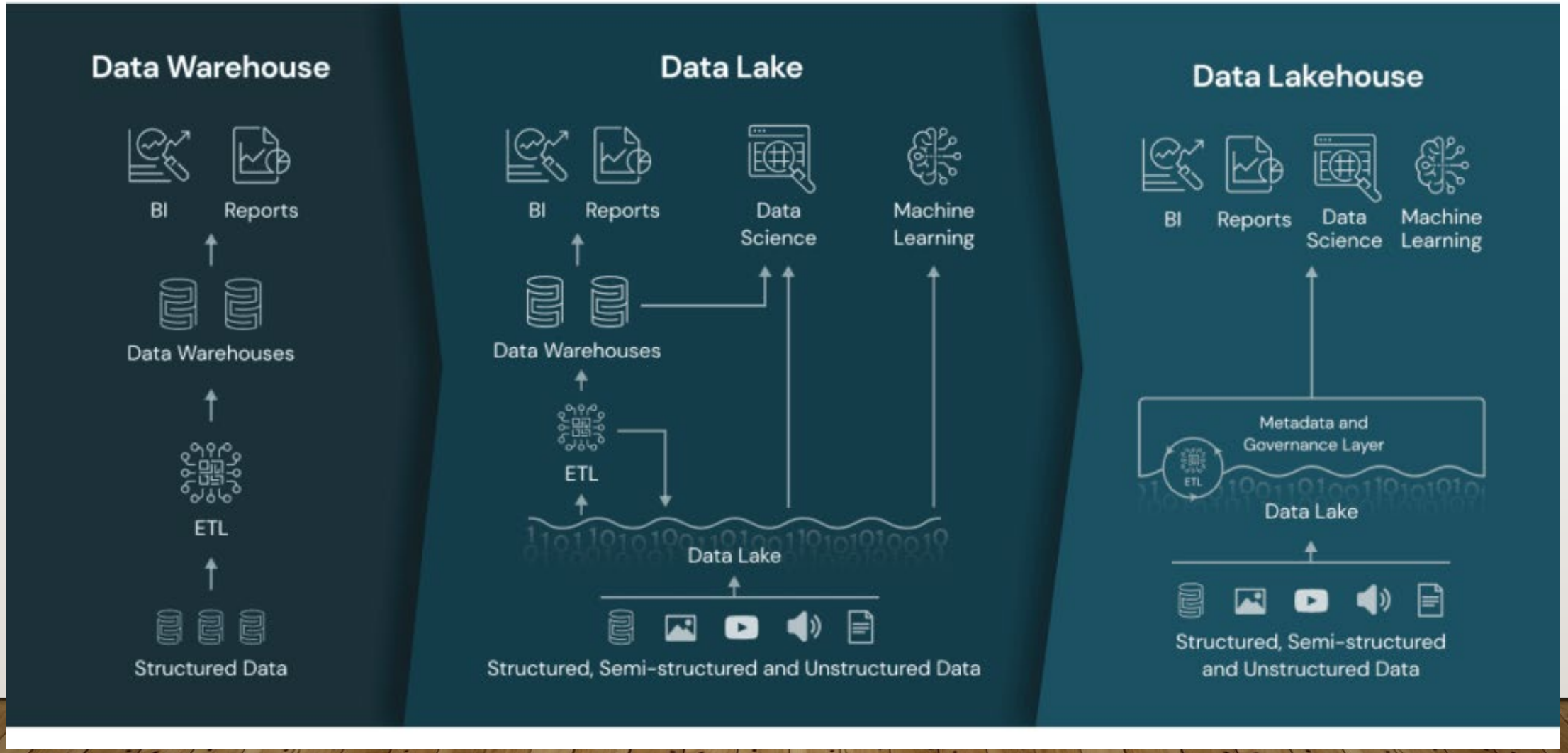
Bringing together the best of both worlds

- Extends or evolves DW architectures
- Store any data in any format
- Durable, available, and exabyte scale
- Secure, compliant, auditable
- Run any type of analytics from DW to Predictive

CSU The California State University
OFFICE OF THE CHANCELLOR



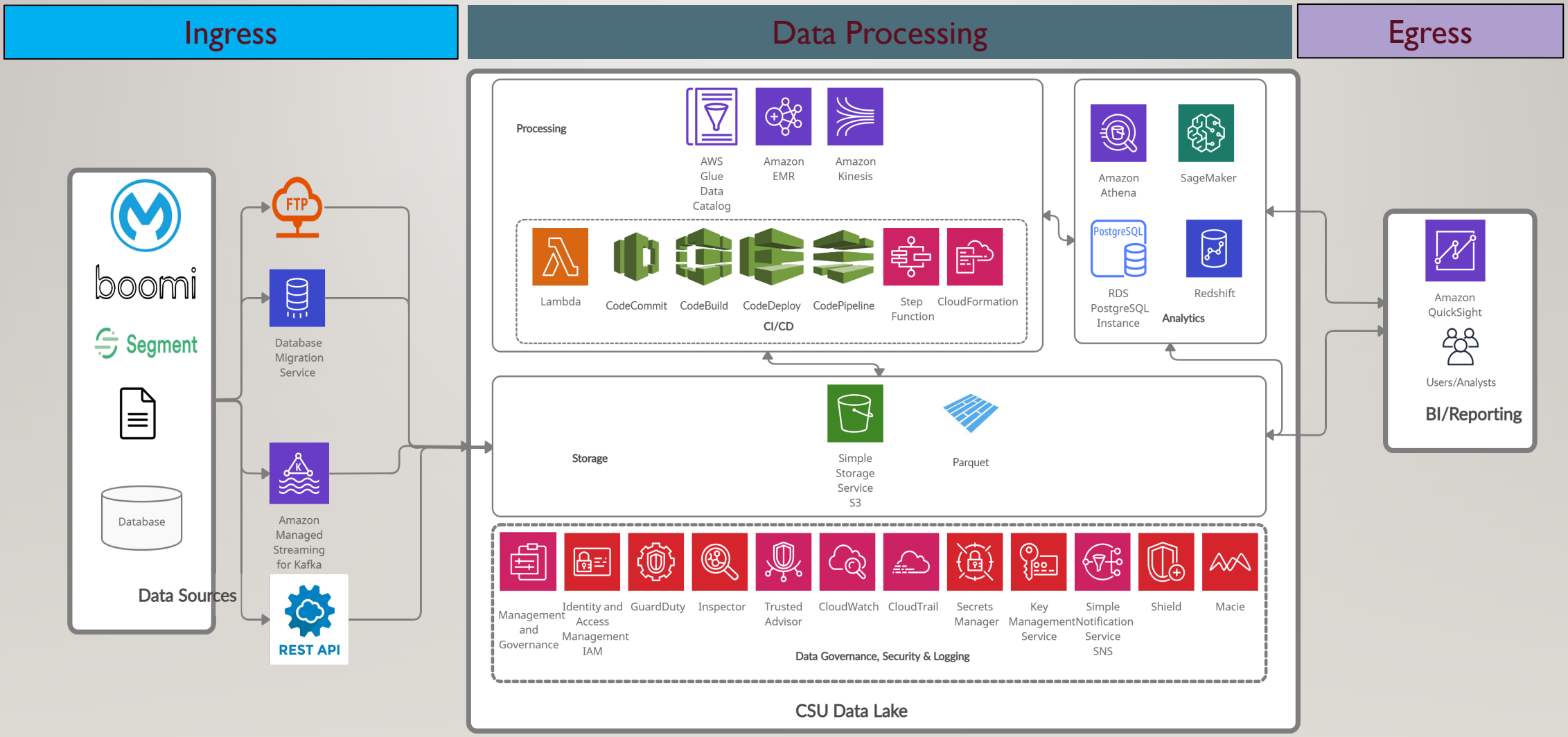
Data Warehouse vs Data Lake vs Data Lakehouse



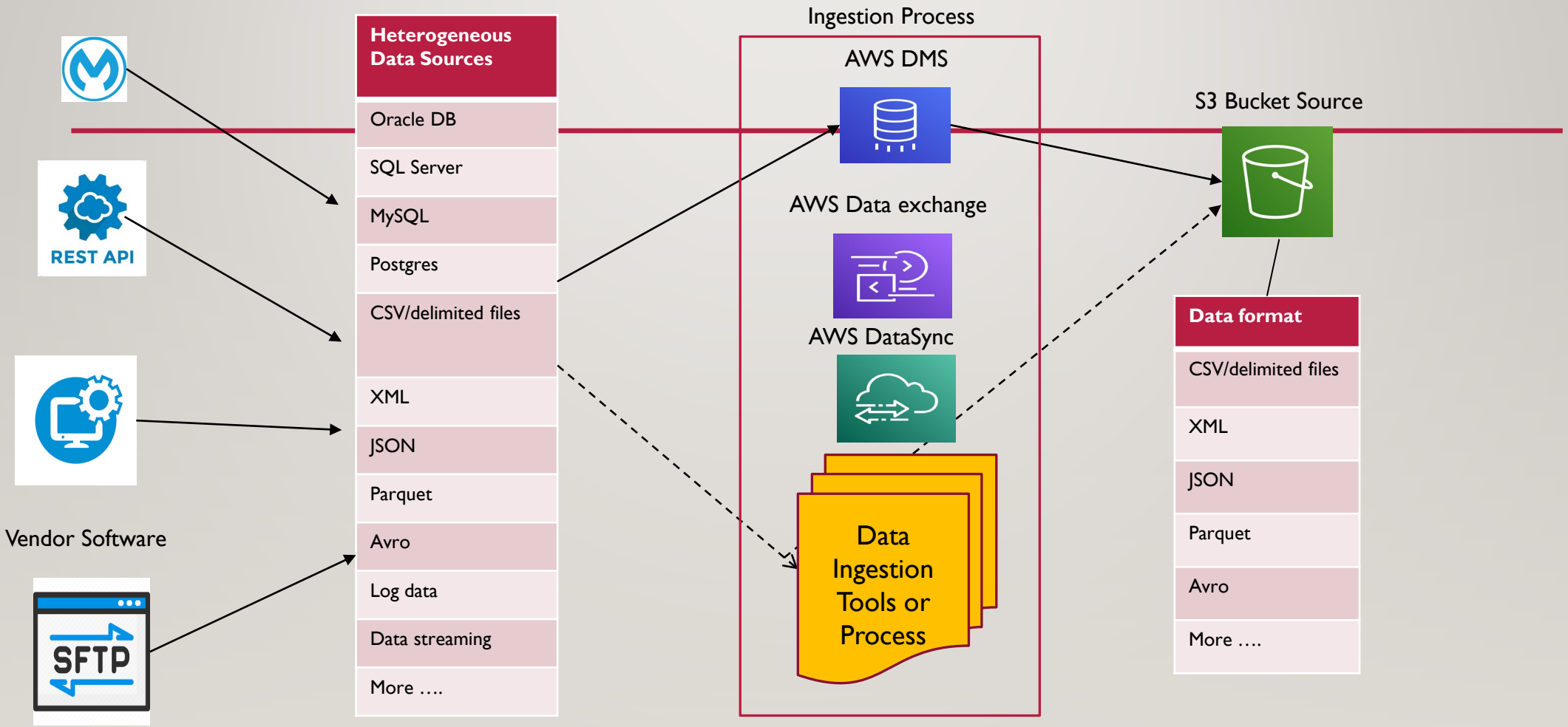
CI/CD & DATAOPS

- A deployment pipeline must be a repeatable and reliable process
  - Use the same process everywhere
- Automate Everything
  - Testing, provisioning, deployment
- Version Control Everything
  - Source code, configuration, build scripts, documentation
- Done means released
- DataOps – DevOps on Data

CSU The California State University
OFFICE OF THE CHANCELLOR

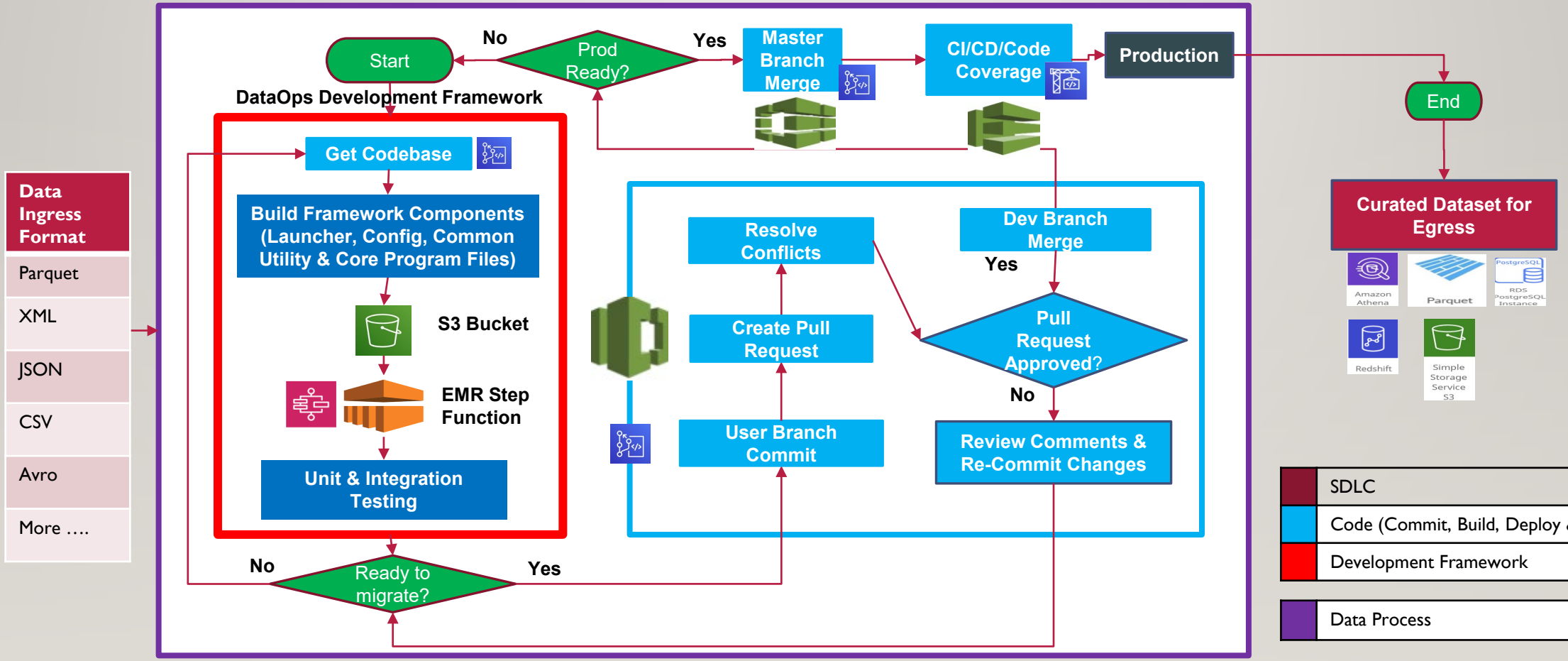


DATA INGRESS

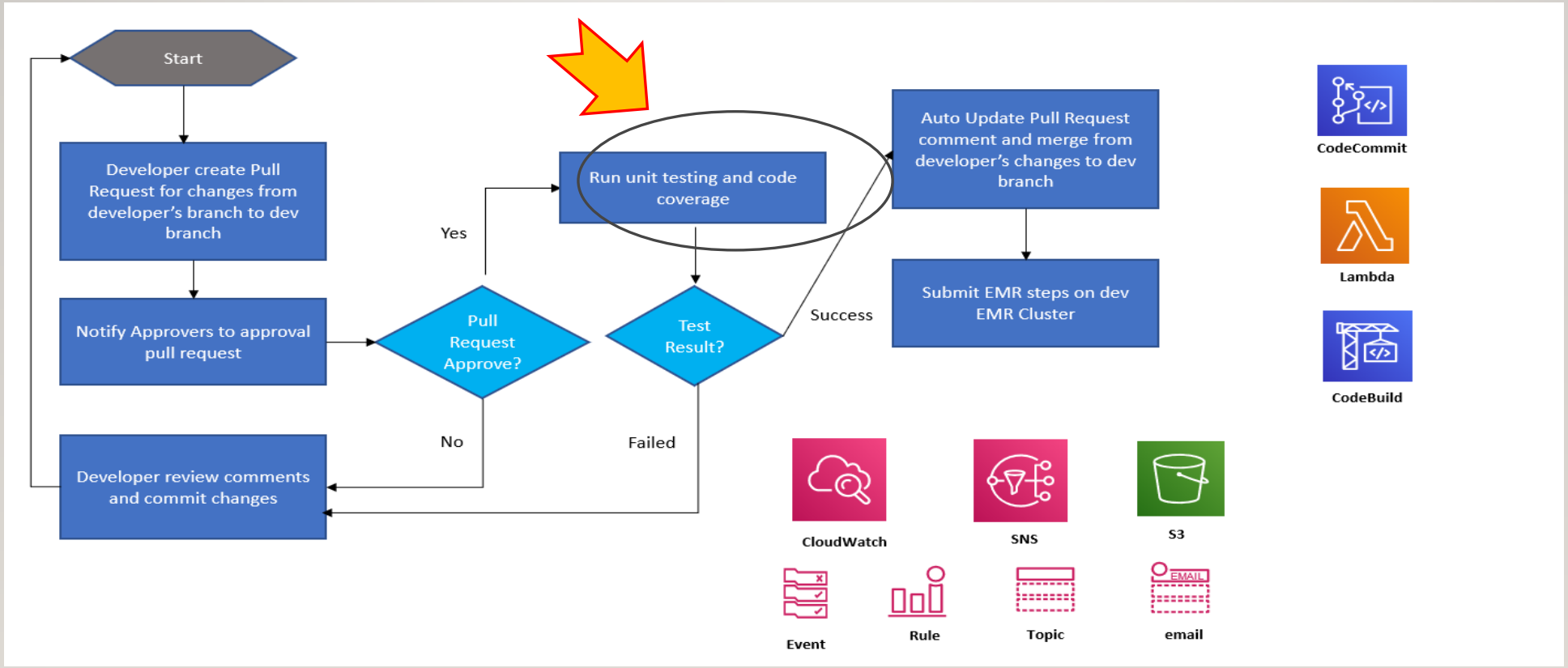


The California State University
OFFICE OF THE CHANCELLOR

DATAOPS DATA PROCESS & DEVELOPMENT FRAMEWORK



DATA PROCESS : TEST AUTOMATION OVERVIEW - TEST EXECUTION INSIDE THE CICD WORK FLOW

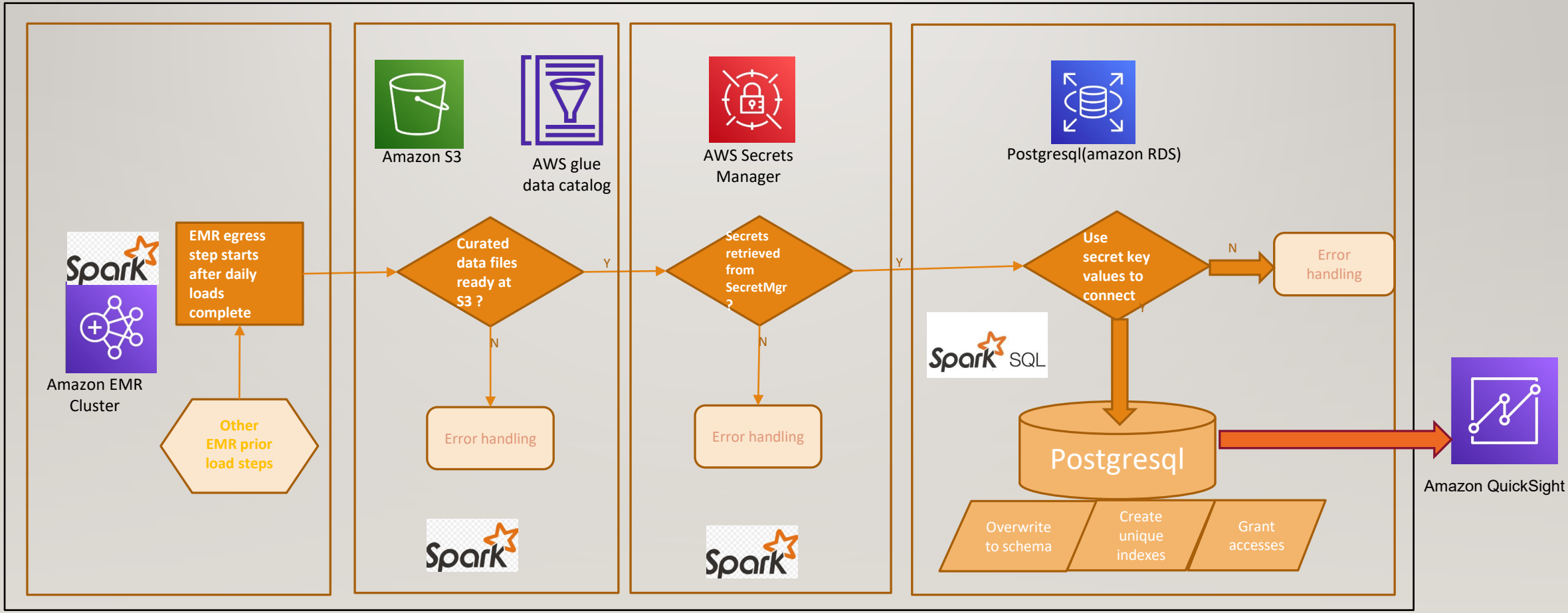


DATA PROCESS : TEST AUTOMATION
BUILD TEST AUTOMATION LEVERAGING PYTEST & CHISPA

- Test cases are automatically run during the CICD process
- Test cases are implemented using pytest, and chispa
- Test coverage will be automatically collected and calculated

The California State University
OFFICE OF THE CHANCELLOR

DATA EGRESS: WORK FLOW



Data Analysis

- In order to be analyzed and useful, data
  - Needs to related to each other
  - Have analytical infrastructure carefully arranged and made available to the end user
- Unless we meet these two conditions, the data lake turns into a swamp, and swamps start to smell after a while.

On to the Data Lakehouse

- How to go from a data lake to a data lakehouse?
- For the most part we have the analytical infrastructure setup with AWS QuickSight sourcing from the various data sources and stage and report on as well as the ability to store structured, semi structured and unstructured data.
- But a key component that would enable true Data Lakehouse is
- Enterprise Data Governance

ATTRIBUTES OF DATA GOVERNANCE

DATA GOVERNANCE CHALLENGES

- Poor data quality costs real money
- Process efficiency is negatively impacted by poor data governance
- Full potential benefits of new systems not be realized because of poor data governance
- Decision making is negatively affected by poor data governance

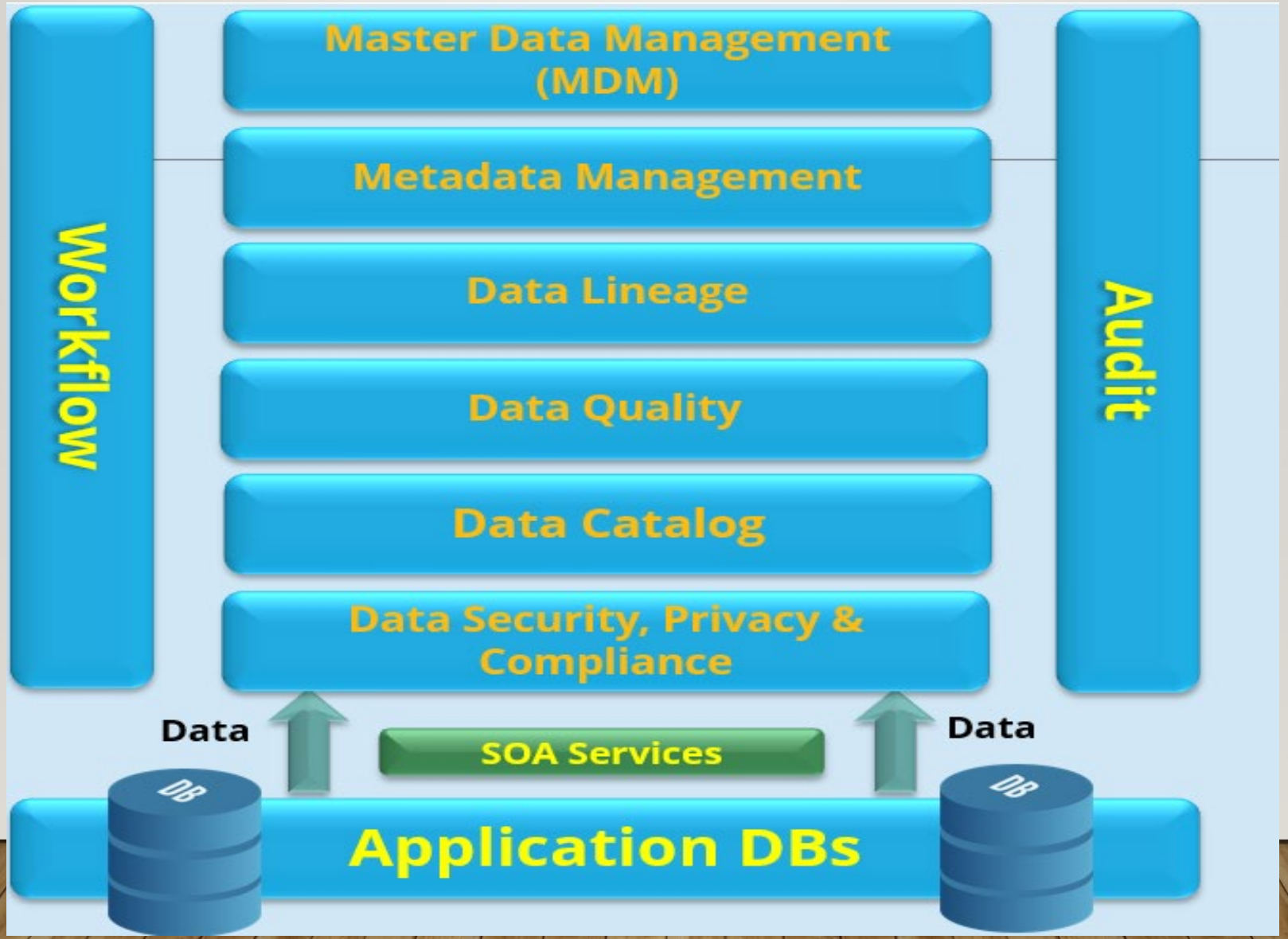
CSU The California State University
OFFICE OF THE CHANCELLOR

DATA GOVERNANCE OBJECTIVES

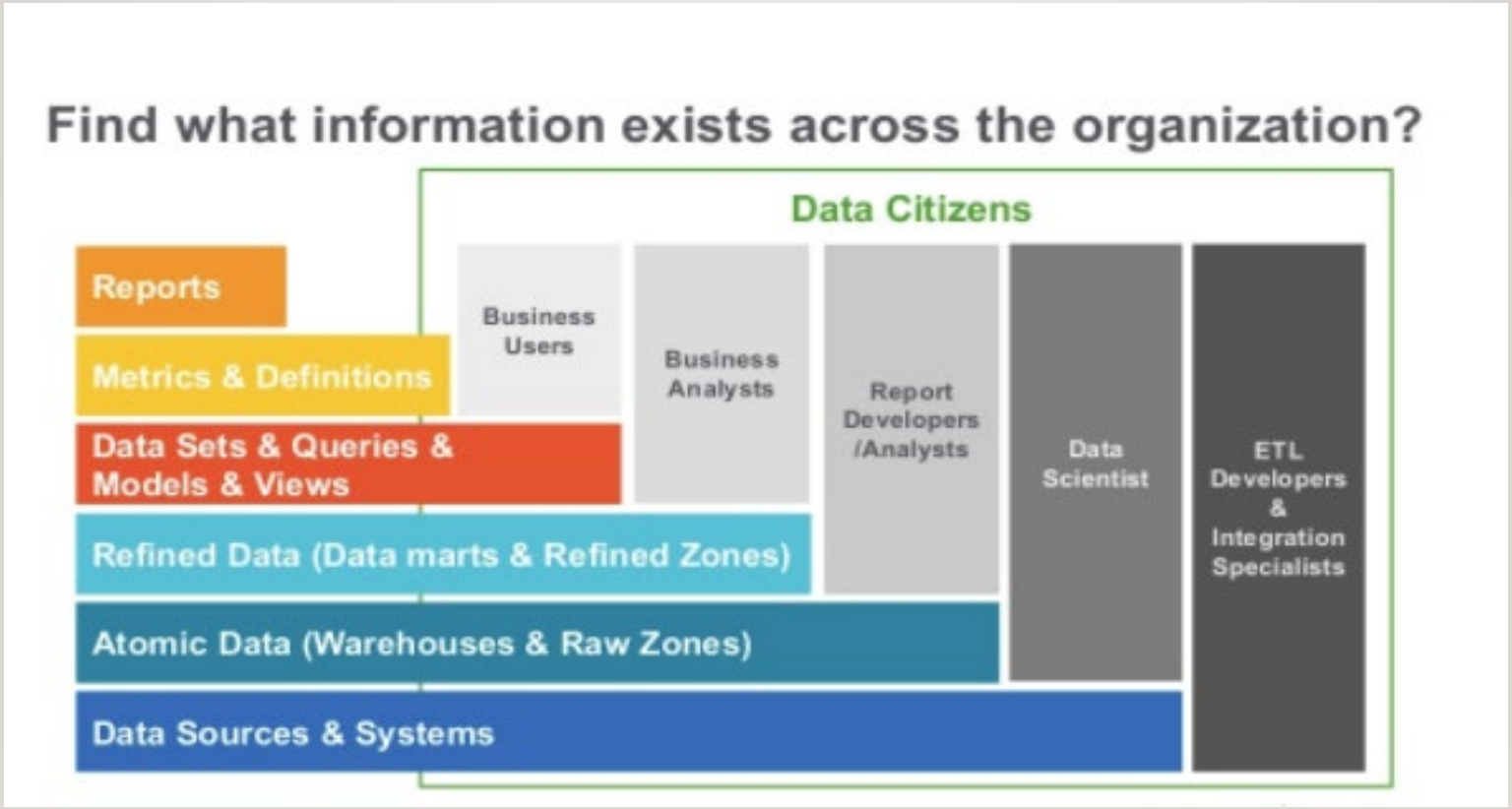
- Guide information management decision-making
- Ensure information is consistently defined and well understood
- Increase the use and trust of data as an organization asset
- Improve consistency of projects across the organization
- Ensure regulatory compliance
- Eliminate data risks

Enterprise Data Governance

CSU The California State University
OFFICE OF THE CHANCELLOR

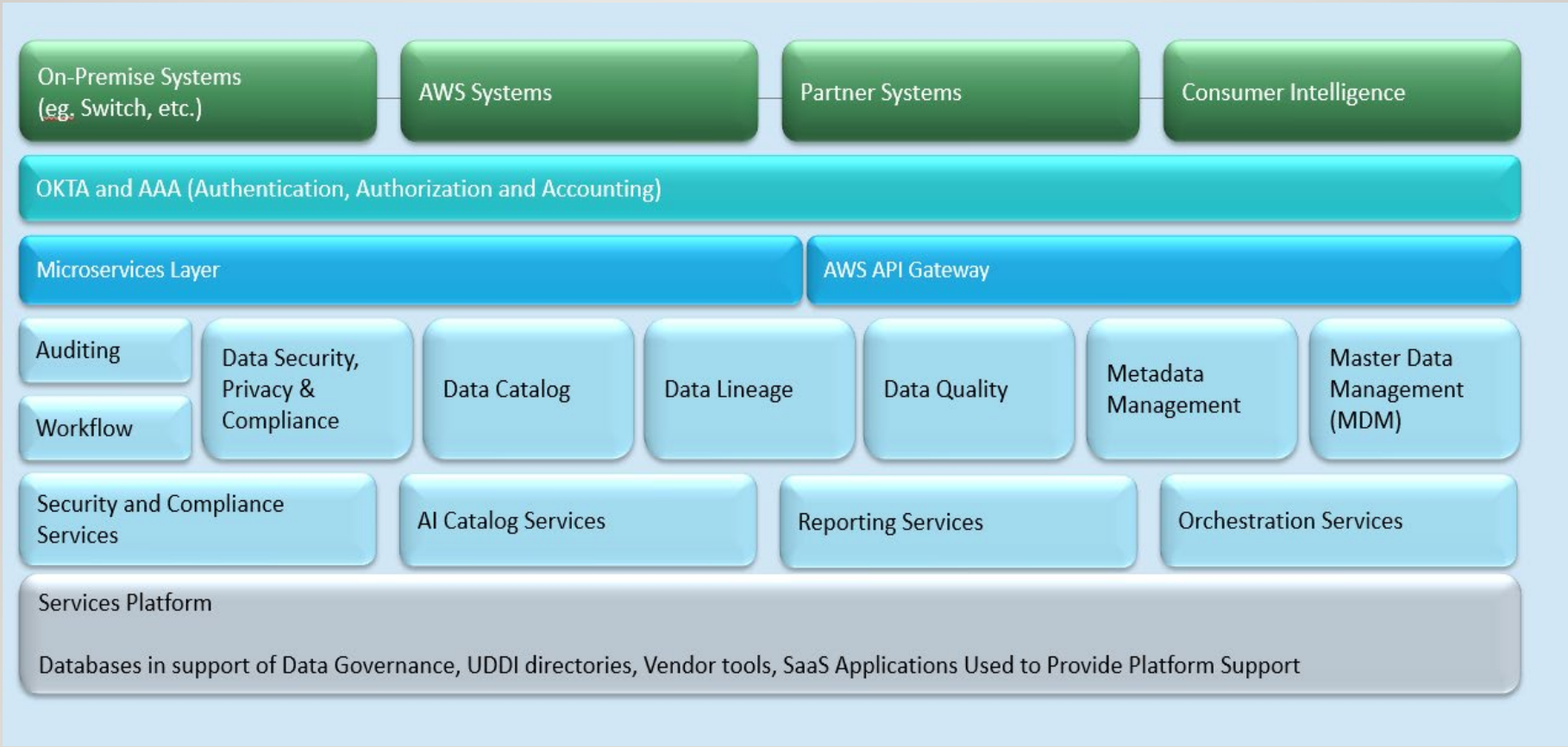


IMPORTANCE OF A DATA CATALOG



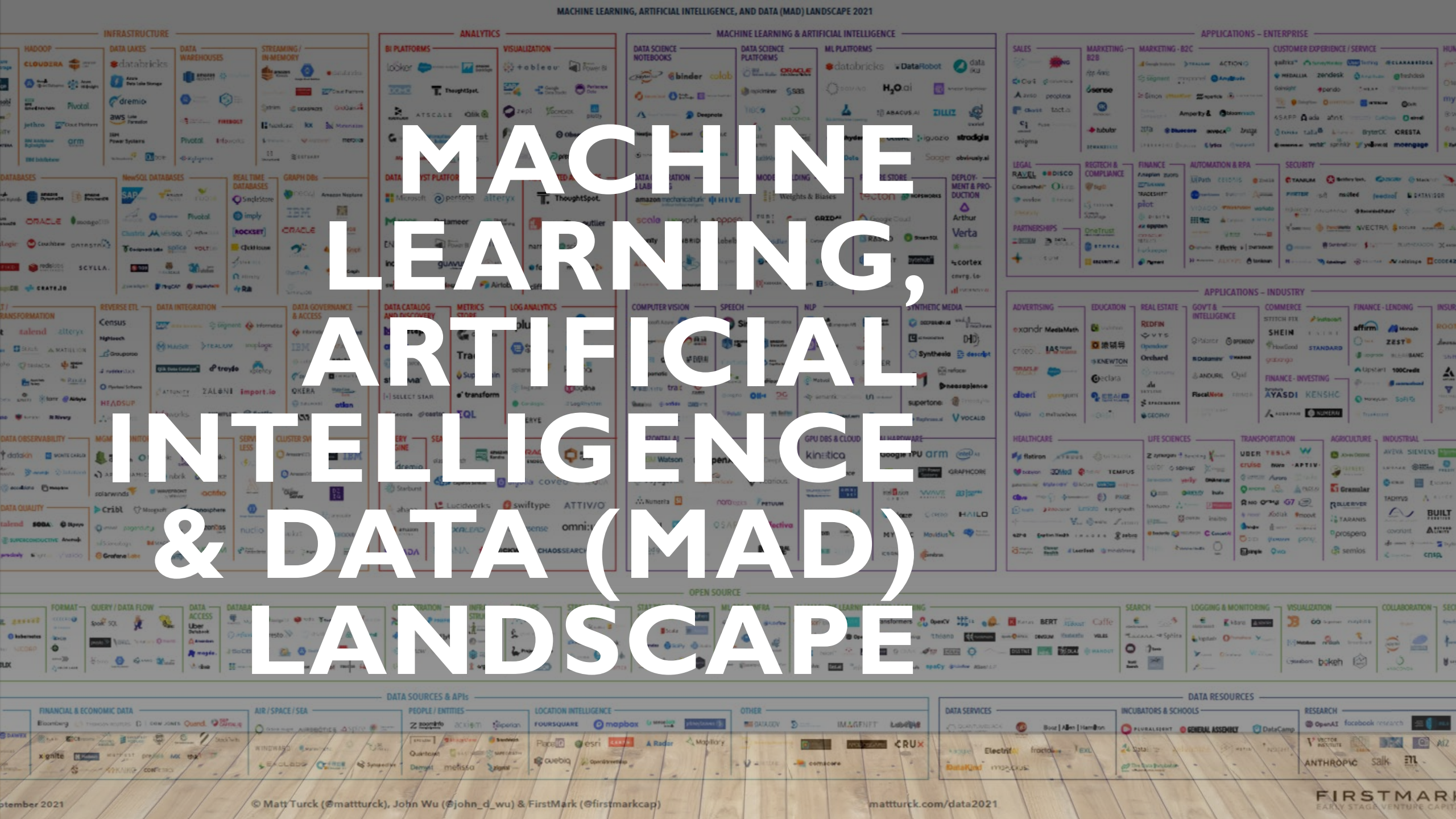
The California State University
OFFICE OF THE CHANCELLOR

CONCEPTUAL ARCHITECTURE



PARTNERSHIP BETWEEN BUSINESS AND IT

- Data management is a shared responsibility between data management professionals within IT and the business data owners representing the interests of data producers and information consumers
- Business data ownership is the concerned with accountability for business responsibilities in data management
- Business data owners are data subject matter experts
- Represent the data interests of the business and take responsibility for the quality and use of data


MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021
INFRASTRUCTURE
ANALYTICS
MACHINE LEARNING & ARTIFICIAL INTELLIGENCE
APPLICATIONS – ENTERPRISE
APPLICATIONS – INDUSTRY
OPEN SOURCE
DATA SOURCES & APIs
DATA RESOURCES
MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA (MAD) LANDSCAPE
© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)
mattturck.com/data2021
FIRSTMARK

CSU The California State University
OFFICE OF THE CHANCELLOR

MAD Landscape(Data)

State of Data Engineering 2022 map

Presented by lakeFS

Ingest SaaS
Stitch
Airbyte
Fivetran
Segment
Rivery
Dataddo
datacoral
MATILLION
SNOWPLOW

Ingest Tech
Google Cloud Pub/Sub
Apache Storm
Flink
amazon Kinesis
beam
Materialize
kafka
Spark
BENEATH
Airbyte
upsolver
CONFLUENT

Object Storage
IBM Cloud Object Storage
wasabi
SEAGATE
MINIO
Alibaba Cloud
Google Cloud Storage
zadara
hadoop
ORACLE CLOUD
amazon S3
filebase
cloudfiles
DigitalOcean
CrowdStorage
SwiftStack
LYVE
ceph
PURESTORAGE

Metastore
HIVE
Cloud Dataproc
Amazon Glue
Azure Purview
CLOUDERA
databricks

Open Table Formats
ONEHOUSE
Apache hudi
Apache orc
Tabular
DELTA LAKE
databricks
ICEBERG

Compute
databricks
RAY
hadoop mapreduce
Cloudwick
DASK
Cloud Dataproc
Azure HDInsight
akka platform
trino
ASCEND.IO
Apache Spark
amazon EMR
CLOUDERA

Analytics Engine
databricks
druid
Google Big Query
dremio
amazon ATHENA
starburst
pentaho
Qubole
ClickHouse
amazon REDSHIFT
FIREBOLT
Starburst
snowflake
VARADA
cloudera IMPALA
pinot
Apache Kylin

Git for Data
lakeFS
Project Nessie

Orchestration
Apache Airflow
Flyte
OOZIE
PREFECT
dagster
Luigi
ASTRONOMER

MLOps End-to-End
Google colab
OctoML
ABACUS.AI
METAFLOW
Disdat
Verta
cnvrg.io
mlflow
snorkel
SELDON
DataRobot
vertex.ai
W&B
HOPSWORKS
Hugging Face
Kubeflow
Google Data Studio
CLEARML
Valohai
neptune.ai
DOMINO
H2O.ai
Pachyderm
Amazon SageMaker
FLOYDHUB
hydrosphere.io
ZenML
Michelangelo
Kedro
comet

Data Centric AI/ML
DVC
Pachyderm
ACTIVELOOP
GRAVITI
DAGsHub

ML Observability
mona
Superwise
fiddler
truera
deepchecks
WHYLABS
Arthur
galileo
GANTRY
arize
ROBUST INTELLIGENCE

Notebooks
HEX
jupyter
noteable
count
Deepnote
Apache Zeppelin

Analytics Workflow
dataform
dbt
databricks
Querybook

Discovery & Governance
MARQUEZ
BigID
Collibra
magda
Apache Atlas
Meta
Amundsen
OKERA
Alation
data.world
ckan
Meta Cat
Cloud Dataproc
Amazon Glue
ckan
IMMUTA
atlan
DataHub
IBM Watson
boomi
Acryl Data
Platform
SELECT STAR

Quality
Bigeye
OwlDQ
MONTE CARLO
unravel
lightup
Databand
Datafold
Collibra
WHYLABS
SODA
redata
griffin
Metaplane
awslabs/deequ
TORO
acceldata
elementary OS
GriFFin
HoloClean

CSU The California State University
OFFICE OF THE CHANCELLOR

References

- DSouza, S., Fung, M., & Repaka, B. (2020, November 20). *Blogs*. Amazon. Retrieved November 1, 2021, from https://aws.amazon.com/blogs/publicsector/modern-data-engineering-higher-ed-dataops-data-lake/.
- Barth, A. (2020, October 26). *Owning your own (data) Lake House*. SlideShare. Retrieved October 13, 2021, from https://www2.slideshare.net/sawjd/owning-your-own-data-lake-house.
- Inmon, B. (2021). *Building the data lakehouse*. Technics Publications.
- Turck, M. (2021, September 28). *Red hot: The 2021 Machine Learning, AI and Data (MAD) landscape*. Matt Turck. Retrieved October 13, 2021, from https://mattturck.com/data2021/.
- Orr, E. (2022, June 22). *The State of Data Engineering 2022*. Lakefs. Retrieved July 22, 2022, from https://lakefs.io/the-state-of-data-engineering-2022/

- August 13th, 2022 at USC
- Full day of conference talks focused on data with tracks in
  - Data Engineering,
  - AI/ ML/ Data Science,
  - Data Infrastructure & Security,
  - Data 4 Good,
  - BI/ Visualizations/ Use Cases,
  - Emerging Tech.
- For more info go to https://www.dataconla.com
- Use complimentary code(DCLA202219SCALE@lug). Valid for 20. expires July 31st.

Q&A