Modern Data Engineering – Concepts, Best Practices & Applications

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

- 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
About California State University

• Public University
• Largest 4-year degree program in the nation with nearly half a million students and 50,000 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
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
DATA INGRESS

Ingestion Process

Heterogeneous Data Sources
- Oracle DB
- SQL Server
- MySQL
- Postgres
- CSV/delimited files
- XML
- JSON
- Parquet
- Avro
- Log data
- Data streaming
- More ….

AWS DMS

AWS Data exchange

AWS DataSync

S3 Bucket Source

Data format
- CSV/delimited files
- XML
- JSON
- Parquet
- Avro
- Log data
- Data streaming
- More ….

Vendor Software

REST API

Ingestion Process

Data Ingestion Tools or Process

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

Start

Developer create Pull Request for changes from developer's branch to dev branch

Notify Approvers to approval pull request

Developers review comments and commit changes

Run unit testing and code coverage

Pull Request Approve?

Test Result?

Yes

Auto Update: Pull Request comment and merge from developer's changes to dev branch

No

Failed

Submit EMR steps on dev EMR Cluster

Success

CloudWatch

SMS

S3

Lambda

CodeCommit

CodeBuild
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
DATA EGRESS: WORK FLOW

Amazon EMR Cluster

EMR egress step starts after daily loads complete

Other EMR prior load steps

Amazon S3

Curated data files ready at S3?

Y

N

Error handling

AWS glue data catalog

Secrets retrieved from SecretMgr?

Y

N

Error handling

AWS Secrets Manager

Use secret key values to connect

Postgresql (amazon RDS)

Overwrite to schema

Create unique indexes

Grant accesses

Postgresql

Error handling

Amazon QuickSight
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
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

- Master Data Management (MDM)
- Metadata Management
- Data Lineage
- Data Quality
- Data Catalog
- Data Security, Privacy & Compliance

Workflow

Audit

Data

SOA Services

Application DBs
IMPORTANCE OF A DATA CATALOG
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 & DATA (MAD) LANDSCAPE
References

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.