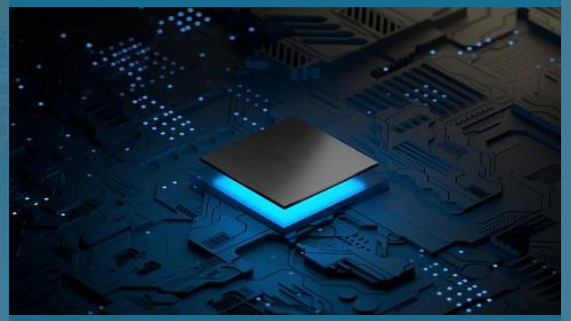
Sandia National Laboratories

Sandia Insights – A Data Sciences Architecture and Framework



National Laboratories Information Technology (NLIT) 2022

October 17, 2022

Dr. Patrick Carlson - Sandia National Lab

Dr. Ian Brooks – Cloudera Government

U.S. DEPARTMENT OF ENERGY | National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Presenters

Dr. Patrick Carlson

- PhD graduate in Human-Computer Interaction (HCI) from Iowa State University
- Worked at Sandia National Labs as a Data Scientist since 2017
- Leads the Sandia Data Sciences Community of Practice (CoP)
- Interested in the technical aspects of Data Science such as data engineering, programming, analysis, modeling, machine learning, productionization, ml-ops, data-ops, and more

Dr. Ian Brooks

- PhD graduate in Computer Science from the University of North Texas
- Focused on Big Data solutions since 2015
- Previous industry roles include Software Engineer, Data Architect, and Data Scientist
- At Cloudera, he is a Principal Solutions Engineer on the Public Sector team and a Machine Learning SME

Why Sandia Insights

- Reduce duplication and rework
- Faster turnaround time

Data Scientists and Data Engineers

- Rapid access to data
- Robust tooling
- Common environment with libraries preinstalled
- HPC and model training

Data Governance

- Data security
- Centralized access controls
- Central data catalog management

Leadership

- Data-driven decision making

Vision for Sandia Insights

- Empower Data Scientists and Analysts through easy access to vetted enterprise datasets
- Secure and document data through Data Governance and a data catalog
- Productionize and operationalize models, use data-ops and ml-ops practices
- Build out a pipeline and best practices for the entire data-lifecycle
- Focus on people and methods, not specific tools

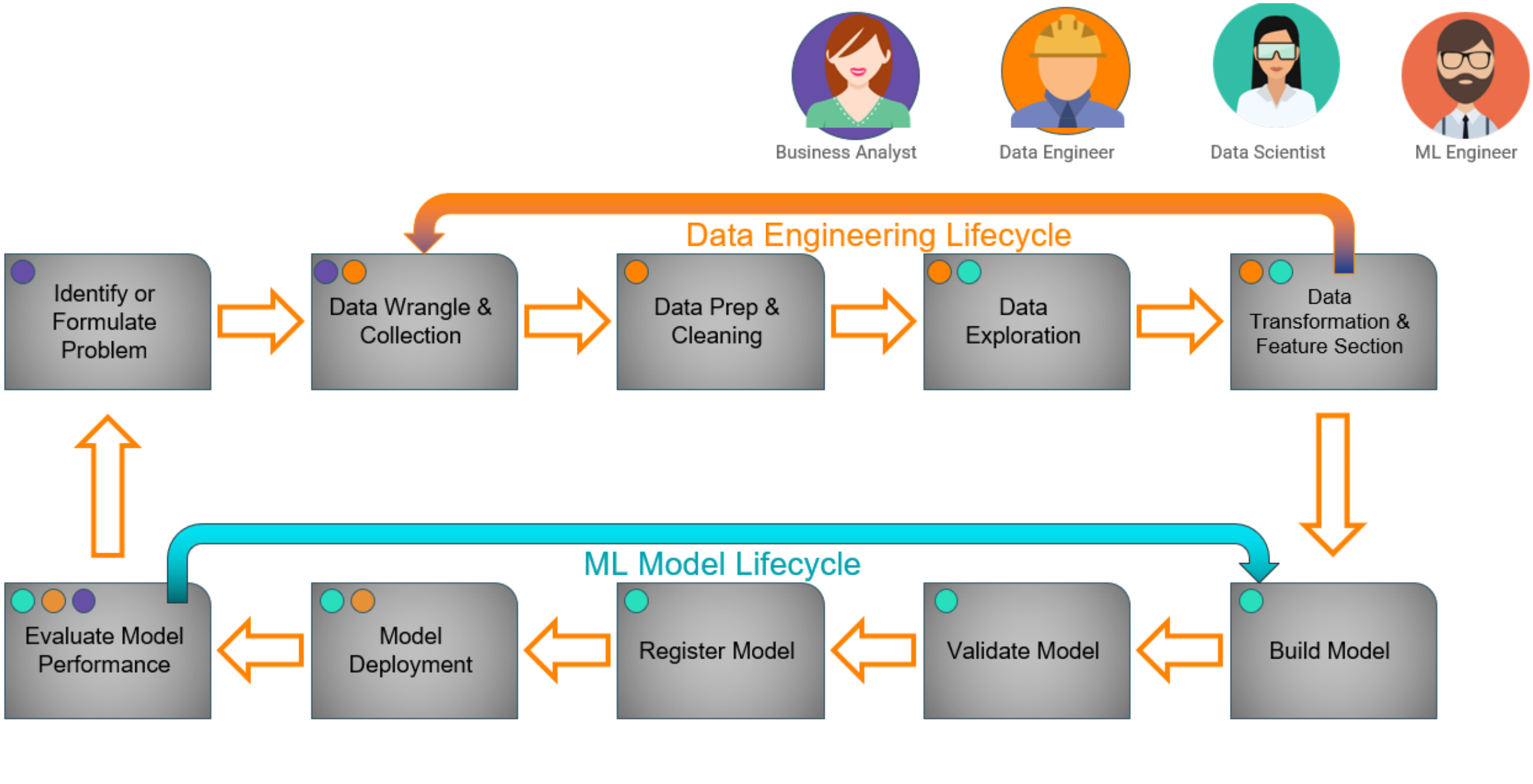
History of Sandia Insights

- 2015 – Tableau Server
- 2016 – Analytics for Sandia Knowledge (ASK)
  - Custom developed catalog
  - Tableau visualizations and reports
  - Vetted enterprise datasets via Data Central
  - Expertise finder and co-author network
- 2018 – CKAN data catalog
- 2018 – RStudio Connect (previously called Shiny Server)
- 2019 – Cloudera installation
- 2020 – Collibra data catalog

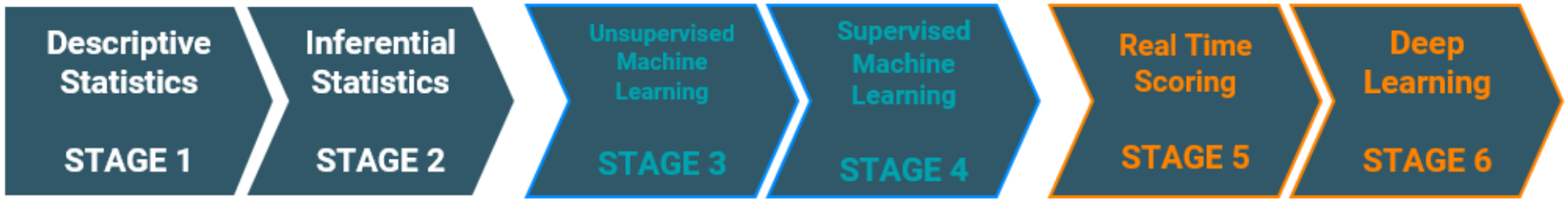
Current State

Data Lifecycle

Advanced Analytics Data Lifecycle

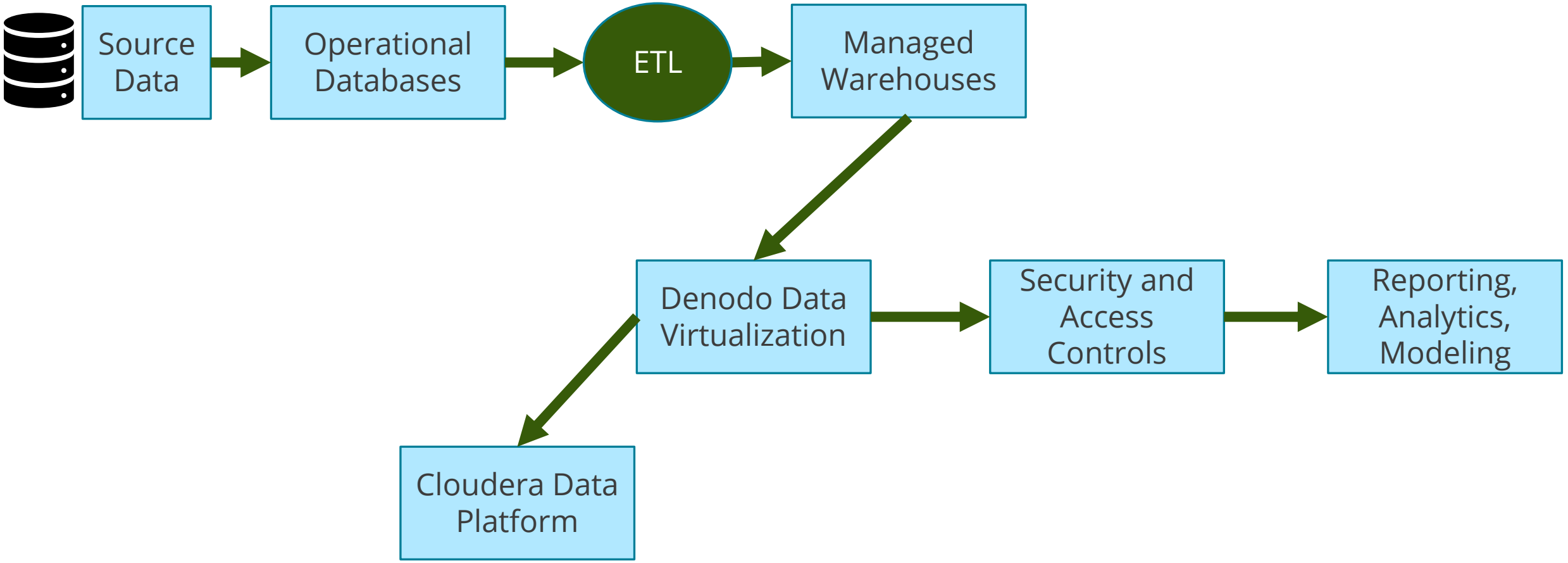


Progression of Enterprise Data Science Practices



	Phase 1	Phase 2	Phase 3
Goals	Gain Insight Into Data	Enable Proactive Approach	Enable Sub Second Decision Making
Deliverables	Reporting / Charts	ML Models	DL Models and Real Time Scoring Engine
Tools	SQL, CLI, and BI	Notebooks, IDE and ML Libraries	Model Management, Streaming Data Management, RT Scoring Engines, DL Libraries, GPUs

Current Data Pipeline



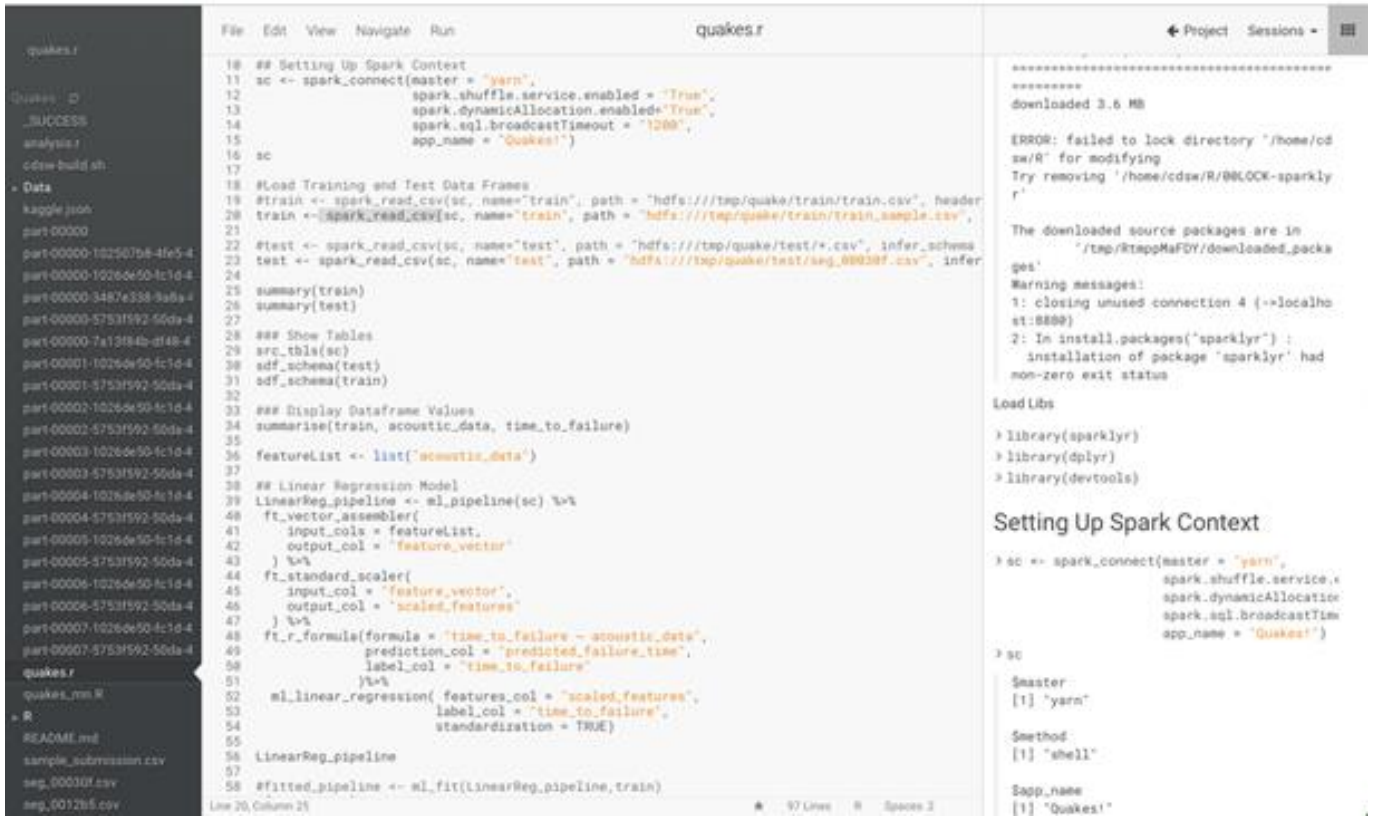
Challenges

- Existing legacy ETL process
- Batch ETL process is well suited for slowly changing datasets or tables (once a week updates)
- Batch ETL process ties up system for many hours, prevents usage of data for reporting and analytics
- Existing process is rigid and prevents agile development

Future State

Productionization of Data Science and Collaboration

- ml-ops – productionizing machine learning applications
- Cloudera Data Science Workbench (CDSW) – containerized environment for Data Scientists, collaborative development
- Future MLFlow support for experiment tracking and hyper-parameter tuning

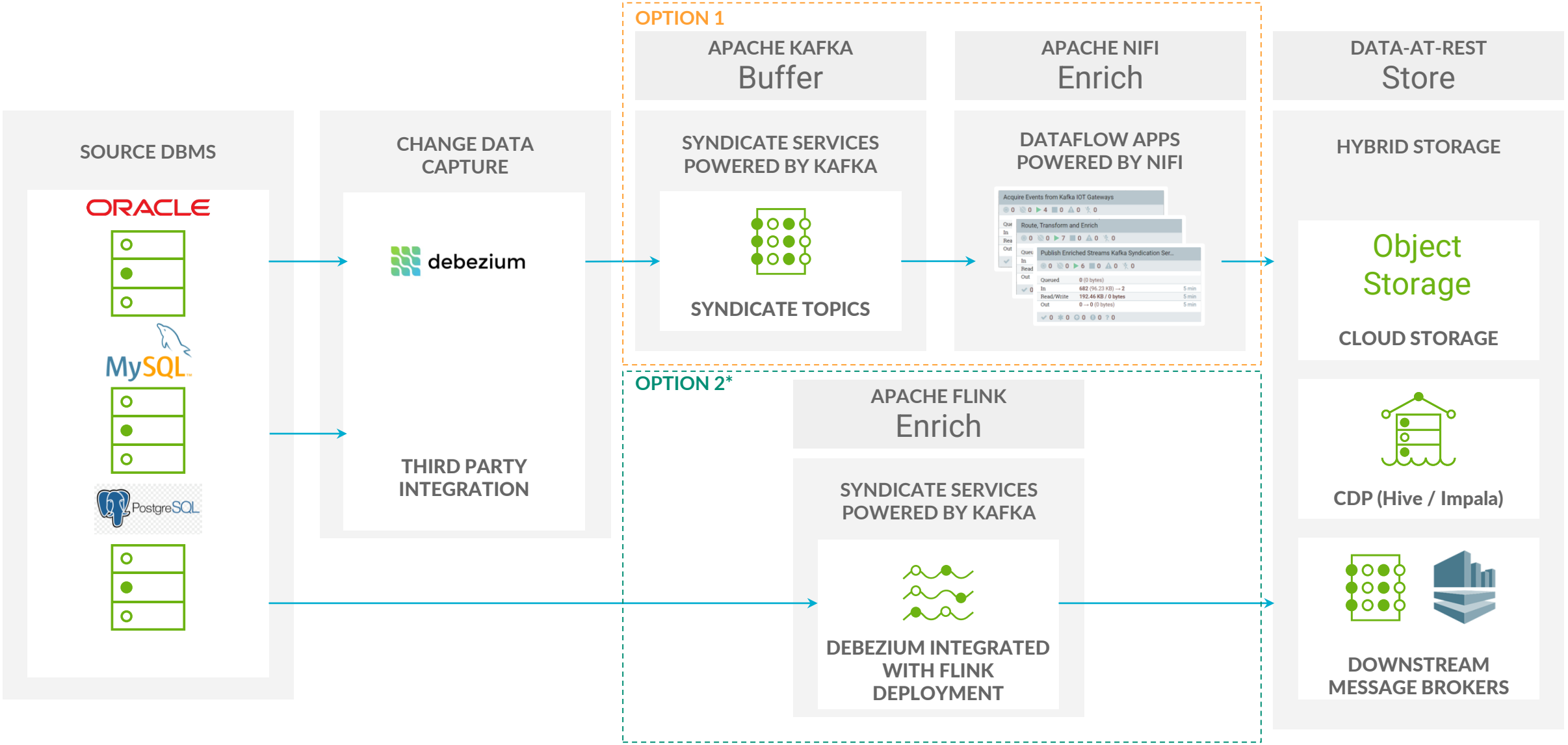


mlflow™ dask APACHE Spark™ Scala python R

Data Governance and Cataloging

- Document more datasets with descriptions and metadata
- Leverage Denodo as singular integration point for structured data and access controls
- Tagging data in Denodo and Cloudera (Apache Atlas)
- Test attribute-based access controls in Denodo and Cloudera (Apache Ranger)
- Working on labels and sensitivity categories to address Controlled Unclassified Information (CUI) requirements

Change Data Capture (CDC)



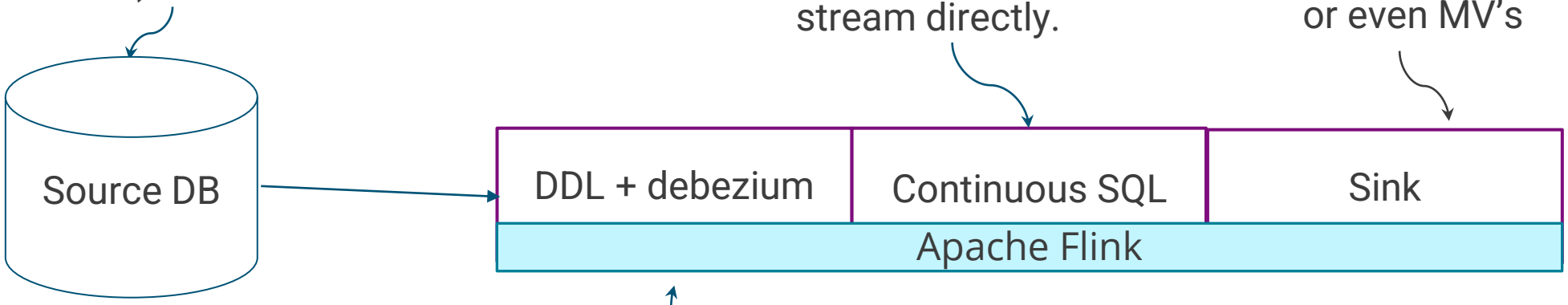
Integrated Change Data Capture

Key Takeaway; Run CDC right from Flink
note: no Kafka between Flink and DB, no connectors, etc.

Capture source changes in native change log (WAL, Logminer, etc)

SQL queries on CDC stream directly.

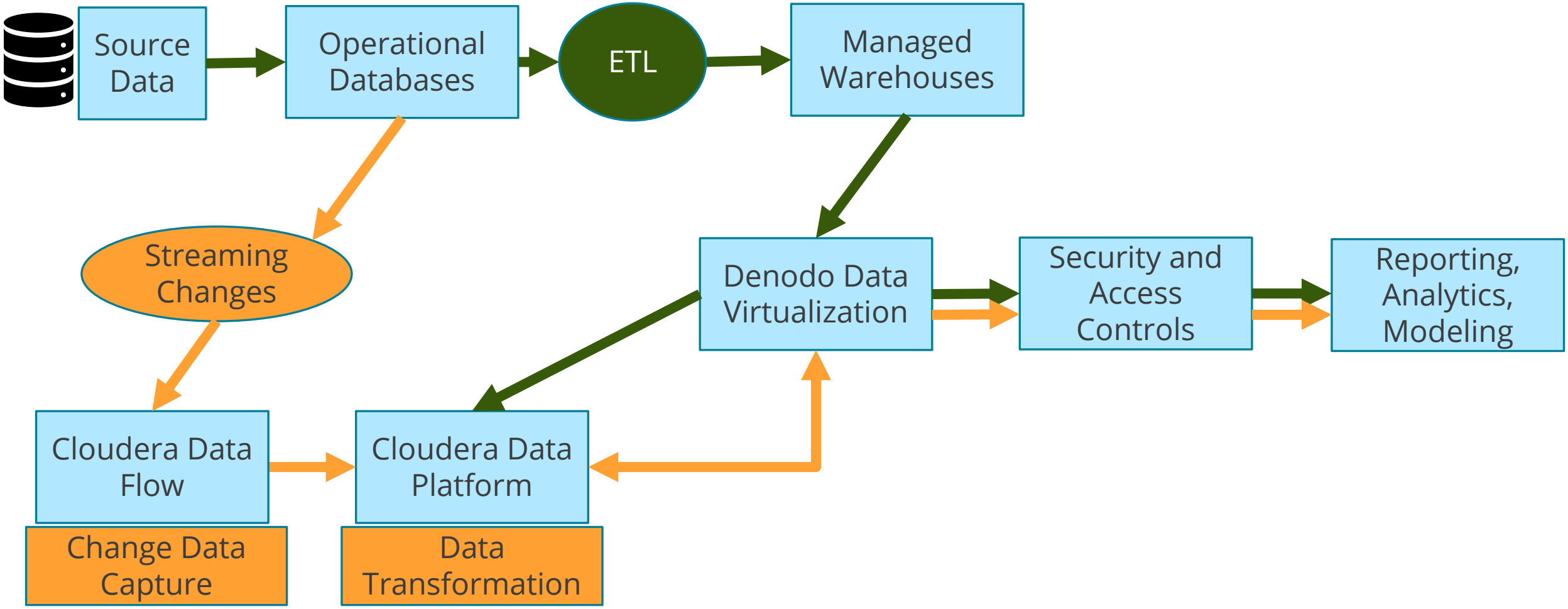
Route output to any sink, including Kafka or even MV's



Debezium as a library running *inside* Flink, capturing changes as a direct source

- Oracle, PostgreSQL, MySQL, DB2, MSSQL
- Logminer (open) and Xstream (licensed) interfaces for Oracle
- Uses DDL Interface directly, one DDL configuration per CDC stream.
- Can specify query on capture

Sandia Automated Data Pipeline



Demo or Video of CDC

Impact of Automated Data Pipelines

- Rapid access to current data enables reporting and analytics access to quickly changing datasets or tables
- Providing robust tooling to augment and not interfere with existing data pipelines such as ETL
- More agile development of future data pipelines
- Ensure end-to-end security and governance throughout entire pipeline
- Provide leadership with real-time dashboards and decision-making opportunities

Lessons Learned and Best Practices

Challenges	Actions
Culture of “ownership” of data	Implement Data Governance
Pockets of Data Scientists around the lab	Data Sciences Community of Practice (CoP)
Lack of leadership support and direction	Get a Chief Data Office (CDO) or Deputy Chief Data Officer
Lack of documentation around data	Implement a data catalog
Data architecture is old	Understand architectural paradigms: ETL, ELT, data warehouse, data lake, lambda and kappa architectures, etc.

PRIORITIZE:

People
Processes
Architectures
Culture
Results

>

Tools
Software
One-Off Projects

Comments or Questions?

pcarlso@sandia.gov

https://carlsonp.github.io
https://www.linkedin.com/in/patrick-carlson-13753628

ibrooks@cloudera.com

https://github.com/BrooksIan
https://www.linkedin.com/in/ianrbrooksphd/