Washington State Department of HEALTH

LESSONS LEARNED FROM IMPLEMENTATION OF A CLOUD ANALYTIC ENVIRONMENT

Peter Dieringer, MPH

Agenda

Background

History

COVID-build

Post-COVID

Future States

Lessons Learned

Washington State Immunization Information System (WAIIS)

- Voluntary web-based lifetime immunization registry for Washington residents of all ages.
  - Over 180 million immunization records
  - 13 million unique individuals
- System users
  - 2,500 health organizations with 5,900 facilities exchanging data
  - 19,000 individual authorized entities
- Established as a two-county project in 1994 and was rolled out to all of Washington in 2004

CEDAR

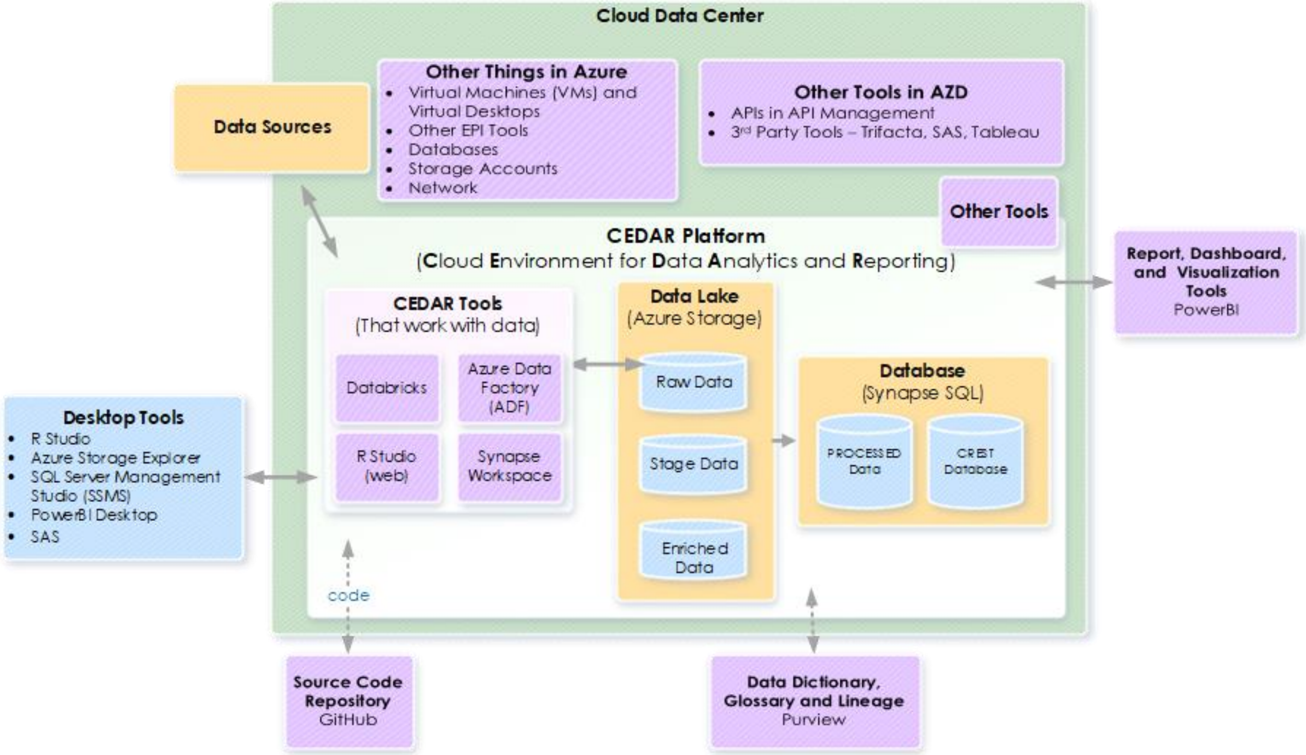
Cloud **E**nvironment for **D**ata **A**nalytics and **R**eporting

WAIIS first system to fully onboard to CEDAR

Cloud storage and computing

Rapidly established in 2020

Part of larger DOH Cloud Data Center



Cloud Data Center

Data Sources

Other Things in Azure
- Virtual Machines (VMs) and Virtual Desktops
- Other EPI Tools
- Databases
- Storage Accounts
- Network

Other Tools in AZD
- APIs in API Management
- 3rd Party Tools – Trifacta, SAS, Tableau

Other Tools

CEDAR Platform
(**C**loud **E**nvironment for **D**ata **A**nalytics and **R**eporting)

CEDAR Tools
(That work with data)

Databricks | Azure Data Factory (ADF) | R Studio (web) | Synapse Workspace

Data Lake
(Azure Storage)

Raw Data | Stage Data | Enriched Data

Database
(Synapse SQL)

PROCESSED Data | CREST Database

Report, Dashboard, and Visualization Tools
PowerBI

Desktop Tools
- R Studio
- Azure Storage Explorer
- SQL Server Management Studio (SSMS)
- PowerBI Desktop
- SAS

code

Source Code Repository
GitHub

Data Dictionary, Glossary and Lineage
Purview

Office of Immunization

- **Pre-COVID**
  - Assessment team
    - 5 Staff
  - Data Exchange
    - 3 Staff

COVID Expansion

- Assessment team
  - 20 Staff
- Data Exchange
  - 4 Staff
- Informatics
  - 18 Staff

Analytic and Reporting Needs

Pre-COVID

- School Reporting
- Childhood vaccination
- CC4 Reporting
- CC3
- Annual reports/dashboards

COVID

- Daily CDC reporting
- Daily COVID Dashboards and Reports
- Local Health Jurisdiction analytic files
- Data quality reports
- Vaccine Ordering and Inventory reports
- Vaccine Allocation
- Data Requests
- HEDIS Matching

Pre-COVID

- Direct access to production WAIIS
- Typical analytic workflow:
  - SQL Developer to generate files then SAS to analyze data
- Large queries would run overnight and off hours to prevent impacts to WAIIS
- CC4 processing took ~2 weeks to complete
  - Batches of files nightly
  - SAS to combine and manipulate data

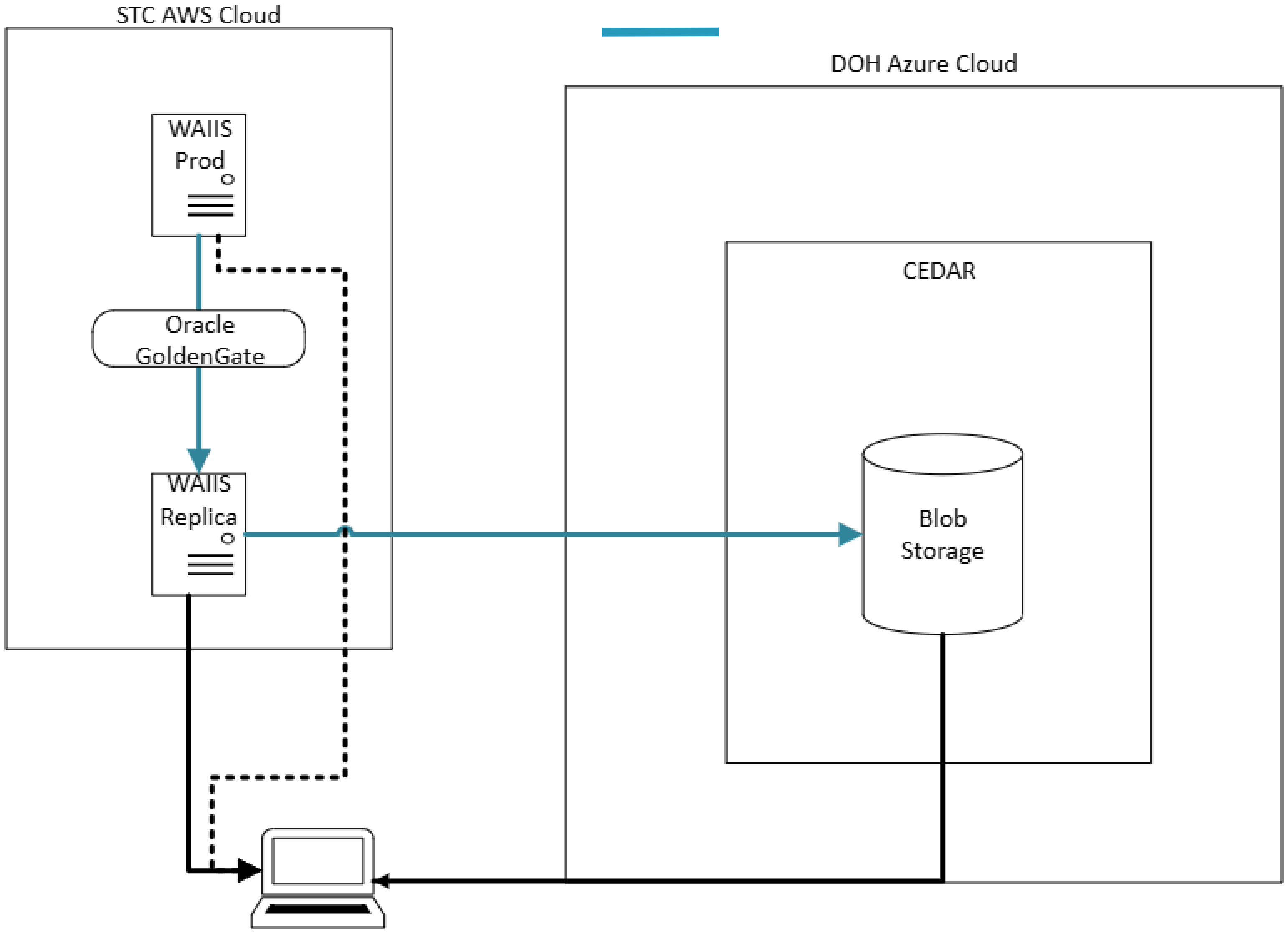
Pre-CEDAR

- COVID vaccine roll-out
- Immediate need for COVID analytic datasets and history of CDC reporting
- Creation of COVID vaccine repository
  - SQL Database
  - COVID19_vax_admin table
    - Daily append of COVID administration data
    - History of CDC reporting
  - Nightly query of WAIIS
- Need for access to real-time WAIIS data for WAVerify
  - Real-time replica database

Issues

- Manual processing of files
- Reporting team would generate reporting file
  - Submit to CDC
  - IT append reporting file to COVID Vaccine Repository
- All teams would wait for process to complete prior to generating analytic datasets

First Iteration of CEDAR



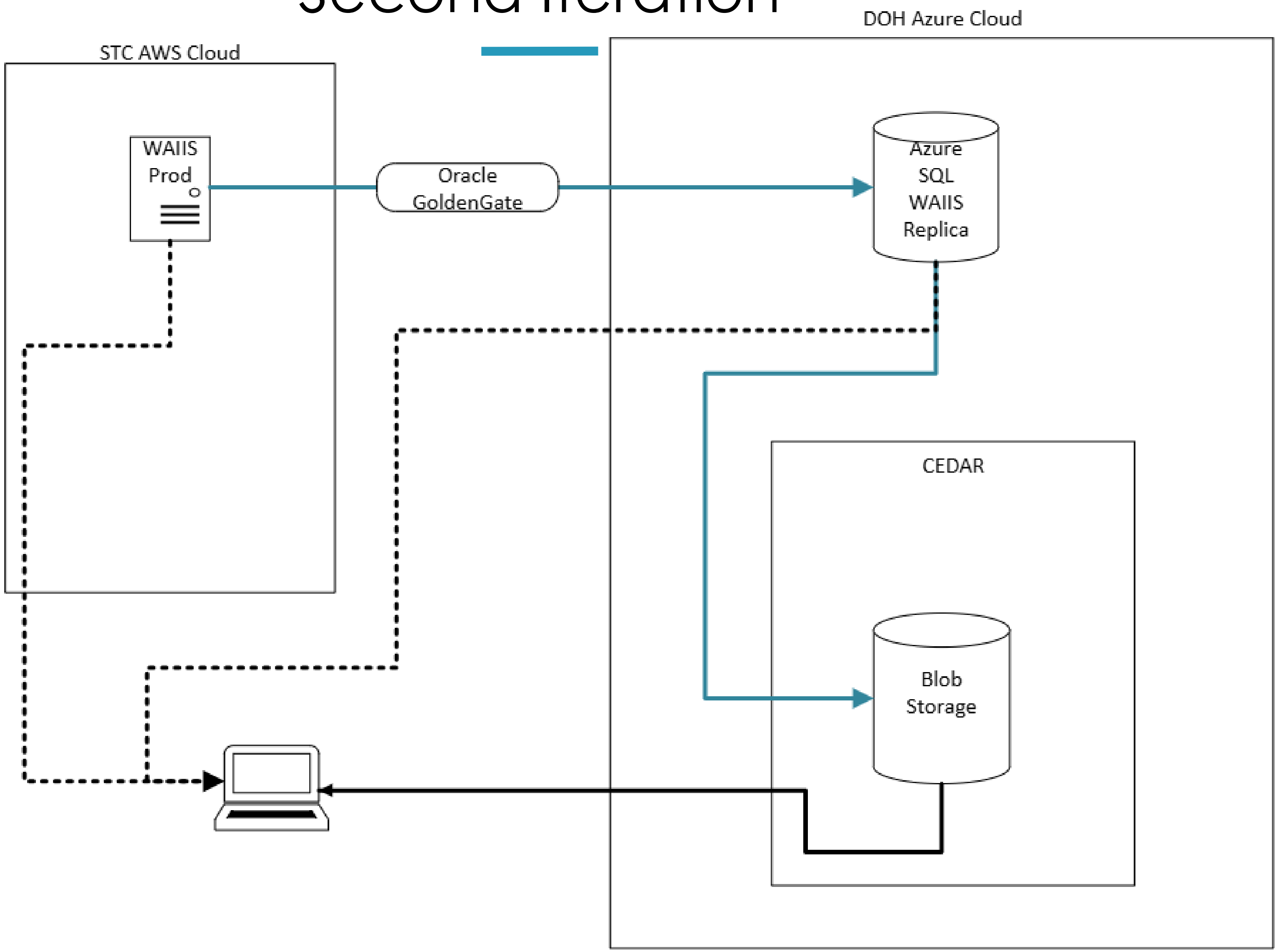
First Iteration of CEDAR

- Daily “kill and fill” of IIS tables
  - Managed by IT
- Analytic tables
  - Managed by Immunization Informatics
  - COVID19_vax_event
- Folder access:
  - Assessment read access to raw and processed data folders, write to outputs folder
  - Informatics read access to raw, write to processed and outputs folder

Issues

- Confusion between raw and processed folders
  - What is the source of truth?
  - Troubleshooting code and processing issues
- User account access
  - Overwhelmed IT services with account management
- Pulling large amounts of stale data everyday
- Missing IIS tables

Second Iteration



Second Iteration

- Implementation of Change Data Feed
- Redesigned folder and access structure
- Assessment
  - read access to processed data
  - write access to output data
- Immunization Informatics
  - read access to raw
  - write to processed data
  - write to output

Issues

- Databricks Unity Catalog is not enabled
  - Required for many of Databricks' new functionality
- Limited access to Azure real-time replica data
  - Concerns of slow-down for WAVerify
  - Direct queries to WAIIS Prod for real-time data
- Duplication of data in Raw and Data Domains

Future State (Cloud 2.0)

- Architecture to support Databricks Unity Catalog
- Adoptions of Medallion Architecture
- Posit Workbench
- Charge-back for costs
- Increased control of data access and sharing

Lessons Learned

- Steep learning curve for new technologies
- Prioritize needs of end-users
- Continued maintenance and monitoring is required
- SDLC and CI/CD do have a place
- Plan for the future

Steep Learning Curves

- Most staff did not have experience with cloud tools
- Consider is it worth training on new systems or making old systems work?
  - Adapting SAS queries to PySpark and Spark SQL in Databricks
  - Using SAS to connect to Databricks via ODBC
- Troubleshooting isn't always straightforward
  - Python vs Spark vs SAS errors
  - IT staff learning at the same time
- "Need to know"
  - Not all staff need to understand datalakes, parquet files, delta files, and distributed computing

Prioritize Needs of End Users

- Central IT, Informatics, Assessment, Programmatic staff all have different needs
- Centralized IT may not understand different teams and their needs
- “How can we get Assessment the data they need?”
- “Epis should do Epi work”
- Not everyone wants to learn a new coding language

Continued Maintenance and Monitoring

- Nothing is stable
- You cannot prepare for every error
- More teams are involved in pushing updates to systems
    - Vendors, IT, etc.

SDLC and CI/CD

- Software Development Life Cycle
- Continuous Integration and Continuous Delivery
- Many public health staff do not have experience or training with these methods
- Does not make sense for all processes, but can greatly increase speed of delivery

Plan For The Future

- Scalability and Flexibility
  - Focus on COVID and Assessment needs limited future states
    - Adding WAIIS tables
    - Limited access to data for programmatic staff
- Balance IT security with business need
  - How can we secure data without being a roadblock
- Variable naming and datatype standards
  - Minimize changes to code for different environments and tables

Peter Dieringer, MPH

Immunization Informatics Manager

WA Department of Health

@WADeptHealth

Washington State Department of
HEALTH

To request this document in another format, call 1-800-525-0127. Deaf or hard of hearing customers, please call 711 (Washington Relay) or email doh.information@doh.wa.gov.