ASU DIVES HEADFIRST INTO DATA LAKE SOLUTIONS

MOST INNOVATIVE UNIVERSITY LAUNCHES DEVILDL TO IMPROVE THE AVAILABILITY OF DATA ACROSS THE ASU CAMPUS

Disclaimer: For visioning purposes only, document may not reflect current state of project

(Tempe, AZ) December 2019 - Arizona State University's Technology Office announced today the launch of the ASU Devil DLData Lake – a one-stop shop that empowers advanced data storage, sharing, security and analysis for users.

ASU's faculty, staff, students and partners previously struggled to leverage data effectively. Data was scattered across multiple services, many of which were difficult or impossible to query. The launch of DevilDL, a new analytics platform that automates the gathering of data from disparate data sources into one central location, gives customers a consolidated solution for their reporting and analytical needs.

"ASU is continually finding new ways to innovate and improving the data collaboration capabilities that will power so many other solutions," said Terrence Bite, Head of Data Operations for ASU. "We think the new data lake will foster collaboration across the university and good things will happen as a result. We have already identified a number of use cases staff is eager to implement, because they are now possible."

DevilDL aggregates data from internal and external sources such as SalesForce, Workday, parking sensors, waste sensors, blue light pole emergency data, and much more, into one data repository. The UTO team has developed an automated service that connects the target data source and the data lake to facilitate the flow of data. The custom middleware utilizes the combination of application interfaces and various AWS services, such as AWS Lake Formation, to generate the data lake. All of the data is cataloged and made available in a way that business analysts across the ASU organization can retrieve information from and develop meaningful insights to help make better decisions across the university.

"The new data lake makes my job so much easier says," Less Server, Director in ASU's Registrar's Office. "Prior to DevilDL, I was constantly running into roadblocks because the data I needed either wasn't available or it would take weeks to access it. Now, with diverse and detailed data readily available, I have a bird's eye view of the university which more effectively helps to shape university decisions. DevilDL enables me to gather the data I need in minutes. The extra time saved allows me to leave work in time for my son's soccer practice !"

By leveraging DevilDL, ASU can store all of its structured and unstructured data at any scale. With access to an unprecedented amount of data at their fingertips, ASU employees ranging from analysts to big data scientists are armed with information to make better decisions across the university in areas like retention and graduation rates. The data can be used to run different types of analytics - from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decision making.

We are excited about the release of this new product and the impact it will have on our students and the community at large. To stay up-to-date with the latest achievements, please follow the progress at devildl.asu.edu.

###

ASU Smart City Cloud Innovation Center

DATA LAKE INNOVATION CHALLENGE

Note - FAQ's are used to support the Press Release and identify and answer key questions and answers to support the solution development. The Data Lake Challenge was a test so the FAQ answers are not provided – however the questions themselves should be helpful in designing and developing a data lake.

FREQUENTLY ASKED QUESTIONS (FAQs)

1. How will we prioritize what data to put in the data lake?
2. What is the data governance model?
3. What middle ware service?
4. What is the workflow for capturing data lineage?
5. How is the data protected?
6. Who can access the data?
7. Is there any data scrubbing needed?
8. How do we notify users of the data?
9. How do you know what types of data there is to work with?
10. Which data sets will we start with?
11. Who decides priority and/or relevance?
12. What metadata do we require for adding data to lake?
13. How do we capture that metadata? (workflow, tech)
14. Do we enforce "no metadata, no lake"?
15. How do users put QA'd/enriched/etc data back in lake for others? (workflow, tech)
16. How do we govern data put back in by users that might be sensitive? (SSN, PCI, HIPPA)
17. Do we bill users for storage over some free limit?
18. Do we need data profiling tool to catch sensitive data?
19. Do we allow sensitive data in the lake? (ISO made us remove SSN from warehouse)
20. Do we allow sensitive admin data into the lake? (HR docs -emp evals, performance improvement plans, letters of reprimand, etc)
21. How do we determine authorized users/entitlement for user generated data?
22. Mask sensitive data how? (IE SSN not necessary for analytics on rest of dataset in CSV file)
23. Do we need column/row level permissions? (Mechanism for enforcement on S3 files)
24. Do we need MDM?
25. How do we support GDPR, CCPA compliance?
26. Allow fishing expeditions? IE broad data access without stated project purpose
27. Create anonymized dataset for public researchers to use? (Kaggle contests, etc)
28. Purchased consumer reporting data governance (Equifax, Experian, etc)
29. Who can say no to putting data in lake? (Data hostage takers)

30. How to document state of data. (Raw->cleansed->QA'd->transformed->normalized->enriched)
31. How do we decide to ingest vs connecting with virtual tool?
32. Who or what level is product/program manager?
33. Should data always go to lake first before other entities that need that data?(they pull from lake)
34. Is the data warehouse part of "the lake"? (If no, is DW data pushed into lake?)
35. Criteria for adding data to data warehouse (who decides)
36. Is ingestion responsibility centralized or distributed to data producers?
37. How to measure success?
38. Data storage stages s3->glacier->?
39. Enable DL metrics (object access, total assets, asset size, unique users, etc)
40. Can DL replace BigQuery?
41. Monitoring (automation run amuck, malicious actors)
42. Change mgmt (user adds/removes/changes columns/data elements, software updates add/remove/change columns/data elements
43. What is minimum MVP that can be called a data lake?
44. Is there any data that should NOT be stored in the data lake?
45. Are there any limitations as to how much data a person/team can store in the lake?
46. What is the process for pulling data from the lake?
47. What does a user need to do to request access to data from the lake?
48. What format/s can the data be exported out in?
49. What access control mechanism will be in place for the data in the lake? (Ex. Can you mark something public/private/top secret/etc.)
50. Is the data updated in real-time?