2021-01-22

Agenda

1. Roll Call (by timezone - East to West)
2. Scribe Shout-out - It’s easy to scribe: How To Scribe Itana Calls Guide
3. **Ashish Pandit - UCSD Integration and Data Strategy.**
4. Itana Org Updates (if any)
   a. Wiki Refresh Working Group
   b. Women in EA Working Group
   c. New2EA Working Group
   d. API Working Group
   e. Business Architecture Working Group
   a. Coaching and Mentoring
   b. Working Group Updates
   c. Steering Committee Update

Attendees

1. Rupert Berk
2. Jim Phelps (UW) He/Him (Host)
3. Ashish Pandit (Co-host)
4. Brian DeMeulle (he/him/his) (Co-host)
5. Christopher Eagle (Co-host)
6. J.J. Du Chateau (Wisconsin) (Co-host)
7. Scott Lee (Co-host)
8. Mona Zarel Guerra
9. Betsy Draper
10. Betsy Reinitz-EDUCAUSE (she/her)
11. Christopher Stanley
12. Dana Miller
13. Dave Berry - University of Edinburgh (He/Him)
14. Dave Bunten
15. Dawn Hemminger
16. Diego Bartholomew
17. Erin Marogian
18. Garrett King (CMU)
19. Greg Charest
20. Irene Walsh (UC Berkeley)
21. Jake Knowlden
22. Jens Haeusser
23. John A Mobley
24. Louis King
25. Mark Grinter - Kansas State University
26. Mary Stevens
27. Miriam Clark
28. Nick Austin - Kansas State University
29. Paul H Prestin
30. Rich Cropp (Penn State)
31. Stelios Bourmpoulas (UM)
32. Stephanie Stansell - Columbia College

UCSD Integration and Data Strategy

Materials
Near Term: Enterprise Renewal.

Streaming first approach is part of the UCSD Enterprise Renewal Strategy.

Convinced execs to invest in the middleware layer.

Big rules.

1. Cloud first.
2. Open source wherever possible.
3. Free the data.

Long Term: Digital Transformation Playbook.

Value proposition: Better, faster, cheaper.

Ent App ops.
Data Ops
DataEng.
ML Ops.
Designing it for the future, for ML.

In Data Eng layer, opportunities during pandemic to join data quickly and provide value.
Focus on automation and observability.

What is streaming?

Moving away from batch.

Where are we now and where are we going?

Long term, needed data pipelines.
Seeing more and more use cases being thrown on the platform.
Activity Hub Curated View Design Concepts.

What is iPaaS?

What we call our platform, borrowed from the industry term.

Kafka.
- Fault tolerant
- Starting with smaller cluster

Nifi.
- Logistics
- Different rates of produces.
- Most important: data provenance.

WSO2 API manager.
- Going with fit for purpose.
- Looking to move to a SaaS.

GoAnywhere.
- File transfer.
- Supports legacy transfers.

Airflow.
- Looking into this. POC stage.
- Dependency management.
- Easy retries.
- Log viewing
- Pattern catalog

Target Operating Model.

Always adjusting the talent pools, along with process and tech.

Started with Nifi.
- POC.
- Took to execs for sponsorship.
- Started deploying
- Define processes.
- Training.

With each iteration on the roadmap, we had to show value: First iteration focused on speed; the next, quality; now, cost. High costs, so need to get those down.

Version 1 - 3, now working on 4.

From a handful of people. Now 150 workloads. 100 APIs. 75 people.

Pure open source early on, but that also hindered us.

Developers wanted to run current release.

Brought in service contracts to support.

Not always chasing latest release in OSS.

Instead, using a vendor supported package: Cloudera Nifi, Confluent Kafka.

Service Maturation.

First thing we needed was observability e.g. Nifi.

Today, we swivel chair because so many tools.

Need to improve our observability.

Want to move towards AI ops.

Learn from the environment to fine tune it.

Example: tried to shut down dev and QA for hana.
Brought in Airflow.

Was able to do this with Airflow, but it was difficult because users were not transparent about when they use it.

Had to look at logs for usage patterns, to get to more just in time computing.

Also, need to trim costs.

Trying to achieve just in time compute.

2 UCSD campuses using activity hubs.

Published RACIs, MOUs, SLAs.

In the past, had operational level agreements, but had to up the game.

Ashish.

CoE.

Need to foster collaboration with all the developers.

Want to expose information to campus developers.

Questions.

What are use cases that drive the streaming approach? What are real-time event needs?

- Example: LMS - Is a student in trouble?
- Example: Covid - are there water quality issues on campus?
- More mobile, more IOT for the future.

--------

Chat

Brian - any idea of the total budget for the integration / data layer(s)?

Brian DeMeulle (he/him/his) to Everyone (11:22 AM)

Construction costs ~$2.5M. Run costs currently $500k annual, but we're working on approaches to optimize and shrink that down to $300k.

Mona Zarei Guerra to Everyone (11:22 AM)

I like "design for future" mindset.

Nick Austin - Kansas State University to Everyone (11:22 AM)

I missed it. What is the ESB or ETL tool you are using to move data?

Also, can you provide 1 or 2 examples of the projects that are needing to utilize Spark?

Irene Walsh (UC Berkeley) to Everyone (11:22 AM)

for the big rule 'cloud first', was the expectation to use public cloud or hybrid cloud to maintain some focus on onprem infrastructure?

Brian DeMeulle (he/him/his) to Everyone (11:23 AM)

Public cloud initially, but we have a hybrid cloud operating model so we have flexibility to move workloads as needed.

ETL is performed through Kafka/NiFi

For remaining batch, some light ETL is available through MFT

Louis King to Everyone (11:25 AM)

Are you querying across the activity hubs or are queries limited to the activity hub?

Brian DeMeulle (he/him/his) to Everyone (11:26 AM)

We blend data across the activity hubs and materialize views for the blended data sets

Louis King to Everyone (11:27 AM)

How are you governing access to the materialized data and ensuring that it is consistent to the business rules of the enterprise purpose built systems?

Brian DeMeulle (he/him/his) to Everyone (11:27 AM)
Data access requests are vetted with the data owners/stewards, who authorize access.

Dave Berry - University of Edinburgh (He/Him) to Everyone (11:28 AM)

How long do you keep data for?

Brian DeMeulle (he/him/his) to Everyone (11:29 AM)

In the Activity Hubs, currently we are keeping everything, but we are planning on tiering out older data for cost and performance

Garrett King (CMU) to Everyone (11:29 AM)

Any tie ins with Identity & Access Management (provisioning identity data, directory services, role based access,)

Brian DeMeulle (he/him/his) to Everyone (11:30 AM)

There is some connection with our legacy IAM. We're in the process of deploying a new IAM architecture which will provision data access based on roles.

J.J. Du Chateau (Wisconsin) to Everyone (11:31 AM)

All of the diagram arrows point towards use of data, presumably READ use. How are the APIs or this infrastructure used to Create, Update or Delete data?

Brian DeMeulle (he/him/his) to Everyone (11:33 AM)

We have existing APIs for create/update into our enterprise systems. These diagrams are for data pipelines and downstream data consumption, which is why they don’t have those others APIs shown. There is no create/update allowed into the Activity Hubs.

Nick Austin - Kansas State University to Everyone (11:37 AM)

For the limited access API cataloging, do you still utilize the data steward process for access to these API's and how do you maintain the security

Me to Everyone (11:38 AM)

Is the pattern catalog publicly available? URL?

Jim Phelps (UW) He/Him to Everyone (11:38 AM)

Is your Patterns & Cookbooks site open to the world to read?

Jim Phelps (UW) He/Him to Everyone (11:39 AM)

Ha! Jinx

Brian DeMeulle (he/him/his) to Everyone (11:39 AM)

Yes. We still require approval even if programmatic access is requested for downstream apps. We allow role-based access and manage the security controls through an annual “renewal” process.

I don’t believe we have published the patterns yet. But we can look into sharing those.

Jim Phelps (UW) He/Him to Everyone (11:41 AM)

Brian - Org and Talent would be interesting to hear about

Brian DeMeulle (he/him/his) to Everyone (11:42 AM)

Sure. Happy to do that. Either during this or at another session.

Jim Phelps (UW) He/Him to Everyone (11:46 AM)

Vendor Package = WS02?

Brian DeMeulle (he/him/his) to Everyone (11:46 AM)

API management

Also, Confluent Kafka

Cloudera NiFi

Russell Connacher (he/him) to Everyone (11:48 AM)

SAP HANA seems much more expensive than say Redis, is it that much better for your use cases?

Brian DeMeulle (he/him/his) to Everyone (11:48 AM)

Currently, yes. However, we are going to be performing some experiments with alternate platforms to manage ongoing cost.

And see how the performance is on those alternate platforms
Garrett King (CMU) to Everyone (11:56 AM)

Have to sign off, thanks everyone/Ashish/Brian/Scott, very useful

Russell Connacher, UC Berkeley (he/him) to Everyone (11:56 AM)

Has there been any student push-back over privacy over the LM activity streams?

Brian DeMeulie (he/him/his) to Everyone (11:59 AM)

No. This was all vetted through our privacy office. And our privacy office is paranoid.

Russell Connacher, UC Berkeley (he/him) to Everyone (11:59 AM)

(yes, they tend to be)