ThinkChicago

8/3

- Thoughts so far:
    - Workbench may still be too complicated for this use case. We can't count on the users understanding consoles/terminals
    - If we were fully engaged in the event planning, we could've made things easier – for example, providing a pre-created database for the 2FM data instead of CSVs and PDF files
    - Hackathons are open-ended with compressed time frames – if we do these in the future, we might want to be closer to the planning.
- http://www.thinkchicago.net/#agenda
    - 9:45am - 12:30pm Civic Tech Challenge Activity

8/2 The main event

- Prompts provided by organizers
- 2pm
    - Received email about Slack invite problems
    - Students couldn't join Slack because of domain restriction. Resolved and created new invite link
- 3pm activity started
- 4pm
    - Requests for 2FM data
    - Data was accessible via /shared directory via Linux console, but not in File Manager
    - symlinked data to each home directory and uploaded table diagrams to Slack
    - Added Postgres tutorial
    - ElasticSearch crash look backoff due to OOM, updated spec for service.
- 5pm
    - Mostly quiet
    - 10 teams running various services (3 studio, 2 mysql, 7 cloud9all, 1 jupyter, 1 elasticsearch)

8/1 Redeploy

- Opted to redeploy workshop1 instance because of problems with ETK workshop. Increased docker volume size, increased Gluster volume size.
- Ran into deployment problems, apparently a Nebula issue?
- By 5pm, all OK.
- Tasks completed:
    - Redeploy
    - Resize docker volumes to 100GB
    - Tutorial examples for Cloud9
    - Updated nodeJS
    - All-in-one Cloud9 container
    - Enabled Nagios monitoring
    - Disabled ElasticSearch/ELK
    - Cache images
    - Confirm certs
    - Added Google analytics
    - Updated nginx max-body-size
    - Added 503 handling and default backend
    - Re-downloaded main datasets

TODO:

- Finalize data
- ~~Finalize catalogs~~
    - ~~Add defaultPath to dev environments~~
    - ~~Add extra ports to Cloud9 environments~~
- Add users
- Add monitoring
- Disable ELK – since we don't actually use the log data and it eats resources?
- Documentation
    - https://nationaldataservice.atlassian.net/wiki/display/NDSC/ThinkChicago
- Disable sign-up?

7/28:

- Upgraded instance to 1.0.12 (released 7/27)
- Scaled up instance, adding 3 compute nodes.
- Total resources: 72cores, 192GB RAM
- Added volumes (bricks) to Gluster to handle additional data

- Created Slack organization and #workbench channel.

7/24:

- Created custom catalog and UI
- Began downloading data
    - Data is larger than expected, will need to scale up storage
- Sent notice to ThinkChicago team of instance availability

7/21:

- Requested wildcard DNS and certs

7/20:

- Deployed initial instance
- Requested wildcard certs for workbench1.nationaldataservice.org

Original notes

What do we know:

- Civic Tech Challenge
- The dates for ThinkChicago are Wednesday August 2-4th. It was mentioned on the 6/22 call that they would have a 1/2 day hackathon (previously said most of the group work for the tech challenge would take place on the 2nd & 3rd.)
- ~200 students in teams of 10 with one developer (15-20 teams)
- "App" development to solve civic problems
- Previous events, students used personal computers with no specific prompts. In response, the organizers want to provide resources (via Workbench) and prompts (email from Amelia) and example data.
- Originally opted for dedicated VMs, but have since accepted Workbench model.

Current thinking:

- Instance of Labs Workbench deployed in NDS Hackathon space with sufficient resources for 20+ concurrent users
- Undecided:
    - Skinned instance for ThinkChicago/NCSA/NDS
- Containerized development environments
    - Possibly Cloud9 kitchen-sink (all languages)
    - Mobile development tools?
- Storage space
- Domain will likely be hackathon1.nationaldataservice.org to follow NCSA security's recommendation for short-term events.

Things we need to do:

- Address some of the open issues/lessons learned from PI4
    - Find long-term solution for email registration issue
    - Improve login/password recovery. Put username in approval email, allow login/recover by email address.
    - Workload characterization/sizing
- Deploy dedicated cluster
- DNS and TLS
- Workload characterization
    - Unfortunately, workload is totally unknown. We need to determine dataset size, anticipated development tasks, resource requirements for containers (i.e., Cloud9) for typical usage scenario.
    - This will allow us to size the cluster, set user quotas, and modify container resource constraints (CPU/RAM)
- What to do with the data?
    - Organizers mentioned that most data is available via API
    - We have the ability to mount data into Workbench now. Need to decide whether this makes sense, instead of requiring each user to download the same dataset.
- Notify Nebula team.
- Documented plan:
    - We need to be clear about what we're delivering and what level of support we're offering.

Planning with Mike:

- Wanted us to work through prompts to give them an idea of what to give them
    - A couple of development environments + tools enabling integration and support of database applications
    - Ruby thing was Hydra

- Deploy instance
- DNS TLS = hackathon1.nationaldataservice.org
- Gluster as shared volume for data
- Disable approval or pre-register?
    - 10-20 logins?
- Open issues:
    - Email issues?
    - New release/build
- Plan
    - Build new release?
    - Deploy instance
    - DNS/TLS – 7/21
    - Catalog customization – 7/28
        - IDEs
        - Databases
    - Support options:
        - Email support
        - Live chat/Slack
        - Appear.in/Video
    - Retention: up 1 week before, down 1 week after
    - Split support work
    - Remove gitter

System scaling:

- 20 people running IDEs
- Can be scaled similar to PI4
    - master1
    - node[1:4]
    - glfs on node[1:2]
    - lma node 3
    - loadbal node 1
    - flavor_small: m1.large
    - flavor_medium: n-rd1.large
    - Start with master1 + 2 compute – all-in-one

From organizers

- Below are some of the sample data sets we will be incorporating into the tech challenge. These are a few of the larger data sets, makes sense to load crimes, the historical load of all DIVVY trips and DIVVY availability, and taxi trips:
- https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
- https://data.cityofchicago.org/Transportation/Divvy-Trips/fg6s-gzvg
- https://data.cityofchicago.org/Transportation/Divvy-Bicycle-Stations-Historical/eq45-8inv
- https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew
- I think any tools you have that would help the students be able to open, visualize, and potentially interact with this data would be very useful to our students. I welcome any thoughts or recommendations you and your team may have.