DataStax

Hack-The-Planet

Holiday Hackathon

2021

GitHub Archive Integrations With Astra

Hack-The-Planet Team

George Cross

Geni Gomez

James Colvin

Steven Matison


Hack the Planet!

Use Case

As a business DataStax employs countless resources who are interacting with GitHub on business, personal, and community projects. Within our team of DA's we asked ourselves: "How could we know what our peers were working on?"

For this Holiday Hackathon we are going to try and solve this business challenge.

What projects are our peers working on in Github?

Who contributed to the repo apache/cassandra?

Who contributes to the repo stargate/stargate?

DataStax

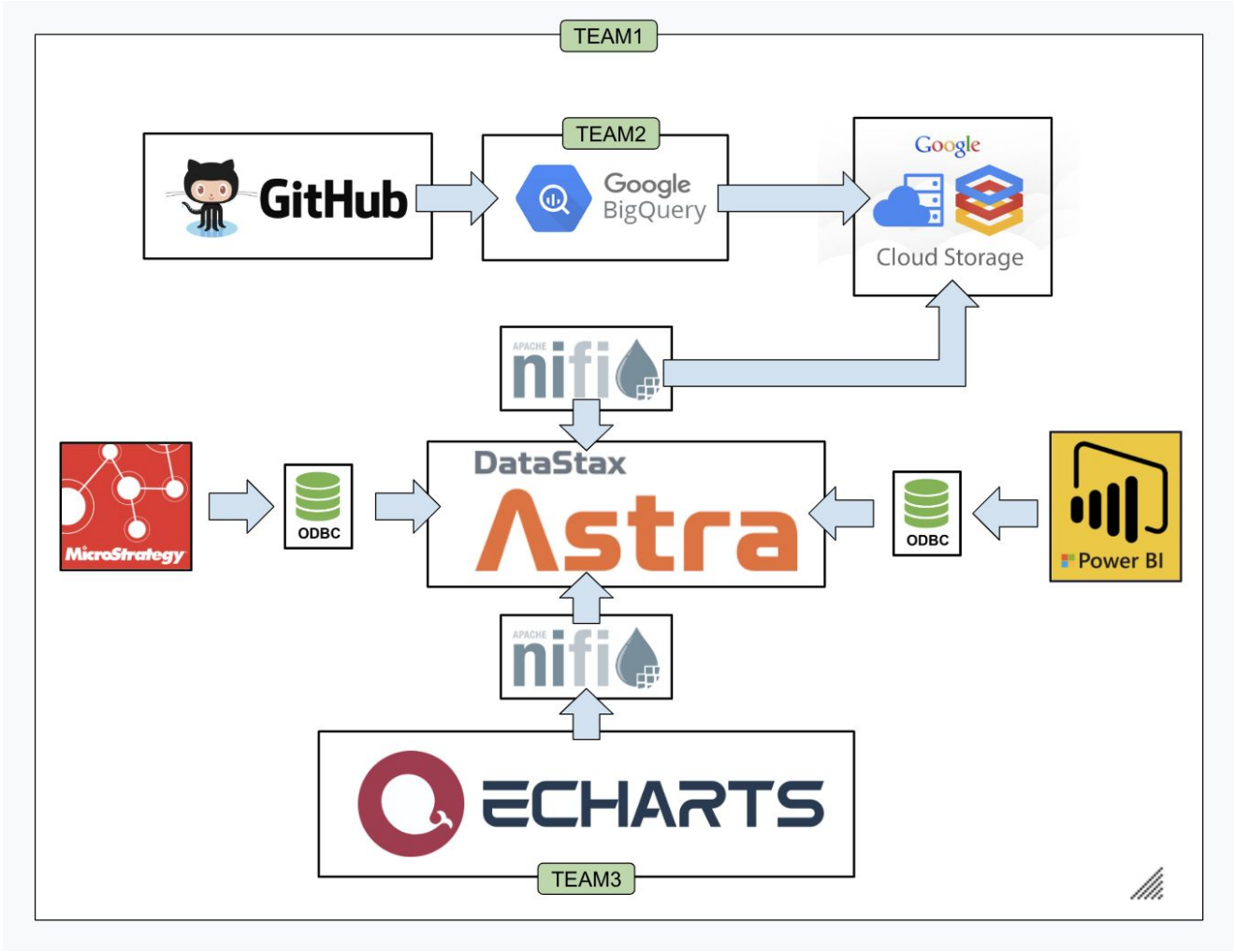
DS

GH Archive

Open-source developers all over the world are working on millions of projects: writing code & documentation, fixing & submitting bugs, and so forth. GH Archive is a project to **record** the public GitHub timeline, **archive it**, and **make it easily accessible** for further analysis.

https://www.gharchive.org/

Technologies Used


TEAM1
TEAM2
GitHub
Google
BigQuery
Google
Cloud Storage
APACHE
nifi
MicroStrategy
ODBC
DataStax
Astra
ODBC
Power BI
APACHE
nifi
ECHARTS
TEAM3

DataStax

DataStax Astra

DataStax Astra is at the heart of our hackathon providing the persistent data store behind all raw data inspection, data modeling, prototyping, and final visualizations.

DataStax

StarGate

StarGate provides API Gateway to Astra with functionality to ingest JSON and CSV data with Document API and REST API.

DataStax

GCP:
BigQuery & Cloud Storage

Google GCP provides access to Public Data Set for GitHuB Archive. BigQuery is used to build refined queries against the GitHub Archive dataset and GCS is used to store the results.

DataStax

Apache NiFi

Wrapping the heart of our project (astra) NiFi is used to automate the data pipeline from GCS to Astra. Additionally NiFi is used to provide a constantly authorized endpoint for eChart visualizations.

BI Tools
Microstrategy & PowerBI

BI Tools were used to create external access to Astra for Data Modeling and BI Tool based visualizations.

DataStax

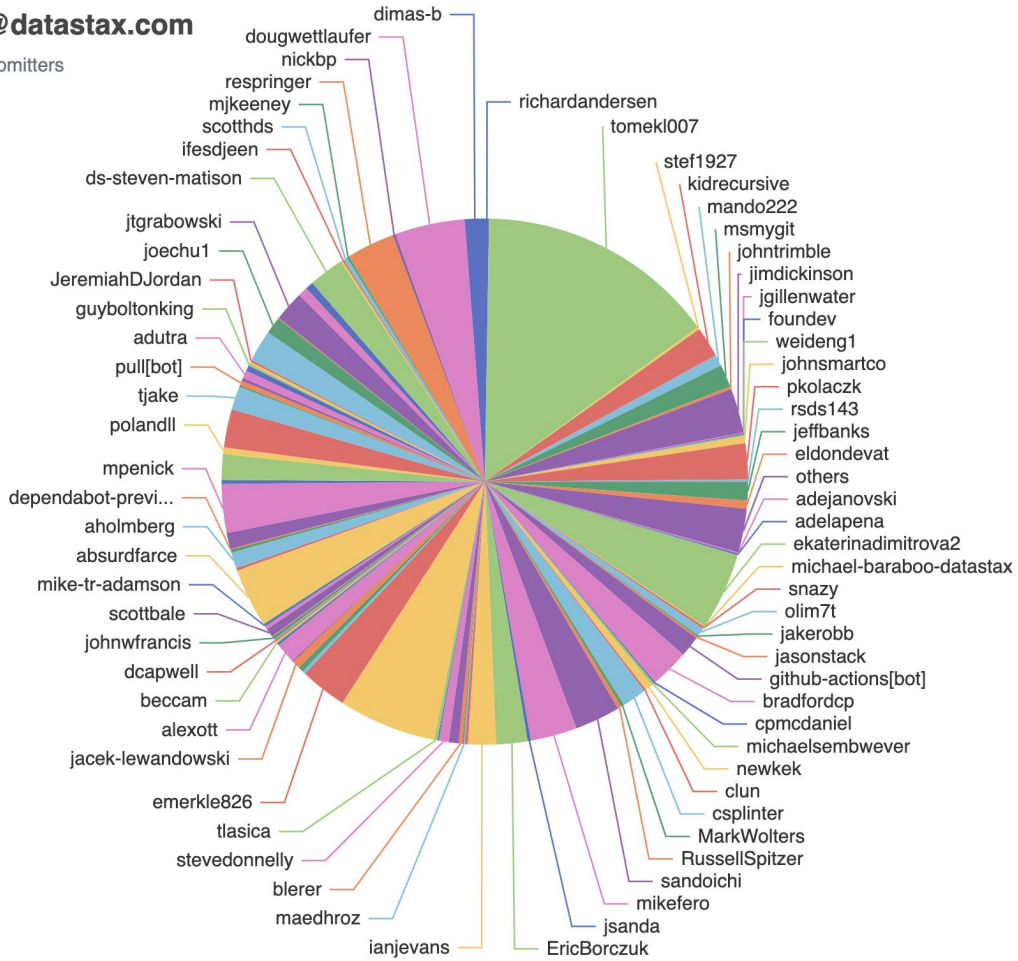
Apache eCharts

A current incubating project at apache.org, eCharts is used to integrate with Astra to provide robust charts, graphs, and powerful data visualizations without drivers, access, and authorization concerns.

eChart Visualizations

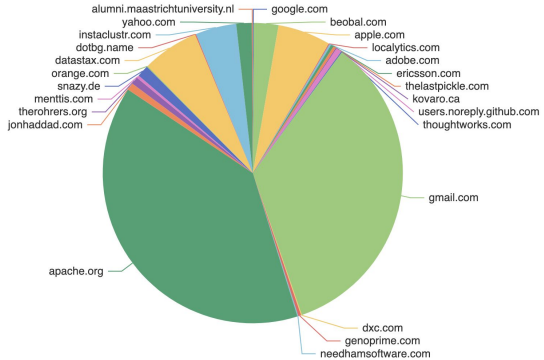
@datastax.com

Comitters



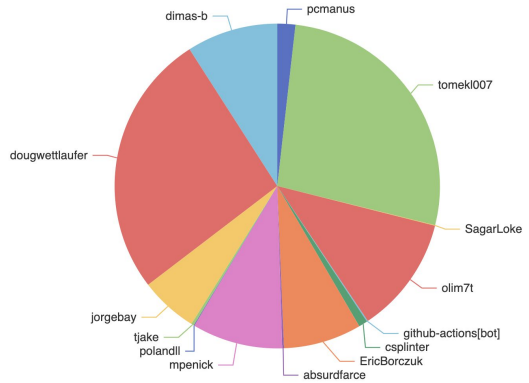
Apache Cassandra

Commit Orgs



DataStax

Stargate Comitters



Live eCharts Visualization URLs

Chart 1:

http://makeopensourcegreatagain.com/gharchive_graph1_csv.html

Chart 2:

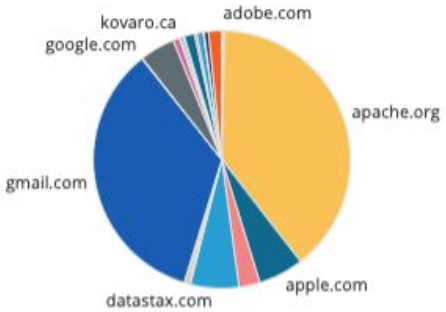
http://makeopensourcegreatagain.com/gharchive_graph2_csv.html

Chart 3:

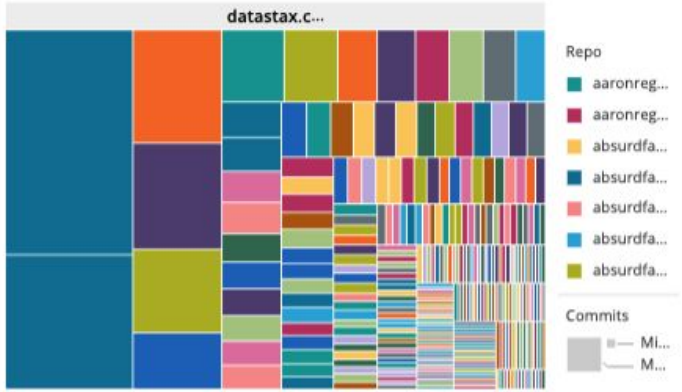
http://makeopensourcegreatagain.com/gharchive_graph3_csv.html

BI Tool Visualizations

Apache/Cassandra Commits by Domain



DataStax.com Committers

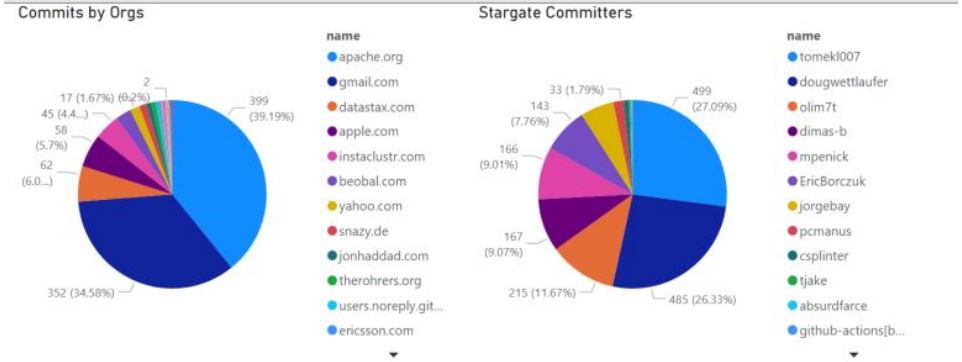


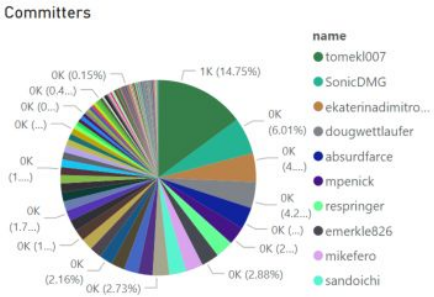
Stargate/Stargate Committers

Email	Login	Commits
0304bcc0dd0fa6af3b697e790c70cd7ceb0420c3@gmail.com	dimas-b	1
	dougwettlaufer	1
	tjake	1
17973dcf2e10b67f63417c88ca4c460b2c9439f2@users.noreply.github.com	dimas-b	7
	dougwettlaufer	30
	EricBorczuk	4
	jorgebay	2
	mpenick	2
	olim7t	6
	tomekl007	6
1fd4df949ecb58c169dc4d21d09b574a86c2406f@gmail.com	dimas-b	3

Power BI

ODBC





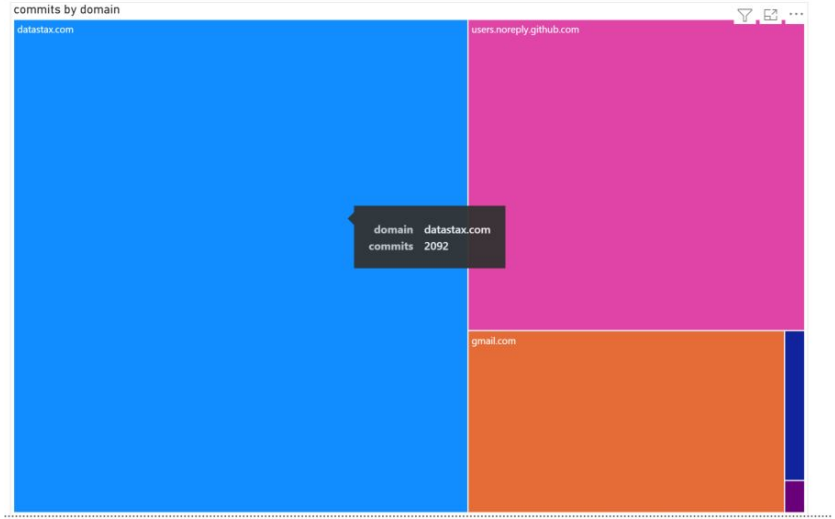
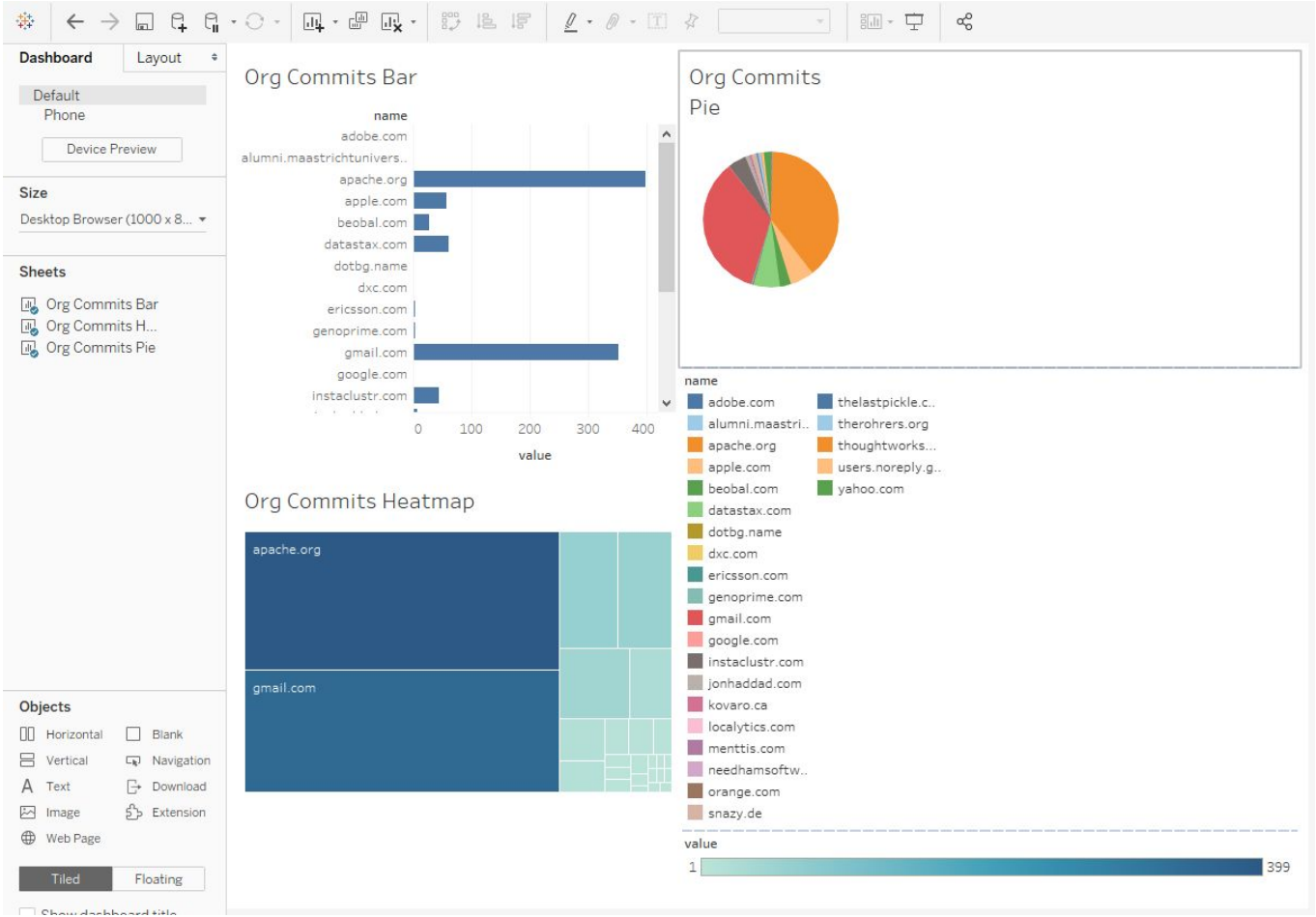



tableau
SERVER


ODBC


Dashboard
Layout
Default
Phone
Device Preview
Size
Desktop Browser (1000 x 8...
Sheets
Org Commits Bar
Org Commits H...
Org Commits Pie
Objects
Horizontal
Blank
Vertical
Navigation
Text
Download
Image
Extension
Web Page
Tiled
Floating
Org Commits Bar
name
adobe.com
alumni.maastrichtunivers..
apache.org
apple.com
beobal.com
datastax.com
dotbg.name
dxc.com
ericsson.com
genoprime.com
gmail.com
google.com
instaclustr.com
0
100
200
300
400
value
Org Commits Pie
name
adobe.com
alumni.maastri..
apache.org
apple.com
beobal.com
datastax.com
dotbg.name
dxc.com
ericsson.com
genoprime.com
gmail.com
google.com
instaclustr.com
jonhaddad.com
kovaro.ca
localytics.com
menttis.com
needhamsoftw..
orange.com
snazy.de
thelastpickle.c..
therohrers.org
thoughtworks...
users.noreply.g..
yahoo.com
value
1
399
Org Commits Heatmap
apache.org
gmail.com

How did we do this?

Team Approach

During this Hackathon our team worked independently on the following areas:

- Team 1 - Infrastructure & Automation
- Team 2 - Data Modeling & Aggregations
- Team 3 - Visualization

At the Infrastructure & Automation level we provided a fully automated raw data ingestion platform based on Apache NiFi. The data pipeline is pulling data from a GCS bucket in csv or json formats and executing the appropriate Stargate API endpoints to insert JSON (document api - no schema) or CSV (rest api - inferred schema).

On the Data Modeling & Aggregation side we viewed raw data using BigQuery, both ODBC & JDBC Drivers (raw ingested data), and Astra Studio. These tools helped us inspect the raw data to formulate final normalized tables and schemas we needed to support automated data pipeline for the final chart aggregations.

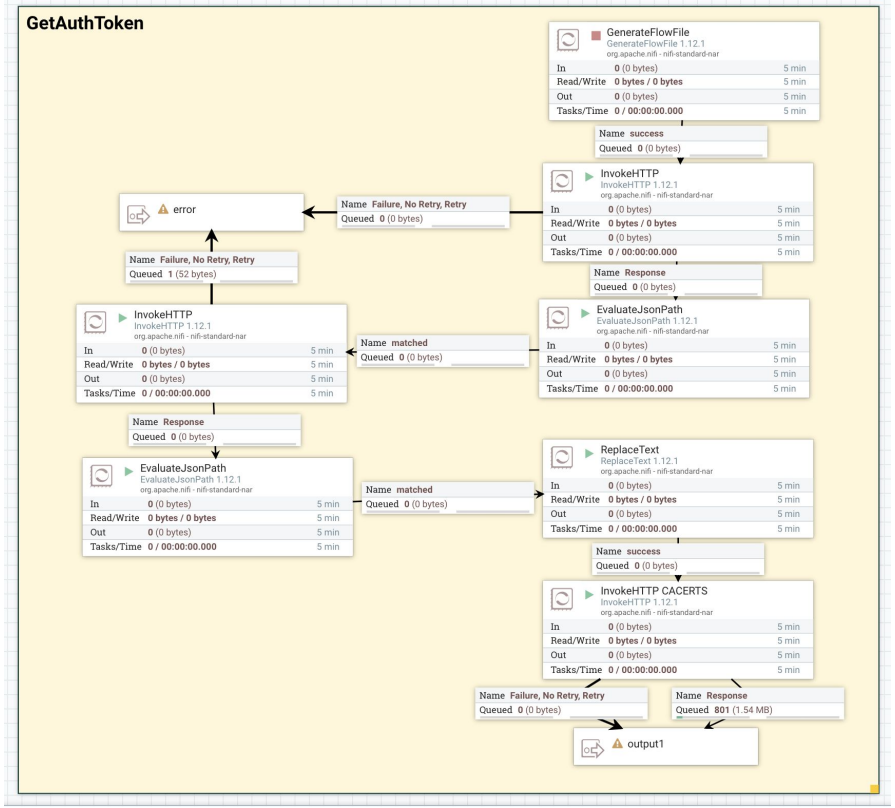
On the Visualization side we leveraged an incubating project at apache.org called eCharts which is accepting of the JSON output of our normalized data sources. Additionally, BI Tool Teams created similar visualizations to prove equal BI Tool capability.

BigQuery Queries

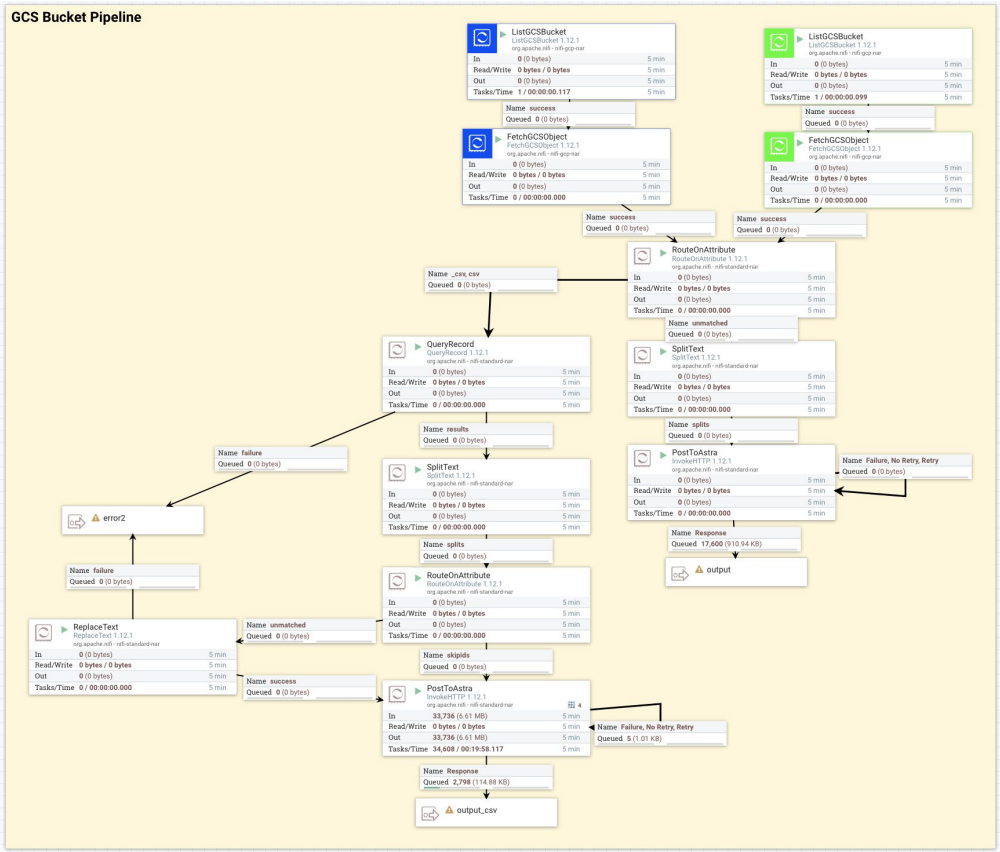
```
// master raw dataset
SELECT repo.name AS repo, actor.login AS login, actor.avatar_url AS avatar
      ,JSON_EXTRACT_SCALAR(payload, '$.commits[0].author.email') AS email
      ,LEFT(JSON EXTRACT SCALAR(payload, '$.commits[0].author.email'),INSTR(JSON_EXTRACT_SCALAR(payload, '$.commits[0].author.email'),'@')-1) AS author
      ,RIGHT(JSON EXTRACT SCALAR(payload, '$.commits[0].author.email'),LENGTH(JSON EXTRACT SCALAR(payload, '$.commits[0].author.email')) - INSTR(JSON_EXTRACT_SCALAR(payload, '$.commits[0].author.email'),'@')) AS domain
FROM `githubarchive.month.202*`
WHERE type = 'PushEvent' AND (repo.name IN('apache/cassandra','stargate/stargate') OR JSON_VALUE(payload,'$.commits[0].author.email') LIKE '%@datastax.com');


// graph 1
SELECT domain AS name, sum(commits) AS value FROM `gharchive-301514.dset_gharchive.gharchive_master` WHERE repo = 'apache/cassandra' GROUP BY name ORDER BY value desc;
// graph 2
select case when rowid < 100 then name else 'others' end name, sum(value) value FROM (SELECT row_number() over (order by value desc) rowid, name, value FROM (SELECT login AS name, sum(commits) AS value FROM `gharchive-301514.dset_gharchive.gharchive_master` WHERE domain = 'datastax.com' GROUP BY name )) group by name;
// graph 3
SELECT login AS name, sum(commits) AS value FROM `gharchive-301514.dset_gharchive.gharchive_master` WHERE repo = 'stargate/stargate' GROUP BY name ORDER BY value desc;
```

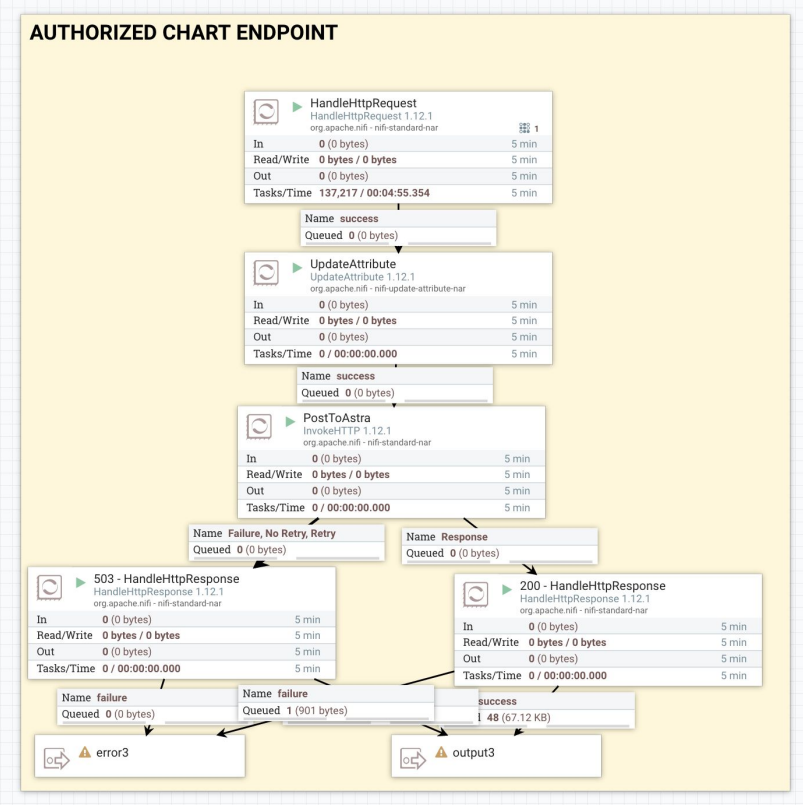
Apache NiFi Always On Auth Token



Apache NiFi GCS Bucket Pipeline



Apache NiFi eCharts EndPoint



Apache eCharts Sample Code & Data Graph 1

```
<script type="text/javascript">
var myChart = echarts.init(document.getElementById('main'));
$.get('http://192.241.141.13:8081/gharchive_graph1_csv').done(function (data) {
   myChart.setOption({
      title: {
         text: 'Apache Cassandra',
                subtext: 'Commit Orgs'
      },
      tooltip: {},
             series: [
      {
         name: 'org domain',
         type: 'pie',
         radius: '55%',
         center: ['40%', '50%'],
         data: data.rows,
         emphasis: {
            itemStyle: {
               shadowBlur: 10,
               shadowOffsetX: 0,
               shadowColor: 'rgba(0, 0, 0, 0.5)'
            }
         }
      }
   ]
   });
});
   </script>
```

```
{
    "count": 25,
    "rows": [
        {
            "name": "google.com",
            "value": 1
        },
        {
            "name": "beobal.com",
            "value": 27
        },
       ...[result set trimmed for presentation]
    ]
}
```

Apache eCharts Sample Code & Data Graph 2

```
<script type="text/javascript">
var myChart = echarts.init(document.getElementById('main'));
$.get('http://192.241.141.13:8081/gharchive_graph21_csv').done(function (data) {
   myChart.setOption({
      title: {
         text: '@datastax.com',
                subtext: 'Comitters'
      },
      tooltip: {},
            series: [
      {
         name: 'github user',
         type: 'pie',
         radius: '55%',
         center: ['40%', '50%'],
         data: data.rows,
         emphasis: {
            itemStyle: {
               shadowBlur: 10,
               shadowOffsetX: 0,
               shadowColor: 'rgba(0, 0, 0, 0.5)'
            }
         }
      }
   ]
   });
});
   </script>
```

```
{
    "count": 100,
    "pageState": "B2RpbWFzLWIA8H///5sA",
    "rows": [
        {
            "name": "richardandersen",
            "value": 18
        },
        {
            "name": "tomekl007",
            "value": 1050
        },
       ...[result set trimmed for presentation]
    ]
}
```

Apache eCharts Sample Code & Data Graph 3

```
<script type="text/javascript">
var myChart = echarts.init(document.getElementById('main'));
$.get('http://192.241.141.13:8081/gharchive_graph3_csv').done(function (data) {
    myChart.setOption({
        title: {
            text: 'DataStax',
            subtext: 'Stargate Comitters'
        },
                tooltip: {},
        series: [
        {
            name: 'github user',
            type: 'pie',
            radius: '55%',
            center: ['40%', '50%'],
            data: data.rows,
            emphasis: {
                itemStyle: {
                    shadowBlur: 10,
                    shadowOffsetX: 0,
                    shadowColor: 'rgba(0, 0, 0, 0.5)'
                }
            }
                }
    ]
    });
});
    </script>
```

```
{
    "count": 14,
    "rows": [
        {
            "name": "pcmanus",
            "value": 33
        },
        {
            "name": "tomekl007",
            "value": 499
        },
       ...[result set trimmed for presentation]
    ]
}
```

Next Steps?

DataStax

GitHub

1. Build a front end application capable of deeper drill down into data sets with live query-able interactive visualizations.
2. Track all commits to Apache Cassandra by engaging DataStax resources with personal, gmail, or apache.org accounts.
3. Expand visualizations and datasets for:
    a. Forks
    b. Issues
    c. PRs
4. Create weekly contests for:
    a. Top Committers
    b. Most Repos Touched
    c. Most Lines of Code

“Mess with the best, die like the rest…”

Dade “Zero Cool” “Crash Override” Murphy

Hackers (The Movie)

DataStax

Thank You!