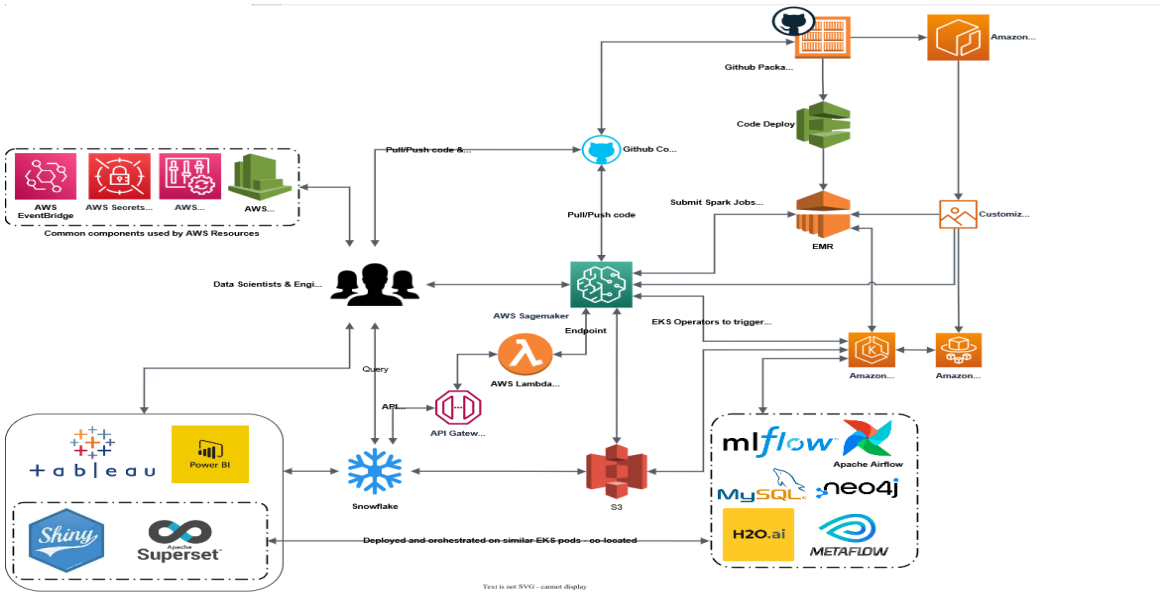
Sandeep Bansal Project Portfolio: CNN

Project 1: *Snowflake AWS Integration min sample size Automation.*



Phase I:

1. Implement lambda handler and modify test event.
2. Attach Lambda layer to lambda function in order to utilize python libraries
3. Design AWS API Gateway endpoint for snowflake to access
4. Build API method utilizing secure Resource policy
5. Design API stage to accept incoming traffic via snowflake

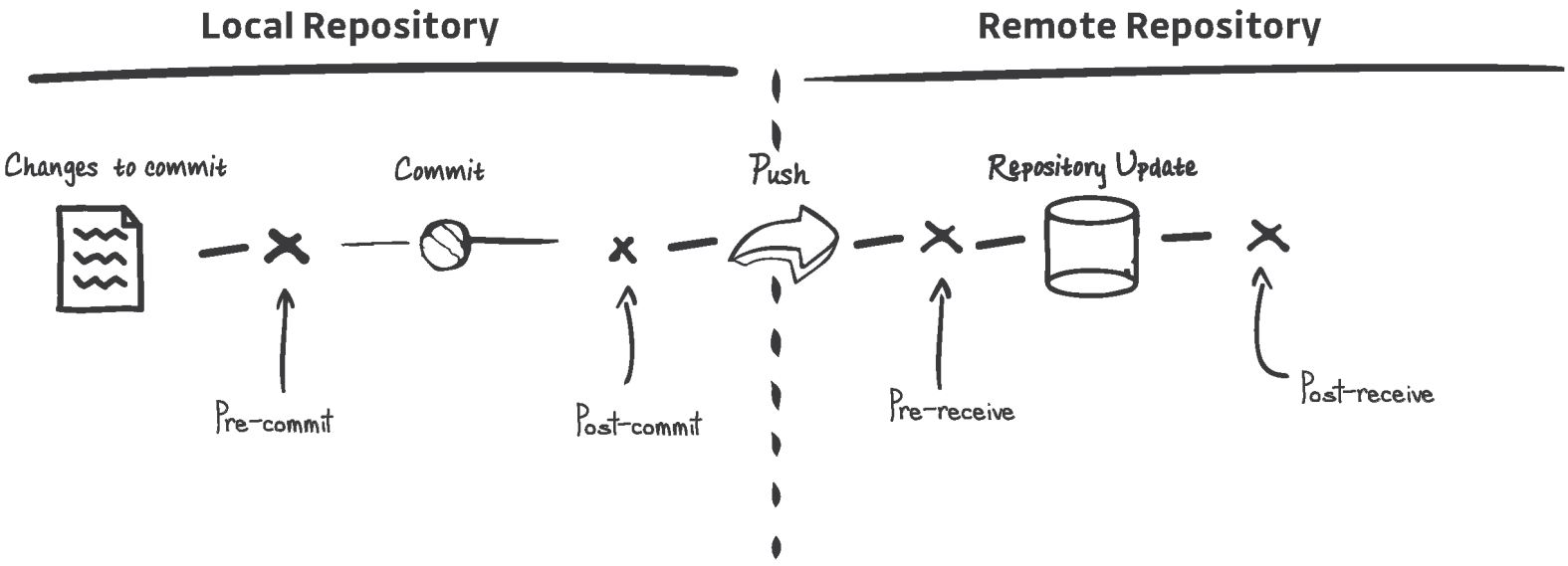
Phase II:

1. Modify and Secure API Gateway with Resource Policy
2. Construct IAM role for Snowflake use
3. Build an API Integration in Snowflake
4. Create and call an external function in Snowflake
5. Edit IAM role policy's for Snowflake external function access
6. Generate results from AWS Lambda and API Gateway to Snowflake.

Phase III:

1. Build and Launch an EC2 instance within CNN environment.
2. Create S3 bucket to hold wheel files for python dependencies.
3. Launch Instance with Amazon Linux 2 CLI.
4. Establish a Docker container on EC2 instance.
5. Place wheel files from s3 bucket to Docker container.
6. Design an Amazon Elastic Container Registry.
7. Register docker files using ECR repository push commands
8. Deploy and test lambda function within CNN environment.

Project 2: Automate git hooks to establish team best practices.



As a data science leader, I'd like to be able to easily read the commit history for a repo to understand what has happened. I'd like a commit-msg hook that checks that a commit starts with one of these 5 messages:

```
data: [...]
analyze: [...]
model: [...]
deploy: [...]
fix: [...]
```

Phase I:

1. Establish a set of team guidelines to ensure git hub commits are consistent across data science pods
2. Each git hub commit message must either contain the following comments: data, analyze, model, deploy, fix. This solidifies a team wide standard of best practice for pushing code to git hub. This also gives team lead or director an easy way to review everyone's code by being able to quickly identify code through the pre-commit git hook.

Phase II:

1. Write customized bash script for pre-commit hook.
2. Complete Quality Assurance tests to ensure that it works for each team guideline established in phase I.
3. Upon completion add it to turner git hub repository template for team members begin using.

Project 3: *Create Whitepaper Guide for CNN Digital to share & use*

Self-developed Project: This project was self -identified & self-directed in order to fill a CNN digital need for documentation.

Phase 1:

1. Establish a document that can be used for years to come by anyone who is working on AWS for CNN Digital. This documentation will clearly outline what projects have been done to ensure clear communication and synergy for all teams on CNN Digital.
2. Create easy to read find information with code snippits embedded to ensure global understanding of each project.
3. After creating this document, I thought I would add the projects and tools I have utilized on AWS in order to pass on to someone else who decides to create so they can also document.

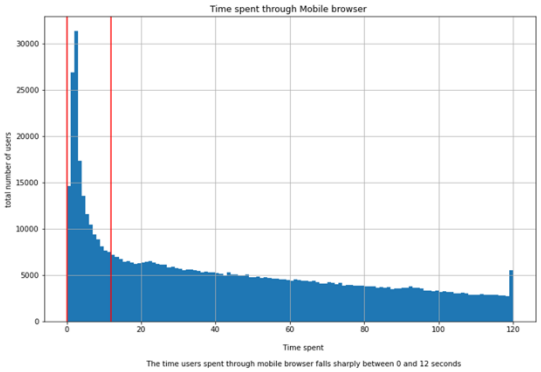
Phase 2:

1. ML-OPS Part I:
    a. Create Lambda Python script to use as trigger results into Snowflake.
    b. Upload flat file of most up to date Python modules into s3 bucket in order for lambda function to successfully run.
    c. Create API Gateway to trigger lambda function to return results to external function.
2. ML OPS Part II:
    a. Create IAM role and policy to ensure secure snowflake external function access
    b. Write Snowflake API integration script which is used to request access to API gateway of AWS.
    c. Create Trust Relationship and Resource Policy as another security check.
3. ML OPS Part III:
    a. Create EC2 instance in correct subnet and select correct security group that follows engineering team security guidelines.
    b. Install Docker container on EC2 instance in order to install correct flat files from s3 bucket so Python modules will run successfully in Lambda function.
    c. Register EC2 instance with correct docker files to ECR in order for it to be used across AWS.
    d. Deploy Image and Test Lambda function.

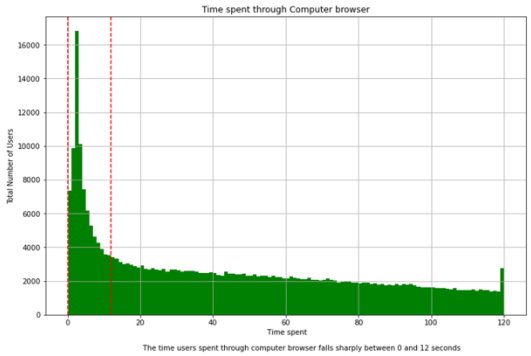
Project 4: Analyze users who bounce after one pageview.

Time spent on page by platform

Mobile Browser



Desktop



Between 0 and 2 minutes

	article	homepage
mobile-browser	64%	46%
computer-browser	60%	56%

Phase 1:

1. Get all users who visit the site (have a pageview) but take no action (no click events, no second pageview.
2. Break them out by whether they landed on an article, or they landed on the homepage.
3. Using the "component on" event, figure out when the last timestamp ran for the user.
4. Determine if these users are consuming content (the article on an article page, the headlines on the homepage) based on how much time they're on site.

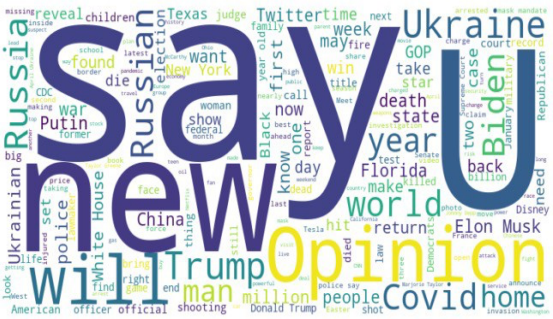
Phase 2:

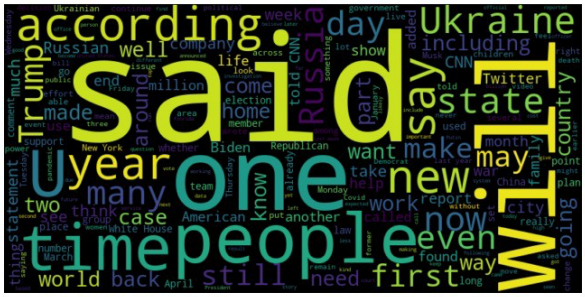
1. Determine threshold in seconds that defines a user who has one pageview on homepage or article.
2. Identify the most popular article urls visited for users who viewed them and then immediately bounced.
3. Compare and contrast time spent on mobile browser and desktop for users who had one pageview and met threshold criteria.
4. Discover the total number of users who bounced after one pageview, met the time in seconds threshold on both mobile browser and desktop.
5. Conclude what percentage of users who meet threshold in seconds are on article in comparison to homepage for both platforms.

Headline & Average time spent

Top 10 most popular articles	Headline Character Count
One of five Texas family members charged in US Capitol riot pleads guilty.	73
Kentucky legislature overrides governor's veto of sweeping abortion bill.	72
'90 Day Fiancé's' Anna-Marie and Mursel Mistanoglu welcome baby born via surrogate in Ukraine.	93
Hailey Beiber says she underwent procedure to close hole in heart after mini-stroke	87
Pete Davison to star in new comedy series 'Bupkis'	51
Video show moments a driver lost control of a tractor-trailer that narrowly avoided hitting an Ohio School bus	111
Dolly Parton now says she'd accept a spot in Rock & Role Hall of fame if she's chosen	89
Villanova head coach Jay Wright Announces retirement	56
Gilbert Godfriend comedian and actor has died	47
'Avatar 2' debuts footage from movie at CinemaCon and gets official title	73

CNN





Phase 1:

1. Predict average read time of users consuming content across CNN Digital articles within given time frame.
2. Used NLP and Regression Analysis to assess Headline quality by character count, and popularly for given article.

Insights:

1. Article Headline character count reveals -3.6% correlation with average duration time
2. Article Body total word count exhibits 48% correlation with duration time
3. Most Popular Article duration time roughly: 3 minutes
4. Most Popular Top Headline character count: 72 words
5. For every one word increase in Headline total word count we expect a decrease in average duration time of -0.94 seconds.