

# Lightning Talk

Team sdmay20-43

Adrian Hamill  
Megan Miller  
Benjamin Carland  
Yi-Hsien Tan  
Diego Realpe

# High Level Overview

- Left Path
  - Download Repositories
  - Create json files with Metadata
  - Convert into a Custom object used by boa researchers
- Right Path
  - Create Queries for repository data
- Create Visuals out of Query Result



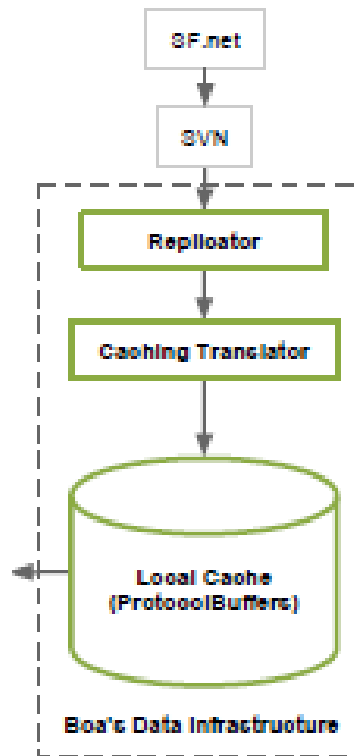
# Repository Data

- Analytics data
  - Traffic Data
  - Timeline of Pushes
  - Size of Pushes
  - Ect.
- Source-code data
  - For code quality analysis



# Custom Object Conversion

- The BOA analysis can't happen on raw data.
- Based on research from ISU's BOA labs, code needs to pass a conversion.
- Standard object to run queries against
  - Object is in Hadoop Sequencefile Format
  - Contain messages with relevant data
  - Can add custom modifications to the object for more measurements



# Query Creation

- Boa queries are used to answer wide variety of GitLab repository mining questions.
- Requirement:
  - Create suitable queries for repositories of COMS/SE309 projects.
  - Creative.
  - Evaluate code qualities.

What are the ten most used programming languages?

```
1 # Counting the 10 most used programming languages
2 p: Project = input;
3 counts: output top(10) of string weight int;
4
5 foreach (i: int; def(p.programming_languages[i]))
6   counts << p.programming_languages[i] weight 1;
```

[Run Example](#) | [Published Results](#)

How many projects use more than one programming language?

```
1 # Counting the number of projects written in more than one languages
2 p: Project = input;
3 counts: output sum of int;
4
5 if (len(p.programming_languages) > 1)
6   counts << 1;
```

[Run Example](#) | [Published Results](#)

# Displaying Data

- Need to analyze data after it's gathered
- R Language
  - Statistical computing and graphics
  - Linear & nonlinear modeling
- R Environment
  - Primarily written in C, Fortran, and R
  - Free GNU package

