

**MICHIGAN STATE**  

---

**UNIVERSITY**

# Project Plan

## Amazon Data Hub

### The Capstone Experience

Team Amazon

Josh Barnett

Austin Cozzo

Dan Farat

Cameron Nejman

Robert Ramirez

Department of Computer Science and Engineering  
Michigan State University

Spring 2020



*From Students...  
...to Professionals*

# Functional Specifications

- Currently, Data Scientists waste a lot of time doing research on finding the “right” dataset
  - Datasets are often vague, old, narrow, too narrow, or too large
- Amazon Data Hub (ADH) will be used to assist in the process of finding useful datasets
  - Will be achieved through the catalog of datasets, the extraction of metadata, and the generation of keywords

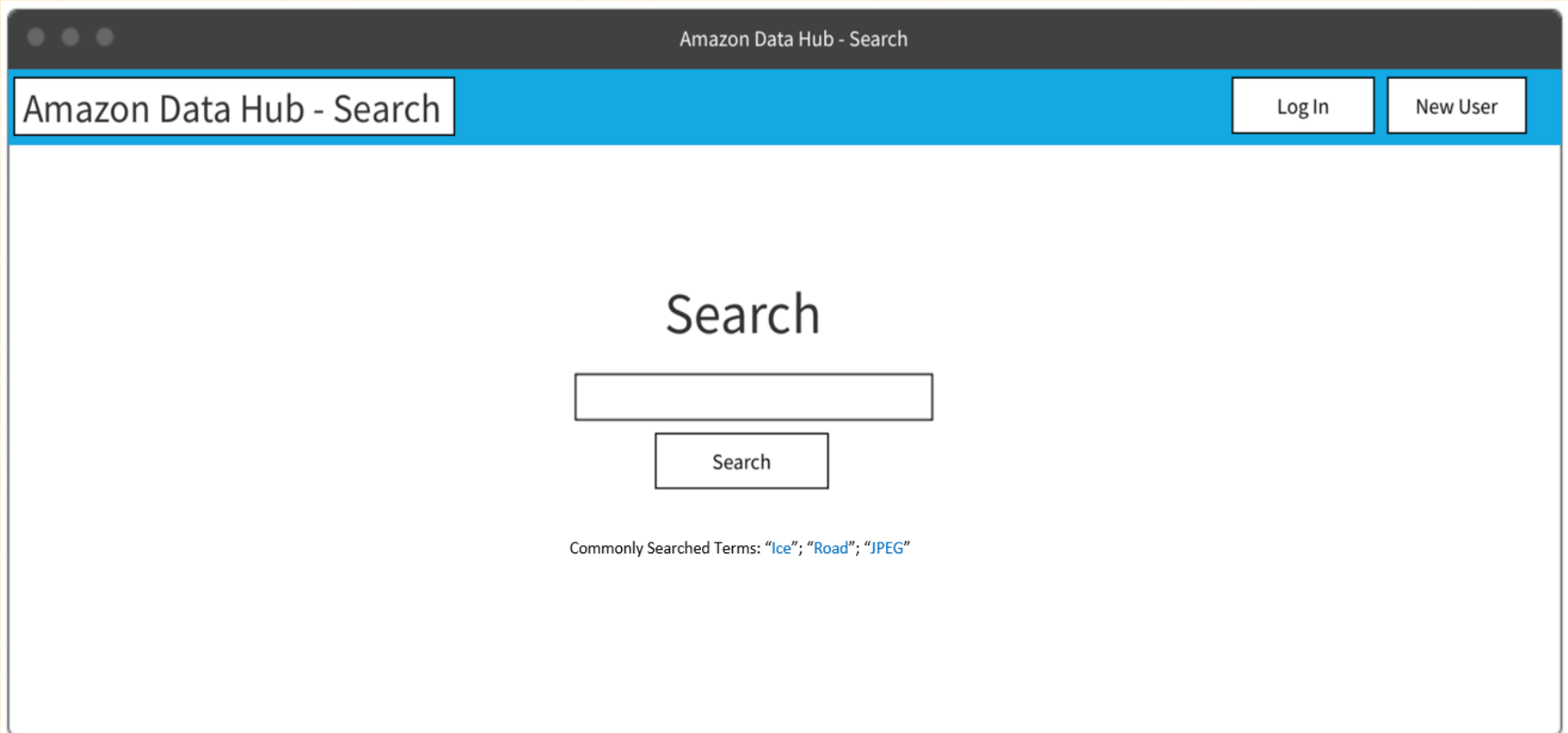


# Design Specifications

- When a user uploads a dataset to the ADH, it will begin the processing operations
  - launches metadata extraction, storage and keyword generation processes
- The ADH will also allow users to search for datasets related to user and system generated keywords
- Related datasets can also be 'linked' together
  - Users can navigate through related datasets using these links



# Screen Mockup: Search Page



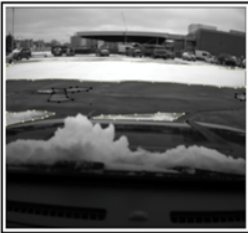
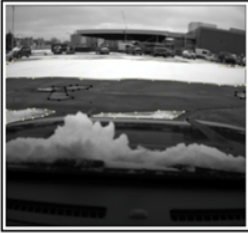
# Screen Mockup: Search Results

Amazon Data Hub - Results

Amazon Data Hub - Results Log Out Profile

Results

Results: 2

<p>Title: Annotated Road Ice Uploader: demo_user_1 Tags: .png (Filetype), .json (Filetype), Binary (class number), Annotated (Misc), "Annotated Road Ice V2" (Link), Car (Implicit Object), 1.0 (Version)</p>	<p>Creation: Jan. 1, 2020</p>	
<p>Title: Annotated Road Ice V2 Uploader: demo_user_1 Tags: .png (Filetype), .json (Filetype), Binary (class number), Annotated (Misc), "Annotated Road Ice" (Link), Car (Implicit Object), 2.0 (Version)</p>	<p>Creation: Jan. 20, 2020</p>	

# Screen Mockup: Dataset Home

Amazon Data Hub - Dataset

Amazon Data Hub - Dataset Log Out Profile

## Annotated Road Ice

Creation: Jan. 1, 2020

Uploader: demo\_user\_1


Tags:

- .png (Filetype)
- .json (Filetype)
- Binary (class number)
- Annotated (Misc)
- Car (Implicit Object)
- 1.0 (Version)
- Ice (Explicit Object)

Links:

[Annotated Road Ice V2](#)

Download





# Screen Mockup: Upload Screen

Amazon Data Hub - Upload

Log Out Profile

Choose File...

Title:

Uploader: demo\_user\_1

Tag:

Tag Type:

Link to Dataset:

All tags:

Tag	Tag Type
Annotated	Misc
Ice	Explicit Object
2.0	Version

All dataset links:

Links
Annotated Road Ice

# Technical Specifications

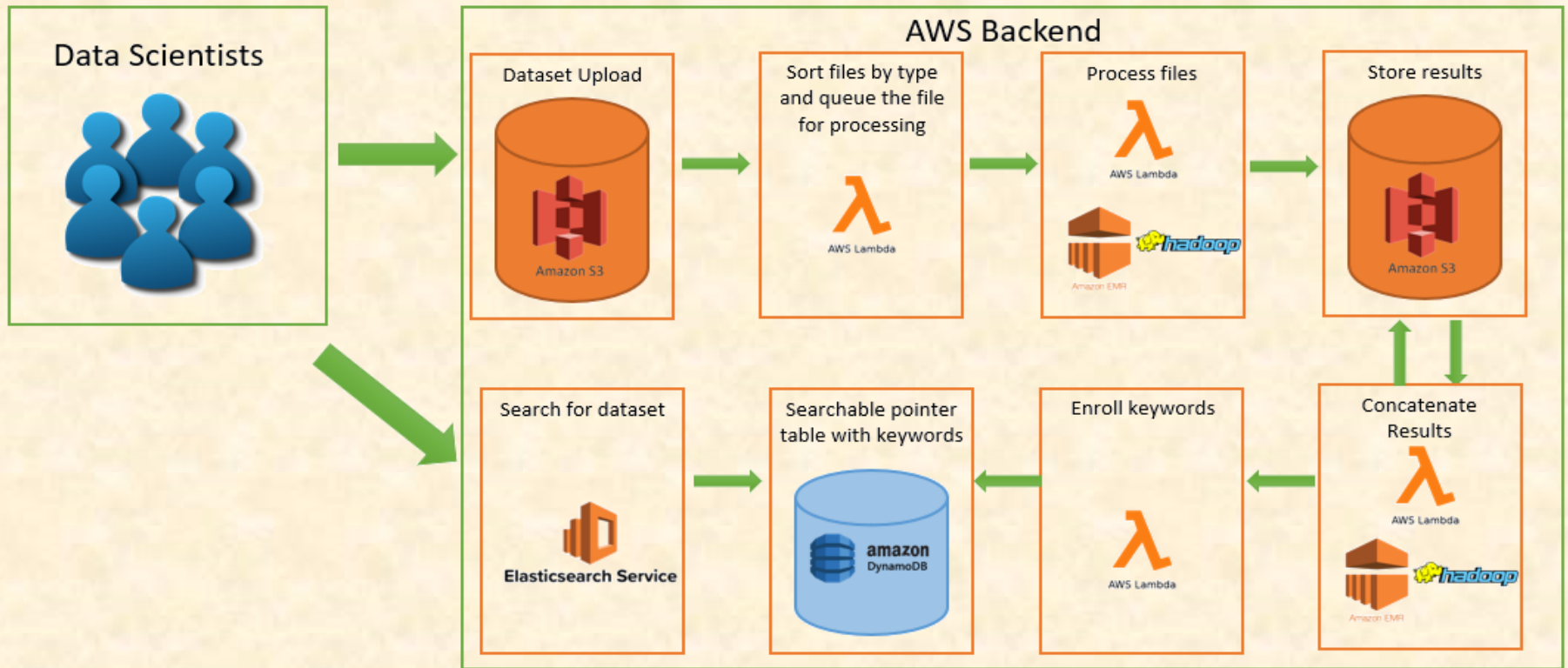
---

- Storage of datasets and results
  - AWS S3, DynamoDB
- Front-end design and functionality
  - Flask and React
  - AWS ElasticBeanstalk
- Back-end data processing
  - AWS Elastic MapReduce, Lambda, Step functions, Rekognition, Transcribe, ElasticSearch





# System Architecture



# System Components

- Application Front/Backend
  - AWS Elastic Beanstalk
  - Flask
  - React
- AWS Processing Backend
  - AWS: S3 Buckets, DynamoDB, Rekognition, Transcribe, Elastic Search
  - AWS Lambda, Elastic MapReduce (multifaceted)
  - File Extractor
  - Text File Processing



# Risks

- Dataset Size and Scalability
  - Intended use of ADH is for datasets of all sizes
  - Processing will slow down considerably with larger datasets
  - Schedule an EMR cluster to be periodically launched
- Dataset Variability
  - ADH must be able to accept datasets of most common types
  - Processing functions will be developed for as many file types as possible
- Cost Vs. Efficiency
  - Utilizing AWS distributed services is necessary, but will quickly accumulate charges
  - Working closely with our client we will be able to find the best middle ground for Amazon's internal needs



# Questions?

---

?

?

?

?

?

?

?

?

?

