

Deploying Machine Learning Application on Edge Devices, a Study of Challenges and Solutions

Marcus Bedeau

Mentor: *Luis Robaina, MITRE*

Instructor/Faculty: *Seyedmasoud Sadjadi, PhD, Florida International University*

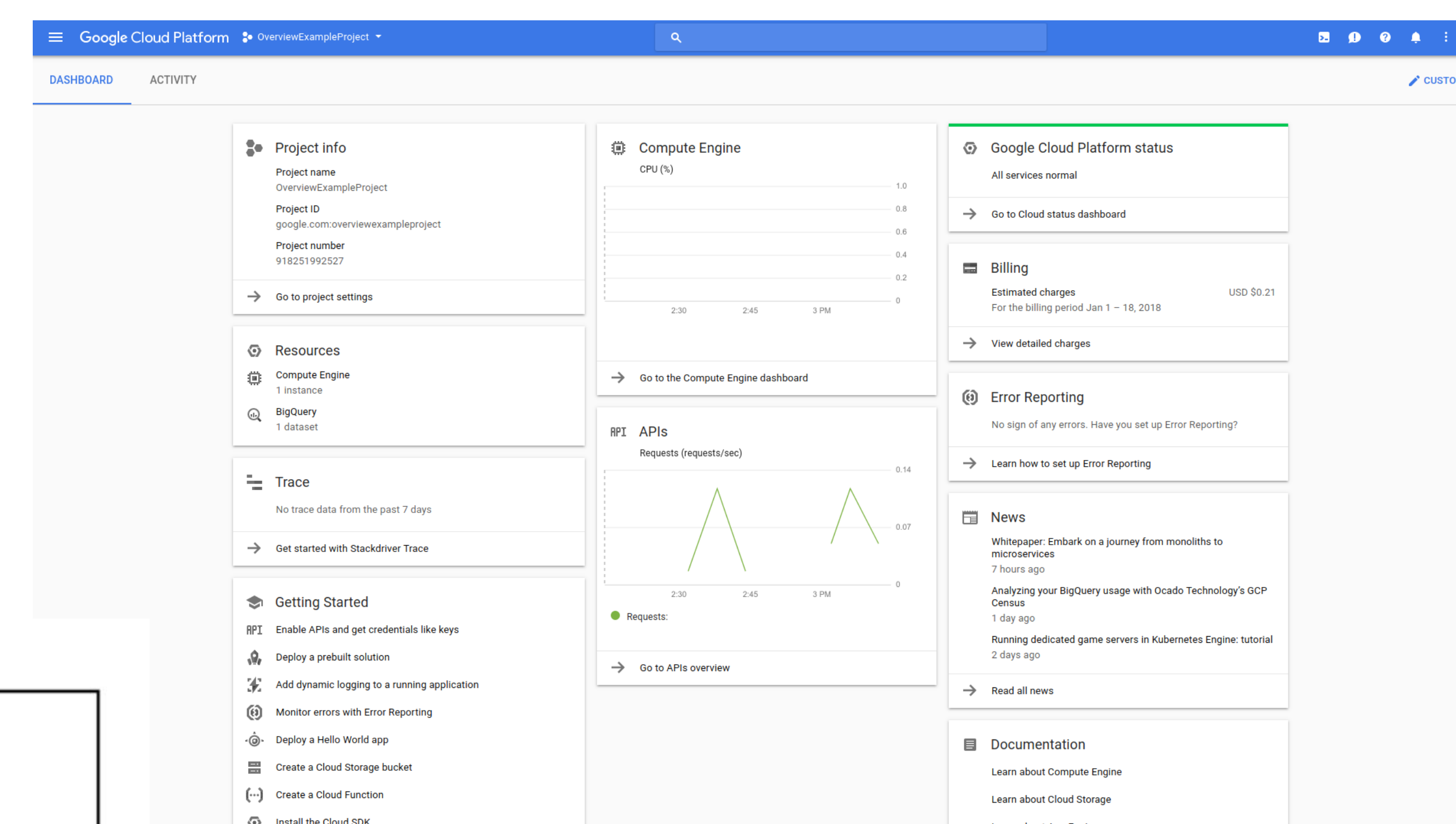
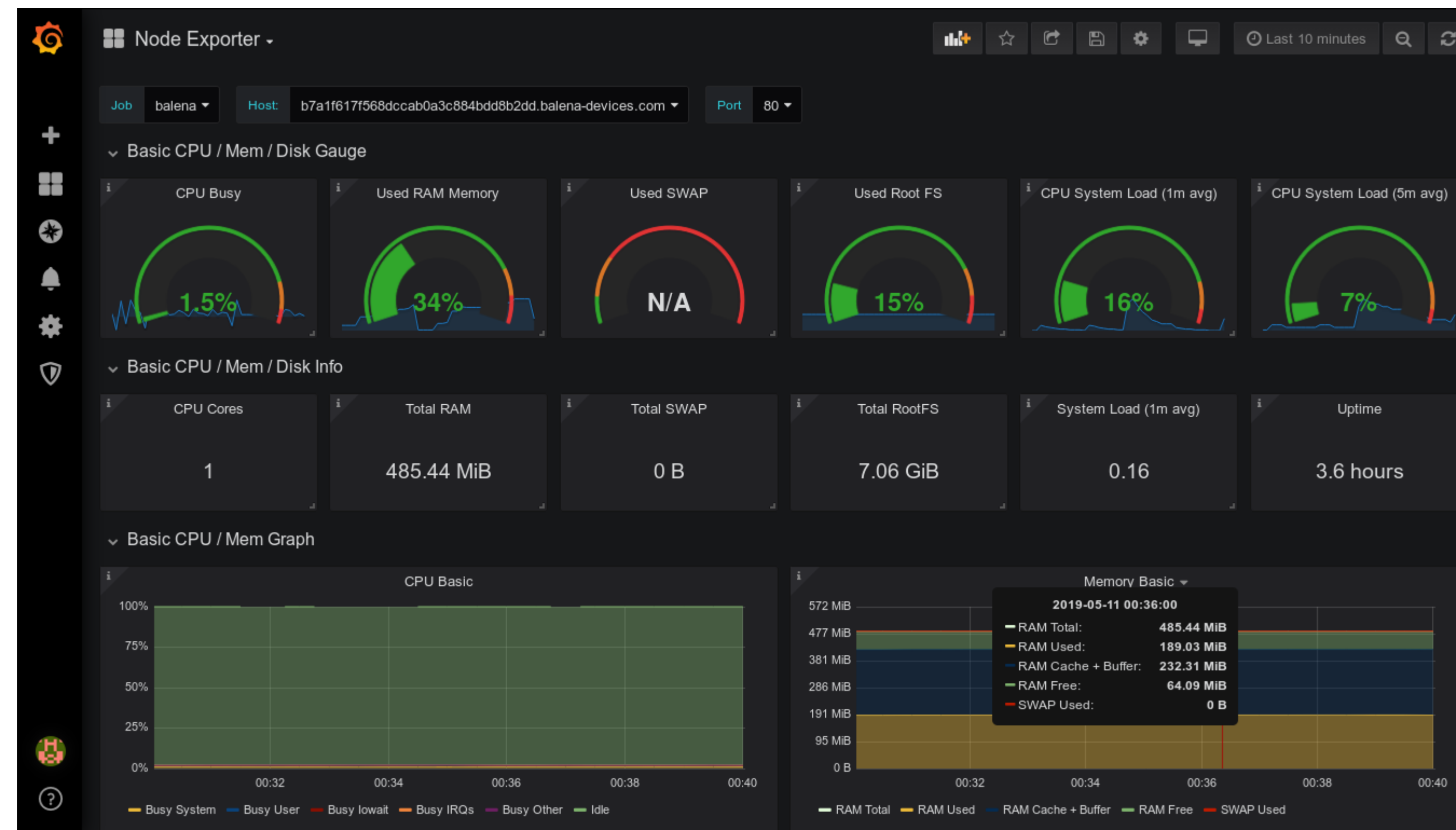
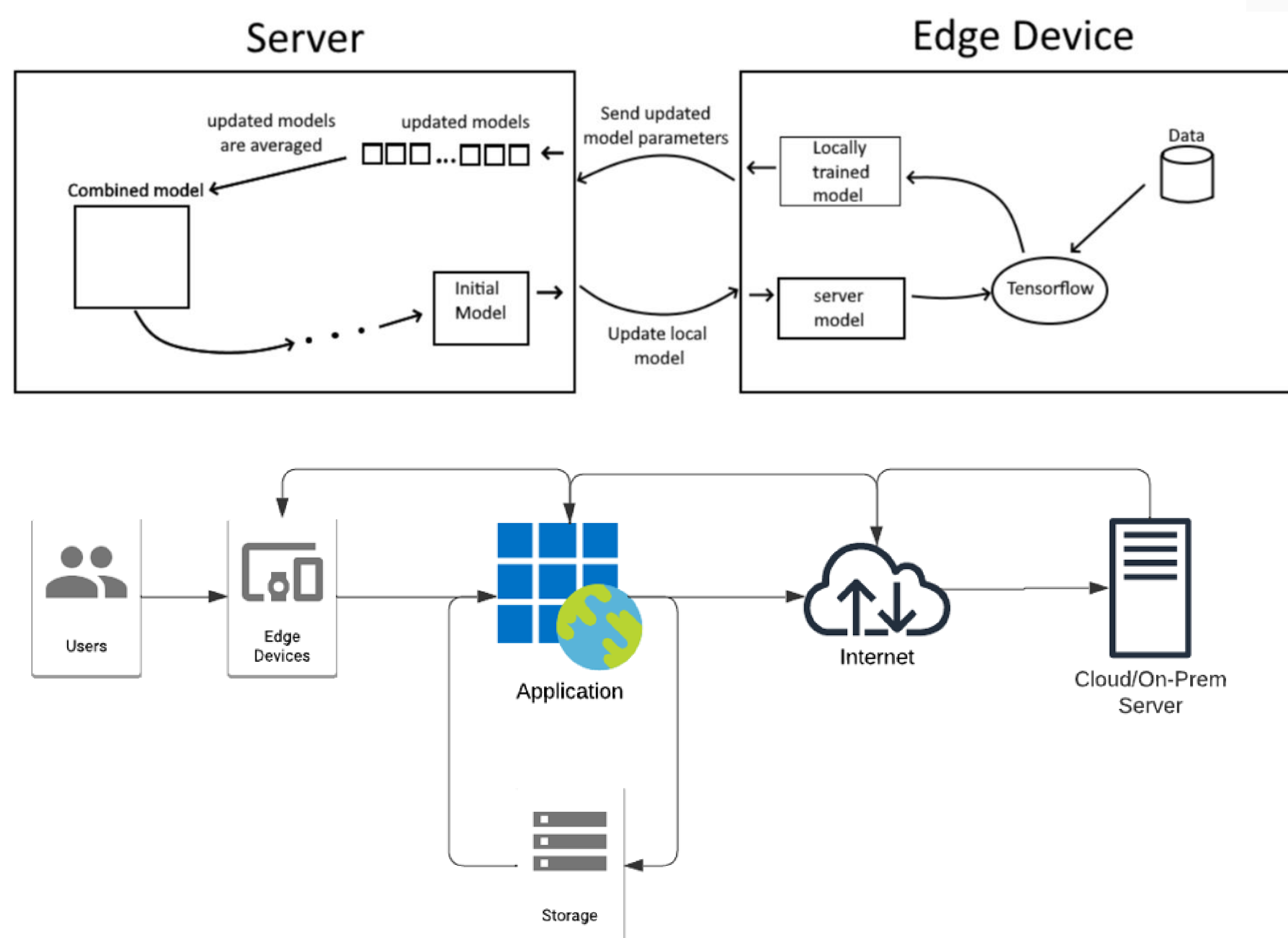
PROBLEM

As the community of AI/ML practitioners finds exciting new applications of these techniques we are faced with the challenge of building software systems whose behavior is dictated by data and code. Pioneers on these efforts argue that operationalizing ML systems requires its own engineering discipline (MLOps), and techniques previously applied to build traditional software systems do not cover the complexity of this new type of software system built on data. One exciting real-life application of AI/ML is deploying artificial intelligence on edge devices. This project aims to understand the following aspects of bringing intelligence to edge devices

REQUIREMENTS

The requirements we needed to achieve where to create a research paper on the deployment and monitoring aspect of the application. We had to design and research the best practices on deploying a machine learning model onto different edge devices. So, finding the best tools and resources and researching the ways they are currently being implemented within previous projects. We also had to design the monitoring aspect so that if there were issues it could be relayed back in real time to the team to start working on a solution. So having these goals in mind we had to design a plan to have a successful deployment.

SYSTEM DESIGN



4 Deployment

Deployment overview

When developing an application there are certain steps that must be taken to ensure the successful rollout of a new application. When developing there are a few great best practices that are used throughout the industry. It starts with knowing the requirements, building a design, development, testing, deployment, and maintenance. This is the general cycle the industry uses for new applications. The first stage is getting the requirements needed from whoever is making the request. This includes the usage of the apps if there are any restrictions and how they need it to be used. Once all the requirements are accepted the team would create a design and give it back to the requester and if the design is acceptable they then move onto the deployment stage. This is where all the things in the design step get built all while keeping communication to the requester. Once development is complete and approved the testing phase begins. Making sure the application works as designed and making sure there are no security issues that were missed in development. After these stages are complete then comes the deployment phase. This stage is key because it is where the application is being.

Deployment types

There are a few deployment types that are regularly used throughout the industry. Having it hosted on the cloud, on premises, and a hybrid approach. The first being hosted on the cloud would be using servers that are provided by a third party company and use a pay as you go approach. When having it deployed in the cloud you have all the resources and benefits the cloud has to offer. Such as nearly endless compute power and storage as well as the elasticity of having all the resources on demand. Cloud Deployment A fully cloud deployment entails the application being fully hosted on a cloud provider. The benefits that come with being on the cloud are the elasticity and the lower upfront cost to get the application deployed. The cloud has many tools and resources that assist in reaching edge devices and are scalable to deal with the increased activity. Each cloud platform has its own tools as well as third party softwares that can integrate with applications. Since we are planning on deploying a machine learning model the compute power of the cloud can decrease the time it would take to get results. The cons of a fully cloud approach is that in some of the use cases of the application having access to the internet is not possible and would render it useless as of now.

On Premise Deployment With an on Premise Deployment the upfront cost of the servers to deploy the application would be higher. The On premise approach would be good for government requirements where certain information can not be stored on a shared server. This approach has been used the longest but there has been a shift to cloud infrastructure. A lot of the same issues arise here if the edge device can not reach the internet the application would be useless.

CURRENT SYSTEM

Currently there are no systems in place that we are trying to build. Having machine models run on the amount of edge devices we are trying to achieve has not been done before. The infrastructure of hosting a model in the cloud and relaying it to edge devices has been done and have many examples but there has only been a small subset of teams that worked on utilizing the resources on edge devices to make this feasible. By combining all of these requirements we will be one of the first to have a deployment such as this.

IMPLEMENTATION

We researched many methods on seeing what type of infrastructure is needed to deploy this application. We needed to research all the different options with deploying it to the cloud as well as to an on-premises environment. We also had to research different monitoring tools such as Prometheus to allow the team to have a way to gain insight to the model they deployed. So, the current projection we recommend is having the application be deployed directly onto the edge device and still function without internet accesses. Storing the data within the edge device and once the device can communicate with the internet that information be sent back to the monitoring software which can also push updates to the application.

SUMMARY

After researching we found that the best means of deployment would be having a hybrid approach with a mix of on-premise and cloud infrastructure. The application will be containerized and installed directly on the edge device and have means of working with and without the internet. There will also be a Prometheus supported monitoring dashboard that will allow us to oversee the results of the model. As well as having a way to push updates over the air to the edge devices so they are up to date. There are many things still to be tested such as having a test device that would be able to demo this implementation.