Tech Stack: Python scripting, Flask, MLflow, CircleCI(CI/CD), DVC, AWS

**\*\*Q1. Tell me about your current project.\*\***

The project is called Wafer Fault Detection. The duration of the project was 6 months.

It was designed to monitor and detect a fault in wafers while going through the manufacturing line so a manual check of every wafer won't be necessary and the detected wafers can be replaced immediately and effectively.

Multiple sensors are embedded in a single wafer. The project's task is to classify the wafers as Good or Bad based on the values of the sensor received.

**\*\*Q2. What was the size of the data?\*\***

The data had 590 features and each feature had 4100 values.

**\*\*Q3. What was the data type?\*\***

All data were of float data type as they were values from sensors.

**\*\*Q4. What was the team size and distribution?\*\***

team - 25 members :

3 - DS

2 - interns

3 Functional DS

2 UI/UX

3 analytics side

5 python developer

3 leads

2 OPs

2 admin

\*\*Q5. What Hadoop distribution were you using?\*\*

\*\*Q6. What is the version of distribution?\*\*

\*\*Q7. What was the size of the cluster?\*\*

\*\*Q8. How many nodes were there in all the Dev, UAT, and Prod

environments?\*\*

**\*\*Q9. How were you creating and maintaining the logs?\*\***

The logs are maintained using MongoDB. The logging starts with the start of the application. The

start time of the application gets logged. After that, there are loggings for entry and exits to the

individual methods. There are loggings for the error scenarios and exception blocks as well.

\*\*Q10. What techniques were you using for data pre-processing for\*\*

\*\*various data science use cases and visualization?\*\*

\*\*Q11. How were you maintaining the failure cases?\*\*

**\*\*Q12. What kind of automation have you done for data processing?\*\***

We had a full-fledged ETL pipeline in place for data extraction and pre-processing.

\*\*Q13. Have you used any scheduler?\*\*

**\*\*Q14. How are you monitoring your job?\*\***

There are logging set-ups done. We regularly monitor the logs to see for any error scenarios.

**\*\*Q15. What were your roles and responsibilities in the project?\*\***

My responsibilities consisted of gathering the dataset, labelling the dataset for the model training,

training the model on the prepared dataset, deploying the trained model to the cloud, monitoring the

deployed model for any issues and providing QA support before deployment.

**\*\*Q16. What was your day to day task?\*\***

My day to day tasks involved completing the JIRA tasks assigned to me, attending the scrum

meetings, participating in design discussions and requirement gathering, doing the requirement

analysis, data validation, Unit test for the models, etc.

**\*\*Q17. In which area you have contributed the most?\*\***

I contributed the most to data validation, pre-processing and model training areas. Also, we did a lot of brainstorming for finding and selecting the best algorithms for our use cases. After that, we identified and finalized the best practices for implementation, scalable deployment of the model, and best practices for seamless deployments as well.

**\*\*Q18. In which technology you are most comfortable?\*\***

I'm comfortable working with ML technology currently. But I'm also eagerly learning about deep learning and enjoying that too!

**\*\*Q19. How do you rate yourself in big data technology?\*\***

I have worked often in big data computing technology with ample knowledge in distributed and

cluster-based computing. But my focus and extensive contribution have been as a data scientist.

**\*\*Q20. In how many projects you have already worked on?\*\***

I have worked on 2 projects during my internship. My major project during my final year of Engineering was also ML-based for which I have also published a paper. Apart from these, quite a few minor projects. I'm also currently creating a website for the NGO that I work for using Django.

**\*\*Q21. How were you doing deployment?\*\***

The deployment was done on AWS EC2 instance and CircleCI was used for CI/CD pipeline.

**\*\*Q22. What kind of challenges have you faced during the project?\*\***

The biggest challenge that we face is in terms of obtaining a good dataset, cleaning it to be fit for

feeding it to a model, and then labelling the prepared datasets. Labelling is a rigorous task and it burns a lot of hours. Then comes the task of finding the correct algorithm to be used for that business case. Then that model is optimized.

**\*\*Q23.What will be your expectations?\*\***

It’s said that the best learning is what we learn on the job with experience. I expect to work on new

projects which require a broad set of skills so that I can hone my existing skills and learn new things

simultaneously.

**\*\*Q24. What is your future objective?\*\***

The field of data science is continuously changing. Almost daily, there is a research paper that

changes the way we approach an AI problem. So, it makes it exciting to work on things that

are new to the entire world. My objective is to learn new things as fast as possible and try and

implement that knowledge to the work that we do for better code, robust application and in turn, a

better user/customer experience. I am also interested in improving my soft skills and becoming better at making the non-technical stakeholders understand the information for business decisions with ease.

\*\*Q25. Why are you leaving your current organization?\*\*

\*\*Q26. How did you do Data validation?\*\*

\*\*Q27. How did you do Data enrichment?\*\*

**\*\*Q28. How would you rate yourself in machine learning?\*\***

Well, honestly, my 10 and your 10 will be a lot different as we have different kinds of experiences.

On my scale of 1 to 10, I’ll rate myself as an 8.2.

\*\*Q29. How would you rate yourself in distributed computation?\*\*

**\*\*Q30. What are the areas of machine learning algorithms that you**

**already have explored?\*\***

I have explored various machine learning algorithms like Linear Regression, Logistic Regression, L1 and L2 Regression, Polynomial Regression, Multi Linear Regression, Decision Trees, Random

Forests, PCA, XG Boost, CAT Boost, ADA Boost, Gradient Boosting, K-Means, K-Means ++, KNN, SVM, SVR, Naïve Bayes, DBScan, Hierarchical clustering, TFIDF, Word to Vec, Bag of words, Doc to Vec, Kernel Density Estimation are some of them.

**\*\*Q31. In which part of machine learning have you already worked**

**on?\*\***

I have worked on both supervised and unsupervised machine learning approaches and building

different models using the as per the user requirement.

**\*\*Q32. How did you optimize your solution?\*\***

Well, model optimization depends on a lot of factors.

 Train with better data(increase the quality), or do data pre-processing steps more

efficiently.

 Increase the quantity of data used for training.

 Increase the number of epochs for which the model was trained

 Tweak the hyperparameter range for tuning the models

 Try and use multithreaded approaches, if possible.

 Reducing Learning rates in plateau regions optimizes the model even further.

\*\*Q33. How much time did your model take to get trained?\*\*

\*\*Q34. At what frequency are you retraining and updating your

model?\*\*

**\*\*Q35. In which mode have you deployed your model?\*\***

I have deployed the model both in cloud environments.

**\*\*Q36. What is your area of specialization in machine learning?\*\***

I have worked on various algorithms. So, It’s difficult to point out one strong area. Let’s have a

discussion on any specific requirement that you have, and then we can take it further from there.