My name is Sivasankar Pendlimarri, I am having overall 6.2 years of IT experience out of that in 3 + years in big data technologies. I started my career as an SQL developer then I migrated to big data, from a big data point of view I used tools like Hadoop, Sqoop, Hive, Spark, Pyspark, AWS S3, Ec2, and EMR for multiple domains like healthcare, Retail and banking domain projects.

Coming to my last organization, Sopra Steria India Pvt Ltd. I was working on the OneBaufi project, It’s a banking domain project. In this project, I am getting data from Oracle DB. I am using Sqoop to import data from Oracle DB to HDFS, Once the import is done, we are defining it as a raw zone. Based on the raw zone a table has 1000s of columns which is used to process the data.

For that, we are defining the refined zone by using some potential columns like First name, Last name, DOB, address, Card Number, CVV, Transaction type, Amount and Gender like that. Before going to the refined zone, we need to standardization of data also which means different clients are given the data with the same meaning with different orientations, like gender is a column, some clients are mentioning gender as 'GE’, some clients mention gender as 'GEN', some clients are mentioned gender as 'Gen' by using a regular expression. Then refine zone is defined then we are going to MDM (Metadata Management) layer. In this zone, we try to assign each customer with a hash (#) value, the hash value is nothing but a unique key value for that particular customer. So based on the unique key value we try to identify the customer. If he/she has the old customer then we append the data to existing hash value data, if he/she is a new customer, we try to create a new hash value for that.

Then the data is moving to the final zone as a production zone, In the production zone, we are scheduling the jobs using DAGs in Airflow as the scheduler. If there is any issue, we will try to connect with the source team to resolve the issues (Source file not found, Data mismatch). Finally, the data is moving to Down stream team (Data Science) to analyze and predict the data.