Good Morning

First of all, thank you so much for giving me this opportunity to introduce myself.

My name is Samadhan Mali. Basically, I belong to Nashik, Maharashtra.

I completed my Bachelor of Engineering from Savitribai Phule Pune University in the year 2020 with a CGPA of 8.21.

Concern to my professional experience,

Currently I am working with DGCT INFOSOFT, Nagpur as Data engineer from last more than 2 years.

Actually, I have 2.8 years of experience in data engineering. I worked on both domains, i.e., data injection and data processing, along with the AWS cloud.

Till now, I had completed two projects in the e-commerce and retail domains, where I mostly used Pyspark APIs along with Amazon web services.

My technical competencies are Hadoop ecosystem in that hdfs, mr, sqoop, hive, yarn, hbase, my sql, oracle for programming language python for processing Spark. Along with this I have experience in AWS in that I used Ec2, EMR, S3, RDS aLso I am sounds with Athena, Redshift, Lambda.

Also I used Git, She’ll scripting Airflow as well.

I would like to tell about my project. My project name is OMS datalake implementation and analysis and this is e-commerce domain, In that my role was bigdata developer and responsibility was to understand client requirements, develop logic on that after do unit testing after that it goes to UAT team for testing purpose, once approved from UAT team, we need to put it on production cluster, then we need to check it working fine or not, if not working fine need any changes in code or optimization then need to fix it and deployed it again on production cluster.

After that we need extract reports, like validation reports. That’s all about me

I would like to tell about my project flow, Data available on source RDS using spark, we Imported data and need to store into hbase.

After that as per clients requirements did some transformations and store that data in hive staging. So in hive there are two types of table used, internal in Staging and external in Hive Final. In hive staging their have some duplicates so we did deductions and store data in hive final. After that extract the reports and stored in AWS S3

Thank u so much sir/Madam.

When it comes to project flow Initially client made as all data his historical data available in the postgre so from the postgre we dump the data into hbase so we reading purpose used pyspark API using this pyspark dump the historical data into hbase

Basically historical data broadly classified into two types the first one was the reference data and 2nd one was the transactional data so after having the data in hbase we can perform some kind of transformation as per totally basis on the client requirements so broadly speaking we perform left joins on reference table and transactional table .... using the pyspark we perform the left join on pyspark dataframe and after that we dump that transforms data into hive internal table

Why we use hbase over here ....so has to have that flexibility or liberty of using the dynamic schema possibility and also hbase supports a versioning so these the main two features of hbase and 3rd was backup

After having the data in hive internal table , so in hive we have created two types of one is enternal table and 2nd one is external table, in internal table have some duplicate data so we're perform some kind of duplication or we can say compaction so has to eliminate the Bulls and all duplicate kind of data from our database , after removing duplicate data then we moving this data into external tables

Why we use external table over here is if by any accidental mistake by developer or server mistake if data get lost or if any table gets drop then even in that case we have backup in HDFS. After having the data into external table then we perform our main login which we have to find outstanding limit of credit card and outstanding limit of amount of credit card of particular customer

Finally after calculating the limit ...we dump the data into S3 for report generation so this all about the historical data we recieved on the initial basis,then we comes to delta data we has to come on the daily basis like roughly it was around 20gb and it's comes in the form of CSV file in HDFS in certain paths so we perform same logic for delta data as like historical data. This was the main flow of ccda project.