# Anusha Jallipalligoanusha2023@gmail.com

 8624020970

# Summary

* I’maseniorbigdataprofessionalwith10yearsofexperienceindataengineeringonmultiplepubliccloudplatforms(GCP,AWS&Azure).
* IhaveextensivebackgroundindistributedstorageandprocessingframeworkslikeHadoop&Spark (Cloudera, Hortonworks, Databricks) and have delivered numerous big data and datalakemodernizationsolutionsforcustomersacrosstheglobe.
* I’mhighlyproficientinprogrammingusingadvancedjava&pythonandhaveexperiencebuilding reliable APIs and web services using high quality coding and testingmethodologies.
* I’mequallyproficientinwritinghighqualitySQLinteractingwithseveraldatabases(likePostgres, MySQL, SQL Server, Oracle etc) and modern data warehouse systems (likeHive/SparkSQL,BigQuery,Redshift,etc).
* I have designed and implemented several batch, streaming and event driven datapipelinesorchestratedoncloudreadingfromavarietyofsourcesystems,transforming(ELT & ETL patterns) and landing in high value targets like BigQuery, CloudSQL, AzureADLS,etc.
* I have extensive experience in designing, implementing and deploying data applicationsacrossingestion,ELT/ETLandconsumptionlayersusingmodernclouddataengineeringservices.
* Ihavestrongexpertiseinautomation,workfloworchestrationanddeploymentofdatapipelinesusingagileSDLCandDevOpsprinciples.
* WorkedonendtoendmigrationofrealtimeingestionframeworkfromkafkatoBigQueryonGCPPlatform.
* Azure/ADLSandAzureDataFactorypipelinesanddataengineering/preprocessingforMLmodels.
* ExperienceinGCPdataflowpipelines.
* ExperienceinAWSserverlessandAPIendpointdevelopment.
* InvolvedindevelopingRestfulWebServiceswithpythonFlask,java(jersey).
* DevelopedAsyncrestserviceusingOozieandjava(jersey).
* InvolvedincreatingJunittestcasesforcodecoverage.
* DevelopthetestautomationscriptstovalidatethebuildsintheCI/CDpipeline.
* Experiencedindevelopingend-to-endIOT(InternetofThings)applicationsusingSpark.
* EfficientinconfiguringMapReduce,Hive,Pig,SqoopandshellorchestrationswithOozie.
* ExperiencedonHadoopdistributionslikeCloudera.
* HandsonworkingknowledgeonNoSQLdatabases.

# Education:

* BachelorofTechnology,AndhraUniversity(SIRCRREDDYCOLLEGEOFENGINEERING,)–2012

# Skills:

* **BigDataTools:**BigDataHadoopEcosystem,ApacheSpark,MapReduce,Pyspark,Hive,Kafka,Flume,Oozie,Zookeeper,Sqoop,HBase.
* **Languages:**Java,Python(Flask,CherryPy,FastAPI,SQLAlchemy),ShellScripting
* **CloudTools:**GCP(BigQuery,Dataflow,Kafka,CloudComposer,Airflow),AWS(S3,Lambda,APIGateway,Cloudwatch,cloudlogs)andAzure(ADFPipelines,ADLS,Databricks)
* **Frameworks:**SpringbootwithHibernate,JerseyRestAPI’s
* **VersionControls:**GitandBitbucket
* **DatabasesandTools:**PostgreSQL,MySQL,SQLServer,Oracle,Hive,Impala,HBase
* **TestingTools:**Junit,MockitoandPowerMockito
* **BuildandDeploy:**MavenandJenkins(CI/CD),SwaggerUI,Insomnia

# ProfessionalExperience:

**Client:Cigna Oct2023-TillDate**

# Role:SeniorDataEngineerResponsibilities:

* + Spearheadedateamofdataengineers,providingmentorship,guidance,andtechnicalleadershiptodrivesuccessfulprojectdeliveries.
  + Workedonreadingandwritingmultipledataformatsliketextfiles,JSONandParquetonHDFSusingPySpark.
  + Configuredreal-timedataingestionfromKafka,seamlesslyprocessedwithSparkStreaming,andstoredinHDFS.ActivelyoversawSparkjobprogressusingtheSparkApplicationMaster,ensuringeffectiveloggingandmonitoring.
  + Deployedmachinelearningmodelsforreal-timepredictionservingusingplatformslikeTensorFlowServing,Kubernetes,orcustomAPIendpoints.
  + ConfiguredDockernetworkingtofacilitatecommunicationbetweencontainersandservices,ensuringseamlessdataflowinreal-timeapplications.
  + Designedandimplementedscalabledataarchitectures,enablingtheprocessingandanalysisofmassivedatasetswithoptimalperformance.
  + Designedsolutionsforprocessinghigh-volumedatastreams,includingingestion,processing,anddataprovisioningusingHadoopecosystems.
  + ConductedperformancetuningandoptimizationofApacheBeamandGCPDataflowjobs,maximizingthroughputandminimizingprocessingtimesforlarge-scaledatasets.
  + ImplementeddataqualitycheckswithinApacheBeamandGCPDataflowpipelines,ensuringtheaccuracyandreliabilityofprocesseddata.
  + ImportthedatafromdifferentsourceslikeHDFS/HBaseintoSparkRDDandperformcomputationsusingPySparktogeneratetheoutputresponse.
  + LeveragedAzureservices,includingAzureDatabricks,forefficientdataingestionandprocessing.
  + Developedreal-timeanomalydetectionmodelstoidentifyandalertonunusualpatternsordeviationsinstreamingdata,helpingdetectissuesastheyoccur.
  + Designedandmaintainedtestdatasets,ensuringdataaccuracyandrelevance,anddevelopedprocessesfordataextraction,transformation,andloading(ETL)forUAT.
  + Successfully harnessed pivotal technologies such as Kafka, Spark Streaming, HDFS, AWSLambdas,andGoogleCloudPlatform,optimizingdataprocessingandmanagementacrossdiverselandscapes.
  + UtilizeDataBuildTool(dbt)tostreamlinedatatransformationandmodelingprocessesforenhanceddataanalysis.
  + ProficientlyoptimizedSQLqueriesanddataprocessingjobsinBigQuerytoimprovequeryperformance,reducecosts,andenhancedataretrievalspeed.
  + Collaboratewithcross-functionalteamsbyintegratingdbtwithversioncontrolsystemslikeGit,promotingeffectiveteamworkandversiontracking.
  + Successfullyledthemigrationoflegacydatasystemstomoderndataplatforms,achievingimprovedperformance,reliability,andeaseofmaintenance.

**Environment:**HadoopEcosystem(HDFS,Hive,Impala,HBase,Zookeeper,Sqoop).Spark,Spark-SQL,Spark-Streaming,Kafka,Java,Python,GCPDataflow,BigQuery

# Accenture,India June2021–Aug2022

**Sr.DataEngineer**

# GMI(GeneralMills)—ODSReal-timeingestion

Customerneededaningestionframeworktobedesignedandimplementedwithabilitytoperformreal-time data ingestion from the ODS layer, streaming into GCP BigQuery. In addition, dynamicschema evolution and change data capture (CDC) were extremely important features to supporttheirdownstreambusinesssystemswithleastdisruption.

# Responsibilities:

* As a senior data engineer lead on the project, I was responsible for leading the customerdiscussionsincludingdiscovery,assessment,designandimplementationoftheingestionframework.
* Developeddatamappingdocuments(DMDs)betweensourceSAPandtargetGCPBigQuerysystems.
* I was leading a team of 8 engineers and was instrumental in collaboratively developingtransactionDataFlowpipelinesreadingfromKafkatopicsandstreamingtoGCPBigQuery.
* IwasalsoresponsiblefordevelopingDataflowjobsforhandlingschemaevolutioninconjunctionwithBigQuerysupportedschemaevolutionfeatures.
* I was collaborating with customer teams to implement CDC operations using stored procsandhaddevelopederrorhandlinganddeadletterworkflowsusingDataflowandKafkaforproducingandconsumingefficientandreliableresultsinBigQuery.
* Developedcomposerworkflows(AirflowDAGs)forschedulingandlaunchingdataflowjobs.
* Worked with customer & Google SMEs and optimized several Dataflow job patternsinvolvingreadingfrommultipleKafkatopicsandwritingtorespectiveBigQuerytables.

**Environment:**Java,Python,Kafka,Dataflow,BigQuery,Airflow(Composer),Jenkins,BigQuerystored-procs

# Accenture,India Jan2021–May2021

**SeniorDataEngineer**

# GMI(GeneralMills)–IDF

CustomerneededDataIngestionFramework(IDF)thatpermitstherapidconsumptionofdatafromseveral different types of data sources by allowing the user to focus on metadata or the “what”instead of the “how” with respect to ingesting various types of data. By leveraging an approach todata management that minimizes manual code generation and maximizes component reuse, theframeworkneededthefollowing4features:

* + SingleframeworktoperformalldataingestionsconsistentlyintoGCPdatalake
  + Enablestrackingmetrics,eventsandnotificationsforalldataingestionactivities
  + Singleconsistentmethodtocapturealldataingestionalongwithtechnicalmetadata,dataquality,datalineageandgovernance
  + ProvideaproperdatagovernancewithSearch&CataloguetofinddatawithinDataLake

# Responsibilities:

* Asaseniordataengineerontheproject,Iwasleadingateamof4engineersinimplementationofIDFframeworkcomponents.
* Iwasresponsiblefordevelopingtheframeworkdatabasestructurebydesigningandimplementingdatamodels(PostgresqlonGCPCloudSQL)usingSQLAlchemy.
* Iwasinvolvedindevelopment,testing&productiondeploymentofcoreingestionandcommonserviceAPIsintheIDFframework.
* Idevelopedover50RestfulAPIsspanningacrossingestion,job/processtracking,lineage,dataqualitymodulesusingFastAPI(Python)andSwagger.
* TheAPIsweredeployedonGKEcontainersandIhelpedintegrateRediscacheandvaultfeatures.

**Environment:**Python,FastAPI,GCPCloudSQL,PostgresQL,SQLAlchemy,Redis,Swagger

# Accenture,India Apr2020-Dec

**2020**

# SeniorDataEngineer

**IntientAnalyticsStudio–Accenture**

AccentureneededtobuildGooglecloud-basedAnalyticsStudiomodulesforitspatientanalyticsplatform called INTIENT. This studio required developing several Rest API services supportingseveralfunctionsacrosstheINTIENTplatformincludingsourceintegration,dataingestion,datasearch&

preparation,featureengineering,modeldevelopment,training&calibrationandreportgenerationonGoogleCloudPlatform.

# Responsibilities:

* Asaseniordataengineer,Iwasresponsibleforcollaboratingwithateamof8engineersanddeveloped several Rest APIs for onboarding user personas & groups, provisioning relevantGCP services (AI-platform python/R notebooks, BigQuery, dataflow, dataproc, etc) anddeployingIAMrolemappingsasperIntientAnalyticStudiorequirements.
* Iwasinvolvedinend-to-enddevelopment,testingandproductionimplementationofAPIsonGKE.
* IhelpedtheteaminimplementingthebackendCloudSQL(Postgres)databasemodules,JavaSpring Boot applications and Okta integrations and performed CRUD operations on userprofiles,andhibernatemappingsforDBinteractions.
* Iwasalsoresponsibleforimplementingatest&qualityframeworkfortheAPIsusingSwagger,Jacoco,SonarQubeandJunit/Mockito.
* Ihaddevelopedseveralhighlevelsandlow-leveltechnicaldesignandimplementationcollateralstoguidethedeveloperteams.

**Environment:** Java with Spring boot and Hibernate, CloudSQL (PostgresQL), GKE, GCP Services(Dataproc,BigQuery,Dataflow,AIPlatformnotebooks,AutoML,Dataprep),JenkinsCI/CD,GitlabandConfluence

# Accenture,India Nov2019–Mar2020

**SeniorDataEngineer**

# UnileverLivewire2.0(country-SouthAfrica)

Unilever needed to implement a data ingestion framework from several internal and externalsystemsintoacentralAzuredatalakeforfurtherdownstreamapplicationconsumptionforbuildingdata driven insights and high impact models. This project involved creating a value chain to helpaddress the challenges of acquiring and ingesting data from multiple sources which providecompetitiveadvantagetoUnilever.

# Responsibilities:

* As a senior data engineer, I contributed by providing thought leadership to the team inbuildingandimplementingareliabledataingestionframework(LiveWire2.0)withreusableingestiontemplates.
* IwasresponsibleforingestingdatafromexternalfilesourcestoAzureDataLake(ADLS)usingAzureDataFactory(ADF)pipelines.
* Developedanddeployedseveralbatchpipelines,orchestratedandscheduledtheseasperSLA

guidelines.

* DevelopedPySparkmodulesinDatabricksfordataquality&validations,deduplicationandperformtransformationsontheingestedrawdatasetsinADLS.

**Environment:**AzureDataFactory(ADF),AzureDataLakeService(ADLS),DatabricksNotebooks,PySpark,SparkSQL

# Accenture,India Dec2018-Oct2019

**SeniorDataEngineer**

# RoyalCaribbeanCruiseLine–PriceRecommendation&ElasticityEngine

The goal of the project is to build a Price Recommendation Engine (PRE) Automation Platform toautomatemorefrequentandprecisepricechangesearlyinthebookingwindow(WeekstoSail>52weeks), where historically sailings have not been actively managed. In early experiments, thisapproach has reduced variance from Track and variance in book position across sailings. TheoverallPRE-Automationsolutionhasthefollowingtracks:

* + A price recommendation engine (i.e. the “PRE”) comprised of business logic coded inpythonthatconvertsvariousinputs–includingconfigurableparameters,Track,elasticitiesandhistoricalpricing&bookingdata–intoasuggestedpricechange
  + Datafoundationcomprisedofrequiredtablesandsupportingdatapipelines
  + Machine-learningbasedonpriceelasticitymodels
  + APIlayerstoconnectunderlyingtablesandrecommendationswithfutureweb-basedinterfacesoranotherUI(i.e.‘Autopilot’)

# Responsibilities:

Asaseniordataengineer,Iwascollaboratingwithateamof12engineerstodevelopthefollowingfeaturesonAzure&AWScloudplatforms.

* + PriceRecommendationEngine(PRE)–Idevelopeddailyandweeklydataingestion(historicalprice,booking,track&elasticitydatafromOracletoAzuredatalake),

pre-processingandfeatureengineeringpipelinesonAzureDatabricksusingPySparkandSparkSQL.

* + PRE-Data Foundation – I was responsible to build and automate data pipelines to Ingest(historical ship sailing and new booking records pricing, booking, track, FIT, Inventory,callforecastdatasetsfromOracletoADLS),Transform,executeElasticityandPRE-Models

pre-processingusingDatabricksandpersistthePRE-modeloutputbacktoOracledatabase.IhaddevelopedADFpipelines,orchestratedandscheduledtheseasperSLAguidelines.

* + PriceElasticityModels–Iwasresponsiblefordevelopingdatapre-processingpipelines(ship sailing & cross product price datasets) in Azure ADF to support regression andmixedeffectsupervisedmodels.
  + API Layer – I have built lightweight back end Services and API layer to retrieve and updateParameter tables data in Oracle EDSSP database. I have developed java REST APIs and AWSLambdafunctionswithAPIGateway.IhavealsocontributedinfunctionaltestingoftheAPIsanddeploymentofapplicationusingCloudFormationTemplates(CFT)

**Environment:**Java,AWS(S3,Lambda,APIGateway,CFT),Azure(ADF,Datapipeline,DatabricksNotebooks,ADLS),PySpark,SparkSQL

# Kogentix(AcquiredbyAccenture),India Jan2017-Nov2018BigDataEngineer

**DMLE(DIGITALMAPPINGLEARNINGENVIRONMENT)-NIELSEN:**

CustomerneededtobuildaDigitalMappingLearningEnvironment(DMLE)tosupportautomated

ML model training, testing and predictions (across categories like Matching, Linking, PredictiveModeling&Clustering)withsupportfromRESTAPIservicestoread/fetch/writedatatoandfromHDFS and Hive on Cloudera Hadoop. There was a need to build automation via several APIs tolaunch the model experimentations including training, test and model inference jobs and trackmodel runs in an asynchronous and consistent manner. Customers also needed data ingestionworkflowstobedevelopedfromOracletoHiveanduseElasticSearchandKibanaforindexingandvisualizationofproductdata.

# Responsibilities:

* As a lead Big Data Engineer, I was responsible to develop several REST APIs with PythonCherryPy/Flask&Java(JerseyAPI)toread/fetch/writedatafromHDFSandHiveandalsotolaunchandtrackMLjobs(training,testandbatchpredictions)asOozieworkflowsinanasynchronousmanner.
* Iwasresponsibleforimplementingthetestframework(Junit&Mockito)andSwaggerdocumentation.
* I was involved in developing SQOOP jobs to fetch Oracle data into Cloudera HiveandimplementedOozieworkflowsforscheduleddataingestionintoHivetables.
* CreatedElasticSearchindexesonHivedataanddevelopedsearchquerypatternstovisualizedatausingKibana.

**Environment:**ShellScripting,Cronjobs,ClouderaCDH5.10.2,HDFS,HIVE,HBaseSqoop,Oozie,javaRestAPI’s,CherryPy

# BimarianInformationTechnologies,India Jan 2015 -Dec2016SoftwareEngineer

**Neva–FogHornSystems**

Nevaisabusinessprocessintelligenceproductfortheconnectedenterprisethatcomprisesofself-curated, consolidated digitization of knowledge that required crawling all public/privatestructured/free-text sources, organize knowledge in the context of user’s past interactionswith

BPMsystemsandrespondtoqueriesthroughpredicatereasoning(NLP-based).TheinitialusecasesweretargetedtouseNLP,andAItechniquestoimproveuserexperienceandTAT(Turnaroundtime)for customer support on servicenow.com via question answering system that can bring up the mostrelevant"answers"toaquestioninagivendomain.

# Responsibilities:

* Iwastaskedtobuildacrawlandindexingframeworktofetchpublicwebinformation(servicenow.com,applesupport,dellsupport),clean,transform,mergeandstoreitinHBase.
* ImplementedcrawlingusingScrapyframeworkinPythonandusedpandastocleanseandmergethedataintoHBase.
* DevelopedSolrRESTAPIstoindexHBasedataandperformconditionalandfreeformsearchoperations.
* DevelopedslackintegrationstoenablesearchfunctionalityonindexingfromHBasetopowerNeva’sincidentmanagementsystem.

**Environment:**Scrapy,Python,HBase,Solr,Slack,ClouderaCDH5.7.1,JavaRestAPI’s

# BimarianInformationTechnologies,India Jan 2015 - Dec 2016SoftwareEngineer

**IOTPlatform-FogHornSystems**

The goal was to develop an end-to-end IOT application platform for FogHorn systems to processmassiverealtimedatacollectedfromedgedevices(FogHornCollectors).Lambdaarchitecturewasdeemed necessary for channeling edge device data through multiple REST API's for real timestreaming and batch data handling. Customers needed multi-tenant implementation to servevarious industrial segments and a channeled approach for tenant management. The IoT platformwas applied to multiple industry use cases - PumpCavitation detection, Parkinson's HealthCare,vehicletrackingandfoodprocessing.

# Responsibilities:

* Iwasresponsiblefordevelopingvariouscomponentmanagementplatforms.
* Developed spark streaming based consumer pipelines JSON sensor data from Kafka topics(producedbyedgedevices)andperformingestionofrawandprocesseddataintoHBasetables.
* Iwasinvolvedindeveloping(Jersey)basedRESTAPIstopowerIoTdashboardsinamicrobatchmannerreadingfromHBasetables.
* ImplementedOozieworkflowtolaunchbatchSparkSQLjobshistoricalaggregationsondashboards.
* Developed end-to-end workflow Kafka topics, consumer spark pipelines, batch andstreamingdashboardAPIsandbatchanalyticsusingSparkSQLacrossseveralusecasestopredictanomaliesinincomingdata.

**Environment:**ApacheSpark,SparkStreaming,SparkSQL,HBase,Oozie,ClouderaCDH5.4.8

# PrincetonITServices,India Sep 2013 -Dec 2014SoftwareEngineer

**EnergyManagementInformationSystem:**

The goal of this in-house project was to develop a Hadoop based Energy Management InformationSystem(EMIS)thatcaningest,processandrunpredictiveenergy,demandandpowerfactormodelsonstreamingsmartenergymeterdatapackets.Severalanalyticandpredictivemodelswereneededto provide insight into domestic and industrial consumer patterns like power/water consumption,quality,billing,wastage,demandfluctuationsetc.

# Responsibilities:

* Asasoftwareengineer,Iwascollaboratingwithpeerengineersindevelopingtheingestion,ELT,pre&amp;postprocessingdatapipelinesforpredictiveMLmodelsandanalyticreportgenerationusingSQLonClouderaHadoopplatform.
* IhaddevelopedingestionpipelinestostreamincomingJSONdatamessagesfromsmartmetersandpersisttheminHBasetables.
* I have developed Hive tables on HBase using Serde and implemented several HiveQLs forgeneratingbatchanalyticsreportslikeenergyoverview,averagepowerfactor(pf),demand&amp; utilization, load factor, peak demand at various granular schedules via Oozieorchestration.
* IwasalsoinvolvedinbenchmarkingqueriesacrossvariousframeworkslikeHive,Impala&amp;PigLatinscripts.

**Environment:**Shellscripting,Hive,Impala,HBase,PigLatinScripts,ClouderaCDH5.4.8asHadoopdistribution