Anusha Jallipalli
anusha.jallipalli@gmail.com
(774)506-9576

Summary

- I'm a senior big data professional with 10 years of experience in data engineering on multiple public cloud platforms (GCP, AWS & Azure).
- I have extensive background in distributed storage and processing frameworks like Hadoop & Spark (Cloudera, Hortonworks, Databricks) and have delivered numerous big data and data lake modernization solutions for customers across the globe.
- I'm highly proficient in programming using advanced java & python and have experience building reliable APIs and web services using high quality coding and testing methodologies.
- I'm equally proficient in writing high quality SQL interacting with several databases (like Postgres, MySQL, SQL Server, Oracle etc) and modern data warehouse systems (like Hive/Spark SQL, BigQuery, Redshift, etc).
- I have designed and implemented several batch, streaming and event driven data pipelines orchestrated on cloud reading from a variety of source systems, transforming (ELT & ETL patterns) and landing in high value targets like BigQuery, CloudSQL, Azure ADLS, etc.
- I have extensive experience in designing, implementing and deploying data applications across ingestion, ELT/ETL and consumption layers using modern cloud data engineering services.
- I have strong expertise in automation, workflow orchestration and deployment of data pipelines using agile SDLC and DevOps principles.
- Worked on end to end migration of real time ingestion framework from kafka to BigQuery on GCP Platform.
- Azure/ADLS and Azure Data Factory pipelines and data engineering/pre processing for ML models.
- Experience in GCP dataflow pipelines.
- Experience in AWS serverless and API endpoint development.
- Involved in developing Restful Web Services with python Flask, java(jersey).
- Developed Async rest service using Oozie and java(jersey).
- Involved in creating Junit test cases for code coverage.
- Develop the test automation scripts to validate the builds in the CI/CD pipeline.
- Experienced in developing end-to-end IOT (Internet of Things) applications using Spark.
- Efficient in configuring MapReduce, Hive, Pig, Sqoop and shell orchestrations with Oozie.
- Experienced on Hadoop distributions like Cloudera.
- Hands on working knowledge on NoSQL databases.

Education:

- Bachelor of Technology in Computer Science and Engineering, Andhra University – 2012

Skills:

- **Big Data Tools:** Big Data Hadoop Ecosystem, Apache Spark, MapReduce, Pyspark, Hive, Kafka, Flume, Oozie, Zookeeper, Sqoop, HBase.
- **Languages:** Java, Python(Flask, CherryPy, FastAPI, SQLAlchemy), Shell Scripting
- **Cloud Tools:** GCP(BigQuery, Dataflow, Apache Beam, Kafka, Cloud Composer, Airflow, Dataproc), AWS(S3, Lambda, API Gateway, Cloud watch, cloud logs) and Azure (ADF Pipelines, ADLS, Databricks)
- **Frameworks:** Spring boot with Hibernate, Jersey RestAPI's
- **Version Controls:** Git and Bitbucket

- **Databases and Tools:** PostgreSQL, MySQL, SQL Server, Oracle, Hive, Impala, HBase
- **Testing Tools:** Junit, Mockito and PowerMockito
- **Build and Deploy:** Maven and Jenkins(CI/CD), Swagger UI,Insomnia
- **File and table formats:** Working knowledge on popular file formats(JSON, Text, CSV, ORC, Avro, Parquet) and table formats (Hive tables, Delta, Iceberg, Hudi)

Certifications:

- Google Cloud Certified Professional Data Engineer
  (https://www.credential.net/8f146020-271f-48da-be32-9bbeb2a83bc1?key=6a9953ae37f7dc8c69929abbdde4dd4 e07698a259 0188ed3f7f1bae4555e9505)

Professional Experience :
Client: Cigna **Mar 2024 - Till Date**
Role: Senior Data Engineer (Lead)

Responsibilities:

- Spearheaded a team of data engineers, providing mentorship, guidance, and technical leadership to drive successful project deliveries.
- Worked on reading and writing multiple data formats like text files, JSON and Parquet on HDFS using PySpark.
- Configured real-time data ingestion from Kafka, seamlessly processed with Spark Streaming, and stored in HDFS. Actively oversaw Spark job progress using the Spark Application Master, ensuring effective logging and monitoring.
- Deployed machine learning models for real-time prediction serving using platforms like TensorFlow Serving, Kubernetes, or custom API endpoints.
- Configured Docker networking to facilitate communication between containers and services, ensuring seamless data flow in real-time applications.
- Designed and implemented scalable data architectures, enabling the processing and analysis of massive datasets with optimal performance.
- Designed solutions for processing high-volume data streams, including ingestion, processing, and data provisioning using Hadoop ecosystems.
- Conducted performance tuning and optimization of Apache Beam and GCP Dataflow jobs, maximizing throughput and minimizing processing times for large-scale data sets.
- Implemented data quality checks within Apache Beam and GCP Dataflow pipelines, ensuring the accuracy and reliability of processed data.
- Import the data from different sources like HDFS/HBase into Spark RDD and perform computations using PySpark to generate the output response.
- Leveraged Azure services, including Azure Databricks, for efficient data ingestion and processing.
- Developed real-time anomaly detection models to identify and alert on unusual patterns or deviations in streaming data, helping detect issues as they occur.
- Designed and maintained test data sets, ensuring data accuracy and relevance, and developed processes for data extraction, transformation, and loading (ETL) for UAT.
- Successfully harnessed pivotal technologies such as Kafka, Spark Streaming, HDFS, AWS Lambdas, and Google Cloud Platform, optimizing data processing and management across diverse landscapes.
- Utilize Data Build Tool (dbt) to streamline data transformation and modeling processes for enhanced data analysis.
- Proficiently optimized SQL queries and data processing jobs in BigQuery to improve query performance, reduce costs, and enhance data retrieval speed.
- Collaborate with cross-functional teams by integrating dbt with version control systems like Git, promoting effective teamwork and version tracking.

- Successfully led the migration of legacy data systems to modern data platforms, achieving improved performance, reliability, and ease of maintenance.

Environment: Hadoop Ecosystem(HDFS, Hive, Impala, HBase, Zookeeper, Sqoop). Spark, Spark-SQL, Spark-Streaming, Kafka, Java, Python, GCP Dataflow, BigQuery

Accenture, India **June 2021 – Aug2022**
Sr. Data Engineer
(Lead)

GMI (General Mills)— ODS Real-time ingestion
Customer needed an ingestion framework to be designed and implemented with ability to perform real-time data ingestion from the ODS layer, streaming into GCP BigQuery. In addition, dynamic schema evolution and change data capture (CDC) were extremely important features to support their downstream business systems with least disruption.

Responsibilities:

- As a senior data engineer lead on the project, I was responsible for leading the customer discussions including discovery, assessment, design and implementation of the ingestion framework.
- Developed data mapping documents (DMDs) between source SAP and target GCP BigQuery systems.
- I was leading a team of 8 engineers and was instrumental in collaboratively developing transaction Data Flow pipelines reading from Kafka topics and streaming to GCP BigQuery.
- I was also responsible for developing Dataflow jobs for handling schema evolution in conjunction with BigQuery supported schema evolution features.
- I was collaborating with customer teams to implement CDC operations using stored procs and had developed error handling and dead letter workflows using Dataflow and Kafka for producing and consuming efficient and reliable results in BigQuery.
- Developed composer workflows (Airflow DAGs) for scheduling and launching dataflow jobs.
- Worked with customer & Google SMEs and optimized several Dataflow job patterns involving reading from multiple Kafka topics and writing to respective BigQuery tables.

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

Accenture, India **Jan 2021 – May 2021**
Senior Data Engineer (Lead)

GMI (General Mills) – IDF
Customer needed Data Ingestion Framework (IDF) that permits the rapid consumption of data from several 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 to data management that minimizes manual code generation and maximizes component reuse, the framework needed the following 4 features:

- Single framework to perform all data ingestions consistently into GCP data lake
- Enables tracking metrics, events and notifications for all data ingestion activities
- Single consistent method to capture all data ingestion along with technical metadata, data quality, data lineage and governance
- Provide a proper data governance with Search & Catalogue to find data within DataLake

Responsibilities:

- As a senior data engineer on the project, I was leading a team of 4 engineers in implementation of IDF framework components.
- I was responsible for developing the framework database structure by designing and implementing data models (Postgresql on GCP CloudSQL) using SQLAlchemy.
- I was involved in development, testing & production deployment of core ingestion and common service APIs in the IDF framework.
- I developed over 50 Restful APIs spanning across ingestion, job/process tracking, lineage, data quality modules using FastAPI (Python) and Swagger.
- The APIs were deployed on GKE containers and I helped integrate Redis cache and vault features.

Environment: Python, FastAPI, GCP CloudSQL, PostgresQL, SQLAlchemy, Redis, Swagger

Accenture, India **Apr 2020 - Dec 2020**
Senior Data Engineer (Lead)

Intient Analytics Studio – Accenture
Accenture needed to build Google cloud-based Analytics Studio modules for its patient analytics platform called INTIENT. This studio required developing several Rest API services supporting several functions across the INTIENT platform including source integration, data ingestion, data search &
preparation, feature engineering, model development, training & calibration and report generation on Google Cloud Platform.

Responsibilities:

- As a senior data engineer, I was responsible for collaborating with a team of 8 engineers and developed several Rest APIs for onboarding user personas & groups, provisioning relevant GCP services (AI-platform python/R notebooks, BigQuery, dataflow, dataproc, etc) and deploying IAM role mappings as per Intient Analytic Studio requirements.
- I was involved in end-to-end development, testing and production implementation of APIs on GKE.
- I helped the team in implementing the backend CloudSQL (Postgres) database modules, Java Spring Boot applications and Okta integrations and performed CRUD operations on user profiles, and hibernate mappings for DB interactions.
- I was also responsible for implementing a test & quality framework for the APIs using Swagger, Jacoco, SonarQube and Junit/Mockito.
- I had developed several high levels and low-level technical design and implementation collaterals to guide the developer teams.

Environment: Java with Spring boot and Hibernate, CloudSQL (PostgresQL), GKE, GCP Services (Dataproc, BigQuery, Dataflow, AI Platform notebooks, AutoML, Dataprep), Jenkins CI/CD, Gitlab and Confluence

Accenture, India **Nov 2019 – Mar 2020**
Senior Data Engineer

Unilever Livewire 2.0 (country - South Africa)
Unilever needed to implement a data ingestion framework from several internal and external systems into a central Azure data lake for further downstream application consumption for building data driven insights and high impact models. This project involved creating a value chain to help address the challenges of acquiring and ingesting data from multiple sources which provide competitive advantage to Unilever.

Responsibilities:

- As a senior data engineer, I contributed by providing thought leadership to the team in building and implementing a reliable data ingestion framework (LiveWire 2.0) with reusable ingestion templates.
- I was responsible for ingesting data from external file sources to Azure Data Lake (ADLS) using Azure Data Factory (ADF) pipelines.
- Developed and deployed several batch pipelines, orchestrated and scheduled these as per SLA guidelines.
- Developed PySpark modules in Databricks for data quality & validations, deduplication and perform transformations on the ingested raw datasets in ADLS.

Environment: Azure Data Factory (ADF), Azure Data Lake Service (ADLS), Databricks Notebooks, PySpark, SparkSQL

Accenture, India **Dec 2018 - Oct 2019**
Senior Data Engineer

Royal Caribbean Cruise Line – Price Recommendation & Elasticity Engine

The goal of the project is to build a Price Recommendation Engine (PRE) Automation Platform to automate more frequent and precise price changes early in the booking window (Weeks to Sail >52 weeks), where historically sailings have not been actively managed. In early experiments, this approach has reduced variance from Track and variance in book position across sailings. The overall PRE-Automation solution has the following tracks:

- A price recommendation engine (i.e. the "PRE") comprised of business logic coded in python that converts various inputs – including configurable parameters, Track, elasticities and historical pricing & booking data – into a suggested price change
- Data foundation comprised of required tables and supporting data pipelines
- Machine-learning based on price elasticity models
- API layers to connect underlying tables and recommendations with future web-based interfaces or another UI (i.e. 'Autopilot')

Responsibilities:

As a senior data engineer, I was collaborating with a team of 12 engineers to develop the following features on Azure & AWS cloud platforms.

- Price Recommendation Engine (PRE) – I developed daily and weekly data ingestion (historical price, booking, track & elasticity data from Oracle to Azure data lake), pre-processing and feature engineering pipelines on Azure Databricks using PySpark and SparkSQL.
- 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, call forecast datasets from Oracle to ADLS), Transform, execute Elasticity and PRE-Models pre-processing using Databricks and persist the PRE- model output back to Oracle database. I had developed ADF pipelines, orchestrated and scheduled these as per SLA guidelines.
- Price Elasticity Models – I was responsible for developing data pre-processing pipelines (ship sailing & cross product price datasets) in Azure ADF to support regression and mixed effect supervised models.
- API Layer – I have built lightweight back end Services and API layer to retrieve and update Parameter tables data in Oracle EDSSP database. I have developed java REST APIs and AWS Lambda functions with API Gateway. I have also contributed in functional testing of the APIs and deployment of application using Cloud Formation Templates (CFT)

Environment: Java, AWS (S3, Lambda, API Gateway, CFT), Azure (ADF, Data pipeline, Databricks Notebooks, ADLS), PySpark, SparkSQL

Kogentix(Acquired by Accenture), India Jan 2017 -Nov 2018 Big Data Engineer

DMLE (DIGITAL MAPPING LEARNING ENVIRONMENT)-NIELSEN:
Customer needed to build a Digital Mapping Learning Environment (DMLE) to support automated ML model training, testing and predictions (across categories like Matching, Linking, Predictive Modeling & Clustering) with support from REST API services to read/fetch/write data to and from HDFS and Hive on Cloudera Hadoop. There was a need to build automation via several APIs to launch the model experimentations including training, test and model inference jobs and track model runs in an asynchronous and consistent manner. Customers also needed data ingestion workflows to be developed from Oracle to Hive and use ElasticSearch and Kibana for indexing and visualization of product data.

Responsibilities:

- As a lead Big Data Engineer, I was responsible to develop several REST APIs with Python CherryPy/Flask & Java (Jersey API) to read/fetch/write data from HDFS and Hive and also to launch and track ML jobs (training, test and batch predictions) as Oozie workflows in an asynchronous manner.
- I was responsible for implementing the test framework (Junit & Mockito) and Swagger documentation.
- I was involved in developing SQOOP jobs to fetch Oracle data into Cloudera Hive and implemented Oozie workflows for scheduled data ingestion into Hive tables.
- Created ElasticSearch indexes on Hive data and developed search query patterns to visualize data using Kibana.

Environment: Shell Scripting, Cron jobs, Cloudera CDH 5.10.2, HDFS, HIVE, HBase Sqoop, Oozie, java Rest API's, CherryPy

Bimarian Information Technologies, India Jan 2015 -Dec 2016 Software Engineer

Neva – FogHorn Systems
Neva is a business process intelligence product for the connected enterprise that comprises of self-curated, consolidated digitization of knowledge that required crawling all public/private structured/free-text sources, organize knowledge in the context of user's past interactions with
BPM systems and respond to queries through predicate reasoning (NLP-based). The initial use cases were targeted to use NLP, and AI techniques to improve user experience and TAT (Turnaround time) for customer support on servicenow.com via question answering system that can bring up the most relevant "answers" to a question in a given domain.

Responsibilities:

- I was tasked to build a crawl and indexing framework to fetch public web information (servicenow.com, apple support, dell support), clean, transform, merge and store it in HBase.
- Implemented crawling using Scrapy framework in Python and used pandas to cleanse and merge the data into HBase.
- Developed Solr REST APIs to index HBase data and perform conditional and free form search operations.

- Developed slack integrations to enable search functionality on indexing from HBase to power Neva's incident management system.

Environment: Scrapy, Python, HBase, Solr, Slack, Cloudera CDH 5.7.1, Java Rest API's

Bimarian Information Technologies, India **Jan 2015 - Dec 2016**
Software Engineer

IOT Platform - FogHorn Systems
The goal was to develop an end-to-end IOT application platform for FogHorn systems to process massive real time data collected from edge devices (FogHorn Collectors). Lambda architecture was deemed necessary for channeling edge device data through multiple REST API's for real time streaming and batch data handling. Customers needed multi-tenant implementation to serve various industrial segments and a channeled approach for tenant management. The IoT platform was applied to multiple industry use cases - PumpCavitation detection, Parkinson's HealthCare, vehicle tracking and food processing.

Responsibilities:

- I was responsible for developing various component management platforms.
- Developed spark streaming based consumer pipelines JSON sensor data from Kafka topics (produced by edge devices) and perform ingestion of raw and processed data into HBase tables.
- I was involved in developing (Jersey) based REST APIs to power IoT dashboards in a micro batch manner reading from HBase tables.
- Implemented Oozie workflow to launch batch SparkSQL jobs historical aggregations on dashboards.
- Developed end-to-end workflow Kafka topics, consumer spark pipelines, batch and streaming dashboard APIs and batch analytics using SparkSQL across several use cases to predict anomalies in incoming data.

Environment: Apache Spark, Spark Streaming, Spark SQL, HBase, Oozie, Cloudera CDH 5.4.8

Princeton IT Services, India **Sep 2013 -Dec 2014**
Software Engineer

Energy Management Information System:
The goal of this in-house project was to develop a Hadoop based Energy Management Information System (EMIS) that can ingest, process and run predictive energy, demand and power factor models on streaming smart energy meter data packets. Several analytic and predictive models were needed to provide insight into domestic and industrial consumer patterns like power/water consumption, quality, billing, wastage, demand fluctuations etc.

Responsibilities:

- As a software engineer, I was collaborating with peer engineers in developing the ingestion, ELT, pre &amp; post processing data pipelines for predictive ML models and analytic report generation using SQL on Cloudera Hadoop platform.
- I had developed ingestion pipelines to stream incoming JSON data messages from smart meters and persist them in HBase tables.
- I have developed Hive tables on HBase using Serde and implemented several HiveQLs for generating batch analytics reports like energy overview, average power factor(pf), demand &amp; utilization, load factor, peak demand at various granular schedules via Oozie orchestration.
- I was also involved in benchmarking queries across various frameworks like Hive, Impala

&amp; Pig Latin scripts.

Environment: Shell scripting, Hive, Impala, HBase, Pig Latin Scripts, Cloudera CDH 5.4.8 as Hadoop distribution