**Vijay**

**Sr Data Engineer**

**Phone: (510)-854-1314**

**PROFESSIONAL SUMMARY**

* Nearly **8 years** of IT expertise in with a strong background in designing and implementing scalable data pipelines,cost-effective solutions on **GCP**.
* **Masters in Business Analytics and Information Systems** from **University of South Florida** (May 2021) and Bachelors in Aerospace Engineering from Amrita University (May 2017)
* Experienced in Infrastructure Development and Operations involving **GCP Cloud platforms,Big Query,Pub/Sub,GCS(Google Cloud Storage),Composer,Cloud Function,DataFlow,DataProc,Big Table,Vertex AI**.
* Excellent grasp of **Data Engineering lifecycle** – **Use Case Design** (Defining the business use case, data availability check), **Data Ingestion & Storage** (Fetching Data from APIs,databases,logs) and **Data Processing** (ETL (Extract, Transform, Load) processes, data cleansing, data enrichment)**, Data Security and Governance** (Access control, encryption, auditing, compliance monitoring), **Data Analytics and Exploration** (Running queries, building dashboards, conducting exploratory data analysis (EDA)).
* Good knowledge on GCP service accounts, billing projects, authorized views, datasets, GCS buckets and gsutil commands
* Hands on experience in developing spark applications using **PySpark DataFrame,RDD,Spark SQL.**
* Extensive experience in developing kafka producers and kafka consumers for streaming millions of events per minute on streaming data using **pyspark**,**python** and **spark structured streaming**.
* Designed and implemented data processing pipelines using Spark and Scala.
* Worked on various automation tools like Git,Terraform,Ansible.Experienced in fact dimensional modeling(star schema,snowflake schema),SCD(Slowly Changing Dimensions) - Type 1 & Type 2.
* Expertise in creating,debugging,scheduling and monitoring jobs using airflow and oozie.
* Knowledge on working with various components of Hadoop Eco System **-**HDFS,Map Reduce,YARN,HBase,Hive,Spark,Pig,Sqoop,Flume,Zookeeper,Kafka.
* Experienced in implementing partitioning and bucketing in Hive, knowledge on designing managed and external tables in Hive for performance.
* Hands on experience working with structured and semi structured data storage formats such as parquet,ORC,CSV,JSON.
* Proficient in SQL databases **MySQL, Oracle, SQL Server** and no-SQL databases
* Advanced working SQL knowledge and experience working with Sub Queries, Stored Procedures, Triggers, Cursors, and Window Functions on **MySQL.**
* Excellent working knowledge in **UNIX** and **Linux shell** environments using command line utilities.
* Experienced in **Agile Methodologies**, Scrum stories and sprints experience in an AWS based environment.
* Skilled in Collaborating with Cross Functional teams including developers, project managers, data scientists to ensure smooth migration process.
* Strong experience in managing projects include planning, defining project scope and deliverables, managing project timelines, **coordinating with stakeholders**, overseeing execution, managing project risks, **tracking progress** and ensuring the project success
* Ability to **communicate effectively with stakeholders** at all levels through successful presentations of data-driven initiatives and strategies

**Technical Skills:**

|  |  |
| --- | --- |
| **Programming Languages** | Python-3.10 & 2.7,SQL,pyspark, Java,Scala |
| **Python Libraries** | Pandas, NumPy, Beautiful Soup,Scikit Learn, Matplotlib, Pickle, SciPy, PyTables, regex, Sklearn, Tensorflow, os, log,Pytorch |
| **DBMS** | MySQL,Oracle,Teradata,MongoDB,PostgreSQL |
| **IDE's/ Development Tools** | Eclipse,Jupyter notebook, Visual Studio, PyCharm, R Studio and Sublime Text |
| **Version Control** | GIT (GitHub), Jenkins |
| **Deployment Tools** | Docker, Kubernetes |
| **Methodologies** | Agile, Scrum and Waterfall |
| **Reporting & Documentation Tools** | Tableau,Looker,Power BI, MS-Word, and MS-PowerPoint. |
| **Operating systems** | Linux/Unix, Windows, Mac |
| **Cloud Environment** | GCP Services – Big Query,GCS,DataProc,DataFlow,Cloud Composer,Cloud Function,Pub/Sub,Big Table,Vertex AI,Vertex AI Vector Search |
| **Big Data Tools** | Hadoop Eco System : Map Reduce,Spark 2.3,Airfloe,Nifi,HBase,Hive,Pig,Sqoop,Kafka,Oozie,Hadoop,Scala,Scalding |

**Professional Experience**

**Client: American Express - New York City, NY (Hybrid)**    **June 2024 – Present**

**Role: Senior Data Engineer(Global Loyalty and Benefits)**

* Responsible in supporting engineers and product managers in rapid development and deployment in agile fashion.
* Troubleshot production issues, formulated action plans, conducted root cause analysis, created and updated support incidents, and collaborated with support teams to ensure the stability and performance of Amex next-gen big data platform.
* Automated manual data validations using Bash and SQL, reducing manual effort, saving engineering time, and improving efficiency by 70%.
* Responsible for setting up audit tables in **BigQuery** to certify the data received from upstream systems before it was integrated into our next-gen big data platform. This process ensured that the data met the necessary quality standards and could be trusted by the fintech team. To enable real-time data validation and processing, I utilized Google Cloud's Pub/Sub for messaging and Cloud Functions for automated execution, streamlining the process and ensuring timely updates for end users.
* I collaborated with the Data Governance Lead and Product Owner to conduct a proof of concept (POC) to build a functionality that automatically generates SQL queries using LLMs and Generative AI, aimed at supporting ad-hoc requests from various business users and assisting clients handling millions of dollars in transactions, ensuring timely and efficient data retrieval and decision-making.

**Environment:** GCP, Big Query,Kafka,Pub/Sub,Cloud Function,Stratergic thinking and Collaboration.

**Client: Univision - New York City, NY (Hybrid)**    **Aug 2022 – May 2024**

**Role: Senior Data Engineer**

* Engaged in the entire spectrum of the Software Development Life Cycle (**SDLC**), encompassing meticulous requirements gathering, in-depth analysis of customer specifications, and the intricate design and development of robust data engineering pipelines.
* Integrated data from multiple sources like amazon, apple, dish, Facebook, Samsung, Verizon, media ocean, Roku and wide orbit. Write DBT models for data aggregation, orchestrate it via GCP Cloud Composer(Airflow) and move it to big query.
* Implemented techniques (windows functions, common table expressions, UDFs, etc.) for data investigation, transformation, analysis, and optimizing for scalability and ease of maintenance.
* Implemented incremental model for multiple databases to ingest data from GCS buckets to GCP BigQuery.
* Worked closely with Product Owners (PO) for requirements, Jira stories and source to target mappings (STM). Analyzed STM’s based on source databases.
* Developed custom Airflow operators by enhancing the existing Airflow operators to perform validations and execute multiple GCP BigQuery SQLs with a single operator.
* Developed the cloud functions (python) to perform feature definition file validations and then push the files into the GCS buckets.
* Implemented Data visualizations using Looker based on client needs, involved in Scheduling and sharing Looks and dashboards, Table calculations, Looker expressions, Custom and advanced filters, Impacts of pivoting.
* Implemented persistent derived tables, caching policies, version control and assess code quality, utilized SQL Runner for data validation.

**Environment:** GCP, Big Query, DBT, Python, Airflow, Looker.

**Client: Bed Bath and Beyond - Union, NJ**    **Nov 2021 – Aug 2022**

**Role: Senior Datawarehouse Engineer**

* Migrated Netezza/Teradata Analytics to Google cloud Platform (GCP) with orchestration and pipeline development using Apache Airflow
* Re-architect and optimize GCP Big Query operational and analytical data sets using best practices
* Created datasets by integrating from multiple sources to enable Analytics ensuring data quality (Accuracy, Completeness & Timeliness)
* Partnered with IT to analyze the source data (EDW/DW) and develop automated data processes for data mart builds
* Implemented processes that support campaign management for marketing initiatives such as customer segmentation
* Collaborated with the Analytics community to understand their data needs and translate them into deliverables in the form of technical specs
* Worked with external vendors to share analytics to identify demographics for better targeting such as college population.

**Environment:** GCP, Cloud SQL, Big Query, Teradata, Python, Airflow.

**Client: Citi Bank - Tampa, FL**    **Jul 2020 – Nov 2021**

**Role: Data Engineer**

**Responsibilities:**

* Experience in building multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate task among the team.
* Using g-cloud function with Python to load data in to Big query for on arrival csv files in GCS bucket.
* Process and load bound and unbound data from Google pub/subtopic to Big query using cloud Dataflow with Python.
* Designed Pipelines with Apache Beam,Google Cloud Dataflow and orchestrated jobs into GCP.
* Developed and Demonstrated the POC, to migrate on-prem workload to Google Cloud Platform using GCS, Big Query, Cloud SQL and Cloud DataProc.
* Documented the inventory of modules, infrastructure, storage, components of existing On-Prem data warehouse for analysis and identifying the suitable technologies/strategies required for Google Cloud Migration.
* Design, development and implementation of performing ETL pipelines using python API (pySpark) of Apache Spark.
* Worked on GCP POC to migrate data and applications from On-Prem to Google Cloud.
* Exposure on IAM roles in GCP.
* Create firewall rules to access Google data procs from other machines.
* Process and load bound and unbound Data from Google pub/subtopic to Big query using cloud Dataflow with Python.
* Setup GCP Firewall rules to ingress or egress traffic to and from the VM's instances based on specified configuration and used GCP cloud CDN (content delivery network) to deliver content from GCP cache locations drastically improving user experience and latency.

**Environment:** GCP, Cloud SQL, Big Query, Cloud DataProc, GCS, Cloud SQL, Cloud Composer

**Client: PWC,Tampa,FL**    **Dec 2019 – Jun 2020**

**Role: Data Engineer**

* Worked on implementing scalable infrastructure and platform for large amounts of data ingestion, aggregation, integration, analytics in Hadoop using Spark and Hive.
* Worked on developing streamlined workflows using high-performance API services dealing with large amounts of structured and unstructured data.
* Worked on processing un-structured data in JSON format to structured data in parquet format by performing several transformations using Pyspark.
* Developed Spark applications using spark libraries to perform ETL transformations and thereby eliminating the need for utilizing ETL tools.
* Developed the end-to-end data pipeline in spark using python to ingest, transform and analyses data.
* Developed Databricks ETL pipelines using notebooks, Spark Data frames, SPARK SQL, and python scripting.
* Created Hive tables using HiveQL, then loaded the data into Hive tables and analyzed the data by developing Hive queries.
* Created and executed Unit test cases to validate transformations and process functions are working as expected.
* Worked on scheduling Control-M workflow engine to run multiple jobs.
* Written shell scripts to automate application deployments.
* Implemented solutions to switch schemas based on the dates so that the transformation would be automated.
* Developed custom functions and UDFs in python to incorporate methods and functionality of Spark.
* Developed data validation scripts in Hive and Spark and perform validation using Jupiter Notebook by spinning up the query cluster in AWS EMR.
* Executed Hadoop and Spark jobs on AWS EMR using data stored in Amazon S3.
* Implemented Spark RDD transformations to map business analysis and apply actions on top of transformations.
* Worked on Data serialization formats for converting complex objects into sequence bits by using Parquet, Avro, JSON, CSV formats.
* Implemented spark using scala and spark SQL for faster testing and processing of data.
* Implemented spark using scala and utilizing data frames and Spark SQL API for faster processing of data.
* Expertise in implementation of spark using scala and spark SQL for faster testing and processing of data responsible to manage data from different sources

**Environment:** Hadoop, Hive, Zookeeper, Sqoop, Databricks,Spark, Scala, Python, SQL, Bit bucket, AWS, Linux.

**Client: Thomson Reuters,Bengaluru, India**    **June 2016 – Jan 2019**

**Role: Big Data Engineer**

**Responsibilities:**

* Created and managed nodes that utilize Java jars and python, shell scripts for scheduling jobs to customize Data ingestion.
* Developed Pig Scripts for change data capture and delta record processing between newly arrived data and already existing data in HDFS.
* Developed and performed Sqoop import from Oracle to load the data into HDFS.
* Created Partitions, Buckets based on State to further process using Bucket based Hive joins.
* Created Hive tables to store the processed results in a tabular format.
* Scheduled MapReduce jobs in production environment using Oozie scheduler and Autosys.
* Developed Kafka producer and brokers for message handling.
* Imported the data to Hadoop using Kafka and implemented the Oozie job for daily imports.
* Configured Kafka ingestion pipeline to transmit the logs from web server to Hadoop.
* Worked with POC’s for stream processing using Apache NIFI.
* Worked on Hortonworks Hadoop Solutions with Real-time Streaming using Apache NIFI.
* Analyzed Hadoop logs using Pig scripts to oversee the errors caused by the team.
* Performed MySQL queries for efficient retrieval of ingested data using MySQL workbench.
* Implemented data ingestion and transformation using automated workflows using Oozie.
* Created and generated audit reports to notify security threat and track all user activity using various Hadoop components.
* Designed various plots showing HDFS analytics and other operations performed on the environment.
* Worked with Infra team for testing the environment after patches, upgrades and migration.
* Developed multiple Java scripts for delivering end-to-end support while maintaining product integrity.
* Enhanced and optimized Spark/Scala/ pyspark jobs to aggregate, group and run data mining tasks using the Spark framework.
* To meet specific business requirements wrote UDF's in Scala and Pyspark.

Environment: HDFS, Hive, MapReduce, Pig, Spark, Kafka, Sqoop, Scala, Oozie, Maven, GitHub, Java, Python, MySQL, Linux.

**Education:**

* Bachelors in aerospace engineering,Amrita University, May 2017
* Masters in business Analytics and Information Systems,University of South Florida, May 2021.