s

**Yoshitha Varre**

**GCP Data Engineer**

**510-854-1314**

**Summary:**

* More than **8+ years** of IT experience in Cloud Data Engineering, Hadoop, **Big Data, Data warehouse, system analysis** gathering requirements supporting project management and Legacy Application Development and Support.
* Sustaining the **Big Query**, **PySpark** and Hive code by fixing the bugs and providing the enhancements required by the Business User.
* Working with **GCP** cloud using in **GCP Cloud storage, Dataproc, Data Flow, Big Query, Cloud Composer, Cloud Pub/Sub.**
* Proficient in creating tables, views, stored procedures, datasets in Big Query. Strong hands-on experience on **SQL** and **PL-SQL**.
* Expertise in creating pipelines in google **dataflow to handle streaming data**.
* Created DataMart’s using Pyspark, Big query job config and **Airflow** big query operators as per the reporting need.
* Created Airflow dags to schedule batch jobs, trigger Big Query or Pyspark operator to refresh the DataMart’s. Utilized Airflow variable to control the **Airflow job tasks** and the logic.
* Collaborated with product teams to build Looker dashboards that streamlined product performance tracking and reporting.
* Expert in working with cloud **PUB/SUB** to replicate data real-time from source system to **GCP Big Query**.
* Good knowledge on **GCP service accounts, billing projects, authorized views, datasets, GCS buckets and gsutil commands.**
* Created partition tables for the daily and monthly data snapshots.
* Good knowledge on Tableau workbooks, custom SQL, filters, calculated columns, performance recording. Worked closely with tableau consultants to create dashboards, reports, self-service analytics.
* Created tool to calculate data statistics using **Apache Spark 2.0** to read **Parquet files** and generate output in **Json format**.
* Strong performance tuning skills in tuning **SQL Queries and stored procedures**.
* Proficient programming and design skills using **Java, SQL and Python.**
* Experience in Service now to create the Incidents, Change Requests and Service tasks.
* Experience in working with Google Production Support team and handled the Incidents based on the SLA’s.
* Involved with Business team to solve the Incidents or Jira according to the team policies.

**Certification:**

**Certified** in Google Cloud Platform **Professional Data Engineer**.

https://googlecloudcertified.credential.net/profile/456ff3dc719e5379e8d50f01f3aeab962025f89f

**TECHNICAL SKILLS:**

* **Technology:** Apache Spark, Hadoop Ecosystem, Oracle, Teradata, Visual Basic 6.0, Apache Airflow, Terraform, GCP
* **Programming Languages:** Python, C, C++, Java
* **Hadoop Platforms:** Cloudera (5.6, 5.7, 5.8)
* **Databases:** Oracle, DB2, Teradata, Postgres, PL SQL, DB2
* **Big Data Ecosystem:** Spark, HDFS, Map Reducing, Hive, Sqoop, Kafka, RabbitMQ, HBase, Python Pandas.
* **Operating Systems:** Windows, UNIX, Mac OS
* **Software Applications:** Application Software, -E-Commerce Software, Database Systems, Web Portal Software, Data Warehousing
* **Data Modeling:** Star Schema Modelling, Snowflake Modeling
* **Version Control Tools:** GitHub, VSS, Perforce, SVN, TFS
* **Cloud Computing:** GCP, Big query, Dialog Flow, Jenkins, CI/CD Google Colab, Dataflow, Cloud Composer, Cloud Pub/Sub, Big table, Data Store, Cloud Function.

**PROFESSIONAL EXPERIENCE:**

**Client: Dollar Tree**

**Role : GCP Data Engineer Oct 2024 – Till date**

**Responsibilities:**

* Designed and implemented scalable, near real-time data pipelines using Google Cloud Run, Cloud SQL, Cloud Scheduler, and Firestore to process and analyze customer behavior data for online retail stores. This provided actionable insights into shopping trends, inventory management, and personalized marketing strategies.
* Integrated Google Analytics 360 with BigQuery to automate data extraction and reporting, enabling real-time dashboards for marketing teams to optimize and personalize campaigns dynamically.
* Developed and deployed data-driven marketing strategies on Google Cloud Platform (GCP), leveraging machine learning models to predict customer behavior, improve targeting accuracy, and optimize ad spend in real time.
* Architected and built cloud-native applications that automatically scale based on demand, significantly reducing infrastructure costs while ensuring high availability and performance during peak retail events (e.g., Black Friday, Cyber Monday).
* Implemented enterprise-grade security measures using Google Cloud IAM, Data Loss Prevention (DLP), and VPC Service Controls to safeguard retail applications and protect sensitive customer data.
* Led end-to-end execution of digital transformation projects, from requirement gathering to solution design, cloud implementation, deployment, and post-launch monitoring, ensuring seamless integration and operational efficiency.
* Integrated DevOps and CI/CD pipelines to streamline deployment processes, enabling faster iteration and delivery of digital retail solutions while maintaining system reliability and security.
* Conducted cost optimization and financial analysis using GCP Cost Management, BigQuery, and predictive analytics to monitor and reduce cloud expenses, leading to significant savings for retail clients.
* Delivered executive-level presentations and technical workshops on cloud strategy, data analytics, and AI-driven retail innovations, ensuring alignment with business objectives and driving digital transformation initiatives.

**Client: Macy’s Aug 22 – Oct 2024**

**Role: GCP Data Engineer**

**Responsibilities:**

* Hands on experience in **Google Cloud as Cloud Storage, Dataflow, Cloud Composer, Big query, Cloud Functions, Cloud Pub sub**.
* Writing Python scripts to load the data from big query to big query using Dataflow and Composer.
* Involving with different teams and back tracking the flows and experience in solving the critical issues.
* Identified and documented strategies, tools and phases in migration to **Google Cloud Platform.**
* Working experience with Support team and took the responsibility for the issues in production.
* Regularly presented **Looker**-generated reports to executive stakeholders, facilitating data-driven decision-making.
* Experience in solving the priority issues and involving in SOC calls while there is any production issues.
* Responsible for Google Production Support for environments and application related issues.
* Involving with business teams and different teams to solve the critical issues.
* Handling all priority incidents created by the end users and provide the solution on time via Service Now
* Utilized **Looker** to analyze customer behavior data, leading to a 15% increase in customer retention rates.
* Analyzing the different databases (**Teradata and Big Query**) from which the data is loading into the multiple reports and fixing the issues in the reports if any.
* Experienced in handling the jobs in Control\_M and Airflow UI.

**Environment:** GCP, Cloud SQL, Big Query, Cloud Dataproc, GCS, Cloud SQL, Cloud Composer, Hadoop, Hive, Teradata, SAS, Teradata, Spark, Python, Java, SQL Server, Polaris and Omega, Service Now, Confluence.

**Client:** Verizon **Jan 22 – Aug 22**

**Role:** GCP Data Engineer

**Responsibilities:**

* Migrating an On-prem database to **Google cloud Big Query.**
* Built Data pipelines using Apache Beam framework in **GCP** for **ETL** related jobs for different airflow operators.
* Used Cloud products like **GCS, Cloud functions, Data Flow, Big Query and Cloud composer.**
* Loading data into **GCS** buckets that is transferred from on-prem databases.
* Built data pipelines in Airflow in **GCP** for ETL related jobs using different airflow operators.
* Exposure to IAM roles in GCP. Project-wide IAM policies and Firewall rules on GCP to restrict unwanted access from public internet and individuals.
* Used Big Query command line utilities to load data into Big Query tables for arrival of csv/text files in GCS bucket.
* Used Cloud Shell SDK in GCP to configure the services Dataproc, Google Cloud Storage and Big Query.
* Created Airflow Scheduling scripts in Python.
* Scheduling the dataflow from BQ staging file to BQ destination table using Cloud Composer.
* Used g-cloud Functions using **python** to load Data into **Big Query** for on arrival files in GCS Bucket.

**Environment:** GCS Buckets, Big Query, Cloud Composer, Cloud Storage, Cloud Functions, Airflow, MySQL, Python, IAM Security, VPC Configuration.

**Client:** JPMC, Houston, TX **May 20 – Dec 21**

**Role:** Data Engineer

**Responsibilities:**

* Developed ELT jobs using Apache beam to load data into **Big Query** tables.
* Designed Pipelines with Apache Beam, Kubeflow, Dataflow and orchestrated jobs into **GCP.**
* Experience in processing unbound Data from Google pub/sub topic to Big query using cloud Dataflow with Python.
* Experience in writing simple and complex SQL scripts to check and validate Dataflow in various applications.
* Developed and Demonstrated the POC, to migrate on-prem workload to Google Cloud Platform using **GCS, Big Query, Cloud SQL and Cloud Dataproc.**
* Identified and documented strategies, tools and phases in migration to Google Cloud Platform.
* 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**.
* Experience in solving the priority issues and involving in SOC calls while there is any production issues.
* Involving with different teams and back tracking the flows and experience in solving the critical issues.
* Developed cloud functions to trigger the cloud composer to spin up the DataProc cluster.
* Process and load bound and unbound Data from Google pub/sub topic 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.
* Experience in analyzing the different databases (Teradata and Big Query) from which the data is loading into the multiple reports and fixing the issues in the reports if any.
* Involving with different teams and back tracking the flows and experience in solving the critical issues.
* Troubleshooting production issues under client defined SLA’s.
* Having experience in creating the Priority Incidents, Change Requests and Service Requests in Service Now. Experience in creating the Jira’s.

**Environment:** GCP, Cloud SQL, Big Query, Cloud Dataproc, GCS, Cloud SQL, Cloud Composer, Informatica Power Center 10.1, Talend 6.4 for Big Data, Hadoop, Hive, Teradata, SAS, Teradata, Spark, Python, Java, SQL Server, Service Now, Confluence.

**Client:** ADP, Alpharetta, GA **Nov 19 – Apr 20**

**Role:** Data Engineer/GCP

**Responsibilities**

* Worked with ecommerce, Sales & Ops, Manufacturing, customer support team to implement projects like Data warehouse, Data Engineering**,** Data integration automation**,** process design, API enablement, Analytics, Data quality etc.
* Developed data pipelines to implement enterprise data warehouse in **Google cloud.**
* Developed ingestion layer in google data storage for manufacturing team to process daily 200GB data.
* Developed **Pyspark** scripts using Data frames/Spark SQL and RDD in Spark for Data Aggregation, queries.
* Created various Google Cloud Terraform modules for reusability into different projects, some of the modules include Compute Engine, Compute Templates, Cloud SQL, Big Query, VPC, Pub/Sub, External, and Internal Load Balancers.
* Deployed Cloud Functions to trigger services like Pub/Sub, Cloud Storage, and Data store for real-time processing of messages on Pub/Sub or process files on Cloud Storage when a Compute Engine modifies or writes to it and uploads the data and messages to Data store.
* Made use of Cloud Functions with **Python** to update the Instance Templates and Managed Instance Groups (MIG) to accommodate the new image versions and patching support for the existing Compute Engine Instances.
* Configured Server less VPC Access Connector for server less environments (Cloud Functions and App Engine) to access an Internal Load Balancer which has backend service on a Managed Instance Group without exposing to the public internet.
* Created a workflow to move Cloud Asset Inventory to **Big Query** by creating cron jobs for scheduled exports using App Engine and importing into Big Query using Dataflow templates to better understand the utilization of Google resources.
* Created Organization, Project-wide IAM policies, and Firewall rules on **GCP** to restrict unwanted access from public internet and individuals.
* Managed GKE and Compute Engine version upgrades and patching regularly to keep the infrastructure up to date and maintain the applications without downtime during regular hours.
* Created Kubernetes YAML configuration files to deploy Services, Ingress for the pods, and created Gitlab CI templates to deploy APIs on the **Kubernetes** Clusters.
* Improved daily jobs performance using data cleaning, query optimization and table partitioning.
* Experience in Extraction, Transformation and Loading of Data from JSON, XML, Flat Files, Excel, SQL Server and API endpoint.
* Created a process to extract email attachments and send required information from Big Query.

**Environment**: GCP (Cloud Functions, Compute Engine, Dataproc, Dataflow, Pub/Sub, Cloud SQL, Big table, Big Query, GKE), Kubernetes, Docker, GitLab CI, Gitlab SAST, Terraform, Stack Driver, Oracle, XML files, Flat files, JSON, Hadoop.

**Client:** Avon Technologies (India) Pvt. Ltd**. Aug 16 – July 19**

**Role:** Data Engineer

**Responsibilities:**

* Profiling huge volumes of data to identify the patterns the structure of data, helping in gaining quick insights and working back and forth with the team to extract the required data.
* Created custom packages to help work with the data present in the databases by secured authentication methods, easy to interact with various databases.
* Experienced in managing large datasets using parallel processing engines/systems.
* Demonstrated expertise utilizing ETL tools, including **Informatica** and **ETL** package design, and RDBMS systems like SQL Server, Oracle.
* Executed Python scripts to automate validation of data which improved validation speed and accuracy by 35%.
* Implemented best methodologies and practices that enhanced product definition, release processes and customization of applications as per the client requirements.
* Develop the ETL mappings for XML, .CSV, .TXT sources and loading the data from these sources into relational tables with **Talend ETL** Developed Joblets for reusability and to improve performance.
* Optimizing the report generation from various data sources including but not limited to MySQL, PostgreSQL, redshift reduced the time taken to take key decisions by various teams by 25%.
* Optimized and automated **ETL** processes for better streamlining
* Involved in SDLC requirements gathering, analysis, design, development, and testing of application using **Agile** (**SCRUM**) methodologies.
* Actively participated in UAT with business users to ensure adherence of system with business requirements.

**Environment:** Python, ETL, SQL, MySQL, Google, XML files, Flat files, JSON, Hadoop, Oracle.

**Other Projects:**

**Hostel Stack:**

* Developed an application where people can even use it in offline. We have mainly concentrated on engineering students’ group, which helps them to find the hostels near their college. We even provided the option to choose the hostel in their desired location. This project was implemented using Android SDK, Eclipse Ganymede IDE.

**ATM Card Fraud Detection Machine:**

* Created a Machine Model that classifies legit and fraud accounts upon providing transactional data.
* The Machine model was trained, tested with a Kaggle data set of 284,807 bank transactions containing data of European cardholders.
* The project was implemented using R language and R studio and provided a web interface using Shiny package in R.
* The project also specifies the brief description about the accuracies and error detection rate of various Machine Learning algorithms.

**Better Reviews:**

* Better reviews are a web application useful for different service-based and product-based organizations. These organizations can register with the application, and they can post the updates of the products and services that they are providing.
* Updates like product launches, home services, event management, restaurants, etc. People can check services can visit the application and they can search for the service providers.
* A few service providers will be available, and people can select any of the service products and can request the quote for the service they require.
* The system will provide the same information to the service providers and then the customer will be getting the quote.

**Education:**

**Bachelor’s Degree in Computer Science:** Jawaharlal Nehru Technological University, India.

**Master’s Degree in Computer Science:** Governor’s State University, IL, USA.