**DIVYASRI PUSALA**

**Mobile Number: +1 (603)-318 9075**

**E Mail ID: pusala66@gmail.com**

**Summary:**

* 10+ years of experience in IT, extensive experience in **AWS (Amazon Web Services) Cloud** and **GCP (Google Cloud Platform)**.
* Experience with a solid understanding of Business Requirement Gathering, Business Process Flow, Business Process Modeling, and Business Analysis.
* Expertise in Data Integrity, Data Quality, Business Intelligence, Database Management, and Data Warehousing across the entire project life cycle.
* Worked extensively with **Apache Hadoop ecosystems** like **HDFS**, **Spark**, **Flume**, **Kafka**, **Airflow,** and **Hive**.
* Built batch and streaming pipelines using AWS services like **S3, RDS, RedShift, Lambda, Glue, EMR, Ec2, Athena, Step Functions, CloudWatch, SageMaker, IAM roles, QuickSight,** and many other services.
* Similarly used **GCP** services like **Cloud storage**, **BigQuery**, **Cloud Composer**, **PubSub**, **Cloud Monitoring**, **Cloud Functions, Cloud Compute Engine Data Proc,** and **Power BI.**
* Used Storage services like **HDFS, AWS S3,** and **GCP Cloud Storage**.
* Extensive experience in Data Warehouse services like **Amazon Redshift**, **Google BigQuery,** and **Snowflake**.
* Used orchestration tools like **Apache Airflow** and **Step Function**.
* Used container services like **Kubernetes** and **Docker**.
* Have knowledge of both the technical and functional aspects of the projects. Well-versed in handling SQL Database and writing SQL queries to test the data.
* Experience in Data Warehouse development working with **Data Migration**, Data Conversion, and **ETL** using **Microsoft SQL Server** Integration Services and **SQL Server**.
* Extensive experience in developing complex data extract, applications, and ad-hoc queries as requested by internal and external customers using **SQL**.
* Experience with dashboard/report design with **Tableau**, **Power BI,** and **QuickSight**.
* Hands-on experience in Unix shell scripting.
* Maintain the production of Business Intelligence products and solutions, including resolving production issues and returning quickly to priority problems.
* Troubleshoot and resolve data issues impacting extract delivery.
* Created different interactive views and reporting Dashboards by combining multiple views using **Tableau Desktop**.
* Designing and developing a prototype for reports and gaining approval from the client for further development. Created Parameterized **Crystal** Reports with conditional formatting and used sub-reports.
* Excellent interdepartmental communication: teamwork and coordination with senior-level executives to meet and exceed project goals; efficient operations under time-sensitive and fluctuating timelines.

**Skill Matrix:**

|  |  |
| --- | --- |
| Big Data/Hadoop Ecosystem | HDFS, MapReduce, HIVE, PIG, Sqoop, Flume, Oozie, Spark, Kafka, Airflow. |
| Opensource libraries | Scikit-learn, NumPy, SciPy, OpenCV, Deep learning, NLP, Keras, Matplotlib (for visualization) |
| NoSQL Databases | MongoDB, Cassandra, HBase |
| Data Analysis Skills | Data Cleaning, Data Visualization, Feature Selection, Pandas |
| Programming Languages | Python, SQL, Scala, Python, PL/SQL, Linux shell scripts. |
| Database | Oracle 11g/10g, DB2, Microsoft SQL, MySQL, Teradata |
| Cloud Ecosystem | AWS and GCP |
| Automation and scheduling | Crontab, Airflow, Google Cloud Composer, AWS Step Function |
| Reporting tools | AWS QuickSight Tableau, POWER BI. |
| DataLake | AWS S3, Apache HDFS, Google Cloud Storage, |
| Data Warehouse | AWS Redshift, Snowflake, Google BigQuery |

**PROFESSIONAL** **EXPERIENCE:**

|  |
| --- |
| **Client: Fonteva, Arlington, VA MAY 2021 – TILL DATE**  **Role: Big Data Engineer** |

**Responsibilities:**

* Collaborating with technical, application, and security leads to delivering reliable and secure Big Data infrastructure tools using cutting-edge technologies like Spark, Container services, and AWS services.
* Developed data processing pipelines in Spark and other big data technologies.
* Designed and deployed high-performance systems with reliable monitoring, logging practices, and dashboards.
* Worked with information Security teams to create data policies and develop interfaces and retention models and deployed the solution to production.
* Designed, Architected, and Developed solutions leveraging big data technology **(Open Source, AWS)** to ingest, process, and analyze large, disparate data sets to exceed business requirements.
* Used AWS services like **AWS S3, AWS RDS, AWS Redshift, AWS Athena, AWS Lambda, AWS Ec2, AWS EMR, AWS IAM, AWS Step Functions, AWS CloudWatch, AWS Glue, AWS QuickSight, AWS EKS, etc.,**
* Used Orchestration services like **Apache Airflow** and created several dags as part of batch and streaming pipelines.
* Developed **Python** and **PySpark** code for **ETL** and used them in **AWS** services like **AWS Lambda, Glue, Ec2, EMR**, etc.
* Used **AWS Lambda, AWS Glue, AWS Athena, AWS S3, AWS RedShift,** and **AWS EMR** for Batch and streaming pipelines.
* Designed and developed the **Data Warehouse in Snowflake.**
* Connected both batch and streaming pipelines to **Snowflake** using **Kafka**.
* Created a POC and MVP using **DBT** as a transformation layer to load data into Data Warehouse.
* Used Terraform to deploy **AWS services** and developed Terraform code for modules and resources.
* Implemented Continuous Integration Continuous Delivery **CICD**.
* Developed POC and MVP using GCP (Google Cloud Services).

**Technologies:**

|  |
| --- |
| AWS S3, Redshift, Athena, RDS, Lambda, Glue, EMR, API Gateway, Step Function, Cloud Watch Events, Cloud Watch Trails, SNS, AWS Boto3 SDK, Confluent Kafka, Spark, PySpark, SparkSQL, Terraform, Vault, Snowflake, etc., |

|  |
| --- |
| **Client: Discover Financial Services, Chicago, IL MAY 2019 – APRIL 2021** **Role: Data Engineer** |

**Responsibilities:**

* Built Real-time and batch pipelines using **GCP services.**
* Modeled and developed a new data warehouse from scratch and then migrated that data warehouse to **BigQuery**. Automated data pipelines and quality control checks.
* Architected and implemented a solution to migrate the data platform from **Hadoop** to **Google cloud Platform.**
* Worked on developing continuous integration (CI), continuous delivery (CD), and continuous training (CT) for the ML system using **Cloud Build.**
* Worked heavily with the business to determine reporting requirements and to explore the data with each other.
* Used Cloud functions, Cloud **pub-sub** and **BigQuery** to build streamline dashboard to monitor services.
* Created jobs using **Cloud composer** (**Airflow DAG**) to migrate data from **Data Lake** (cloud storage) to transform it using DataProc and ingest in **BigQuery** for further analysis.
* Member of the Enterprise Architecture Committee and presented multiple times to this group and other business leaders.
* Built batch and streaming jobs using **GCP** Services like **BigQuery**, **Pub/Sub**, **DataProc**, **Dataflow**, **Cloud Run**, **Compute engine** and **Cloud Composer**.
* Helped to implement **PowerBI** in the organization. Developed PowerBI reports for every department in the company.
* Developed **PySpark** scripts for ETL using **DataProc**.
* Good experience with **GCP Stack driver** Trace, Profiler, Logging and Error Reporting, Monitoring
* Setup and Implement Continuous Integration and Continuous Delivery (CI & CD) Process stack using   **GIT**, **Jenkins**
* Experience in developing APIs in cloud.
* Good Knowledge on **IAM** rules and **Cloud Security.**
* Designed **PowerBI** data visualizations using cross tabs, maps, scatter plots, pie charts etc.

**Technologies:**

|  |
| --- |
| Google Cloud Storage, DataFlow, DataProc, Cloud function, Kubernetes, Big Query, Cloud run, Cloud Composer, Airflow, Hive, Docker, Hive, HDFS, Compute Engine, etc., |

|  |
| --- |
| **Client: Chase Bank, New York City, NY JUN 2017 - APR 2019**  **Role: Data Engineer** |

**Responsibilities:**

* Collaborated with business, analytical teams, and data scientists to improve efficiency, increase the applicability of predictive models, and help translate ad-hoc analyses into scalable data delivery solutions.
* Collaborated with the DevOps team to integrate innovations and algorithms into a production system.
* Worked with the DevOps team to create and manage deployment workflows for all scripts and code.
* Develop and maintain scalable data pipelines that will ingest, transform, and distribute data streams and batches within the **AWS S3** and **Snowflake** using **AWS Step Function, AWS Lambda, AWS Kinesis, AWS Glue, and AWS EMR**.
* Created batch pipelines using **AWS S3, AWS Lambda, AWS Glue, AWS EMR, AWS Athena, AWS RedShift, AWS RDS,** etc.,
* Used **AWS** services like **AWS S3, AWS RDS,** and **AWS** **Redshift** for storage.
* Created **Apache Airflow** dags to ingest data from sources like APIs, Servers, and Databases to transform using **PySpark** in **Glue** and **EMR** and loaded data into Data Warehouse like **AWS RedShift**.
* Created Data pipelines using **AWS services** like **S3**, **Glue**, **EMR**, **Lambda**, **Athena**, **IAM**, etc.,
* Created reports and dashboards that provide information on metrics, usage, trends, and behaviors using **AWS** **services** like **S3**, **Lambda**, **Athena,** and **QuickSight**.
* Orchestrated pipelines and dataflow using **Apache Airflow** and **Step Function**.
* Created reports and dashboards using **AWS services** like **Lambda**, **Glue**, **Step Function,** and **QuickSight**.
* Created monitoring service using **AWS CloudWatch**, **AWS Lambda**, **AWS Glue**, **AWS Step Function, Grafana,** and **ElasticSearch**.
* Created Airflow dags to extract, transform and load data into Data Warehouse.
* Developed and deployed **Kubernetes** pods to extract, transform, and load data.
* Used **Docker** and **Kubernetes** for Data Pipelines and ETL Pipelines.
* Used **Hadoop** ecosystems like **Apache HDFS, Apache Spark, Apache Hive, Apache Airflow,** and **Apache Kafka**
* Facilitated the developed and deployed proof-of-concept machine learning systems.

**Technologies:**

|  |
| --- |
| Apache Hadoop, Hive, Kafka, AWS S3, Redshift, Athena, RDS, Lambda, Glue, EMR, EC2, API Gateway, StepFunction, Cloud Watch Events, Cloud Watch Trails, SNS, AWS Boto3 SDK, Spark, PySpark, SparkSQL, Terraform, Vault, etc., |

|  |
| --- |
| **Client: MYNTRA, BGR, IND JAN 2015 – NOV 2016**  **Role: Jr Data Engineer** |

**Responsibilities:**

* Built Real-time and batch pipelines using **GCP services.**
* Modeled and developed a new data warehouse from scratch and then migrated that data warehouse to **BigQuery**.  Automated data pipelines and quality control checks.
* Architected and implemented a solution to migrate the data platform from **Hadoop** to **Google Cloud Platform.**
* Worked on developing continuous integration (CI), continuous delivery (CD), and continuous training (CT) for the ML system using **Cloud Build.**
* Worked heavily with the business to determine reporting requirements and explore the data with each other.
* Used Cloud functions, Cloud **pub-sub,** and **BigQuery** to build a streamlined dashboard to monitor services.
* Created jobs using **Cloud Composer** (**Airflow DAG**) to migrate data from **Data Lake** (cloud storage) to transform it using DataProc and ingest it in **BigQuery** for further analysis.
* Member of the Enterprise Architecture Committee and presented multiple times to this group and other business leaders.
* Built batch and streaming jobs using **GCP** Services like **BigQuery**, **Pub/Sub**, **DataProc**, **Dataflow**, **Cloud Run**, **Compute Engine,** and **Cloud Composer**.
* Helped to implement **PowerBI** in the organization.  Developed PowerBI reports for every department in the company.
* Developed **PySpark** scripts for ETL using **DataProc**.
* Good experience with **GCP Stack driver** Trace, Profiler, Logging and Error Reporting, Monitoring
* Setup and Implement Continuous Integration and Continuous Delivery (CI & CD) Process stack using   **GIT**, **Jenkins**
* Experience in developing APIs in the cloud.
* Good Knowledge of **IAM** rules and **Cloud Security.**
* Designed **PowerBI** data visualizations using cross tabs, maps, scatter plots, pie charts, etc.

**Technologies:**

|  |
| --- |
| Google Cloud Storage, DataFlow, DataProc, Cloud function, Kubernetes, Big Query, Cloud run, Cloud Composer, Airflow, Hive, Docker, Hive, HDFS, Compute Engine, etc., |

|  |
| --- |
| **Client: United Healthcare, Hyderabad, India JUNE 2012 – DEC 2014**  **Role: Hadoop Developer** |

**Responsibilities:**

* Installed, Configured, and Maintained Apache Hadoop clusters for application development and **Hadoop** tools like **Hive, Pig, Zookeeper, Kafka,** and **Sqoop**.
* Integrated HDP clusters with Active Directory and enabled Kerberos for Authentication.
* Installed and Configured **Sqoop** to import and export the data into Hive from Relational databases.
* Administering large **Hadoop** environments building and supporting cluster setup, performance tuning, and monitoring in an enterprise environment.
* Monitoring the **Hadoop** cluster functioning through **MCS** and working on NoSQL databases including **HBase**.
* Used Hive and created **Hive** tables and was involved in data loading and writing Hive UDFs and worked with the **Linux** server admin team in administering the server hardware and operating system.
* Configured **Spark Streaming** to receive real-time data from **Kafka** and store the stream data to **HDFS**.
* Designed and Developed data mapping procedures ETL-Data Extraction, Data Analysis, and Loading process for integrating data using R programming.
* Implemented Partitioning, **Dynamic Partitions**, and **Buckets** in **HIVE**.
* Developed storytelling dashboards in **Tableau** Desktop and published them on **Tableau** **Server** which allowed end users to understand the data on the fly with the usage of quick filters for on-demand needed information.
* Close monitoring and analysis of the **MapReduce** job executions on the cluster at the task level and optimized **Hadoop** cluster components to achieve high performance.

**Technologies:**

|  |
| --- |
| Apache Hadoop, HDFS, Kafka, Flume, Airflow, Hive, Docker, Hive, MapReduce, SQL, Python, Spark, SparkSQL, etc., |