**Jayanth Gundagoni**

**jayanthgundagoni8118@gmail.com**

**312-767-4254**

**Sr. Data Engineer**

|  |
| --- |
| **PROFESSIONAL SUMMARY:** |
| * Over 9+ years of experience as a Sr. Data Engineer, adept at designing and implementing robust data solutions.
* Proficient in programming languages, including Python, Scala, R, SQL, and Bash, with a strong data manipulation and analysis foundation.
* Extensive expertise in cloud platforms such as AWS, Azure, and GCP, and skilled in managing data warehouses like Snowflake, Google BigQuery, and RedShift.
* Specialized in big data technologies, including Spark, PySpark, Hadoop, Hive, Pig, Sqoop, and Impala, for efficient data processing and analytics.
* Experienced in working with various databases such as SQL Server, CosmosDB, PostgreSQL, Oracle, Aurora, and DynamoDB.
* Well-versed in ETL tools like Informatica, Nifi, ADF, SSIS, and AWS Glue for seamless data extraction, transformation, and loading.
* Proficient in data processing and analytics libraries such as Pandas, NumPy, PyTorch, Scikit-Learn, TensorFlow, and Matplotlib.
* Skilled in workflow orchestration using AWS Step Functions and Apache Oozie, ensuring smooth data pipeline execution.
* Expertise in data modeling and visualization tools, including Erwin, QuickSight, Dataiku, Power BI, and Looker.
* Hands-on experience with containerization and orchestration tools like Openshift, Docker, Kubernetes, ECS, and EKS.
* Familiarity with infrastructure such as code through AWS CloudFormation and Terraform for efficient deployment and management.
* Proficient in using integrated development environments (IDEs) such as Eclipse, VS Code, and IntelliJ.
* Adept at monitoring and logging using CloudWatch, Prometheus, Splunk, Grafana, and Azure Monitor for proactive system management.
* Strong focus on security and identity management with expertise in IAM, KMS, JWT, and Azure AD.
* Well-versed in version control and collaboration tools like GitHub, GIT, Bitbucket, and JIRA.
* Experienced in project management methodologies, including Agile, Scrum, and Kanban.
* Demonstrated capability to collaborate effectively with cross-functional teams, promoting transparent communication and exchanging insights to improve overall project efficiency.
* Strong interpersonal skills, facilitating productive collaboration with stakeholders, data scientists, and other engineering teams to achieve common data-driven goals.
* Experience in changing over existing AWS infrastructure to Server less architecture (AWS Lambda, AWS Kinesis) through the
* creation of a Server less Architecture using AWS Lambda, API gateway, Route 53, S3 buckets
* Hands on experience in migrating on premise ETLs to Google Cloud Platform (GCP) using cloud native tools such as BIG query, Cloud Data Proc, Google Cloud Storage, Compose.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Worked parallelly in both GCP and AWS Clouds coherently.
* Experience in GCP Dataproc, GCS, Cloud functions, Big Query, Azure Data Factory Data Bricks.
* Experience in building efficient pipelines for moving data between GCP and Azure using Azure Data Factory.
* Experience in building power bi reports on Azure Analysis services for better performance when comparing that to direct query using GCP Big Query.
* Extensive use of cloud shell SDK in GCP to configure/deploy the services Data Proc, Storage, and Big Query
 |

**Technical Skills:**

| Big Data Technologies | Hadoop, MapReduce, Spark, HDFS, Sqoop, YARN, Oozie, Hive, Impala, Zookeeper, Apache Flume, Apache Airflow, Cloudera, HBase |
| --- | --- |
| Programming Languages | Python, PL/SQL, SQL, Scala, C, C#, C++, T-SQL, PowerShell Scripting, JavaScript, Java. |
| Cloud Services | AWS, GCP, Azure. |
| Databases | MySQL, SQL Server, Oracle, MS Access, Teradata, and Snowflake. |
| NoSQL DataBases | MongoDB, Cassandra DB, HBase. |
| Monitoring tool | Apache Airflow. |
| Visualization & ETL tools | Tableau, Informatica, Talend, SSIS, and SSRS. |
| Version Control & Containerization tools | Jenkins, Git, and SVN. |
| Operating Systems | Unix, Linux, Windows, Mac OS. |

**PROFESSIONAL EXPERIENCE:**

**Client: Apple, New York May 2023 – Tll Date**

**Role:**  **Sr. GCP Data Engineer**

**Responsibilities**:

* Participated in sprint planning sessions to define sprint goals, prioritize user stories, and estimate the effort required for data engineering tasks. Managed and groomed the Agile backlog, ensuring that it contained well-defined and prioritized data engineering tasks.
* Proficient in designing, implementing, and maintaining data solutions using various GCP services, including Big Query, Dataflow, Data prep, Pub/Sub, and Storage (e.g., Cloud Storage, Bigtable).
* Developed real-time data processing solutions using Pub/Sub and Dataflow for immediate data ingestion and analysis.
* Strong proficiency in Python for scripting, automation, and data manipulation tasks within the GCP environment.
* Developed custom Python scripts to automate data pipeline orchestration, data quality checks, and monitoring.
* Proficient in writing complex SQL queries to retrieve, manipulate, and analyze data stored in GCP databases and data warehouses.
* Developed optimized SQL scripts for data extraction, transformation, and loading (ETL) processes within GCP services like Big Query.
* Utilized Python libraries such as pandas, NumPy, and SciPy for data analysis, data transformation, and statistical computations.
* Developed and maintained data ingestion and transformation workflows using Python and PySpark.
* Designed and implemented complex data pipelines in PySpark to extract, transform, and load (ETL) data from various sources into GCP data lakes and data warehouses.
* Leveraged PySpark's distributed computing capabilities to process large volumes of data in parallel, ensuring efficient data processing.
* Extensive experience with Google Cloud Composer and Apache Airflow for orchestrating complex data pipelines and ETL processes on GCP.
* Designed and developed data pipelines to efficiently extract, transform, and load (ETL) data from diverse sources into GCP.
* Managed data storage using Google Cloud Storage, including object storage, bucket configuration, and data lifecycle management.
* Implemented data archiving and tiering strategies to optimize storage costs.
* Implemented data quality checks and data validation processes to maintain data accuracy and consistency across GCP services.
* Utilized Dataflow for stream and batch data processing, enabling real-time analytics and data transformation.
* Utilized GCP services such as Bigtable and Datastore for NoSQL data storage and processing.
* Implemented scalable data processing solutions with GCP Dataflow for analyzing and transforming large datasets.
* Proficient in using Git for version control, collaborating with cross-functional teams, and managing codebase changes in data engineering projects.
* Implemented data governance practices in GCP, including data lineage tracking, metadata management, and audit trails for compliance.
* Ensured data security and privacy compliance by implementing encryption, access control, and data masking techniques across GCP services.
* Implemented data warehousing solutions in GCP, utilizing BigQuery for scalable and cost-effective storage and querying of large datasets.
* Proficient in deploying and managing applications and data solutions in Google Kubernetes Engine (GKE) for containerized workloads.
* Experience in collaborating with cross-functional teams, including data scientists, analysts, and business stakeholders, to understand data requirements and deliver data engineering solutions that meet business needs.
* Proven track record of successfully delivering data engineering projects on GCP, meeting project timelines, quality standards, and customer expectations.

**Environment:** GCP, Python, PySpark, SQL, Big Query, Dataflow, Google storage buckets, ETL, Airflow, Composure, Pub/Sub, Pandas.

**Client: Credit Suisse, Morrisville, NC March 2021 to April 2023**

**Role: Senior Data Engineer**

**Responsibilities:**

* Installed, configured, and maintained Data Pipelines, developed Data Pipeline with Kafka and Spark.
* Developed processes for loading data into Snowflake, designed data modeling, and optimized views for Tableau reporting.
* Implemented Azure cloud solutions using HDInsight, Azure Databricks, Event Hubs, Cosmos DB, cognitive services, KeyVault, and Treasure Data (CDP).
* Led a team of data engineers in designing and implementing a real-time data pipeline, reducing data processing time.
* Authored Python (PySpark) Scripts for custom UDFs, data labeling, and cleaning tasks.
* Evaluated Snowflake design considerations, redesigned views to increase performance, and conducted unit testing.
* Proficient in the analysis, design, and development of solutions utilizing Microsoft Azure technology stack.
* Developed and deployed multi-source data ingestion pipelines using Azure Databricks, with a focus on performance and scale.
* Developed a data warehouse model with Azure Databricks for over 100 datasets.
* Managed SAP HANA security activities, loading data into SAP HANA Database using SLT, BODS, and DXC.
* Developed data warehouse model in Snowflake for over 100 datasets using WhereScape, created reports in Looker based on Snowflake connections.
* Leveraged ETL tools such as Talend Open Studio, SAS Data Management, and Python for process improvements.
* Worked on Apache Hadoop components, scheduled jobs using Airflow scripts, connected Tableau for interactive reports.
* Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation, and aggregation.
* Conducted tuning for SQL and PL procedures, Informatica objects, and views.
* Used AWS services like EC2 and S3 for data processing and storage, maintained Hadoop cluster on AWS EMR.
* Experienced in branching, tagging, and maintaining version control using GIT, SVN, and TFS.
* Administered production, development, and test environments on various platforms (Windows, Ubuntu, Red Hat Linux, SUSE Linux, CentOS, Solaris).
* Designed and developed ETL workflows and datasets in Alteryx for BI reporting tools, specifically for Tableau Data Extract.
* Worked on Dimensional and Relational Data Modeling using Star and Snowflake Schemas, OLTP/OLAP.
* Developed automation regressing scripts for ETL process validation between multiple databases using Python.

**Environment:** AWS, EC2, ETL, Pyspark, Snowflake, Kafka, Spark, Lambda, Talend Open Studio, SAS Data Management, Hadoop, Tableau, Python, Hive, SQL, Oracle, Alteryx, Informatica, scheduling tool, Shell scripting, Azure Databricks, Linux, Treasure Data (CDP).

**Client: Anheuser-Busch, St Louis, MO Jan 2019 – Feb 2021**

**Role: Senior Data Engineer**

**Responsibilities:**

* Utilize Azure SQL Database for relational data storage and processing.
* Collaborating with the SessionM platform to capture and analyze customer interactions and behaviors within the loyalty program. Integrating SessionM data into the data pipeline for further processing and analysis.
* Setting up and maintaining the Treasure Data (Customer Data Platform) to consolidate customer data from various sources, allowing for a unified view of customer information. Ensuring data is synced and available for analysis in real-time.
* Excelled in Teradata with Azure, contributing to the seamless integration of data engineering processes.
* Shared data securely across within the organization using Azure Data Share by simplifing data sharing across Azure subscriptions.
* Applied advanced techniques in PySpark to optimize data retrieval operations, ensuring streamlined processing and enhanced performance of the data pipeline.
* Implemented PySpark optimizations to fine-tune queries and operations, resulting in reduced processing time and improved efficiency in generating insights and reports.
* Possess valuable experience in the health industry, providing insights into specific sector nuances.
* Applied knowledge of Azure technologies to optimize data workflows and enhance overall efficiency.
* My project involved using MongoDB, I am responsibilities may include designing the database schema, optimizing data retrieval operations, and ensuring data security and privacy.
* Implementing data quality checks and validation mechanisms to ensure the accuracy and reliability of the data used for analysis. Ensuring compliance with data governance policies and best practices.
* Tuning and optimizing the data pipeline and database queries to enhance performance and reduce processing time, allowing for faster insights and reporting.
* Mentored and Lead junior data engineers, fostering their technical growth and career development.
* Setting up monitoring and logging systems to track data pipeline performance and troubleshoot issues proactively. Identifying bottlenecks and addressing them to maintain data availability and reliability.
* Working closely with data analysts and other stakeholders to understand their data requirements and providing the necessary data support. Documenting the data engineering processes, data schemas, and architecture for easy comprehension and future reference.
* Lead the development team in designing, building, and maintaining data pipelines and infrastructure.
* Expertise in using Snowflake features such as SQL, stored procedures, and functions.
* Working with stakeholders to prioritize data requests based on business impact and urgency. This ensures that critical requests are addressed first while managing resources effectively.
* Implement Azure DevOps for continuous integration and continuous delivery (CI/CD) pipelines, enabling automated testing and deployment of data engineering solutions.

**Environment**: Azure DevOps, Azure SQL Database, Azure Databricks, Azure Data Share, Azure Data Factory, SessionM, Treasure Data (CDP) and MongoDB

**Client: Evoke Technologies, Hyderabad, India Aug 2016 – Nov 2018**

**Role: AWS Data Engineer**

**Responsibilities:**

* Created Lambda functions and assigned IAM roles to run **Python** scripts with a variety of triggers (SQS, Event Bridge, SNS)
* Involved in the development of multiple applications using AWS Stack **(S3, EMR, Redshift, Glue, CloudWatch, and SNS**), with a focus on high availability, fault tolerance, and auto-scaling.
* Developed a Python script to connect to **REST APIs** and extract data to **AWS S3**.
* Created a Python script to move data from on-premises to AWS S3.
* Transfer data from on-premises to AWS storage buckets.
* Designed and executed a migration strategy to move the Data Warehouse from an Oracle platform to AWS Redshift.
* Implemented POC’s to replace EMQ broker with **RabbitMQ** broker with **MQTT** plug-in enabled and deployed to the environments with Network load balancer.
* Kafka is used for log accumulation, such as gathering physical log documents from servers and storing them in a central location such as **HDFS** for processing.
* Configured **Spark** Streaming to receive continuous data from **Kafka** and store the stream data in **HDFS.**
* Proficient in AWS services such as VPC, Glue Pipelines, Glue Crawler, Cloud front, **EC2, ECS, EKS, Elastic bean stalk, Lambda, S3, Storage gateway, RDBS, Dynamo db, Redshift, Elastic Cache, DMS, SMS, Data Pipeline, IAM, WAF, Artifacts, API gateway, SNS, SQS, SES,** Auto Scaling, Cloud Formation, Cloud Watch and cloud Trail.
* worked on optimizing volumes and **EC2** instances, as well as setting up multiple VPC instances. Elastic Beanstalk was used to deploy applications on AWS, and **Route53** was implemented and configured for AWS Web Instances.
* Hands on experience with Requests, **NumPy, Matplotlib, SciPy, PySpark, and Pandas Python** libraries during the development lifecycle, as well as experience developing APIs for the application using **Python, Django, MongoDB, Express, ReactJS, and NodeJS**.
* Knowledge about setting up **Python REST API Framework** using **Django.**
* worked with a variety of databases, including **Oracle**, **SQL Server**, **Teradata,** and **Cassandra**.
* Experience in Performance Tuning and Debugging of existing ETL processes.

**Environment:** Python, EC2, S3, IAM, EMR, Redshift, Glue, CloudWatch, SNS, AWS S3, Oracle, Kafka, HDFS, Lambda, NumPy, PySpark, Pandas Python, Django, MongoDB, Express, ReactJS, NodeJS, Teradata, Cassandra.

**Client: HITACHI- Hyderabad, India Feb 2014 – July 2016**

**Role: AWS Data Engineer**

**Responsibilities:**

* Migrated data from on-premises to **AWS S3 storage buckets.**
* Developed python scripts to hit REST API’s and extract data to AWS S3.
* Created **YAML** files for each data source and including glue table stack creation.
* Developed Lambda functions and assigned IAM roles to run **python** scripts along with various **triggers** (SQS, Event Bridge, SNS).
* Create external tables with partitions using **Hive, AWS Athena**, and **Redshift.**
* Developed Spark Programs using Scala and Java **API's** and performed transformations and actions on RDD's.
* Document the complete process flow to describe program development, logic, testing, and implementation, application integration, coding.
* Performed Fit Gap analysis for all the data migration requirements.
* Analyze and define researcher's strategy and determine system architecture and requirement to achieve goals and developed multiple **Kafka** Producers and Consumers from as per the software requirement specifications.
* Configured Spark Streaming to get ongoing information from Kafka and store the stream information to HDFS.
* Created data models in Splunk using pivot tables by analyzing vast amounts of data and extracting key information to suit various business requirements.
* Involved in writing custom Map-Reduce programs using java API for data processing.
* Worked on a python script to extract data from **Netezza** databases and transfer it to **AWS S3.**
* Used Kafka for log accumulation like gathering physical log documents off servers and places them in a focal spot like **HDFS** for handling.
* Worked on Ingesting data by going through cleansing and transformations and leveraging AWS **Lambda**, AWS Glue and Step Functions.
* Developed and executed a migration strategy to move Data Warehouse from an Oracle platform to AWS Redshift.

**Environment:** AWS S3, AWS Athena YAML, python, Redshift, Hive, Kafka, Lambda, Netezza, Redshift, Redshift, Java API, Oracle