Sravani Gajula

Data Engineer

Mail:sravanigajula04@gmail.com

**Professional Summary**

* 10+ years of experience as a Data Engineer, and experienced in Design, Development, and Implementation of Django, Flask web framework sand client-server technologies-based applications, RESTFUL services, AWS, C, and SQL.
* Skilled in Python with proven expertise in using new tools and libraries like NumPy, SciPy, matplotlib, PyTest, python-twitter, Pandas, requests, urllib2, etc.
* Good experience in developing web applications implementing Model View Control architecture using Django and Flask web application frameworks.
* Experience in containerizing applications using Docker to ensure consistency across development, testing, and production environments.
* Proficiency in using Docker Compose to define and manage multi-container Docker applications, particularly for complex data processing workflows.
* Experience with Kubernetes as a container orchestration tool to automate the deployment, scaling, and operation of application containers across clusters of hosts.
* Experience in integrating Apache Kafka with other components of the data ecosystem, such as Apache Spark, Hadoop/HDFS, databases (e.g., MySQL, PostgreSQL), and messaging systems (e.g., RabbitMQ), to facilitate data ingestion, transformation, and streaming analytics.
* Knowledge of Kafka's fault tolerance mechanisms, such as replication, retention policies, and data recovery strategies, to ensure data durability and availability.
* Experienced in working with various Python Integrated Development Environments like PyCharm, Spyder, PyStudio, PyDev, and Sublime.
* Knowledge of big data technologies like Hadoop, Spark/pySpark, etc.
* Extensively worked on Spark Streaming and Apache Kafka to fetch live stream data.
* Strong expertise in the development of web-based applications using Python, Django, HTML, XML, Angular JS, CSS, REST APIs, JavaScript, JSON, and jQuery.
* Experience in Agile Methodologies, Scrum stories, and sprints experience in a Python-based environment, along with data analytics, data wrangling, and Excel data extracts.
* Experienced with setup, configuration, and maintenance of stack (Elasticsearch, Logstash, and Kibana).
* Strongly follow PEP-8 coding standard and test a program by running it across test cases to ensure validity and effectiveness of code using PyChecker and Pylint.
* Experience with both relational (SQL) and non-relational (NoSQL) databases like MongoDB, Redis, Oracle, MySQL, SQLite, PostgreSQL, Greenplum, etc. using ORM and ODM.
* Hands-on Experience in AWS EC2, S3, Redshift, EMR, RDS, S3, Lambda, AMI, VPC, Elastic IP Addresses, Load Balancer. Cloud system design, implementation, and troubleshooting (AWS, Rackspace, Google, AZURE).
* Experience in using various version control systems like CVS, Git, GitHub, and deployment using Heroku.
* Experience in integrating Apache Solr with data pipelines and ETL processes to ingest and index data from various sources, such as databases (e.g., MySQL, PostgreSQL), log files, and streaming platforms (e.g., Kafka), ensuring data consistency and relevance in search results.
* Proficiency in designing efficient data models for Aerospike databases, considering data access patterns, query performance, and scalability requirements.
* Experience in administering Aerospike NoSQL databases, including installation, configuration, and ongoing maintenance to ensure high availability and performance.
* Write, maintain, and improve automation scripts in Python and BASH.
* Experience working knowledge in UNIX and Linux shell environments using command line utilities, also setting up development environments in Linux/OSX.
* Proficient in writing SQL Queries, Stored procedures, functions, packages, tables, views, and triggers using relational databases like Oracle, DB2, and MySQL.
* Experience in working with Python ORM Libraries including Django ORM.
* Experience in Test Driven Development and Behavior Driven Development methodologies for consulting firms and enterprise projects

**Technical Skills**

Operating Systems: Windows 98/2000/XP/7,8, Mac OS and Linux CentOS, Debian, Ubuntu

Programming Languages: Python, R, C, C++

Web Technologies: HTML/HTML5, CSS/CSS3, XML, jQuery, JSON, Bootstrap, Angular JS

Python Libraries/Packages: NumPy, SciPy, Boto, Pickle, PySide, PyTables, Data Frames, Pandas, Matplotlib, SQLAlchemy, HTTPLib2, Urllib2, Beautiful Soup, Py Query

Statistical Analysis Skills: A/B Testing, Time Series Analysis, Marko

IDE: PyCharm, PyScripter, Spyder, PyStudio, PyDev, IDLE, NetBeans, Sublime Text, Visual Code

Cloud Computing: AWS, Azure, Rackspace, OpenStack

Databases/Servers: MySQL, SQLite3, Cassandra, Redis, PostgreSQL, CouchDB, MongoDB,TerraData, Apache Web Server 2.0, NginX, Tomcat, JBoss, WebLogic

ETL: Informatica 9.6, Data Stage, SSIS.

Web Services/ Protocols: TCP/IP, UDP, FTP, HTTP/HTTPS, SOAP, Rest, Restful

Miscellaneous: Git, GitHub, SVN, CVS

Build and CI tools: Docker, Kubernetes, Maven, Gradle, Jenkins, Hudson, Bamboo

SDLC/Testing Methodologies: Agile, Waterfall, Scrum, TDD

**W.R.Berkley, Chesterfield, MO Sep 2023 to till the date.**

**Data Engineer**

**Responsibilities:**

* Requirement gathering for ongoing projects by working closely with subject matter experts and product selection leads.
* Sqoop jobs for Migrating data from sources such as Oracle, and SQL Server to Azure Data Factory.
* Developed and managed data pipelines using Azure Data Factory for scheduling ETL processes, data ingestion, and transformation, integrating with Snowflake for cloud data warehousing.
* Utilized Azure Synapse Analytics and Azure Data Lake Storage to facilitate data synchronization and storage for large-scale datasets.
* Implemented data pipelines using SQL queries to handle large-scale data transformation and aggregation.
* Used Azure Synapse and Snowflake to build a scalable data warehouse, enabling advanced analytics and reporting. Python and spark programs on EMR clusters for data integration.
* Implemented data synchronization between on-premises Oracle databases and Snowflake using Azure Data Factory and Azure Synapse Analytics, ensuring seamless data flow and real-time data access.
* Developed complex SQL joins, subqueries, and indexing to enhance data retrieval efficiency, resulting in faster query performance.
* Implemented Delta Lake architecture on Databricks for improved data reliability and performance.
* Developed Snowflake procedures, views, and performance-tuned queries to enhance data processing efficiency and support business intelligence initiatives.
* Implemented Azure Active Directory (AAD) and Azure Key Vault for secure data access management and governance.
* Developed and deployed scalable and resilient applications using Azure Fabric to manage microservices architecture in cloud environments.
* DAG jobs in Apache Airflow for scheduling.
* Automated data pipelines using Airflow, integrating with Snowflake and Azure Data Factory for seamless data orchestration across platforms.
* Developed custom Python operators and scripts for Airflow to enhance data pipeline functionality and error handling.
* Developed data processing pipelines in Scala using Apache Spark, enabling real-time analytics for high-volume streaming data.
* Leveraged Spark's parallel processing capabilities to optimize ETL workflows, reducing data processing times by 50%.
* Implemented data transformations in Scala, performing complex operations like data cleansing, aggregation, and feature extraction for machine learning models.
* Leveraged Scala for data processing and manipulation within Apache Spark, enhancing data transformation and analytics capabilities.
* Solution design document creation and Architecture for an Enterprise project.
* Data engineering tasks involved applying statistical analysis on a high-volume input file (30 GB) and run it as a parallel process in a high-performance cluster.
* Knowledge of Docker orchestration tools like Docker Swarm or Kubernetes to deploy and manage containers at scale, ensuring high availability and reliability of data services.
* Skill in defining and managing Kubernetes Pods, Deployments, StatefulSets, and DaemonSets for running data processing jobs and services efficiently.
* Proficiency in designing, implementing, and managing Apache Kafka-based data streaming architectures for real-time data processing and analytics.
* Gurobi Optimization on python for Predictive analysis.
* Developed custom Airflow operators and sensors to integrate with proprietary systems, improving workflow efficiency by 30%.
* Implemented row-level security and dynamic data masking in Snowflake, ensuring GDPR compliance for sensitive data.
* Skill in optimizing Solr queries, leveraging features like faceting, highlighting, filtering, and sorting to enhance search performance and user experience.
* Skill in optimizing Aerospike database performance by tuning configurations, indexing strategies, and data storage options for low-latency data access and high throughput.
* Data modeling in MS Visio of all the data sources ingested into a web application.
* Jira for project management and GitHub for code reposition.
* Power BI and Tableau for Data Visualization and Analytics.
* Environment: AWS (EC2, S3, Lambda, SQS & SNS, Glue, Athena, AWS Amplify, Workspace, Aurora DB etc.), Boto 3, R, SQL Server, EMR, Hadoop, Oracle, Spark, Python, Kubernetes, Dockers, Kafka, Json, XML&CSV files, GitHub, Jira, Aerospike.

**Toyota Washington, DC Dec 2021 to Sep 2023**

**Data Engineer**

Responsibilities:

* Worked creating, developing on retail sales, POS, Corporate Forecast data by developing & deploying spark pipelines for ensuring continuous delivery of data from various cross-functional teams like Omni-channel, Demand-Supply, Global Logistics and deliver enterprise cleansed datasets to BI Engineering & Data Science teams.
* Collaborate Solution Architects, Principle Engineers, Data Scientists, DevOps team, Data Governance & Business Analysts to understand the precise business needs of the acceptance criteria and ensuring the deliverables of data products leveraged for Corporate Forecasts and Forecast Accuracy biases by Business & Product owners.
* **Data Stack:** Azure Databricks, ADLS, ADF, AAS, DAX, Azure Automation Accounts, Azure Active Directory(AD), Azure IAM security Groups, Pyspark, Spark SQL, Azure Data warehouse (ADW), Power BI, DAX coding, MSBI, SSAS, CI/CD and Production Support.
* Developed and managed ETL pipelines on Azure Databricks, utilizing PySpark for data transformation and integrating with Snowflake to enhance data warehousing and analytics capabilities.
* Utilized Azure Fabric for service orchestration, enabling seamless integration and management of distributed applications.
* Implemented monitoring and diagnostics in Azure Fabric to enhance application performance and reliability.
* Developed Spark applications using Scala, optimizing ETL processes to handle large datasets efficiently.
* Implemented data ingestion pipelines using Hadoop to process sales data from multiple sources, providing insights into customer behavior and sales trends.
* Utilized Hive for querying large datasets, enabling fast analytics and report generation across distributed data.
* Designed and managed data storage on HDFS, ensuring scalability and fault-tolerance for high-volume sales data processing.
* Orchestrated data workflows using Azure Data Factory, facilitating seamless data movement between Azure Data Lake Storage (ADLS) and Snowflake for efficient storage and retrieval. Familiarity with incorporating Docker into CI/CD pipelines to automate testing, integration, and deployment of data pipelines and analytics solutions.
* Utilized Azure Analysis Services (AAS) for developing tabular models and enhancing data analytics capabilities.
* Built scalable data integration pipelines to ingest, cleanse, and transform data from various sources into Snowflake, supporting cross-functional teams such as Omni-channel, Demand-Supply, and Global Logistics.
* Employed DAX coding and Power BI to create dynamic dashboards and reports for business stakeholders.
* Skill in creating Docker images, Dockerfiles, and Docker registries to facilitate efficient development, testing, and deployment of data pipelines and applications.
* Performed end-to-end delivery of pyspark ETL pipelines on Azure-databricks to perform the transformation of data orchestrated via Azure Data Factory (ADF) scheduled through Azure automation accounts and trigger them using Tidal Schedular.
* Created Azure AAS Models from scratch by designing the Dimensional tables derived from fact table and Normalize data to create a tabular model on top which PowerBI reports are generated by business users.
* Created security groups through CI/CD process and associated the object-IDs to the user groups based on the business domain filters.
* Used Enterprise GitHub and Azure DevOps Repos for version control. Good understanding of branching strategies while collaborating with peer groups and other teams on shared repositories.
* **Environment:** Azure Databricks, Azure Databases, Azure Devops, Azure Repos, Pyspark, Delta-lake, Azure Data-warehouse,Data Modeling, Tidal Scheduler, Azure Data Factory(ADF), Data Lake Storage (ADLS), Analysis Services (AAS), Databricks (DBRX), PowerBI, SQL Server Management Studio (SSMS),Azure Automation Accounts, Runbooks, Webhooks, SparkSQL.

**Caterpillar Inc., Morton, IL, USA Dec 2019 – Nov 2021**

**Snowflake Data Engineer**

**Responsibilities:**

* Responsible for designing, implementing, validating, and maintaining data pipelines using Apache Airflow.
* Orchestrated Airflow DAGs to schedule the Ingestions, ETL jobs and business reports using YAML files on a Managed Airflow platform.
* Worked extensively on Snowflake cloud data warehouse implementation on AWS.
* Full load and delta load of data objects from (AWS S3, SQL Server, Oracle, Teradata) to snowflake cloud using Airflow pipelines.
* Used COPY, LIST, PUT and GET commands for validating the internal stage files.
* Implemented complex Snow SQL scripts in snowflake cloud data warehouse for business analysis and reporting.
* Developed snowflake procedures for executing branching and looping
* Implemented Alteryx workflows as Snowflake views, redesigned the existing views in snowflake for performance enhancement.
* Performed data transformations using Spark data frames in PySpark, EMR, and Airflow.
* Understanding of Docker security best practices, such as container hardening, image scanning, and role-based access control (RBAC), to ensure data integrity and compliance.
* Skill in defining and managing Kubernetes Pods, Deployments, StatefulSets, and DaemonSets for running data processing jobs and services efficiently.
* Understanding of Kafka security features such as SSL encryption, authentication mechanisms (e.g., SASL), authorization (e.g., ACLs), and secure multi-tenancy practices to protect data integrity and confidentiality in distributed environments.
* Scripted Snowflake views to feed as inputs for PowerBI and Tableau reports.
* Defined roles and privileges required to access different database objects.
* Supported production environment and debug issues using CloudWatch logs.
* Create and load staging tables based on the logic defined views using distribution, Sort Key’s for optimal performance.
* Environment: AWS, S3, DynamoDB, Snowflake, Lambda, Cloud Watch, EMR,Python, PowerBI, Tableau, SSIS cubes, Airflow, Spark, GitHub, Docker, Kubernetes, Kafka, Jenkins, Jira.

**Cigna Health, Overland Park, KS Sep 2017 to Dec 2019**

**Data Engineer**

**Responsibilities:**

* Developed entire frontend and backend modules using Python on Django Web Framework.
* Used update strategy to effectively migrate data from source to target.
* Designed ETL Process using Informatica to load data from Flat Files, and Excel Files to target Oracle Data Warehouse database.
* Interacted with the business community and database administrators to identify the Business requirements and data realties.
* Created various transformations according to the business logic like Source Qualifier, Normalizer, Lookup, Stored Procedure, Sequence Generator, Router, Filter, Aggregator, Joiner, Expression and Update Strategy.
* Created Informatica mappings using various Transformations like Joiner, Aggregate, Expression, Filter and Update Strategy.
* Improving workflow performance by shifting filters as close as possible to the source and selecting tables with fewer rows as the master during joins.
* Used connected and unconnected lookups whenever appropriate, along with the use of appropriate caches.
* Engaged with various Azure services, enhancing our data engineering capabilities and ensuring efficient data management and analytics.
* Designed and implemented ETL pipelines using SQL to extract data from multiple sources, transform it to meet business requirements, and load it into a centralized data warehouse.
* Developed data models and schemas for data warehousing solutions, enabling efficient data storage, retrieval, and reporting.
* Automated ETL processes using SQL scripts, scheduling tasks to ensure timely and accurate data loading, minimizing data lag by 40%.
* Developed and managed ETL processes using Azure Data Factory, automating data ingestion and transformation workflows. This allowed for seamless integration of data from multiple sources into our data warehouse, improving data accessibility for analytics.
* Used Azure Data Lake Storage to store and manage large volumes of unstructured data. By implementing data ingestion pipelines, ensured that data was readily available for analytics and reporting, facilitating better decision-making across the organization.
* worked with Azure Synapse Analytics to enable advanced analytics and data integration. Developed data pipelines that connected various data sources, allowing for comprehensive data analysis and reporting capabilities.
* Perform Maintenance, including managing Space, Remove Bad Files, Remove Cache Files and monitoring services.
* Set up Permissions for Groups and Users in all Development Environments.
* Migration of developed objects across different environments.
* Experienced in Agile Methodologies and SCRUM Process.
* Maintained program libraries, user’s manuals and technical documentation.
* Worked in development of applications especially in UNIX environment and familiar with all its commands.
* Handling the day to day issues and fine tuning the applications for enhanced performance.
* Implement code in Python to retrieve and manipulate data.
* Experience in using Kubernetes-native monitoring tools (e.g., Prometheus, Grafana) and logging solutions (e.g., Fluentd, Elasticsearch, Kibana) to monitor cluster health and application performance.
* Environment: Python,Kubernetes, Django, MySQL, Linux, Informatica Power Centre 9.6.1, PL/SQL, HTML, XHTML, CSS, AJAX, JavaScript, Apache Web Server, NO SQL, jQuery.

**Genpact Pune India July 2013 to Aug 2016**

**Software Developer**

**Description**

We worked for an insurance client with Improved coordination between data teams during the cloud transition project and for ongoing work in the data center of excellence (CoE)

Collaborated with the client to build a data CoE and expert team to Implemented and orchestrated the latest technology in python, Instilled best practices and project and data management to improve workflows

Created a robust and scalable enterprise data warehouse (EDW) and organization data service (ODS) with a cloud-based data lake, all governed by best practices and data frameworks

**Responsibilities:**

* Involved in Design, Development and Support phases of Software Development Life Cycle (SDLC).
* Designed and developed components using Python with Django framework. Implemented code in python to retrieve and manipulate data. Developed Python Framework using Django to perform scan software unit monitoring.
* Involved in development of the enterprise social network application using Python, Twisted, and Cassandra and responsible for setting up Python REST API framework and spring frame work using Django.
* Developed Rest API's using python with flask and Django framework and done the integration of various data sources including RDBMS, Shell Scripting, Spreadsheets, and Text files.
* Experienced in Object Oriented Analysis in developing Python Back-End framework using various design patterns. Used Python programming and Django for the backend development, Bootstrap and Angular for frontend connectivity and MongoDB for database.
* Developed web sites using Django, Python, XHTML, CSS, and JavaScript. Developed and designed e-mail marketing campaigns using HTML and CSS. Tested and implemented applications built using Python.
* Developed views and templates with Python and Django’s view controller and templating language to create a user-friendly website interface.
* Developed of Python APIs to dump the array structures in the Processor at the failure point for debugging. Handling Web applications - UI security, logging, backend services.
* Applied breadth of knowledge in programming (Python), Descriptive, Inferential, and Experimental Design statistics, advanced mathematics, and database functionality (SQL, Hadoop).
* Developed Spark/Scala, Python for regular expression (regex) project in the Hadoop/Hive environment with Linux/Windows for big data resources. Used clustering technique K-Means to identify outliers and to classify unlabeled data.

**Education:**

* Bachelor's degree, India
* Masters of Computer Science, US