**SWATI RANI SUBUDHI**

**TALASU**

**raj@abntechcorp.com** 

 **484-841-7747**

 **Tracy, CA 95391**

 **www.linkedin.com/in/swati-rani-subudhi-talasu**

**PROFESSIONAL SUMMARY**

* Around 12+ years of IT experience as a Dataengineer with capacity for creating, implementing, and documenting data models for enterprise-level applications. Background in ETL Data pipeline, Data Visualization, Data Warehousing and Data Lake.
* Learn about accelerating data processing by running Pyspark tasks on a Kubernetes cluster.
* Involved in the creation of RDDs and data frames from the necessary HDFS files in order to transform Hive
* Queries into different Spark actions and transformations.
* Practical knowledge of dimensional modeling with the use of star and snowflake models.
* Extensive knowledge of the Big Data Ecosystem, specifically using the Hadoop framework and related technologies including HDFS, MapReduce, HIVE, PIG, HBASE, STORM, YARN, OOZIE, SQOOP, Airflow and Zookeeper. Also, includes work with SparkCore, SparkSQL, Spark Streaming, Scala and Kafka.
* Solid background in converting other databases to Snowflake. Comprehensive understanding of the Snowflake Database's schema and table architecture.
* CI/CD (continuous integration & deployment) pipelines were set up, developed, and maintained. Automation was used in environments and applications. GIT, Terraform, and Ansible were some of the automation technologies.
* Skilled in transactional modeling, fact dimensional modeling (Star schema, Snowflake schema) and SCD
* (Slowly changing dimension).
* A solid background in data modeling, data pipelines, and SQL and NoSQL databases. Participating in the creation and automation of ETL pipelines from start to finish using Python andSQL
* The DAGs were constructed using Airflow to perform jobs sequentially and in parallel after installing and configuring Apache Airflow for workflow management.
* Knowledge of version control tools like GitLab, SVN Code Commit, and Bitbucket to arrange the versions and settings of the code.
* Using SparkContext, Spark-SQL, Dataframe API, Spark Streaming, MLlib, Pair RDDs, and working specifically with PySpark and Scala, an expert with Spark improved the performance and optimized the current Hadoop techniques.
* Used cloud shell SDK in GCP to configure the services Data Proc, Storage, Big Query.
* Developed Groovy Domain classes and scripts to automate repetitive tasks, significantly reducing manual workload and improving operational efficiency
* A solid working understanding of command-line utilities in the UNIX and Linux shell environments.
* Skilled in using Agile Methodologies, Scrum stories, sprints, daily standup meetings, and pair programming to achieve high-quality deliverables on schedule
* Extensive knowledge in Dimensional Data Modeling, Relational Data Modeling, StarSchema/Snowflake
* Modeling, FACT & Dimensions Tables, Physical & Logical Data Modeling, and Data Analysis.
* Having expertise in the whole software development life cycle (SDLC), designing scalable platforms, object- oriented programming, database design, and agile approaches.
* A solid working knowledge of Docker's image format, registry, and Kubernetes for container-based deployments.
* Amazon S3 buckets were used to install AWS Lambda code. In order to receive events from your S3 bucket, you must first create a Lambda Deployment function.
* Experience with databases like DynamoDB, S3 Buckets, MySQL, and Elastic Cache on the AWS Cloud with good knowledge of IaaS, PaaS, andSaaS.
* Excellent understanding of Sql Basics and Working knowledge of Teradata and Oracle Database.
* Hands-on experience in working with Cloudera, Azure HDInsight, and AWS cloud
* Hands on experience on Unified Data Analytics with Databricks, Databricks Workspace User Interface, Managing Databricks Notebooks, Delta Lake with Python, Delta Lake with Spark SQL.
* Good knowledge of NoSQL databases like MongoDB, Redis, and ApacheCassandra.
* Effective team work requires strong leadership, strong work ethics, and solid communication abilities. Agile and Waterfall methodologies are used for implementation and support.
* Knowledgeable of EC2, Cloud Watch, Cloud Formation, and managing security groups on AWS, as well as automating, configuring, and deploying instances on these platforms.
* Strong Python PEP Guidelines compliance and experience with Linux Bash scripting
* Being skilled in creating SQL queries, stored procedures, functions, packages, tables, views, and triggers for relational databases like Oracle, DB2, MySQL, and MSSQLServer.

**EDUCATION:**

San Jose State University

San Jose, CA  , 12/2011

*Master of Science*: Software Engineering, Enterprise Software

**CERTIFICATIONS**

*AWS Solutions Architect Associate Certification*

*Certified Snowflake SnowPro Core*

# Technical Skill Set:

|  |  |
| --- | --- |
| ETL Tools | AWS Glue, Airflow, Spark, Sqoop, Flume, Apache Kafka, Spark Streaming |
| NoSQL Databases | MongoDB, Cassandra, Amazon DynamoDB, HBase |
| Data Warehouse | AWS RedShift, Google Cloud Storage, Snowflake, Teradata, |
| Tools | PyCharm, Visual Studio, R Studio, Power BI, Tableau, SAS Studio,Gephi, Eclipse, Putty, Mainframes, Excel, Jupyter Notebook, Azure Databricks |
| Web Development | HTML, XML, JSON, CSS, JQUERY, JavaScript |
| Monitoring Tools | Splunk, Chef, Nagios, ELK |
| Source Code Management | JFrogArtifactory, Nexus, GitHub, Code Commit |
| Containerization | Docker &DockerHub, Kubernetes, OpenShift |
| Hadoop Distribution | Cloudera, Hortonworks, MapR, AWS EMR, GCP DataProc |
| Programming and Scripting | Spark Scala, Python, Java, MySQL, PostgreSQL, Shell Scripting, Pig, HiveQL |
| AWS | EC2, S3, Glacier, Redshift, RDS, EMR, Lambda, Glue, CloudWatch, Recognition,Kinesis,CloudFront,Route53,DynamoDB,CodePipeline, EKS, Athena, QuickSight |
| Hadoop Tools | HDFS, HBase, Hive, YARN, MapReduce, Pig, HIVE, Apache Storm, Sqoop, Oozie, Zookeeper, Spark, SOLR, Atlas |
| Build & Development Tools | Jenkins, Maven, Gradle, Bamboo |
| Methodologies | Agile/Scrum, Waterfall |

**Hitachi Vantara, Tracy, CA. 02/2022 - Current**

**Sr. AWS Data Engineer**

* Performed actions and transformations on RDDs, data frames, and datasets using SparkSQL and Spark Streaming Contexts while working on a Scala code base connected to Apache Spark.
* Making sure the development process adheres to the SDLC (Software Development Life Cycle) principles
* Worked on the design of Restful Software services and stored the data in MySQL and PostgreSQL
* databases.
* Developed a plan and executed a successful migration of databases from on-premises to the AWS cloud.
* Creating data pipelines from many sources to Snowflake.
* Create a snowflake warehouse strategy and set it up to use PUT scripts to migrate a terabyte of data from S3 into Snowflake.
* By creating a customized read/write Snowflake utility function in Scala, data was transferred from an AWS S3 bucket to Snowflake.
* Developed Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for Analyzing& transforming the data to uncover insights into the customer usage patterns
* In Snowflake, create an external S3 stage to be used for data migration.
* Performed actions and transformations on RDDs, data frames, and datasets using SparkSQL and Spark Streaming Contexts while working on a Scala code base connected to Apache Spark.
* To keep up with business standards, used and configured a variety of AWS services, including RedShift, EMR, EC2, and S3.
* Conducted data blending, prepared data for Tableau consumption using Alteryx and SQL, and published data sources to the Tableau Server.
* For data extraction, processing, and aggregation from various file formats, developed Spark applications utilizing Pyspark and Spark-SQL
* Made Spark Streaming tasks in Python to retrieve JSON files from AWS S3 buckets and read messages from Kafka.
* Responsible for converting manual report systems into fully automated CI/CD data pipelines that feed data from several marketing platforms into an AWS S3 data lake.
* The loading of data into Parquet hive tables from Avro hive tables after creating partitioned and bucketed Hive tables in Parquet File Formats using Snappy compression.
* In AWS, Lambda, and other requirements were deployed to automate EMR Spin jobs.
* Created, scheduled, and monitored Data Pipelines with Apache Airflow
* Worked with Jira and Confluence and had data visualization skills using Matplotlib and the Seaborn package.
* Involved in creating a test environment utilizing Docker containers and setting up the containers using Kubernetes.
* Scheduled Spark processes/applications in the AWS EMR cluster.
* Involved in processing huge datasets of diverse types, including structured, semi-structured, and unstructured data. Worked on creating tables in Hive, MYSQL utilizing SQOOP, and processing data, such as importing and exporting databases to the HDFS.
* Knowledge of big data analytics, including Spark, Hive, Pig, Sqoop, HBase, Oozie, Impala, and Kafka, as well as MapReduce programming
* Written multiple MapReduce Jobs using Java API, Pig and Hive for data extraction, transformation and aggregation from multiple file formats including Parquet, Avro, XML, JSON, CSV, ORCFILE and other compressed file formats
* Working on building data pipelines, end to end ETL process for ingesting data in GCP.
* Collaborated seamlessly with multidisciplinary teams to devise and execute Groovy-driven solutions aimed at enhancing software development processes and code quality
* To create dashboards and workbooks using PowerBI, action filters, parameters, and calculated sets were created.
* Creating a data pipeline using Spark, Scala, and Apache Kafka to ingest data from a CSL source and store it in an HDFS secured folder.
* Install and configure Apache Airflow for AWS S3 bucket and create dags to run the Airflow
* Built user-friendly website interfaces using Python and Django view controllers and templating language.
* The creation and development of UNIX shell scripts for task scheduling. Additionally, PL/SQL scripts were written for removing and rebuilding indexes, as well as pre-and post-session shell scripts.

**Environment**: Python, SQL, Amazon Web Services (EC2, S3, RDS, Elastic Load Balancing, Elastic Search, Amazon MQ, Lambdas, SQS, IAM, Google Cloud Platform (GCP), CloudWatch, EBS, CloudFormation), Pandas, NumPy, seaborn, SciPy, Matplotlib, Scikit-learn, NLTK, Docker, Machine Learning, Snowflake, Hadoop, Hive, Sqoop, Yarn, HDFS, Flume, PySpark, Spark SQL, Tableau, SAS Visual Analytics

**First Republic Bank ,SANTA CLARA ,CA 01/2021 - 02/2022**

**AWS Data Engineer**

* Utilizing both organized and unstructured data, reports and dashboards were produced.
* ETL jobs were created to load server data and data from other sources into buckets and move S3 data into the data warehouse.
* Used PySpark, to join different tables and then used EMR to perform analytics on top of them.
* Designed and developed a Scala pipeline to transform and store data in the Parquet format and allow HIVE.
* Developing controllers that perform data transformation tasks in accordance with business needs.
* Used spark streaming to transform data in real-time.
* Created several Python scripts that do SQL injection, permission checks, and analysis to discover SQL query vulnerabilities.
* By creating a customized read/write Snowflake utility function in Scala, data was transferred from an AWS S3 bucket to Snowflake.
* Participated in the creation of apps in Scala and Spark. Created Python Spark Streaming tasks to download JSON files from AWS S3 buckets and read messages from Kafka.
* Designed Python functions and Spark programs that conduct operations and transformations on data sets.
* Actively gathered project planning and development needs from business customer SMEs.
* To automate EMR Spin for Data Lake jobs, Lambda and other dependencies were deployed into AWS.
* Configure spark jobs for continuous integration and deployment to EMR clusters. Spark tasks/applications scheduled in the AWS EMR cluster.
* Spark SQL was used to connect to Hive to read data, and distributed processing was used to make the system very scalable.
* Using Tableau/OBIEE, we wrote data and loaded it into a Data Lake environment (SNOWFLAKE) that was accessed by business users and data scientists.
* Participated in the loading of data from the LINUX file system to the HDFS.
* Created an ETL pipeline to use in extracting historical logs from various sources and storing them in an S3 data lake.
* Developed a plan and moved databases to the AWS cloud from on-premises locations effectively.
* Made several dashboards in Tableau for various business requirements.
* Published dashboards, reports, and spreadsheets with interactive data visualization on Looker.
* For effective data access, partitioning, dynamic partitions, and buckets were implemented in HIVE.
* Added test scripts to aid continuous integration and test-driven development.
* Multi-terabyte, complete end-to-end Data Warehouse infrastructure designed and constructed from the ground up.
* Confidential Redshift for managing massive amounts of data Daily millions of records.
* ETL jobs were designed and created to extract data from the Salesforce replica and load it into the Redshift data mart.
* Made a Splunk dashboard to store the logs for the entire data import process.
* Created unit test plans for the Spark code used in the CICD process.
* Loaded JSON data into SparkSQL, created a Schema RDD, and loaded it into Hive tables while handling structured data with Spark SQL.

**Environment**: HDFS, Hive, Scoop, Pig, Oozie, Amazon Web Services (AWS), SNS, Spark SQL, GLUE, Python, GIT, Jenkins, Snowflake, SQL, Redshift, Schema RDD.

**Cisco , San Jose, CA 04/2016 - 12/2020**

**Data Engineer**

* Executed upgrades and maintenance of Cisco Digital Customer Registry (DCR) for legal hierarchies as well as descriptive and firmographic customer master data.
* Developed a single source of truth for Cisco Customer Master Data to drive accurate data acquisition and manage customer relationship information and bookings.
* Created core DCR components that correctly assign or create a customer in the hierarchy.
* Led data management, system performance, and customer experience initiatives.
* System Migration and Optimization
* Developed tools and methods to facilitate the decommissioning of Hadoop and the migration of all jobs and associated programs to Snowflake cloud-based data storage solution; developed and maintained the programs in Oracle. Obtained SnowPro Core Certification and completed the transition on time. Drastically reduced runtime by 70%.
* Programmed and maintained ETL data pipelines for incoming data originating from various sources.
* Scrubbed existing data in multiple phases. Automated cleanup process for current data and decreased incoming data pollution by 95%
* Business Intelligence
* Maintained DCR data with industry changes such as mergers, acquisitions, and divestitures. Provided NRT data to various organizations within Cisco.
* Designed Customer Partner Account Management View, linking ERP account sites, DCR Customer sites, and partner sites, supporting transactions across Cisco commerce lifecycle.
Created and maintained Customer Account View used to track customer success.

**Cisco – San Jose, CA 07/2012 - 04/2016**

**Software Engineer**

**Role: Python Developer**

* Designed, develop, test, deploy and maintain the website.
* Developed entire frontend and backend modules using Python.
* Developed Python scripts to update content in the database and manipulate files.
* Rewrite existing Java application in Python module to deliver certain format of data
* Developed entire frontend and backend modules usingPython.
* Generated property list for every application dynamically using Python.
* Designed and developed the UI of the website using HTML, XHTML, AJAX, CSS and JavaScript.
* Wrote Python scripts to parse XML documents and load the data in database.
* Generated property list for every application dynamically using Python.
* Designed and developed data management system using MySQL.
* Creating unit test/regression test framework for working/new code.
* Responsible for search engine optimization to improve the visibility of the website.
* Responsible for debugging and troubleshooting the web application.