**SURESH BODDU**

sureshboddu9@gmail.com

https://www.linkedin.com/in/suresh-boddu-062a6196

+1 737-222-9032

Austin, Texas, USA




## Professional Summary:

- Around 10 years of experience in **Data Engineering** with expertise in **Big Data technologies, Data Pipelines, SQL/NoSQL, Cloud based RDS, Distributed Database, Serverless Architecture, Data Mining,** and cloud technologies **like Snowflake, AWS EMR, Redshift, Lambda, Step Functions, Cloud Watch.**
- Adapt at **implementing E2E solutions** on Big Data using Hadoop framework, executed, and designed big data solutions on multiple distribution systems.
- Expertise in designing data intensive applications using **Hadoop Ecosystem, Big Data Analytical, Cloud Data engineering, Data Warehouse, Data Visualization, Reporting**, and Data Quality solutions.
- Expertise in Big Data processing using Hadoop, Hadoop Ecosystem (**Map Reduce, Spark, Python, Scala, Hive, HBase, Mongo DB**) implementation, maintenance, ETL and Big Data analysis operations.
- Involved in the development of real time streaming applications using **PySpark, Apache Flink, Kafka, Hive** on distributed Hadoop Cluster
- 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 understanding of Spark Architecture with Databricks, Structured Streaming, setting up AWS with Databricks, databricks workdspace for business analytics, manage clusters in databricks.
- Experience in working on Spark- Core, Data Frame, Dataset, SQL, Delta Lake and Structured Streaming.
- Experience in working on different big data warehouses in **Snowflake** and **Star** schemas.
- Strong experience in writing **Hive UDF**, **Generic UDF's** to incorporate complex business logic into **Hive Queries**.
- Involved in designing the data model in Hive for migrating the **ETL** process into **Hadoop** and to load data into **Hadoop** environment.
- Involved in converting Hive/SQL queries into Spark transformations using **Spark Data frames and Python**.
- Strong experience in Python, SQL, Teradata, **SAS**, **QlikView, Tableau and Power BI** and worked on Feature Scaling, Feature Engineering, Modeling and Evaluation with Python.
- Worked extensively with **Dimensional modeling**, **Data migration**, **Data cleansing**, **Data profiling**, and **ETL** Processes features for data warehouses.
- Experience in importing and exporting data using **Sqoop** from **HDFS** to **Relational Database Systems** (**RDBMS**), **Teradata** and vice versa.
- Experience in developing ETL applications in AWS which include creating data pipeline using **AWS Glue, EMR (Managed Hadoop), PySpark, Python, Scala, Redshift, S3, Athena, EC2, Lambda.**
- Expertise in Creating, Debugging, Scheduling and Monitoring jobs using Airflow.
- Good working experience on Spark (spark streaming, spark SQL) with Scala and Kafka. Worked on reading multiple data formats on HDFS using Python. Managed Error Handling, Performance Tuning, Error Logging clustering and High Availability
- Experience implementing Cloud based **Linux OS in AWS** to Develop Scalable Applications with **Python.**
- Experience on Shell scripting to automate various activities.
- Capable of using AWS utilities such as EMR, S3 and cloud watch to run and monitor **Hadoop and spark jobs** on **Amazon Web Services (AWS)**.
- Worked with and maintained **data warehouses** in **snowflake** and **star** schemas.
- Expert in big data ecosystem using Hadoop, Spark, Kafka with column-oriented big data systems such as Vertica and Cassandra
- Expertise in design and development of various web and enterprise applications using Type safe technologies like **Scala.**
- Install and configure Apache Airflow for AWS S3 bucket and create dags to run the Airflow
- Experience in developing workflows using Flume Agents with multiple sources like Web Server logs, **REST API** and multiple sinks like **HDFS sink**.
- Expert in Transact-SQL (DDL, DML, & DCL) like views, functions, procedures and writing complex ad-hoc queries for project maintenance.
- Successfully design and deploy data processing solutions on **AWS Databricks**, enabling data ingestion, transformation, and processing workflows effective.
- Expert in **Pentaho & Talend ETL** tools for designing ETL jobs in the process of building Data warehouses and Data Marts.

## Technical Skills:

| | |
|---|---|
| **Big Data Technologies** | Python, Spark, Databricks, AWS Glue, Hadoop, HDFS, Hive, HBase, Flume |
| **Databases** | My SQL, Oracle |
| **Cloud DataWarehouse** | Snowflake, Redshift |
| **NoSQL Databases** | MongoDB, HBase, Dynamo DB, Oracle NoSQL Database |
| **ETL/Integration Tools** | Pentaho, Talend, Informatica |
| **Version Control** | Git hub, AWS CodeCommit |
| **Languages** | Python, Spark, Core Java, Scala |
| **Scripting** | Shell Scripting |
| **Visualization Tools** | Tableau, IBM Cognos |
| **Cloud Platforms** | AWS and Azure |

## Education:

**Master of Computer Applications – JNTUK University, Kakinada, India: 2013**
**Bachelor of Science (Computers) – Andhra University, India: 2010**

**Certifications:**

| | |
|---|---|
| Databricks Certified Data Engineer Associate | https://scl.io/1q6yZ9S |
| AWS Certified Solutions Architect – Associate | https://www.credly.com/badges/b443bdce-f739-4863-929b-3446e55c362e |
| SnowPro Core Certification - Issued by Snowflake | https://www.credly.com/badges/562f61a9-6969-4785-ba5b-a63c57e4d46e |

**Work Experience:**

**Global Business Intelligent Project | Apple Inc | Tata Consultancy Services** **August 2021 – Present**
**Sr Data Engineer**
**Responsibilities:**

- Collaborated with cross-functional teams to gather and analyze business requirements for data-driven projects, ensuring alignment with organizational objectives.
- Designed, developed, and maintained robust data pipelines using Python, incorporating best practices for data extraction, transformation, and loading (**ETL**).
- Developed spark jobs using Spark-SQL in **Databricks notebooks** for data extraction, transformation, and aggregation from multiple file formats for analyzing, transforming the data and loading into **databricks DBFS**.
- Responsible for estimating the cluster size, monitoring, and troubleshooting of the **Spark databricks cluster**.
- Developed PySpark jobs to ingest data data from aws S3 to **Delta tables in Databricks**.
- Optimized Pyspark jobs on databricks, which yielded a significant amount of cost reduction.
- Developed Spark applications using **Pyspark** and **Spark-SQL** for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns in Databricks.
- 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.
- Utilized advanced **SQL and Snowflake** queries to extract and manipulate data from diverse sources, addressing specific business needs and data analysis requirements.
- Played a key role in monitoring and optimizing the performance of applications and infrastructure by leveraging **Datadog**, ensuring efficient resource utilization.
- Demonstrated expertise in **Amazon Web Services (AWS)** cloud services, including **EC2, EBS, S3, VPC, and Elastic Load Balancer**, for scalable and cost-effective data solutions.
- Utilized Spark SQL API in **PySpark** to extract and load data and perform SQL queries.
- Leveraged Lambda functions to automate infrastructure provisioning and management, enhancing system reliability and reducing manual intervention.
- Successfully set up and managed databases in **AWS using RDS**, implementing robust backup and recovery strategies for S3 bucket storage.
- Involved in converting Hive/SQL queries into Spark transformations using Spark Data frames and Pyspark.
- Effectively managed Amazon EC2 clusters, deploying and maintaining files within buckets, and optimizing data storage and retrieval processes.
- Responsible for data extraction and data ingestion from different data sources into Snowflake by creating ETL pipelines using **PySpark.**
- Developed Kafka consumer API in Scala for consuming data from Kafka topics.
- Install and configure Apache Airflow for AWS S3 bucket and create dags to run the Airflow.
- Extensively utilized AWS services like **AppSync, S3, Lambda, ECS, Fargate, DynamoDb, Cloudwatch, CodePipeline, EKS**, and others to develop comprehensive data solutions.
- Deployed and thoroughly tested different modules within **Docker** containers and **GIT**, streamlining development workflows and ensuring code consistency.
- Developing Spark scripts, UDFS using both Spark DSL and Spark SQL query for data aggregation, querying, and writing data back into RDBMS through Sqoop.
- Implemented automated **CI/CD pipelines** using Jenkins and Ansible, fostering a culture of continuous integration and deployment, and achieving efficient software delivery.
- Stored the log files in AWS S3. Used versioning in S3 buckets where the highly sensitive information is stored.
- Proficiently wrote and executed **MySQL** database queries from Python, utilizing the **Python-MySQL** connector and **MySQL DB** package for seamless data retrieval and manipulation.
- Utilized Postman **API** for visualizing query results, simplifying data validation and testing processes.
- Implemented and configured AWS Security Groups to facilitate the deployment and management of **AWS EC2 instances**, ensuring a secure and compliant infrastructure.

**Environment:** Pyspark, Databricks, Kafka, Hive, Apache Spark, Scala, Snowflake, Python, AWS Services (Lambda, EMR, Autoscaling), Github, Restful web service

**P&G Pampers Datamart | P&G, UK | Wipro** **January 2020 to August 2021**
**Sr Data Engineer**
**Responsibilities:**

- Translate business propositions into quantitative queries and collect/clean the necessary data.
- Build scalable databases capable of **ETL** processes using **SQL and Spark**.
- Ingested data from different data sources like Smart Button Loyalty Platform (SFTP), P&G's Customer Portal (Mongo DB), Campaigns (Amazon S3) and User activity data from MySQL database.
- Evaluate the workflow and increase the efficiency of data pipelines that process over 50 TB of data daily.
- Utilize MongoDB to create NoSQL databases that harvests data from a variety of sources.
- **Key Achievement**: Developed a data pipeline using Delta Lake that led to a process optimization and corresponding revenue increase of 19%.
- Used AWS-CLI to suspend an AWS Lambda function. Used AWS CLI to automate backups of ephemeral data-stores to S3 buckets, EBS.
- Moving this partitioned data onto the different tables as per as business requirements.

- Develop and deploy the outcome using **spark** code in Hadoop cluster running on **AWS**.
- Setting up the work schedule using oozie and identifying the errors in the logs, rescheduling/resuming the job.
- Involved in Designing and Developing Enhancements product features.
- Extracting data from data warehouse (**MySQL**)on to the Spark RDD's.
- Working on Stateful Transformations in Spark Streaming.
- Worked on Ingesting data by going through cleansing and transformations and leveraging **AWS Lambda, SAS, AWS Glue and Step Functions**.
- Used Scala function, dictionary, and data structure (array, list, map) for better code reusability.
- Good hands-on experience on Loading data onto Hive from Spark RDD's.
- Using decision tree as a model evaluation for both classification and regression.
- Stored data in **AWS S3** like HDFS and performed EMR programs on data stored.
- Collaborated with the infrastructure, network, database, application, and BI teams to ensure data quality and availability.
- Worked and learned a great deal from AWS Cloud services **like EC2, S3, EBS, RDS and VPC.**
- Creating external tables and moving the data onto the tables from managed tables.
- Created monitors, alarms, and notifications for **EC2** hosts using Cloud Watch, Cloud trail and SNS.

**Environment**: PySpark, SQL, Amazon EMR, MySQL, Step Function, CloudWatch, Snowflake, Redshift, Lambda

**SOMOS TFN Registry | Ericsson, US | Wipro** **March 2019 – December 2019**
**Data Engineer**
**Responsibilities:**

- Participated in all phases including Analysis, Design, Coding, Testing and Documentation and gathered requirements and performed Business Analysis.
- Responsible for development, support, maintenance and implementation of a complex project module. Configured and deployed instances on **AWS** environments.
- Worked as an independent team member, capable of applying judgment to plan and execute your tasks.
- Responded to technical queries / requests from team members and customers.
- Responsible to coach, guide and mentor junior members in the team.
- Involved in requirements gathering with help of BA's and Module leads.
- Attended regular SCRUM meetings which are related to tasks status, backlog grooming, participating in sprint release planning and retro.
- Participated in new databases creation and will update/modify DB with alter script using DB Patch.
- Implemented complex reports ETL's. Prepared Unit Test Cases and executed them.
- Understand the functional Specification and end user requirements.
- Created the jobs and transformations as per business requirement.
- Setting up of different environments. Thoroughly understand bugs, analyze the bugs and fixed as per the new requirement.
- Error handling implemented in all ways of work. Daily monitoring the **Pentaho** servers.
- Fixed many Incidents and Defects raised by Customer.
- Supported for Preproduction and Production deployments.
- Set up and migration of ETL code from Wipro data center to **AWS**.

**Environment**: Pentaho 8.3, 9.1, Oracle, Snowflake, AWS, Tableau, Linux OS.

**Warranty Solution Product| Bluebird, Premium 2000+, Bluestar | Mize** **April 2014 – March 2019**
**Data Engineer**
**Responsibilities:**

- Worked closely with the business analysts to convert the Business Requirements into Technical Requirements and preparing low and high-level documentation.
- Initiated the Customer Central product implementations using **Pentaho Data Integration and Cognos**.
- Involved in Designing the **SRS** with Activity Flow Diagrams using **UML**.
- Configured and deployed instances on **AWS** environments.
- As per as business requirements we use **Talend** to integrate the data on cloud and make it accessible to the offshore team.
- Involved in requirements gathering with help of BA's and Module leads.
- Participated in end to end activities (i.e. Star schema design).
- Created **AWS EC2** windows instances to host Pentaho ETL server.
- Taking care of Support projects deployments activity.
- Resolving Issues which were raised by customers and monitored all tickets through the Quick base tool.
- Conducting the Meetings with Business Analysts for requirements clarification.
- Attended regular SCRUM meetings which are related to tasks status, backlog grooming, participating in sprint release planning and retro.
- Implemented 24by7 ETL running process with deletion ETL's concept.
- Developed existing and new enhancement implementations using **Kettle and Cognos**.
- Calculated warranty reserves using MIS (month in service) concept. Implemented Delete ETL's jobs.
- Participated in new databases creation and will update/modify DB with alter script using DB Patch.
- Calculated earnings from service plans using MIS (month in service) concept.
- Prepared database scripts and implemented ETL's.
- Design and development of BI Reports, all these reports were based on various types of reports.

- Worked on Defect Fixing and formatting issues.
- Loaded master data and transaction data from their legacy system to our product database.
- Participated in user meetings, gathered Business requirements & specifications for the Data-warehouse design. Translated the user inputs into **ETL design docs**.
- Defined, and documented the technical architecture of the **Data Warehouse**, including the physical components and their functionality.
- Analyze and gather user requirements and create necessary documentation of their data migration.
- Designed ETL architecture to process many files and created **High-level design, low-level design documents.**
- Work alongside clients to develop strategies for migration of their business data across platforms utilizing **Microsoft SQL Server.**
- Estimate schedules for data modeling activities and complete them on time, adhering to predetermined specifications and quality standards.
- Used Informatica to extract, transform and load data from **SQL Server to Oracle databases.**
- Involved in the creation of Informatica mappings to extract data from oracle, Flat Files to load into the Stage area.
- Worked data mapping, data cleansing, program development for loads, and data verification of converted data to legacy data.
- Worked on **Master Data Management (MDM)** for maintaining the customer information and for the ETL rules to be applied.
- Building, publishing customized interactive reports and dashboards, report scheduling using Tableau server.
- Created action filters, parameters, and calculated sets for preparing dashboards and worksheets in **Tableau**.

**Environment:** Pentaho Data Integration, Talend, MySQL, Cognos 10.2.2, Tableau, AWS, EC2, Windows OS.