**Balakrishna M** 

**Email:** **raj@abntechcorp.com**

**Phone: 484-841-7747**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Professional Summary:**

* Over 13+ 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 **AWS EMR, Redshift, Lambda, Step Functions, Cloud Watch.**
* Hands on experience with different AWS services like **S3, EC2, EMR, SNS, SQS, Lambda, Redshift, Data pipeline, Athena, AWS Glue, S3 Glacier, Cloud Watch, Cloud Formation, IAM**, AWS Single Sign-On, Key

Management Service, AWS Transfer for **SFTP**, **VPC**, **Code** **Commit**, **Code** **Build**.

* Adapt at **implementing E2E solutions** on **Big** **Data** using **Hadoop** **framework**, executed, and designed big data solutions on multiple distribution systems.
* Developed and deployed serverless applications using **AWS** **Lambda**, enabling efficient and cost-effective execution of code in response to various triggers and events.
* Managed and optimized **S3** **buckets** for scalable and secure storage of data, implementing versioning, encryption, and lifecycle policies to meet data governance requirements and optimize costs.
* Designed and implemented data processing pipelines using Amazon **EMR** (**Elastic** **MapReduce**) and **Apache** **Hive**, enabling distributed data querying and analysis on large datasets stored in S3.
* Configured and managed real-time data streaming solutions using **Amazon** **Kinesis**, facilitating the collection, processing, and analysis of streaming data from diverse sources for real-time insights and decision-making.
* Leveraged **AWS** **Lambda** to automate data processing tasks and workflows, integrating with **S3** and **EMR** to orchestrate batch and streaming data processing jobs efficiently.
* Utilized **AWS** **Glue** for data cataloging and **ETL** (**Extract**, **Transform**, **Load**) processes, enabling seamless integration and transformation of data between different data sources and target systems.
* Implemented scalable and fault-tolerant architectures on AWS, leveraging services like **S3**, **EMR**, and **Lambda** to ensure high availability and reliability of data processing and analytics pipelines.
* Designed and implemented security controls and best practices for **AWS** **services**, **including** **IAM** **policies**, **encryption**, and **access** **controls**, to ensure data confidentiality, integrity, and availability.
* Implemented monitoring and alerting solutions using **Amazon** **CloudWatch**, enabling real-time visibility into AWS resources and applications to detect and respond to operational issues proactively.
* Designed and implemented disaster recovery strategies using AWS services like **S3** **cross**-**region** replication and **EMR** cluster snapshots, ensuring data resilience and business continuity in the event of outages or disasters.
* Used Python to work with JSON files to store data for testing Django websites.
* Designed and implemented data warehousing solutions using **Snowflake**, **leveraging** its scalable architecture and separation of compute and storage layers for efficient data storage and analytics.
* Orchestrated data processing pipelines using **Kafka** **Connect** and **Kafka** **Streams**, enabling reliable and scalable data movement and transformation between Kafka topics and downstream systems.
* Expertise in Big Data processing using Hadoop, Hadoop Ecosystem (**Map Reduce, Spark, Scala, Hive, HBase, Mongo DB**) implementation, maintenance, ETL and Big Data analysis operations.
* Designed and optimized data pipelines using **Apache** **Hive** for **batch** **processing,** data transformation, and analytics on large-scale datasets.
* Developed **HiveQL** **queries** and scripts for data extraction, transformation, and loading (ETL) processes from various data sources into **Hive** **tables**.
* Architected and implemented data lake solutions using Apache **Hudi** for data ingestion, transformation, and storage on Apache Hadoop or cloud-based platforms like **Amazon** **EMR** or Azure HDInsight.
* Integrated **Apache** **Iceberg** with Apache Spark or **Apache** **Flink** for data processing, transformations, and analytics on **Iceberg**-**managed** **tables**.
* Developed ETL (Extract, Transform, Load) processes using **Apache** **Iceberg** to ingest, transform, and load data from various sources into Iceberg-managed tables, ensuring data consistency and data versioning.
* Integrated **Databricks** **Delta** **Lake** with **Apache** **Spark** for data processing, transformations, and analytics on **Delta** **Lake** **tables**, leveraging ACID transactions and data versioning.
* Experience in working on different big data warehouses in **Snowflake** and **star** schemas.
* Involved in designing the data model in Hive for migrating the **ETL** process into **Hadoop** and to load data into **Hadoop** environment.
* Hands-on use of Spark and Scala APIs to compare the performance of Spark with Hive and SQL, and Spark SQL to manipulate Data Frames in Scala
* 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.
* Proficiency in SQL across several dialects (we commonly write MySQL, PostgreSQL, Redshift, SQL Server, and Oracle).
* Worked on **AWS** **Redshift** and **RDS** for implementing models and data on **RDS** and **Redshift**.
* Expertise in **Creating**, **Debugging**, **Scheduling** and **Monitoring** jobs using **Airflow**.
* Experience implementing Cloud based **Linux OS in AWS** to Develop Scalable Applications with **Python.**
* Experience on Shell scripting to automate various activities.
* Experience in working with NoSQL databases like HBase and Cassandra
* Architected cloud-native solutions using **Docker** containers and **Kubernetes** on **AWS**, **leveraging** services like **AWS** **EKS**, **Amazon** **ECS**, **AWS** **Fargate**, and **AWS** **Lambda** for containerized workloads.
* 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)**.
* Experience in developing workflows using Flume Agents with multiple sources like Web Server logs, **REST API** and multiple sinks like **HDFS sink**.

**Technical Skills:**

|  |  |
| --- | --- |
| **Big Data Technologies** | Hadoop, HDFS, Hive, HBase, Spark, Flume. |
| **Databases** | My SQL, Oracle, Microsoft SQL SERVER, Microsoft Azure SQL,  |
| **NoSQL Databases** | MongoDB, HBase, Dynamo DB, Oracle NoSQL Database |
| **Integration Tools** |  Informatica, Autosys |
| **Version Control** |  Git hub, Jira |
| **Languages** | Python, R, SAS, Teradata, MySQL, NoSQL, PySpark, Scala, Spark SQL |
| **Visualization Tools** | Tableau, Power BI, Matplotlib and QlikView/ Qlik sense, IBM DB2 |
| **Cloud Platforms** | AWS  |

**Education:**

* Masters in software engineering from International Technological University, 2016. San Jose, CA
* Bachelor of Technology in Electronics and Communication from Kakatiya University, 2007

**Work Experience:**

**HHSC |Dallas, Texas Nov 2021 – Present**

**Sr Data Engineer**

Responsibilities:

* Participated in all phases including Analysis, Design, Coding, Testing and Documentation and gathered requirements and performed Business Analysis.
* Developed **ETL** (**Extract**, **Transform**, **Load**) processes using **AWS** **Glue** to automate data ingestion and transformation workflows, improving data quality and efficiency.
* Utilized **AWS** **Glue** for **schema** **discovery** and **data** **cataloging**, facilitating metadata management and improving data governance practices.
* Implemented data pipelines using **AWS** **Glue** and **S3** to **ingest**, **process**, and transform large volumes of data for analytics and reporting purposes, improving decision-making processes.
* Configured **AWS** **Redshift** **external** **tables** for seamless integration with external data sources, enhancing data accessibility and analysis capabilities.
* 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
* Build data pipelines in Airflow/Composer for orchestrating ETL related jobs using different airflow operators
* Designed and implemented serverless architectures using **AWS** **Lambda** to execute code in response to events, reducing operational overhead and optimizing resource utilization.
* Managed and optimized **S3** **buckets** for efficient storage and retrieval of data, ensuring high availability, durability, and security of stored assets.
* Proficient in designing and implementing **Event** **Bridge** pipelines to **ingest**, **transform**, and **route** **events** from various sources such as applications, **AWS** **services**, and third-party systems.
* Experienced in integrating **Event** **Bridge** with AWS services such as **AWS** **Lambda**, **Amazon** **SNS**, **Amazon** **SQS**, **Amazon** **S3**, and **AWS** **Step** **Functions** to **orchestrate** event-driven architectures and serverless workflows.
* Integrated AWS Lambda with **AWS** **DynamoDB** to build real-time data processing pipelines, enabling low-latency data ingestion and analysis for time-sensitive applications.
* Orchestrated complex workflows using **AWS** **Step** **Functions**, enabling the coordination of multiple **AWS** **services** and microservices into scalable and reliable applications.
* Created Presentations for GIT, SQL and Python
* Leveraged **AWS** **Step** Functions for **workflow** **automation** and state management in serverless applications, enabling error handling, retries, and parallel processing of tasks.
* Automated infrastructure provisioning and configuration management using **AWS** **Cloud Formation** templates, ensuring consistency and repeatability of deployments across environments.
* Implemented serverless **CI/CD** pipelines using AWS **Code Pipeline** and AWS **Code Build**, enabling automated build, test, and deployment workflows for cloud-native applications.
* Leveraged **AWS** **IAM** **policies** and roles to enforce least privilege access control and ensure the security and compliance of AWS resources and services.
* Using Tableau/OBIEE, we wrote data and loaded it into a Data Lake environment (SNOWFLAKE) that was accessed by business users and data scientists
* Implemented serverless monitoring and logging solutions using AWS **CloudWatch**, enabling real-time visibility into application performance, resource utilization, and operational metrics.
* Designed and implemented real-time data ingestion pipelines using **Amazon** **Kinesis** **Data** **Streams** to process large volumes of streaming data.
* 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
* Streaming data analysis using Dataflow templates by leveraging Cloud Pub/Sub service
* Configured data **partitioning**, **sharding**, and **replication** within **Kinesis** **Data** **Streams** for efficient data processing and high availability.
* By creating a customized read/write Snowflake utility function in Scala, data was transferred from an AWS S3 bucket to Snowflake.
* Integrated Kinesis **Data** **Streams** with **AWS** **Lambda** functions and Amazon Kinesis Data Analytics for real-time data processing, transformation, and analytics.
* Experience in using various operators in composer/airflow and have using the google cloud client libraries in python for big query & storage hooks.
* Designed and implemented data delivery pipelines using **Amazon** **Kinesis** **Data** **Firehose** to efficiently ingest, transform, and deliver streaming data to destinations such as **Amazon** **S3,** **Amazon** **Redshift**.
* Configured data transformation, compression, and encryption settings within **Kinesis** **Data** **Firehose** to optimize data storage, reduce costs, and ensure data security during transit and at rest.
* Integrated **Kinesis** **Data** **Firehose** with AWS Lambda functions for custom data transformations, validations, and enrichment before delivering data to target destinations.

Environment**:** AWS Lambda, Redshift, Glue, Step Function, Event bridge**,** SQL**,** Kafka,S3, EMR,Hadoop, Hive, HDFS, PySpark, Apache Spark, Oracle, Python, Restful web service,Kinesis.

**State street, Charlotte, NC Aug 2019 – Oct 2021**

**Sr Data Engineer**

Responsibilities:

* Participated in all phases including **Analysis, Design, Coding, Testing and Documentation** and **gathered requirements** and performed Business Analysis.
* Developed Entity-Relationship diagrams and modeling **Transactional Databases** and Data Warehouse using ER/ Studio and Power Designer.
* Maintained data pipeline up-time of 99.9% while **ingesting streaming and transactional data** across 7 different primary data sources using **Spark, Redshift, S3, and Python**.
* **Ingested data from disparate data sources** using a combination of **SQL**.
* **Google Analytics API, and Salesforce API** using **Python to create data views** to be used in BI tools like **Tableau.**
* Designed and implemented data lake solutions using **Hive** **external** **tables** on **Amazon** **S3**, providing efficient storage and query capabilities for large-scale datasets.
* Managed data **schemas**, partitioning strategies, and metadata for **Hive** **external** **tables** on **S3** to optimize data organization, query performance, and data governance.
* Integrated **Hive** **external** **tables** with other data processing frameworks and **ETL** tools for data transformations, aggregations, and analytics.
* Working with two different datasets one using **HiveQL**
* Experience with AWS services related to AI/ML highly desirable, particularly EMR, AWS Lambda, Sage maker, Kinesis, Machine Learning, Lex, Polly, Recognition, DynamoDB, S3.
* Writing **MapReduce code using python** to get rid of certain security issues in the data.
* Synchronizing both the **unstructured and structured data using Pig and Hive** on business prospectus.
* Used Pig Latin at **client-side cluster and HiveQL at server-side cluster**.
* Having experience working on **Google Cloud Big data** Technologies like Data Proc, Data Flow, Big Query and GCP Storage, and having knowledge on pub sub.
* Process and load bound and unbound Data from Google pub/sub topic to big query using Cloud Dataflow with Python.
* Wrote scripts and indexing strategy for a migration to Confidential Redshift from SQL Server and MySQL databases
* Importing the complete data from **RDBMS to HDFS** cluster using Sqoop.
* Worked on AWS Redshift and RDS for implementing models and data on RDS and Redshift.
* Creating external tables and moving the data onto the tables from managed tables.
* Performing the subqueries in **Hive and partitioning and bucketing** the imported data using HiveQL.
* Created monitors, alarms, and notifications for **EC2 hosts using Cloud Watch, Cloud trail and SNS.**
* As per as business requirements we use Talend to integrate the data on cloud and make it accessible to the offshore team.

Environment**:** Hadoop, Alteryx, Hive, HDFS, PySpark, HBase, Apache Spark, Scala, Oracle, Python, Scala, GCP, Restful web service.

**First American, Princeton, NJ Mar 2018 – Jul 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.
* Database design, data modeling and maintenance using **SQL**/Snowflake
* Maintain existing database and pipelines making sure data is being refreshed according to set frequency
* Maintain scripts / job queues in sync with IT updates on server etc.
* Maintain access management to dashboards / SharePoint sites
* Implement support structure for underlying **DB query scripts**, job queues, usage tracking etc.
* Leverage business and analytics acumen to deliver insights, root causes, and corrective actions
* Perform analysis on interesting observations to identify insights, possible root causes, and corrective actions
* Developed views and templates with **Python** view controller and templating language to create a user-friendly website interface.
* Wrote **HIVE UDF's** as per requirements and to handle different schemas and xml data.
* Implemented **ETL** code to load data from multiple sources into **HDFS.**
* Developed **data pipeline using Python**, Hive to load data into data link. Perform data analysis data mapping for several data sources.
* Loaded data into S3 buckets using **AWS** **Glue** and **PySpark**. Involved in filtering data stored in **S3** **buckets** using **Elasticsearch** and loaded data into **Hive** **external** **tables**.
* Used **AWS EMR** to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service **(Amazon S3) and Amazon DynamoDB.**
* Designed new Member and Provider booking system which allows providers to book new slots, with sending out the **member leg and provider Leg directly to TP** through Datalink.
* Analyze various type of raw file like **Json, Csv, Xml with Python** using Pandas, NumPy etc.
* Automated the existing scripts for performance calculations using scheduling tools like **airflow**.
* Designed and developed the core data pipeline code, involving work in **Java** and **Python.**
* Good knowledge on **Partitions, bucketing** concepts in **Hive** and designed both Managed and **External tables** in Hive for optimized performance.
* Hands on experience on fetching the live stream data from DB2 to **HBase** table using **Spark Streaming**.
* Worked on **NoSQL** databases including **HBase.**

Environment**:** Map Reduce, HDFS, Hive, HBase, Python, SQL, Scala, Spark, Apache Kafka, Play, AWS.

**Bank of West |San Ramona, CA Aug 2016 – Mar 2018**

**Data Engineer**

Responsibilities:

* 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 a large number of 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.
* Rewrote SSIS packages in Spark using Spark SQL, Pyspark and executed in AWS EMR
* 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 also 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.
* Designed and Created data cleansing, data conversion, validation, and External loading scripts for **Teradata warehouse** using **Informatica ETL tool.**
* Involved in error handling, performance tuning of mappings, testing of **Stored Procedures and Functions**, Testing of **Informatica Sessions**, and the **Target Data.**
* Possess expertise with relational database concepts, stored procedures, functions, triggers, and scalability analysis. Optimized the SQL and PL/SQL queries by using different tuning techniques like using hints, parallel processing, and optimization rules.

Environment**:** Informatica, Oracle 11g, SQL Server 2008, SPSS, PL/SQL, Teradata, UNIX Shell Scripting, Visual Studio, Tableau, MS Excel, MS Access, Windows XP.

**Careator Technologies Pvt Ltd Hyderabad, India Nov 2012 to May 2015**

**Role: Big Data Engineer**

Responsibilities:

**•** By constructing a data pipeline with Sqoop, Hive, and PySpark, the Data Lake in Hadoop was

 maintained.

• Participated in the development of a data pipeline and carried out analytics utilizing the AWS stack

 (EMR, EC2, S3, RDS, Lambda, Kinesis, Athena, SQS, Redshift, and ECS).

• Prior to importing the data into HBase, Spark was used to process it. Scala was used to develop real-

 time spark jobs as well as batch jobs.

• Used PySpark to read and write data from HDFS in a variety of formats, including JSON, ORC, and

 Parquet.

• Used Python to work with JSON files to store data for testing Django websites.

• Created shell scripts for post-processing to streamline the file transfer process.

• Used Python to work on data cleaning and pre-processing in order to do feature engineering and data

 imputation techniques for the dataset's missing values.

• Developed Data Quality Scripts with SQL and Hive to verify the quality of the data and the success of

 the das ta load.

• Made several data visualizations using Tableau and Python.

• Setup Python Boto3 to enable CLI access to a few AWS services.

• In charge of creating scalable distributed data solutions utilizing Amazon EMR's EMR cluster

 environment.

• Proficient at using Spark Streaming with Kafka as a data pipeline technology to write core jobs and

 real-time processing applications.

• Improved Map Utilize various compression techniques to reduce jobs to make the most of HDFS.

• Used Spark's in-memory processing power to manage big datasets on S3 Data Lake.

• Involved in the Teradata to Snowflake Object Migration.

• Built a Snow pipe for continuous data loading and used NiFi to schedule various Snowflake tasks.

• Involved a lot of Hive effort, including creating the tables and loading data into them using Spark

 Streaming to consume event data from Kafka.

• Created an admin API to control and see brokers, topics, and other Kafka objects.

• Participated in the development, construction, testing, and deployment to a distributed Hadoop

 cluster.

• Built a common learner data model that uses Spark Streaming APIs to conduct transformations and

 actions instantly, obtaining data from Kafka in almost real-time and persisting it to Cassandra.

• Spark code and Spark-SQL/streaming were developed to speed up testing and data processing.

• Spark application development experience with RDD and Data frames.

• Created a Python script to compare daily files from one S3 bucket to daily files from another bucket.

• Added a PySpark, SQL module that gives the Spark session access to efficient data queries.

• Utilized the Spark/Data bricks framework and Stream Sets to implement close to real-time data

 processing.

• Made Hive queries that, by contrasting recent data with reference tables and historical measures,

 enabled market analysts to identify emerging trends.

• Thorough familiarity with configuration management platforms like Bit bucket/GitHub and Bamboo

 (CICD).

Environment: HDFS, Hive, Scoop, Pig, Oozie, Amazon Web Services (AWS), Python.

**Client: SMS Software Private Limited, India. Jun 2009 – Nov 2012**

**Role: Python Developer**

Responsibilities:

* Designed front end and backend of the application using Python on Django Web framework.
* Used HTML, CSS, AJAX, JSON designed and developed the user interface of the website.
* Developed views and templates with Python and Django view controller and template language to create a user-friendly website interface.
* Used JavaScript and JSON to update a portion of a webpage.
* Used Java Persistence API (JPA) for managing relational data mapping
* Handled on issues related to conversion of Java to AJAX.
* Developed frontend and backend modules using Java, Spring MVC Web Framework
* Developed consumer-based features and applications using Python, Django, HTML and Test-Driven Development (TDD).
* Rewrite existing Python/Django modules to deliver certain format of data.
* Build SQL queries for performing various CRUD operations like create, update, read and delete.
* Improved the coding standards, code reuse. Increased performance of the extended applications by making effective use of various design patterns.
* Worked on a large-scale distributed computing environment, monitoring data nodes prioritize jobs for processing function.
* Worked extensively with Bootstrap, JavaScript, and jQuery to optimize the user experience.
* Used Python and Django to interface with the jQuery UI and manage the storage and deletion of content.
* Carried out various mathematical operations for calculation purpose using python libraries.
* Skilled in using Collections in Python for manipulating and looping through different user defined objects.
* Engaged in Design, Development, Deployment, Testing and Implementation of the application.

Environment: Python 2.7, Django 1.6, HTML, CSS, XML, MySQL, JavaScript