**Krupa**

**Sr. Big Data Engineer**

**Mail: shivalika943@gmail.com**

**Phn: (210) 960-3692**

**SUMMARY**

* 9 years of IT industry experience in dealing with Apache Hadoop components like HDFS, MapReduce, Databricks Spark, Hive, Pig, Sqoop, Oozie, Cassandra, MongoDB, Amazon Web Services and Google Cloud Platform.
* Strong hands-on experience with AWS services, that include EMR, S3, EC2, RDS, Glue, DynamoDB, Athena and GCP services include Big Query, Cloud SQL, Cloud Functions, Cloud Composer, Cloud Storage and Dataproc.
* Strong experience and knowledge of NoSQL databases such as MongoDB, DynamoDB, and Cassandra.
* Experience in ingesting the data using Sqoop from HDFS to Relational Database Systems like Oracle, SQL Server and vice-versa.
* In-depth experience in automating various Databricks Spark, Hive, Bash and Scala applications using Airflow and Oozie.
* Installed and configured Apache Airflow for workflow management and created workflows in python, created the DAG’s using Airflow to run jobs sequentially and parallelly.
* Good experience in Tableau for Data Visualization and analysis on large data sets, drawing various conclusions.
* Tune the performance of the Spark applications to efficiently run the ER model on a national level. The performance tuning includes tracking the spark metrics, maintaining the costs, and re-designing the low performing code.
* Created HBase tables to load large sets of structured, semi-structured and unstructured data coming from UNIX, NoSQL, and a variety of portfolios.
* Worked on AWS Cloud – used S3 buckets for storing the data, EMR – for spinning up the cluster and running the spark jobs, Athena – for creating external tables, IAM – for authentication, AWS GLUE – to get data from sources.
* Experience on working with Structured, Unstructured data with various file formats such as CSV, Avro, Parquet, JSON, XML, ORC, Sequence.
* Experience in writing Python custom parsers for JSON and XML formatted data.
* Strong working experience with SQL and NoSQL databases, data modeling and data pipelines. Involved in end-to-end development and automation of ETL pipelines using SQL and Python.
* Proficiency in developing SQL with various relational databases like Oracle, SQL Server for Support of Data Warehousing and Data Integration Solutions.
* Have experience of working on Spark logs to identify skewed data, improve shuffle performance and join performance.
* Develop effective working relationships with client teams to understand and support requirements,
* develop tactical and strategic plans to implement technology solutions, and effectively manage client expectations.
* Flexible, enthusiastic, and project-oriented team player with solid communication and leadership skills to develop creative solutions for challenging client needs.
* Experienced in using Agile methodologies including extreme programming, SCRUM and Test-Driven Development (TDD).

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| Programming skills | Python, Scala, Java. |
| Hadoop | HDFS, MapReduce, Yarn, PySpark, Spark-Scala, SparkSQL, Hive, HBase, Airflow, Oozie, Sqoop, Hue, ZooKeeper, Jupyter. |
| Operating Systems | Windows, Unix, Linux, Ubuntu. |
| Cloud Technologies | AWS, S3, EMR, EC2, Lambda, Redshift, Glue, Athena, RDS, GCP, BigQuery, Cloud Functions, Cloud SQL, Cloud Storage, Cloud Composer, Dataproc. |
| Reporting Tools | Tableau, PowerBI. |
| Databases | MySQL, MongoDB, Cassandra, Teradata, DynamoDB, Oracle, PostgreSQL. |
| IDE Tools | Eclipse, IntelliJ, NetBeans. |
| Version Controls | GIT, SVN. |

**EDUCATION:** Bachelor of Technology - Computer Science - JNTUH, 2014

**PROFESSIONAL EXPERIENCE**

**Gathi Analytics / Fiser** Brookfield, WI **Dec 2021 – present.**

**Sr. Big Data Engineer**

Description: Mazda North American Operations, which includes Mazda Motor of America, Inc., is Mazda Motor Corporation’s North American arm, and constitutes the largest component of that company outside Japan. The company has its headquarters in Irvine, California.

Responsibilities:

* Worked extensively on Databricks Spark, RDDs, Spark SQL and Datasets and ability to manage the development independently.
* Worked closely with the Mazda data and developed a thorough grasp of the automotive industry.
* Successfully oversaw the creation, rollout, and upkeep of Mazda’s incentive program for North American dealers.
* A solid understanding of how to build complicated data pipelines utilizing transformations, aggregations, cleaning, and filtering.
* In-depth familiarity with and practical expertise using Apache Hadoop’s HDFS, Hive, and Sqoop components.
* Well versed with object storage capabilities - S3 for AWS and extensive working knowledge of using Amazon S3 for daily ingestions and writing file-based triggers.
* Understanding of AWS Lambda and experience developing applications that use Lambda to directly trigger S3 and Cloud Watch.
* Create a Pyspark application to track the progress of file ingestions and scheduled runs in the data lake and provide a daily monitoring report.
* Ability to analyze massive data collections that are both structured and semi-structured. Data in Text and Avro file formats were processed, and transformations were carried out using Spark RDD/Dataframes.
* Experience in converting data from a custom format to CSV format. Additionally, solid expertise in utilizing Spark tasks to clean such data.
* Have extensive experience in development of PySpark apps both via Jupyter and IDEs. Familiarity with loading and manipulating a variety of datasets, including CSV, JSON, Parquet, and Avro.
* Perform Hive querying on an ad-hoc basis to resolve customer queries and to generate reports as required.
* Experience running Hive queries using the Hive shell, Impala, and Hue. Extensive knowledge in writing complicated Hive queries using joins, nested queries, aggregations, and window functions.
* Good knowledge of Hive partitioned tables, map join, bucketing and gained a good understanding of dynamic partitioning.
* Extensive knowledge in using Airflow and Oozie to automate PySpark, Hive, Bash, and Python applications.
* Experience in importing and exporting the data from teradata to HDFS using Sqoop.
* Acquired understanding of Mazda’s internal products and actively participated in the Mazda Dealer Communication and Incentives systems.
* Performed integration testing and unit testing following development and took part in code reviews.
* Work with business analysts to convert the needs from the business world into the world of technology.
* Collaborate on the design and implementation of technical requirements with a variety of technical specialists, architects, and developers.

Environment: PySpark, Airflow, Databricks Spark 2.2, Python 3, Scala 2.10.5, MongoDB, Apache Hive, HDFS, Oozie, AWS, Maven, IntelliJ, GIT

**McKesson Corporation** Irving, Texas **Marc 2019 – Nov 2021**

**Big Data Engineer**

Description: Wells Fargo; Company is an American multinational financial services company headquartered in San Francisco, California, with central offices throughout the United States. It is the world’s fourth-largest bank by market capitalization and the fourth largest bank in the US by total assets.

Responsibilities:

* Strong understanding and practical experience in developing Spark applications with Python.
* Developed Scala scripts, UDFs using both Data frames/SQL in Spark for Data Aggregations.
* Exploring with Spark for improving the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL, Data Frame, and pair RDDs.
* Experience in importing and exporting data using Sqoop from Relational Database Systems to HDFS and vice-versa.
* Knowledge in converting Hive or SQL queries into Spark transformations using Python.
* Experience in creating Sqoop jobs with incremental load to populate Hive External tables.
* Experience in migrating Python and Shell scripts to Spark-Scala code to improve the performance.
* Good experience in writing shell scripts to support additional functionality for the application.
* Good experience in storing metadata in SQL, to track the status of ingestion in data lake.
* Developed Spark, Python, and shell applications to track the small files in the cluster and merge them.
* Created a statistics report that depicts the number of small files merged and current memory space occupied.
* Strong expertise in internal and external tables of HIVE and created Hive tables to store the processed results in a tabular format.
* Firm knowledge on HDFS commands to perform basic to advanced activities.
* Provide production support to resolve the issues for the Java applications deployed.
* Strong knowledge on the process of creating complex data pipelines using transformations, aggregations, cleansing and filtering.
* Good understanding of Partitioning, Bucketing, Join optimizations and query optimizations in Hive.
* Good experience in communicating with offshore teams with daily status calls.
* Experience in dealing with production issues and good exposure to the banking domain.
* Practical experience in developing Spark applications in IntelliJ with Maven.
* Involved in loading data from the UNIX file system to HDFS and developed UNIX scripts for job scheduling, process management and for handling logs from Hadoop.
* Developed test cases for Unit testing and performed integration and system testing.

Environment: Hadoop 2.7.3, Java 1.7, Spark 1.6.0, SparkSQL, Python, Scala 2.10.5, MongoDB, Apache Pig 0.12.0, Apache Hive 1.1.0, HDFS, Sqoop, Maven, IntelliJ, UNIX Shell scripting, Oracle 11g/10g, Linux, SVN.

Cross Sense Analytics, Farmington Hills, MI Dec 2018 – April 2019

Sr Big Data Engineer/ Engagement Consultant

## Responsibilities:

* Design, Plan, Implement and Responsible for Sales, Finance Data pipelines to Data Lake (S3) and Data Warehouse (Redshift),
* Process Data, using EMR and
* Participate in the design, architect and implementation of CCPA on S3 and redshift.
* Write code in Java for the pipelines to ingest data from different sources to flow through S3 and Redshift.
* Deploy code to dev, int and prod.
* Collaborate with different teams to communicate, negotiate and implement end to end solutions.
* Use EMR to process heavy load batch processing.
* Support BI team for analytics reports.
* Implemented system wide monitoring and alerts.
* Installed & configured Hive, Impala, Oracle BigData Discovery, Hue, Apache Spark, Tika, Tika Tesseract, Sqoop, Spark sql etc.
* Importing and exporting data into MapRFS and Hive using Sqoop.
* Used Bash Shell Scripting, Sqoop, AVRO, Hive, Impala, HDP, Pig, Java, Map/Reduce daily to develop ETL, batch processing, and data storage functionality.
* Responsible for developing data pipeline using Flume, Sqoop and Pig to extract the data from weblogs and store in HDFS.
* Worked on loading all tables from the reference source database schema through Sqoop. Worked on designed, coded and configured server side J2EE components like JSP, AWS and JAVA.

Environment: AWS Redshift, S3, Java Spring, PostgreSQL, MS SQL, Python, AWS EMR, Github, Jenkins, Veracode, Scala, Talend, Jaws

**Carelon global solutions, India | April 2016 – Nov 2017**

**Data Engineer**

Description: Carelon Global Solutions (CGS) is a healthcare solutions company that is simplifying complex operational processes to improve the health of the healthcare system.

Responsibilities:

* Designed, built, and maintained **CI/CD** data pipelines from various sources into a **data warehouse**.
* Developed **SQL** **queries** to support the extraction and transformation of data for business reporting needs.
* Performed Optimizations and performance tuning in **Spark** and **Hive**.
* Used **Apache** **Airflow** to programmatically author, schedule, and monitor workflows.
* Used **GCS** as the repository to store the data, which then loaded into **Big Query** for analytical purposes.
* Transforming data into the appropriate format, schema, and structure to meet business requirements.
* Collaborated with other teams such as data science, business intelligence, and software engineering to meet business needs.
* Used **Cloud SQL** with **MySQL** and **SQL Server** as the database system, and utilized **GCS** as the storage solution for data integration and management.
* Worked with orchestration tools like Airflow and comprehensive understanding of **cloud Composer** in **GCP**.
* Used Google **Cloud Functions** for parsing of CSV and JSON source file to parquet file.
* Familiar with the **Google Cloud Dataflow** and can use it to create and manage batch processing jobs
* Developed and maintained data models to support data analysis and reporting by business analysts and data scientists.
* Experienced in writing scripts to import and export the data from RDBMS to **HDFS** and vice-versa.

Environment: Hadoop 2.6.2, HDFS 2.6.2, Map Reduce 2.9.0, Hive 1.1.1, GCP, Sqoop 1.4.4, Apache Spark 2.0, Nifi, ETL, Oozie, Java 7, Python 3.2,

**Tera Soft, Hyderabad, India | May2014 – March 2016**

**Python Developer**

Description: Tera Soft is a globally recognized IT Solutions & Staffing Firm that has been assisting our clients for over 20 years in maximizing their technological investments with highly effective staffing solutions. Orion’s specialized Delivery model evokes the highest caliber of talent: Onshore, Offshore, and Nearshore.

Responsibilities:

* Followed Agile methodology with daily SCRUM meetings, sprint planning, showcases and retrospective and used JIRA as a tracking tool.
* Designed and developed the UI of the website using HTML, AJAX, CSS and JavaScript.
* Designed and developed data management system using MySQL.
* Expertise working with AWS cloud services like EMR, S3, EC2, Redshift, cloud watch, cloud formation, security groups, for big data development.
* Rewritten existing Python/Django modules to deliver a certain format of data.
* Used Django Database APIs to access database objects.
* Wrote Python scripts to parse XML documents and load the data in the database.
* Expertise in writing Constraints, Indexes, Views, Stored Procedures, Cursors, Triggers, and User Defined function.
* Improved code reuse and performance by making effective use of various design patterns and refactoring code base.
* Developed entire frontend and backend modules using Python on Django Web Framework.
* Deploying, managing, and operating scale, highly available, and fault tolerant systems to AWS like SQS, S3, Kinesis, SNS.
* Developed views and templates with Python and Django view controllers and templating language to create a user-friendly website interface.
* Used a test-driven approach for developing the application and implemented the unit tests using Python Unit test framework.
* Worked on translation of web pages to different languages as per client requirements.
* Performed troubleshooting, fixed and deployed many Python bug fixes of the two main applications that were a main source of data for both customers and internal customer service team.

Environment: Python, Django, MySQL, Linux, HTML, XHTML, SVN, CSS, AJAX, Bugzilla, JavaScript, Apache Web Server, Selenium