RAJASHEKAR V

Sr. Data Engineer

+1 (919) 659-8898 rajashekarv2203@gmail.com LinkedIn: linkedin.com/in/rajashekar-v-50aa56262
Certification: DP-203

PROFESSIONAL SUMMARY:

6+ years of experience as **Big Data Engineer** with expertise on cloud technologies like **AWS**, **Azure** to implement enterprise wide **ETL** pipelines using Python, Spark,PySpark, Scala and SQL. Expert in building ETL enterprise platforms from scratch like architecting, designing, developing, and maintaining production pipelines by following industries best practices.

- Proficient with **Spark Core, Spark SQL and Spark Streaming** for processing and transforming complex data using in-memory computing capabilities written in **Pyspark.**
- In-depth knowledge of and **Pyspark** Architecture and cloud-based platforms.
- Experience in end-to-end **data engineering** including data **ingestion**, data **cleansing**, data **transformations**, data **validations**/**auditing** and **feature** engineering.
- Experience in Developing **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.
- Experienced with **Snowflake** cloud data warehouse and **AWS S3** bucket for integrating data from multiple source system which include loading nested **JSON** formatted data into **snowflake** table.
- Proficient in using **Snowflake Clone** and **Time Travel** as well as In-depth knowledge of **Snowflake Database**, Schema and Table structures.
- Expertise in AWS services including **S3, EC2, SQS, RDS, EMR, Kinesis, Lambda, Step Functions, Glue, Redshift, DynamoDB, Elasticsearch, Service Catalog, CloudWatch and IAM.**
- Worked on **ETL** Migration services by developing and deploying **AWS Lambda** functions for generating a serverless data pipeline which can be written to **Glue Catalog** and can be queried from **Athena.**
- Build pipelines in **Azure data factory** to move data from on prem to **Azure** SQL Datawarehouse, from Amazon **S3** buckets to **Azure** blob storage.
- Worked extensively on Azure Cloud Services such as Azure Data Lake Storage **(ADLS),** Azure **Synapse,** Azure **Databricks**, Azure Synapse Sql, Azure Data Factory for building Data Lakes on Azure cloud platform.
- Development level experience in Microsoft Azure providing data movement and scheduling functionality to cloud-based technologies such as **Azure Blob** Storage and **Azure SQL** Database.
- Experienced on Migrating **SQL database** to **Azure Data Lake, Azure Data Lake Analytics, Azure SQL Database, Data Bricks and Azure SQL Data warehouse** and controlling and granting database access and Migrating On premise databases to **Azure Data Lake** store using **Azure Data factory**.
- Expertise in Azure services including **Blob Storage, ADLS, ADF, Synapse, Azure SQL Server, Azure Databricks. VM, Azure HDInsight, Azure functions, Azure Event Bridge.**
- Experience in working with Hadoop ecosystem components like **HDFS**, **Map Reduce**, **Spark**, **HBase**, **Oozie**, **Hive**, **Sqoop**, **Pig**, **Flume** and **Kafka.**
- Experienced in using **Spark Context, Spark SQL, Data Frame, Pair RDD's,** and **Spark YARN.**
- Developed **Spark Applications** that can handle data from various **RDBMS** (Cloud) and Streaming sources.
- Good hands-on experiencing working with various Hadoop distributions mainly Cloudera (**CDH**), Hortonworks (**HDP**) and Amazon **EMR**.
- Expert in dealing with the **Hive** data warehouse tool, including constructing tables, portioning and bucketing data, developing and optimizing **HiveQL** queries.
- Extensively used Apache **Sqoop** for efficiently transferring the bulk data between **Apache Hadoop** and relational databases.
- Excellent understanding of NoSQL databases like **HBase**, **Cassandra**, **MongoDB**.
- Solid experience in working with csv, text, **Avro**, **parquet**, orc, **Json** formats of data.
- Hands on experience with continuous integration and automation using **Jenkins** and version control tools such as **GIT**, **SVN.**
- Expertise in development of various reports, dashboards using various **Power BI, Tableau Visualization.**
- Expertise in all phases of System Development Life Cycle Process **(SDLC),** Agile Software Development, **Scrum** Methodology and defining user stories and driving the agile board in **JIRA** during project execution.

- Have good interpersonal, communication skills, strong problem-solving skills, explore/adopt to new technologies with ease and a good team member.

TECHNICAL SKILLS:

Cloud Technologies	**AWS, Microsoft Azure, Snowflake, Data Bricks.**
Programming Languages	**Python, Scala, Pyspark.**
ETL Tools	**Informatica, DataStage, Airflow.**
Big data tools	**Spark, hive, Sqoop, Kafka, Spark, MapReduce, HDFS, PIG, Oozie, Impala, Zookeeper.**
Databases	**Oracle, Teradata, MySQL, Azure SQL, Postgre SQL, Maria DB.**
NoSQL Databases	**HBase, Cassandra, MongoDB.**
CI/CD Tools	**Jenkins, GitHub, Jira, git, GitLab, GitHub.**
Visualization tools	**Tableau, PowerBI, Matplotlib.**
IDE's	**PyCharm, Jupyter Notebooks, Visual Code.**

Education:

Masters Of Science In Information Systems **Aug 2021- Dec 2022**
Northwest Missouri State University **GPA - 3.8**

Professional Experience

Blue Cross Blue Shield. **Sep'2021- till date**
Sr. Data Engineer
Responsibilities:

- Developed data processing tasks using **PySpark** such as reading data from external sources, merging the obtained data, performing data enrichment, and loading into data warehouses.
- Performed the transformations and actions on the imported data from **AWS S3** using PySpark.
- Developed robust and scalable data integration pipelines to transfer data from S3 bucket to **Redshift** database using **Python** and **AWS Glue**.
- Scheduled **Airflow DAGs** to run multiple **Hive** and **Spark** jobs, which independently run with time and data availability.
- Scheduled Apache **Airflow DAGs** to export the data to AWS S3 buckets by triggering to invoke an AWS **lambda** function.
- Designed the **snowflake** schema to integrate the tables in **snowflake** for over 100's of tables from the existing databases.
- Using **Informatica** Cloud **Data Integration** to integrate healthcare data from disparate sources, such as remote patient monitoring devices and patient portals, into a central repository hosted on **AWS**. Support real-time analytics and inform clinical decision-making, such as identifying patients who are at risk for complications or adverse events.
- Implemented
- **Informatica PowerCenter** to extract and transform patient data from various sources, such as electronic medical records **(EMRs)** and claims systems. Load the transformed data into a centralized data warehouse hosted on AWS. Analyze patient health outcomes, identify trends and patterns in healthcare utilization, and support population health management initiatives. Automate data validation and cleansing processes to ensure data accuracy.
- Developed **API** integration pipeline for the data transformation and cleansing from the web **api's** for the faster and access of data to the stakeholder using the **AWS glue** and **Pandas.**
- Implemented the error handling and logging mechanisms to capture and handle the **API's** errors, data inconsistency due to low network connectivity using the **Splunk**.
- Designed Spark-based real-time data ingestion and real-time analytics and implemented **AWS Lambda** functions to drive real-time monitoring dashboards from the system.
- Responsible for building scalable distributed data solutions using an **EMR** cluster environment with Amazon **EMR**.

- Created alarms and notifications for **EC2** hosts using **Cloud Watch, Cloud Trail**, and **SNS.**
- Created **S3** buckets managing policies for S3 buckets and Glacier for storage and backup on AWS.
- Collected the data from the edge device databases, exported in **CSV** format, and stored in **AWS S3** buckets.
- Developed **Lambda** functions to create ad-hoc tables to add schema and structure to data in S3 and performed data validation, filtering, sorting, and transformations for every data change in a **Dynamo DB** table and load the transformed data to **Postgres database**.
- Experienced in developing 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.
- Developed **MapReduce** that extract, transform, and aggregate data from a variety of file formats, including **XML, ISON, CV, Avro,** and other compressed file formats and process **Avro, Parquet** files to map-side joins.
- Optimized the report and dashboard performance by tuning and implementing optimization strategies.
- By using CI\CD tools like **Git** for automate the building, testing and deployment of code changes. And deploying them on for production if they pass the test cases.
- Created reports with complex calculations, designed dashboards for analyzing **POS** data and developed visualizations, and worked on Ad-hoc reporting using **Tableau**. Inventing new metrics, conducting root cause analysis, testing hypotheses, and developing visualization solutions using Tableau and.

Environment: PySpark, Python, Airflow, Snowflake, Informatica, AWS Lambda, AWS EMR, AWS EC2, AWS S3, AWS Redshift, AWS Glue, SQL, Hive, Git, Tableau.

Fidelity Investments **July 2019 -June 2021**
Data Engineer
Responsibilities:

- Developed the **spark** applications to perform various data cleansing validation, transformation, and summarization activities according to the requirement.
- Responsible for developing data pipelines involving ingesting raw Json files, transactional and user profile information from on prem data warehouses and processing them using **spark.**
- Developed **Spark** applications for data extraction, transformation, and aggregation from numerous file formats for analysis using **Pyspark** and **Spark-SQL**.
- Written **Kafka** producers for streaming real time Json messages to Kafka topics and processed them using **spark streaming** and performed streaming inserts to Synapse SQL.
- using **Informatica** successfully integrated the financial data from various sources, such as banking systems, trading platforms, and accounting software, into a centralized data repository & storage hosted on Azure.
- By using **Informatica**, the financial data such as **CRM** systems and billing software, into a central data repository & storage in Azure for supporting the real-time financial analytics and inform business **decision-making**, such as identifying new revenue streams or optimizing financial operations.
- Created **Azure Databricks** notebooks with SQL, Python, and notebooks that are automated with jobs.
- Automated launch of **Azure Databricks** Runtimes and autoscaling the clusters and submitted **spark** jobs to **Azure Databricks clusters**.
- Deployed an **Azure Databricks** workspace to an existing virtual network that has public and private subnets and properly configured network security groups.
- Involved in automation of Azure Cloud Infrastructure and deployment of Data pipelines to **Azure Data Factory**.
- Worked **Azure SQL** as external **hive** meta store for **Azure HD Insight** clusters so that metadata is persisted across multiple clusters.
- Expertise in Migrating SQL database to **Azure Data Lake, Azure data lake Analytics, Azure SQL Database, Data Bricks, and Azure SQL** data warehouse Controlling and Providing database access, and Migrating On-premises databases to Azure Data Lake Store via Azure Data Factory **(ADF).**
- Developed the scalable Pipelines in **ADF** utilizing Linked Services/Datasets/Pipeline/ in order to Extract, Transform, and Load data from many sources such **Azure SQL, Blob storage,** Azure SQL Data warehouse, write-back tool, and backward.
- Ingested Data to Azure Services like **Azure Data Lake, Azure blob Storage, Azure SQL** Datawarehouse and processing the data with **Azure Databricks**.
- Utilized **Spark-Synapse Sql** Connector for writing the processed data from spark to Synapse SQL directly.

- Designed and maintained interactive reports in **Power BI**, built on top of Azure **Synapse/Azure Data Warehouse, Azure Data Lake, Azure SQL**. As a Power BI admin creating workspaces, designing security including row level security for various reports.
- Utilized **ADLS** as Data Lake and ensured all the processed data is written to ADLS directly from spark and **hive** jobs.
- Developed daily process to do incremental import of data from **Teradata** into **Hive** tables using **Sqoop.**
- Involved in creating centralized **Data Lake** on Azure Cloud Platform.
- Produced and provisioned several **Azure Databricks** clusters required for batch and continuous streaming data processing and deployed the necessary cluster libraries.
- Worked extensively on performance tuning of **Spark application** to improve job execution times and troubleshooting failures.
- Worked on different file formats like **Text, Avro, Parquet, JSON** and Flat files using Spark.
- Experience in using to analyze data from multiple sources and creating reports with Interactive Dashboards using **power BI.**

Environment: Spark, python, Kafka, Infromatica, Azure Databricks, Azure Data Factory, Azure SQL, Azure Blob storage, Azure HDInsight, Azure Synapse Analytics, Hive, Teradata, Power BI, Git.

Virtusa. **Jan 2018 – June 2019**
Big Data Developer

Responsibilities:

- Involved in importing and exporting data between **Hadoop** Data Lake and Relational Systems like **Oracle**, **MySQL** using **Sqoop**.
- Using **DataStage** extracted data from various financial data sources such as transactional databases, trade platforms, market data feeds. The extracted data then loaded into a centralized data warehouse like **oracle** where it can be analyzed and used for reporting and analysis.
- Integrated data using **DataStage** from multiple sources such as customer data, financial data, and regulatory data, and transform it into a standardized format that can be easily stored in **RDDs** for future analysing purpose.
- Involved in developing **spark** applications to perform ELT kind of operations on the data.
- Modified existing **MapReduce** jobs to Spark transformations and actions by utilizing **Spark RDDs**, Data Frames and **Spark SQL API's**.
- Utilized Hive partitioning, Bucketing and performed various kinds of **joins** on **Hive** tables.
- Involved in creating **Hive** external tables to perform **ETL** on data that is produced on daily basis.
- Validated the data being ingested into Hive for further filtering and cleansing.
- Developed **Sqoop** jobs for performing incremental loads from **RDBMS** into **HDFS** and further applied Spark transformations.
- Loaded data into **hive tables** from spark and used **Parquet** columnar format.
- Created **Oozie** workflows to automate and productionize the data pipelines.
- Migrating Map Reduce code into Spark transformations using **Spark** and **Scala**.
- Collecting and aggregating large amounts of log data using **Apache Flume** and staging data in **HDFS** for further analysis.
- Experience using **Impala** for data processing on top of HIVE for better utilization.
- Did a **Poc** on **GCP** cloud services and feasibility of migrating onprem setup to GCP cloud and utilizing various services in **GCP like Dataproc**, **Big Query**, Cloud Storage etc.,
- Designed, documented operational problems by following standards and procedures using **JIRA.**
- Created Hive tables, loading and analyzing data using hive scripts. Implemented Partitioning, Dynamic Partitions, Buckets in HIVE.
- Developed workflows in **Oozie** and scheduling jobs in Mainframes by preparing data refresh strategy document & Capacity planning documents required for project development and support.
- Developed and implemented **Hive scripts** for transformations such as evaluation, **filtering,** and aggregation.

Environment: CDH, Hadoop, Hive, Impala, Oracle, Spark, Pig, Sqoop, Oozie, Map Reduce, GIT, Confluence, Jenkins, DataStage, Jira.

Global Logic **Jan 2017 – Dec 2017**

Python Developer

Responsibilities:

- Developed rest API's using python with flask and **Django framework** and done the integration of various data sources including Java, JDBC, **RDBMS**, **Shell Scripting**, Spreadsheets, and Text files.
- Developed data platform from scratch and took part in requirement gathering and analysis phase of the project in documenting the business requirements.
- Analyzed **SQL** scripts and designed the solutions to implement using **PySpark**.
- Experience with **Pandas** and **NumPy** packages for data manipulations &analysis.
- Experience with unit testing strategies for all **Python frameworks**.
- Executed various **MySQL database** queries from Python using **Python-MySQL** connector and **MySQL** database **packages**.
- Perform Data Cleaning, features scaling, features engineering using pandas and **NumPy** packages in **Python**.
- Performed preliminary data analysis using descriptive statistics and handled anomalies such as removing duplicates and imputing missing values.
- Extensively worked on oracle and **SQL** server and wrote complex SQL queries to query ERP system for data analysis purpose.
- Developed **Python** scripts to migrate data from **Postgres DB to SQL Server.**
- Used **Pandas**, **NumPy**, **Seaborn**, **Matplotlib** in **Python** for developing data pipelines and various machine learning algorithms.
- Design and engineer REST **APIs** and/or packages that abstract feature extraction and complex prediction/forecasting algorithms on time series data.
- Worked in development of applications and worked on **Jenkins** and deployed the project into Jenkins using **GIT** version control system.

Environment: Python, VS code, Jupiter Notebook, PyCharm, Django, RDBMS, Shell scripting, SQL, Pyspark, Pandas, NumPy, Seaborn, Matplotlib, MySQL, Postgres, UNIX, Jenkin, GIT.