**Krishna**

**SENIOR Big Data ENGINEER**

**Email : krishnam9009@gmail.com**

**Ph: 202 643 4469**

**­PROFESSIONAL SUMMARY:**

* Over **9** years of professional IT experience and over **7+** years of **Big Data Ecosystem** experience in ingestion, storage, querying, processing and analysis of **big data/ Azure Databricks/cloud Technologies.**
* Technical expertise in all phases of SDLC (Software Development Life Cycle) which includes a major concentration on Big Data Technologies in highly scalable end-to-end Hadoop Infrastructure analyzing frame works, various **Relational Databases, NoSQL** **Databases, Spark and Java/J2EE technologies** with top most software practices.
* Working knowledge of cloud services specifically Microsoft Azure data pipeline services as-in Azure Data factory, Azure blob storage, Azure delta lake Gen 2 storage, Azure data explorer, Python, Azure Event Hub, Azure DevOps, Azure Logic Apps and Databricks using spark.
* In depth understanding/Knowledge of Hadoop Architecture and various components such as **HDFS, YARN, Name Node, Data Node, Resource Manager, Node Manager**.
* Worked on Databricks Notebooks with languages like Scala, Python and SQL. Implementation of REST based Databricks APIs to create Standard and High Concurrent Clusters, Creating, Scheduling and Executing Jobs and Running Standalone Notebooks etc.
* Experience in building, maintaining multiple **Hadoop** clusters of different sizes and configuration.
* In-depth knowledge of Hadoop and Spark, experience with data mining and stream processing technologies (Kafka, Spark Streaming, Akka Streams).
* Experience 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.
* Experience in importing and exporting data between **HDFS** and Relational Database Management systems using **Sqoop**. Good knowledge in using job scheduling and monitoring tools like **Oozie**.
* Experienced in managing **Azure PaaS** in the complete web application lifecycle which includes building, testing, deploying, managing, and updating.
* Experience in Implementation of **CI/CD pipeline** for the **Azure cloud-based analytical** data ecosystem using Azure DevOps (VSTS) GIT as versioning controlling and hosted pipelines for build and release.
* Experienced in using **Docker** for automating the deployment of applications inside software containers and virtualized deployments using **Docker, with Docker images, Docker Hub and Docker registries.**
* Experience in working with **Cloudera (CDH4 & 5), and Hortonworks** and **AWS Amazon EMR**, **Lambda, Kinesis** data stream to fully leverage and implement new Hadoop features.
* Extensive experience in developing **PIG** Latin Scripts and using **Hive** Query Language for data analytics and working with Cloudera, Hortonworks, and **Microsoft Azure HDINSIGHT Distributions** and also optimizing **Hive Queries** by tuning configuration parameters.
* Experience in building pipelines using **Azure Data Factory** and moving the data into **Azure Data Lake Store.**
* Worked on AWS SAAS, PAAS, Hybrid Cloud and Google cloud platforms
* Exposure and development experience in Microservices Architectures best practices, **Java Spring Boot Framework, Docker, Kubernetes Jenkins** and **Python**.
* Experienced configuring **Amazon EC2 instances** that launch behind a load balancer, monitor the health of **Amazon EC2 instances, deploy Amazon EC2 instances** using command line calls and troubleshoot the most common problems with instances.
* Implement **AWS Data Lake leveraging** **S3, terraform, vagrant/vault, EC2, Lambda,** **VPC, and IAM** in performing data processing and storage while writing complex SQL queries, analytical and aggregate functions on views in **Snowflake data warehouse** to develop near real time visualization using Tableau Desktop/Server 10.4 and Alteryx.
* Experience with **Python, SQL on AWS cloud platform**, better understanding of Data Warehouses like **Snowflake and Data-bricks platform, etc.**
* Experience in implementing **OLAP multi-dimensional** cube functionality using **Azure SQL Data Warehouse,** Azure Data Factory, ADLS, Databricks, SQL DW.
* Excellent understanding and Knowledge of NOSQL databases like **HBase, Cassandra**.
* Hands on experience in Sequence files, RC files, Combiners, Counters, Dynamic partitions, Bucketing for best practice and performance improvement.
* In depth understanding of Spark Architecture including **Spark Core, Spark SQL, Data Frames and Spark Streaming for** developing **Spark Programs** for Batch and Real-Time Processing**.**
* Good understanding of **Oracle data dictionary** and normalization techniques. Experienced in **Amazon redshift database usage, Oracle 12c/11g/10g/9i/8i systems**.
* Experience in managing Hadoop clusters and services using **Cloudera Manager**. Proficient in using Cloudera Manager, an end to end tool to manage Hadoop operations.
* Knowledge of ETL methods for data extraction, transformation and loading in corporate-wide ETL Solutions and Data warehouse tools for reporting and data analysis.
* Experience in scheduling and monitoring jobs using **Oozie and Zookeeper.**
* Experienced in writing Map Reduce programs **& UDF's for both Pig & Hive in java.**
* Experience in collecting log files and error messages across clusters to extract data and to copy into **HDFS using Apache Flume.**
* Experience in database design using **PL SQL** to write stored Procedures, Functions, Triggers and writing queries for Oracle 10g using Data Modelling techniques to find the results based on **SQL and PL/SQL queries**.
* Experienced with code versioning and dependency management systems such as **Git, SVT, and Maven.**
* Experience working with different databases, such as **Oracle, SQL Server, MySQL** and writing stored procedures, functions, joins, and triggers for different Data Models.
* Hands on experience with **Shell Scripting** and UNIX.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Hadoop Distributions** | Apache, Cloudera CDH4, Hortonworks and CDH 5 |
| **Microsoft Azure Databricks** | Multi-Language Notebooks in Scala and Python, Azure Data Lake Store, Azure Key Vault, Standard and High Concurrent Clusters. Databricks APIs, Job setup with Library Installation on Automated Cluster. |
| **Big Data Ecosystem** | Apache Hadoop (HDFS/Map Reduce), Hive, Pig, Sqoop, Zookeeper, Oozie, Hue, Spark, Spark SQL, Apache Kafka |
| **NoSQL Databases** | HBase, Cassandra, Mongo DB |
| **Languages** | SQL, Python, Scala, Core Java, PL SQL, AZURE PowerShell, PySQL |
| **Java / J2EE Technologies** | Core Java, Servlets, Hibernate, Spring, Struts, JMS, EJB |
| **Cloud** | AWS (SaaS, PaaS, IaaS), Lambda, Hybrid, Kinesis, S3, AZURE, Google Cloud BigQuery Platform, Databricks, Azure Data Factory |
| **Visualization Tools** | Power BI, Tableau Desktop and Server, TIBCO Spot-Fire, QlikView, MicroStrategy, Information Builders,  |
| **Databases** | Oracle 10g/9i/8i, DB2, MySQL, MS-SQL Server |
| **Application Servers**  | WebLogic, Web Sphere, JBoss, Tomcat |
| **Development Tools** | Eclipse, Rational Rose |
| **Build Tools** | Jenkins, Maven, ANT |
| **Software Engineering** | Agile/Scrum & Waterfall Methodology |

**PROFESSIONAL EXPERIENCE:**

**Fidelity Investments -Merrimack, NH Feb 2019 - Till date**

**Senior Big Data Engineer**

**RESPONSIBILITIES:**

* Participate in design and development of **Big Data analytical applications**
* Design, support and continuously enhance the project code base, continuous integration pipeline, etc.
* Provide and implement data pipeline solutions / proof of concepts involving **Microsoft Azure cloud services as-in Databricks, Data factory, PySpark.**
* Involved in Requirement gathering, Business Analysis and translated business requirements into Technical design in **Google cloud/Big query, Hadoop and Big Data.**
* Collaborate with Data Warehouse Implementation teams, BI Administrators, Developers and Analysts for successful development of BI reporting and analytic solutions.
* Deployment of Scala Library on Databricks by setting up new Jobs and configuring Automated Cluster to Install Library by adding dependency.
* Implementation of Scala Framework in all **Python Databricks Notebooks** by adding New Cells before and after HIVE queries with %Scala keyword to monitor the dataload and perform Audit checks with log results to set the Notebook Status as either Success or Failure.
* Extract Transform and Load data from Sources Systems **to Azure Data Storage** services using a combination of Azure Data Factory, T-SQL, Spark SQL and U-SQL Azure Data Lake Analytics . Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in **Azure Databricks.**
* Worked with cutting-edge **Google Cloud Big query** technologies to deliver next-Gen Cloud solutions.
* **Created Pipelines** in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Developed JSON Scripts for deploying the Pipeline in **Azure Data Factory (ADF**) that process the data using the **Sql Activity.**
* Creating pipelines, data flows and complex data transformations and manipulations using **ADF and PySpark with Databricks.**
* Define BI standards, guidelines and best practices related to the Google cloud BigQuery platform for clients, client services, and technical teams.
* Maintain accurate and complete technical documents on confluence and Jira for the data Migration from legacy system to Google cloud Big Query platform.
* Developed **Spark/Scala, Python** for regular expression (regex) project in the Hadoop/Hive environment with Linux/Windows for big data resources.
* Developing Spark programs using **Scala API’S** to compare the performance of Spark with Hive and SQL and implemented **Spark using Scala and Sparks SQL** for Faster Testing and Processing of data. Developed analytical component using **Scala, Spark and Spark Streaming.**
* Worked on **NoSQL / SQL, Microservice, RESTful API development**
* Designed and created **Hive** external tables using shared eta store instead of derby partitioning, dynamic portioning and buckets.
* Involved in analyzing business, system and data mapping requirements and developing **ETL** data pipelines for real-time streaming of data using **Kafka** and **Spark**.
* Developed **PySpark scripts** using **Python on Azure HDInsight** for Data Aggregation, Validation, verifying its performance over MR jobs and to extract the data from the web server output files to load into HDFS.
* Built pipelines to move hashed and un-hashed data from Azure Blob to Data lake and wrote **AZURE POWERSHELL scripts** to copy or move data from local file system to HDFS Blob storage
* Involved in **SQOOP implementation** which helps in loading data from various RDBMS sources to **Hadoop systems.**
* Provided automation and deployment of applications inside software containers by providing additional layer of abstraction and automation of operating system level virtualization on Linux using **Dockers, Kubernetes.**
* Responsible for high performance of data architecture and design including **Star** **Schemas, Snowflake Schemas, and Dimensional Modeling.**
* Involved in developing of Confidential Data Lake and in building Confidential Data Cube on **Microsoft Azure HDINSIGHT cluster.**
* Created pipelines to move data from on-premise servers to **Azure Data Lake.**
* Worked on the pros and cons, and best practices of implementing Data Lake, using **Microsoft Azure Data Lake Storage Gen2, and Databricks**
* Experience with container systems like Docker and container orchestration like EC2 Container Service, **Kubernetes, worked with Terraform.**
* Involved in Analyzing system failures, identifying root causes, and recommended course of actions, Documented the systems processes and procedures for future references.
* Databricks Framework to trigger Configured Notebooks in parallel with required parameters.
* Involved in developing Json Scripts for deploying the Pipeline in **Azure Data Factory** (ADF) that process the data using the Cosmos Activity.
* Involved in Data Migration process using **Azure** by integrating with **Github repository and Jenkins.**
* Involved in Configuring **Hadoop cluster** and **Hadoop installation,** Commissioning, Decommissioning, Load Balancing, Troubleshooting, Monitoring and, debugging Configuration of multiple nodes using Hortonworks platform.
* Work inside the team of industry experts on the cutting-edge **Big Data technologies** to develop solutions for deployment at massive scale.
* Developed spark applications in **python(PySpark)** on distributed environment to load huge number of CSV files with different schema in to **Hive tables.**
* Exported the analyzed data to the relational databases **(MySQL, Oracle**) using **Sqoop** from HDFS and accessing Kafka cluster to consume data to **Hadoop** and analyzing the data by performing **Hive queries and running Pig scripts**
* Hands on experience with Angular and React for SPA development
* Used Spark for interactive queries, processing of streaming data and integration with popular **NoSQL** database for huge volume of data.
* Involved in converting **Hive/SQL** queries into Spark transformations using Spark RDDs, Python. Involved in converting map reduce programs into Spark transformations using **Spark RDD in Scala.**

**ENVIRONMENT: Apache Hadoop, Apache Kafka, Spark RDD, Hive, AWS, Cloud, TIBCO, Angular, Python, PySpark, EC2, Kubernetes, Docker, Pig, Oozie, Azure, Azure data factory, Azure Databricks, Zookeeper, Spark, Scala, Cloudera CDH 4/5 Distribution, Eclipse, SQL, NoSQL**

**SCIF, Pleasanton, CA Nov 2017- Jan 2019**

**Role: Big Data / Cloud Engineer**

**RESPONSIBILITIES:**

* Involved in integration of Hadoop cluster with **Spark** engine to perform **BATCH** operations.
* Migrated the data coming from different sources into Spark Cluster through **Spark-Submit** job.
* Developed a **Python Script** to load the CSV files into the S3 buckets and created AWS S3buckets, performed folder management in each bucket, managed logs and objects within each bucket.
* Developed pipelines in Azure Data Factory and built multiple Data Lakes. Built how the data will be received, validated, transformed and then published
* Used Spark API over Hortonworks Hadoop YARN to perform analytics on data in Hive.
* Worked on Databricks APIs to Create and Run Jobs on ad-hoc basis or from oozie Workflow.
* Worked on HIVE Query Optimization by implementing SPARK concepts.
* OOZIE Workflow with Coordinator with File check before triggering Job in Databricks via CLI commands
* Involved in designing a responsive, elastic, and resilient data software architecture that can power large volume data processing pipelines and complex data analysis.
* Involved in building proof of concept using modern Big Data technologies and convert into production-grade implementation.
* Involved in Creating **PySpark frame** to bring data from DB2 to Amazon S3.
* Implemented **Spark using Scala and Spark** SQL for faster testing and processing of data.
* Involved in running all the hive scripts through hive, Impala, **Hive on Spark and some through Spark SQL**
* Architect & implement medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NoSQL DB).
* Uses AWS Lambda function to process records in an Amazon Kinesis data stream and develop hybrid cloud solutions to enhance, harden and support our service delivery processes.
* Developed various scripts to load the CSV files into the S3 buckets and created **AWS S3buckets.**
* Imported data from **AWS S3** into **Spark RDD** and performed transformations and actions on RDD's.
* Develop programs/scripts/data model to extract and store required data from the source systems to produce web-based control reporting using Tibco.
* Deploy **Kubernetes in both AWS using Kinesis and Google cloud**. Setup cluster, replicator. Deploy multiple containers in a pod.
* Involved in migrating team with the current Linux environment to **AWS/RHEL Linux** environment and used auto scaling feature.
* Involved in configuring and monitoring **Amazon Web Services resources** as well as involved in deploying the content cloud platform to Amazon Web Services using EC2, S3 and EBS.
* Leveraged **AWS Lambda cloud services such as EC2**, auto scaling and VPC to build secure, highly scalable and flexible systems that handled load on the servers.
* Applied transformations on the data loaded into **Spark RDDs** and done in memory data computation to generate the output response.
* Data sources are extracted, transformed and loaded to generate **CSV data files** with **Python programming and SQL queries.**
* Developed multiple POCs using **PySpark** and deployed on the Yarn cluster, compared the performance of Spark, with Hive and SQL.
* Worked with **Avro** and **Parquet** file formats and used various compression techniques to leverage the storage in **HDFS**.
* Utilized **Spark Core, Spark Streaming and Spark** SQL API for faster processing of data instead of using MapReduce in Java.
* Load the data into **Spark RDD** and do in memory data Computation to generate the Output response.
* Developed Python Script to import data SQL Server into HDFS & created Hive views on data in HDFS using Spark.
* Involved in converting **MapReduce programs** into Spark transformations using Spark RDD in Scala.
* Involved in Implementing various AWS environment for provisioning of Linux servers and services implemented by the providers.
* Experience in **AWS (Lambda, Kinesis, S3) environment** including AWS Storage and content Delivery, Databases, Networking, Management Tools, Security & Identity etc.
* Analyzed the SQL scripts and designed the solution to implement using **PySpark.**
* Used **Spark-SQL** to Load JSON data and create Schema RDD and loaded it into Hive Tables and handled structured data using SparkSQL.
* Analyzed the SQL scripts and designed it by using **PySpark SQL** for faster performance.
* Involved in giving architecture guidance for selected Business Units, assessing migration feasibility and deployment strategy based on **AWS Architecture best practices.**
* Used hive to analyze the **partitioned** data and compute various metrics for reporting.
* Reduced the latency of spark jobs by tweaking the spark configurations and following other performance and **Optimization** techniques.
* Involved in using **Apache Splunk** add-ons to enhance data ingestion and analysis of log data.
* Used **Oozie** workflow engine to manage interdependent Hadoop jobs and to automate several types of Hadoop jobs such as Java map-reduce, Hive, Pig.
* Worked on various production issues during the month end support and provide resolutions without missing any SLA.

**ENVIRONMENT: Apache Hadoop, Apache Spark, Spark RDD, Scala, AWS, EC2, Apache Splunk, MapReduce, Hive, Pig, Oozie, Zookeeper, Python, PySpark, Tibco, SQL, Linux.**

**Paypal - San Jose, CA Jan 2017- Nov 2017**

**Role: Hadoop / Big Data Developer**

**RESPONSIBILITIES:**

* Worked with the source team to understand the format and delimiters of the data file.
* Running Periodic **Map-Reduce** jobs to load data from **Cassandra** into Hadoop.
* Involved in creating **Hive tables,** loading with data and writing **Hive queries** which will invoke and run Map-Reduce jobs in the backend.
* In depth understanding/ knowledge of **Hadoop architecture** and various components such as HDFS, application manager, node master, resource manager name node, data node and map reduce concepts.
* Involved in developing a **Map Reduce framework** that filters bad and unnecessary records.
* Involved heavily in setting up the **CI/CD pipeline** using Jenkins, Maven, Nexus, GitHub, and AWS.
* Involved in moving all log files generated from various sources to HDFS for further processing through flume.
* Written the **HIVE queries** to extract the data processed.
* Developed data pipeline using flume, Sqoop, pig and map reduce to ingest customer behavioral data and purchase histories into HDFS for analysis.
* Used Spark-SQL to load JSON data and create schema RDD and loaded it into Hive tables handled structured data using **Spark SQL**
* Created **Hbase tables** to store variable data formats of data coming from different legacy systems.
* Used **HIVE to do transformations**, event joins and some pre-aggregations before storing the data onto HDFS.
* The Hive tables created as per requirement were **internal** or **external** tables defined with appropriate static and dynamic partitions, intended for efficiency.
* Experience in **analyzing** Cassandra database and comparing it with other open-source **NoSQL** databases to find which one of them better suits the current requirements.
* Loading the output data into Cassandra using Bulk Load.
* Implemented the workflows using **Apache Oozie** framework to automate tasks.
* Developing design documents considering all possible approaches and identifying best of them.
* Written Map Reduce code that will take input as log files and parse the and structures them in tabular format to facilitate effective querying on the log data.
* Transformed the data using Hive, Pig for BI team to perform visual analytics according to the client requirement.
* Developed **scripts** and automated data management from end to end and sync up b/w all the Clusters.
* Implemented **Fair schedulers** on the Job Tracker to share the resources of the cluster for the Map Reduce jobs given by the users.

**ENVIRONMENT: Cloudera CDH 3/4 Distribution, Tibco, HDFS, MapReduce, Cassandra, Hive, Oozie, Pig, Shell Scripting, MySQL.**

**DSW, Irving, TX Nov 2015 – Dec 2016 Role: Hadoop Developer**

**RESPONSIBILITIES:**

* Explored and used **Hadoop ecosystem** features and architectures.
* Worked with business team to gather their requirements and new support features.
* Configured **Sqoop** and developed scripts to extract data from **MySQL** into **HDFS**.
* Wrote programs using scripting languages like Pig to manipulate data.
* Involved in creating **Hive** tables, loading structured data and writing hive queries which will run internally in map reduce way.
* Monitoring the running Map-Reduce programs on the cluster.
* Implemented the workflows using **Apache Oozie** framework to automate tasks.
* Reviewed the **HDFS usage** and system design for future scalability and fault-tolerance.
* Prepared Shell scripts to get the required info from the logs.
* Responsible to manage data coming from different sources.

**ENVIRONMENT: Apache Hadoop, Eclipse, MySQL, Pig, Hive, Sqoop, Linux, Oozie, Shell Scripting.**

**Analytix Data Services, Hyderabad, India Jun 2012 - Dec 2014 Role: Junior SQL Developer**

**RESPONSIBILITIES:**

* Created stored **procedures** to transform the data and worked for various needs of the transformations while loading the data.
* Designing and creating **SQL** tables, indexes and sequences.
* Conducted logical and physical database design including data modeling, maintenance and problem diagnosis.
* Involved in creating and modifying several UNIX **shell scripts** according to the changing needs of the project and client requirements.
* Involved in writing SQL queries, joins, **DDL**, **DML** and user defined functions (UDF) to implement business logic.
* Responsible for **performance tuning** and query optimization.
* Analyze end user database needs and provide efficient solutions.
* Performed **backup/restore**, database objects such as tables, procedures, constraints, Indexes and views

**ENVIRONMENT: DB2, SQL Server, Shell Scripting, UNIX, PL SQL**