**Big Data Engineer || Data Engineer || ETL Developer || Hadoop || Azure || GCP ||**

**Mani Krishna Prasad**

**Email: krishna@adven-it.com**

**PH:** (972 8538231

Summary

* Overall **8+ years** of professional experience in Information Technology and around 5 years of expertise in BIGDATA using HADOOP framework and Analysis, Design, Development, Testing, Documentation, Deployment, and Integration using SQL and Big Data technologies.
* Expertise in using major components of Hadoop ecosystem components like HDFS, YARN, MapReduce, Hive, Impala, Pig, Sqoop, HBase, Spark, Spark SQL, Kafka, Spark Streaming, Flume, Oozie, Zookeeper, Hue.
* Good understanding of distributed systems, HDFS architecture, Internal working details of MapReduce and Spark processing frameworks.
* Knowledge of ETL methods for data extraction, transformation and loading in corporate-wide ETL Solutions and Data Warehouse tools for reporting and data analysis.
* Deployed the Big Data Hadoop application using Talend on cloud AWS (Amazon Web Services) and on Microsoft Azure.
* Expertise in deploying cloud-based services with Amazon Web Services (Databases, Migration, Compute, IAM, Storage, Analytics, Network & Content Delivery, Lambda and Application Integration).
* Excellent Proficiency in storage, compute and networking services with implementation experience in data engineering using key AWS services such as VPC, EC2, S3, ELB, AutoScalingGroup(ASG), EBS, RDS, IAM, EFS, CloudFormation, Redshift, DynamoDB, Glue, Lambda, Step Functions, Kinesis, Route 53, CloudWatch, CloudFront, CloudTrail, SQS,SNS, SES, AWS Systems Manager etc.
* Developed data ingestion modules using AWS Step Functions, AWS Glue and Python modules.
* Develop data set processes for data modelling, and Data mining. Recommend ways to improve data reliability, efficiency, and quality.
* Created a multi-threaded Java applications running on master node for pulling other data feeds (Json and XML) into S3.
* Handle streaming data in real time with Kafka, Flume and Spark Streaming. Knowledge on Flink
* Developed and deployed various Lambda functions in AWS with in-built AWS Lambda Libraries and deployed Lambda Functions in Scala with custom Libraries.
* Optimized Hive tables utilizing partitions and bucketing to give better execution Hive QL queries.
* Used Spark-SQL to read data from hive tables and perform various data cleansing, data validations, transformations, and aggregations as per down stream business team requirements.
* Responsible for loading processed data to AWS Redshift table for allowing Business reporting team to build dashboards.
* Created views in AWS Athena to allow secure and streamlined data analysis access to downstream business teams.
* Worked extensively with Data Science team to help productionalize machine learning models and to build various feature datasets as needed for data analysis and modelling.
* Experience implemented AWS Kinesis Firehose to sink data directly to Redshift and S3.
* Experience in converting AWS existing infrastructure to server less architecture (AWS Lambda, Kinesis) and deployed AWS Cloud formation
* Worked extensively in automating launching of EMR clusters and terminating EMR cluster as soon as jobs are finished.
* Excellent understanding of Spark Architecture with Databricks, Structured Streaming. Setting Up AWS and Microsoft Azure with Databricks workspace for Business Analytics, Manage Clusters in Databricks for Managing the Machine Learning Lifecycle.
* Hands on experience on Python programming PySpark implementations in AWS EMR, building data pipelines infrastructure to support deployments for Machine Learning models, Data Analysis and cleansing and using Statistical models with extensive use of Python, Pandas, Numpy, Visualization using Matplotlib, Seaborn and Scikit packages for predictions, xgboost.
* Worked on data visualization and analytics with research scientist and business stake holders.
* Experience in complete project life cycle (design, development, testing and implementation) of Client Server and Web applications.
* Excellent programming skills with experience in Java, C, SQL, and Python Programming.
* Worked on various programming languages using IDEs like Eclipse, NetBeans, and IntelliJ, Putty, GIT.
* Experienced in working in SDLC, Agile and Waterfall Methodologies.
* Excellent experience in designing and developing Enterprise Applications for J2EE platform using Servlets, JSP, Struts, Spring, Hibernate and Web services.

# Skills

|  |  |
| --- | --- |
| * **BigData/Hadoop Technologies**
 | AWS EMR, S3, EC2-Fleet, Spark-2.2, 2.0 and 1.6, Hortonworks HDP, Hadoop, Mapreduce, Pig, Hive, Apache Spark, SparkSQL, Informatica Power Center 9.6.1/8.x, Kafka, NoSQL, Elastic Mapreduce(EMR), Hue,YARN, Nifi, Impala, Sqoop, Solr, OOZie. |
| * **Languages**
 | Java, Scala, SQL, UNIX shell script, JDBC, Python, Perl. |
| * **Cloud Environment**
 | AWS (Amazon Web Services), Microsoft Azure, GCP ( Google ClouPlatform) |
| * **Operating Systems**
 | All versions of Windows, UNIX, LINUX, Macintosh HD, Sun Solaris |
| * **Web Design Tools**
 | HTML, CSS, JavaScript, JSP, jQuery, XML |
| * **Development Tools**
 | Microsoft SQL Studio, IntelliJ, Azure Databricks, Eclipse, NetBeans. |
| * **Public Cloud**
 | EC2, IAM, S3, Auto scaling, CloudWatch, Route53, EMR, RedShift |
| * **Databases**
 | Oracle 10g, 11g, 12c, Microsoft SQL Server 2008,2010/2012, MySQL 4.x/5.x, DB2, Teradata, Netezza |
| * **NO SQL Databases**
 | Cassandra, HBase, MongoDB, MariaDB |
| * **Development Methodologies**
 | Agile/Scrum, UML, Design Patterns, Waterfall |
| * **Build Tools**
 | Jenkins, Toad, SQL Loader, PostgreSql, Talend, Maven, ANT, RTC, RSA, Control-M, Oozie, Hue, SOAP UI |
| * **Reporting Tools**
 | MS Office (Word/Excel/Power Point/ Visio/Outlook), Crystal reports XI, SSRS, cognos. |

**PROFESSIONAL EXPERIENCE**

**E Trade, Arlington, VA Sep 2022 - present**

**Sr. AWS Data Engineer**

**Roles & Responsibilities:**

* Design, create and implement RDBMS as well as NoSQL database, build views, indexes, and stored procedures.
* Involved in designing and deploying multi-tier applications using all the AWS services like (EC2, Route53, S3, RDS, Dynamo DB, SNS, SQS, IAM) focusing on high-availability, fault tolerance, and auto-scaling in AWS Cloud Formation.
* Supporting Continuous storage in AWS using Elastic Block Storage, S3, Glacier. Created Volumes and configured Snapshots for EC2 instances.
* Data modeling of the product information, customer features, build data warehouse solution to support BI activities.
* SQL queries on RDBMS such as MySql/Postgres and HiveQL on Hive tables for data extraction and preliminary data analysis. Database Migration from access to SQL Server.
* Build data pipelines including data ingestion, data transformation such as aggregation, filtering, cleaning, and data storage.
* Implement One time Data Migration of Multistate level data from SQL server to Snowflake by using Python and SnowSQL.
* Worked on scheduling all jobs using Airflow scripts using python. Adding different tasks to DAG’s and dependencies between the tasks.
* Data ingestion from SQL and NoSQL database and multiple data formats such as XML, JSON, and CSV.
* Data ingestion of real-time customer behavioral data into HDFS using Flume, Sqoop, Kafka, and data transformation using Spark Streaming.
* Perform ETL operations using Scala Spark and PySpark under IntelliJ with Java and PyCharm with Python respectively.
* Implement and execute the parallel processing of Map-Reduce job utilizing Java for the log data from the servers.
* Monitor and health check of the data warehouse by providing failover solutions and disaster recovery solutions in a cost-effective manner.
* Leverage Yarn for large-scale distributed data as well as troubleshoot and resolve Hadoop cluster performance issues.
* Expertise in Creating, Debugging, Scheduling and Monitoring jobs using Airflow for ETL batch processing to load into Snowflake for analytical processes.
* Developed ETL’s using PySpark. Used both Dataframe API and Spark SQL API.
* Perform data management and data query using Spark and deal with streaming data using Kafka to make sure data transfers and processes in a fast and reliable manner.
* Leverage AWS S3 as storage solution for HDFS, AWS Glue as the ETL solution and AWS kinesis as the data streaming solution to deploy the data pipeline on cloud.
* Migrate data warehouse from RDBMS to AWS Redshift and analyze log data using AWS Athena on S3. Maintain Hadoop cluster using AWS EMR.
* Data cleansing, data manipulation, data wrangling using Python to eliminate invalid datasets and reduce prediction error.
* Conducted A/B test on metrics such as customer retention, acquisition, sales revenue, and volume growth to assess the performance of products.
* Leveraged Pandas, Numpy and Seaborn for exploratory data analysis.
* Extend Hive functionality by using User Defined Functions including UDF, UDTF, and UDAF.
* Developed predictive modeling using Python packages such as SciPy and scikit-learn as well as Mixed-effect models and time series models in R based on business requirements.
* Stage the API or Kafka Data(in JSON file format) into Snowflake DB by FLATTENing the same for different functional services.
* Carried out Dimension Reduction with PCA and Feature Engineering with Random Forest to capture key features for predicting annual sales and best purchased product using Python and R.
* Created Hive integrated Tableau dashboards and reports to visualize the time series of purchase value to keep track of the business metrics as well as deliver business insights to stakeholders.
* Work with Git for version control, Maven for Java project build, test and deploy.

**Environment:** Spark, Python, Scala, Kafka,AWS, EC2, SQL, Hive, AWS, Java, Oracle, Glue, Athena, S3, Parquet, PowerBI, Data Studio, Tableau, Oozie, Kafka, Hbase, Data bricks, EMR.

**Kroger, Cincinnati, OH Jan 2021 - Jul 2022**

**Sr Azure Data Engineer**

**Responsibilities:**

* As a Big Data/Hadoop Developer worked on Hadoop eco-systems including Hive, MongoDB, Zookeeper and Spark
* Developed multiple MapReduce jobs in Java for data cleaning and preprocessing.
* Involved in various phases of development analyzed and developed the system going through Agile Scrum methodology.
* Handled importing data from different data sources into HDFS using Sqoop and also performing transformations using Hive, Impala, MapReduce and then loading data into HDFS.
* Designed and Developed Custom Hqls in Hive.
* Good knowledge in writing hive Query and assign schemas and create Hive tables.
* Good knowledge on workflow/schedulers on Oozie.
* Worked with application teams to install operating system, Hadoop updates, patches, version upgrades as required.
* Specified the cluster size, allocating Resource pool, Distribution of Hadoop by writing the specification texts in JSON File format.
* Developed Spark Applications by using Scala, Java, and Implemented Apache Spark data processing project to handle data from various RDBMS and Streaming sources.
* Wrote Hive Queries for analyzing data in Hive warehouse using Hive Query Language (HQL).
* Created and maintained Technical documentation for launching Hadoop Clusters and for executing Hive queries.
* Handled importing of data from various data sources, performed transformations using Hive, Impala, MapReduce, loaded data into HDFS and extracted the data from MySQL into HDFS using Sqoop.
* Used Spark SQL on data frames to access hive tables into spark for faster processing of data.
* Configured Spark streaming to receive real time data from the Kafka and store the stream data to HDFS using Scala.
* Responsible for developing data pipeline using flume, Sqoop and Pig to extract the data from weblogs and store in HDFS.
* 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 In Azure Databricks.
* Developed Pig Latin scripts to extract data from the web server output files to load into HDFS.
* Developed data pipeline using MapReduce, Flume, Sqoop and Pig to ingest customer behavioral data into HDFS for analysis.
* Used Different Spark Modules like Spark core, Spark SQL, Spark Streaming, Spark Data sets and Data frames.
* Used Spark for interactive queries, processing of streaming data and integration with popular NoSQL database for huge volume of data.
* Responsible for implementing monitoring solutions in Ansible, Terraform, Docker, and Jenkins.
* Wrote Hive Queries for analyzing data in Hive warehouse using Hive Query Language (HQL).
* Created and maintained Technical documentation for launching Hadoop Clusters and for executing Hive queries.
* Handled importing of data from various data sources, performed transformations using Hive, MapReduce, loaded data into HDFS and extracted the data from MySQL into HDFS using Sqoop.
* Used Spark SQL on data frames to access hive tables into spark for faster processing of data.
* Involved in full phases of cleaning, designing, developing, testing and deploying the Datasets into table.
* Maintain Hadoop, Hadoop ecosystems, and database with updates/upgrades, performance tuning and monitoring.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
* Prepared data analytics processing, and data regression for availability of analytics results to visualization systems, applications, or external data stores.

**Environment**: Red Hat Enterprise Linux, HDP, Azure, Hadoop, Map Reduce, HDFS, Hive, Shell Script, SQOOP, Python, PostgreSQL, spark, airflow, snowflake.

**T-Mobile, Bellevue, WA Aug 2020 - Dec 2021**

**Aws Data Engineer**

**Responsibilities:**

* Evaluated suitability of Hadoop and its ecosystem to the project and implementing, validating with various proof of concept (POC) applications to eventually adopt them to benefit from the Big Data Hadoop initiative
* Estimated the Software & Hardware requirements for the Name node and Data nodes in the cluster.
* Experience in migrating existing databases from on premise to AWS Redshift using various AWS services
* Developed the Pysprk code for AWS Glue jobs and for EMR.
* Installed and configured Hadoop Map Reduce, HDFS, developed multiple Map Reduce jobs in java and Scala for data cleaning and preprocessing
* Developed Java Map Reduce programs for the analysis of sample log file stored in cluster
* Implemented Spark using Python and Spark SQL for faster testing and processing of data.
* Developed Spark scripts using Python for Data Aggregation, Validation and verified its performance over MR jobs.
* Imported data using Sqoop to load data from MySQL to HDFS on regular basis.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala
* Used IAM to create new accounts, roles and groups and polices and developed critical modules like generating amazon resource numbers and integration points with S3, Dynamo DB, RDS, Lambda and SQS Queue
* Reviewing the explain plan for the SQLs in snowflake
* Developed ETL parsing and analytics using Python/Spark to build a structured data model in Elastic search for consumption by the API and UI.
* Developed ETL jobs using Spark -Scala to migrate data from Oracle to new Cassandra tables.
* Used Spark -Scala (RDD’s, Data frames, Spark Sql) and Spark - Cassandra -Connector API's for few tasks (Data migration, Business report generation etc.)
* Created Partitions, Buckets based on State to further process using Bucket based Hive joins.
* Created an e-mail notification service upon completion of job for the team which requested for the data.
* Implemented security to meet PCI requirements, using VPC Public/Private subnets, Security Groups, NACLs, IAM roles, policies, VPN, WAF, Trust Advisor, Cloud Trail etc. to pass penetration testing against infrastructure
* Defined job work flows as per their dependencies in Oozie.
* Played a key role in productionizing the application after testing by BI analysts.

**Environment**: MapReduce, Hive, Sqoop 1.4.4, Oozie 4.2, Python, Scala, Spark 2.3, Kafka, Ambari, Cassandra, Linux, AWS EMR, S3, Storm

**Option Care Health, Cleveland, Ohio. Aug 2019 - July 2020**

**Data Engineer**

**Responsibilities:**

* As a Big Data/Hadoop Developer worked on Hadoop eco-systems including Hive, MongoDB, Zookeeper, Spark Streaming with MapR distribution.
* Developed multiple MapReduce jobs in Java for data cleaning and preprocessing.
* Involved in various phases of development analyzed and developed the system going through Agile Scrum methodology.
* Involved in designing the row key in HBase to store Text and JSON as key values in HBase table and designed row key in such a way to get/scan it in a sorted order.
* Used Cloud watch logs to move application logs to S3 and create alarms based on a few exceptions raised by applications.
* Used Kibana, which is an open source based browser analytics and search dashboard for Elastic Search.
* Maintain Hadoop, Hadoop ecosystems, and database with updates/upgrades, performance tuning and monitoring.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
* Prepared data analytics processing, and data egress for availability of analytics results to visualization systems, applications, or external data stores.
* Builds large-scale data processing systems in data warehousing solutions, and work with unstructured data mining on NoSQL.
* Responsible for design and development of Spark SQL Scripts based on Functional Specifications.
* Used AWS services like EC2 and S3 for small data sets processing and storage.
* Provisioning of Cloudera Director AWS instance and adding Cloudera manager repository to scale up Hadoop Cluster in AWS.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, and Scala.

**Environment**: Hadoop 3.0, Spark 2.3, MapReduce, Java, MongoDB, HBase 1.2, JSON, Hive 2.3, Zookeeper 3.4, AWS, MySQL, Scala 2.12, Python, Cassandra 3.11, HTML5, JavaScript

**Pexway, Hyderabad, India June 2015 - Aug 2018**

# Role: SQL Server BI Developer

**Responsibilities:**

* Analyzing and understanding the existing reporting environment.
* Worked closely with the Business analysts for requirements gathering, designing the database and creating workflow for report.
* Created one central SSIS Master Package to execute multiple child packages with Control Flow and Data Flow. Utilized Parent Package and XML Configuration to dynamically pass variable values on runtime.
* Participated in identifying data migration issues and resolved them.
* Extensively used variables, break point, check point, logging, package configuration and event handler in SSIS packages to meet the business needs.
* Involved in performance tuning to optimize queries and enhance the performance of databases, SQL queries, and stored procedures using SQL Profiler, Execution Plan and Index Tuning Wizard.
* Created Complex SSAS Cubes with multiple fact measure groups, and multiple hierarchies based on the OLAP reporting needs.
* Built MDX queries for Analysis Services & Reporting Services.
* Designed, developed, created and tested PivotTable/PivotChart reports based on OLAP cubes and offline cubes.
* Created multiple partitions and aggregations for the different measure groups for improving performance of the cubes.
* Resolving the SSAS cube connectivity and data issues as and when needed.
* Interacted with Business Users to help them understand to generate reports/look at the business data with various drill down options with Excel via connecting to SQL Server Analysis Services (SSAS).

**Environment:** SQL Server 2012, SQL Server Integration Services (SSIS), Reporting Services (SSRS), TFS, Pivot Tables, MS Visual Studio.Net, C#, SQL Profiler, Windows 2003/2007 Server OS.