**Likith Email:** Likithkishore93@gmail.com

**Hadoop/Bigdata Developer Phone: +1(425)324-6217**

**Summary:**

* 8+ years of IT experience in Software Development with 5+ years’ work experience as Big Data /Hadoop Developer with good knowledge of Hadoop framework.
* Expertise in Hadoop architecture and various components such as HDFS, YARN, High Availability, Job Tracker, Task Tracker, Name Node, Data Node, and MapReduce programming paradigm.
* Experience with all aspects of development from initial implementation and requirement discovery, through release, enhancement and support (SDLC & Agile techniques).
* Having 4+ years of experience in Design, Development, Data Migration, Testing, Support and Maintenance using Redshift Databases.
* Having hands on experience on NoSql databases like HBASE,HIVE
* Having hands on experience on streaming data using kafka & spark scala streaming API
* Having 4+ years of experience on Apache Hadoop technologies like Hadoop distributed file system (HDFS), Map Reduce framework, Hive, PIG, Python, Sqoop, Oozie, HBase, Spark, Scala and Python.
* 3+ years of experience in AWS cloud solution development using Lambda, SQS, SNS, Dynamo DB, Athena, S3, EMR, EC2, Redshift, Glue, and CloudFormation.
* Experience in using Microsoft Azure SQL database, Data Lake, Azure ML, Azure data factory, Functions, Databricks and HDInsight.
* Having hands on experience in versioning using bitbucket
* Working experience in Big data on cloud using AWS EC2 & Microsoft Azure, and handled redshift & Dynamo databases with huge amount of data 300 TB.
* Extensive experience in migrating on premise Hadoop platforms to cloud solutions using AWS and Azure.
* Migrated an existing on-premises application to AWS.
* Experience in design and developing various cloud based ETL pipelines on GCP.
* Hands on experience in installing, configuring, and using Hadoop ecosystem components like Hadoop MapReduce, HDFS, HBase, Hive, Sqoop, Pig, Zookeeper, Storm, Spark, Kafka, and Flume
* Possess good communication, interpersonal, analytical skills and a go-getter personality.
* 3+ years of experience in writing python as ETL framework and Pyspark to process huge amount of data daily.
* Strong experience in implementing data models and loading unstructured data using HBase, Dynamo Db and Cassandra.
* Having hands on experience in Application Deployment using CICD pipeline.
* Experience in implementing Spark using Scala and Spark SQL for faster processing of data.
* Strong experience in extracting and loading data using complex business logic’s using Hive from different data sources and built the ETL pipelines to process tera bytes of data daily.
* Experienced in transporting, and processing real time event streaming using Kafka and Spark Streaming.
* Hands on experience with importing and exporting data from Relational databases to HDFS, Hive and HBase using Sqoop.
* Experienced in processing real time data using Kafka 0.10.1 producers and stream processors and implemented stream process using Kinesis and data landed into data lake S3.
* Designed and developed spark pipelines to ingest real time event-based data from Kafka and other message queue systems and processed huge data with spark batch processing into data warehouse hive.
* Experienced in creating and analyzing Software Requirement Specifications (SRS) and Functional Specification Document (FSD).
* Working with Bootstrap twitter framework to Design single page application.
* Strong working knowledge developing Cross Browser Compatibility (IE, Firefox, Safari, Chrome etc.) for dynamic web applications.
* Experienced in developing the unit test cases using Junit, Mockito, Scala Test.
* Capable of organizing, coordinating and managing multiple tasks simultaneously.
* Excellent communication and inter-personal skills, self-motivated, organized and detail-oriented, able to work well under deadlines in a changing environment and perform multiple tasks effectively and concurrently.
* Strong analytical skills with ability to quickly understand client’s business needs. Involved in meetings to gather information and requirements from the clients.

**Technical Skills:**

|  |  |
| --- | --- |
| **Hadoop** | Hadoop, Map Reduce, HIVE, PIG, Impala SQOOP, HDFS, HBASE, Oozie, Spark, Pyspark, Scala and Mongo DB |
| **Cloud Technologies:** | AWS Kinesis, Lambda, EMR, EC2, SNS, SQS, Dynamo DB, Step Functions, Glue, Athena, CloudWatch, Azure Data Factory, Azure Data Lake, Functions, Azure SQL Data Warehouse, Databricks and HDInsight |
| **DBMS:** | Amazon Redshift, Postgres, Oracle 9i, SQL Server, IBM DB2 And Teradata |
| **ETL Tools:** | Data Stage, Talend and ABInitio |
| **Deployment Tools:** | Git, Jenkins, Terraform and CloudFormation |
| **Programming Language:** | Python, Scala, PL/SQL |
| **Scripting:** | Unix Shell and Bash scripting |

**Education:**

* Bachelor’s in Computer Science BIET College of Engineering – 2014

**Professional Experience:**

**NCFAST (North Carolina Families Accessing Services Through Technology Apr 20 to till date**

**Hadoop Developer**

**Client Description:** Medicaid managed care in a way that advances high-value care, improves population health, engages and supports providers, and establishes a sustainable program with predictable costs. At the core of these efforts is the goal to improve the health of North Carolinians through an innovative, whole person centered, well-coordinated system of care, which addresses both medical and non-medical drivers of health. In Medicaid Managed Care, people enroll in a health care plan that is run by a prepaid health plan or “PHP.”

**Responsibilities:**

* Written ETL jobs in using spark data pipelines to process data from different source to transform data to multiple targets.
* Created streams using Spark and processed real time data into RDDs & data frames and created analytics using SPARK SQL.
* Creating Test Automation Framework in Python.
* Created control structures
* Designed Redshift based data delivery layer for business intelligence tools to operate directly on AWS S3.
* Implemented kinesis data streams to read real time data and loaded into data s3 for downstream processing.
* Created framework for Data Profiling
* Created framework for data encryption
* AWS Infrastructure setup on EC2 and S3 API implementation for accessing S3 bucket data file.
* Designed “Data Services” to intermediate data exchange between the Data Clearinghouse and the Data Hubs.
* Written ETL flows and MapReduce to process data from AWS S3 to dynamo DB and HBase.
* Involved in the ETL phase of the project & Designed and analyzed the data in oracle and migrated to Redshift and Hive.
* Use Spark to process live Streaming data using Apache Kafka.
* Created ETL framework to hydrate the data lake using pyspark
* Developed Spark using Scala and Spark SQL for faster processing of data.
* Developed Spark streaming jobs to process terabytes of XML format data using Scala.
* Created Databases and tables using Redshift and dynamo DB and written complex EMR scripts to process Tera bytes of data into AWS S3 cluster.
* Performed real time analytics on transactional data using python to create statistical model for predictive and reverse product analysis.
* Develop Scala Source Code to process heavy RAW JSON data.
* Involving in client meetings and explaining the views to supporting and gathering requirements.
* Working in an agile methodology, understand the requirements of the user stories
* Prepared High-level design documentation for approval.
* Created ETl Framework using azure databricks
* Also, data visualization software tableau, quick sight and Kibana are used as part of bringing new insights from data extracted and better representation of data.
* Created ETL pipeline using azure datafactory.
* Designed data models for dynamic and real-time data with intention to be used by various applications with OLAP and OLTP needs

**Environment:** Pyspark, Scala, Python, AWS, tableau, shell-scripting, Apache Kafka, JIRA, azure datafactory, azure databricks.

**Vanguard, Malvern, PA Oct 2018 – Mar 2020**

**Big Data Engineer**

**Client Description:** The Vanguard Group, Inc. is an American registered investment advisor based in Malvern, Pennsylvania with about $6.2 trillion in global assets under management.

**Responsibilities:**

* Full life cycle of the project from Design, Analysis, logical and physical architecture modeling, development, Implementation, testing.
* Designed and developed automation test scripts using Python
* Creating Data Pipelines using Azure Data Factory.
* Automating the jobs using Python.
* Creating tables and loading data in Azure MySQL Database
* Creating Azure Functions, Logic Apps for Automating the Data pipelines using Blob triggers.
* Analyze SQL scripts and design the solution to implement using Pyspark
* Developed Spark code using Python (Pysaprk) for faster processing and testing of data.
* Used SparkAPI to perform analytics on data in Hive
* Optimizing and tuning Hive and spark queries using data layout techniques such as partitioning, bucketing or other advanced techniques.
* Used Scala collection framework to store and process the complex information.
* Data Cleansing, Integration and Transformation using PIG
* Involved in exporting and importing data from local file system and RDBMS to HDFS
* Designing and coding the pattern for inserting data into Data lake.
* Moving the data from On-Prem HDP clusters to Azure
* Building, installing, upgrading or migrating petabyte size big data systems
* Fixing Data related issues.
* Monitoring the functioning of big data and messaging systems like Hadoop, Kafka, Kafka Mirror makers to ensure they operate at their peak performance at all times.
* Created Hive tables, and loading and analyzing data using hive queries
* Reading and translating data models, data querying and identifying data anomalies and provide root cause analysis.
* Analyzing real time data using spark streaming and Kafka.
* Implementing Kafka Produce and consumer API Using Scala.
* Shell scripting to automate the orchestration of spark jobs
* Deploying Applications using CICD pipeline.
* Performance Tuning of Spark jobs.
* Performance Tuning of Hive.
* Tracking batch processing using HBASE.
* Migrating data from external sources to S3 buckets in AWS.
* Engage in project planning and delivering to commitments.
* POC’s on new technologies (Snowflake) that is available in the market to determine the best suitable one for the Organization needs.

**Environment:** Azure MySQL Database, Pyspark, Scala, Python, AWS, Hive, Performance Tuning, API’s,Kafka, CICD pipeline, Java

**Blue Cross Blue Shield Jacksonville-FL Sep 2015 to Aug 2018**

**Big Data Engineer**

**Blue Cross Blue Shield is the largest health insurer serving more than seven million people nationwide. It offers health care products and insurance services.**

**Responsibilities**

* Set up and built AWS infrastructure with various services available by writing cloud formation templates (CFT) in json and yaml.
* Developed Cloud Formation scripts to build EC2 on demand
* With the help of IAM created roles, users and groups and attached policies to provide minimum access to the resources.
* Updating the bucket policy with IAM role to restrict the access to user.
* Configured AWS Identity Access Management (IAM) Group and users for improved login authentication.
* Created topics in SNS to send notifications to subscribers as per the requirement.
* Involved in full life cycle of the project from Design, Analysis, logical and physical architecture modeling, development, Implementation, testing.
* Moving data from Oracle to HDFS using Sqoop
* Data profiling on critical tables from time to time to check for the abnormalities
* Created Hive Tables, loaded transactional data from Oracle using Sqoop and Worked with highly unstructured and semi structured data.
* Developed MapReduce (YARN) jobs for cleaning, accessing and validating the data.
* Created and worked Sqoop jobs with incremental load to populate Hive External tables
* Scripts were written for distribution of query for performance test jobs in Amazon Data Lake.
* Developed optimal strategies for distributing the web log data over the cluster importing and exporting the stored web log data into HDFS and Hive using Sqoop.
* Apache Hadoop installation & configuration of multiple nodes on AWS EC2 system
* Working on CDC (Change Data Capture) tables using Spark Application to load data into Dynamic Partition Enabled Hive Tables.
* Analyzed the SQL scripts and designed the solution to implement using Pyspark
* Responsible for developing data pipeline with Amazon AWS to extract the data from weblogs and store in HDFS.
* Uploaded streaming data from Kafka to HDFS, HBase and Hive by integrating with storm.
* Supporting data analysis projects by using Elastic MapReduce on the Amazon Web Services (AWS) cloud performed Export and import of data into s3.
* 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.
* Integrated Oozie with the rest of the Hadoop stack supporting several types of Hadoop jobs out of the box (such as Map-Reduce, Pig, Hive, and Sqoop) as well as system specific jobs (such as Java programs and shell scripts).
* Designed and Implemented Partitioning (Static, Dynamic) Buckets in HIVE.
* Monitored workload, job performance and capacity planning using Cloudera Manager.
* Involved in build applications using Maven and integrated with CI servers like Jenkins to build jobs.

**Environment:** AWS EC2, CDC (Change Data Capture), Sqoop. Amazon Data Lake, Oracle, MapReduce, Pyspark

**Augusta HI-TECH Soft Sol, India Jun 2014 – July 2015**

**Software Engineer**

**Description**: Augusta is a focused product development, software services and technology consulting company.

* Utilized Oracle Designer 6i to perform data modelling
* Documented Tech Specs for the proposed database design
* Devised PL/SQL packages and procedures for the back end processing of the proposed database design
* Delivered PL/SQL training session for co-workers to educate about the latest PL/SQL features, PL/SQL performance tuning
* Facilitated management of database
* Designed tables, synonyms, sequences, views, PL/SQL stored procedures and triggers
* Performed testing and code review
* Conducted performance tuning of the overall system by eliminating redundant joins, creating indexes, removing redundant code. Developed UNIX shell scripts to perform a nightly refresh of the test system from Production databases. Monitored user profiles, roles and privileges for the Sybase database
* Maintaining technical, functional documentation and all the deliverables.

**Environment:** Oracle, Forms Developer, Report Developer, HP UX 11i, UNIX SUN Solaris 5.8, Oracle 9i, Putty, WINSCP, Pl/SQL, SQL\*Plus, Unix scripting