**AKASH VERMA**

**Data Engineer**

**| Phone: 972 848 4595| Email: akashhadoop01@gmail.com**

**PROFESSIONAL SUMMARY:**

* Around 9 years of work experience in IT, which includes experience in Data Engineering and Implementation of Hadoop, Spark and cloud Data warehousing solutions.
* Extensive experience in developing Kafka producers and Kafka consumers for streaming millions of events per minute on streaming data using Pyspark, Python & Spark Streaming.
* Significant experience in Scala, Python & Shell languages.
* Experience in Spark eco-system, core, SQL, Streaming modules**.**
* Extensive experience in AWS Big Data Services, S3, ECS (Elastic Container Service), EMR Spark.
* Experienced in using security components of AWS IAM, KMS, VPC, Route 53 and Hashicorp Vault**.**
* Experience in configuring networking AWS Services, ELB, NLB, ALB for TCP & HTTP applications.
* Implemented automatic CI/CD pipelines with Jenkins to deploy Micro services in AWS ECS for streaming data, Python jobs in AWS Lambda, Containerized deployments of Java & Python.
* Strong experience production ailing end to end data pipelines on Hadoop platform.
* Good hands-on experience on DevOps Stack like Terraform, Vault, Jenkins, Ansible, Boto3, Docker, Elastic Container Service (ECS), CloudFormation and System Manager.
* Extensive experience in ETL tools like Teradata Utilities, Informatica, Oracle.
* Hands on experience with AWS Messaging & streaming systems, Kinesis data streams, SQS, SNS.
* Production deployments are continuously monitored & handled using observability tools, AWS CloudWatch, Datadog.
* Hands on experience in Azure Cloud, Azure DevOps, Azure Data Factory, Azure Data Lake Storage, Azure Cosmos NO SQL DB, Azure HD Insight Big Data Technologies (Hadoop and Apache Spark) and Data bricks.
* Expertise in NoSQL databases and hands on work experience in writing applications on NoSQL databases like Cosmos DB.
* Experienced in Data Bricks, Hive SQL, Azure CI/CD pipeline, Delta Lake, Hadoop file system, Snowflake.
* Projects implements using agile methodology, with scrum management in JIRA.
* Experience in working with different operating systems Windows, Mac, UNIX, and LINUX, EC2, Ubuntu, CentOS.
* Experience in source code & build management with Git & Enterprise GitHub with Jenkins, Artifactory.
* Experience working with NoSQL database technologies, including MongoDB, Cassandra and HBase.
* Experience in Amazon web services (AWS) and Google Cloud Platform (GCP).
* Extensive experience in creating batch & streaming data pipelines in AWS & Big Data infrastructure.
* Skilled in working with Hive data warehouse and snowflake modelling
* Strong understanding of real time streaming technologies Spark and Kafka.
* Strong understanding of Logical and Physical data base models and entity-relationship modelling.
* Strong understanding of Message Queues and streaming systems architectures, ZeroMQ, SQS, Kinesis, Kafka, Spark.
* Strong experience in Azure Cloud platforms.
* Developed Microservices solution for streaming data in AWS ECS.
* Developed REST API Service based on masking data using Python Flask, for Dev & QA regions testing for team utilities initiatives.
* Possess excellent communication, interpersonal and analytical skills along with positive attitude.
* Experience in dealing with Apache Hadoop components like HDFS, MapReduce, HiveQL, HBase, Pig, Sqoop, Big Data and Big Data Analytics.
* Hands on experience in developing ETL jobs in Hadoop eco-system using Oozie & NiFi, Streaming and batch jobs in AWS ECS & Lambda, EMR Spark and docker containers.
* Spark, ML lib, Python scripts in Unified data bricks
* Hands on experience in installing, configuring and using echo system components like Hadoop, MapReduce, HDFS, HBase, Zookeeper, Hive, Sqoop and Pig with Hortonworks Data Platform.

**TECHNICAL SKILLS:**

|  |  |  |
| --- | --- | --- |
| **Programming/Scripting Languages** | Scala, Java, Python, C, C++, SQL, Shell. | |
| **Big Data, Streaming & MQ** | Hadoop, MapReduce, HDFS, Hive, Pig and Sqoop, Spark 2.0, Kafka, Kinesis Firehose & Data Streams, SQS, SNS, ZeroMQ, Athena. | |
| **AWS** | EC2, S3, ECS, ECR, Lambda, CloudWatch, ELB, NLB, ALB, VPC, Route 53, IAM, KMS. | |
| **Other tools** | Microsoft Office tools, VSTS, VM ware, IoT, Git, Enterprise GitHub, Intellij, PyCharm, Maven and SBT. | |
| **Databases:** | NoSQL, OracleSQL, MSQL, RDBMS, Apache Cassandra, HBase, Snowflake |
| **Big data Eco System** | HDFS, Oozie, Zookeeper, Spark, Spark streaming, Kafka, NiFi |
| **UNIX Tools:** | Apache, Yum, RPM |
| **FILE FORMATS** | Txt, XML, JSON, Avro, Parquet, ORC, Protobuf |
| **Cloud Computing** | AWS |
| **Visualization and Reporting Tools** | Tableau, Microsoft Power BI, AWS QuickSight. |
| **Methodologies** | Agile, UML, Design Patterns |

**PROFESSIONAL EXPERIENCE:**

**Client: T-Mobile Seattle, WA**

**July 18 – Current**

**Role: Big Data Engineer**

**Responsibilities:**

* Worked with Hadoop 2.x version and Spark 2.x (Python and Scala).
* Used Spark for interactive queries, processing of streaming data and integration with NoSQL database for huge volume of data.
* Experienced in handling large datasets using Partitions, Spark in-memory capabilities, Broadcasts in Spark, Effective &amp; efficient Joins, Transformations and other during ingestion process itself.
* Developed custom ETL solutions, batch processing and real-time data ingestion pipeline to move data in and out of Hadoop using Python and shell scripting.
* Responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and Pyspark.
* Created monitors, alarms, notifications and logs for Lambda functions, Glue Jobs, EC2 hosts using CloudWatch
* Used AWS Glue for the data transformation, validate and data cleansing.
* Used python Boto 3 to configure the services AWS glue, EC2, S3.
* Actively sought open source/third-party tools (eg. Deequ, Glue, etc.) and developed proofs of concepts to better our data ingestion processes.

And Both Deequ and Glue were integrated into our big data ecosystem.

* Used AWS glue catalog with crawler to get the data from S3 and perform sql query operations
* Created ETL/Talend jobs both design and code to process data to target databases
* Created Talend jobs to load data into various Oracle tables. Utilized Oracle stored procedures and wrote few Java codes to capture global map variables and use them in the job
* Implemented IOT streaming with Databricks Delta tables and Delta Lake to enable ACID transaction logging.
* Developed and deployed the outcome using spark and Scala code in Hadoop cluster running on GCP.
* Launched multi-node Kubernetes cluster in Google Kubernetes Engine (GKE) and migrated the dockerized application from AWS to GCP.
* Developed pipeline for POC to compare performance/efficiency while running pipeline using the AWS EMR Spark cluster and Cloud Dataflow on GCP.
* Expertise in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate task among the team.
* Designed star schema in Big Query.
* Used Cosmos DB for storing catalog data and for event sourcing in order processing pipelines.
* Designed and developed user defined functions, stored procedures, triggers for Cosmos DB.
* Created Airflow Scheduling scripts in Python.
* Experienced in setting up workflow using Apache Airflow and Oozie workflow engine for managing and scheduling Hadoop jobs.
* Loaded salesforce Data every 15 min on incremental basis to BIGQUERY raw and UDM layer using SOQL, Google DataProc, GCS bucket, HIVE, Spark, Scala, Python, Gsutil and Shell Script.
* Used rest API with Python to ingest Data from and some other site to BIGQUERY.
* Built a program with Python and Apache beam and execute it in cloud Dataflow to run Data validation between raw source file and Bigquery tables.
* Developed pipeline for POC to compare performance/efficiency while running pipeline using the AWS EMR Spark cluster and Cloud Dataflow on GCP.
* Processed and loaded bound and unbound Data from Google pub/sub topic to Bigquery using cloud Dataflow with Python.
* Devised simple and complex SQL scripts to check and validate Dataflow in various applications.
* Built a Scala and spark based configurable framework to connect common Data sources like MYSQL, Oracle, Postgres, SQL Server, Salesforce, Bigquery and load it in Bigquery.
* Experienced in Developing Spark applications using Spark - SQL, Pyspark and Delta Lake 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.
* Written AWS Lambda code in Python for nested Json files, converting, comparing, sorting etc.
* Explored with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark SQL, Data Frame, Pair RDD and Spark on YARN.
* Worked with Sqoop import and export functionalities to handle large data set transfer between Oracle databases and HDFS.
* Designed, developed, and tested the dimensional data models using Star and Snowflake schema methodologies under the Kimball method.
* Extracted large datasets from Teradata using utilities like fastexport, bteq and Load into Verticas.
* Worked on query optimization and ETL loading on Teradata, Vertica DB performance tuning
* Migrated from Lucas framework to Confidential framework using Talend DI and Vertica analytic server.
* Developed a detailed project plan and helped manage the data conversion migration from the legacy system to the target snowflake database.
* Developed Spark jobs to clean data obtained from various feeds to make it suitable for ingestion into Hive tables for analysis.
* Worked on importing data into HBase using HBase Shell and HBase Client API.
* Used Cassandra (CQL) with Java APIs to retrieve data from Cassandra tables.
* Developed Python-based API (RESTful Web Service) to track revenue and perform revenue analysis.
* Designed, developed Hadoop eco system components, installation, configuration, supporting and monitoring of Hadoop clusters using Apache, Cloudera distributions, Azure data bricks and AWS.
* Developed Spark applications using Spark - SQL in Databricks for data extraction.
* Used SOAP REST (JAX-RS) web services to provide/consume the Web services from/to down-stream systems.
* Developed Custom Input Formats in Spark jobs to handle custom file formats.
* Configured Oozie workflow to run multiple Hive jobs which run independently with time and data availability.
* Utilized Hive tables and HQL queries for daily and weekly reports. Worked on complex data types in Hive like Structs and Maps.
* Developed HiveQL queries for trend analysis and pattern recognition on user data.
* Helped this regional bank streamline business processes by developing, installing and configuring Hadoop ecosystem components that moved data from individual servers to HDFS.
* Spark streaming is used for a single frame work to satisfy all their processing needs.
* Imported data from AWS S3 into Spark RDD, performed transformations and actions on RDDs.
* Worked with Hadoop distribution of Hortonworks.
* Created HBase tables to load large sets of structured, semi-structured and unstructured data coming from UNIX, NoSQL and a variety of portfolios.
* Supported code/design analysis, strategy development and project planning.
* Created reports for the BI team using Sqoop to export data into HDFS and Hive.
* Developed multiple Spark jobs in Scala & Python for data cleaning and pre-processing.
* Assisted with data capacity planning and node forecasting.
* Used Azure Portal, Azure PowerShell, Storage Accounts, and Azure Data Management.
* Migrated databases to cloud platform SQL Azure and as well the performance tuning.
* Build and maintain the environment on Azure IAAS, PAAS.
* Collaborated with the infrastructure, network, database, application and BI teams to ensure data quality and availability.
* Developed Informatica mappings, Lookups, Reusable Components, Sessions, Work Flows etc. (on ETL side) as per the design documents/communication
* Fine-tuned Informatica maps for performance optimization
* Developed complex mappings and SCD type-I, Type-II and Type III mappings in Informatica to load the data from various sources using different transformations like Source Qualifier.
* Designing ETL processes using Informatica to load data from Flat Files, Oracle and Excel files to target Oracle Data Warehouse database.

**Environment:**  Spark, Spark SQL, HBase, Hive, Oozie, Informatica, HQL, Sqoop, snowflake Oozie, Java, Scala, Python, Shell scripting, Maven, GIT, Tableau.

**Client: BCBS, Dallas Tx**

**Aug 2017 to July 2018**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Experienced in writing Spark Applications in Scala and Python (PySpark).
* Imported Avro files using Apache Kafka and did some analytics using Sparking Scala.
* Extracting real time data using Kafka and Spark streaming by Creating D streams and converting them into RDD, processing it and stored it into Cassandra.
* Configured, deployed and maintained multi-node Dev and Test Kafka Clusters.
* Using Spark-Streaming APIs to perform transformations and actions on fly for building the common learner data model which gets the data from Kafka in near real time and persists into Cassandra.
* Created Databricks notebooks using SQL, Python and automated notebooks using jobs.
* Created Spark clusters and configuring high concurrency clusters using Azure Databricks to speed up the preparation of high-quality data.
* Designed and constructed AWS Data pipelines using various resources in AWS including AWS API Gateway to receives response from AWS lambda and retrieve data from snowflake using lambda function and converted the response into Json format using Database as Snow Flake, DynamoDB, AWS Lambda function and AWS S3.
* Good experience in Snowflake data warehouse, developed data extraction queries, automatic ETL for data loading from Data Lake.
* Used Azure data factory and data Catalog to ingest and maintain data sources
* Used Informatica features to implement Type 2 changes in slowly changing dimensions to update the data in dimension tables.
* Tuned performance of Informatica session for large data files by increasing block size, data cache size and target-based commit interval.
* Worked on **Kafka** and REST API to collect and load the data on Hadoop file system also used Sqoop to load the data from relational databases
* Used Scala sbt to develop Scala coded spark projects and executed using spark-submit.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
* Developed the batch scripts to fetch the data from AWS S3storage and do required transformations in Scala using Spark framework.
* Data validation between Hive target tables and Redshift source tables using automation scripts.
* Building the Cassandra nodes using AWS & amp; setting up the Cassandra cluster using Ansible automation tools
* Worked and learned a great deal from Amazon Web Services (AWS) cloud services like EC2, S3, EMR, EBS, RDS and VPC.
* Developed Scala scripts, UDF’s using both Data frames/SQL and RDD/MapReduce in Spark for Data Aggregation, queries and writing data back into RDBMS through Sqoop.
* Developed Oozie Bundles to Schedule Pig, Sqoop and Hive jobs to create data pipelines.
* Developed Hive queries to do analysis of the data and to generate the end reports to be used by business users.
* Used spark and spark-SQL to read the parquet data and create the tables in hive using the Scala API.
* Design solution for various system components using Microsoft Azure.
* Configures Azure cloud services for endpoint deployment.
* Designed end to end Azure cloud-based analytics dashboard for a state government for showing real time updates for the recently their state assembly elections 2016. Solution utilized PowerBI, Enterprise Gateway and Azure SQL Server.
* Written generic extensive data quality check framework to be used by the application using impala.
* Experience in NoSQL Column-Oriented Databases like Cassandra and its Integration with Hadoop cluster.
* Worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using python (Pyspark).
* Developed script to load the data into Redshift from Hive tables.
* Hands on experience in AWS Cloud in various AWS services such as Redshift cluster, Route 53 domain configuration.
* Providing guidance to the development team working on Pyspark as ETL platform.
* Used Pyspark jobs to run on Kubernetes Cluster for faster data processing
* Automated resulting scripts and workflow using Apache Airflow and shell scripting to ensure daily execution in production.
* Installed and configured Apache Airflow for S3 bucket and Snowflake data warehouse and created dags to run the Airflow.
* Deployed application to GCP using Spinnaker (rpm based)
* Launched multi-node Kubernetes cluster in Google Kubernetes Engine (GKE) and migrated the dockerized application from AWS to GCP.
* Develop and deploy the outcome using spark and Scala code in Hadoop cluster running on GCP.
* Worked on Docker based containers for using Airflow.

**Environment:** Hadoop, Hive, Impala, Oracle, Spark, Python, Pig, Sqoop, snowflake, Oozie, Map Reduce, GIT, HDFS, Cassandra, Apache Kafka, Storm, Linux, Solr, Confluence, Jenkins.

**Client: TMW Systems, Chennai, IN**

**April 2015 to Nov 2016**

**Hadoop/Spark Developer**

**Responsibilities:**

* Handling the importing of data from various data sources (media, MySQL) and performing transformations using Hive, MapReduce.
* Developed multiple Map Reduce jobs in Java for data cleaning and pre-processing
* Ran Pig scripts on Local Mode, Pseudo Mode, and Distributed Mode in various stages of testing.
* Configured Hadoop cluster with Name node and slaves and formatted HDFS.
* Performed Importing and exporting data from SQL server to HDFS and Hive using Sqoop.
* End-to-end involvement in data ingestion, cleansing, and transformation in Hadoop.
* Logical implementation and interaction with HBase.
* Implemented Apache PIG scripts to load data from and to store data into Hive.
* Created HBase tables to load large sets of structured, semi-structured and unstructured data coming from UNIX, NoSQL and a variety of portfolios.
* Used Impala to read, write and query the Hadoop data in HDFS from Cassandra and configured Kafka to read and write messages from external programs.
* Optimizing existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD's.
* Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate task among the team.
* Implemented Spark using Scala and Spark SQL for faster testing and processing of data.
* Create a complete processing engine, based on Cloudera distribution, enhanced performance.
* Developed data pipeline using Flume, Spark and Hive to ingest, transform and analysing data.
* Wrote Flume configuration files for importing streaming log data into MongoDB with Flume.
* Designed and Modified Database tables and used HBASE Queries to insert and fetch data from tables.
* Developing and supporting Map-Reduce Programs running on the cluster.
* Implemented the file validation framework, UDFs, UDTFs and DAOs.
* Preparation of Technical architecture and Low -level design documents.
* Tested raw data and executed performance scripts.

**Environment**: Linuxsuse12, eclipse photon(64bit), jdk1.8.0, Hadoop2.9.0, flume 1.7.0, HDFS, MapReduce, Pig0.16.0, Spark, Hive 2.0, Apache-Maven3.0.3

**Client: Oracle Hyderabad, India**

**June 2012 - April 2015**

**Role: Java Developer**

**Responsibilities:**

* Participated in gathering business requirements and converting them into detailed design documents.
* Developed Screens by using Swing components like JFrame, JButton. JTextfield.
* Project was developed following Agile and Scrum methodologies.
* Developed views using JSPs and struts tags. Using Tiles framework, improved UI flexibility and providing single point of maintenance.
* Developed application using JavaScript for Web pages to add functionality, validate forms, communicate with the server.
* Written Java program to retrieve data from HDFS and providing REST services.
* Skills gained on web-based REST API, SOAP API, and Apache for real-time data streaming.
* Extensively used Core Java concepts like Collections, Exception Handling, and Generics during development of business logic.
* Extensively written Core Java and Multi-Threading code in application.
* Provided connections using JDBC to the database and developed SQL queries to manipulate the data on DB.
* Written JDBC statements, prepared statements and callable statements in Java, JSPs and Servlets.
* Developed application components using JSPs, EJB's, Value Objects and model layer logic.
* Performed CRUD operations like Update, Insert and Delete data in Oracle.
* Created Functional Test cases and achieved bug fixes.
* Code review and function testing for better client interface and usability.
* Participation in meeting with team, senior management and client stakeholders.
* Worked with Web logic application server set up & deployed the application on it. Have good working background with J2EE Frameworks like Servlets, JSP.
* Responsible for building scalable distributed data solutions using Data tax Cassandra.
* Implemented DAO layer using IBATIS and wrote queries for persisting demand core banking related information from the backend system using Query tool
* Developed web-based presentation using JSP and Servlet technologies and implemented MVC pattern using STRUTS framework.
* Used Spring Rest Controllers and Services classes to support migration to Spring framework
* Experienced in developing the UNIX Shell Scripts and PERL Scripts to execute the scripts and manipulate files and directory.html
* Used Ajax to communicate with the server to get the asynchronous response.
* Developed code for Web services using XML, SOAP and used SOAP UI tool for testing the services.
* Developed SOAP and WSDL files for Web Services interacting with business Logic.
* Used Log4J for logging the user events. Maven for compilation and building JAR, WAR and EAR files.
* Used JUnit for unit testing and Continuum for integration testing.

**Environment: Java 1.4, HTML, CSS, JSP 2.0, Servlets, Struts, EJB, JDBC, SQL, Oracle, Swing, Eclipse, MS Office, Windows, JPA Annotations, WebLogic.**

**Educational details:**

* Masters- Gannon University, C.SC, 2019
* Bachelors- U.P Technological University, IT, 2012