EMPLOYEE NAME

Email: [gmail@gmail.com](mailto:gmail@gmail.com)

Phone: (999)-999-9999

Hadoop Big Data Engineer and Developer with skills in legacy Hadoop Ecosystems, Cloudera Hadoop, Hortonworks Hadoop, Amazon Web Services and Microsoft Azure. Skilled in the use of Spark, Spark Streaming, Spark SQL, Kafka, python and Scala.

Professional Summary

* 5+ years’ experience in Hadoop Big Data
* Proven success in team leadership, focusing on mentoring team members and managing tasks for efficiency.
* Worked with various stakeholders for gathering requirements to create solutions
* Harvesting customer data (Unix)
* Writing transformers to prepare customer data for ingestion into the in-house platforms (Groovy)
* Training models on customer data (NLP Model training interface)
* Setting up in-house company software tools on different environments (Unix,elastic search)
* Converting lexicons from other systems to regexes and deploying them on company platforms to demonstrate that the platforms ability to handle legacy methods. (Java,Unix).
* Leasing with other teams in the bank to bridge the gap between datasets
* Loading data onto Hadoop into a data warehouse
* Writing Scala Transformations to cleanse data, using functional programming for immutable data (Scala/Spark)
* Writing Transformations for data insight, understanding shape of data, predicting and extracting business value from data (Scala/Spark)
* Merging incompatible datasets using fuzzy logic (String approximation, elastic joins in Java/Scala/Spark).
* Creating and extending data Schema for new and existing datasets
* Creating and extending relational models
* Testing Scala code relating to message queuing and testing it on ACTIVE MQ
* Automating data processes making use of a larger team’s Kafka messaging system for receiving data and automatically triggering transformations (Scala/Spark/Kafka)
* Creating polishing data for user friendly demos in Hadoop Hive (Data Mart)

Technical Skills

|  |  |
| --- | --- |
| IDE: | Jupyter Notebooks (formerly iPython Notebooks), Eclipse, IntelliJ, PyCharm |
| PROJECT METHODS: | Agile, Scrum,Test-Driven Development, Unit Testing |
| HADOOP DISTRIBUTIONS: | Hadoop,Cloudera Hadoop, Hortonworks Hadoop |
| CLOUD PLATFORMS: | Amazon AWS, GCP |
| CLOUD SERVICES: | Databricks |
| CLOUD DATABASE & TOOLS: | Apache Hbase, SQL |
| PROGRAMMING LANGUAGES: | Spark, Spark Streaming, Java, Python, Scala, PySpark |
| SCRIPTING: | Hive, MapReduce, SQL, Spark SQL, Shell Scripting |
| CONTINUOUS INTEGRATION (CI-CD): |  |
| VERSIONING: | Git, GitHub |
| PROGRAMMING METHODOLOGIES: | Object-Oriented Programming, Functional Programming |
| FILE FORMAT AND COMPRESSION: | CSV, JSON, Avro, Parquet, ORC |
| FILE SYSTEMS: | HDFS |
| ETL TOOLS: | Apache Camel, Flume, Kafka,Sqoop |
| DATA VIZIUALIZATION TOOLS: | Tableau |
| SEARCH TOOLS: | Apache Lucene, Elasticsearch |
| SECURITY: | Kerberos |
| AWS: |  |
| Data Query: | Spark SQL, Data Frames |

Professional Experience

**TESCO, BIG DATA DEVELOPER  June 2019 – Present**

• Implemented solutions for ingesting data from various sources and processing the Data-at-Rest utilizing Big Data technologies such as Hadoop, Map Reduce Frameworks, HBase, Hive.

• Exploring with the Spark for improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.

• Experienced developing and maintaining ETL jobs.

• Performed data profiling and transformation on the raw data using Pig, Python, and oracle

• Experienced with batch processing of data sources using Apache Spark.

• Developing predictive analytics using Apache Spark Scala APIs.

• Created Hive External tables and loaded the data into tables and query data using HQL.

• Used Sqoop to efficiently transfer data between databases and HDFS and used Flume to stream the log data from servers.

• Developed Spark code using Scala and Spark-SQL for faster testing and data processing.

• Imported millions of structured data from relational databases using Sqoop import to process using Spark and stored the data into HDFS in CSV format.

• Developed Spark streaming application to pull data from cloud to Hive table.

• Used Spark SQL to process the huge amount of structured data.

• 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 Spark code using Scala and Spark-SQL/Streaming for faster processing of data.

• Load the data from different sources such as HDFS or HBase into Spark RDD and implement in memory data computation to generate the output response.

• Extracted files from Mongo DB through Sqoop, placed in HDFS, and processed.

• Developed complete end to end Big-data processing in hadoop eco system.

• Created automated python scripts to convert the data from different sources and to generate the ETL pipelines.

• Converted some existing hive scripts to Spark applications using RDD's for transforming data and loading into HDFS.

• Extensively worked on Text, ORC, Avro and Parquet file formats and compression techniques like snappy, Gzip and zlib.

• Extensively used Hive optimization techniques like partitioning, bucketing, MapJoin and parallel execution.

• Creation, configuration and monitoring Shards sets. Analysis of the data to be shared, choosing a shard Key to distribute data evenly.

• Developed scripts for collecting high frequency log data from various sources and integrating it into HDFS using Flume; staging data in HDFS for further analysis.

• Worked with different data science teams and provided respective data as required on an ad-hoc request basis

• Hands-on experience in Microsoft Azure Cloud Services (PaaS & IaaS), Storage, Active Directory, Application Insights, Internet of Things (IoT), Azure Search, Key Vault, Visual Studio Online (VSO) and SQL Azure.

• Responsible for ensuring Systems & Network Security, maintaining performance and setting up monitoring using Cloud Watch and Nagios.

• Experience in working on version controller tools like GitHub (GIT), Subversion (SVN) and software builds tools like Apache Maven, Apache Ant.

• Extensively worked on CI/CD pipeline for code deployment by engaging different tools (Git, Jenkins, CodePipeline) in the process right from developer code check-in to Production deployment

**G-RESEARCH, BIG DATA ENGINEER**  **Jan 2017 – May 2019**

* Implemented Kafka producers and consumers applications on Kafka cluster setup with help of Zookeeper
* Have knowledge on partition of Kafka messages and setting up the replication factors in Kafka Cluster.
* Configured Kafka producer with API endpoints using JDBC Autonomous REST Connectors
* Configured a multinode –multibroker cluster of 10 Nodes and 30 brokers for consuming high volume, high velocity data
* Implemented parser, query planner, query optimizer, and native query execution using replicated logs combined with indexes
* Developed distributed query agents for performing distributed queries against shards
* Developed producer /consumer scripts to process JSON response in python and formatted the response into a data frame using a schema containing, country code, artist name, number of plays and genre to parse the JSON
* Experience in Kafka cluster’s units, integrations and performance tests
* Developed DBC/ODBC connectors between Hive and Spark for the transfer of the newly populated data frame
* Developed scripts for collecting high frequency log data from various sources and integrating it into HDFS using Flume; staging data in HDFS for further analysis
* Proficient in writing complex queries the API into Apache Hive on Hortonworks Sandbox
* Utilized HiveQL to query the data to discover trends from week to week
* Configure and deploy production ready multinode hadoop services Hive, Sqoop, Flume, Oozie on the Hadoop cluster with latest patches
* Created Hive queries to summarizing and aggregating business queries by comparing Hadoop data with historical metrics
* Loaded into ingested data into Hive Managed and External tables
* Wrote custom user define functions (UDF) for complex Hive queries (HQL)
* Performed upgrades, patches and bug fixes in Hadoop in a cluster environment
* Developed shell scripts to automate workflows to pull data from various databases into Hadoop framework for users to access the data through Hive based views
* Developed Hive Queries for analyzing data in Hive warehouse using Hive Query Language
* Used Cloudera Manager for installation and management of single-node and multi-node Hadoop cluster
* Developed shell scripts for automating the process of data loading

**BARCLAYS, BIG DATA DEVELOPER JAN 2015 - DEC 2016**

* Hands on experience in installing, configuring HDFS and launching Hadoop jobs
* Transformed the data by applying ETL processes using Hive with large sets of structured, semi structured and unstructured data
* Writing Hive Queries and UDF’s on different datasets and joining them.
* Responsible for Optimizing and tuning Hive to improve performance and solve related issues with good understanding of Joins, Group and aggregation
* Used Sqoop for data transfer between MS SQL and HDFS
* Interacted with Team and Analysis, Design and Develop database using ER Diagram, Normalization and relational database concept
* Involved in Design, Development and testing of the system
* Developed SQL Server Stored Procedures, Tuned SQL Queries (using Indexes and Execution Plan)
* Developed User Defined Functions and created Views
* Created Triggers to maintain the Referential Integrity
* Extensively used Joins and sub-Queries to simplify complex queries involving multiple tables and also optimized the procedures and triggers to be used in production.
* Performance Tuning in SQL Server 2000 using SQL Profiler and Data Loading
* Installing SQL Server Client-side utilities and tools for all the front-end developers/programmer
* Created DTS package to schedule the jobs for batch processing.
* Involved in performance tuning to optimize SQL queries using query analyzer
* Created indexes, Constraints and rules on database objects for optimization.

**VODAFONE, HADOOP DEVELOPER MAR 2014 – DEC 2014**

* Worked with technology and business groups for Hadoop migration strategy.
* Research and recommend suitable technology stack for Hadoop migration considering current enterprise architecture.
* Validated and Recommended on Hadoop Infrastructure and data center planning considering data growth.
* Transferred data to and from cluster, using Sqoop and various storage media such as flat files.
* Developed MapReduce programs and Hive queries to analyze sales pattern and customer satisfaction index over the data present in various relational database tables.
* Orchestrated hundreds of Sqoop scripts, Pig scripts, Hive queries using Oozie workflows and sub-workflows
* Developed pig scripts for replacing the existing home loans legacy process to the Hadoop
* Worked extensively in performance optimization by adopting/deriving at appropriate design patterns of the MapReduce jobs by analyzing the I/O latency, map time, combiner time, reduce time etc.
* Developed Pig scripts in the areas where extensive coding needs to be reduced. ● Followed agile methodology for the entire project

Education & Training

**University of Brighton, UK - BSC (Hons.) Mathematics and Computing - 2007**

Certifications

* Hadoop 101 - Cognitive Class
* Big Data 101 - Cognitive Class
* AZ-900 Microsoft Azure Fundamentals Certified