|  |
| --- |
| **Omar Farooqui**  **Sr. Hadoop/Big Data Developer**  Email: omar.farooqui718@gmail.com  Contact No: 331-239-3273 |

**PROFESSIONAL SUMMARY**

* Over 9+ years of software development with 4+ years as Hadoop developer in Big Data/ Hadoop/Spark technology development.
* Experience in developing applications that perform large scale distributed data processing using big data ecosystem tools like HDFS, YARN, Sqoop, Flume, Kafka, MapReduce, Pig, Hive, Spark, Spark SQL, Spark Streaming, HBase, Cassandra, MongoDB, Mahout, Oozie, and AWS.
* Good functional experience in using various Hadoop distributions like Hortonworks, Cloudera, and EMR.
* Good understanding in using data ingestion tools- such as Kafka, Sqoop and Flume.
* Experienced in performing in-memory real time data processing using Apache Spark.
* Good experience in developing multiple Kafka Producers and Consumers as per business requirements.
* Extensively worked on Spark components like Spark SQL, MLlib, GraphX, and Spark Streaming.
* Configured Spark Streaming to receive real time data from Kafka and store the stream data to HDFS and process it using Spark and Scala.
* Experience in spinning up different Azure resources using ARM templates
* Experience in setting up Azure Big data environment using Azure HD Insight
* Experience in Amazon AWS cloud Administration and actively involved highly available, Scalability, cost effective and fault tolerant systems using multiple AWS services.
* Experience with an in-depth level of understanding in the strategy and practical implementation of AWS Cloud-Specific technologies including IAM, EC2, EMR, SNS, RDS, Redshift, Athena, Dynamo DB, Lambda, Cloud Watch, Auto-Scaling, S3, and Route 53.
* Developed quality code adhering to Scala coding standards and best practices.
* Experienced in the Hadoop ecosystem components like Hadoop Map Reduce, Cloudera, Hortonworks, HBase, Oozie, Flume, Kafka, Hive, Scala, SPARK SQL, Data Frames, SQOOP, MySQL,Unix commands, Cassandra, MongoDB, Tableau tool and related Big data tools.
* Hands on developing and debugging YARN (MR2) Jobs to process large Datasets.
* Experience in support of IBM Mainframe applications - MVS, COBOL, JCL, PROCs, VSAM, File aid, JCL, SQL and DB2.
* Hands on experiences with Hadoop stack. (HDFS, Map Reduce, YARN, Sqoop, Flume, Hive-Beeline, Impala, Tez, Pig, Zookeeper, Oozie, Solr, Sentry, Kerberos, Centrify DC, Falcon, Hue, Kafka, Storm).
* Experienced in working with Hadoop/Big-Data storage and analytical frameworks over Azure cloud
* Experience in migrating map reduce programs into Spark RDD transformations, actions to improve performance.
* Worked on standards and proof of concept in support of CDH4 and CDH5 implementation using AWS cloud infrastructure.
* Involved in collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFS for further analysis.
* Extensive working experience with data warehousing technologies such as HIVE.
* Good experience on partitions, Bucketing concepts. Designed and managed them and created external tables in Hive to optimize performance.
* Expertise in writing Hive and Pig queries for data analysis to meet the business requirement.
* Extensively worked on Hive and Sqoop for sourcing and transformations.
* Extensive work experience in creating UDFs, UDAFs in Pig and Hive.
* Involved in deploying applications on Azure. Involved in setting big data cluster using Azure HDInsight
* Good experience in using Impala for data analysis.
* Experience on NoSQL databases such as HBase, Cassandra, MongoDB, and DynamoDB.
* Implemented CRUD operations using CQL on top of Cassandra file system.
* Manage and review HDFS data backups and restores on Production cluster.
* Experience in creating data-models for client's transactional logs, analyzed the data from Cassandra tables for quick searching, sorting, and grouping using the Cassandra Query Language (CQL).
* Expert knowledge on MongoDB data modeling, tuning, disaster recovery and backup.
* Hands on experience on Ad-hoc queries, Indexing, Replication, Load balancing, Aggregation in MongoDB.
* Extending HIVE and PIG core functionality by using custom User Defined Function's (UDF), User DefinedTable-Generating Functions (UDTF) and User Defined Aggregating Functions (UDAF) for Hive and Pig.
* Expertise in relational databases like MySQL, SQL Server, DB2, and Oracle.
* Great understanding on Solr to develop search engine on unstructured data in HDFS.
* Experience in cloud platforms like AWS, Azure.
* Working closely with Azure to migrate the entire Data Centers to the cloud using Cosmos DB, ARM templates.
* Extensively worked on AWS services such as EC2 instance, S3, EMR, Cloud Formation, Cloud Watch, and Lambda.
* Expertise to handle tasks in Red Hat Linux includes upgrading RPMS using YUM, kernel, configure SAN Disks, Multipath and LVM file system.
* Good knowledge in understanding the security requirements for Hadoop and integrate with Kerberos authentication and authorization infrastructure.
* Experience on ELK stack and Solr to develop search engine on unstructured data in HDFS.
* Implemented ETL operations using Big Data platform.
* Involved in identifying job dependencies to design workflow for Oozie & YARN resource management.
* Experience working with Core Java, J2EE, JDBC, ODBC, JSP, Java Eclipse, EJB and Servlets.
* Strong experience on Data Warehousing ETL concepts using Informatica, and Talend.

**Education**

* Bachelor's in computer Science, Lincoln College of Technology, IL 2012

**Skills**

* Big Data : Hadoop, HDFS, MapReduce, Pig, Hive, Spark, Kafka, Flume, Sqoop, Impala, Oozie, Zookeeper, YARN, Hue.
* Hadoop Distributions : Cloudera (CDH4, CDH5), Hortonworks, EMR.
* Programming Languages : C, Java, Python, Scala.
* Database : NoSQL, HBase, Cassandra, MongoDB, MySQL, Oracle, DB2, PL/SQL, Microsoft SQL Server.
* Cloud Services : AWS, Azure.
* Frameworks : Spring, Hibernate, Struts.
* Scripting Languages : JSP, Servlets, JavaScript, XML, HTML.
* Java Technologies : Servlets, JavaBeans, JSP, JDBC, EJB.
* Application Servers : Apache Tomcat, Web Sphere, WebLogic, JBoss.
* ETL Tools : Informatica, Talend.

**Work Experience**

|  |
| --- |
| **Sr. Hadoop/Big Data Developer**  **AT&T, Plano, Texas (IBM)**  Dec 2020 – Current |

* Designing and Creating Azure Data Factory (ADF) extensively for ingesting data from different source systems like relational and non-relational to meet business functional requirements.
* 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 at least one Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
* Worked extensively on Hadoop Componiments such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, YARN, Spark and Map Reduce programming
* Converting the existing relational database model to Hadoop ecosystem.
* Designing and Creating Azure Data Factory (ADF) extensively for ingesting data from different source systems like relational and non-relational to meet business functional requirements.
* Exploring with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark   
  Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN
* Extract Real time feed using Kafka and Spark Streaming and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in HDFS.
* Used Spark and Spark-SQL to read the parquet data and create the tables in hive using the Scala API.
* Created pipeline for processing structured and unstructured streaming data using spark streaming and stored the filtered data into S3 as parquet files.
* 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 at least one Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
* Worked with Linux systems and RDBMS database on a regular basis to ingest data using Sqoop.
* Developed Schedulers that communicated with the Cloud based services (AWS) to retrieve the data.
* Strong experience in working with ELASTIC MAPREDUCE and setting up environments on Amazon AWS EC2 instances.
* Ability to spin up different AWS instances including EC2-classic and EC2-VPC using cloud formation templates.
* Collected data using Spark Streaming from AWS S3 bucket in near-real-time and performs necessary Transformations and Aggregations to build the data model and persists the data in HDFS.
* Imported the data from different sources like AWS S3, LFS into Spark RDD.
* Experienced in working with Amazon Web Services (AWS) EC2 and S3 in Spark RDD
* Managed and reviewed Hadoop and HBase log files.
* Worked extensively with importing metadata into Hive and migrated existing tables and applications to work on Hive and AWS cloud.
* Designed and implemented HIVE queries and functions for evaluation, filtering, loading and storing of data.
* Analyze table data and implement compression techniques like Teradata Multivalued compression
* Involved in ETL process from design, development, testing and migration to production environments.
* Involved in writing the ETL test scripts and guided the testing team in executing the test scripts.
* Involved in performance tuning of the ETL process by addressing various performance issues at the extraction and transformation stages.
* Writing Hadoop MapReduce jobs to run on Amazon EMR clusters and creating workflows for running jobs
* Generating analytics reporting on probe data by writing EMR (elastic map reduce) jobs to run on Amazon VPC cluster and using Amazon data pipelines for automation.
* Model complex ETL jobs that transform data visually with data flow or by using compute services Azure Databricks, Azure Blob Storage, Azure SQL Database, Cosmos DB.
* Worked with Elastic MapReduce (EMR) on Amazon Web Services (AWS).
* Have good understanding of Teradata MPP architecture such as Partitioning, Primary Indexes,
* Good knowledge in Teradata Unity, Teradata Data Mover, OS PDE Kernel internals, Backup and Recovery
* Created HBase tables to store variable data formats of data coming from different portfolios.
* Created Partitions, Buckets based on State to further process using Bucket based Hive joins.
* Involved in transforming data from Mainframe tables to HDFS, and HBase tables using Sqoop.
* Creating Hive tables and working on them using HiveQL.
* Building and creating scripts for data modelling, mining for easier access to Azure Logs, App Insights to
* Creating and truncating HBase tables in hue and taking backup of submitter ID
* Developed data pipeline using Kafka to store data into HDFS.
* Used Spark API over Hadoop YARN as execution engine for data analytics using Hive.
* Continuous monitoring and managing the Hadoop cluster through Cloudera Manager.
* Involved in review of functional and non-functional requirements.
* Developed ETL Process using HIVE and HBASE.
* Worked as an ETL Architect/ETL Technical Lead and provided the ETL framework Solution for the Delta process, Hierarchy Build and XML generation.
* Prepared the Technical Specification document for the ETL job development.
* Responsible to manage data coming from different sources.
* Loaded the CDRs from relational DB using Sqoop and other sources to Hadoop cluster by using Flume.
* Experience in processing large volume of data and skills in parallel execution of process using Talend functionality.
* Installed and configured Apache Hadoop, Hive and Pig environment.

**Environment:**Azure Data Factory (ADF v2), Azure Databricks (PySpark), Azure Data Lake, Spark (Python/Scala), Hadoop, HDFS, pig, Hive, Flume, Sqoop, Oozie, Python, Shell Scripting, SQL Talend, Spark, HBase, Elastic search, Linux- Ubuntu, Kafka.

|  |
| --- |
| **Hadoop Developer**  **Ford Motor Company, Dearborn, Michigan**  Mar 2019 – Oct 2020 |

* Responsible for building scalable distributed talend data solutions using Hadoop.
* Involved in start to end process of Hadoop jobs that used various technologies such as Sqoop, PIG, Hive, Spark and Python scripts (for scheduling of jobs) Extracted and loaded data into Data Lake environment.
* Expertise in designing and deployment of Hadoop cluster and different Big Data analytic tools including Pig, Hive, HBase, Oozie, Zookeeper, Sqoop, flume, Kafka, Spark,Impala, Cassandra with Cloudera.
* Developed Spark code using Python and Spark-SQL/Spark Streaming for faster testing and processing of data.
* Involved in migration from Hadoop System to Spark System.
* Primarily involved in Data Migration process using Azure by integrating with GitHub repository and Jenkins.
* Developed Sqoop scripts to import and export data from RDBMS into HDFS, HIVE and handled incremental loading on the customer and transaction information data dynamically.
* Extending HIVE and PIG core functionality by using custom User Defined Function's (UDF), User Defined Table-Generating Functions (UDTF) and User Defined Aggregating Functions (UDAF) for Hive and Pig using python.
* Performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
* Set up Linux Users, and tested HDFS, Hive, Pig and MapReduce Access for the new users.
* Optimized Hadoop clusters components: HDFS, Yarn, Hive, Kafka to achieve high performance.
* Integrated Oozie with the rest of the Hadoop stack supporting several types of Hadoop jobs such as MapReduce, Pig, Hive, and Sqoop as well as system specific jobs such as Java programs and Python scripts.
* Optimizing of existing algorithms in Hadoop using Spark Context, Spark -SQL, Data Frames and Pair RDD's.
* Handled large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations and other during ingestion process itself.
* Used AmazonS3 as a storage mechanism and written python scripts that dump the data into S3.
* Designed, developed, and did maintenance of data pipelines in a Hadoop and RDBMS environment with both traditional and non-traditional source systems using RDBMS and NoSQL data stores for data access.
* Build and evaluated an 18 node HDF NiFi/Kafka cluster in Azure for a specific use case requirement to ingest and process real time Drilling data into NiFi and write to Kafa/ Azure Datalake.
* Development of Spark jobs for Data cleansing and Data processing of flat files.
* Worked on Job management using Fair scheduler and Developed job processing scripts using Oozie workflow.
* Worked with different File Formats like TEXTFILE, SEQUENCEFILE, AVROFILE, ORC, and PARQUET for Hive querying and processing.
* Involved in creating Spark cluster in HDInsight by create Azure compute resources with spark installed and configured.
* Used Spark-Streaming APIs to perform necessary transformations and actions on the fly for building the common learner data model which gets the data from Kafka in near real time and Persists into Cassandra.
* Developed Spark Applications in Scala and build them using SBT.
* Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive.
* Have been working with AWS cloud services (VPC, EC2, S3, EMR, DynamoDB,SNS, SQS).
* Developed Scala scripts, UDAFs using both Data frames/SQL/Data sets and RDD/Map Reduce in Spark
* 1.6 for Data Aggregation, queries and writing data back into OLTP system through Sqoop.
* Involved in creating Hive tables, loading and analyzing data using hive queries.
* Implemented schema extraction for Parquet and Avro file Formats in Hive.
* and analysis.
* Experience in working with Hadoop 2.x version and Spark 2.x (Python and Scala).
* Worked on a POC to compare processing time of Impala with Apache Hive for batch applications to implement the former in project.
* Worked extensively with Git, Sqoop for importing metadata from Oracle.
* Installation & configuration of Apache Hadoop on Amazon AWS (EC2) system.
* Implemented Partitioning, Dynamic Partitions, Buckets in HIVE.
* Used Talend Open Studio for getting the data.
* Worked on Git, Continuous Integration of application using Jenkins.
* Used Reporting tools like Tableau to connect with Hive for generating daily reports of data.
* Collaborated with the infrastructure, network, database, application, and BA teams to ensure data quality and availability.

**Environment:** Hadoop YARN, Spark -Core, Spark Streaming, Spark SQL, Scala, Kafka, Hive, HBase, Pig, Sqoop, MapR, Amazon AWS, Azure, Impala, Cassandra, Tableau, Oozie, Jenkins, Talend, Cloudera, Oracle 12c, RedHat Linux, Python language.

|  |
| --- |
| **Hadoop Developer**  **Century Bank, Medford, Massachusetts**  Jan 2017 – Feb 2019 |

* Installed/Configured/Maintained Apache Hadoop clusters for application development and Hadoop tools like Hive, Pig, HBase, Flume, Oozie Zookeeper and Sqoop.
* Installed and configured Hadoop Map reduce, HDFS, developed multiple Map Reduce jobs in java for data cleaning and pre-processing.
* Collected the logs from the physical machines and the Open Stack controller and integrated into HDFS using Flume.
* Developing the applications using programming languages like Scala and Spark.
* Worked on Data frames and Spark SQL for efficient data querying and analysis.
* Involved in the implementation of the Hadoop cluster on AZURE as a part of POC.
* Developed intranet portal for managing Amazon EC2 servers using Tornado and MongoDB.
* Used Sqoop to migrate the data from MySQL tables into HDFS and Hive DB. Implemented importing all tables into Hive DB, incremental appends and last modified updates etc.
* Experienced in migrating HiveQL into Impala to minimize query response time.
* Developing and running Map-Reduce jobs on YARN and Hadoop clusters to produce daily and monthly reports as per user's need.
* Used Spark API over Horton works Hadoop YARN to perform analytics on data in Hive.
* Developed Spark scripts by using Scala Python commands as per the requirement.
* Developed Spark code and Spark-SQL/Streaming for faster testing and processing of data
* Extensive experience in working with HDFS, Pig, Hive, Sqoop, Flume, Oozie, MapReduce, Zookeeper, Kafka, Spark and HBase. Worked on Text mining project with Kafka.
* Integrated Apache Storm with Kafka to perform web analytics and to perform click stream data from Kafka to HDFS.
* Experience in running Hadoop streaming jobs to process terabytes of xml format data.
* Migrate mongo dB shared/replica cluster form one data center to another without downtime.
* Manage and Monitor large production MongoDB shared cluster environments having terabytes of the data.
* Worked on Importing and exporting data from RDBMS into HDFS with Hive and PIG using Sqoop.
* Highly skilled and experienced in Agile Development process for diverse requirements.
* Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark using Scala, Python.
* Setting up MongoDB Profiling to get slow queries.
* Configuring HIVE and Oozie to store metadata in Microsoft SQL Server.
* Created Sqoop scripts to import/export user profile data from RDBMS (DB2) to Azure Data lake.
* Expertise in deployment of Hadoop Yarn, Spark, and Storm integration with Cassandra, ignite and Kafka etc.
* Designed and developed SparkRDDs, Spark SQLs.
* Experience on implementation of a log producer in Scala that watches for application logs, transform incremental log and sends them to a Kafka and Zookeeper based log collection platform.
* Sqoop jobs, PIG and Hive scripts were created for data ingestion from relational databases to compare with historical data.
* Created HBase tables to load large sets of structured, semi-structured and unstructured data coming from UNIX, NoSQL and a variety of portfolios.

**Environment:** Hadoop, Map Reduce, HDFS, Pig, Hive, Sqoop, Flume, Oozie, Java, Linux, Teradata, Zookeeper, Kafka, Impala, Akka, Apache Spark, Spark Streaming Horton Works, HBase, MongoDB.

|  |
| --- |
| **Hadoop Developer**  **Phillips 66, Houston, Texas**  April 2014 – Dec 2016 |

* Worked on extracting and enriching HBase data between multiple tables using joins in spark.
* Worked on writing APIs to load the processed data to HBase tables.
* Replaced the existing MapReduce programs into Spark application using Scala.
* Built on premise data pipelines using Kafka and Spark streaming using the feed from API streaming Gateway REST service.
* Experienced in writing Sqoop scripts to import data into Hive/HDFS from RDBMS.
* Developed intranet portal for managing Amazon EC2 servers using Tornado and MongoDB.
* Building SSIS packages to create ETL process and load data into SQL Server database for some of the SSRS Reporting requirements.
* Created new database objects like procedures, functions, packages, triggers, indexes, and views using T-SQL in development and production environment for SQL server 2008/2012.
* Developed Hive Queries to analyze the data in HDFS to identify issues and behavioral patterns.
* Extensively used Spark stack to develop preprocessing job which includes RDD, Datasets and Data frame Api's to transform the data for upstream consumption.
* Involved in writing optimized Pig Scripts along with developing and testing Pig Latin Scripts.
* Able to use Python Pandas, NumPy modules for Data analysis, Data scraping and parsing.
* Deployed applications using Jenkins framework integrating Git- version control with it.
* Extracted files from NoSQL database like HBase through Sqoop and placed in HDFS for processing.
* Installed Oozie workflow engine to run multiple Hive and Pig jobs.
* Working with data delivery teams to setup new Hadoop users, Linux users, setting up Kerberos principles and testing HDFS, Hive.
* Installed Hadoop eco system components like Pig, Hive, Hbase and Sqoop in a Cluster.
* Participated in production support on a regular basis to support the Analytics platform.
* Used Rally for task/bug tracking.
* Used GIT for version control.
* Good knowledge on Kafka streams API for data transformation.
* Implemented logging framework - ELK stack (Elastic Search, Logstash& Kibana) on AWS.
* Setup Spark EMR to process huge data which is stored in AWS S3.
* Developed Oozie workflow for scheduling & orchestrating the ETL process.
* Used Talend tool to create workflows for processing data from multiple source systems.
* Created sample flows in Talend, Stream sets with custom coded jars and analyzed the performance of Stream sets and Kafka steams.

**Environment:** MapR, Hadoop, HBase, HDFS, AWS, PIG, Hive, Drill, SparkSQL, MapReduce, Spark streaming, Kafka, Flume, Sqoop, Oozie, Jupyter Notebook, Docker, Kafka, Spark, Scala, HBase, Talend, Python Scripting, Java.

|  |
| --- |
| **Linux System Admin**  **Sears Holdings Corporation, Hoffman Estates, Illinois**  Aug 2013 – Mar 2014 |

* Primarily responsible for keeping the servers up and running as well as providing direct user support for any technical issues related to Linux systems.
* Actively monitoring systems health using monitoring tools and responding to those tickets through the ticketing platform.
* Setting up secure passwordless SSH authentication on servers using SSH key pair.
* Provided support with data migration using tools like tar and gzip followed by SCP for migration.
* Dynamically modify kernel Parameters as requested by clients. Setup cron jobs schedules for various backup and monitoring tasks.
* Tuning and hardening Linux based OS's with enhanced security layers of firewalls.
* Used Bash scripting for day-to-day automation task. Worked in a data center for the Racking and stacking of servers.
* Performed regular installation of patches using RPM and YUM.
* Managing users including creating accounts, controlling password, deleting users, adding users in groups, and assigning permissions and privileges.
* Worked with daily system monitoring, verified the integrity and availability of all hardware and server resources, and reviewed system and application logs.
* Worked with "directory naming" technologies (Active Directory (AD), LDAP etc)
* Hands-on experience with incident, change, problem management and expert in leading troubleshooting efforts and performing root cause analysis (RCA)
* Expert in network administration - Linux routing, network interface configuration, and troubleshooting.
* Configured NIC-bonding on new builds for fault-tolerance, load-balance, and redundancy.
* Managed LVM to create volumes on the volume groups, and file systems, extended logical volumes and file systems as and when needed. Managed file systems like EXT3, EXT4, XFS.

**Environment:**RHEL, AD, SSH, SQL Server, Oracle, OOAD and UML, Windows, Server Builds, HP, DELL, Brocade, Cisco UCS.