|  |
| --- |
| **Arshad Ayub Syed** |
| **Hadoop Developer****Pheonix, AZ**937-900-1105  [**mailarshadsyed**@gmail.com] |

professional Summary

* About 8+ years of professional IT experience which includes about 7 years of experience in Big Data ecosystem related technologies like Hadoop HDFS, Map Reduce, cloudera, Apache Spark, Apache Pig, Hive, Sqoop, HBase, Flume, Oozie.
* Highly skilled and experienced in installing, configuring, and using Big Data core components and **Hadoop Ecosystem**, **Spark**, **HDFS**, MapReduce, YARN, **Hive**, HBase, Pig, **Sqoop**, and **Flume**.
* Experience in working with **MapReduce** programs using Apache Hadoop for working with **Big Data**.
* Developed Apache Spark jobs using **PySpark** and **SparkSQL** in test environment for faster data processing and used **Spark SQL** for querying.
* Strong experience working with conceptual, logical, and physical **data modeling** considering Metadata standards.
* Experience working with **Agile** and **Waterfall** data modeling methodologies.
* Experience in **NoSQL** Column-Oriented Databases like HBase and its Integration with Hadoop cluster.
* Working Knowledge in **Amazon** **AWS, GCP** cloud which includes services like EC2, S3, EBS, ELB, AMI, IAM, Route53, Autoscaling, CloudFront, CloudWatch, Security Groups, and GCS storage, Dataproc, BigQuery.
* Good knowledge and Hands on Experienced of Databases with **SQL Server**, **Oracle, Teradata,** and **DB2** databases.
* In-depth understanding/knowledge of **Hadoop Architecture** and various components such as HDFS, JobTracker, TaskTracker, NameNode, DataNodes, and MapReduce concepts.
* Strong knowledge in working with **UNIX/LINUX** environments, writing shell scripts and SQL Stored Procedures.
* Knowledge of **Airflow** and **Snowflake** big data tools, for scheduler for ETL tasks and a Data warehouse tool.
* Worked with **Oozie** workflow engine to schedule time-based jobs to perform multiple actions.
* Experienced in importing and exporting data from **RDBMS** into **HDFS** using Sqoop.
* Experienced in Development with **Python and Scala programming** languages.
* Knowledge of **continuous integration** software (**Jenkins**) and **version control** systems (**GIT**).
* Experienced with **Linux-**based commands of **BASH** and **Shell/Perl Scripting**.
* Developed and executed data processing workflows using **Databricks** notebooks, optimizing data pipelines for performance and reliability.
* Implemented **workflow** **automation** and **orchestration** using Databricks jobs and clusters, streamlining data processing tasks and job **scheduling**.
* Excellent logical, analytical, and debugging skills.
* Ability to learn, R&D, and adapt new technologies, applications, and products, understanding of business workflow, and delivered outputs in a timely manner.

TECHNICAL SKILLS

|  |  |
| --- | --- |
| **Hadoop Ecosystem** | HDFS, Hive, MapReduce, Oozie, Pig, Sqoop, Flume, Spark, SparkSQL, Kafka |
| **SQL Databases** | MySQL, MS SQL Server |
| **NoSQL Databases** | HBase, Cassandra |
| **Operating Systems** | Linux, Windows |
| **Languages** | Python, Scala, SQL, HiveQL, PySpark, SparkSQL Unix Shellscript, Java, Pig |
| **Version control** | Git, GitHub, SVN |
| **Hadoop Distributions** | MapR, Cloudera, Hortonworks |
| **AWS, GCP cloud** | EC2, S3, Amazon EMR, GCP, GCS Storage, Dataproc, BigQuery |
| **Certifications** | AWS Certified Cloud Practitioner, [PCEP-30-02] PCEP – Certified Entry-Level Python Programmer |

professional Experience

**Petsmart, Phoenix, AZ Sept 2023 – Present**

**Big Data Engineer**

**Roles & Responsibilities:**

* Handled data cleaning, transformation, and feature engineering tasks efficiently using PySpark DataFrame operations and SQL queries.
* Utilized Google Cloud Storage (**GCS**) for efficient storage and management of large-scale datasets, ensuring data integrity and accessibility.
* Leveraged the distributed computing capabilities of **Apache** **Spark** to handle large volumes of data efficiently.
* Implemented complex data **workflows** and **pipelines** using Apache Airflow, and Databricks workflows enabling seamless orchestration and automation of data processing tasks.
* Utilized Databricks **Notebooks** for collaborative development, data visualization, and iterative experimentation.
* Leveraged the interactive environment of Databricks **Notebooks** to prototype, test, and refine data processing and analytics workflows.
* Wrote optimized SQL queries to perform data analysis, aggregation, and filtering operations efficiently.
* Utilized SQL capabilities within PySpark and **Databricks** environments to interact with structured data seamlessly.
* Demonstrated expertise in understanding database architecture and data models, developing complex **Hive/SQL query** code to efficiently access and manipulate data.
* Employed Spark and SparkSQL for data processing tasks, such as reading data from external sources, merging data, performing data enrichment, and loading data into target GCS buckets and **BigQuery** tables on GCP.
* Implement performance tuning techniques to optimize **PySpark** jobs and SQL queries for improved execution times and resource utilization.
* Applied industry best practices for data validation and cleansing on GCP, using custom **Spark**, **SparkSQL**, and SQL functions to maintain data consistency, accuracy, and adherence to established standards.
* Document the project's architecture, data pipelines, algorithms, and methodologies comprehensively.
* Facilitate knowledge sharing within the team by documenting best practices, code snippets, and insights gained during the project.

**Environment:**Linux, BigData, Hive, Pyspark, Spark, SparkSQL, Scala, Pyspark, GCP, Python, GCS buckets, BigQuery, Git, GitHub, Notebook, IntelliJ

**Walmart, Sunnyvale, CA Feb 2022 – Aug 2023**

**Hadoop Developer**

**Roles & Responsibilities:**

* Developed data integration and transformation code pipelines using **Spark** on a Linux environment, leveraging object-oriented programming techniques to create maintainable and scalable solutions.
* Utilized SQL Developer to query source/target tables on Hive, validating SQL queries and implementing Lookup overrides to ensure data consistency and accuracy.
* Demonstrated expertise in understanding database architecture and data models, developing complex **Hive/SQL query** code to efficiently access and manipulate data.
* Employed Spark and SparkSQL for data processing tasks, such as reading data from external sources, merging data, performing data enrichment, and loading data into target GCS buckets and **BigQuery** tables on GCP.
* Optimized data processing performance by employing advanced Spark and Spark features, such as caching, partitioning, and broadcasting, to reduce computational overhead and improve overall efficiency.
* Applied industry best practices for data validation and cleansing on GCP, using custom **Spark**, **SparkSQL**, and SQL functions to maintain data consistency, accuracy, and adherence to established standards.
* Optimized query performance by leveraging data partitioning and bucketing techniques in Hive and BigQuery, reducing data retrieval times and enhancing user experience in data-driven applications.
* Employed **Automic** automation software to streamline and orchestrate complex big data workflows, improving efficiency and reducing manual intervention by automating repetitive tasks, job scheduling, and error handling within the Hadoop ecosystem.
* Utilized Zeppelin and IntelliJ Notebook for iterative development, testing, and visualization of data processing workflows, enabling efficient collaboration and sharing of insights among team members.

**Environment:**Linux, Hadoop, Hive, Spark, SparkSQL, Scala, GCP, Python, GCS buckets, BigQuery, Automic, Maven, Git, GitHub, Zeppelin, Notebook, IntelliJ

**American Express, Phoenix, AZ Nov 2019 – Jan 2022**

**Hadoop Developer**

**Roles & Responsibilities:**

* Designed and implemented scalable, distributed systems leveraging Big Data technologies like **Hadoop**, **Hive**, **SparkSQL,** and scripting languages like **Python** and **Shell Script**.
* Developed and maintained Linux Shell Scripts for automating routine tasks, streamlining data processing workflows, and improving overall system efficiency in a Hadoop-based big data environment.
* Managed Hadoop clusters and **HDFS** storage, ensuring optimal configuration, data organization, and resource allocation to support high-performance data processing and analytics.
* Utilized **Hive** and **SparkSQL** to develop efficient and scalable data processing pipelines, creating custom queries and scripts to extract, transform, and load (ETL) data from various sources.
* Leveraged **PySpark** and Python programming skills to implement complex data transformations, aggregations, and analytical operations, optimizing **Spark** performance and code maintainability.
* Scheduled and monitored recurring data processing tasks using **Crontab**, ensuring timely and reliable execution of critical workflows and data updates.
* Integrated Kafka for real-time data ingestion and streaming, enabling faster data processing and analysis in a distributed and fault-tolerant architecture.
* Utilized Grafana for monitoring and visualizing system metrics and performance data, providing valuable insights to stakeholders and driving data-driven decision-making.
* Used Git and GitHub for version control, effectively managing code changes and updates within a team-based development environment.
* Performed exploratory **data analysis** to determine which questions can be answered effectively with a given dataset.
* Developed scripts and automate the process to create the **tables** and store the data in them and generate the reports on a daily basis.

**Environment:**Linux, ShellScript, MapR, Hadoop, HDFS, Hive, Spark, PySpark, SparkSQL, Python, Crontab, Kafka, Grafana, Git, GitHub.

**State of Ohio, Columbus, OH May ’17 – Oct ‘19**

**Big Data Developer/ Hadoop Developer**

**Roles & Responsibilities:**

* Involved in the configuration of Apache Hadoop clusters for application development and **Hadoop tools** like **Hive**, HBase, Zookeeper, and Sqoop.
* Involved in design, development, integration, deployment/production support & other technical aspects of the development and modification to the applications
* Created Informatica source-to-target mapping using different transformations to implement business rules to fulfill the data integration requirements.
* Importing and exporting data into HDFS and **Hive** using **Sqoop**.
* Experienced in defining job flows. Experienced in managing and reviewing Hadoop log files.
* Load and transform large sets of structured, semi-structured, and unstructured data.
* Assisted in managing and optimizing HBase for real-time read and write access to large datasets in a distributed and fault-tolerant NoSQL database.
* Installed and configured Hive and wrote Hive UDFs.
* Involved in creating Hive tables, loading with data, and writing Hive queries which will run internally in map reduce way.
* Developing Web Services with **Python** programming language.
* Involved in building and maintaining test plans, test suites, test cases, defects and test scripts
* Conducted functional, system, data, and regression testing.
* Integrated Jenkins for continuous integration and deployment, automating the build, test, and deployment processes to ensure seamless delivery of high-quality big data solutions.
* Involved in Bug Review meetings and participated in Biweekly meetings with the management team.

**Environment:** Hadoop, HDFS, Cloudera, Pig, Hive,Python, HBase Spark, SparkSQL, Oozie, Sqoop, Flume, Linux, Java, Maven, Junit, Git, GitHub, Jenkins.

#### **The Hartford, Hartford, CT Jul’15– April ’17**

#### **Hadoop developer**

**Roles & Responsibilities:**

* Involved in the configuration of Apache Hadoop clusters for application development and Hadoop tools like **Hive**, HBase, **Zookeeper,** and **Sqoop**.
* Involved in **shell scripts** to monitor the health check of Hadoop daemon services and respond accordingly to any warning or failure conditions.
* Involved in collecting and aggregating large amounts of log data using Apache Flume and staging data in HBASE/HDFS for further analysis.
* Collected the logs data from web servers and integrated in to **HBASE** using Flume.
* Used Pig & **Python scripting** for pre-processing the data.
* Used Sqoop to import and export data from HDFS to RDBMS and vice-versa.
* Worked on importing and exporting data from Oracle and DB2 into HDFS and **HIVE** using Sqoop.
* Used Hive in partitioning, bucketing and perform different types of joins on Hive tables and implementing Hive serDes like JSON, Avro and REGEX.
* Developed Java-based MapReduce programs and integrated them with Maven and JUnit for efficient dependency management, build automation, and testing.
* Developed Sqoop scripts to import data from relational sources and handled incremental loading.
* Developed scalable, Hadoop-based data processing algorithms using HBase, Pig, Hive, Map Reduce and the Hadoop ecosystem.
* Deployed Hadoop Cluster in Pseudo-distributed and Fully Distributed modes.
* Supported in setting up QA environment and updating configurations for implementing scripts with Pig, Hive and Sqoop.
* Involved in writing scripts to automate the process and generate reports.

**Environment:** Hadoop, MapReduce, Spark, Java, Hive, HDFS, PIG, Sqoop, Kafka, Oozie, Cloudera, Flume, HBase, Zookeeper, CDH4&CDH5, Oracle, PL/SQL, Linux

Education

Bachelors in Computer Science and Engineering,VSU, 2014

Masters in Computer Science, WSU, OH, 2016

CERTIFICATIONS

